Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

SVM Multiclasses, Théorie et Applications

MÉMOIRE

présenté et soutenu publiquement le 28 novembre 2007

pour l'obtention de l'

Habilitation à Diriger des Recherches de l'Université Nancy I

(Spécialité Informatique)

par

Yann Guermeur

Composition du jury

Président :	Monsieur Karl TOMBRE
Rapporteurs :	Monsieur François DENIS Monsieur Stéphane BOUCHERON Monsieur Olivier GASCUEL
Examinateurs :	Madame Michèle SEBAG Monsieur Jean-Paul HATON Monsieur Patrick GALLINARI

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thloria.

Remerciements

Je tiens tout d'abord à remercier Monsieur Denis pour avoir accepté d'être le premier rapporteur de ce travail. J'ai beaucoup apprécié les échanges scientifiques que nous avons eus depuis le lancement de l'action spécifique "Apprentissage et Bioinformatique". J'espère en avoir tiré une leçon sur la rigueur avec laquelle se conduit un travail de recherche.

Monsieur Boucheron est l'auteur de certains des articles qui ont le plus influencé mes travaux sur le calcul de risques garantis. Plus encore, l'article de synthèse qu'il a co-écrit sur la théorie de la discrimination m'apparaît comme une source d'inspiration inépuisable pour mes recherches futures. Je suis donc particulièrement heureux du fait qu'il ait rapporté sur ce manuscrit, en attirant à cette occasion mon attention sur de nouvelles références.

En acceptant d'être rapporteur, Monsieur Gascuel m'a accordé une nouvelle marque de confiance qui m'honore. Je n'oublie pas toute l'aide qu'il m'a apportée au cours des années, me permettant ainsi de franchir des étapes importantes dans ma formation de chercheur. J'ai plaisir à rappeler que la lecture de ses articles a guidé mes premiers pas à la fois en prédiction de la structure secondaire des protéines et en théorie des bornes.

Ces dernières années, j'ai beaucoup profité des conseils de Madame Sebag. La manière très didactique avec laquelle elle transmet la vision globale qu'elle possède de la recherche en apprentissage automatique m'a beaucoup aidé à situer mes travaux par rapport à l'état de l'art tout en dégageant des perspectives nouvelles. C'est donc une joie pour moi d'avoir pu la compter parmi les membres du jury.

Au cours de la période de rédaction de ce manuscrit, Monsieur Haton m'a apporté avec constance toute l'aide que l'on peut espérer d'un parrain scientifique. Je lui suis extrêmement reconnaissant de m'avoir ainsi permis d'atteindre cette nouvelle étape dans des conditions idéales.

Je souhaite exprimer toute ma gratitude à Monsieur Gallinari pour avoir accepté d'évaluer à nouveau, après un intervalle de dix ans, les travaux de son ancien élève. Je suis très sensible à l'amitié qu'il m'a marquée en me permettant de conserver des liens forts avec son équipe. Je souhaite vivement que ceux-ci continuent à se développer dans l'avenir.

Monsieur Tombre m'a fortement encouragé à prépaper cette HDR. Je lui en suis très reconnaissant, de même que d'avoir accepté de présider le jury.

L'HDR sanctionne en particulier la capacité à collaborer avec de jeunes chercheurs. Je mesure la chance que j'ai eue de rencontrer Olivier Teytaud et Régis Vert à l'aube de leur brillant parcours scientifique. Encadrer Emmanuel Didiot, Sumit Kumar Jha et Julien Vannesson s'est également révélé une expérience très enrichissante. Je leur souhaite de suivre la voie tracée par leurs aînés. Plusieurs des principaux résultats que j'ai obtenus en sélection de modèle sont le fruit d'une collaboration avec deux chercheurs post-doctoraux, Frédéric Sur et Emmanuel Monfrini. C'est un plaisir pour moi de saisir cette occasion de leur exprimer ma gratitude. Le jeune collègue avec lequel j'ai le plus collaboré, sans l'avoir encadré, est André Elisseeff. Je lui suis très reconnaissant pour tout ce qu'il m'a permis d'apprendre, en commençant par les subtilités du théorème de Glivenko-Cantelli. Tous ces chercheurs d'avenir allient à leurs qualités scientifiques de grandes qualités humaines, qui méritent d'être soulignées. C'est tout simplement un grand plaisir de travailler avec eux, et je souhaite vivement continuer longtemps à le faire.

Je n'oublie pas non plus ce que je dois à ceux avec lesquels j'ai eu la chance de collaborer au cours de ces dix dernières années. Je pense en particulier aux chercheurs de l'IBCP dirigés par Monsieur Deléage, ainsi qu'aux membres de l'équipe MODBIO-ABC.

Nombreux sont les collègues qui ont relu ce mémoire, ou des manuscrits dont les résultats sont rapportés ici. Je tiens à leur faire part de ma profonde reconnaissance, en distinguant tout spécialement Fabienne Thomarat, Bernard Maigret, Liva Ralaivola et Alain Lifchitz.

Je voudrais pour conclure exprimer toute ma gratitude à ceux sans lesquels je ne serais pas chercheur aujourd'hui, qu'ils m'aient apporté leur aide au cours des années, ou de manière plus ponctuelle. Parmi ceux que je n'ai pas encore eu l'occasion de citer nommément, je pense en particulier à mes parents, à Madame Paugam-Moisy, Monsieur Cosnard, Monsieur Alexandre et Monsieur Bockmayr.

On se lasse de tout sauf d'apprendre.
Virgile

Table des matières

Chapitre 1

Notice de titres et travaux

1.1 Curriculum vitæ

1.1.1 Etat civil et coordonnées

Nom : Guermeur
Prénoms : Yann, Charles, Louis, Marie
Date et lieu de naissance : 19 mai 1967, à Neuilly-sur-Seine, Hauts-de-Seine (92)
Nationalité : Française
Situation de famille : célibataire

Adresse personnelle

4 bis rue A. Bontemps, 78000 Versailles, France

Coordonnées professionnelles

LORIA, Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy cedex, France
Téléphone : 03 83 59 30 18 depuis la France, 00 33 3 83 59 30 18 depuis l'étranger
Adresse électronique : `Yann.Guermeur@loria.fr`
Page web : `http://www.loria.fr/~guermeur/`

Fonctions et établissement actuels

Chargé de recherche (CR1) au CNRS, affecté au LORIA - UMR 7503

Responsable scientifique de l'équipe "Apprentissage et Biologie Computationnelle" (ABC) du LORIA

1.1.2 Titres universitaires

- 1997
Doctorat de l'Université Paris 6, spécialité Informatique, mention très honorable avec félicitations
Directeur de thèse : Patrick Gallinari
Titre : Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines
Date de soutenance : 10 décembre 1997
Lieu de soutenance : Université Paris 6
Composition du jury : Monsieur Lebbe, Président, Madame Paugam-Moisy, Rapporteur, Monsieur Deléage, Rapporteur, Monsieur Gallinari, Directeur de thèse, Madame d'Alché-Buc, Examinateur, Monsieur Nadal, Examinateur

- 1993
DEA "Intelligence Artificielle, Reconnaissance des Formes et Applications" (IARFA), de l'Institut Blaise Pascal (IBP), Université Paris 6, mention bien

- 1991
Diplôme d'ingénieur de l'Institut d'Informatique d'Entreprise (IIE), école du concours commun Centrale-Supélec (l'IIE est devenu en 2006 l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE))

- 1991
"Certificate of Proficiency in English" de l'Université de Cambridge

- 1985-1988
Classes préparatoires au lycée Chaptal à Paris

- 1985
Baccalauréat série C, mention assez bien

1.1.3 Stages, expérience professionnelle

- D'octobre 2003 à ce jour

Chargé de recherche de première classe au CNRS, affecté au LORIA. Membre de l'équipe MODBIO jusqu'au 30 juin 2006, actuellement responsable scientifique de l'équipe ABC.

- De février 1999 à septembre 2003

Maître de conférences à l'Université Henri Poincaré-Nancy 1 (UHP), affecté au site de Saint-Dié de l'IUT de Nancy Brabois, devenu l'IUT de Saint-Dié le premier janvier 2000. La ville de Saint-Dié se nomme à présent Saint-Dié-des-Vosges. Membre de l'équipe CORTEX du LORIA, dirigée par Monsieur Alexandre, jusqu'au 31 décembre 2000. Membre de l'équipe MODBIO, dirigée par Monsieur Bockmayr, après cette date.

- De septembre 1998 à janvier 1999

Stage post-doctoral effectué dans l'équipe connexionniste dirigée par Monsieur Gallinari, au sein du thème Apprentissage et Acquisition des Connaissances - APprentissage Automatique (APA) du Laboratoire d'Informatique de Paris 6 (LIP6), à l'Université Paris 6. Durant ce stage, j'ai participé à la conception d'un système de modélisation statistique des utilisateurs de l'Internet.

- De septembre 1997 à août 1998

ATER à l'Ecole Normale Supérieure de Lyon, membre de l'équipe connexionniste du Laboratoire de l'Informatique du Parallélisme (LIP) dirigée par Madame Paugam-Moisy.

- De septembre 1996 à août 1997

ATER à l'UFR d'Informatique de l'Université Paris 7.

- D'octobre 1994 à août 1996

Moniteur à l'UFR d'Informatique de l'Université Paris 7.

- De mars à mai 1994

Vacataire à l'Institut de Statistique de l'Université Pierre et Marie Curie (ISUP).

- De novembre 1993 à août 1996

Doctorant, allocataire MRE, sous la direction de Monsieur Gallinari, dans l'équipe connexionniste du LAFORIA, puis du LIP6, à l'Université Paris 6.

- D'avril à septembre 1993

Stage de DEA effectué dans l'équipe connexionniste du LAFORIA, à l'Université Paris 6, sous la direction de Monsieur Gallinari. Sujet : "Identification du locuteur, prise en compte de la dynamique par des systèmes récurrents".

- De septembre 1991 à août 1992

Société CEGI, à Paris. Responsable d'une application de gestion destinée aux associations "Profession Sport" dépendant du ministère de la Jeunesse et des Sports.

- De janvier à juin 1991

Institut Polytechnique de Leeds (Grande-Bretagne). Stage de recherche effectué sous la direction de Monsieur Kennedy, dans le cadre du programme ERASMUS. Rédaction du mémoire de fin d'études de l'IIE intitulé "Formation de concepts dans les réseaux neuronaux".

- De juin à octobre 1990

Société EDGETEK, Département Ingénierie, aux Ulis. Stage portant sur la conception et le développement d'un logiciel de test de composants électroniques.

- De juillet à septembre 1989

AEROSPATIALE, Division Engins Tactiques, à Châtillon. Stage consacré à l'analyse et à la réalisation d'un logiciel de gestion des affaires traitées par la division.

1.2 Enseignement

1.2.1 Tableau synthétique des enseignements effectués

Le tableau ci-dessous résume les enseignements que j'ai effectués au premier janvier 2007. L'ensemble représente 1440 heures et 10 minutes en "équivalent TD".

Etablissement	Enseignement	Activité	Etudiants	Vol. horaire
UHP	Bioinformatique	Cours	M2P Génomique et Info.	12
INPL	Bioinformatique	Cours	troisième année	6
UHP	Bioinformatique	Cours	DESS RGTI	20
UHP	Système (UNIX)	Cours	DESS RGTI	6
UHP	Bioinformatique	Cours	Maîtrise biol. (MBCP)	29
UHP	Bioinformatique	Cours	DEA Info.	6
UHP	Système (UNIX)	Cours	DUT SRC 1	16
UHP	Système (UNIX)	Cours	DUT SRC 2	32
UHP	Stat. et probabilités	Cours	DUT SRC 2	92
UHP	Programmation en C	Cours	DUT SRC 2	2
UHP	Programmation en C	Cours	DUT SRC 1	4
UHP	Système (UNIX)	Cours	DUT SRC 1	34
UHP	Programmation en JAVA	Cours	DUT SRC 1	34
ENS	Connexionnisme	Cours	Magistère Info. et Math. 2	16
ENS	Connexionnisme	Cours	Maîtrise Sciences Co.	12
UHP	Bioinformatique	TD	M2P Génomique et Info.	8
UHP	Bioinformatique	TD	Maîtrise biol. (MBCP)	4
UHP	Bases de données (DBASE)	TD	DEUG SV2	60
UHP	Programmation en CAML	TD	DEUG MIAS1	28
UHP	Système (UNIX)	TD	DUT SRC 2	60
UHP	Stat. et probabilités	TD	DUT SRC 2	198
UHP	Programmation en C	TD	DUT SRC 2	44
UHP	Programmation en C	TD	AETP	6
UHP	Programmation en C	TD	DUT SRC 1	24
UHP	Système (UNIX)	TD	DUT SRC 1	80
UHP	Programmation en JAVA	TD	DUT SRC 1	48
ENS	Connexionnisme	TD	Licence Sciences Co.	16
ENS	Fonctionnement (Système)	TD	Magistère Info. et Math. 1	24
ENS	Programmation en C	TD	Maîtrise Sciences Co.	14
P7	Réseaux	TD	Maîtrise d'Informatique	24
P6	Informatique et Programmation	TD	ISUP, première année	32
P7	Programmation en Pascal	TD	DEUG Sciences	200
UHP	Bioinformatique	TP	M2P Génomique et Info.	6
UHP	Programmation en C	TP	DUT SRC 1	40
UHP	Système (UNIX)	TP	DUT SRC 2	8
UHP	Programmation en HTML	TP	DUT SRC 1	32
UHP	Bioinformatique	TP	DESS RGTI	8
UHP	Système (UNIX)	TP	DESS RGTI	4
UHP	Bioinformatique	TP	Maîtrise biol. (MBCP)	35

1.2.2 Détail des enseignements effectués

En tant qu'enseignant-chercheur, j'ai enseigné les principales matières de base de l'informatique : algorithmique, programmation, système, réseaux, bases de données... Ces enseignements ont été dispensés à l'ensemble des publics universitaires (élèves de grandes Ecoles, étudiants de facultés et d'IUT) en premier, deuxième et troisième cycle. Il s'agissait essentiellement d'étudiants en informatique et en biologie. Dès

cette époque, j'ai également donné des cours en lien plus direct avec mon domaine de recherche : statistique et probabilités, apprentissage automatique et bioinformatique. Devenu chercheur, j'ai concentré mon activité d'enseignement sur la présentation des fondements et méthodes de l'apprentissage statistique, avec comme application privilégiée le traitement des données biologiques. Ainsi, depuis l'année universitaire 05-06, je suis responsable de l'UE 3.105 "Apprentissage statistique et fouille de données" de la spécialité "Génomique et Informatique" en deuxième année du parcours professionnel (M2P) du Master "Sciences de la Vie et de la Santé" (SVS) de l'Université Nancy 1. Le support du cours que je donne dans ce cadre est disponible à l'adresse suivante : `http://www.loria.fr/~guermeur/Cours_01.ps`.

1.3 Encadrement d'activités de recherche

1.3.1 Stages de DEA - Master 2 recherche

Durant l'année scolaire 97-98, Hélène Paugam-Moisy et moi avons co-encadré Olivier Teytaud, élève normalien effectuant son stage du DEA Informatique de Lyon (DIL). Olivier a rédigé un mémoire intitulé "Représentations internes dans les réseaux de neurones artificiels". Ce travail a produit des résultats originaux dans deux domaines : le calcul des dichotomies polyédriques par les PMC à unités à seuil et l'étude des capacités de généralisation des SVM. Olivier est actuellement chargé de recherche à l'INRIA.

Durant l'année scolaire 01-02, j'ai encadré Régis Vert, élève de l'Ecole Nationale Supérieure des Mines de Nancy (ENSMN) effectuant son stage du DEA Informatique de Lorraine. Le sujet du stage était la conception et la mise en œuvre de M-SVM dédiées au traitement de séquences biologiques. Le principal résultat de cette étude a été une extension du principe d'alignement de noyaux au cas multiclasse. Ceci a permis à Régis de proposer un nouveau noyau pour la prédiction de la structure secondaire des protéines (voir en particulier la référence [CL02] de ma liste de publications). Après avoir soutenu une thèse en théorie statistique de l'apprentissage en juin 2006, Régis est actuellement chercheur dans le privé.

Durant l'année scolaire 02-03, j'ai encadré le stage du DEA Informatique de Lorraine d'Emmanuel Didiot. Celui-ci a poursuivi les recherches de Régis Vert sur la mise au point d'un noyau dédié au traitement des séquences protéiques. La validation de ces travaux a de nouveau été effectuée en prédiction de la structure secondaire des protéines. Emmanuel achève actuellement une thèse en parole dans l'équipe Parole du LORIA. De mai à juillet 2003, j'ai également encadré Sumit Kumar Jha, étudiant en avant-dernière année de l'IIT de Kharagpur, en Inde. Nous avons travaillé sur l'application d'une M-SVM à la prédiction des ponts disulfures. Sumit est actuellement en thèse à l'Université Carnegie Mellon, où il a conservé comme domaine d'application de ses recherches la prédiction du repliement des protéines.

Durant l'année scolaire 05-06, Nadir T. Mrabet et moi avons co-encadré Levoly Fani, élève de l'Ecole Nationale Supérieure d'Electricité et de Mécanique (ENSEM), à l'occasion de son stage du Master Informatique, dans la spécialité "Services Distribués et Réseaux de Communication" (SDRC), proposée par l'UHP, l'Université Nancy 2 et l'Institut National Polytechnique de Lorraine (INPL). Celui-ci a rédigé un mémoire intitulé "Spiralix, un programme PERL pour définir les positions frontières d'hélices α : optimisation et représentation vectorielle des hélices α". Il travaille à présent en Afrique.

Cette année, j'ai encadré deux étudiants en stage de deuxième année de Master recherche. Aurélie Colas, élève de l'ENSMN, suivait les enseignements du Master Chimie et Physico-chimie Moléculaires (CPM) dans la spécialité "Chimie Informatique et Théorique" (CIT), de l'UHP et de l'INPL. L'intitulé de son travail de stage était : "Mise en œuvre d'une solution efficace au problème de programmation quadratique de grande taille correspondant à l'apprentissage des SVM multiclasses". Julien Vannesson, étudiant du Master Informatique, dans la spécialité "Perception, Raisonnement, Interaction Multimodale" (PRIM), proposée par l'UHP, l'Université Nancy 2 et l'INPL, a travaillé à l'intégration des modules discriminants et génératifs du modèle hybride de prédiction de la structure secondaire des protéines développé dans l'équipe ABC. Il poursuit actuellement ces recherches comme ingénieur.

1.3.2 Thèse

De septembre 2003 à septembre 2006, Alexander Bockmayr et moi avons co-encadré le travail de thèse de Yannick Darcy intitulé "Conception, mise en œuvre et évaluation de machines à noyau dédiées au traitement de séquences biologiques". Le financement de cette thèse de l'UHP par une bourse ministérielle a été obtenu dans le cadre du projet GENOTO3D décrit dans la section 1.4.1. Des raisons personnelles ont conduit Yannick à décider d'interrompre ses recherches pour partir à l'étranger. Celles-ci ont déjà fait l'objet de publications et sont poursuivies par d'autres membres de l'équipe ABC.

1.3.3 Recherches post-doctorales

D'octobre 2004 à août 2005, j'ai encadré le stage post-doctoral du STIC-CNRS de Frédéric Sur, traitant du sujet suivant : "Modèles statistiques hybrides pour la recherche de motifs dans les séquences biologiques". Frédéric Sur occupe à présent un poste de maître de conférences à l'INPL (plus précisément à l'ENSMN).

D'octobre 2005 à mars 2007, j'ai supervisé le stage post-doctoral qu'a effectué Emmanuel Monfrini dans le cadre du projet "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques" financé par le programme Décrypthon (voir la section 1.4.1). Ce stage portait sur la conception et le développement d'un noyau dédié à l'identification des différentes catégories intervenant dans les phénomènes d'épissage alternatif. Pendant les six premiers mois du projet, les bases de données utilisées par Emmanuel ont été assemblées par un ingénieur d'études, Delphine Autard. Celle-ci occupe à présent un poste d'ingénieur d'études permanent à l'INSERM.

1.4 Administration de la recherche et responsabilités collectives

1.4.1 Actions nationales et internationales

De 1999 à 2002, j'ai participé au groupe de travail ESPRIT "Neural and Computational Learning Theory" (NeuroCOLT2) : `http://www.neurocolt.com/`. Ce groupe de travail se poursuit à travers le réseau d'excellence européen "Pattern Analysis, Statistical Modelling and Computational Learning" (PASCAL) : `http://www.pascal-network.org/`. J'ai été membre du comité d'organisation du challenge théorique PASCAL "Type I and type II errors for multiple simultaneous hypothesis testing" (`http://www.lri.fr/~teytaud/risq/risq.html`) proposé par Olivier Teytaud en 2006. Ce challenge s'est achevé par un workshop organisé à Paris en mai 2007.

En 2003, j'ai été membre du comité de pilotage de l'Action Spécifique (AS) du CNRS "Apprentissage et bioinformatique". J'ai animé dans ce cadre un groupe de travail portant sur l'apprentissage et le traitement de séquences : `http://www.loria.fr/~guermeur/GdT/`. J'ai également été membre de l'AS du CNRS intitulée "Machines à vecteurs support et méthodes à noyau".

Je suis membre de la "Fédération des Equipes de Recherche en Apprentissage" (FERA) : `http://www.lri.fr/~proml/wiki/index.php/Main_Page`.

De mars 2003 à novembre 2006, j'ai animé le projet GENOTO3D financé pour trois ans par l'Action Concertée Incitative (ACI) "Masses de Données". Ce projet, dont l'objectif était de prédire la structure tertiaire des protéines par des méthodes issues de l'apprentissage automatique, regroupait six équipes de six laboratoires : le LORIA, l'IBCP, le LIF, l'IRISA, le LIRMM et l'unité MIG de l'INRA. Son site web se trouve à l'adresse suivante : `http://www.loria.fr/~guermeur/ACIMD/`.

Depuis septembre 2003, je participe au projet intitulé "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques". Ce projet, une collaboration avec le laboratoire "Maturation des ARN et Enzymologie Molé-

culaire" (MAEM), UMR 7567 à Nancy, a été financé pendant 18 mois par le programme Décrypthon (`http://www.decrypthon.fr/`). De 2005 à 2006, il a également fait l'objet d'une opération du thème "Bioinformatique et applications à la génomique" du PRST "Intelligence Logicielle", opération dont j'étais co-responsable.

De 2005 à 2006, j'ai été membre du projet intitulé "Modélisation de la protéine FAK (Focal Adhesion Kinase) en vue de l'identification de molécules anti-métastases". Cette collaboration avec l'"équipe de Dynamique des Assemblages Membranaires" (eDAM) du laboratoire "Structure et Réactivité des Systèmes Moléculaires Complexes", UMR 7565 à Nancy, était une opération du thème "Bioinformatique et applications à la génomique" du PRST "Intelligence Logicielle". Elle bénéficie actuellement d'un financement de l'ANR, dans le cadre du projet "Tyrosines kinases de la famille de FAK : bases structurales de la régulation et de la localisation intracellulaire" (FAKs) retenu par l'ANR non-thématique pour les années 2006 à 2008.

1.4.2 Activités éditoriales

Je suis relecteur régulier pour de nombreuses revues, dont le "Journal of Machine Learning Research" (JMLR), Machine Learning, Neurocomputing, IEEE Transactions on Neural Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on Systems, Man and Cybernetics (Part B) (SMCB), Statistics and Computing, la revue RIA et Bioinformatics, ainsi que pour de nombreuses conférences dont IJCAI, ICANN, NIPS, COLT, WABI, ECA, JOBIM et IJCNN.

J'ai également relu des articles soumis à des revues portant sur des thèmes a priori plus éloignés des miens, comme l'"Internet Electronic Journal of Molecular Design" ou "Process Biochemistry".

En 2000, Hélène Paugam-Moisy, André Elisseeff et moi avons organisé une session spéciale de la conférence "International Joint Conference on Neural Networks" (IJCNN) intitulée "Multi-Class Support Vector Machines".

Depuis 2003, je suis membre du comité de programme de la "Conférence francophone d'Apprentissage" (CAp). J'ai été membre du comité de programme de la conférence "Reconnaissance des Formes et Intelligence Artificielle" (RFIA) en 2004.

En 2007, j'ai organisé la session spéciale de la conférence "Applied Stochastic Models and Data Analysis" (ASMDA 2007, `http://www.asmda.com/id7.html`) intitulée "Supervised Prediction with Neural Networks and SVMs" (`http://www.loria.fr/~guermeur/ASMDA_CFP.html`).

1.4.3 Responsabilités collectives

De septembre 1999 à septembre 2003, j'ai été co-directeur des études du département "Services et Réseaux de Communication" (SRC) de l'IUT de Saint-Dié-des-Vosges, qui est une composante de l'Université Nancy 1. J'ai appartenu aux différents jurys d'admission, de passage et de diplôme de cet IUT. De février 2000 à septembre 2003, j'ai également fait partie de la commission de choix. A ce titre, j'ai en particulier rapporté sur les dossiers de candidature aux postes de maîtres de conférences et d'ATER ouverts au recrutement au département. J'ai aussi participé aux travaux de la commission mixte constituée à partir de la commission de choix et de la commission de spécialistes des sections 7, 11, 12 et 71 de l'Université Nancy 1.

Depuis janvier 2006, je suis membre suppléant élu de la commission de spécialistes de la section 27 de l'Université Nancy 1.

En juin 2007, j'ai été nommé membre titulaire de la commission de spécialistes de la section 27 de l'Université Paris 13.

De 2004 à 2006, j'ai été membre du comité des opérations du thème "Bioinformatique et applications à la génomique" du PRST "Intelligence Logicielle". Je poursuis cette activité dans le cadre du thème "Modélisation des Biomolécules et de leurs Interactions" (MBI) du CPER "Modélisations, Informations et Systèmes Numériques" (MISN) qui s'étend de 2007 à 2013.

Je suis le représentant des équipes ABC et CARTE à la Commission "Information Scientifique et Technique" (IST), anciennement Commission Documentation (ComDoc), du LORIA.

1.4.4 Activités d'expertise

De 2003 à 2005, j'ai rédigé des rapports sur plusieurs projets soumis aux ACI "Masse de Données" et IMPBio. Depuis 2006, je suis expert pour l'ANR. Dans ce cadre, j'ai rédigé des rapports sur des projets soumis aux programmes "blanc", "Jeunes chercheuses - jeunes chercheurs" et "Masse de Données" (actuellement "Masse de Données et COnnaissances"). En 2007, j'ai également expertisé un projet soumis au programme "Plates-formes technologiques du vivant" (PFTV).

1.4.5 Jurys de thèses

Le 11 janvier 2006, j'ai participé en tant que membre invité au jury de thèse de Nicolas Sapay. Cette thèse de biologie de l'Université Lyon 1, dirigée par Gilbert Deléage et François Penin, est intitulée "Les peptides d'ancrages à l'interface membranaire, analyses structurales par RMN et dynamique moléculaire et développement d'une méthode de prédiction bioinformatique".

1.4.6 Animation d'équipes de recherche

J'anime actuellement l'activité scientifique de l'équipe ABC du LORIA, anciennement projet MODBIO de l'INRIA Lorraine. Le texte de notre nouveau projet de recherche est disponible à l'adresse suivante : `http://modbio.loria.fr/download/abc2006.pdf`. Il a été présenté à l'assemblée des responsables d'équipes (AREQ) du LORIA le 25 septembre 2006.

1.5 Publications

1.5.1 Chapitres de livres

[CL03] Y. Guermeur et O. Teytaud (2006). Estimation et contrôle des performances en généralisation des réseaux de neurones. Dans *Apprentissage connexionniste*, édité par Y. Bennani, Hermès, Chapitre 10, 279-342.

[CL02] Y. Guermeur, A. Lifchitz et R. Vert (2004). A kernel for protein secondary structure prediction. Dans *Kernel Methods in Computational Biology*, édité par B. Schölkopf, K. Tsuda et J.-P. Vert, The MIT Press, Chapitre 9, 193-206.

[CL01] Y. Guermeur et H. Paugam-Moisy (1999). Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. Dans *Apprentissage automatique*, édité par M. Sebban et G. Venturini, Hermès, Chapitre 5, 109-138.

1.5.2 Journaux internationaux

[JI07] Y. Guermeur (2007). VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research (JMLR)*, **Vol. 8**, 2551-2594.

[JI06] N. Sapay, Y. Guermeur et G. Deléage (2006). Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics*, **Vol. 7**, n° 255.

[JI05] Y. Guermeur, A. Elisseeff et D. Zelus (2005). A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers. *Applied Stochastic Models in Business and Industry (ASMBI)*, **Vol. 21**, n° 2, 199-214.

[JI04] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy et P. Baldi (2004). Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, **Vol. 56C**, 305-327.

[JI03] Y. Guermeur (2002). Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications (PAA)*, **Vol. 5**, n° 2, 168-179.

[JI02] Y. Guermeur, C. Geourjon, P. Gallinari, et G. Deléage (1999). Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, **Vol. 15**, n° 5, 413-421.

[JI01] O. Gascuel *et al.* (groupe SYMENU) (1998). Twelve numerical, symbolic and hybrid supervised classification methods. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, **Vol. 12**, n° 5, 517-571.

Soumis

[JI00] Y. Guermeur (2007). Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. Soumis à la revue *Communications in Statistics*.

1.5.3 Journaux nationaux

[JN01] Y. Guermeur et H. Paugam-Moisy (1999). Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. *Revue Electronique sur l'Apprentissage par les Données (READ)*, **Vol. 3**, n° 1, 17-38.

1.5.4 Conférences internationales avec comité de lecture et publication des actes

[CI09] Y. Guermeur (2007). Scale-sensitive Ψ-dimensions : the capacity measures for classifiers taking values in $\mathbb{R}^Q$. Actes de l'*International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'07)*.

[CI08] Y. Guermeur, M. Maumy et F. Sur (2005). Model selection for multi-class SVMs. *ASMDA'05*, 507-517.

[CI07] Y. Guermeur, A. Elisseeff et D. Zelus (2002). Bound on the risk for M-SVMs. Actes de la conférence *Statistical Learning, Theory and Applications (SLTA'02)*, 48-52.

[CI06] Y. Guermeur, A. Elisseeff et H. Paugam-Moisy (2000). A new multi-class SVM based on a uniform convergence result. Actes de l'*International Joint Conference on Neural Networks (IJCNN'00)*, IEEE, **Vol. IV**, 183-188.

[CI05] H. Paugam-Moisy, A. Elisseeff et Y. Guermeur (2000). Generalization performance of multi-class discriminant models. *IJCNN'00*, IEEE, **Vol. IV**, 177-182.

[CI04] Y. Guermeur, A. Elisseeff et H. Paugam-Moisy (1999). Estimating the sample complexity of a multi-class discriminant model. Actes de l'*International Conference on Artificial Neural Networks (ICANN'99)*, publié par IEE, 310-315.

[CI03] Y. Guermeur, H. Paugam-Moisy et P. Gallinari (1998). Multivariate linear regression on classifier outputs : a capacity study. *ICANN'98*, Perspectives in neural computing, Springer-Verlag, 693-698.

[CI02] Y. Guermeur, F. d'Alché-Buc et P. Gallinari (1997). Optimal linear regression on classifier outputs. *ICANN'97*, LNCS vol. 1327, Springer-Verlag, 481-486.

[CI01] Y. Guermeur et P. Gallinari (1996). Combining statistical models for protein secondary structure prediction. *ICANN'96*, LNCS vol. 1112, Springer-Verlag, 599-604.

1.5.5 Conférences internationales avec comité de lecture (posters)

[PI02] D. Eveillard et Y. Guermeur (2002). Statistical processing of SELEX results. *International Conference on Intelligent Systems for Molecular Biology (ISMB'02).*

[PI01] Y. Guermeur et D. Zelus (2000). Combining protein secondary structure prediction methods with a new multi-category SVM. *ISMB'00.*

1.5.6 Ateliers de travail internationaux avec comité de lecture

[AI02] Y. Guermeur, A. Lifchitz et R. Vert (2003). A hybrid kernel machine for protein secondary structure prediction. Communication invitée au workshop *Kernel Methods in Computational Biology* de la *Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB'03).*

[AI01] Y. Guermeur (1997). An ensemble method for protein secondary structure prediction. *Mathematical Analysis of Biological Sequences (MABS'97).*

1.5.7 Conférences nationales avec comité de lecture et publication des actes

[CN07] Y. Darcy, E. Monfrini et Y. Guermeur (2006). Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes. Actes des *XXXVIII-ièmes Journées de Statistique (JdS'06).*

[CN06] N. Sapay, Y. Guermeur et G. Deléage (2005). Prediction of in-plane amphipathic membrane segments based on an SVM method. Actes des *Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM'05)*, 299-311.

[CN05] D. Eveillard et Y. Guermeur (2002). Traitement statistique des résultats de SELEX. *JOBIM'02*, 277-283.

[CN04] Y. Guermeur et D. Zelus (2001). Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *JOBIM'01*, 97-104.

[CN03] A. Elisseeff, H. Paugam-Moisy et Y. Guermeur (1999). Risque garanti pour les modèles de discrimination multi-classes. Actes des *7-ièmes rencontres de la Société Francophone de Classification (SFC'99)*, 111-118.

[CN02] Y. Guermeur (1998). Combinaison de classifieurs estimant les probabilités *a posteriori* des classes. *SFC'98*, 121-124.

[CN01] Y. Guermeur, F. d'Alché-Buc et P. Gallinari (1997). Combinaison linéaire optimale de classifieurs. *JdS'97*, 425–428.

1.5.8 Rapports de recherche et rapports techniques

[RR08] Y. Darcy et Y. Guermeur (2005). Radius-margin bound on the leave-one-out error of multi-class SVMs. Rapport de recherche RR-5780 de l'INRIA.

[RR07] Y. Guermeur (2004). Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions. Rapport de recherche RR-5314 de l'INRIA (révisé en 2006).

[RR06] E. Gothié, Y. Guermeur, S. Muller, C. Branlant et A. Bockmayr (2003). Recherche des gènes d'ARN non codants. Rapport de recherche RR-5057 de l'INRIA.

[RR05] Y. Guermeur (2002). A simple unifying theory of multi-class support vector machines. Rapport de recherche RR-4669 de l'INRIA.

[RR04] Y. Guermeur, A. Elisseeff et D. Zelus (2002). Bounding the capacity measure of multi-class discriminant models. Rapport technique NC2-TR-2002-123 du groupe de travail ESPRIT NeuroCOLT2.

[RR03] Y. Guermeur (2000). Combining discriminant models with new multi-class SVMs. Rapport technique NC2-TR-2000-086 du groupe de travail ESPRIT NeuroCOLT2.

[RR02] A. Elisseeff, Y. Guermeur et H. Paugam-Moisy (1999). Margin error and generalization capabilities of multi-class discriminant systems. Rapport technique NC2-TR-1999-051 du groupe de travail ESPRIT NeuroCOLT2 (révisé en 2001).

[RR01] Y. Guermeur et H. Paugam-Moisy (1998). Linear ensemble methods for multiclass discrimination. Rapport de recherche 1998-52 du LIP, ENS Lyon.

1.5.9 Logiciels

[L03] Le logiciel "AmphipaSeeK" mettant en œuvre la méthode de prédiction des ancrages membranaires interfaciaux introduite dans l'article [JI06] (voir aussi [CN06]) est disponible en ligne depuis le serveur d'analyse de séquences protéiques NPS@ (`http://npsa-pbil.ibcp.fr/`) du Pôle Bio-Informatique Lyonnais (PBIL Lyon-Gerland, `http://pbil.univ-lyon1.fr/`), à l'adresse suivante : `http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_amphipaseek.html`.

La mise en ligne de cette application est due à Nicolas Sapay.

[L02] Deux versions du logiciel de la M-SVM nommée M-SVM1 dans la référence [JI03], l'une de base, déposée à l'APP sous le numéro IDDN `IDDN.FR.001.170014.000.R.P.2005.000.10000`, l'autre dédiée au traitement des séquences protéiques (voir la référence [CL02]), sont diffusées sous la licence GNU GPL. Elles sont accessibles à partir du site web des "kernel machines", à l'adresse suivante : `http://www.kernel-machines.org`.

[L01] Deux logiciels de prédiction de la structure secondaire des protéines globulaires, HNN et MLRC, décrits respectivement dans les références [CI01] et [JI02], sont disponibles en ligne sur le serveur NPS@ du PBIL, à l'adresse suivante : `http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html`.

Chapitre 2

Machines à vecteurs support multiclasses

2.1 Introduction

Ce chapitre présente l'essentiel de nos contributions à la théorie et à la mise en œuvre pratique des SVM multiclasses. Il ne porte que sur des résultats généraux, c'est-à-dire indépendants de toute application particulière. Le chapitre suivant viendra compléter l'exposé, en détaillant nos réalisations sur un problème ouvert de biologie structurale : la prédiction de la structure secondaire des protéines globulaires.

Les SVM multiclasses sont des modèles de l'apprentissage de conception relativement récente, dont l'étude est actuellement en plein essor. Ceci résulte en premier lieu du fait que la communauté des théoriciens, qui avait jusque dans un passé récent consacré l'essentiel de ses forces au développement de la théorie statistique du calcul des dichotomies, exprime à présent un intérêt de plus en plus marqué pour le cas multiclasse, dont elle perçoit mieux les spécificités. Cette situation nouvelle fait naître un besoin, celui de disposer d'une étude synthétique sur les SVM multiclasses, ou plus généralement l'utilisation de SVM pour la discrimination à catégories multiples. C'est la raison qui nous a conduits à donner pour cadre de la présentation de nos résultats dans le domaine un exposé général tentant de réaliser un état de l'art.

Dans [37, 55], Vapnik et ses co-auteurs ont introduit les machines à vecteurs support (SVM) comme des extensions non linéaires d'un séparateur linéaire : l'hyperplan de marge maximale [227]. Ainsi, à l'image du perceptron [182, 12], ces machines étaient initialement dédiées au calcul des dichotomies. Cependant, si les difficultés rencontrées lors du cheminement conduisant du modèle de Rosenblatt jusqu'au perceptron multicouche (PMC) actuel se sont revélées être de nature essentiellement technique (de ce point de vue, l'introduction de l'algorithme de rétro-propagation du gradient a constitué une étape essentielle), il n'en a pas été de même avec les SVM. De fait, si l'utilisation des PMC pour effectuer des tâches de discrimination à catégories multiples ou de régression multivariée ne pose pas de problème conceptuel particulier, les résultats standard de la théorie statistique de l'apprentissage [65, 229, 106, 39] ne peuvent être étendus de manière triviale afin de spécifier des SVM dédiées à la discrimination à catégories multiples (dans ce qui suit, nous nommerons ces modèles "SVM multiclasses"). En outre, si cette difficulté théorique n'a pas empêché l'apparition dans la littérature de plusieurs modèles de ce type, ceux-ci doivent systématiquement faire face à deux critiques : d'une part, ils sont d'un emploi malaisé, d'autre part, leurs performances ne les distinguent pas significativement des méthodes de décomposition impliquant des SVM biclasses. Ainsi, en dépit de près de dix ans de travail, la théorie et la pratique des SVM multiclasses demeurent encore aujourd'hui des sujets de recherche largement ouverts.

En rédigeant ce chapitre, nous poursuivons deux objectifs. Le premier, évoqué plus haut, consiste à fournir un exposé de nos travaux inscrit dans un état de l'art. Le second consiste à mettre en lumière, domaine par domaine, le fait que les trois théories dont relèvent le plus directement les SVM : théorie statistique des classifieurs à grande marge [229], des machines à noyau [197] et de la régularisation [217, 15], ainsi que les éléments de leur mise en œuvre pratique, s'étendent ou résistent à l'extension

au cas multiclasse. Il s'agit ainsi de mettre en évidence à la fois les propriétés intrinsèques du cas des catégories multiples (ou a contrario le caractère ad hoc de certaines solutions dérivées pour le cas biclasse) et les problèmes sur lesquels la communauté bute actuellement, ce qui permet en définitive de dégager des perspectives de recherche claires.

Chacun des cadres théoriques dont relèvent les SVM fournit un angle particulier pour les considérer. Cependant, en première approximation, il est possible de distinguer deux grandes approches pour leur étude. La première, de nature essentiellement géométrique, consiste à les présenter comme des classifieurs à vaste marge. La seconde, relevant plutôt de l'analyse fonctionnelle, plonge ces machines dans le cadre de la régularisation de Tikhonov. Elle permet de tirer le meilleur parti des propriétés du noyau. Naturellement, ces deux options sont duales, et mettent simplement en évidence le fait qu'un même objectif, le contrôle des performances en généralisation (dans le premier cas, il s'agit plus précisément de la mise en œuvre du principe inductif de minimisation structurelle du risque (SRM) [227]), poursuivi par des chemins différents, peut conduire à des solutions identiques. Dans ce chapitre, nous les adopterons alternativement, en fonction des besoins de l'exposé.

Le plan est le suivant. La section 2.2 est dédiée à l'introduction du cadre théorique dans lequel nous nous plaçons. La section 2.3 présente la première approche mise en œuvre pour effectuer des tâches de discrimination à catégories multiples au moyen de SVM, l'emploi de méthodes de décomposition. Nous considérons ensuite dans la section 2.4 les principaux modèles de SVM multiclasses, en mettant en évidence leurs spécificités. Le calcul de risques garantis fait l'objet de la section 2.5. Les sections suivantes abordent les questions pratiques que sont la résolution du problème de programmation mathématique correspondant à l'algorithme d'apprentissage (section 2.6) et la sélection de modèle (choix du noyau et des valeurs des hyperparamètres, section 2.7).

2.2 Cadre théorique et notations

Nous présentons ici le cadre théorique de la discrimination à catégories multiples ainsi que les familles de fonctions sur lesquelles s'appuient les principaux modèles de SVM multiclasses, familles de fonctions construites à partir d'un noyau symétrique semi-défini positif. Une introduction aux notions de base d'analyse fonctionnelle utilisées dans ce chapitre peut être trouvée dans [49, 45, 225].

2.2.1 Théorie statistique de la discrimination à catégories multiples

Nous nous plaçons dans le cadre de la théorie statistique de l'apprentissage de Vapnik [229]. Parmi les trois types de problèmes auxquels cette théorie s'applique : discrimination, régression et estimation de la (fonction de) densité, seul le premier est considéré ici. Nous nous intéressons plus spécifiquement à des problèmes de discrimination à Q catégories, avec $3 \leq Q < \infty$. Ceux-ci consistent à affecter des objets à leur catégorie. Un objet est représenté par sa description x appartenant à l'*espace de description* $\mathcal{X}$ et l'ensemble des catégories, $\mathcal{Y}$, peut être identifié à l'ensemble des indices des catégories, i.e. $\{1, \ldots, Q\}$. Les catégories sont supposées être indépendantes les unes des autres. Ainsi, l'ordre dans lequel elles se présentent (l'indexation) est arbitraire. Le lien entre objets et catégories est supposé être de nature probabiliste. Plus précisément, $\mathcal{X}$ et $\mathcal{Y}$ sont des espaces probabilisés, et l'espace produit $\mathcal{X} \times \mathcal{Y}$ est muni d'une mesure de probabilité P, fixe mais inconnue. La mesure P caractérise entièrement le problème traité. Dans le domaine de l'apprentissage probablement approximativement correct (PAC) [222], ce cadre standard est connu sous le nom d'*apprentissage de concepts probabilistes* [123]. L'affectation d'une description x à une catégorie y s'effectue au moyen d'une fonction g choisie dans une famille de fonctions $\mathcal{G}$ sur $\mathcal{X}$ qui est donnée.

Présentée de manière idéale, la tâche de sélection de fonction correspondant à l'apprentissage consiste à rechercher dans $\mathcal{G}$ une fonction dont la probabilité d'erreur est aussi proche que possible de celle de la règle de décision de Bayes [84], c'est-à-dire égale à cette probabilité augmentée de l'*erreur d'approximation* [42]. Ce problème statistique se reformule immédiatement comme un problème d'optimisation consistant à minimiser sur $\mathcal{G}$ un critère (une fonctionnelle) correspondant à l'espérance par rapport à la mesure P d'une *fonction de perte* donnée. Cette espérance est nommée le *risque*. Naturellement, P étant inconnue, ce problème ne peut pas être résolu directement. On a recours à une procédure approchée, induisant une

erreur supplémentaire nommée *erreur d'estimation* [42]. Cette procédure s'appuie sur l'utilisation d'un *échantillon d'apprentissage*. Soit (X, Y) un couple aléatoire à valeurs dans $\mathcal{X} \times \mathcal{Y}$ distribué suivant P. On suppose disposer d'un m-échantillon $D_m = ((X_i, Y_i))_{1 \leq i \leq m}$ constitué de copies de (X, Y) indépendantes (échantillon i.i.d.). Cet échantillon est utilisé pour construire un estimateur empirique du risque à la base d'un nouveau critère de sélection de fonction pouvant cette fois, en théorie du moins, être optimisé sur $\mathcal{G}$. Le problème de l'apprentissage se reformule ainsi comme un problème de *M-estimation* [223]. A ce point de l'exposé, les propriétés de la famille de fonctions $\mathcal{G}$ doivent être précisées. Ce chapitre considère principalement des familles de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ vérifiant certaines conditions de mesurabilité qui apparaîtront de manière implicite dans la suite (voir par exemple le chapitre 10 de [72] pour une étude détaillée de la question dans un contexte analogue), plus la contrainte $\sum_{k=1}^{Q} g_k = 0$, dont la signification apparaîtra elle aussi ultérieurement. Il ne s'agit donc pas directement de règles de décision. Dans ce cas, une fonction $g = (g_k)_{1 \leq k \leq Q}$ affecte $x \in \mathcal{X}$ à la catégorie d'indice l si et seulement si on a : $g_l(x) > \max_{k \neq l} g_k(x)$. En cas d'égalité, elle ne se prononce pas, ou affecte l'exemple à une catégorie fictive $*$, ce qui est compté comme une réponse erronée (i.e. contribue au risque). Ceci conduit tout naturellement à choisir pour fonction de perte la fonction indicatrice ℓ définie sur $\mathcal{Y} \times \mathcal{G}(\mathcal{X})$ par

$$\ell(y, g(x)) = \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}}.$$

Le risque de g est alors donné par la formule :

Définition 1 (Risque)

$$R(g) = \mathbb{E}[\ell(Y, g(X))] = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}} dP(x, y).$$

Il représente simplement la probabilité d'erreur correspondant à cette fonction. Précisément, en notant f la règle de décision associée à g (f peut être considérée comme une fonction de $\mathcal{X}$ dans $\{1, \ldots, Q, *\}$), on a :

$$R(g) = P(f(X) \neq Y).$$

Dans la suite, on nommera "ensemble d'apprentissage" une réalisation $((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times \mathcal{Y})^m$ du m-échantillon D_m. On notera m_k le nombre d'exemples de la catégorie k dans cet ensemble. La valeur optimale d'une quantité sera identifiée par l'ajout d'une étoile en exposant.

2.2.2 Du noyau à la machine à noyau multivariée

Soit κ un noyau symétrique semi-défini positif ou *noyau de Mercer* [159, 2] sur $\mathcal{X}$. Soit $\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right)$ l'espace de Hilbert à noyau reproduisant (RKHS) [13, 233, 234, 29] correspondant. L'existence et l'unicité de $\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right)$ sont assurées par le théorème de Moore-Aronszajn [13] (voir aussi le théorème 6.1 de [234] ou le théorème 3 de [29]). Φ est l'une quelconque des fonctions sur $\mathcal{X}$ satisfaisant :

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle, \tag{2.1}$$

où $\langle ., . \rangle$ désigne le produit scalaire de l'espace ℓ_2. Par abus de langage, on parlera de l'*espace de représentation* (*feature space*) pour évoquer l'un quelconque des espaces de Hilbert $\left(E_{\Phi(\mathcal{X})}, \langle ., . \rangle\right)$ engendrés par les $\Phi(\mathcal{X})$. Du fait même de la définition d'un RKHS, $\mathcal{H} = \left(\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right) + \{1\}\right)^Q$ est la famille des fonctions à valeurs vectorielles $h = (h_k)_{1 \leq k \leq Q}$ dont les fonctions composantes sont les combinaisons affines de taille finie de la forme :

$$h_k(.) = \sum_{i=1}^{l_k} \beta_{ik} \kappa(x_{ik}, .) + b_k,$$

où les x_{ik} sont des éléments de $\mathcal{X}$ (les β_{ik} et b_k sont des scalaires), ainsi que les limites de ces fonctions lorsque les ensembles $\{x_{ik} : 1 \leq i \leq l_k\}$ deviennent denses dans $\mathcal{X}$ au sens de la norme induite par le produit scalaire. Il résulte de l'équation 2.1 que la famille $\mathcal{H}$ peut également être considérée comme un modèle affine multivarié sur $\Phi(\mathcal{X})$. Les fonctions composantes des fonctions h prennent alors la forme suivante :

$$h_k(.) = \langle w_k, . \rangle + b_k,$$

où w_k est un élément de $E_{\Phi(\mathcal{X})}$. Ainsi, H_κ apparaît comme l'espace dual de $E_{\Phi(\mathcal{X})}$ (l'espace des formes linéaires sur $E_{\Phi(\mathcal{X})}$). Compte tenu de la définition de $\langle .,. \rangle_{H_\kappa}$, il s'agit même de l'espace de Hilbert dual de $E_{\Phi(\mathcal{X})}$. Dans la suite, pour une fonction h donnée, $\mathbf{w}$ désignera le vecteur $(w_k)_{1\leq k\leq Q}$ de $E^Q_{\Phi(\mathcal{X})}$ et $\mathbf{b}$ le vecteur $(b_k)_{1\leq k\leq Q}$ de $\mathbb{R}^Q$. On notera $\bar{\mathcal{H}}$ l'espace produit H^Q_κ, $\bar{h} = (\langle w_k, .\rangle)_{1\leq k\leq Q}$ ses fonctions (considérées comme des fonctions sur $\Phi(\mathcal{X})$) et $\|.\|_{\bar{\mathcal{H}}}$ sa norme. Par défaut, on prendra

$$\|\bar{h}\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \|\bar{h}_k\|^2_{H_\kappa}} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2} = \|\mathbf{w}\|,$$

où $\|.\|$ désigne la norme de ℓ_2. On munira également $E^Q_{\Phi(\mathcal{X})}$ et donc $\bar{\mathcal{H}}$ d'une seconde norme, $\|.\|_\infty$, définie par $\|\mathbf{w}\|_\infty = \max_{1\leq k\leq Q} \|w_k\|$. Par un faible abus de notations, on pourra aussi considérer que $\bar{\mathcal{H}}$ est le sous-ensemble de $\mathcal{H}$ correspondant aux fonctions vérifiant $\mathbf{b} = 0$. Dans ce cas, ses éléments pourront également être notés h.

Pour établir le lien entre les familles de fonctions évoquées dans cette section et la famille de fonctions univariées sur laquelle s'appuient les SVM biclasses, il faut considérer que dans le cas biclasse, il existe de manière implicite une famille $\mathcal{H}$ de fonctions à valeurs dans $\mathbb{R}^2$. Soit $\tilde{\mathcal{H}}$ la famille des fonctions $\tilde{h}$ réalisables par une SVM. L'opérateur associant à une fonction $\tilde{h}$ de $\tilde{\mathcal{H}}$ une fonction $h = (h_1, h_2)$ de $\mathcal{H}$ est défini par : $h = \left(\tilde{h}, -\tilde{h}\right)$. Nous voyons ainsi apparaître l'intérêt de la contrainte $\sum_{k=1}^{Q} h_k = 0$, qui fait de cette opérateur une bijection. Ainsi, pour plonger une SVM biclasse dans le cadre plus général considéré ici, il suffit de réécrire ses paramètres sous la forme $w_1 = w = -w_2$ et $b_1 = b = -b_2$.

2.3 Méthodes de décomposition

Les méthodes de décomposition permettent d'aborder un problème de discrimination à catégories multiples comme une combinaison de problèmes de calcul de dichotomies.

2.3.1 Approches "un contre tous"

L'approche "un contre tous" est la plus simple et la plus ancienne des méthodes de décomposition. Elle consiste à utiliser un classifieur binaire (à valeurs réelles) par catégorie. Le k-ième classifieur est destiné à distinguer la catégorie d'indice k de toutes les autres. Pour affecter un exemple, on le présente donc à Q classifieurs, et la décision s'obtient en application du principe "winner-takes-all" : l'étiquette retenue est celle associée au classifieur ayant renvoyé la valeur la plus élevée. On cite ordinairement comme plus anciens travaux évoquant l'emploi de cette stratégie avec des SVM [194] (voir aussi [228]). Dans [180], les auteurs soutiennent la thèse selon laquelle cette approche, aussi simple soit-elle, lorsqu'elle est mise en œuvre avec des SVM correctement paramétrées, obtient des performances qui ne sont pas significativement inférieures à celles des autres méthodes de décomposition et des SVM multiclasses actuelles. Il convient cependant de souligner qu'elle implique d'effectuer des apprentissages aux répartitions entre catégories très déséquilibrées, ce qui soulève souvent des difficultés pratiques.

2.3.2 Approches "un contre un"

Une autre méthode de décomposition très naturelle est la méthode "un contre un". Ordinairement attribuée à Knerr et ses co-auteurs [127], elle consiste à utiliser un classifieur par couple de catégories. Le classifieur indicé par le couple (k, l) (avec $1 \leq k < l \leq Q$), est destiné à distinguer la catégorie d'indice k de celle d'indice l. Pour affecter un exemple, on le présente donc à C^2_Q classifieurs, et la décision s'obtient habituellement en effectuant un vote majoritaire ("max-wins voting"). La voix de chaque classifieur peut être pondérée par une fonction de la valeur de la sortie calculée. Une référence de base pour l'analyse statistique de cette stratégie (dans un cadre où les SVM ne sont pas considérées spécifiquement) est [83]. L'auteur y dérive l'approche un contre un dans le cadre de l'estimation du classifieur de Bayes. Sous l'hypothèse que la frontière séparant une catégorie d'une autre peut être moins complexe que celle séparant cette même catégorie de toutes les autres, il y voit un moyen d'obtenir des estimateurs présentant

un biais plus faible qu'avec l'approche un contre tous. Naturellement, le prix à payer est un possible accroissement de la variance de ces estimateurs, compte tenu du fait que les bases d'apprentissage de chacun des classifieurs sont plus petites que l'échantillon initial.

Les premiers articles envisageant l'utilisation d'une méthode un contre un avec des SVM datent de 1998. Dans [155], les auteurs évoquent simplement cette possibilité (section 6), tandis qu'elle est évaluée dans le cadre d'une étude comparative dans [240] (voir aussi [130]). Mais l'extension la plus naturelle des travaux de Friedman est présentée dans [105]. Les auteurs considèrent également des classifieurs binaires estimant les probabilités a posteriori des classes $P(k \mid x, x \in \{k, l\})$, $1 \leq k < l \leq Q$, et la synthèse des sorties s'effectue au moyen d'une méthode nommée "couplage par paire" (*pairwise coupling*, PWC), de manière à produire à nouveau des estimations des probabilités a posteriori $P(k \mid x)$ pour l'ensemble des Q catégories. Le modèle probabiliste sous-jacent est celui de Bradley-Terry [43], et la détermination de la valeur des paramètres s'obtient par minimisation de la divergence de Kullback-Leibler. Notons que cette solution, qui impose de post-traiter les sorties des SVM, semble en contradiction avec le principe même de ces machines : rechercher la frontière de décision optimale (de Bayes) et non une estimation des densités conditionnelles. On voit ainsi se profiler les difficultés inhérentes aux problèmes incorrectement posés au sens d'Hadamard [15], ou au fléau de la dimension (*curse of dimensionality*). Cependant, les résultats expérimentaux fournis sont bons. De nombreux auteurs ont proposé des améliorations de la méthode de base, portant soit sur la façon de réaliser le couplage par paire, soit sur l'estimation des probabilités conditionnelles $P(k \mid x, x \in \{k, l\})$. Parmi les contributions du premier type, on peut en particulier citer [162, 142]. Pour produire des estimations des probabilités a posteriori d'un couple de catégories à partir des sorties de la SVM correspondante, Hastie et Tibshirani proposent de modéliser les densités conditionnelles par des gaussiennes sphériques. m_+ et m_- représentant respectivement les moyennes des exemples d'apprentissage de la classe "positive" et de la classe "négative", les centres de ces gaussiennes sont $b - (m_+ + m_-)/2$ et $b + (m_+ + m_-)/2$, et la variance, commune, est le seul paramètre qui reste à déterminer. Cela revient à modéliser les probabilités a posteriori par des sigmoïdes dont la pente est fonction de la variance des gaussiennes. Dans [174], Platt critique ce modèle qu'il juge trop pauvre. Au lieu d'estimer les densités conditionnelles, il propose d'utiliser un modèle paramétrique s'adaptant directement à la probabilité conditionnelle de la classe "positive" sachant la sortie de la SVM. Il considère spécifiquement le cas où ce modèle est une sigmoïde à deux paramètres A et B :

$$P\left(+1 \mid \tilde{h}(x)\right) = \frac{1}{1 + \exp\left(A\tilde{h}(x) + B\right)}.$$

Les valeurs de A et B sont déterminées par un apprentissage appliquant le principe de maximum de vraisemblance. Cependant, Platt n'évalue pas l'utilisation de cette modification dans le cadre de la mise en œuvre du couplage par paire. L'incorporation de sa méthode d'estimation des probabilités a posteriori des classes dans une méthode de décomposition s'appuyant sur un raffinement du couplage par paire est l'objet d'un travail postérieur de Li et ses co-auteurs [142]. De manière étonnante, la combinaison avec la version originale du couplage par paire est encore postérieure [68]. Les auteurs de ce dernier article n'avaient probablement pas pris connaissance de [142] à l'époque de leurs travaux. La méthode ainsi produite, nommée PWC_PSVM, fournit des résultats expérimentaux supérieurs à ceux des autres méthodes de décomposition classiques. Les auteurs, s'appuyant sur ces résultats, soulignent la qualité des estimations fournies par le modèle de Platt.

Un raffinement très intéressant est introduit dans [10, 11]. Il consiste à utiliser des classifieurs de base produisant non plus deux mais trois réponses différentes, -1, 0 et 1, suivant que le point qui leur est présenté appartient à l'une des deux catégories qu'ils privilégient (réponse -1 ou 1), ou au contraire à l'une des $Q - 2$ autres catégories (réponse 0). Pour appliquer ce schéma avec des SVM, les auteurs spécifient une machine originale, nommée K-SVCR, qui combine celle dédiée à la discrimination et celle dédiée à la régression [204, 229]. Dans cette configuration, l'approche un contre un ne souffre plus de son principal défaut évoqué plus haut, celui d'utiliser des bases d'apprentissage pouvant être significativement plus petites que l'échantillon initial.

La référence la plus récente sur l'analyse et l'évaluation de l'approche un contre un semble être [85]. L'auteur y démontre que, lorsque l'algorithme d'apprentissage du classifieur de base est super-linéaire en la taille de l'ensemble d'apprentissage, ce qui est notamment le cas avec les SVM (voir la section 2.6.1),

l'utilisation d'une décomposition un contre un prend moins de temps que l'utilisation d'une décomposition un contre tous. Le gain est d'autant plus important que le classifieur de base est plus complexe.

Au-delà des arguments théoriques ou empiriques avancés pour justifier l'emploi de méthodes de décomposition, une considération pratique explique leurs bonnes performances. Les experts spécifiant le problème d'apprentissage et effectuant la collecte des données ont souvent à l'esprit le schéma canonique dans lequel les catégories se trouvent être aussi indépendantes que possible les unes des autres. Les membres de la classe 1 sont aussi différents de ceux de la classe 2 que de ceux de la classe 3. Les cas pratiques échappant à ce schéma sont rares.

2.3.3 Utilisation de codes correcteurs d'erreurs

L'introduction de l'emploi de codes correcteurs d'erreurs (ECOC) [171] en apprentissage est ordinairement attribuée à Duda et ses co-auteurs [70]. La référence de base dans laquelle ces codes sont utilisés en discrimination multiclasse pour spécifier des méthodes de décomposition est [67]. Ce cadre, dans une version avancée présentée plus bas, englobe les deux méthodes de décomposition décrites ci-dessus. A la base, le principe consiste à représenter les catégories par des mots binaires de même taille, les "mots codes". En notant N la taille de ces mots, on obtient ainsi une matrice $M = (m_{kl})$ de $\mathcal{M}_{Q,N}(\{0,1\})$ (ou $\mathcal{M}_{Q,N}(\{-1,1\})$) dont les lignes $(M_{k.})_{1\leq k\leq Q}$ sont les mots codes et dont les colonnes $(M_{.l})_{1\leq l\leq N}$ spécifient N dichotomies (partitions en deux super-catégories de l'ensemble des catégories). Chaque dichotomie est calculée par un classifieur. Ainsi, chaque exemple est associé à un vecteur de $\{0,1\}^N$ ou $\{-1,1\}^N$ (suivant l'espace auquel appartient M). Il est affecté à la catégorie correspondant au mot code (vecteur $M_{k.}$) le plus proche de ce vecteur au sens de la distance de Hamming. Naturellement, cette approche est d'autant plus efficace que les mots codes sont plus distants les uns des autres (toujours au sens de la distance de Hamming). C'est là qu'interviennent les ECOC. En fait, il convient de maximiser non seulement la séparation des lignes de la matrice, mais aussi la séparation de ses colonnes. Si cette dernière séparation n'est pas assurée, alors les classifieurs associés à des colonnes proches risquent d'effectuer des erreurs similaires (corrélées). Or, l'utilisation des ECOC n'est efficace que si les erreurs effectuées sur les différents bits sont relativement peu corrélées, de manière que la probabilité d'observer des erreurs simultanées sur un grand nombre de bits soit faible. Les auteurs de [67] décrivent une étude comparative dans laquelle C4.5 et des PMC sont mis en œuvre pour effectuer des tâches de discrimination à catégories multiples suivant trois procédures : l'approche multiclasse directe, la décomposition un contre tous et l'utilisation d'ECOC. Leur conclusion est que la dernière solution donne les meilleures perfomances, même dans le cas des petits échantillons, ceci en dépit du fait qu'elle conduise à la construction de fonctions de décision relativement plus complexes. Le gain n'est pas fonction de la manière dont les mots codes sont associés aux différentes catégories.

Spécifier un système discriminant multiclasse fondé sur une méthode de décomposition nécessite d'effectuer trois choix : nature des classifieurs de base, nature des dichotomies à réaliser au moyen de ces classifieurs et méthode de combinaison des prédictions. Dans [3], les auteurs étudient ce sujet pour les classifieurs à grande marge (SVM et AdaBoost [82, 193]), dans un cadre généralisant ceux des études évoquées jusqu'ici dans la section 2.3. En premier lieu, en considérant une matrice des mots codes dont les coefficients peuvent prendre trois valeurs, $\{-1,0,1\}$, au lieu de deux, ils permettent l'intégration du cas de l'approche un contre un. Elle résulte de la convention suivante : si $m_{kl} = 0$, alors le classifieur binaire d'indice l ne prend pas en compte les exemples d'apprentissage appartenant à la catégorie k. Ensuite, ils considèrent un décodage (une méthode de combinaison des prédictions) fondée sur une fonction de perte, comme une alternative à l'emploi de la distance de Hamming. En notant $g^{(l)}$, pour l allant de 1 à N, les classifieurs binaires, $\tilde{g} = \left(g^{(l)}\right)_{1\leq l\leq N}$, ℓ_{dec} la fonction de perte, dont l'argument est la marge biclasse standard, et $d_{\ell_{\text{dec}}}$ la fonction de décodage, on a :

$$\forall x \in \mathcal{X},\ \forall k \in \{1,\ldots,Q\},\ d_{\ell_{\text{dec}}}\left(M_{k.}, \tilde{g}(x)\right) = \sum_{l=1}^{N} \ell_{\text{dec}}\left(m_{kl} g^{(l)}(x)\right).$$

$d_{\ell_{\text{dec}}}$ est une mesure de dissimilarité entre vecteurs (ℓ_{dec} est une fonction décroissante de son argument).

En conséquence, la catégorie k^* à laquelle est affecté x est celle définie par :

$$k^* = \underset{k}{\operatorname{argmin}}\, d_{\ell_{\text{dec}}}\left(M_{k.}, \tilde{g}(x)\right)$$

toujours sous l'hypothèse que cette expression n'est pas ambiguë. La "perte binaire moyenne" de $\tilde{g}$ sur la suite d'observations $((x_i, y_i))_{1\leq i\leq m}$ a pour formule :

$$\epsilon = \frac{1}{mN}\sum_{i=1}^{m}\sum_{l=1}^{N}\ell_{\text{dec}}\left(m_{y_i l}g^{(l)}(x_i)\right) = \frac{1}{mN}\sum_{i=1}^{m} d_{\ell_{\text{dec}}}\left(M_{y_i.}, \tilde{g}\left(x_i\right)\right).$$

Les deux principaux résultats exposés par les auteurs sont une borne sur l'erreur en apprentissage du système multiclasse en fonction de N, ϵ, $\ell_{\text{dec}}(0)$ et la distance minimale entre deux lignes distinctes de M (théorème 1) ainsi qu'une borne sur l'erreur en généralisation dédiée au cas où la détermination des classifieurs $g^{(l)}$ est fondée sur l'algorithme AdaBoost (théorème 3). Cette borne, comme l'extension de celle de Bartlett évoquée dans la section 2.5.1, s'appuie sur une notion de risque impliquant une marge multiclasse définie de la manière suivante.

Définition 2 (Fonction $\mathcal{M}$) *Soit $\mathcal{M}$ la fonction de $\mathbb{R}^Q \times \{1,\ldots,Q\}$ dans $\mathbb{R}$ définie par :*

$$\forall (v,k) \in \mathbb{R}^Q \times \{1,\ldots,Q\}\,,\ \mathcal{M}(v,k) = \frac{1}{2}\left(v_k - \max_{l\neq k} v_l\right).$$

On désigne par $\mathcal{M}(v,.)$ le maximum pour k appartenant à $\{1,\ldots,Q\}$ des termes $\mathcal{M}(v,k)$.

Définition 3 (Marge multiclasse, définition 7 dans [75] (voir aussi [3])) *Soit $\mathcal{G}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\mathbb{R}^Q$. Pour un couple $(x,y) \in \mathcal{X}\times\mathcal{Y}$ donné, la* marge multiclasse *de $g \in \mathcal{G}$ est définie comme étant égale à $\mathcal{M}\left(g(x), y\right)$.*

Cette notion de marge étend de manière directe au cas multiclasse la notion de marge biclasse standard introduite par Vapnik et Chervonenkis dans les années 60. Allwein et ses co-auteurs définissent la fonction $g = (g_k)_{1\leq k\leq Q}$ de manière que la k-ième composante du vecteur $g(x)$ fournisse une mesure de la similarité entre le vecteur $\tilde{g}(x)$ et le mot code associé à la k-ième catégorie, $M_{k.}$. η étant un réel strictement positif, la formule est la suivante :

$$g(x) = \left(-\frac{1}{\eta}\ln\left(\frac{1}{N}\sum_{l=1}^{N} e^{-\eta m_{kl}g^{(l)}(x)}\right)\right)_{1\leq k\leq Q}. \tag{2.2}$$

On aimerait disposer d'une étude évaluant l'intérêt de mettre en œuvre la machine K-SVCR dans le cadre général considéré par Allwein et ses co-auteurs, c'est-à-dire au-delà de l'application de la seule méthode de décomposition un contre un. A notre connaissance, cette étude demeure à réaliser.

La littérature fournit un certain nombre d'exemples de tâches de discrimination multiclasse du monde réel abordées en mettant en œuvre une méthode de décomposition fondée sur l'emploi de codes correcteurs d'erreurs, avec comme classifieurs de base des SVM biclasses. Parmi celles-ci, on peut en particulier citer la catégorisation de textes, étudiée dans [126]. Les auteurs observent que l'emploi d'ECOC fournit de meilleures performances que la méthode de décomposition un contre tous pour la reconnaissance des catégories faiblement représentées. Ils expliquent le phénomène par le fait que dans le cas où certaines catégories sont nettement moins représentées que d'autres, les SVM tirent profit du rééquilibrage entre exemples positifs et négatifs induit par les agrégations de catégories propres à l'utilisation d'ECOC. Les auteurs de [57] proposent une méthode permettant de spécifier un code (une matrice M) efficace pour un problème multiclasse donné, c'est-à-dire un ensemble donné de classifieurs binaires déjà entraînés. Pour ce faire, ils considèrent des matrices M appartenant à $\mathcal{M}_{Q,N}(\mathbb{R})$ et non plus à $\mathcal{M}_{Q,N}(\{0,1\})$ (la recherche d'une matrice des mots codes optimale dans cet espace discret est un problème NP-complet), et munissent $\mathcal{M}_{Q,N}(\mathbb{R})$ de la norme ℓ_p, en considérant qu'à toute matrice de cet espace peut être associé un vecteur de $\mathbb{R}^{QN}$ correspondant à la concaténation de ses vecteurs lignes. Ils introduisent également un produit scalaire K sur $\mathbb{R}^N$, qui vient se substituer à la distance de Hamming pour définir le classement

des exemples. Par comparaison avec les travaux d'Allwein et ses co-auteurs, le vecteur $g(x)$ a donc pour expression :

$$g(x) = \left(K\left(M_{k.}, \tilde{g}(x)\right)\right)_{1 \leq k \leq Q}$$

au lieu de l'équation 2.2. Ce cadre leur permet de définir la matrice M recherchée comme étant solution optimale du problème d'optimisation sous contraintes suivant :

Problème 1 (Optimisation de la matrice des mots codes)

$$\min_{M \in \mathcal{M}_{Q,N}(\mathbb{R})} \left\{ \lambda \|M\|_p^q + \sum_{i=1}^{m} \xi_i \right\}$$

$$s.c.\ \ K\left(M_{y_i.}, \tilde{g}(x)\right) - K\left(M_{k.}, \tilde{g}(x)\right) + \delta_{y_i,k} \geq 1 - \xi_i, \ \ (1 \leq i \leq m), \ \ (1 \leq k \leq Q),$$

où δ est le symbole de Kronecker ($\delta_{y_i,k}$ est égal à 1 si $k = y_i$ et 0 dans le cas contraire). On remarquera que l'utilité des contraintes de bon classement correspondant à $k = y_i$ est d'imposer la positivité des variables d'écart ξ_i. Les auteurs traitent les cas $(p,q) = (1,1)$ et $(p,q) = (\infty,1)$, pour lesquels le problème 1 est un problème linéaire, ainsi que le cas $(p,q) = (2,2)$, conduisant à la résolution d'un problème de programmation quadratique. L'apprentissage de la SVM multiclasse décrite dans la section 2.4.1.2 apparaît comme un cas particulier de ce dernier problème. Ces travaux s'apparentent à ceux décrits dans [221], qui introduisaient déjà un cadre théorique pour transformer le problème de nature combinatoire consistant à trouver le codage optimal en un problème d'optimisation continue. Ce nouveau problème, découlant de l'application du principe de maximum de vraisemblance, se prêtait à la mise en œuvre de l'algorithme "espérance-maximisation" (EM) [63]. Dans [187], Rätsch et ses co-auteurs proposent une extension des travaux de Crammer et Singer consistant à apprendre à la fois la fonction $\tilde{g}$ et la matrice M. Plus précisément, une procédure itérative fait alterner la résolution de problèmes d'optimisation portant sur $\tilde{g}$ et la résolution de problèmes d'optimisation portant sur M. Ces problèmes consistent en la minimisation d'un risque empirique pénalisé sous des contraintes de bon classement qui sont pratiquement équivalentes à celles du problème 1.

A notre connaissance, la dernière avancée majeure sur l'emploi d'ECOC pour combiner des classifieurs binaires à grande marge est décrite dans [169]. Elle consiste à résoudre le problème du décodage en utilisant une fonction d_{PPF} qui s'appuie non plus sur une fonction de perte mais sur des estimations des probabilités conditionnelles des classes sachant $\tilde{g}$. Précisément,

$$\forall k \in \{1, \ldots, Q\}, \ d_{\text{PPF}}\left(M_{k.}, \tilde{g}(x)\right) = -\log\left(P\left(Y = k | \tilde{g}(x)\right)\right).$$

Comme dans [3], la matrice M est supposée appartenir à $\mathcal{M}_{Q,N}(\{-1,0,1\})$. Soit $\mathcal{O} = \left\{o_p : \ 1 \leq p \leq 2^N\right\}$ l'ensemble des mots codes de taille N sur $\{-1,1\}$ et $O = (O_l)_{1 \leq l \leq N}$ un vecteur aléatoire à valeurs dans cet ensemble. L'expression des probabilités conditionnelles est la suivante :

$$P\left(Y = k | \tilde{g}(x)\right) = \sum_{p=1}^{2^N} P\left(Y = k | O = o_p, \tilde{g}(x)\right) P\left(O = o_p | \tilde{g}(x)\right).$$

A chaque ligne $M_{k.}$ de la matrice M, et donc à chaque catégorie k, est associé l'ensemble $\mathcal{C}_k$ des mots codes $o_p = (o_{pl})_{1 \leq l \leq N}$ vérifiant : $\forall l \in \{1, \ldots, N\}$, $o_{pl} = m_{kl}$ si $m_{kl} \neq 0$. On note $\bar{\mathcal{C}} = \mathcal{O} \setminus \bigcup_{1 \leq k \leq Q} \mathcal{C}_k$. Les composantes O_l du vecteur aléatoire O vérifient des hypothèses d'indépendance permettant de reformuler les probabilités conditionnelles d'intérêt sous la forme :

$$\begin{aligned} P\left(Y = k | \tilde{g}(x)\right) &= \sum_{o_p \in \mathcal{C}_k} \prod_{l=1}^{N} P\left(O_l = o_{pl} | g^{(l)}(x)\right) + \frac{1}{Q} \sum_{o_q \in \bar{\mathcal{C}}} P\left(O = o_q | \tilde{g}(x)\right) \\ &= \prod_{l:\ m_{kl} \neq 0} P\left(O_l = m_{kl} | g^{(l)}(x)\right) + \frac{1}{Q} \sum_{o_q \in \bar{\mathcal{C}}} P\left(O = o_q | \tilde{g}(x)\right). \end{aligned}$$

Comme dans le cas des travaux relatifs au couplage par paire exposés dans la section 2.3.2, les auteurs s'inspirent des travaux de Platt en utilisant des sigmoïdes à deux paramètres pour estimer les probabilités conditionnelles à la base de leur modèle. La formule est la suivante :

$$\forall k \in \{1,\ldots,Q\}\,,\ \forall l \in \{1,\ldots,N\}\ : m_{kl} \neq 0,\ P\left(O_l = m_{kl}|g^{(l)}(x)\right) = \frac{1}{1+\exp\left\{m_{kl}\left(A_l g^{(l)}(x)+B_l\right)\right\}},$$

où les paramètres ajustables A_l et B_l reflètent la pente et l'ordonnée à l'origine des distributions cumulées des fonctions $g^{(l)}$. Cette étude propose également une borne sur l'erreur empirique "leave-one-out" qui est décrite dans la section 2.7.1.

2.3.4 Méthodes fondées sur des graphes de décision

La première méthode de décomposition fondée sur un graphe de décision est la DAGSVM [175]. Comme son nom l'indique, cette méthode s'appuie plus précisément sur un graphe de décision orienté sans cycle (DDAG).

Définition 4 (Graphe de décision orienté sans cycle, définition 1 dans [175]) *Etant donnés un espace $\mathcal{X}$ et un ensemble $\mathcal{F}$ de fonctions de $\mathcal{X}$ dans $\{0,1\}$, la famille $DDAG(\mathcal{F})$ de* graphes de décision orientés sans cycle à Q catégories sur $\mathcal{F}$ *est l'ensemble des fonctions réalisables à partir d'un graphe orienté sans cycle (DAG) enraciné à Q feuilles associées à chacune des classes, où chacun des C_Q^2 nœuds internes est étiqueté avec un élément de $\mathcal{F}$. Les nœuds sont disposés en triangle avec au sommet la seule racine, deux nœuds dans la seconde couche et ainsi de suite jusqu'à la couche finale de Q feuilles. Le i-ième nœud de la couche interne d'indice j ($j < Q$) est connecté aux nœuds (ou feuilles) d'indices i et $i+1$ de la couche d'indice $j+1$.*

Définition 5 (DAGSVM [175]) *Une DAGSVM est un modèle de discrimination multiclasse dont l'architecture est un graphe de décision orienté sans cycle avec pour étiquettes de ses nœuds des SVM biclasses. A un nœud donné est associée une liste de classes auxquelles l'exemple d'intérêt peut appartenir. La SVM correspondante effectue une décision entre les deux catégories aux extrémités de la liste : les catégories 1 et Q pour la SVM située à la racine, 2 et Q pour la SVM située sur le fils gauche de la racine, 1 et $Q-1$ pour la SVM située sur le fils droit de la racine et ainsi de suite. Les nœuds de la couche d'indice $Q-1$ produisent une décision en séparant les deux seules catégories contenues dans leur liste.*

La figure 2.1 représente la DAGSVM correspondant au cas où le nombre de catégories est quatre.

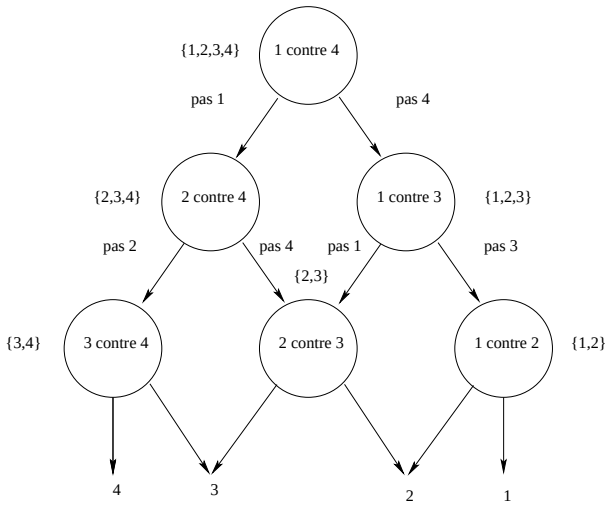

Fig. 2.1 – Architecture d'une DAGSVM à quatre catégories.

En pratique, on s'aperçoit que les classifieurs de base d'une DAGSVM sont exactement les mêmes que dans le cas de la méthode de décomposition un contre un. La différence réside dans la manière de construire la décision à partir de ces classifieurs. Dans la DAGSVM, chaque exemple est présenté à $Q-1$ SVM (une par couche interne du DAG), contre C_Q^2 avec la méthode de décomposition un contre un. Les auteurs étudient leur architecture selon trois critères : performances en généralisation (calcul d'un risque garanti à risque empirique nul, voir la section 2.5.3), performances mesurées sur des problèmes jouet et temps de calcul. Sous l'hypothèse que les exemples d'apprentissage se répartissent de manière équilibrée entre les Q catégories, ils établissent que le temps d'apprentissage de la DAGSVM est inférieur à celui de la méthode de décomposition un contre tous. Le temps d'utilisation en test de la DAGSVM est inférieur à celui des deux méthodes de décomposition, un contre tous et un contre un. Dans le second cas, le gain résulte directement de la manière dont se construit la décision. Cette approche présente cependant un défaut : la décision qu'elle produit peut être affectée par la manière dont sont numérotées les catégories. Supposons ainsi un problème à trois catégories, où pour un exemple donné, le premier classifieur rencontré (1 contre 3) rejette la catégorie 3, le classifieur "2 contre 3" choisissant la catégorie 3 tandis que le classifieur "1 contre 2" choisit la catégorie 2. La DAGSVM affectera alors l'exemple à la catégorie 2. Renumérotons à présent les catégories de la manière suivante : $1 \longrightarrow 2$, $2 \longrightarrow 1$ ($3 \longrightarrow 3$). On s'aperçoit que la machine affecte cette fois l'exemple à la nouvelle catégorie 2, c'est-à-dire à l'ancienne catégorie 1.

Dans [26], les auteurs proposent une méthode de décomposition dont la structure classificatoire ne s'appuie pas sur un DDAG mais sur un dendrogramme. Il la nomment en conséquence "Dendogram-based SVM" (DSVM). Le dendrogramme est construit en calculant les centres de gravité des sous-ensembles de l'ensemble d'apprentissage associés aux différentes catégories, et en appliquant ensuite à ces centres l'algorithme de la classification ascendante hiérarchique (CAH) [41, 133]. Le modèle discriminant s'obtient alors en plaçant sur les nœuds de l'arbre des SVM biclasses. L'architecture correspondante est illustrée par la figure 2.2.

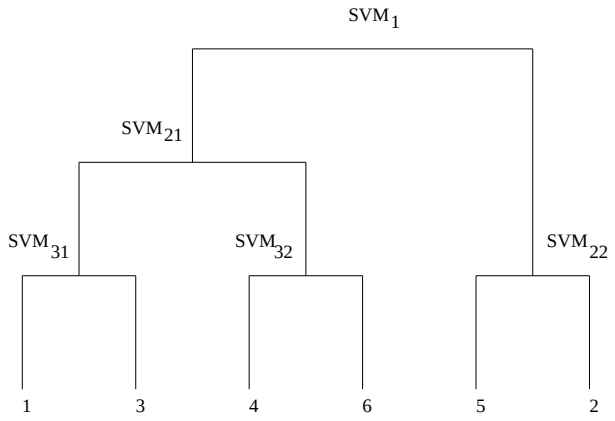

FIG. 2.2 – Architecture d'une DSVM à six catégories.

Dans cet exemple, la SVM d'indice 1 sépare l'ensemble de catégories $\{1,3,4,6\}$ de l'ensemble de catégories $\{2,5\}$, tandis que la SVM d'indice 21 sépare l'ensemble de catégories $\{1,3\}$ de l'ensemble de catégories $\{4,6\}$. Afin d'illustrer les performances de la DSVM, les auteurs fournissent les résultats d'expériences comparatives faisant intervenir les méthodes de décomposition un contre tous et un contre un, ainsi qu'un PMC. Malheureusement, ils n'incorporent à leur étude ni la DAGSVM ni aucune SVM multiclasse. Cependant, les taux de reconnaissance obtenus sont encourageants.

2.4 Principaux modèles de SVM multiclasses

La plupart des modèles de SVM multiclasses partagent la même "architecture" de base. Ils se distinguent simplement par leurs algorithmes d'apprentissage. Cette architecture correspond à la famille de fonctions $\mathcal{H}$ décrite dans la section 2.2.2. Dans la suite, nous nommerons M-SVM ces machines.

2.4.1 M-SVM

En reprenant les notations introduites dans la section 2.2.2, les M-SVM s'appuient sur le modèle affine multivarié donné par :

$$\forall x \in \mathcal{X},\ h(x) = (\langle w_k, \Phi(x)\rangle + b_k)_{1\leq k\leq Q}\,.$$

Plus précisément, elles s'appuient sur la restriction de ce modèle à l'hyperplan d'équation $\sum_{k=1}^{Q} h_k = 0$. Ce potentiel n'est en fait pas entièrement disponible. Il est limité par la nature de l'ensemble d'apprentissage, $((x_i, y_i))_{1\leq i\leq m}$. La forme générale que peut prendre la solution de l'apprentissage en fonction de la classe de fonctions de base et de cet ensemble est donnée par des théorèmes de représentation [125]. Ces théorèmes s'appuient sur la forme générale prise par la fonction objectif minimisée par l'apprentissage. On s'aperçoit ainsi que pour les M-SVM, comme pour les SVM, architecture et algorithme d'apprentissage sont indissociables. Cela nous conduit à revenir brièvement au cas biclasse. Dans ce cas, la fonction objectif est un risque empirique pénalisé où l'indicatrice caractérisant le bon classement est remplacée par une fonction de perte positive, décroissante et convexe, ℓ_{SVM}. La raison de cette substitution est purement pratique. Il s'agit d'obtenir pour algorithme d'apprentissage la résolution d'un problème de programmation convexe. La fonction ℓ_{SVM} habituelle est celle utilisée dans l'article introduisant la SVM à marge douce, [55]. Elle est nommée fonction de perte charnière (*hinge loss*) et a pour expression $\ell_{\mathrm{CV}}\left(y, \tilde{h}(x)\right) = \left(1 - y\tilde{h}(x)\right)_+$, la fonction $(.)_+$ retournant son argument si celui-ci est positif, et 0 dans le cas contraire. Nous en rencontrerons d'autres dans la suite de ce chapitre. La littérature propose plusieurs théorèmes de représentation relatifs aux SVM biclasses (voir en particulier [234, 78, 196]). Ces théorèmes utilisent des hypothèses sur la fonction de perte ℓ_{SVM} et la manière dont le terme de pénalisation s'exprime en fonction de la norme de la fonction de H_κ considérée. Les fonctionnelles utilisées comme fonction objectif par les différentes M-SVM peuvent toutes être exprimées sous une forme similaire à celle du cas biclasse. Ceci permet de donner de ces machines la définition générique suivante.

Définition 6 (M-SVM) *Soit $((x_i, y_i))_{1\leq i\leq m} \in (\mathcal{X} \times \{1, \ldots, Q\})^m$. Une M-SVM à Q catégories est un modèle discriminant à grande marge obtenu en minimisant sur l'hyperplan $\sum_{k=1}^{Q} h_k = 0$ de $\mathcal{H}$ une fonction objectif $J_{M\text{-}SVM}$ de la forme :*

$$J_{M\text{-}SVM}(h) = \sum_{i=1}^{m} \ell_{M\text{-}SVM}(y_i, h(x_i)) + \lambda \left\|\bar{h}\right\|_{\bar{\mathcal{H}}}^2 \tag{2.3}$$

où la fonction de perte $\ell_{M\text{-}SVM}$ est construite autour de la fonction $(.)_+$.

Les deux éléments distinguant les différentes M-SVM sont donc cette fonction et le choix de la norme sur $\bar{\mathcal{H}}$. Cependant, dans tous les cas, la forme générale des fonctions composantes h_k (le théorème de représentation) est la même que dans le cas biclasse, c'est-à-dire que l'on a :

$$\forall x \in \mathcal{X},\ h_k(x) = \sum_{i=1}^{m} \beta_{ik} k(x_i, x) + b_k. \tag{2.4}$$

Nous décrivons à présent les trois principaux modèles de M-SVM, ainsi que les variantes auxquelles ils ont donné naissance.

2.4.1.1 Modèle de Weston et Watkins et ses variantes

La première publication décrivant une SVM multiclasse est [240] (voir aussi [241]). Elle présente un modèle proposé indépendamment par Vapnik et Blanz lors de communications orales légèrement antérieures (communication personnelle de Volker Blanz), puis plus tard par d'autres auteurs sous des formes variées. Si plusieurs justifications à sa spécification ont été avancées, justifications sans lien, au moins explicite, avec le principe inductif de minimisation structurelle du risque, la manière la plus naturelle de l'introduire est probablement de le relier à la notion d'hyperplan de marge maximale. L'extension multiclasse la plus directe de la notion de marge géométrique sur laquelle reposent les SVM standard est la suivante :

Définition 7 (Marges géométriques multiclasses) *Considérons une M-SVM (une fonction de $\mathcal{H}$) dont l'erreur sur l'ensemble d'apprentissage $((x_i, y_i))_{1\leq i\leq m}$ est nulle. γ_{kl}, sa* marge relative aux classes d'indices k et l, *est la distance euclidienne minimale (dans l'espace de représentation) entre un point de l'ensemble d'apprentissage dans l'une ou l'autre de ces classes et l'hyperplan les séparant (d'équation $\langle w_k - w_l, \Phi(x)\rangle + b_k - b_l = 0$). En notant*

$$d_{M\text{-}SVM} = \min_{1\leq k<l\leq Q} \left\{ \min \left[\min_{x_i\ :y_i=k} (h_k(x_i) - h_l(x_i)) \,, \min_{x_j\ :y_j=l} (h_l(x_j) - h_k(x_j)) \right] \right\}$$

et

$$d_{kl} = \frac{1}{d_{M\text{-}SVM}} \min \left[\min_{x_i\ :y_i=k} (h_k(x_i) - h_l(x_i) - d_{M\text{-}SVM}) \,, \min_{x_j\ :y_j=l} (h_l(x_j) - h_k(x_j) - d_{M\text{-}SVM}) \right],$$

l'expression analytique de γ_{kl} est donc :

$$\gamma_{kl} = d_{M\text{-}SVM} \frac{1 + d_{kl}}{\|w_k - w_l\|}.$$

Cette extension s'appuie, comme la définition originale, sur la notion de forme canonique d'un hyperplan séparateur. A la valeur 1 choisie par Vapnik pour caractériser cette forme se substitue la valeur de $d_{\text{M-SVM}}$ (nous verrons dans la suite que cette valeur est effectivement fonction de la M-SVM considérée). L'introduction des termes positifs ou nuls d_{kl} prend alors en compte le fait qu'a priori, un seul hyperplan séparateur se trouve sous la forme canonique. Si les définitions de la marge (analytique) multiclasse (définition 3) et des marges géométriques multiclasses apparaissent comme des extensions naturelles des définitions correspondantes du cas biclasse, à l'inverse, le lien qu'elles entretiennent entre elles est plus difficile à caractériser que dans le cas biclasse. Ainsi, un même exemple peut appartenir à plusieurs marges géométriques, correspondant à des frontières faisant intervenir des catégories n'entrant pas dans l'expression de sa marge multiclasse. La définition 7 laisse apparaître $\sum_{k<l} \|w_k - w_l\|^2$ comme l'un des termes de contrôle susceptibles de généraliser le terme $\|w\|^2$ du cas biclasse (en gardant à l'esprit le fait que le vecteur w peut se réécrire sous la forme $\frac{1}{2}(w_1 - w_2)$). Une fois le pénalisateur choisi, reste à déterminer la fonction de perte rendant compte de l'adéquation des fonctions du modèle aux données d'apprentissage. Une des formulations possibles pour les contraintes de bon classement, qui tende à maximiser, pour chaque exemple d'apprentissage (x_i, y_i), la marge multiclasse $\mathcal{M}(h(x_i), y_i)$, est la suivante :

$$\forall k \neq y_i,\ h_{y_i}(x_i) - h_k(x_i) \geq 1 - \xi_{ik},$$

où les $Q-1$ variables d'écart ξ_{ik} sont positives ou nulles. On obtient ainsi pour fonction de perte :

$$\ell_{\text{WW}}(y, h(x)) = \sum_{k\neq y} (1 - h_y(x) + h_k(x))_+$$

et $d_{\text{WW}} = 1$. Si l'on s'arrête à ces considérations, la fonction objectif qui en découle est :

$$J(h) = \sum_{i=1}^{m} \sum_{k\neq y_i} (1 - h_{y_i}(x_i) + h_k(x_i))_+ + \lambda \sum_{k<l} \|w_k - w_l\|^2. \tag{2.5}$$

L'intérêt de considérer pour classe de fonctions la restriction de $\mathcal{H}$ à l'hyperplan d'équation $\sum_{k=1}^{Q} h_k = 0$ se présente ici sous un nouveau jour. En effet, si la forme quadratique $\sum_{k<l} \|w_k - w_l\|^2$ n'est pas positive non dégénérée, et ne définit donc pas une norme sur $\bar{\mathcal{H}}$, ($\sum_{k<l} \|w_k - w_l\|^2 = 0 \not\Longrightarrow \mathbf{w} = 0$), à l'inverse, elle induit une norme sur le l'hyperplan de $\bar{\mathcal{H}}$ défini par l'équation $\sum_{k=1}^{Q} w_k = 0$. La preuve est apportée par l'équation :

$$\sum_{k<l} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2 - \left\| \sum_{k=1}^{Q} w_k \right\|^2 \tag{2.6}$$

qui établit que sous l'hypothèse $\sum_{k=1}^{Q} w_k = 0$, $\sum_{k<l} \|w_k - w_l\|^2 = Q\|\mathbf{w}\|^2 = Q\left\|\bar{h}\right\|^2_{\bar{\mathcal{H}}}$. La fonction objectif définie par l'équation 2.5 apparaît ainsi pour cet hyperplan comme une instanciation de celle

caractérisant de manière générale les M-SVM, donnée par l'équation 2.3. En choisissant pour expression de la constante de marge douce en fonction du coefficient de pénalisation $C = (2Q\lambda)^{-1}$, on obtient en définitive pour algorithme d'apprentissage le programme quadratique suivant, qui est celui définissant la M-SVM de Weston et Watkins :

Problème 2 (M-SVM de Weston et Watkins, problème primal)

$$\min_{h\in\mathcal{H}} \left\{ \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k\neq y_i} \xi_{ik} \right\}$$

$$s.c. \left\{ \begin{array}{ll} \langle w_{y_i} - w_k, \Phi(x_i)\rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \end{array} \right. .$$

La caractérisation de cette machine comme séparateur à vaste marge, reposant sur l'équation 2.6, a été effectuée dans [93]. Il convient de remarquer qu'il n'est pas utile d'ajouter aux contraintes du problème 2 la contrainte $\sum_{k=1}^{Q} w_k = 0$. Celle-ci est implicite, puisqu'elle est vérifiée par la solution optimale. Pour s'en convaincre, il suffit d'appliquer la dualité lagrangienne. Plus précisément, on peut rechercher le point-selle du lagrangien, correspondant, en application des conditions de Kuhn-Tucker [80, 160], au zéro de son gradient par rapport aux variables primales. Soit α_{ik} le multiplicateur de Lagrange associé à la contrainte de bon classement $\langle w_{y_i} - w_k, \Phi(x_i)\rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}$. Afin de rendre l'exposé plus clair, nous utilisons la notation matricielle $\alpha = (\alpha_{ik})_{1\leq i\leq m, 1\leq k\leq Q}$ pour désigner le vecteur de ces multiplicateurs (le passage de la matrice au vecteur s'obtient naturellement par concaténation des vecteurs lignes). Sont introduites à cette occasion, pour i allant de 1 à m, les pseudo-variables α_{iy_i} toutes égales à 0. On obtient alors pour expression analytique des vecteurs w_k^* :

$$\forall k \in \{1, \ldots, Q\}\,,\ w_k^* = \sum_{x_i\in k}\sum_{l=1}^{Q} \alpha_{il}^*\Phi(x_i) - \sum_{i=1}^{m} \alpha_{ik}^*\Phi(x_i) \tag{2.7}$$

et par conséquent $\sum_{k=1}^{Q} w_k^* = \sum_{i=1}^{m}\sum_{l=1}^{Q} \alpha_{il}^*\Phi(x_i) - \sum_{k=1}^{Q}\sum_{i=1}^{m} \alpha_{ik}^*\Phi(x_i) = 0$. A cette observation, effectuée a posteriori à partir de la solution du problème dual, répond une explication a priori de cette propriété, fondée sur un simple calcul concernant le problème primal.

Proposition 1 *Soit J_{WW} la fonction objectif du problème 2 et h un élément de $\mathcal{H}$ défini par le couple $(\mathbf{w}, \mathbf{b})$, tel que $\sum_{k=1}^{Q} w_k \neq 0$. Soit h' l'élément de $\mathcal{H}$ défini par le couple $(\mathbf{w}', \mathbf{b})$ construit de la manière suivante : $\forall k \in \{1, \ldots, Q\}$, $w'_k = w_k - \frac{1}{Q}\sum_{l=1}^{Q} w_l$. Alors,*

$$\left\{ \begin{array}{l} \sum_{k=1}^{Q} w'_k = 0 \\ J_{WW}(h') < J_{WW}(h) \end{array} \right. .$$

Preuve Notons en premier lieu que les deux fonctions h et h' sont associées aux mêmes variables d'écart : $\forall x \in \mathcal{X}$, $\forall (k,l) \in \{1, \ldots, Q\}^2$, $h'_k(x) - h'_l(x) = h_k(x) - h_l(x)$. De plus, en posant $s_{\mathbf{w}} = \frac{1}{Q}\sum_{l=1}^{Q} w_l$, on a :

$$\sum_{k=1}^{Q} \|w'_k\|^2 - \sum_{k=1}^{Q} \|w_k\|^2 = \sum_{k=1}^{Q} \langle w_k - s_{\mathbf{w}}, w_k - s_{\mathbf{w}}\rangle - \sum_{k=1}^{Q} \langle w_k, w_k\rangle.$$

$$\sum_{k=1}^{Q} \|w'_k\|^2 - \sum_{k=1}^{Q} \|w_k\|^2 = -2Q\,\|s_{\mathbf{w}}\|^2 + Q\,\|s_{\mathbf{w}}\|^2 = -Q\,\|s_{\mathbf{w}}\|^2 < 0.$$

■

La formulation du problème d'apprentissage donnée par le problème 2 est précisément celle retenue dans [240, 229]. Dans [44], les auteurs utilisent un terme de contrôle différent : $\frac{1}{2}\left\{\sum_{k<l} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2\right\}$.

Cependant, là encore, la contrainte $\sum_{k=1}^{Q} w_k = 0$ intervient de manière implicite, et l'équation 2.6 nous indique par suite que cela revient au même (fournit la même solution), pourvu que la constante de marge douce soit modifiée en conséquence. La dimension de l'espace de représentation pouvant être très grande, voire infinie, de même que dans le cas biclasse, dans la majorité des cas, il est préférable de substituer à la résolution du problème 2 celle de son dual de Wolfe [80]. La formulation de ce dual résulte de l'application de la procédure décrite pour obtenir l'équation 2.7 : recherche du zéro du gradient du lagrangien par rapport aux variables primales et exploitation des équations ainsi produites afin d'obtenir un problème ne faisant plus intervenir que les multiplicateurs de Lagrange. En notant $H_{\mathrm{WW}} = \left(h_{ik,jl}^{(\mathrm{WW})}\right)_{1\leq i,j\leq m, 1\leq k,l\leq Q}$ la matrice de $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ de terme général

$$h_{ik,jl}^{(\mathrm{WW})} = \left(\delta_{y_i,y_j} - \delta_{y_i,l} - \delta_{y_j,k} + \delta_{k,l}\right)\kappa(x_i, x_j) \tag{2.8}$$

et 1_{Qm} le vecteurs de $\mathbb{R}^{Qm}$ dont toutes les composantes sont égales à 1, on obtient ainsi :

Problème 3 (M-SVM de Weston et Watkins, problème dual)

$$\min_{\alpha} \left\{ \frac{1}{2}\alpha^T H_{WW}\alpha - 1_{Qm}^T \alpha \right\}$$

$$s.c. \begin{cases} 0 \leq \alpha_{ik} \leq C, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{x_i \in k} \sum_{l=1}^{Q} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0, & (1 \leq k \leq Q-1) \end{cases} .$$

Comme dans le cas biclasse, les seuls multiplicateurs de Lagrange qui deviennent variables duales sont ceux associés aux contraintes de bon classement. Ceux associés à la positivité des variables d'écart interviennent simplement dans la forme prise par les contraintes-inégalités. Compte tenu de la présence des pseudo-variables α_{iy_i}, la matrice H_{WW} et le vecteurs 1_{Qm} ne constituent naturellement pas l'unique solution pour cette formulation algébrique. La résolution du problème 3 fournit l'expression des fonctions composantes $\langle w_k^*, . \rangle$. La valeur des composantes du vecteur $\mathbf{b}^* = (b_k^*)_{1\leq k\leq Q}$ se déduit de l'application des conditions de Kuhn-Tucker et de la contrainte $\sum_{k=1}^{Q} b_k = 0$. En utilisant les variables intermédiaires $\tilde{b}_k$, on obtient ainsi le système linéaire suivant :

$$\begin{cases} \tilde{b}_1 = 0 \\ 0 < \alpha_{ik}^* < C \Longrightarrow \tilde{b}_k - \tilde{b}_{y_i} = \langle w_{y_i}^* - w_k^*, \Phi(x_i)\rangle - 1, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ b_k^* = \tilde{b}_k - \frac{1}{Q}\sum_{l=1}^{Q} \tilde{b}_l, & (1 \leq k \leq Q) \end{cases} . \tag{2.9}$$

En exprimant la matrice H_{WW} dans le cas de deux catégories, on retrouve la matrice hessienne de la SVM biclasse à un facteur multiplicatif 2 près. Ceci est une conséquence du fait que le terme de pénalisation utilisé dans le problème 2 est $\sum_{k=1}^{Q} \|w_k\|^2$ et non le terme qui permettrait d'incorporer directement le cas biclasse (en multipliant également par $\frac{1}{2}$ le membre de gauche des contraintes de bon classement) : $\frac{1}{4}\sum_{k<l} \|w_k - w_l\|^2 = \frac{Q}{4}\sum_{k=1}^{Q} \|w_k\|^2$. En effet,

$$\|w\|^2 = \left\|\frac{1}{2}(w_1 - w_2)\right\|^2 = \frac{1}{4}\left\{\|w_1\|^2 + \|w_2\|^2 - 2\langle w_1, w_2\rangle\right\} = \frac{1}{2}\left(\|w_1\|^2 + \|w_2\|^2\right).$$

A cette différence de détail près, cette M-SVM se réduit bien à la SVM biclasse standard dans le cas de deux catégories. Le lien entre les valeurs des marges géométriques et celles des variables duales à l'optimum est très simple :

$$\frac{1}{Q}\sum_{k<l} \frac{(1+d_{kl})^2}{\gamma_{kl}^2} = \sum_{k=1}^{Q} \|w_k^*\|^2 = \alpha^{*T} H_{\mathrm{WW}} \alpha^*.$$

Il permet de vérifier rapidement que plus C est grand, plus les variables duales sont grandes et plus les marges sont petites. La machine se concentre sur la satisfaction des contraintes de bon classement,

au détriment de la taille des marges. Lorsque la machine est à marge dure, ce qui peut s'exprimer par $C = \infty$, l'application des conditions de Kuhn-Tucker fournit un lien plus simple encore, puisqu'aux égalités précédentes vient s'ajouter la suivante :

$$\frac{1}{Q}\sum_{k<l}\frac{(1+d_{kl})^2}{\gamma_{kl}^2} = 1_{Qm}^T\alpha^*. \tag{2.10}$$

Nous illustrons à présent le comportement de cette M-SVM sur un problème jouet emprunté à [74], afin de caractériser les marges géométriques "douces" qu'elle calcule. Dans ce problème de discrimination (partitionnement) à trois catégories dans le carré unité ($\mathcal{X} = \{x = (x_1, x_2) \ : 0 \leq x_1, x_2 \leq 1\}$), le domaine de la catégorie 1 est défini par $(x_1 > x_2) \wedge (1 - x_1 < x_2)$, celui de la catégorie 2 par $(x_1 < x_2)$ et celui de la catégorie 3 par $(x_1 > x_2) \wedge (1 - x_1 > x_2)$, le symbole $\wedge$ désignant l'opérateur de conjonction entre deux événements. Les points des deux segments de droites restants, constituant les frontières, ne sont affectés à aucune catégorie. Si le classifieur de Bayes correspondant à ce problème est bien linéaire, il n'appartient pas à la classe des fonctions réalisables par la M-SVM utilisant pour noyau le produit scalaire euclidien. En effet, pour que cette M-SVM sépare les classes 1 et 2 d'une part, 2 et 3 d'autre part, de manière optimale, il faut qu'elle vérifie $w_1 = w_3$ et $b_1 = b_3$, ce qui ne lui permet pas de séparer correctement les classes 1 et 3.

La figure 2.3 représente le comportement d'une M-SVM linéaire avec $C = 100$. L'ensemble d'apprentissage est constitué de 100 exemples tirés suivant la loi uniforme sur $\mathcal{X}$.

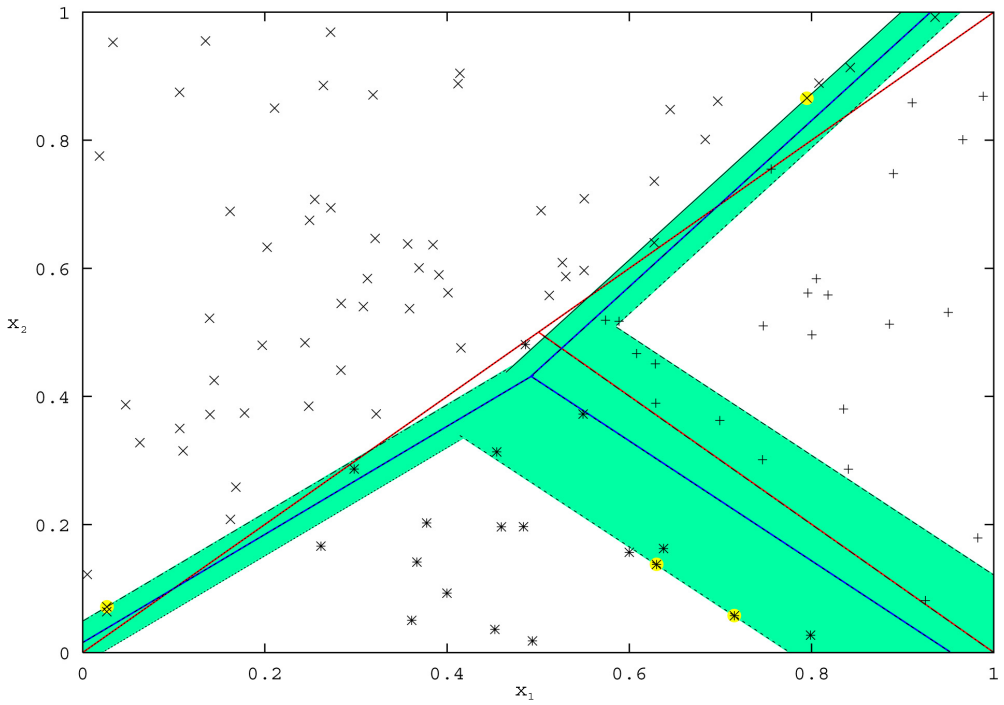

FIG. 2.3 – M-SVM de Weston et Watkins appliquée à un problème de discrimination à trois catégories dans le plan. Les frontières de décision optimales apparaissent en rouge, celles calculées par une M-SVM linéaire en bleu, les marges correspondantes étant matérialisées en vert. Les vecteurs support posés sur une marge sont cerclés de jaune.

On remarque que seuls quatre vecteurs support se trouvent sur une marge (il s'agit d'exemples pour lesquels au moins une variable duale appartient à l'intervalle ouvert $]0, C[$), aucun d'entre eux n'appartenant à la catégorie 1. Ceci ne permet pas de déterminer les surfaces de séparation calculées par la M-SVM en s'appuyant uniquement sur des éléments géométriques simples. Il faut utiliser pour ce faire l'expression analytique des hyperplans dont nous avons décrit ci-dessus l'obtention (équation 2.7 et système linéaire 2.9). Cette expression fait intervenir l'ensemble des vecteurs support (posés ou non sur une marge). La marge séparant les catégories 1 et 3 est nettement plus grande que les deux autres, ce qui démontre que le choix de la norme sur $\mathbf{w}$ joue un rôle important.

Au lieu d'optimiser globalement les marges, on peut souhaiter simplement maximiser la plus petite d'entre elles. On verra à la section 2.5.1 quelle incidence cela peut avoir sur la valeur du risque garanti. Cette possibilité, qui a à notre connaissance été explorée pour la première fois dans [93], ne peut pas être mise en œuvre en utilisant directement pour terme de contrôle $\max_{k<l} \|w_k - w_l\|^2$. En effet, le problème d'optimisation qui en résulte alors n'est pas un problème de programmation convexe. Une solution consiste à remarquer que $\frac{1}{4} \max_{k<l} \|w_k - w_l\|^2$ est majoré par $\max_k \|w_k\|^2 = \|\mathbf{w}\|_\infty^2$, qui est une fonction convexe de $\mathbf{w}$ puisqu'il s'agit d'une norme au carré, et choisir en conséquence cette dernière expression comme terme de pénalisation. Ceci conduit à la résolution du problème d'optimisation suivant :

Problème 4 (M-SVM de Weston et Watkins utilisant la norme $\|.\|_\infty$)

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} t^2 + C \sum_{i=1}^{m} \sum_{k \neq y_i} \xi_{ik} \right\}$$

$$s.c. \begin{cases} \langle w_{y_i} - w_k, \Phi(x_i) \rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \|w_k\| \leq t, & (1 \leq k \leq Q) \end{cases} .$$

Notons en premier lieu qu'il n'est pas possible de proposer une méthode de résolution de ce problème s'appuyant sur l'étude de la version du problème 1 correspondant à $(p, q) = (\infty, 1)$. La "norme infinie" que nous avons définie sur $E^Q_{\Phi(\mathcal{X})}$ est en fait nettement plus difficile à manipuler que la norme infinie (standard) sur $\mathcal{M}_{Q,N}(\mathbb{R})$ ($\|M\|_\infty = \max_{(k,l) \in \{1,\ldots,Q\} \times \{1,\ldots,N\}} |m_{kl}|$). La résolution de ce problème est également nettement plus compliquée que celle du problème 2. La raison en est cette fois la non différentiabilité de la norme $\|.\|_\infty$, qui rend délicate, par exemple, l'application du principe de dualité lagrangienne. L'algorithme que nous avons retenu est une variante de l'algorithme de Frank et Wolfe (voir la section 2.6.2) qui s'appuie sur la détermination, à chaque pas de la descente en gradient, du vecteur w_k vérifiant $\|w_k\| = \max_l \|w_l\|$. L'efficacité de cette procédure reste à établir.

2.4.1.2 Modèle de Crammer et Singer

Dans [56, 57], les auteurs considèrent une architecture simplifiée, qui n'est plus affine (dans l'espace de représentation) mais linéaire. En d'autres termes, ils restreignent leur étude à la classe de fonctions $\bar{\mathcal{H}}$. Pour cette famille, des considérations liées à la notion de marge multiclasse correspondant à la définition 3 les conduisent à proposer pour fonction de perte

$$\ell_{\mathrm{CS}}(y, h(x)) = \max_{1 \leq k \leq Q} \{h_k(x) + 1 - \delta_{y,k}\} - h_y(x) = \left(1 - h_y(x) + \max_{k \neq y} h_k(x)\right)_+$$

($d_{\mathrm{CS}} = 1$) et donc l'algorithme d'apprentissage suivant :

Problème 5 (M-SVM de Crammer et Singer, problème primal)

$$\min_{\bar{h} \in \bar{\mathcal{H}}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \xi_i \right\}$$

$$s.c. \ \langle w_{y_i} - w_k, \Phi(x_i) \rangle + \delta_{y_i,k} \geq 1 - \xi_i, \ \ (1 \leq i \leq m), \ \ (1 \leq k \leq Q),$$

où δ est comme précédemment le symbole de Kronecker. Lorsque κ est le produit scalaire euclidien (Φ est l'identité, et il existe $n \in \mathbb{N}^*$ tel que $\mathcal{X} \subset \mathbb{R}^n$), ce problème est bien un cas particulier du problème 1, obtenu en posant N égal à n et en choisissant pour fonction $\tilde{g}$ l'identité. Les vecteurs lignes de la matrice M sont les vecteurs w_k, le produit scalaire K est le produit scalaire euclidien et (p, q) est égal à $(2, 2)$. Ce problème fait intervenir Q contraintes de bon classement par élément de l'ensemble d'apprentissage. En conservant les notations de la sous-section précédente, les α_{iy_i} apparaissent donc ici comme de véritables multiplicateurs de Lagrange et non des pseudo-variables. L'expression des vecteurs w_k^* en fonction des variables α_{il}^* est :

$$\forall k \in \{1, \ldots, Q\}\,,\ w_k^* = \sum_{x_i \in k} \sum_{l=1}^{Q} \alpha_{il}^* \Phi(x_i) - \sum_{i=1}^{m} \alpha_{ik}^* \Phi(x_i) = \sum_{i=1}^{m} \left(C\delta_{y_i,k} - \alpha_{ik}^*\right) \Phi(x_i). \tag{2.11}$$

Le problème dual est donné par :

Problème 6 (M-SVM de Crammer et Singer, problème dual (formulation initiale))

$$\min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\{ \sum_{k=1}^{Q} \left(C\delta_{y_i,k} - \alpha_{ik}\right) \left(C\delta_{y_j,k} - \alpha_{jk}\right) \right\} \kappa(x_i, x_j) + \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \delta_{y_i,k} \right\}$$

$$s.c. \left\{ \begin{array}{ll} \alpha_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \leq Q) \\ \sum_{k=1}^{Q} \alpha_{ik} = C, & (1 \leq i \leq m) \end{array} \right. .$$

Notons en premier lieu que les contraintes-égalités impliquent que les Q variables duales associées à un exemple d'apprentissage ne représentent en fait que $Q - 1$ degrés de liberté. L'équation 2.11 nous apprend que l'expression analytique des vecteurs w_k^* est la même que dans le cas de la machine de Weston et Watkins. En conséquence, la contrainte $\sum_{k=1}^{Q} w_k^* = 0$ apparaît ici encore de manière implicite. Ce résultat était de nouveau prévisible, la proposition 1 s'étendant directement à la nouvelle machine. Une autre conséquence de l'identité des expressions des vecteurs w_k^* est le fait que la matrice hessienne du problème 6 est la même que celle du problème 3. En notant $\delta = (\delta_{y_i,k})_{1 \leq i \leq m, 1 \leq k \leq Q}$, la forme standard de la fonction objectif est donc donnée par :

$$J_{\mathrm{CS,d}}(\alpha) = \frac{1}{2} \alpha^T H_{\mathrm{WW}} \alpha + \delta^T \alpha. \tag{2.12}$$

Un changement de variables permet de reformuler le problème 6 sous une forme algébrique plus compacte. Soient $\alpha_{i.} = (\alpha_{ik})_{1 \leq k \leq Q}$, $\delta_{y_i,.} = (\delta_{y_i,k})_{1 \leq k \leq Q}$ et $\tau_{i.} = (\tau_{ik})_{1 \leq k \leq Q} = C\delta_{y_i,.} - \alpha_{i.}$, pour i allant de 1 à m. Soit $\tau = (\tau_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$. On obtient alors :

Problème 7 (M-SVM de Crammer et Singer, problème dual (formulation finale))

$$\min_{\tau} \left\{ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \tau_{i.}^T \tau_{j.} \kappa(x_i, x_j) - \sum_{i=1}^{m} \tau_{i.}^T \delta_{y_i,.} \right\}$$

$$s.c. \left\{ \begin{array}{ll} \tau_{ik} \leq C\delta_{y_i,k}, & (1 \leq i \leq m), (1 \leq k \leq Q) \\ 1_Q^T \tau_{i.} = 0, & (1 \leq i \leq m) \end{array} \right. .$$

Crammer et Singer voient deux avantages à l'emploi de cette M-SVM. D'une part, l'utilisation d'une unique variable d'écart par exemple d'apprentissage permet de concentrer les efforts sur l'optimisation de la quantité jouant, pour un couple (x_i, y_i) donné, le rôle central : la différence entre $h_{y_i}(x_i)$ et $\max_{k \neq y_i} h_k(x_i)$ (et non chacun des $h_k(x_i)$ considéré séparément). D'autre part, l'apprentissage se prête à la mise en œuvre d'une méthode de décomposition particulièrement efficace. La raison en est que

les contraintes du problème 7 font intervenir les différents exemples d'apprentissage de manière indépendante. Il est donc possible de passer d'une solution réalisable de ce problème à une autre de meilleure qualité en effectuant l'optimisation par rapport à un seul point. Cette optimisation correspond à la résolution d'un nouveau problème de programmation quadratique, à Q variables et $Q+1$ contraintes (les contraintes du problème 7 relatives au point considéré). L'idée de cette décomposition est à rapprocher de celle se trouvant à la base de l'algorithme "sequential minimal optimization" (SMO) [173], qui, elle, fait intervenir deux points. Les auteurs considèrent l'application d'un algorithme de point fixe afin de résoudre les problèmes quadratiques partiels, et proposent plusieurs méthodes pour choisir les points par rapport auxquels effectuer l'optimisation. Il convient de souligner que si la possibilité d'appliquer la décomposition précitée est appréciable, elle est au moins autant liée au choix de travailler simplement dans $\bar{\mathcal{H}}$, ce qui permet d'éviter de faire apparaître certaines contraintes-égalités, qu'aux propriétés intrinsèques du problème 5 (à la nature de la fonction de perte ℓ_{CS}).

2.4.1.3 Modèle de Lee et co-auteurs

Pour un problème d'apprentissage (une mesure de probabilité P) donné, une méthode de sélection de fonction est dite consistante si le risque de la fonction qu'elle retourne converge en probabilité vers le risque de Bayes lorsque la taille de l'échantillon d'apprentissage tend vers l'infini. Cette définition s'étend au cas où le classifieur de Bayes n'appartient pas à la famille de fonctions considérée en substituant au risque de Bayes le plus petit risque possible sur cette famille, c'est-à-dire le risque de Bayes augmenté de l'erreur d'approximation [42]. La consistance est dite forte si la convergence est presque sûre. Elle est universelle si la convergence a lieu quelle que soit P. Le lecteur souhaitant obtenir plus d'informations sur ce sujet pourra par exemple se reporter à [87, 100]. La question de la consistance du principe inductif de minimisation empirique du risque (ERM) se situe au cœur même de la théorie statistique de l'apprentissage, à travers les trois étapes importantes (*milestones*) atteintes par Vapnik entre la fin des années 60 et le début des années 90 (voir en particulier le chapitre 3 de [229]). On sait que les SVM n'appliquent pas le principe de minimisation empirique du risque mais le principe de minimisation structurelle du risque. Plus précisément, leurs algorithmes d'apprentissage consistent en la minimisation d'une fonction objectif convexe faisant intervenir un terme empirique différent de la fréquence des erreurs et un terme de pénalisation (dépendant de la norme induite par le noyau). Elles appellent donc la formulation de résultats de consistance dédiés. La littérature propose plusieurs résultats de consistance universelle pour les SVM biclasses (voir en particulier [209, 145, 22, 210]). Ceux-ci reposent sur des hypothèses faibles concernant la nature de la fonction de perte ℓ_{SVM}, du noyau, du terme de pénalisation et de la décroissance du coefficient de pénalisation λ avec la taille de l'échantillon d'apprentissage. Le résultat de base est que l'apprentissage de la SVM standard se concentre sur la détermination des x vérifiant $P(+1|x) = P(-1|x) = \frac{1}{2}$. Plusieurs auteurs ont souligné le fait qu'à l'inverse, l'étude de la consistance des algorithmes d'apprentissage des M-SVM soulève des problèmes originaux. On pourra se référer sur ce sujet à [248, 215]. Il apparaît en particulier que les deux M-SVM décrites ci-dessus, pour naturels que soient leurs principes, étendant directement celui de la SVM biclasse, ne convergent pas vers le classifieur de Bayes lorsque la taille de l'ensemble d'apprentissage tend vers l'infini.

Dans [135, 137], Lee et ses co-auteurs proposent un modèle de M-SVM conçu de manière que le principe inductif sur lequel il repose soit universellement consistant. Cette propriété est obtenue en choisissant un codage approprié pour les catégories. Il correspond au choix de représentants des catégories situés sur les sommets d'un polytope régulier de $\mathbb{R}^{Q-1}$ à Q sommets centré sur l'origine de $\mathbb{R}^Q$. Plus précisément, la catégorie d'indice k est représentée par le vecteur de $\mathbb{R}^Q$ dont la l-ième composante est égale à 1 si $l = k$ et à $-\frac{1}{Q-1}$ dans le cas contraire. Le polytope en question est donc un simplexe. Le cas où $Q = 3$ est illustré par la figure 2.4.

Naturellement, cette répartition optimale de points est bien connue en théorie de l'apprentissage. Elle est en particulier utilisée par Vapnik dans [227] afin d'établir une variante du lemme de Sauer-Shelah (voir la section 2.5.1.2). Elle est également à la base de l'extension multiclasse du principe d'alignement noyau-cible proposée dans [232] et que nous avons appliquée à la prédiction de la structure secondaire des protéines (voir la section 3.2). Prolongeant la propriété de sommation à 0 des représentants, les auteurs introduisent, cette fois explicitement, la contrainte standard $\sum_{k=1}^{Q} h_k = 0$. On voit ainsi apparaître pour

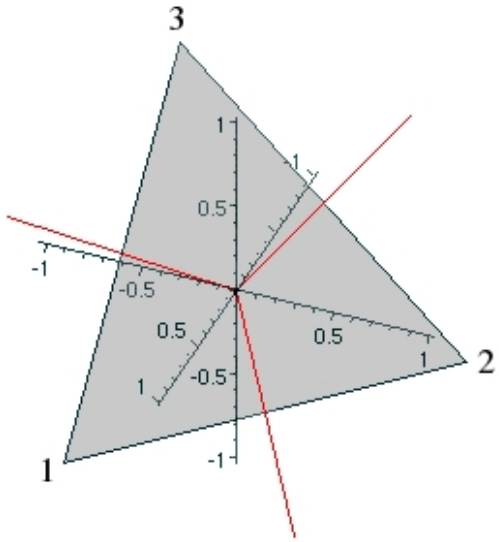

FIG. 2.4 – M-SVM de Lee et ses co-auteurs pour un problème à trois catégories. Les représentants des catégories sont situés sur les sommets d'un triangle équilatéral centré sur l'origine de $\mathbb{R}^3$. Les frontières de décision optimales apparaissent en rouge.

fonction de perte

$$\ell_{\mathrm{LLW}}\left(y, h(x)\right) = \sum_{k \neq y} \left(h_k(x) + \frac{1}{Q-1} \right)_+$$

avec $d_{\mathrm{LLW}} = \frac{Q}{Q-1}$ et l'apprentissage consiste donc à rechercher dans $\mathcal{H}$ la fonction minimisant la fonction objectif suivante :

$$J_{\mathrm{LLW}}\left(h\right) = \sum_{i=1}^{m} \sum_{k \neq y_i} \left(h_k(x_i) + \frac{1}{Q-1} \right)_+ + \lambda \sum_{k=1}^{Q} \|w_k\|^2,$$

sous la contrainte $\sum_{k=1}^{Q} h_k = 0$. Le problème primal d'apprentissage prend ainsi la forme du problème de programmation quadratique suivant :

PROBLÈME 8 (M-SVM de Lee et co-auteurs, problème primal)

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k \neq y_i} \xi_{ik} \right\}$$

$$s.c. \left\{ \begin{array}{ll} \langle w_k, \Phi(x_i) \rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{k=1}^{Q} w_k = 0, \ \ \sum_{k=1}^{Q} b_k = 0 & \end{array} \right. .$$

Il convient de remarquer qu'ici, il est bien nécessaire de prendre en compte explicitement les contraintes $\sum_{k=1}^{Q} w_k = 0$ et $\sum_{k=1}^{Q} b_k = 0$ (la seconde ayant une importance relative, puisqu'elle peut toujours être satisfaite a posteriori). Le raisonnement de la proposition 1 n'est pas transposable, car les contraintes de bon classement ne font plus intervenir les fonctions composantes sous la forme de différences. Cependant, ceci ne complique pas l'expression du problème dual, dont les variables sont, comme dans le cas des deux M-SVM déjà rencontrées, les multiplicateurs de Lagrange associés aux contraintes de bon classement.

En notant $G = (\kappa(x_i, x_j))_{1\leq i,j\leq m}$ la matrice de Gram, $\alpha_{.k} = (\alpha_{ik})_{1\leq i\leq m}$, pour k allant de 1 à Q et $\bar{\alpha} = \frac{1}{Q}\sum_{k=1}^{Q}\alpha_{.k}$, on obtient en effet :

$$\forall k \in \{1, \ldots, Q\}\,,\ w_k^* = \sum_{i=1}^{m}\left(\frac{1}{Q}\sum_{l=1}^{Q}\alpha_{il}^* - \alpha_{ik}^*\right)\Phi(x_i)$$

(on vérifie que l'on a bien $\sum_{k=1}^{Q} w_k^* = 0$). Le problème dual est le suivant :

Problème 9 (M-SVM de Lee et co-auteurs, problème dual)

$$\min_{\alpha}\left\{\frac{1}{2}\sum_{k=1}^{Q}(\alpha_{.k} - \bar{\alpha})^T G\,(\alpha_{.k} - \bar{\alpha}) - \frac{1}{Q-1}1_{Qm}^T\alpha\right\}$$

$$s.c.\left\{\begin{array}{ll} 0 \leq \alpha_{ik} \leq C, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ 1_m^T(\alpha_{.k} - \bar{\alpha}) = 0, & (1 \leq k \leq Q) \end{array}\right. .$$

En notant $H_{\mathrm{LLW}} = \left(h_{ik,jl}^{(\mathrm{LLW})}\right)_{1\leq i,j\leq m, 1\leq k,l\leq Q}$ la matrice de $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ de terme général

$$h_{ik,jl}^{(\mathrm{LLW})} = \left(\delta_{k,l} - \frac{1}{Q}\right)\kappa(x_i, x_j),$$

on obtient la fonction objectif de ce problème sous sa forme standard, c'est-à-dire :

$$J_{\mathrm{LLW,d}}(\alpha) = \frac{1}{2}\alpha^T H_{\mathrm{LLW}}\alpha - \frac{1}{Q-1}1_{Qm}^T\alpha.$$

Pour cette machine, les auteurs établissent d'une part un théorème de représentation exprimant le fait que la solution du problème 8 est de la forme (2.4) (théorème 1), et d'autre part un résultat de consistance universelle (lemme 1). Ils discutent de plus, de manière plus sommaire que Zhang, Tewari et Bartlett, les liens existant entre la machine de Weston et Watkins et le classifieur de Bayes, en fournissant, avec le lemme 2, une preuve que pour cette machine, la fonction h minimisant $\mathbb{E}\left[\ell_{\mathrm{WW}}(Y, h(X))\,|X = x\right]$ peut ne pas être associée à une fonction de décision correspondant en x au classifieur de Bayes (i.e. peut ne pas affecter x à la catégorie k vérifiant $P(k|x) > \max_{l\neq k} P(l|x)$). La portée de ce lemme est cependant limitée, comme le prouve le résultat suivant.

Proposition 2 *x étant un élément quelconque de $\mathcal{X}$ et ℓ_{WW} la fonction de perte associée à la M-SVM de Weston et Watkins, la fonction h minimisant sur $\mathcal{H}$ $\mathbb{E}\left[\ell_{WW}(Y, h(X))\,|X = x\right]$ vérifie nécessairement :*

$$P(k|x) = \max_{1\leq l\leq Q} P(l|x) \Longrightarrow h_k(x) = \max_{1\leq l\leq Q} h_l(x).$$

Preuve Considérons un x tel que $P(k|x) > \max_{l\neq k} P(l|x)$ et une fonction $h^{(1)}$ de $\mathcal{H}$ telle que $h_k^{(1)}(x) < \max_{1\leq l\leq Q} h_l^{(1)}(x)$. On établit que $h^{(1)}$ ne minimise pas $\mathbb{E}\left[\ell_{\mathrm{WW}}(Y, h(X))\,|X = x\right]$ sur $\mathcal{H}$. Sans perte de généralité, on fait l'hypothèse $k = 1$ et $h_2^{(1)}(x) = \max_{1\leq l\leq Q} h_l^{(1)}(x)$. Considérons la fonction $h^{(2)}$ déduite de $h^{(1)}$ de la manière suivante : $h_1^{(2)}(x) = h_2^{(1)}(x)$, $h_2^{(2)}(x) = h_1^{(1)}(x)$ et $h_k^{(2)}(x) = h_k^{(1)}(x)$ pour k allant de 3 à Q.

$$\mathbb{E}\left[\ell_{\mathrm{WW}}\left(Y, h^{(1)}(X)\right)|X = x\right] - \mathbb{E}\left[\ell_{\mathrm{WW}}\left(Y, h^{(2)}(X)\right)|X = x\right] =$$

$$\sum_{k\neq 1}\left(1 - h_1^{(1)}(x) + h_k^{(1)}(x)\right)_+ P(1|x) + \sum_{k\neq 2}\left(1 - h_2^{(1)}(x) + h_k^{(1)}(x)\right)_+ P(2|x)$$

$$-\sum_{k\neq 2}\left(1 - h_2^{(1)}(x) + h_k^{(1)}(x)\right)_+ P(1|x) - \sum_{k\neq 1}\left(1 - h_1^{(1)}(x) + h_k^{(1)}(x)\right)_+ P(2|x) =$$

$$\left(\sum_{k\neq 1}\left(1-h_1^{(1)}(x)+h_k^{(1)}(x)\right)_+-\sum_{k\neq 2}\left(1-h_2^{(1)}(x)+h_k^{(1)}(x)\right)_+\right)(P(1|x)-P(2|x))>0.$$

■

On remarque ainsi que la seule faiblesse susceptible de poser un problème est le fait que l'on a uniquement $h_k(x)=\max_{1\leq l\leq Q}h_l(x)$ et pas $h_k(x)>\max_{l\neq k}h_l(x)$, si bien que des cas d'ex æquo sources d'indéterminations peuvent a priori survenir.

2.4.1.4 M-SVM "à coût quadratique"

Les sous-sections précédentes ont donné une description des trois M-SVM proposées à ce jour. Pour toutes ces machines, la contribution des variables d'écart à la fonction objectif est linéaire. Notons ξ le vecteur de ces variables. On a $\xi=(\xi_{ik})_{1\leq i\leq m,1\leq k\leq Q}$, avec $\xi_{iy_i}=0$, $(1\leq i\leq m)$ dans le cas des M-SVM de Weston et Watkins et Lee et co-auteurs, $\xi=(\xi_i)_{1\leq i\leq m}$, dans le cas de la M-SVM de Crammer et Singer. Ces notations étant posées, la contribution des variables d'écart à la fonction objectif est dans tous les cas égale à $C\|\xi\|_1$ (en conservant à l'esprit le fait qu'elles sont positives). La SVM biclasse standard possède une variante dite "de norme 2", pour laquelle la contribution empirique à la fonction objectif est quadratique. Il s'agit précisément de $C\|\xi\|_2^2$. Avec cette variante, la contrainte de positivité des variables d'écart devient superflue (voir en particulier le chapitre 7 de [201]). Ce modèle possède une propriété utile : il rend possible, au moyen d'un changement de noyau adéquat, la reformulation de l'algorithme d'apprentissage d'une SVM à marge douce comme l'algorithme d'apprentissage d'une SVM à marge dure. Un avantage immédiat de cette propriété est qu'elle permet l'utilisation de bornes "rayon-marge" (voir la section 2.7.1) afin de déterminer la valeur de la constante de marge douce, C. Il apparaît donc intéressant d'étudier les propriétés des M-SVM "de norme 2", de manière en particulier à déterminer si la contribution du terme empirique à la matrice hessienne peut à nouveau se reformuler comme un changement de noyau. Si l'on utilise simplement le carré de la norme euclidienne, la réponse sur ce dernier point est négative. Le tableau 2.1 résume la situation.

Modèle	Hessien de base	Terme empirique	Modification du hessien
SVM	$y_iy_j\kappa(x_i,x_j)$	$C\sum_{i=1}^m\xi_i^2$	$+\frac{1}{2C}y_iy_j\delta_{i,j}$
M-SVM WW	$\left(\delta_{y_i,y_j}-\delta_{y_i,l}-\delta_{y_j,k}+\delta_{k,l}\right)\kappa(x_i,x_j)$	$C\sum_{i=1}^m\sum_{k\neq y_i}\xi_{ik}^2$	$+\frac{1}{2C}\delta_{i,j}\delta_{k,l}$
M-SVM CS	$\left(\delta_{y_i,y_j}-\delta_{y_i,l}-\delta_{y_j,k}+\delta_{k,l}\right)\kappa(x_i,x_j)$	$C\sum_{i=1}^m\xi_i^2$	$+\frac{1}{2C}\delta_{i,j}$
M-SVM LLW	$\left(\delta_{k,l}-\frac{1}{Q}\right)\kappa(x_i,x_j)$	$C\sum_{i=1}^m\sum_{k\neq y_i}\xi_{ik}^2$	$+\frac{1}{2C}\delta_{i,j}\delta_{k,l}$

TAB. 2.1 – Hessiens de la SVM et des M-SVM dans leurs versions "de norme 1" et "de norme 2".

On observe que dans le cas de la SVM biclasse, les deux contributions au terme général du hessien, celle provenant de $\|w\|^2$, $y_iy_j\kappa(x_i,x_j)$, et celle provenant de $C\sum_{i=1}^m\xi_i^2$, $\frac{1}{2C}y_iy_j\delta_{i,j}$, s'écrivent comme le produit de la valeur d'une fonction noyau par un terme multiplicatif commun : y_iy_j. Il est donc possible de mettre ce terme en facteur des deux noyaux. La somme de deux noyaux étant encore un noyau, comme on s'en aperçoit en considérant l'implication

$$\left(\sum_{i=1}^n\sum_{j=1}^n a_ia_j\kappa_1(x_i,x_j)\geq 0\ \wedge\ \sum_{i=1}^n\sum_{j=1}^n a_ia_j\kappa_2(x_i,x_j)\geq 0\right)$$

$$\Longrightarrow\sum_{i=1}^n\sum_{j=1}^n a_ia_j\left(\kappa_1(x_i,x_j)+\kappa_2(x_i,x_j)\right)\geq 0,$$

tout se passe comme si l'on mettait en œuvre une machine à marge dure munie du noyau somme. Dans le cas d'espèce, ce noyau, $\tilde{\kappa}$, est défini par :

$$\forall(i,j)\in\{1,\ldots,m\}^2\ ,\ \tilde{\kappa}(x_i,x_j)=\kappa(x_i,x_j)+\frac{1}{2C}\delta_{i,j}. \tag{2.13}$$

Dans le cas des M-SVM, les deux facteurs sont toujours différents, ce qui empêche de faire apparaître un noyau somme. La question de l'existence d'une matrice M_ξ symétrique semi-définie positive telle qu'en utilisant pour fonction objectif de l'une des SVM multiclasses $\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\xi^T M_\xi \xi$, on obtienne la transition recherchée vers le cas à marge dure, a trouvé une réponse positive dans [161]. Elle correspond, pour la M-SVM de Lee et co-auteurs, au choix d'une matrice $M_\xi = \left(m_{ik,jl}^{(\xi)}\right)_{1\leq i,j\leq m, 1\leq k,l\leq Q}$ telle que

$$m_{ik,jl}^{(\xi)} = \left(\delta_{k,l} - \frac{1}{Q}\right)\delta_{i,j}. \tag{2.14}$$

On obtient alors pour problème primal d'apprentissage :

Problème 10 (M-SVM de Lee et co-auteurs "à coût quadratique", problème primal)

$$\min_{h\in\mathcal{H}}\left\{\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\xi^T M_\xi \xi\right\}$$

$$s.c.\left\{\begin{array}{ll} \langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}, & (1\leq i\leq m), (1\leq k\neq y_i\leq Q) \\ \sum_{k=1}^{Q} w_k = 0, \ \sum_{k=1}^{Q} b_k = 0 & \end{array}\right. .$$

Le dual de Wolfe a pour expression :

Problème 11 (M-SVM de Lee et co-auteurs "à coût quadratique", problème dual)

$$\min_{\alpha}\left\{\frac{1}{2}\alpha^T \tilde{H}_{LLW}\alpha - \frac{1}{Q-1}1_{Qm}^T\alpha\right\}$$

$$s.c.\left\{\begin{array}{ll} \alpha_{ik}\geq 0, & (1\leq i\leq m), (1\leq k\neq y_i\leq Q) \\ 1_m^T(\alpha_{.k} - \bar{\alpha}) = 0, & (1\leq k\leq Q) \end{array}\right. .$$

où la matrice $\tilde{H}_{\mathrm{LLW}}$ se déduit de H_{LLW} en remplaçant le noyau κ par le noyau $\tilde{\kappa}$ défini sur l'ensemble d'apprentissage par l'équation 2.13.

Il est important de remarquer que le fait que Ker (M_ξ) soit un sous-espace de $\mathbb{R}^{Qm}$ de dimension m ne pose pas de difficulté. En effet, le vecteur ξ appartient à un sous-espace de $\mathbb{R}^{Qm}$ défini par les contraintes $\xi_{iy_i} = 0$, $(1\leq i\leq m)$. Or, l'intersection de ce sous-espace avec le noyau Ker (M_ξ) est réduite au vecteur nul. Tout se passe donc en pratique comme si M_ξ était symétrique définie positive. Notons encore que la démarche conduisant à la formulation de l'équation 2.14 est principalement calculatoire. Du point de vue de la qualité du classement, il est plus difficile de justifier l'utilisation de la matrice M_ξ que celle des normes ℓ_1 et ℓ_2.

Au-delà de la question de la borne "rayon-marge", nous voyons à cette étude comparative un intérêt plus grand, qui est celui de mettre en évidence une des difficultés nouvelles induites par le passage au cas multiclasse. Nombreux sont les résultats du cas biclasse reposant directement sur le choix de $\{-1, 1\}$ plutôt que $\{1, 2\}$ comme ensemble des "étiquettes" des catégories. Citons en particulier certains résultats fondés sur l'utilisation de moyennes de Rademacher (voir la section 2.5.1.4), ou les simplifications qu'obtient Chapelle dans la mise en œuvre de son algorithme d'apprentissage construit autour de la méthode de Newton-Raphson (voir la section 2.6.1). Dans le cas multiclasse, l'impossibilité d'étendre cette idée fait apparaître dans les formules des termes qui sont beaucoup plus difficiles à manipuler. Nous retrouverons cette difficulté dans la section 2.5 consacrée aux performances en généralisation.

2.4.2 Des SVM multiclasses qui ne sont pas des M-SVM

Il n'existe à notre connaissance que trois SVM multiclasses qui ne soient pas des M-SVM.

2.4.2.1 LS-SVM multiclasse

Comme la SVM "de norme 2", La "SVM des moindres carrés" (LS-SVM) [213] utilise pour contribution empirique à la fonction objectif le carré de la norme euclidienne du vecteur ξ (les auteurs évoquent à ce propos une similitude avec la régression "ridge", similitude qui mériterait d'être précisée). La seule différence entre les deux machines est que la seconde reprend les contraintes de bon classement de la première sous la forme de contraintes-égalités. Ainsi, l'apprentissage correspond à la résolution du problème d'optimisation suivant :

Problème 12 (Apprentissage de la LS-SVM)

$$\min_{\tilde{h}\in\tilde{\mathcal{H}}} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^2 \right\}$$

$$s.c.\ \ y_i\left(\langle w, \Phi(x_i)\rangle + b\right) = 1 - \xi_i,\ (1 \leq i \leq m),$$

qui se réduit à la résolution d'un système linéaire.

L'extension multiclasse de cette machine est décrite dans [214]. Elle repose sur l'emploi d'une matrice de mots codes $M = (m_{kl}) \in \mathcal{M}_{Q,N}(\{-1,1\})$. La classe de fonctions utilisée est la même que dans le cas des M-SVM, à cette différence près que le nombre de fonctions composantes est égal à N (taille des mots codes) et non Q et que la somme de ces fonctions n'est plus supposée être nulle. L'algorithme d'apprentissage correspond au problème de programmation convexe suivant :

Problème 13 (Apprentissage de la LS-SVM multiclasse)

$$\min_{h\in\mathcal{H}} \left\{ \frac{1}{2}\sum_{k=1}^{N} \|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k=1}^{N} \xi_{ik}^2 \right\}$$

$$s.c.\ \ m_{y_i k}\left(\langle w_k, \Phi(x_i)\rangle + b_k\right) = 1 - \xi_{ik},\ (1 \leq i \leq m), (1 \leq k \leq N).$$

Il se réduit encore à la résolution d'un système linéaire. Les auteurs ne donnent pas la formule de décodage utilisée pour affecter les descriptions aux catégories en fonction de la valeur du vecteur $(h_k(x))_{1\leq k\leq N}$. Ils ne discutent pas davantage la question du choix de la matrice M. Le lecteur en est réduit à constater qu'ils traitent un problème à quatre catégories en utilisant des fonctions à deux composantes, avec pour codage des catégories l'ensemble des vecteurs de $\{-1,1\}^2$.

2.4.2.2 Modèle d'Anguita et co-auteurs

L'architecture et l'algorithme d'apprentissage du modèle introduit dans [8, 9] sont exactement les mêmes que ceux de la SVM biclasse. La prise en compte du caractère multiclasse de la tâche à effectuer résulte d'une simple reformulation des données. Ici encore, c'est une matrice de mots codes qui est utilisée. Cette matrice correspond au codage binaire canonique des catégories. On a donc $M = (m_{kl}) \in \mathcal{M}_{Q,Q}(\{-1,1\})$ avec $m_{kl} = 2\delta_{k,l} - 1$. A chaque exemple d'apprentissage (x_i, y_i) viennent se substituer Q nouveaux exemples, $(x_{(i-1)Q+k}, y_{(i-1)Q+k})$, pour k allant de 1 à Q. $x_{(i-1)Q+k}$ est la concaténation de x_i et de $M_{.k}$, et $y_{(i-1)Q+k}$ est égal à 1 si $k = y_i$, et à -1 dans le cas contraire. Les données de test engendrent également Q vecteurs suivant le même principe. Notons $\left(x^{(k)}\right)_{1\leq k\leq Q}$ la suite des concaténations de x et des vecteurs $M_{.k}$. x est affecté à la catégorie k^* vérifiant :

$$k^* = \operatorname*{argmax}_k \left\{ \langle w, \Phi(x^{(k)})\rangle + b \right\}.$$

Deux remarques s'imposent. Tout d'abord, le biais b n'intervient pas dans la décision et son utilité n'apparaît donc pas clairement. Ensuite, le choix du noyau revêt ici une importance accrue. Ainsi, il n'est

pas possible d'utiliser cette SVM avec le noyau linéaire (le produit scalaire euclidien dans l'espace où vivent les données reformulées). En effet, dans cette configuration, la valeur de la fonction de décision calculée en x serait indépendante de x. Si le principe de cette machine est nettement plus simple que celui des M-SVM, son apprentissage ne l'est pas autant. On sait en effet que le problème de programmation quadratique à résoudre pour entraîner une SVM biclasse fait intervenir, dans sa formulation duale, une variable par exemple d'apprentissage. Les transformations opérées ici spécifient donc un problème de programmation quadratique à Qm variables, c'est-à-dire autant que dans le cas de la M-SVM de Crammer et Singer et m de plus que dans le cas des M-SVM de Weston et Watkins et Lee et ses co-auteurs. Naturellement, la différence favorable est le fait de conserver une contrainte-égalité unique, ce qui joue souvent un rôle majeur dans la résolution. L'exemple le plus significatif sur ce point est l'algorithme SMO et ceux qui s'appuient sur lui, comme LASVM [36]. Il est possible de diminuer le nombre de paramètres de cette machine en choisissant, comme dans le cas de la LS-SVM multiclasse, une matrice de mots codes différente. N peut alors au mieux prendre pour valeur le plus petit entier vérifiant $2^N \geq Q$. Cependant, les performances expérimentales s'en trouvent dégradées.

2.4.2.3 Modèle de Tsochantaridis et co-auteurs

La description du modèle proposé par Anguita et ses co-auteurs fournit une transition naturelle pour présenter cette machine. Introduite dans [218] (voir aussi [219]), elle s'appuie sur un cadre théorique différent de celui que nous avons décrit à la section 2.2.1. Dans ce cadre, l'ensemble $\mathcal{Y}$ n'est plus réduit à un ensemble d'indices de catégories, mais représente un ensemble d'éléments structurés (séquences, arbres, treillis, graphes...) qui peut être infini dénombrable [16]. La classe de fonctions de décision $\mathcal{F}$ utilisée, formellement paramétrée par w ($\mathcal{F} = \{f_w\}$), est définie de $\mathcal{X}$ dans $\mathcal{Y}$ à partir d'une classe $\mathcal{G}$ de fonctions g_w de même paramétrage et définies de $\mathcal{X} \times \mathcal{Y}$ dans $\mathbb{R}$. Le lien entre les deux familles est donné par l'équation suivante :

$$f_w(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} g_w(x, y).$$

La nature particulière de l'ensemble $\mathcal{Y}$ appelle un changement de la fonction de perte et donc du risque à minimiser. En notant $\Delta_{\mathcal{Y}}$ la fonction de perte de $\mathcal{Y} \times \mathcal{Y}$ dans $\mathbb{R}_+$, on obtient pour risque :

$$R_{\Delta_{\mathcal{Y}}}(f_w) = \int_{\mathcal{X} \times \mathcal{Y}} \Delta_{\mathcal{Y}}\left(y, f_w(x)\right) dP(x, y).$$

Les auteurs considèrent plus particulièrement le cas où $\mathcal{G}$ est un modèle linéaire opérant dans un espace de représentation défini par $\Phi_{\mathcal{X},\mathcal{Y}}$, c'est-à-dire une famille de fonctions du type :

$$g_w(x, y) = \langle w, \Phi_{\mathcal{X},\mathcal{Y}}(x, y) \rangle.$$

Ceci leur permet de définir un ensemble de SVM multiclasses dont les contraintes de bon classement constituent des variantes de celles de la M-SVM de Crammer et Singer (une contrainte par exemple), variantes prenant en compte la fonction de perte $\Delta_{\mathcal{Y}}$. Dans tous les cas, le problème d'apprentissage demeure un problème de programmation quadratique que les auteurs proposent de résoudre en utilisant une méthode des plans sécants (voir par exemple le chapitre 5 de [160]). Cette méthode possède l'avantage d'assurer la convergence tout en nécessitant uniquement le calcul et le stockage d'une petite partie du gradient. La M-SVM de Crammer et Singer peut elle-même être obtenue dans ce cadre, en posant w égal à la matrice de vecteurs lignes w_k et $\Phi_{\mathcal{X},\mathcal{Y}}(x, y) = \Phi(x) \otimes (\delta_{y,k})_{1 \leq k \leq Q}$, où $\otimes$ désigne le produit tensoriel ou produit de Kronecker.

2.4.3 Discussion

En présentant les principales SVM multiclasses et M-SVM, nous avons évoqué les motivations de leurs auteurs. Celles-ci sont de natures très diverses, les considérations liées à la consistance côtoyant celles liées au temps de résolution du problème d'optimisation définissant l'apprentissage, ou à l'extension des notions de marges biclasses. Cependant, elles reposent globalement sur une même idée, celle de tirer parti des résultats et de l'expertise disponibles dans le cas biclasse. Dans ces conditions, on aurait pu s'attendre

à ce que les premiers articles traitant de SVM multiclasses démontrent une borne sur les performances en généralisation des systèmes affines multivariés faisant intervenir une notion de dimension de Vapnik-Chervonenkis (VC) généralisée. Nous savons déjà que cela ne fut pas le cas. Les SVM multiclasses furent développées sans lien explicite avec le principe inductif de minimisation structurelle du risque. La section suivante est consacrée aux différentes manières d'établir ce lien a posteriori.

2.5 Bornes sur le risque

Dans les sections précédentes, nous avons introduit les méthodes de décomposition avant les SVM multiclasses, respectant en cela globalement l'ordre chronologique. L'étude des performances en généralisation des différents modèles nous invite à retenir l'ordre inverse, de manière à faire apparaître la théorie multiclasse comme une extension naturelle de la théorie biclasse. Cela permet en particulier d'organiser l'introduction des différents schémas de construction des bornes et les concepts associés suivant la même progression.

2.5.1 M-SVM

Dans cette section, nous considérons deux types de bornes : des lois fortes des grands nombres uniformes classiques, fondées sur des extensions du théorème de Glivenko-Cantelli, et une borne s'appuyant sur la notion de moyenne de Rademacher. En suivant le cheminement standard de construction des risques garantis (voir par exemple [71], le chapitre 2 de [176], le chapitre 12 de [65] ou le chapitre 4 de [229]), Bartlett a démontré dans [19] une borne (correcte aux constantes près), dédiée aux séparateurs biclasses à grande marge. Cette borne s'applique également aux SVM. Nous avons fondé la théorie VC des M-SVM sur l'extension de cette borne au cas multiclasse.

2.5.1.1 Résultat de convergence uniforme de base

Afin de formuler le théorème correspondant, nous devons au préalable introduire quelques définitions et notations. Il s'agit de caractériser le pouvoir discriminant, au sens de la marge multiclasse donnée par la définition 3, des fonctions de la famille $\mathcal{G}$. Cela passe par l'application d'un ensemble d'"opérateurs de marge" destinés à extraire, dans différents contextes, l'information pertinente pour réaliser la caractérisation.

Définition 8 (Opérateur Δ, définition 6 dans [75]) *$\mathcal{G}$ étant une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ et $\mathcal{M}$ la fonction de la définition 2, soit Δ l'opérateur sur $\mathcal{G}$ défini par :*

$$\begin{aligned} \Delta \ : \ & \mathcal{G} \longrightarrow \Delta\mathcal{G} \\ & g \mapsto \Delta g = (\Delta g_k)_{1\leq k\leq Q} \end{aligned}$$

$$\forall x \in \mathcal{X},\ \Delta g(x) = (\mathcal{M}\,(g(x),k))_{1\leq k\leq Q}\,.$$

Dans un souci de simplification des notations, nous avons écrit Δg_k au lieu de $(\Delta g)_k$. Dans la suite, cette simplification sera réitérée implicitement pour d'autres opérateurs.

Définition 9 (Opérateur Δ^*, définition 7 dans [95]) *$\mathcal{G}$ étant une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ et $\mathcal{M}$ la fonction de la définition 2, soit Δ^* l'opérateur sur $\mathcal{G}$ défini par :*

$$\begin{aligned} \Delta^* \ : \ & \mathcal{G} \longrightarrow \Delta^*\mathcal{G} \\ & g \mapsto \Delta^* g = (\Delta^* g_k)_{1\leq k\leq Q} \end{aligned}$$

$$\forall x \in \mathcal{X},\ \Delta^* g(x) = (\max\,(\Delta g_k(x), -\mathcal{M}\,(g(x),.)))_{1\leq k\leq Q}\,.$$

Remarque 1 *Si $\mathcal{M}\,(g(x),.) > 0$, $\Delta g(x)$ possède une unique composante strictement positive, dans le cas contraire il n'en a aucune. Considérons le premier cas et posons $k^* = \operatorname{argmax}_{1\leq k\leq Q} \Delta g_k(x) = \operatorname{argmax}_{1\leq k\leq Q} g_k(x)$ ($\Delta g_{k^*}(x) = \mathcal{M}\,(g(x),.)$).*

$$\forall x \in \mathcal{X},\ \left\{ \begin{array}{ll} si\ \mathcal{M}\,(g(x),.) > 0, & \Delta^* g(x) = ((2\delta_{k,k^*} - 1)\,\Delta g_{k^*}(x))_{1\leq k\leq Q} \\ si\ \mathcal{M}\,(g(x),.) = 0, & \Delta^* g(x) = 0 \end{array} \right. .$$

On remarque que l'opérateur Δ, dont l'application permet de restituer, à une constante additive près, l'intégralité du comportement de son argument, fournit plus d'informations que l'opérateur Δ^*. Ce dernier opérateur permet simplement d'identifier la sortie la plus haute, ainsi que la différence entre cette sortie et la seconde plus haute (en supposant qu'il n'y ait pas d'ex æquo). En pratique, il fournit à peine plus d'informations que ce qui est nécessaire et suffisant pour calculer le risque à marge (voir la définition 10 ci-dessous). Considérons en particulier le cas d'un exemple x pour lequel on connaît la catégorie, y. Alors la valeur de la marge multiclasse $\mathcal{M}(g(x), y)$ est fournie par $\Delta g_y(x)$, tandis qu'a priori, aucune des composantes de $\Delta^* g(x)$ ne lui est égale. Cette différence disparaît lorsque $Q = 2$, car les deux opérateurs s'avèrent alors être identiques. En reprenant les notations de la section 2.2.2, on a encore dans ce cas $\tilde{h}(x) = \Delta h_1(x) = -\Delta h_2(x)$. Dans ce qui suit, $\Delta^\#$ remplace Δ et Δ^* dans les formules vérifiées par les deux opérateurs de marge. Naturellement, la première d'entre elles est une reformulation de la définition du risque.

Proposition 3 *Soit $\mathcal{G}$ une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$. Le risque d'une fonction g de $\mathcal{G}$, donné par la définition 1, peut être reformulé comme suit :*

$$R(g) = \mathbb{E}\left[\mathbb{1}_{\{\Delta^\# g_Y(X) \leq 0\}}\right].$$

Ces définitions étant posées, la notion de risque à marge multiclasse s'obtient comme une extension directe de la notion biclasse.

Définition 10 (Risque à marge γ, définition 8 dans [75]) *Soient $\mathcal{G}$ une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ et $\gamma \in \mathbb{R}_+^*$. Le* risque à marge γ *d'une fonction g de $\mathcal{G}$ est défini par :*

$$R_\gamma(g) = \mathbb{E}\left[\mathbb{1}_{\{\Delta^\# g_Y(X) < \gamma\}}\right].$$

Définition 11 (Risque empirique à marge γ) *Soient $\mathcal{G}$ une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ et $\gamma \in \mathbb{R}_+^*$. g étant un élément de $\mathcal{G}$, son* risque empirique à marge γ *sur un m-échantillon est défini par :*

$$R_{\gamma,m}(g) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{\Delta^\# g_{Y_i}(X_i) < \gamma\}}.$$

En conservant à l'esprit l'objectif de ne retenir que l'information utile pour la tâche à effectuer, on observe que la valeur du risque à marge n'est pas affectée si les fonctions composantes $\Delta^\# g_k$ sont seuillées de manière à prendre leurs valeurs dans l'intervalle $[-\gamma, \gamma]$. A l'inverse, ce seuillage peut s'avérer bénéfique. En effet, certaines méthodes utilisées pour borner la capacité des classes de fonctions exploitent directement le fait que leur codomaine soit un compact de la forme $[-M_\mathcal{G}, M_\mathcal{G}]^Q$. C'est en particulier le cas des méthodes fondées sur une discrétisation telles que celle que nous décrirons dans la section suivante (voir aussi [4] pour le cas biclasse). Ceci conduit à faire usage de l'opérateur suivant.

Définition 12 (Opérateur π_γ, d'après [19]) *$\mathcal{G}$ étant une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ et $\gamma \in \mathbb{R}_+^*$, soit π_γ l'opérateur sur $\mathcal{G}$ défini par :*

$$\begin{aligned} \pi_\gamma \ : \ & \mathcal{G} \longrightarrow \pi_\gamma \mathcal{G} \\ & g \mapsto \pi_\gamma g = (\pi_\gamma g_k)_{1 \leq k \leq Q} \end{aligned}$$

$$\forall x \in \mathcal{X},\ \pi_\gamma g(x) = (\mathit{signe}\,(g_k(x)) \cdot \min\left(|g_k(x)|, \gamma\right))_{1 \leq k \leq Q}.$$

où la fonction signe retourne 1 si son argument est positif ou nul, et -1 dans le cas contraire. En conservant les notations ci-dessus, $\Delta_\gamma^\#$ désigne l'opérateur $\pi_\gamma \circ \Delta^\#$ et $\Delta_\gamma^\# \mathcal{G}$ la famille des fonctions $\Delta_\gamma^\# g$ pour g appartenant à $\mathcal{G}$.

Les mesures de capacité standard pour les classifieurs à marge sont des nombres de couverture. Leur définition repose sur celles des ϵ-couvertures et des ϵ-réseaux.

Définition 13 (ϵ-couverture et ϵ-réseau, d'après les définitions 1 et 2 dans [129]) *Soit (E, ρ) un espace pseudo-métrique. e appartenant à E et r à $\mathbb{R}_+^*$, $B(e, r)$ désigne la boule ouverte de centre e et de rayon r dans E. Soit E' un sous-ensemble de E et ϵ un élément de $\mathbb{R}_+^*$. Un sous-ensemble $\overline{E'}$ de E est un* ϵ-réseau *de E' si et seulement si :*

$$E' \subset \bigcup_{e \in \overline{E'}} B(e, \epsilon).$$

$\bigcup_{e \in \overline{E'}} B(e, \epsilon)$ constitue alors une ϵ-couverture *de E'. $\overline{E'}$ est un* ϵ-réseau propre *de E' s'il est inclus dans E'.*

Définition 14 (Nombre de couverture [129]) *Soit (E, ρ) un espace pseudo-métrique, E' un sous-ensemble de E et ϵ un réel strictement positif. Si E' possède un ϵ-réseau de cardinalité finie, alors son* nombre de couverture *$\mathcal{N}(\epsilon, E', \rho)$ est le plus petit des cardinaux de ses ϵ-réseaux. Si un tel réseau n'existe pas, alors le nombre de couverture est ∞. On notera $\mathcal{N}^{(p)}(\epsilon, E', \rho)$ les nombres de couverture calculés en ne considérant que des ϵ-réseaux propres.*

Dans la suite, la pseudo-métrique dont seront munies les classes de fonctions considérées est la suivante.

Définition 15 (Pseudo-métrique d_{x^n}) *Soit $\mathcal{G}$ une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}^Q$ et $n \in \mathbb{N}^*$. Etant donnée une suite $x^n = (x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$, on définit la pseudo-métrique d_{x^n} sur $\mathcal{G}$ de la manière suivante :*

$$\forall (g, g') \in \mathcal{G}^2,\ d_{x^n}(g, g') = \max_{1 \leq i \leq n} \|g(x_i) - g'(x_i)\|_\infty .$$

A ce point de l'exposé, il est important de rappeler pourquoi il convient de travailler avec une pseudo-métrique empirique de ce type plutôt qu'avec une métrique fonctionnelle standard. L'idée de base des bornes VC est de caractériser la capacité d'une famille de fonctions de cardinalité infinie en calculant la trace de ces fonctions, éventuellement après discrétisation dans le cas où celles-ci sont à valeurs réelles, sur un ensemble de points de cardinalité finie. L'ensemble des traces ainsi obtenues permet de regrouper les fonctions en classes d'équivalence, et de passer de ce fait d'un ensemble infini de fonctions à un ensemble fini. Sur l'ensemble fini, établir des résultats de convergence uniforme ne pose pas de problème. Il suffit d'appliquer la fameuse "union bound". Naturellement, en ne considérant le comportement des fonctions qu'en un nombre fini de points, on ne peut vérifier la propriété de séparation des métriques, c'est-à-dire que l'on a $g \neq g' \not\Longrightarrow d(g, g') \neq 0$. Cependant, c'est précisément ce qui rend possible la constitution des classes d'équivalence. L'application de la discrétisation aux familles de fonctions considérées ici est décrite dans la section 5 de [95].

Définition 16 *$\mathcal{G}$, n et ϵ étant définis comme précédemment, on note :*

$$\mathcal{N}^{(p)}(\epsilon, \mathcal{G}, n) = \max_{x^n \in \mathcal{X}^n} \mathcal{N}^{(p)}(\epsilon, \mathcal{G}, d_{x^n}),$$

où le maximum est utilisé à la place d'un supremum pour souligner le fait que nous faisons implicitement l'hypothèse que tous les ϵ-réseaux considérés ici sont de cardinalité finie.

L'introduction de ces définitions permet de formuler le résultat de convergence uniforme suivant, étendant au cas multiclasse le corollaire 9 de [19] (la preuve effectue également un emprunt à celle du théorème 4.1 de [229]).

Théorème 1 (Théorème 22 dans [95]) *Soit $\mathcal{G}$ la famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\mathbb{R}^Q$ que peut réaliser un classifieur à Q catégories à grande marge. Soient $\Gamma \in \mathbb{R}_+^*$ et $\delta \in]0, 1[$. Avec une probabilité au moins égale à $1 - \delta$, uniformément pour toute valeur de γ dans $]0, \Gamma]$, le risque de toute fonction g de $\mathcal{G}$ est borné supérieurement de la manière suivante :*

$$R(g) \leq R_{\gamma,m}(g) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}\left(\gamma/4, \Delta_\gamma^\# \mathcal{G}, 2m\right)\right) + \ln\left(\frac{2\Gamma}{\gamma\delta}\right)\right)} + \frac{1}{m}. \tag{2.15}$$

Cette borne "d'un seul côté" est valable quelle que soit la mesure de probabilité P (on parle d'une borne indépendante de la distribution ou "distribution-free"). L'intérêt pratique d'un tel risque garanti repose entièrement sur la possibilité de dériver un majorant de bonne qualité sur le nombre de couverture qu'il fait intervenir. Pour ce faire, la littérature propose deux grandes familles de stratégies. Elles font l'objet des deux sections suivantes.

2.5.1.2 Utilisation de dimensions VC étendues

L'article fondateur de Vapnik et Chervonenkis sur l'extension du théorème de Glivenko-Cantelli (voir par exemple [176, 224]) aux processus indexés par des familles de fonctions à valeurs booléennes [231] a mis en évidence, pour les classifieurs binaires, le rôle central joué par la fonction de croisssance dans la caractérisation de la convergence uniforme, en probabilité (la convergence est en fait presque sûre) et pour toute mesure de probabilité P, du risque empirique vers le vrai risque. A part dans des cas simples, comme celui des PMC à unités à seuil, cette fonction est difficile à borner directement. Cependant, plusieurs auteurs [231, 191, 202] ont démontré indépendamment le résultat suivant : lorsque la dimension VC est finie, et pour une taille d'échantillon m supérieure à cette dimension, la fonction de croissance est bornée par une fonction croissant polynomialement avec m, le degré du polynôme étant précisément la dimension VC. On nomme ordinairement le lemme correspondant lemme de Sauer-Shelah. Ainsi, pour les classifieurs à valeurs binaires, l'étude des performances en généralisation peut se réduire au calcul d'une borne sur la dimension VC.

Le pendant standard du couple (fonction de croissance, dimension VC) pour les classifieurs à vaste marge calculant des dichotomies est le couple (nombres de couverture, dimension *fat-shattering* [122, 123]). Pour cette dernière dimension, les résultats théoriques de base ont été établis dans [4]. Cette référence fournit en particulier l'extension correspondante du lemme de Sauer-Shelah (lemme 3.5). C'est précisément ce résultat qui permet à Bartlett de prolonger le corollaire 9 de [19] de manière à produire des bornes sur les performances en généralisation des PMC biclasses (à unité de sortie unique) faisant intervenir les modules des poids des connexions. Comme le théorème 4.6 de [21] (entre autres) fournit une borne sur la dimension fat-shattering d'un séparateur linéaire, il est aisé de regrouper les éléments permettant de borner (avec une forte probabilité) l'erreur en généralisation d'une SVM biclasse.

Il existe également des dimensions dédiées aux systèmes discriminants multiclasses prenant leurs valeurs dans des ensembles finis, en premier lieu l'ensemble des catégories $\{1, \ldots, Q\}$. Parmi celles-ci, on peut en particulier citer la dimension graphique [73, 164], ainsi que la dimension de Natarajan [164]. Toutes ces dimensions ont été munies dans [24] d'un cadre théorique unificateur. Elles apparaissent ainsi comme des Ψ-dimensions. Là encore, des lemmes de Sauer-Shelah généralisés sont fournis (voir aussi [108]), ainsi que des bornes sur les dimensions VC généralisées.

En résumé, la théorie standard propose des résultats pour les classifieurs calculant des dichotomies (avec ou sans marge), ainsi que pour les modèles multiclasses à valeurs dans l'ensemble des catégories, c'est-à-dire sans marge. Elle s'avère ainsi applicable, en particulier, pour différentes familles de PMC biclasses (voir par exemple [23, 206, 188, 128, 207, 19, 12]) ou multiclasses à unités à seuil [200]. Naturellement, ceci ne permet pas de traiter de manière satisfaisante le cas des M-SVM (non plus que celui des PMC multiclasses à fonctions d'activation dérivables). Pour de telles machines, il convient de spécifier des mesures de capacité qui constituent à la fois des Ψ-dimensions "à marge" et des dimensions fat-shattering "multivariées". C'est précisément ce que nous avons fait dans [95].

Définition 17 (Ψ-dimensions, [24]) *Soit $\mathcal{F}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans l'ensemble fini $\{1, \ldots, Q\}$. Soit Ψ une famille d'applications ψ de $\{1, \ldots, Q\}$ dans $\{-1, 1, *\}$, où le symbole $*$ représente une valeur prise par défaut. Un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$, de $\mathcal{X}$ est dit être* Ψ-pulvérisé *par $\mathcal{F}$ s'il existe une application $\psi^n = \left(\psi^{(i)}\right)_{1 \leq i \leq n}$ dans Ψ^n telle que, pour tout vecteur v_y de $\{-1, 1\}^n$, il existe une fonction f_y dans $\mathcal{F}$ satisfaisant*

$$\left(\psi^{(i)} \circ f_y(x_i)\right)_{1 \leq i \leq n} = v_y.$$

La Ψ-dimension *de $\mathcal{F}$, notée Ψ-dim($\mathcal{F}$), est le cardinal du plus grand sous-ensemble de $\mathcal{X}$ Ψ-pulvérisé par $\mathcal{F}$, si ce cardinal est fini, et l'infini dans le cas contraire.*

Remarque 2 *Soient $\mathcal{F}$ et Ψ les familles de fonctions définies ci-dessus. En étendant la définition de la dimension VC, VC-dim, de manière qu'elle s'applique aux familles de fonctions prenant leurs valeurs dans $\{-1, 1, *\}$, ce qui ne change rien en pratique, on dispose de la caractérisation suivante des Ψ-dimensions :*

$$\Psi\text{-}dim(\mathcal{F}) = VC\text{-}dim\left(\{(x, \psi) \mapsto \psi \circ f(x) : f \in \mathcal{F}, \psi \in \Psi\}\right).$$

La Ψ-dimension la plus utilisée est la dimension graphique.

Définition 18 (Dimension graphique, [73]) *Soit $\mathcal{F}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\{1, \ldots, Q\}$. La* dimension graphique *de $\mathcal{F}$, G-dim($\mathcal{F}$), est la Ψ-dimension de $\mathcal{F}$ dans le cas particulier où $\Psi = \{\psi_k : 1 \leq k \leq Q\}$, l'application ψ_k prenant la valeur 1 si son argument est égal à k et la valeur -1 dans le cas contraire. En reformulant cette définition dans le contexte de la discrimination multiclasse, les fonctions ψ_k sont les fonctions indicatrices des catégories.*

Naturellement, les Ψ-dimensions sont liées aux méthodes de décomposition décrites dans la section 2.3, et de ce point de vue, la dimension graphique correspond à la méthode un contre tous. La Ψ-dimension correspondant à la méthode un contre un a été proposée par Natarajan.

Définition 19 (Dimension de Natarajan, [164]) *Soit $\mathcal{F}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\{1, \ldots, Q\}$. La* dimension de Natarajan *de $\mathcal{F}$, N-dim($\mathcal{F}$), est la Ψ-dimension de $\mathcal{F}$ dans le cas particulier où $\Psi = \{\psi_{k,l} : 1 \leq k \neq l \leq Q\}$, l'application $\psi_{k,l}$ prenant la valeur 1 si son argument est égal à k, la valeur -1 si son argument est égal à l, et la valeur $*$ partout ailleurs.*

Définition 20 (Dimension fat-shattering, [123]) *Soit $\mathcal{G}$ une famille de fonctions sur $\mathcal{X}$ à valeurs réelles. Pour $\gamma \in \mathbb{R}_+^*$, un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de $\mathcal{X}$ est dit être* γ-pulvérisé *par $\mathcal{G}$ s'il existe un vecteur $v_b = (b_i) \in \mathbb{R}^n$ tel que, pour tout vecteur $v_y = (y_i)$ de $\{-1, 1\}^n$, il existe une fonction g_y dans $\mathcal{G}$ satisfaisant*

$$\forall i \in \{1, \ldots, n\}, \; y_i \left(g_y(x_i) - b_i\right) \geq \gamma.$$

La dimension fat-shattering à marge γ, *ou* P_γ dimension, *de la famille $\mathcal{G}$, P_γ-dim($\mathcal{G}$), est le cardinal du plus grand sous-ensemble de $\mathcal{X}$ γ-pulvérisé par $\mathcal{G}$, si ce cardinal est fini, et l'infini dans le cas contraire.*

Ces définitions étant posées, nous proposons d'introduire la notion de marge dans les Ψ-dimensions de la manière suivante.

Définition 21 (γ-Ψ-dimensions, définition 28 dans [95]) *Soit $\mathcal{G}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\mathbb{R}^Q$. Soit Ψ une famille d'applications ψ de $\{1, \ldots, Q\}$ dans $\{-1, 1, *\}$, où le symbole $*$ représente une valeur prise par défaut. Pour $\gamma \in \mathbb{R}_+^*$, un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de $\mathcal{X}$ est dit être* γ-Ψ-pulvérisé *(*Ψ-pulvérisé avec une marge γ*) par $\Delta^\# \mathcal{G}$ s'il existe une application $\psi^n = \left(\psi^{(i)}\right)_{1 \leq i \leq n}$ dans Ψ^n et un vecteur $v_b = (b_i)$ de $\mathbb{R}^n$ tels que, pour tout vecteur $v_y = (y_i)$ de $\{-1, 1\}^n$, il existe une fonction g_y dans $\mathcal{G}$ satisfaisant*

$$\forall i \in \{1, \ldots, n\}, \left\{ \begin{array}{lrlrl} si\ y_i = & 1, & \exists k : \psi^{(i)}(k) = & 1 & \wedge\ \Delta^\# g_{y,k}(x_i) - b_i \geq \gamma \\ si\ y_i = & -1, & \exists l : \psi^{(i)}(l) = & -1 & \wedge\ \Delta^\# g_{y,l}(x_i) + b_i \geq \gamma \end{array} \right. .$$

La γ-Ψ-dimension, *ou* Ψ-dimension à marge γ, *de $\Delta^\# \mathcal{G}$, notée Ψ-dim($\Delta^\# \mathcal{G}, \gamma$), est le cardinal du plus grand sous-ensemble de $\mathcal{X}$ γ-Ψ-pulvérisé par $\Delta^\# \mathcal{G}$, si ce cardinal est fini, et l'infini dans le cas contraire.*

On vérifie aisément que cette définition, reformulée dans le cas biclasse, correspond exactement à celle de la dimension fat-shattering. Nous avons vu qu'il existait deux Ψ-dimensions principales, la dimension graphique et la dimension de Natarajan. La première est inspirée de la méthode de décomposition un contre tous, tandis que la seconde est inspirée de la méthode de décomposition un contre un. Cette dernière approche est celle qui se prête le mieux à une extension directe de résultats biclasses. Pour cette raison, parmi les γ-Ψ-dimensions, celle qui apparaît la plus utile en pratique est la dimension de Natarajan à marge.

Définition 22 (Dimension de Natarajan à marge γ, définition 29 dans [95]) *Soit $\mathcal{G}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $\mathbb{R}^Q$. Pour $\gamma \in \mathbb{R}_+^*$, un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de $\mathcal{X}$ est dit être* γ-N-pulvérisé *(N-pulvérisé avec une marge γ) par $\Delta^\# \mathcal{G}$ s'il existe un ensemble*

$$I(s_{\mathcal{X}^n}) = \{(i_1(x_i), i_2(x_i)) \ : 1 \leq i \leq n\}$$

de n couples d'indices distincts dans $\{1, \ldots, Q\}$ et un vecteur $v_b = (b_i)$ de $\mathbb{R}^n$ tels que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^n$, il existe une fonction g_y de $\mathcal{G}$ satisfaisant

$$\forall i \in \{1, \ldots, n\}, \left\{ \begin{array}{lll} si\ y_i = & 1, & \Delta^\# g_{y,i_1(x_i)}(x_i) - b_i \geq \gamma \\ si\ y_i = & -1, & \Delta^\# g_{y,i_2(x_i)}(x_i) + b_i \geq \gamma \end{array} \right. .$$

La dimension de Natarajan à marge γ *de la famille $\Delta^\# \mathcal{G}$, N-dim$(\Delta^\# \mathcal{G}, \gamma)$, est le cardinal du plus grand sous-ensemble de $\mathcal{X}$ γ-N-pulvérisé par $\Delta^\# \mathcal{G}$, si ce cardinal est fini, et l'infini dans le cas contraire.*

Nous avons vu que la spécification d'une dimension VC généralisée n'a de sens que si l'on dispose pour elle de deux résultats : un lemme de Sauer-Shelah étendu et une borne sur sa valeur pour les classifieurs d'intérêt. De ce point de vue, l'utilisation de l'opérateur π_γ, qui ne présentait que des avantages dans la dérivation du théorème 1, soulève ici une difficulté. Il est en effet délicat de gérer la non-linéarité qu'il introduit dans le cadre du calcul d'une borne sur une dimension VC généralisée. C'est la raison pour laquelle nous avons donné une définition des γ-Ψ-dimensions faisant intervenir les familles de fonctions $\Delta^\# \mathcal{G}$ et non $\Delta_\gamma^\# \mathcal{G}$. La manière la plus naturelle d'effectuer la transition entre ces deux familles consiste à opérer au niveau des nombres de couverture (établir un lien au niveau des γ-Ψ-dimensions demeure un problème ouvert). Profitant du fait que π_γ est une fonction lipschitzienne de rapport 1, on démontre immédiatement le lemme suivant.

Lemme 1 *Soit $\mathcal{G}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}^Q$, $(\gamma, \epsilon) \in \left(\mathbb{R}_+^*\right)^2$ et $n \in \mathbb{N}^*$. Alors,*

$$\mathcal{N}^{(p)}(\epsilon, \Delta_\gamma^\# \mathcal{G}, n) \leq \mathcal{N}^{(p)}(\epsilon, \Delta^\# \mathcal{G}, n).$$

Comme nous l'avons vu dans la section précédente, le prix à payer pour l'utilisation de ce lemme dans le cadre du passage par une dimension VC généralisée est l'ajout d'une hypothèse supplémentaire sur la famille $\mathcal{G}$, celle d'un codomaine borné. Plus précisément, on suppose qu'il existe $M_\mathcal{G} \in \mathbb{R}_+^*$ tel que les fonctions g, et par voie de conséquence les fonctions $\Delta^\# g$, prennent leurs valeurs dans $[-M_\mathcal{G}, M_\mathcal{G}]^Q$. Sous ces hypothèses, les seules valeurs du paramètre de marge γ correspondant à une situation non dégénérée (un risque à marge différent de 1) sont celles qui sont inférieures ou égales à $M_\mathcal{G}$. En conséquence, nous considérons dans la suite que Γ est égal à $M_\mathcal{G}$. Dans ces conditions, on dispose du lemme de Sauer-Shelah généralisé suivant.

Lemme 2 (Lemme 39 dans [95]) *Soit $\mathcal{G}$ une famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $[-M_\mathcal{G}, M_\mathcal{G}]^Q$. Pour toute valeur d'$\epsilon$ dans $]0, M_\mathcal{G}]$ et toute valeur entière de n satisfaisant $n \geq$ N-dim$(\Delta\mathcal{G}, \epsilon/6)$, on dispose de la borne suivante :*

$$\mathcal{N}^{(p)}(\epsilon, \Delta^* \mathcal{G}, n) < 2 \left(n\, Q^2 (Q-1) \left\lfloor \frac{3M_\mathcal{G}}{\epsilon} \right\rfloor^2 \right)^{\left\lceil d \log_2 \left(e n C_Q^2 \left(2 \left\lfloor \frac{3M_\mathcal{G}}{\epsilon} \right\rfloor - 1 \right) / d \right) \right\rceil}$$

où $d =$ N-dim$(\Delta\mathcal{G}, \epsilon/6)$ et e est la base des logarithmes népériens.

La caractéristique fondamentale de ce résultat réside dans le fait que l'opérateur de marge figurant dans le nombre de couverture doit être l'opérateur Δ^*, ceci alors qu'il est possible de substituer Δ^* à Δ dans l'expression de la dimension de Natarajan à marge. L'utilisation de Δ^* est en effet nécessaire pour étendre au cas multiclasse le lien entre séparation de fonctions au sens de la pseudo-métrique utilisée et capacité de pulvérisation, lien qui est à la base des lemmes de Sauer-Shelah. Cette caractéristique du cas multiclasse (rappelons que dans le cas biclasse, $\Delta = \Delta^*$) est analysée en détail dans la section 5.3 de [95] (voir en particulier le lemme 35 et la remarque 36 de cet article). En combinant le résultat de convergence uniforme de base et le lemme de Sauer-Shelah généralisé, on obtient en définitive la borne suivante sur les performances en généralisation des systèmes discriminants multiclasses à grande marge.

Théorème 2 (Théorème 40 dans [95]) *Soit $\mathcal{G}$ la famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$ que peut réaliser un classifieur à Q catégories à grande marge. Soit $\delta \in]0,1[$. Avec une probabilité au moins égale à $1-\delta$, uniformément pour toute valeur de γ dans $]0, M_{\mathcal{G}}]$, le risque de toute fonction g de $\mathcal{G}$ est borné supérieurement de la manière suivante :*

$$R(g) \leq R_{\gamma,m}(g) +$$

$$\sqrt{\frac{2}{m}\left(\ln\left(4\left(2m\,Q^2(Q-1)\left\lfloor\frac{12M_{\mathcal{G}}}{\gamma}\right\rfloor^2\right)^{\left\lceil d\log_2\left(emQ(Q-1)\left(2\left\lfloor\frac{12M_{\mathcal{G}}}{\gamma}\right\rfloor-1\right)/d\right)\right\rceil}\right)+\ln\left(\frac{2M_{\mathcal{G}}}{\gamma\delta}\right)\right)+\frac{1}{m}}$$

où $d = N\text{-}dim\,(\Delta\mathcal{G}, \gamma/24)$.

L'analyse de cette borne en fonction de ses deux principaux paramètres, m et Q, souligne une fois de plus les difficultés originales auxquelles nous confronte le cas multiclasse. Sa dépendance au nombre de catégories Q est en effet entièrement liée au choix de la Ψ-dimension sous-jacente, ici la dimension de Natarajan. Pour présenter les choses de manière rapide, le choix de cette dimension, reposant sur le principe de la méthode de décomposition un contre un, fait naturellement apparaître dans les formules le coefficient du binôme C_Q^2. En utilisant pour γ-Ψ-dimension la dimension graphique à marge, nous aurions vu se substituer à ce coefficient le nombre de catégories, Q. Ce résultat est donc très relatif. En fait, en prolongeant le raisonnement ci-dessus, on conclut assez simplement que l'équivalent du Lemme 2 ou du théorème 2 qui fait intervenir la plus petite fonction de la γ-Ψ-dimension employée est justement celui correspondant à la dimension graphique à marge. Se repose alors la question du choix de la γ-Ψ-dimension. Nous avons eu l'occasion de souligner les avantages de nature méthodologique présentés par la méthode de décomposition un contre un. Le second théorème de cette section est une borne sur la dimension de Natarajan à marge des M-SVMs. Nous ne savons pas actuellement comment dériver une borne similaire pour la dimension graphique à marge. La difficulté principale tient au fait que le domaine complémentaire du domaine associé à une catégorie par un classifieur linéaire multiclasse n'est pas convexe.

Le terme de contrôle du risque garanti apparaît également comme un $O\left(\frac{\ln(m)}{\sqrt{m}}\right)$. Ce taux de décroissance en fonction de la taille de l'échantillon d'apprentissage est moins bon que celui de la borne VC standard, qui était initialement en $\sqrt{\frac{\ln(m)}{m}}$ (voir le théorème 6.7 de [227]), avant de trouver sa forme optimale en $\sqrt{\frac{1}{m}}$ (voir par exemple le théorème 3.4. de [39], ainsi que la discussion dans [153]). Des améliorations peuvent naturellement être apportées afin de combler ce fossé, par exemple en changeant de pseudo-métrique.

Le théorème 2 fournit une borne "distribution-free" correspondant, uniformément pour toute valeur de γ dans l'intervalle $]0, M_{\mathcal{G}}]$, à une convergence en probabilité "d'un seul côté" de la forme

$$\lim_{m\longrightarrow+\infty}\sup_P \mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)>\epsilon\right)=0.$$

En fait, on dispose d'un résultat plus fort, dans la mesure où la convergence s'avère être presque sûre. Dans un but de simplification, nous exprimons ce résultat pour une valeur de γ donnée.

Proposition 4 (Résultat de convergence presque sûre) *Soit $\mathcal{G}$ la famille de fonctions sur un domaine $\mathcal{X}$ à valeurs dans $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$ que peut réaliser un classifieur à Q catégories à grande marge. Si, pour γ appartenant à $]0, M_{\mathcal{G}}]$, $N\text{-}dim\,(\Delta\mathcal{G}, \gamma/24) < \infty$, alors pour tout ϵ appartenant à $\mathbb{R}_+^*$,*

$$\lim_{m\longrightarrow+\infty}\sup_P \mathbb{P}\left(\sup_{n\geq m}\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,n}(g)\right)>\epsilon\right)=0.$$

Le théorème 2 s'applique pour des familles de fonctions à valeurs dans un domaine borné de la forme $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^Q$. Afin de l'appliquer aux M-SVM, il convient donc d'introduire des hypothèses sur $\mathcal{X}$ et la fonction noyau κ, qui se traduisent par des hypothèses sur $\Phi(\mathcal{X})$, ainsi que des restrictions sur

les paramètres $\mathbf{w}$ et $\mathbf{b}$. Ces hypothèses et restrictions trouveront également leur utilité lors des étapes ultérieures de la construction du risque garanti. Elles constituent en fait une extension directe du cadre dans lequel est ordinairement traité le cas biclasse, cadre qui distingue la sélection de modèle (pas de restriction a priori sur les valeurs que peuvent prendre les paramètres du noyau κ et C, sauf dans le cas des approches bayésiennes) et la sélection de fonction (contraintes a priori sur $\mathbf{w}$ et $\mathbf{b}$).

Hypothèses 1 *Pour majorer la capacité d'une M-SVM à Q catégories, les hypothèses et contraintes suivantes sont introduites :*

1. *$\Phi(\mathcal{X})$ est inclus dans la boule fermée de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$;*
2. *le vecteur $\mathbf{w}$ satisfait la condition $\|\mathbf{w}\|_\infty \leq \Lambda_w$;*
3. *le vecteur $\mathbf{b}$ appartient à $[-\beta, \beta]^Q$.*

Dans ces conditions, il résulte de l'inégalité de Cauchy-Schwarz que dans le cas d'une M-SVM, le paramètre $M_{\mathcal{G}}$ peut être instancié de la manière suivante : $M_{\mathcal{H}} = \Lambda_w \Lambda_{\Phi(\mathcal{X})} + \beta$ ($M_{\bar{\mathcal{H}}} = \Lambda_w \Lambda_{\Phi(\mathcal{X})}$).

Borner les dimensions VC à marge fait essentiellement appel à des résultats d'algèbre. Dans le cas des M-SVM, cette étape est donc simplifiée si elle est précédée d'une transition de $\mathcal{H}$ vers $\bar{\mathcal{H}}$, afin de travailler uniquement dans des espaces de Hilbert. Ici encore, la transition s'effectue plus simplement au niveau des nombres de couverture. Elle est alors fournie par le lemme suivant.

Lemme 3 *Soit $\mathcal{H}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous l'hypothèse que $\mathbf{b}$ appartient à $[-\beta, \beta]^Q$ et $\bar{\mathcal{H}}$ sa restriction aux fonctions vérifiant $\mathbf{b} = 0$. Soient $\epsilon \in \mathbb{R}_+^*$ et $n \in \mathbb{N}^*$. Alors,*

$$\mathcal{N}^{(p)}\left(\epsilon, \Delta^{\#}\mathcal{H}, n\right) \leq \left(2\left\lceil \frac{\beta}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}^{(p)}\left(\epsilon/2, \Delta^{\#}\bar{\mathcal{H}}, n\right).$$

Ce lemme peut être légèrement amélioré en exploitant l'hypothèse supplémentaire $\sum_k h_k = 0$. Il se présente alors comme une généralisation directe du résultat biclasse. Il permet en définitive de calculer la complexité en échantillon des M-SVM par application d'un dernier résultat sur la dimension de Natarajan à marge de la famille de fonctions $\Delta\bar{\mathcal{H}}$.

Théorème 3 (Théorème 48 dans [95]) *Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous les hypothèses 1 et la contrainte supplémentaire $\mathbf{b} = 0$. Alors, pour toute valeur strictement positive d'ϵ, on dispose de la borne suivante :*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \epsilon\right) \leq C_Q^2 \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2. \tag{2.16}$$

Il convient de souligner le fait qu'ici, c'est l'opérateur de marge Δ qui doit être utilisé. Plus précisément, c'est son utilisation qui permet d'étendre au cas multiclasse le lemme 4.2 de [21]. Cette observation achève l'explication de la forme prise par le lemme 2.

Ce théorème, reformulé dans le cas $Q = 2$, correspond à la borne sur la dimension fat-shattering d'un séparateur linéaire fournie par le théorème 4.6 de [21] (voir aussi la remarque 1 dans [103]) :

$$P_\epsilon\text{-dim}\left(H_\kappa\right) \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2.$$

Dans le cas général, il nous apprend que la dimension de Natarajan à marge d'une M-SVM à Q catégories peut être bornée supérieurement par une borne uniforme sur les dimensions fat-shattering de ses hyperplans séparateurs (définis par l'équation $\langle w_k - w_l, \Phi(x)\rangle = 0$) multipliée par le nombre de ces hyperplans, C_Q^2. Il convient de noter que la preuve s'appuie directement sur l'idée fondant la définition des Ψ-dimensions, qui est de simuler la mise en œuvre d'une méthode de décomposition, et en tirer parti pour exploiter des résultats biclasses. En ce sens, la discussion rejoint celle du théorème 2 : ce résultat est entièrement lié au principe de la méthode de décomposition un contre un. Les termes C_Q^2 et $\|\mathbf{w}\|_\infty$ apparaissent dans (2.16) parce que toutes les paires de catégories sont considérées indépendamment les unes des autres et qu'elles jouent toutes le même rôle (ce qui appelle l'utilisation d'une borne

sur $1/2 \max_{1\leq k<l\leq Q} \|w_k - w_l\|$). Bien entendu, prendre en compte le fait que les C_Q^2 classifieurs binaires ne sont pas indépendants, dans la mesure où ils s'appuient sur un ensemble commun de Q vecteurs w_k, devrait engendrer une borne de meilleure qualité. Ici apparaît une fois encore le besoin de produire des solutions originales pour le cas multiclasse, au lieu de simples extensions de résultats biclasses. Un autre problème ouvert est celui qui consiste à harmoniser les considérations pratiques, incitant à utiliser la norme ℓ_2 sur $E^Q_{\Phi(\mathcal{X})}$, et la majoration de la mesure de capacité, incitant au contraire à utiliser la norme $\|.\|_\infty$. De nouveau, on peut s'appuyer sur la grande différence entre les marges observées sur la figure 2.3 pour établir que ce problème est d'importance.

Un grand nombre de travaux ont porté sur la majoration des nombres de couverture associés à la SVM biclasse par des méthodes ne faisant pas intervenir de dimension VC généralisée. L'idée de base consiste à tirer le meilleur parti de la nature du noyau et du domaine dans lequel vivent les données, ce qui est difficile à effectuer dans le cadre précédent. L'extension multiclasse de ces travaux doit naturellement fournir des alternatives intéressantes à la borne résultant de la combinaison des lemmes 2 et 3 et du théorème 3, dont la preuve repose essentiellement sur des résultat valables pour des familles de fonctions plus générales que $\mathcal{H}$. Nous exposons à présent la première étude de ce type.

2.5.1.3 Utilisation des nombres d'entropie de l'opérateur d'évaluation

A la fin des années 90, Williamson et ses co-auteurs ont dérivé un ensemble de résultats permettant de borner directement les nombres de couverture associés à des machines à noyau binaires, en particulier les SVM standard. Ces résultats sont exposés dans [245, 243, 244, 197, 102]. Ils sont à l'origine des travaux de même nature que nous avons effectués dans le cas multiclasse. Le théorème 1 a été établi pour les deux opérateurs de marge, Δ et Δ^*. Dans le cas de l'utilisation d'une dimension VC généralisée, le lemme 2 impose de travailler sur les nombres de couverture de $\Delta^*_\gamma \mathcal{G}$ (ou $\Delta^* \mathcal{G}$). Les résultats de la présente section peuvent être formulés avec l'un ou l'autre des opérateurs. Cependant, afin de nous conformer à l'exposition effectuée dans [98], nous les exprimons ici pour le seul opérateur Δ. Pour borner directement $\mathcal{N}^{(p)}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m))$, comme dans le cas de l'utilisation de la dimension de Natarajan à marge, une simplification préliminaire apparaît utile. Elle est fournie par le lemme suivant :

Lemme 4 *Soit $\mathcal{G}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}^Q$. Soient $\epsilon \in \mathbb{R}^*_+$ et $n \in \mathbb{N}^*$. Alors,*

$$\mathcal{N}^{(p)}(\epsilon, \Delta\mathcal{G}, n) \leq \mathcal{N}^{(p)}(\epsilon, \mathcal{G}, n).$$

Appliquer successivement les lemmes 1, 3 et 4, permet de substituer à l'étude de la capacité de la famille de fonctions $\Delta_\gamma \mathcal{H}$ celle de $\bar{\mathcal{H}}$ (plus précisément, sa restriction à l'hyperplan d'équation $\sum_{k=1}^Q w_k = 0$ sous les hypothèses 1). Cette transition effectuée, il devient possible d'exploiter l'idée développée dans [243] (voir également [244]), consistant à borner les nombres de couverture de $\bar{\mathcal{H}}$ en fonction des nombres d'entropie de l'opérateur d'évaluation correspondant, et tirer parti ensuite des résultats disponibles sur les nombres d'entropie des opérateurs linéaires entre espaces de Banach. Comme dans le cas des ϵ-couvertures, ϵ-réseaux et nombres de couverture, il est difficile d'identifier la référence de base introduisant la notion de nombres d'entropie. Dans le premier cas, nous avions dirigé le lecteur vers un célèbre article de Kolmogorov et Tihomirov, présentant l'intérêt supplémentaire d'exposer plusieurs résultats sur la capacité de familles de fonctions très utiles en apprentissage. Nous employons ici comme référence le traité de base sur l'entropie et la compacité des opérateurs, [49], sur lequel nous aurons d'autres occasions de nous appuyer dans la suite.

Définition 23 (Nombres d'entropie d'un ensemble et d'un opérateur [49]) *Soit (E, ρ) un espace pseudo-métrique. Soit E' un sous-ensemble borné de E et $n \in \mathbb{N}^*$. Le n-*ième nombre d'entropie *de E', $\epsilon_n(E')$, est défini comme étant le plus petit réel ϵ tel qu'il existe un ϵ-réseau de E' de cardinalité au plus n. Soient E et F deux espaces de Banach. $\mathfrak{L}(E, F)$ désigne l'espace de Banach de tous les opérateurs (linéaires bornés) de E dans F muni de sa norme habituelle. Soit U_E la boule unité fermée de E. Le n-*ième nombre d'entropie *de $S \in \mathfrak{L}(E, F)$ est défini par :*

$$\epsilon_n(S) = \epsilon_n(S(U_E)).$$

Dans cette double définition apparaissent à la fois un espace pseudo-métrique et des espaces de Banach. Naturellement, le passage d'un espace pseudo-métrique (complet) (E, ρ) à un espace semi-normé (complet) $(E, \|.\|_E)$ s'obtient en posant $\|e\|_E = \rho(e, 0)$, le passage inverse s'obtenant en posant $\rho(e, e') = \|e - e'\|_E$. On désigne par ℓ_p^n l'espace vectoriel des suites de réels de longueur n muni de la norme $\|.\|_p$.

Définition 24 (Opérateur d'évaluation) *Pour $n \in \mathbb{N}^*$, soit $x^n \in \mathcal{X}^n$. L'*opérateur d'évaluation *S_{x^n} sur $\bar{\mathcal{H}}$ est défini par :*

$$\begin{array}{rccc} S_{x^n} : & \bar{\mathcal{H}} & \longrightarrow & \ell_\infty^{Qn} \\ & \bar{h} = (w_k)_{1 \le k \le Q} & \mapsto & S_{x^n}\left(\bar{h}\right) = (\langle w_k, \Phi(x_i) \rangle)_{1 \le i \le n,\, 1 \le k \le Q}. \end{array}$$

Lemme 5 *Considérons la restriction de $\bar{\mathcal{H}}$ aux fonctions vérifiant $\|\mathbf{w}\|_\infty \le \Lambda_w$. Soit $\mathcal{U}$ sa restriction aux fonctions vérifiant $\|\mathbf{w}\|_\infty \le 1$. Soient $\epsilon \in \mathbb{R}_+^*$ et $n \in \mathbb{N}^*$. Alors,*

$$\mathcal{N}^{(p)}(\Lambda_w \epsilon, \bar{\mathcal{H}}, n) \le \mathcal{N}^{(p)}(\epsilon, \mathcal{U}, n).$$

Naturellement, ce résultat est indépendant du choix de la norme. $E_{\Phi(\mathcal{X})}^Q$ n'a été muni ici de la norme $\|.\|_\infty$ que dans le but de rendre les résultats de cette sous-section directement comparables à ceux de la sous-section précédente. Pour établir le lien entre $\mathcal{N}^{(p)}(\epsilon, \mathcal{U}, n)$ et les nombres d'entropie de S_{x^n}, il convient en premier lieu d'identifier une difficulté. Depuis la formulation du théorème 1, nous n'avons considéré que des ϵ-réseaux propres. Or, la définition des nombres d'entropie ne se fonde pas sur une telle restriction. Cette difficulté est surmontée grâce au lemme suivant, qui découle directement de l'inégalité triangulaire.

Lemme 6 *Soit $\mathcal{G}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}^Q$. Soient $\epsilon \in \mathbb{R}_+^*$ et $n \in \mathbb{N}^*$. Alors,*

$$\mathcal{N}^{(p)}\left(2\epsilon, \mathcal{G}, n\right) \le \mathcal{N}\left(\epsilon, \mathcal{G}, n\right).$$

Ce résultat n'est bien entendu pas spécifique aux familles de fonctions $\mathcal{G}$. La question de savoir s'il est possible de développer l'ensemble du raisonnement qui suit en ne considérant que des ϵ-réseaux propres, de manière à éviter, comme dans le cas de l'utilisation de la dimension de Natarajan à marge, l'emploi de ce lemme, reste ouverte. Le lien entre $\mathcal{N}(\epsilon, \mathcal{U}, n)$ et les nombres d'entropie de S_{x^n} est fourni par la proposition suivante.

Proposition 5 *Soient $\epsilon \in \mathbb{R}_+^*$ et $n \in \mathbb{N}^*$.*

$$\sup_{x^n \in \mathcal{X}^n} \epsilon_p(S_{x^n}) \le \epsilon \Longrightarrow \mathcal{N}(\epsilon, \mathcal{U}, n) \le p.$$

Différents résultats sont disponibles pour borner $\epsilon_p(S_{x^n})$, suivant que l'espace de représentation $E_{\Phi(\mathcal{X})}$ est de dimension finie ou non. Le premier cas est le plus simple. Il permet l'utilisation de la proposition suivante.

Proposition 6 (Proposition 1.3.1 dans [49]) *Soient E et F des espaces de Banach et $S \in \mathfrak{L}(E, F)$. Si S est de rang r, alors pour $n \in \mathbb{N}^*$,*

$$\epsilon_n(S) \le 4\|S\| n^{-1/r},$$

la norme de S étant définie de manière canonique comme $\|S\| = \sup_{e \in E\ :\|e\|_E = 1} \|S(e)\|_F$.

Cette situation, étudiée dans [75, 74], fournit une borne par combinaison des lemmes 1, 3, 4, 5 et 6 (naturellement, différents enchaînements sont possibles), et des propositions 5 et 6. Elle est donnée par le théorème suivant.

Théorème 4 (D'après le théorème 3 de [98]) *Soit $\mathcal{H}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous les hypothèses 1. Si la dimension de l'espace $E_{\Phi(\mathcal{X})}$ est finie et égale à d, alors, pour toute valeur strictement positive de γ, on dispose de la borne suivante :*

$$\mathcal{N}^{(p)}\left(\gamma/4, \Delta_\gamma \mathcal{H}, 2m\right) \le \left(2\left\lceil \frac{8\beta}{\gamma} \right\rceil + 1\right)^Q \cdot \left(\frac{64 \Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\gamma}\right)^{Qd}. \tag{2.17}$$

La preuve de cette borne repose sur le fait que sous l'hypothèse $\dim\left(E_{\Phi(\mathcal{X})}\right) = d$, pour tout $n \in \mathbb{N}^*$, le rang de l'opérateur S_{x^n} est majoré par la dimension de son domaine, dimension égale à Qd (puisque les vecteurs w_k appartiennent à $\mathbb{R}^d$). Cette dimension peut être ramenée à $(Q-1)d$ en utilisant l'hypothèse supplémentaire $\sum_k h_k = 0$. Dans le cas contraire, ce qui se produit par exemple si le noyau est un noyau RBF (gaussien), la seule borne sur le rang disponible est une borne sur la dimension du codomaine de l'opérateur, Qn. François Denis a observé que ce dernier majorant, pour $n = 2m$, se substituant à d dans l'inégalité (2.17), produit une borne sur les nombres de couverture inutilisable dans le risque garanti donné par (2.15). En effet, en employant cette borne, la borne sur le terme de contrôle ne tend plus vers 0 lorsque m tend vers l'infini. Il est alors nécessaire de recourir à un résultat plus complexe, le théorème de Maurey-Carl.

Théorème 5 (Théorème de Maurey-Carl (lemme 6.4.1 dans [49])) *Soient H un espace de Hilbert et S un opérateur appartenant à $\mathfrak{L}(\ell_1^n, H)$ ou $\mathfrak{L}(H, \ell_\infty^n)$. Alors, pour tout couple d'entiers (k, n) vérifiant $1 \leq k \leq n$, on a*

$$e_k(S) \leq c\left(\frac{1}{k}\log_2\left(1+\frac{n}{k}\right)\right)^{1/2}\|S\|, \tag{2.18}$$

où le nombre d'entropie dyadique $e_k(S)$ *est égal à $\epsilon_{2^{k-1}}(S)$ et c est une constante universelle.*

En substituant le théorème 5 à la proposition 6 dans l'enchaînement des résultats conduisant au théorème 4, on obtient en définitive la borne recherchée.

Théorème 6 (D'après le théorème 2 de [98]) *Soit $\mathcal{H}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous les hypothèses 1. Alors, pour toute valeur strictement positive de γ, on dispose de la borne suivante :*

$$\mathcal{N}^{(p)}(\gamma/4, \Delta_\gamma\mathcal{H}, 2m) \leq \left(2\left\lceil\frac{8\beta}{\gamma}\right\rceil+1\right)^Q \cdot 2^{\frac{16c\Lambda_w\Lambda_{\Phi(\mathcal{X})}}{\gamma}\sqrt{\frac{2Qm}{\ln(2)}}-1}.$$

Cette majoration du nombre de couverture d'intérêt fait intervenir la même norme sur $E^Q_{\Phi(\mathcal{X})}$ que celle impliquant la dimension de Natarajan à marge. Elle ne permet donc pas davantage de caractériser précisément l'influence du pénalisateur commun aux formulations de base des trois M-SVM : $\sum_{k=1}^Q \|w_k\|^2$. L'incorporer dans la borne sur le risque donnée par le théorème 1 fournit pour taux de décroissance du terme de contrôle $\sqrt{\frac{1}{\sqrt{m}}}$. Cette borne est donc, au moins pour une taille de l'échantillon d'apprentissage suffisamment grande, moins bonne que celle issue de l'emploi de la dimension de Natarajan à marge. Naturellement, l'utilisation pratique de ce résultat, par exemple pour mettre en œuvre une méthode de sélection de modèle, nécessite la détermination de la valeur de la constante universelle c. Si l'on compare les résultats exposés ci-dessus aux sources d'inspiration provenant du cas biclasse, on s'aperçoit que beaucoup reste à faire, en particulier pour exploiter les propriétés du noyau, autrement qu'à travers son influence sur les marges ou sur la valeur de $\Lambda_{\Phi(\mathcal{X})}$. Dans [245, 244, 197] la famille des fonctions réalisables par une machine à noyau est considérée comme étant engendrée par un opérateur intégral induit par le noyau, et des propriétés de cet opérateur sont utilisées afin de borner les nombres de couverture d'intérêt. L'étude est poursuivie dans [102], où le cas d'un noyau RBF est plus particulièrement considéré. Les auteurs obtiennent une borne sur les nombres de couverture en fonction de la variance σ^2 du noyau. Plus précisément, le logarithme des nombres de couverture apparaît comme un "grand o" de l'inverse de σ (équation 30). Tous ces travaux demeurent encore à étendre au cas multiclasse.

2.5.1.4 Utilisation d'une moyenne de Rademacher

La borne établie dans cette section s'appuie de nouveau sur les hypothèses 1 et la contrainte supplémentaire $\mathbf{b} = 0$ (on travaille avec $\bar{\mathcal{H}}$). Elle est directement inspirée des résultats biclasses exposés dans [20] et les sections 3 et 4 de [39]. Elle fait également appel au contenu des chapitres 4 de [134] et [201]. Nous débutons en rappelant la définition d'une moyenne de Rademacher. On nomme suite de Bernoulli ou suite de Rademacher une suite $(\sigma_i)_{1\leq i\leq n}$ de variables aléatoires indépendantes vérifiant : $\forall i \in \{1, \ldots, n\}$, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$.

Définition 25 (Moyenne de Rademacher) *Pour $n \in \mathbb{N}^*$, soient $\mathcal{A}$ un ensemble borné de vecteurs $a = (a_i)_{1\leq i\leq n}$ appartenant à $\mathbb{R}^n$ et $(\sigma_i)_{1\leq i\leq n}$ une suite de Rademacher. La* moyenne de Rademacher associée à $\mathcal{A}$, $\mathcal{R}_n(\mathcal{A})$, *est définie par :*

$$\mathcal{R}_n(\mathcal{A}) = \mathbb{E}\sup_{a\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i a_i\right|.$$

L'utilisation de moyennes de Rademacher se trouve au centre de nombreux résultats en théorie des processus empiriques et théorie des bornes. On dispose pour ces moyennes d'un vaste ensemble de propriétés, au nombre desquelles se trouve le principe de contraction, dont la première formulation correspond au théorème 4.12 de [134], et qui nous sera utile dans la suite.

Théorème 7 (Principe de contraction, d'après le théorème 4.15 de [201]) *$\mathcal{A}$ et n étant définis comme dans la définition 25, soit ϕ une fonction lipschitzienne de rapport L_ϕ telle que $\phi(0) = 0$. Notons*

$$\phi(\mathcal{A}) = \left\{(\phi(a_i))_{1\leq i\leq n} \ : (a_i)_{1\leq i\leq n} \in \mathcal{A}\right\}.$$

Alors :

$$\mathcal{R}_n(\phi(\mathcal{A})) \leq 2L_\phi \mathcal{R}_n(\mathcal{A}).$$

Comme dans le cas du théorème 1, la construction de la borne exposée dans cette section s'appuie sur la définition d'un nouveau risque et l'utilisation d'une inégalité de concentration. En nous inspirant de la fonction de perte de la M-SVM de Crammer et Singer, nous considérons le risque suivant :

$$\tilde{R}(h) = \mathbb{E}\left[(1 - \Delta h_Y(X))_+\right].$$

$\tilde{R}_m(h)$ désigne le risque empirique correspondant, mesuré sur un m-échantillon. La preuve du théorème 1 faisait intervenir l'inégalité de Hoeffding [112]. L'inégalité de concentration à la base de notre borne s'appuyant sur une moyenne de Rademacher est l'inégalité des différences bornées.

Théorème 8 (Inégalité des différences bornées, [157]) *Pour n appartenant à $\mathbb{N}^*$, soit $(T_i)_{1\leq i\leq n}$ une suite de n variables aléatoires indépendantes à valeurs dans un ensemble $\mathcal{T}$. Soit g une fonction de $\mathcal{T}^n$ dans $\mathbb{R}$ telle qu'il existe une suite de constantes positives $(c_i)_{1\leq i\leq n}$ vérifiant :*

$$\forall i \in \{1,\ldots,n\}, \quad \sup_{(t_i)_{1\leq i\leq n}\in\mathcal{T}^n, t_i'\in\mathcal{T}} |g(t_1,\ldots,t_n) - g(t_1,\ldots,t_{i-1},t_i',t_{i+1},\ldots,t_n)| \leq c_i.$$

Alors, pour toute valeur strictement positive de τ, la variable aléatoire $g(T_1,\ldots,T_n)$ satisfait les inégalités suivantes :

$$\mathbb{P}\{g(T_1,\ldots,T_n) - \mathbb{E}g(T_1,\ldots,T_n) > \tau\} \leq e^{-\frac{2\tau^2}{c}}$$

et

$$\mathbb{P}\{\mathbb{E}g(T_1,\ldots,T_n) - g(T_1,\ldots,T_n) > \tau\} \leq e^{-\frac{2\tau^2}{c}}$$

où $c = \sum_{i=1}^n c_i^2$.

Ces définitions et résultats de base étant posés, nous démontrons la borne suivante.

Théorème 9 (Théorème 6 dans [94]) *Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une M-SVM à Q catégories sous les hypothèses 1 et la contrainte supplémentaire $\mathbf{b} = 0$. Soit $K_{\bar{\mathcal{H}}} = \Lambda_w\Lambda_{\Phi(\mathcal{X})} + 1$. Avec une probabilité au moins égale à $1-\delta$, le risque de toute fonction $\bar{h}$ de $\bar{\mathcal{H}}$ est borné supérieurement de la manière suivante :*

$$R(\bar{h}) \leq \tilde{R}_m(\bar{h}) + \frac{4}{\sqrt{m}} + \frac{4Q(Q-1)\Lambda_w}{m}\sqrt{\sum_{i=1}^{m}\kappa(X_i,X_i)} + K_{\bar{\mathcal{H}}}\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{2.19}$$

Preuve Etant donné que $\mathbb{1}_{\{\Delta^{\#}\bar{h}_Y(X)\leq 0\}} \leq \left(1-\Delta\bar{h}_Y(X)\right)_+$, la proposition 3 implique la majoration suivante :

$$R(\bar{h}) \leq \tilde{R}(\bar{h}).$$

En conséquence,

$$R(\bar{h}) \leq \tilde{R}_m(\bar{h}) + \sup_{\bar{h}'\in\bar{\mathcal{H}}}\left(\tilde{R}(\bar{h}') - \tilde{R}_m(\bar{h}')\right). \tag{2.20}$$

La suite de cette preuve est dédiée à la majoration du supremum du processus empirique apparaissant dans (2.20). Notons Z le couple aléatoire (X,Y) et Z_i ses copies constituant le m-échantillon D_m : $D_m = (Z_i)_{1\leq i\leq m}$. Cette simplification des notations étant posée, l'inégalité des différences bornées s'applique au supremum qui nous intéresse, en posant $n=m$, $(T_i)_{1\leq i\leq n} = D_m$ (i.e., $T_i = Z_i$), $g(T_1,\ldots,T_n) = \sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right)$. Du fait des hypothèses du théorème 9, les fonctions $\bar{h}$ de $\bar{\mathcal{H}}$ et donc les fonctions $\Delta\bar{h}$ de $\Delta\bar{\mathcal{H}}$ prennent leurs valeurs dans $[-M_{\bar{\mathcal{H}}}, M_{\bar{\mathcal{H}}}]^Q$ avec $M_{\bar{\mathcal{H}}} = \Lambda_w\Lambda_{\Phi(\mathcal{X})}$. En conséquence, la fonction de perte associée au risque $\tilde{R}$ prend ses valeurs dans l'intervalle $[0, M_{\bar{\mathcal{H}}}+1]$. Notons $K_{\bar{\mathcal{H}}} = M_{\bar{\mathcal{H}}}+1$. On peut alors choisir la suite $(c_i)_{1\leq i\leq m}$ de la manière suivante : $\forall i \in \{1,\ldots,m\}$, $c_i = \frac{K_{\bar{\mathcal{H}}}}{m}$. Comme nous souhaitons simplement majorer le supremum, c'est la première inégalité qui est employée. On obtient le résultat suivant : avec une probabilité au moins égale à $1-\delta$,

$$\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) \leq \mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) + K_{\bar{\mathcal{H}}}\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{2.21}$$

L'introduction d'un échantillon fantôme $D'_m = ((X'_i, Y'_i))_{1\leq i\leq m} = (Z'_i)_{1\leq i\leq m}$, possédant les mêmes propriétés que l'échantillon initial D_m, et indépendant de celui-ci, permet d'appliquer une symétrisation. Soient $\tilde{R}_{D_m}(\bar{h})$ et $\tilde{R}_{D'_m}(\bar{h})$ les risques empiriques fonctions respectivement de D_m et D'_m (on a donc $\tilde{R}_{D_m}(\bar{h}) = \tilde{R}_m(\bar{h})$). Alors,

$$\mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) = \mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\mathbb{E}\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) = \mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\mathbb{E}\left[\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})|D_m\right]\right).$$

Le supremum étant convexe, l'application de l'inégalité de Jensen (voir par exemple le théorème A.18 de [65]) donne :

$$\mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\mathbb{E}\left[\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})|D_m\right]\right) \leq \mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right).$$

Par substitution dans (2.21), on en déduit la majoration suivante :

$$\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) \leq \mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) + K_{\bar{\mathcal{H}}}\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}.$$

$$\begin{aligned}\mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) &= \mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\sum_{i=1}^m\left(\left(1-\Delta\bar{h}_{Y'_i}(X'_i)\right)_+ - \left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right)\right]\\ &\leq \mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^m\left(\left(1-\Delta\bar{h}_{Y'_i}(X'_i)\right)_+ - \left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right)\right|\right].\end{aligned}$$

Notons D_{2m} l'union des échantillons D_m et D'_m. A ce niveau, l'introduction d'une pondération par des variables de Rademacher peut être envisagée comme l'application d'une permutation sur D_{2m}, permutation constituée uniquement de transpositions entre éléments de mêmes indices dans l'échantillon initial et l'échantillon fantôme. Naturellement, une telle permutation préserve la mesure produit P^{2m}. On a donc :

$$\mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) \leq \mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^m\sigma_i\left(\left(1-\Delta\bar{h}_{Y'_i}(X'_i)\right)_+ - \left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right)\right|\right].$$

$$\mathbb{E}\sup_{\bar{h}\in\bar{\mathcal{H}}}\left(\tilde{R}_{D'_m}(\bar{h})-\tilde{R}_{D_m}(\bar{h})\right)\leq 2\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right|\right]. \tag{2.22}$$

Soulignons que dans la formule 2.22, le terme de gauche correspond à une espérance par rapport à D_{2m}, tandis que le terme de droite correspond à une espérance par rapport à $\sigma=(\sigma_i)_{1\leq i\leq m}$. En notant ϕ_+ la fonction qui à t associe $(t)_+$ et a le vecteur aléatoire de terme général $a_i=1-\Delta\bar{h}_{Y_i}(X_i)$, le terme de droite de la formule 2.22, au coefficient 2 près, se réécrit sous la forme suivante :

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\phi_+(a_i)\right|\right]=\mathcal{R}_m\left(\phi_+(\mathcal{A})\right)$$

où $\mathcal{A}$ est l'ensemble des vecteurs a obtenus lorsque $\bar{h}$ décrit la famille $\bar{\mathcal{H}}$ (plus précisément sa restriction induite par les hypothèses faites). La fonction ϕ_+ étant lipschitzienne de rapport 1, l'application du théorème 7 fournit une majoration de cette dernière expression en fonction de la moyenne de Rademacher associée à $\mathcal{A}$ qui est donnée par :

$$\begin{aligned}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right|\right]&\leq 2\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(1-\Delta\bar{h}_{Y_i}(X_i)\right)\right|\right]\\&\leq 2\left(\mathbb{E}\left[\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\right|\right]+\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|-\sum_{i=1}^{m}\sigma_i\Delta\bar{h}_{Y_i}(X_i)\right|\right]\right)\\&\leq 2\left(\frac{1}{\sqrt{m}}+\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\Delta\bar{h}_{Y_i}(X_i)\right|\right]\right).\end{aligned}$$

$$\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\Delta\bar{h}_{Y_i}(X_i)\right|\right]=\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\frac{1}{2}\left(\bar{h}_{Y_i}(X_i)-\max_{k\neq Y_i}\bar{h}_k(X_i)\right)\right|\right].$$

Les calculs effectués jusqu'à présent ne sont propres aux M-SVMs qu'en ce qu'ils utilisent, pour la détermination de la valeur de la constante $K_{\bar{\mathcal{H}}}$, les hypothèses 1. La suite des calculs s'appuie directement sur le fait que la famille $\bar{\mathcal{H}}$ est construite autour d'un RKHS induit par le noyau κ. Pour toute fonction $\bar{h}$ de $\bar{\mathcal{H}}$ et tout couple (X_i,Y_i) de D_m, notons K_i^* la variable aléatoire correspondant au plus petit indice de catégorie tel que $\bar{h}_{K_i^*}(X_i)=\max_{k\neq Y_i}\bar{h}_k(X_i)$. On a donc $K_i^*=\operatorname{argmax}_{k\neq Y_i}\bar{h}_k(X_i)$, lorsque cette dernière expression est définie sans ambiguïté. Cela permet d'écrire :

$$\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\left(\bar{h}_{Y_i}(X_i)-\max_{k\neq Y_i}\bar{h}_k(X_i)\right)\right|\right]=\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\left(\bar{h}_{Y_i}(X_i)-\bar{h}_{K_i^*}(X_i)\right)\right|\right].$$

Notons à présent $\mathcal{P}_m$ l'ensemble des applications p_m de $\{1,\ldots,m\}$ dans $\{1,\ldots,Q\}^2$ telles que le couple $p_m(i)$ est toujours constitué de deux valeurs différentes. Ces applications définissent donc des partitions de $\{1,\ldots,m\}$ en $Q(Q-1)$ classes. Alors, en s'appuyant sur ces partitions, on peut éliminer des calculs les couples aléatoires (Y_i,K_i^*), propres au cas multiclasse, de la manière suivante :

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\left(\bar{h}_{Y_i}(X_i)-\bar{h}_{K_i^*}(X_i)\right)\right|\right]&\leq\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sup_{p_m\in\mathcal{P}_m}\left|\sum_{k\neq l}\sum_{i\,:\,p_m(i)=(k,l)}\sigma_i\left(\bar{h}_k(X_i)-\bar{h}_l(X_i)\right)\right|\right]\\&\leq\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sup_{p_m\in\mathcal{P}_m}\sum_{k\neq l}\left|\sum_{i\,:\,p_m(i)=(k,l)}\sigma_i\langle h_k-h_l,\kappa(X_i,.)\rangle\right|\right].\end{aligned} \tag{2.23}$$

Du fait des hypothèses formulées sur la famille de fonctions $\bar{\mathcal{H}}$, l'inégalité de Cauchy-Schwarz fournit la majoration suivante : $\forall \bar{h} \in \bar{\mathcal{H}}$, $\forall p_m \in \mathcal{P}_m$, $\forall (k,l) \in \{1,\ldots,Q\}^2$, $k \neq l$,

$$\frac{1}{2}\left|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \langle h_k - h_l, \kappa\left(X_i,.\right)\rangle\right| = \frac{1}{2}\left|\langle h_k - h_l, \sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \kappa\left(X_i,.\right)\rangle\right|$$

$$\leq \Lambda_w \left\|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \kappa\left(X_i,.\right)\right\|.$$

Par suite,

$$\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sup_{p_m\in\mathcal{P}_m}\sum_{k\neq l}\left|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \langle h_k - h_l, \kappa\left(X_i,.\right)\rangle\right|\right] \leq \Lambda_w \mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sup_{p_m\in\mathcal{P}_m}\sum_{k\neq l}\left\|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \kappa\left(X_i,.\right)\right\|\right]$$

$$= \Lambda_w \mathbb{E}\left[\sup_{p_m\in\mathcal{P}_m}\sum_{k\neq l}\left\|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \kappa\left(X_i,.\right)\right\|\right] \leq \Lambda_w \sum_{k\neq l}\mathbb{E}\left[\sup_{p_m\in\mathcal{P}_m}\left\|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \kappa\left(X_i,.\right)\right\|\right]. \quad (2.24)$$

En conséquence, pour achever la dérivation de la borne, il reste à trouver une expression majorant uniformément les expressions de la forme :

$$\mathbb{E}\left\|\sum_{i\in\mathcal{I}_m} \sigma_i \kappa\left(X_i,.\right)\right\|,$$

où $\mathcal{I}_m$ est un sous-ensemble de $\{1,\ldots,m\}$. Les auteurs de [39] utilisent pour cela l'inégalité de Khintchine (voir par exemple [134]), mais l'inégalité de Jensen fournit ici encore le résultat désiré.

$$\mathbb{E}\left\|\sum_{i\in\mathcal{I}_m} \sigma_i \kappa\left(X_i,.\right)\right\| = \mathbb{E}\left[\langle \sum_{i\in\mathcal{I}_m} \sigma_i \kappa\left(X_i,.\right), \sum_{j\in\mathcal{I}_m} \sigma_j \kappa\left(X_j,.\right)\rangle^{\frac{1}{2}}\right] \quad (2.25)$$

Par application de l'inégalité de Jensen, le terme de droite de l'équation 2.25 est majoré comme suit :

$$\mathbb{E}\left[\langle \sum_{i\in\mathcal{I}_m} \sigma_i \kappa\left(X_i,.\right), \sum_{j\in\mathcal{I}_m} \sigma_j \kappa\left(X_j,.\right)\rangle^{\frac{1}{2}}\right] \leq \left(\mathbb{E}\left[\sum_{i\in\mathcal{I}_m}\sum_{j\in\mathcal{I}_m} \sigma_i\sigma_j \kappa\left(X_i,X_j\right)\right]\right)^{\frac{1}{2}} = \left(\sum_{i\in\mathcal{I}_m} \kappa\left(X_i,X_i\right)\right)^{\frac{1}{2}}.$$

Notons au passage que ce raisonnement est précisément celui permettant d'établir le résultat partiel utilisé plus haut : $\mathbb{E}\left[\frac{1}{m}\left|\sum_{i=1}^m \sigma_i\right|\right] \leq \frac{1}{\sqrt{m}}$.

$$\forall \mathcal{I}_m \subset \{1,\ldots,m\},\ \mathbb{E}\left\|\sum_{i\in\mathcal{I}_m} \sigma_i \kappa\left(X_i,.\right)\right\| \leq \left(\sum_{i\in\mathcal{I}_m} \kappa\left(X_i,X_i\right)\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^m \kappa\left(X_i,X_i\right)\right)^{\frac{1}{2}}$$

On obtient donc en résumé :

$$\forall p_m \in \mathcal{P}_m,\ \forall (k,l) \in \{1,\ldots,Q\}^2,\ k\neq l,\ \mathbb{E}\left[\sup_{p_m\in\mathcal{P}_m}\left\|\sum_{i\,:\,p_m(i)=(k,l)} \sigma_i \kappa\left(X_i,.\right)\right\|\right] \leq \left(\sum_{i=1}^m \kappa\left(X_i,X_i\right)\right)^{\frac{1}{2}}.$$

Par substitution dans le terme de droite de (2.24), puis dans le terme de droite de (2.23), nous obtenons :

$$\frac{1}{2}\mathbb{E}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^m \sigma_i\left(\bar{h}_{Y_i}\left(X_i\right) - \max_{k\neq Y_i}\bar{h}_k\left(X_i\right)\right)\right|\right] \leq Q(Q-1)\Lambda_w\left(\sum_{i=1}^m \kappa\left(X_i,X_i\right)\right)^{\frac{1}{2}}.$$

En rassemblant tous les résultats partiels, on obtient la borne 2.19, ce qui achève la démonstration. ■

Notons que $\sum_{i=1}^{m} \kappa(x_i, x_i)$ est la trace de la matrice de Gram de la M-SVM, si bien que cette borne peut être estimée très aisément à partir de l'ensemble d'apprentissage. Nous avons donné de la borne 2.19 une formulation aussi proche que possible de celles des résultats biclasses similaires (voir par exemple le théorème 4.12 de [201]), c'est-à-dire faisant apparaître dans le terme de contrôle les termes empiriques $\kappa(X_i, X_i)$. Naturellement, les hypothèses 1 permettent de simplifier l'écriture, en utilisant la majoration $\frac{1}{m}\sqrt{\sum_{i=1}^{m} \kappa(X_i, X_i)} \leq \sqrt{\frac{1}{m}}\Lambda_{\Phi(\mathcal{X})}$. L'utilisation d'une moyenne de Rademacher fournit donc un risque garanti dont le terme de contrôle décroît en $\sqrt{\frac{1}{m}}$. Ce taux, qui est également celui de la borne VC standard, est optimal. Il reste cependant une marge de progression dans la détermination des constantes. Une question ouverte consiste à déterminer si le terme $Q(Q-1)$ peut être remplacé par C_Q^2. En reprenant les notations de la preuve, il s'agit d'établir s'il est possible de traiter globalement les deux ensembles $\{i : p_m(i) = (k,l)\}$ et $\{i : p_m(i) = (l,k)\}$. Un autre avantage de cette approche par rapport à celles décrites dans les deux sections précédentes, reposant sur le théorème 1, est de faciliter la formulation de bornes dédiées aux différentes M-SVMs. En reprenant les notations du théorème 9, il suffit d'adapter la définition du risque $\tilde{R}$ afin qu'elle prenne en compte la fonction de perte $\ell_{\text{M-SVM}}$ appropriée, tout en restant compatible avec la contrainte $R(\bar{h}) \leq \tilde{R}(\bar{h})$ et l'application du principe de contraction.

2.5.2 Autres SVM multiclasses

La famille des fonctions réalisables par la version multiclasse de la LS-SVM est exactement la même que celle correspondant à la mise en œuvre d'une méthode de décomposition fondée sur l'emploi de la SVM biclasse standard et d'une matrice de mots codes. Il est donc possible de borner le risque de cette SVM multiclasse en s'appuyant sur les résultats disposibles pour ces méthodes de décomposition.

L'étude des performances en généralisation de la SVM multiclasse d'Anguita et co-auteurs peut s'appuyer sur le calcul d'une borne sur sa dimension de Natarajan à marge. Nous détaillons ci-dessous ce calcul. Soit x un élément de $\mathcal{X}$. En reprenant les notations de la section 2.4.2.2, nous désignons par $\left(x^{(k)}\right)_{1\leq k\leq Q}$ la suite de ses reformulations correspondant aux différentes catégories. Soit $\tilde{\mathcal{X}}$ l'espace des données reformulées : $\tilde{\mathcal{X}} = \left\{x^{(k)} : x \in \mathcal{X},\ k \in \{1,\ldots,Q\}\right\}$. On dispose alors du résultat suivant, qui représente une variante du lemme 4.2 de [21] :

Lemme 7 *Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une SVM multiclasse d'Anguita et co-auteurs à Q catégories sous les contraintes $\|w\| \leq \Lambda_w$ et $b = 0$. Soit $\epsilon \in \mathbb{R}_+^*$. Si un sous-ensemble $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ de $\mathcal{X}$ est N-pulvérisé avec une marge ϵ par $\Delta\bar{\mathcal{H}}$, alors pour toute partition de cet ensemble en deux sous-ensembles s_1 et s_2, on dispose de la majoration suivante :*

$$\left\| \sum_{x_i \in s_1} \left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) - \sum_{x_i \in s_2} \left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) \right\| \geq \frac{2n}{\Lambda_w}\epsilon. \tag{2.26}$$

Preuve Supposons que $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ soit un sous-ensemble de $\mathcal{X}$ N-pulvérisé avec une marge ϵ par $\Delta\bar{\mathcal{H}}$. Soit $(I(s_{\mathcal{X}^n}), v_b)$ un témoin de cette pulvérisation avec $I(s_{\mathcal{X}^n}) = \{(i_1(x_i), i_2(x_i)) : 1 \leq i \leq n\}$ et $v_b = (b_i)_{1\leq i\leq n}$. Par définition, pour tout vecteur $v_y = (y_i)$ de $\{-1,1\}^n$, il existe une fonction $\bar{h}_y$ dans $\bar{\mathcal{H}}$ caractérisée par le vecteur w_y telle que :

$$\forall i \in \{1,\ldots,n\},\ \begin{cases} \text{si } y_i = \quad 1, & \Delta\bar{h}_{y,i_1(x_i)}(x_i) - b_i \geq \epsilon \\ \text{si } y_i = \ -1, & \Delta\bar{h}_{y,i_2(x_i)}(x_i) + b_i \geq \epsilon \end{cases}.$$

Par définition de $\bar{\mathcal{H}}$ et de l'opérateur de marge Δ, ceci est équivalent à :

$$\forall i \in \{1,\ldots,n\},\ \begin{cases} \text{si } y_i = \quad 1, & \frac{1}{2}\left(\langle w_y, \Phi\left(x_i^{(i_1(x_i))}\right)\rangle - \max_{k\neq i_1(x_i)} \langle w_y, \Phi\left(x_i^{(k)}\right)\rangle\right) - b_i \geq \epsilon \\ \text{si } y_i = \ -1, & \frac{1}{2}\left(\langle w_y, \Phi\left(x_i^{(i_2(x_i))}\right)\rangle - \max_{k\neq i_2(x_i)} \langle w_y, \Phi\left(x_i^{(k)}\right)\rangle\right) + b_i \geq \epsilon \end{cases}$$

et implique par suite

$$\forall i \in \{1,\dots,n\}, \begin{cases} \text{si } y_i = \quad 1, & \frac{1}{2}\langle w_y, \Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\rangle - b_i \geq \epsilon \\ \text{si } y_i = -1, & \frac{1}{2}\langle w_y, \Phi\left(x_i^{(i_2(x_i))}\right) - \Phi\left(x_i^{(i_1(x_i))}\right)\rangle + b_i \geq \epsilon \end{cases}. \tag{2.27}$$

Considérons à présent une partition quelconque de $s_{\mathcal{X}^n}$ en deux sous-ensembles s_1 et s_2 et le vecteur v_y de $\{-1,1\}^n$ tel que $y_i = 1$ si $x_i \in s_1$ et $y_i = -1$ si $x_i \in s_2$. Il résulte du système 2.27 que :

$$\frac{1}{2}\langle w_y, \sum_{x_i \in s_1}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right)\rangle - \sum_{x_i \in s_1} b_i + \frac{1}{2}\langle w_y, \sum_{x_i \in s_2}\left(\Phi\left(x_i^{(i_2(x_i))}\right) - \Phi\left(x_i^{(i_1(x_i))}\right)\right)\rangle + \sum_{x_i \in s_2} b_i$$
$$\geq n\epsilon$$

ce qui se simplifie en

$$\frac{1}{2}\langle w_y, \sum_{x_i \in s_1}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) - \sum_{x_i \in s_2}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right)\rangle - \sum_{x_i \in s_1} b_i + \sum_{x_i \in s_2} b_i \geq n\epsilon.$$

Réciproquement, considérons le vecteur v_y tel que $y_i = -1$ si $x_i \in s_1$ et $y_i = 1$ si $x_i \in s_2$. On obtient alors

$$\frac{1}{2}\langle w_y, -\sum_{x_i \in s_1}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) + \sum_{x_i \in s_2}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right)\rangle + \sum_{x_i \in s_1} b_i - \sum_{x_i \in s_2} b_i \geq n\epsilon.$$

En conséquence, si $\sum_{x_i \in s_1} b_i - \sum_{x_i \in s_2} b_i \geq 0$, il existe une fonction $\bar{h}_y$ de $\bar{\mathcal{H}}$ telle que

$$\frac{1}{2}\langle w_y, \sum_{x_i \in s_1}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) - \sum_{x_i \in s_2}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right)\rangle \geq n\epsilon \tag{2.28}$$

tandis que si $\sum_{x_i \in s_1} b_i - \sum_{x_i \in s_2} b_i < 0$, il existe une autre fonction $\bar{h}_y$ de $\bar{\mathcal{H}}$ telle que

$$\frac{1}{2}\langle w_y, -\sum_{x_i \in s_1}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) + \sum_{x_i \in s_2}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right)\rangle \geq n\epsilon. \tag{2.29}$$

L'application de l'inégalité de Cauchy-Schwarz à (2.28) et (2.29) fournit dans les deux cas :

$$\frac{1}{2}\|w_y\| \left\| \sum_{x_i \in s_1}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) - \sum_{x_i \in s_2}\left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) \right\| \geq n\epsilon,$$

résultat dont la validité est donc indépendante du signe de $\sum_{x_i \in s_1} b_i - \sum_{x_i \in s_2} b_i$. En définitive, la minoration 2.26 découle de cette borne, compte tenu de l'hypothèse $\|w\| \leq \Lambda_w$. ■

Lemme 8 (Lemme 4.3 dans [21]) *Considérons une suite de points $(t_i)_{1 \leq i \leq n}$ d'un espace de Hilbert telle que $\max_{1 \leq i \leq n} \|t_i\| \leq \Lambda_t$. Alors il existe une partition de $\{t_i : 1 \leq i \leq n\}$ en deux sous-ensembles s_1 et s_2 telle que :*

$$\left\| \sum_{t_i \in s_1} t_i - \sum_{t_j \in s_2} t_j \right\| \leq \sqrt{n}\Lambda_t.$$

Notre borne sur la dimension de Natarajan à marge de la SVM multiclasse d'Anguita et co-auteurs est une conséquence directe des lemmes 7 et 8.

Théorème 10 *Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une SVM multiclasse d'Anguita et co-auteurs à Q catégories sous les contraintes $\|w\| \leq \Lambda_w$ et $b = 0$. Notons $\tilde{\mathcal{X}}$ l'espace des données reformulées. Si $\Phi\left(\tilde{\mathcal{X}}\right)$ est inclus dans la boule de rayon $\Lambda_{\Phi(\tilde{\mathcal{X}})}$ centrée sur l'origine de $E_{\Phi(\tilde{\mathcal{X}})}$, alors pour toute valeur strictement positive d'ϵ, on dispose de la borne suivante :*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \epsilon\right) \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\tilde{\mathcal{X}})}}{\epsilon}\right)^2.$$

Preuve Soit $s_{\mathcal{X}^n}$ un sous-ensemble de $\mathcal{X}$ de cardinalité n N-pulvérisé avec une marge ϵ par $\Delta\bar{\mathcal{H}}$. Soit $(I(s_{\mathcal{X}^n}), v_b)$ un témoin de cette pulvérisation. En application du lemme 7, la minoration 2.26 est vérifiée pour toute partition de $s_{\mathcal{X}^n}$ en deux sous-ensembles s_1 et s_2. De plus, d'après le lemme 8, il existe au moins une de ces partitions pour laquelle :

$$\left\| \sum_{x_i \in s_1} \left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) - \sum_{x_i \in s_2} \left(\Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right)\right) \right\| \leq$$

$$\sqrt{n} \max_{1 \leq i \leq n} \left\| \Phi\left(x_i^{(i_1(x_i))}\right) - \Phi\left(x_i^{(i_2(x_i))}\right) \right\|. \tag{2.30}$$

$\Phi\left(\tilde{\mathcal{X}}\right)$ étant inclus dans la boule de rayon $\Lambda_{\Phi(\tilde{\mathcal{X}})}$ centrée sur l'origine de $E_{\Phi(\tilde{\mathcal{X}})}$, le terme de droite de (2.30) est majoré par $2\sqrt{n}\Lambda_{\Phi(\tilde{\mathcal{X}})}$. On obtient donc :

$$\frac{2n}{\Lambda_w}\epsilon \leq 2\sqrt{n}\Lambda_{\Phi(\tilde{\mathcal{X}})},$$

soit encore

$$n \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\tilde{\mathcal{X}})}}{\epsilon}\right)^2.$$

Poser n égal à sa valeur maximale, N-dim $\left(\Delta\bar{\mathcal{H}}, \epsilon\right)$, fournit en définitive le résultat annoncé. ■

2.5.3 Méthodes de décomposition

Dans [175], les auteurs soulignent qu'il n'existe pas de borne dédiée à la méthode de décomposition un contre tous, non plus qu'à la méthode de décomposition un contre un. Ils proposent à l'inverse pour la DAGSVM un risque garanti à risque empirique nul (ce que Vapnik nomme le cas favorable). Ils s'appuient pour ce faire sur une variante de la borne standard de la dimension fat-shattering des séparateurs linéaires déjà rencontrée dans la section 2.5.1.2, borne donnée par le théorème 4.6 de [21]. Le résultat est le suivant :

Théorème 11 (Théorème 1 dans [175]) *Soit $\mathcal{G}$ la famille des fonctions réalisables par une DAGSVM à Q catégories. Sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule fermée de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$ et que la fonction g sélectionnée par l'apprentissage ne commet aucune erreur sur le m-échantillon d'apprentissage, alors avec une probabilité au moins $1 - \delta$, on dispose de la borne suivante :*

$$R(g) \leq \frac{130\Lambda^2_{\Phi(\mathcal{X})}}{m}\left(\sum_{n=1}^{C_Q^2} \frac{1}{\gamma_n^2}\log_2(4em)\log_2(4m) + \log_2\left(\frac{2(2m)^{C_Q^2}}{\delta}\right)\right),$$

où $(\gamma_n)_{1 \leq n \leq C_Q^2}$ est le vecteur des marges géométriques (biclasses) de l'ensemble des SVM associées à g.

La preuve, comme souvent en théorie des bornes, s'obtient par combinaison de plusieurs résultats standard. En l'occurence, les emprunts sont faits à l'approche utilisée dans [28] pour étudier les arbres de décision combinant des perceptrons, et à [227] pour l'utilisation d'un échantillon fantôme. Crammer et Singer démontrent dans [57] que ce résultat peut être étendu de manière immédiate afin de fournir une borne pour leur M-SVM.

Corollaire 1 (Corollaire 2 dans [57]) *Soit $\bar{\mathcal{H}}$ la famille des fonctions réalisables par une M-SVM de Crammer et Singer à Q catégories. Sous l'hypothèse que $\Phi(\mathcal{X})$ est inclus dans la boule fermée de rayon $\Lambda_{\Phi(\mathcal{X})}$ centrée sur l'origine de $E_{\Phi(\mathcal{X})}$ et que la fonction $\bar{h}$ sélectionnée par l'apprentissage ne commet aucune erreur sur le m-échantillon d'apprentissage, alors avec une probabilité au moins $1-\delta$, on dispose de la borne suivante :*

$$R(\bar{h}) \leq \frac{130\Lambda^2_{\Phi(\mathcal{X})}}{m}\left(Q\sum_{k=1}^{Q}\|w_k\|^2\log_2(4em)\log_2(4m)+\log_2\left(\frac{2(2m)^{C_Q^2}}{\delta}\right)\right).$$

Ce résultat n'est pas directement comparable à celui que produit la combinaison des théorèmes 2 et 3, dans la mesure où il repose sur l'hypothèse d'un risque empirique nul, qui est connue comme étant favorable du point de vue de la décroissance de la borne en fonction de m (voir par exemple le chapitre 4 de [229]). Crammer et Singer fournissent également, avec le corollaire 3 de [57], une borne sur le risque associé à une catégorie particulière, risque défini comme la probabilité d'observer pour cette catégorie un "faux positif" ou un "faux négatif".

2.5.4 Discussion

Les travaux relatifs aux bornes multiclasses apparaissent bien moins avancés que ceux portant sur le cas biclasse. Beaucoup d'idées demeurent à étendre, comme la condition de bruit de Mammen-Tsybakov [220] et son exploitation pour obtenir un meilleur taux de convergenge. On trouve plus de résultats relatifs aux M-SVM qu'aux méthodes de décomposition, ce qui peut sembler étonnant si l'on considère que la pratique favorise ces dernières. Parmi les bornes dédiées aux M-SVM, celle s'appuyant sur une moyenne de Rademacher est actuellement la meilleure. Un important défi consiste à obtenir une borne VC dont le terme de contrôle décroisse en $\sqrt{\frac{1}{m}}$. Plus précisément, il serait très instructif d'établir si une telle borne peut être obtenue en changeant simplement les normes dans la preuve du théorème 2. Cela devrait dans tous les cas faire apparaître des problèmes techniques originaux.

2.6 Programmation des SVM multiclasses

Les éléments de programmation mathématique évoqués dans cette section peuvent être trouvés dans [160, 80].

2.6.1 Etat de l'art

Depuis que les SVM biclasses ont été introduites, leur programmation a fait l'objet de nombreux travaux. L'enjeu est d'importance, dans la mesure où le temps de calcul requis pour l'apprentissage de ces machines (résolution du problème de programmation quadratique correspondant), est au moins proportionnel à m^2, lorsque C est petit, et m^3 lorsque C devient grand [36]. Il apparaît donc nécessaire, pour traiter des ensembles d'apprentissage de très grande taille, cas qui se présente de plus en plus souvent en pratique, de disposer d'algorithmes dédiés, utilisant des heuristiques afin de produire des solutions approchées de bonne qualité. Les plus anciens ne manqueront pas de se demander à cette occasion pourquoi tant de chercheurs s'acharnent à utiliser un modèle conçu pour les petits échantillons dans un régime plus proche de l'asymptotique, où les PMC, par exemple, donnent souvent entière satisfaction. Nous avons déjà évoqué plus haut l'algorithme SMO de Platt. Il s'agit d'une version extrême des méthodes de décomposition qui sont décrites par exemple dans [117] ou le chapitre 7 de [58]. Récemment, deux algorithmes d'apprentissage ont retenu l'attention de la communauté. LASVM [36] est un algorithme en ligne conçu pour traiter avec un faible temps d'exécution des problèmes de grande taille. S'appuyant sur SMO, il permet d'obtenir, après une seule passe sur les données, une solution de bonne qualité. Dans [50], Chapelle propose de résoudre le problème primal de manière approchée, au moyen d'un algorithme appliquant la méthode de Newton-Raphson. L'intérêt qu'il voit à son algorithme réside dans le fait qu'à temps de calcul égal, la qualité de la solution approchée qu'il fournit doit être supérieure à celle que fournirait un algorithme résolvant le dual de Wolfe. Cette analyse doit cependant être nuancée, dans la

mesure où le lien entre la qualité d'une solution approchée du problème d'apprentissage, au sens de la valeur de la fonction objectif du primal $J_{\text{M-SVM}}$, et la performance (le risque) du système discriminant correspondant, est complexe. Ces deux algorithmes font également l'objet de chapitres dans un nouveau livre [38] regroupant l'état de l'art en matière de mise en œuvre de machines à noyau sur des ensembles d'apprentissage de grande taille.

Comme toujours, le passage au cas multiclasse soulève des difficultés originales. SMO, et donc LASVM, s'étendent mal à cette situation, du fait de la complexité plus grande du problème quadratique à résoudre, liée au nombre accru de contraintes-égalités. L'algorithme de Chapelle est également difficile à étendre, car il s'appuie sur d'autres spécificités du cas biclasse. A notre connaissance, la seule tentative dans ce sens ayant conduit à une solution opérationnelle est présentée dans [250]. Elle porte sur la SVM multiclasse de Tsochantaridis et co-auteurs. Un autre algorithme d'apprentissage efficace pour cette machine, reposant sur une méthode de descente en gradient stochastique, est décrit dans [35]. Nous avons vu à la section 2.4.1.2 que Crammer et Singer ont proposé un algorithme de décomposition particulièrement efficace pour leur M-SVM. Cependant, cet algorithme s'appuie directement sur la contrainte supplémentaire $\mathbf{b} = 0$. Il ne peut donc pas être étendu pour s'appliquer aux deux autres M-SVM. Le problème est réel, dans la mesure où ajouter un prédicteur fictif à valeur constante et une composante aux vecteurs w_k afin de compenser l'absence de vecteur $\mathbf{b}$ a pour conséquence une diminution de la taille des marges géométriques. On trouvera dans [1] une alternative à l'algorithme de Crammer et Singer. De nature incrémentale, elle s'appuie cette fois sur l'algorithme SMO. La première étude comparative portant sur la mise en œuvre des méthodes de décomposition et des M-SVMs est rapportée dans [113]. Elle considère la M-SVM de Weston et Watkins et celle de Crammer et Singer, pour lesquelles elle propose un algorithme d'apprentissage incorporant une méthode de décomposition. Dans le cas de la première machine, les auteurs simplifent cependant la tâche, en utilisant le subterfuge évoqué plus haut, consistant à compenser l'absence de vecteur $\mathbf{b}$ par l'ajout d'un prédicteur fictif et d'une composante aux vecteurs w_k. Ils ne fournissent pas de comparaison entre leur méthode de décomposition et celle de Crammer et Singer. Les machines développées par Chih-Jen Lin et les membres de son équipe sont regroupées dans une librairie nommée LIBSVM, qui est librement téléchargeable. D'autres logiciels, librairies et boîtes à outils sont disponibles sur internet. Citons en particulier la boîte à outils "The Spider", de Weston, Elisseeff, Bakir et Sinz, qui propose une mise en œuvre de la M-SVM de Weston et Watkins utilisant une méthode de point intérieur [165]. La boîte à outils MATLAB de Canu et ses collaborateurs [47] fournit une autre implémentation de cette M-SVM, appliquant cette fois un algorithme de type contrainte active. Dans la section suivante, nous présentons notre algorithme d'apprentissage, inspiré de celui que propose André Elisseeff dans sa thèse [74].

2.6.2 Méthode de directions admissibles

Notre logiciel dédié à la M-SVM de Weston et Watkins (`http://www.kernel-machines.org`), décrit dans [99], utilise pour résoudre le problème de programmation quadratique correspondant à l'apprentissage un algorithme itératif qui est une variante de la méthode de Frank et Wolfe [81] (voir aussi le chapitre 5 de [160]). L'idée principale consiste à linéariser le problème de manière à limiter les besoins en termes de mémoire. L'algorithme met également en œuvre une méthode de décomposition. Dans un but de simplification, nous présentons la décomposition séparément.

La méthode de Frank et Wolfe s'applique à des problèmes avec contraintes linéaires de la forme

$$\min_t f(t)$$

$$\text{s.c.} \begin{cases} At = b \\ t \geq 0 \end{cases} .$$

Il s'agit d'une méthode itérative qui engendre une suite $\left(t^{(n)}\right)_{n\in\mathbb{N}}$ de solutions admissibles telle que pour tout n, $t^{(n+1)}$ est déterminée à partir de $t^{(n)}$ en deux étapes.

1. On résoud le programme linéaire LP $\left(t^{(n)}\right)$ donné par :

$$\min_u \left\{ \nabla f \left(t^{(n)}\right)^T u \right\}$$

$$\text{s.c.} \left\{ \begin{array}{l} Au = b \\ u \geq 0 \end{array} \right. .$$

2. Soit $u^{(n)}$ un point extrême du polytope défini par les contraintes qui soit solution optimale de LP $\left(t^{(n)}\right)$. $t^{(n+1)}$ est choisie de façon à minimiser f sur le segment $\left[t^{(n)}, u^{(n)}\right]$.

L'application de cette méthode à la M-SVM de Weston et Watkins est immédiate. On peut en particulier choisir pour point de départ $\alpha^{(0)} = 0_{Qm}$. En notant $J_{\mathrm{WW,d}}$ la fonction objectif du problème 3, le programme linéaire à résoudre à l'itération $n+1$ est le suivant :

Problème 14

$$\min_\zeta \left\{ \nabla J_{WW,d} \left(\alpha^{(n)}\right)^T \zeta \right\}$$

$$s.c. \left\{ \begin{array}{ll} 0 \leq \zeta_{ik} \leq C, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{x_i \in k} \sum_{l=1}^{Q} \zeta_{il} - \sum_{i=1}^{m} \zeta_{ik} = 0, & (1 \leq k \leq Q-1) \end{array} \right. .$$

Soit $\theta^{(n)} \in [0,1]$ le coefficient de la combinaison convexe optimale entre $\alpha^{(n)}$ et $\zeta^{(n)}$, i.e.

$$\theta^{(n)} = \underset{\theta \in [0,1]}{\operatorname{argmin}} J_{\mathrm{WW,d}} \left((1-\theta)\, \alpha^{(n)} + \theta \zeta^{(n)} \right).$$

Son expression analytique est donnée par :

$$\theta^{(n)} = \min \left\{ \frac{\nabla J_{\mathrm{WW,d}}(\alpha^{(n)})^T \delta^{(n)}}{\delta^{(n)T} H_{\mathrm{WW}} \delta^{(n)}},\ 1 \right\}$$

où $\delta^{(n)} = \alpha^{(n)} - \zeta^{(n)}$.

Les principales difficultés rencontrées lorsque l'on tente de résoudre directement le problème 3, que ce soit par la méthode de Frank et Wolfe ou tout autre algorithme standard, découlent de la manipulation de la matrice hessienne H_{WW}. D'une part, dans de nombreux cas, cette matrice est trop grande pour être stockée en mémoire, dans la mesure où elle appartient à $\mathcal{M}_{Qm,Qm}(\mathbb{R})$. D'autre part, le calcul de ses termes, donnés par l'équation 2.8, peut prendre beaucoup de temps si le noyau est complexe, puisqu'il repose sur le calcul des termes $\kappa(x_i, x_j)$ de la matrice de Gram. Un moyen naturel de surmonter la première difficulté, parfois au détriment de la seconde, consiste à appliquer une méthode de décomposition. Cette approche était déjà mise en œuvre dans les premiers travaux portant sur les SVM [37]. La méthode de "chunking" utilisée par les auteurs fut introduite dans [227] pour le cas d'un modèle linéaire. Les principales méthodes de décomposition introduites par la suite (voir en particulier le chapitre 7 de [58]) consistent à résoudre le problème dual en figeant les valeurs d'une partie des variables. Ce cadre général est à présent détaillé dans le cas du problème 14.

Pour simplifier les notations, mais sans perte de généralité, nous faisons l'hypothèse que l'ensemble de travail est constitué des variables duales α_B associées aux N_B premiers exemples de l'ensemble d'apprentissage, les valeurs des variables duales α_H associées aux $N_H = m - N_B$ derniers exemples étant fixées. La fonction objectif du problème 3 se réécrit alors sous la forme suivante :

$$J_{\mathrm{WW,d}}(\alpha) = \frac{1}{2} \begin{pmatrix} \alpha_B \\ \alpha_H \end{pmatrix}^T \begin{pmatrix} H_{BB} & H_{BH} \\ H_{HB} & H_{HH} \end{pmatrix} \begin{pmatrix} \alpha_B \\ \alpha_H \end{pmatrix} - 1_{Qm}^T \begin{pmatrix} \alpha_B \\ \alpha_H \end{pmatrix}.$$

Cette fonctionnelle peut de nouveau être réécrite comme suit :

$$J_{\mathrm{WW,d}}(\alpha) = \frac{1}{2} \alpha_B^T H_{BB} \alpha_B - \left(1_{QN_B}^T - \alpha_H^T H_{HB}\right) \alpha_B + \frac{1}{2} \alpha_H^T H_{HH} \alpha_H - 1_{QN_H}^T \alpha_H.$$

Nous obtenons donc :

$$\nabla J_{\mathrm{WW,d}}(\alpha_B) = H_{BB}\alpha_B + H_{BH}\alpha_H - 1_{QN_B} = \begin{pmatrix} H_{BB} & H_{BH} \end{pmatrix} \alpha - 1_{QN_B}.$$

Cette dernière formule met en évidence le fait que la décomposition laisse inchangée l'expression du gradient de la fonction objectif par rapport aux variables de l'ensemble de travail. Par suite, le temps requis pour calculer le gradient partiel (partie principale de la première étape de l'algorithme de Frank et Wolfe) est égal à $\frac{N_B}{m}$ fois le temps requis pour calculer le gradient entier. Le gain concernant le temps requis pour calculer le pas optimal $\theta^{(n)}$ (seconde étape de l'algorithme) est plus grand encore, puisque le nouveau dénominateur à calculer est :

$$\left\{\zeta_B^{(k)} - \alpha_B^{(k)}\right\}^T H_{BB} \left\{\zeta_B^{(k)} - \alpha_B^{(k)}\right\}.$$

Le nombre de termes de cette forme quadratique est proportionnel à N_B^2 au lieu de m^2.

L'efficacité d'une méthode de décomposition dépend à l'évidence de la manière dont l'ensemble de travail est sélectionné. Le lecteur intéressé pourra trouver dans [195, 58] un panorama des possibilités les plus utilisées dans le cas biclasse. La littérature multiclasse est pratiquement muette sur le sujet et nos propres investigations n'ont pas encore fourni de résultat significatif. L'heuristique mise en œuvre dans notre logiciel a pour but principal le contrôle du nombre de variables actives, de manière à limiter le temps de calcul et la complexité de la solution produite.

2.7 Sélection de modèle

Dans le cas des SVM multiclasses, comme dans celui des SVM biclasses, la sélection de modèle correspond au choix du noyau et à la détermination de la valeur des hyperparamètres, qui sont de deux types : la constante de marge douce C et les paramètres du noyau. Néanmoins, la multiplicité des catégories appelle, ici encore, des solutions dédiées. Dans le cas biclasse, une grande partie des travaux portant sur le choix du noyau des SVM ou de l'analyse discriminante à noyau étudie la combinaison linéaire de noyaux [124]. A notre connaissance, les seuls travaux multiclasses de ce type sont décrits dans des articles en cours de relecture. Actuellement, la littérature se concentre exclusivement sur les méthodes de décomposition et les M-SVM. Pour ces machines, elle traite en premier lieu la détermination de la constante de marge douce et la paramétrisation de noyaux gaussiens. Dans cette section, la sélection de modèle est abordée comme un problème d'optimisation dont la fonction objectif est un majorant ou une estimation d'un risque empirique. La manière dont cette optimisation est réalisée en fonction des hyperparamètres considérés n'est pas détaillée. Cependant, le choix de la constante de marge douce est l'application privilégiée.

Initialement, la méthode de choix pour effectuer une sélection de modèle avec des SVM était la validation croisée. On connaît depuis longtemps les problèmes que son utilisation peut soulever (voir en particulier la référence de base sur le sujet, [212], ou [46, 106, 27] pour des résultats plus récents). Cependant, des résultats positifs sont également disponibles, comme celui de Luntz et Brailovsky [148, 229, 230] stipulant que la procédure "leave-one-out" produit un estimateur de l'erreur en généralisation presque sans biais.

Lemme 9 ([148]) *Considérons un modèle d'apprentissage associé à une classe de fonctions $\mathcal{G}$ et un m-échantillon d'apprentissage. Soit $\mathcal{L}_m$ le nombre des erreurs du modèle résultant de la mise en œuvre sur l'échantillon d'apprentissage d'une procédure de validation croisée "leave-one-out". Alors,*

$$\mathbb{E}R\left(g_{m-1}\right) = \mathbb{E}\left(\frac{\mathcal{L}_m}{m}\right), \tag{2.31}$$

où $R\left(g_{m-1}\right)$ est le risque (la probabilité d'erreur en généralisation) d'une fonction de $\mathcal{G}$ sélectionnée par apprentissage sur un échantillon de taille $m-1$.

L'estimateur est presque sans biais dans la mesure où le risque $R\left(g_{m-1}\right)$ correspond à une taille de l'échantillon d'apprentissage égale à $m-1$ et non m. Notons qu'une formulation alternative de ce lemme

remplace position à position les fonctions sélectionnées par l'apprentissage par celles minimisant le risque empirique. Dans [76], les auteurs proposent un panorama des arguments justifiant l'utilisation de l'erreur "leave-one-out" en apprentissage. D'un point de vue purement pratique, cette configuration extrême de la validation croisée s'avère très coûteuse en temps de calcul. Ces observations ont conduit de nombreux auteurs à proposer des bornes supérieures sur l'erreur empirique "leave-one-out" des SVM.

2.7.1 Bornes sur l'erreur empirique "leave-one-out"

Parmi les bornes proposées pour le cas biclasse, on peut en particulier citer celles de Jaakkola et Haussler [116], de Wahba et ses co-auteurs [235] de Opper et Winther [168], ainsi que la "span bound" de Chapelle et Vapnik [230] (voir [51] pour un état de l'art sur le sujet). La dernière est reconnue comme étant la plus fine. La plus utilisée est probablement la borne "rayon-marge", qui représente un bon compromis entre qualité et temps de calcul. Plus précisément, elle s'avère presque aussi efficace que la "span bound" pour déterminer la valeur des hyperparamètres, tandis que les calculs qu'elle nécessite sont moins lourds. La formulation de cette borne fait intervenir une SVM à marge dure. Elle ne s'étend donc aux M-SVM à marge douce que dans des versions se ramenant au cas de la marge dure, c'est-à-dire pour lesquelles la contribution empirique à la fonction objectif, quadratique, est convenablement choisie (voir la section 2.4.1.4).

2.7.1.1 Borne "rayon-marge"

Théorème 12 ([229]) *Considérons une SVM biclasse à marge dure sur un domaine $\mathcal{X}$. En notant $\gamma = \frac{1}{\|w\|}$ sa marge géométrique sur l'ensemble d'apprentissage $((x_i, y_i))_{1 \leq i \leq m}$, on obtient pour majorant de $\mathcal{L}_m$:*

$$\mathcal{L}_m \leq \frac{\mathcal{D}_m^2}{\gamma^2}, \tag{2.32}$$

où $\mathcal{D}_m$ est le diamètre de la plus petite boule de l'espace de représentation contenant les vecteurs support.

On remarquera qu'ici, la boule de rayon minimal considérée n'est pas supposée être centrée sur l'origine de $E_{\Phi(\mathcal{X})}$. La détermination de $\mathcal{D}_m$ s'obtient par résolution d'un problème de programmation quadratique similaire à celui correspondant à l'apprentissage d'une SVM (voir par exemple l'algorithme de "support vector clustering" (SVC) [25]). Nous avons indiqué que cette borne conjugue deux avantages, celui d'être simple et celui d'être efficace pour choisir les valeurs des hyperparamètres. Sa simplicité ne réside pas seulement dans le peu de difficulté de son calcul, mais également dans la facilité avec laquelle se calcule son gradient par rapport aux hyperparamètres. Cela permet de réduire l'algorithme de sélection de modèle à une descente en gradient (voir [51]).

2.7.1.2 Extension multiclasse de Wang et co-auteurs

Dans [236], Wang et ses co-auteurs proposent deux extensions multiclasses de la borne "rayon-marge" dédiées à la méthode de décomposition un contre un. Il s'agit d'extensions au sens large, dans la mesure où les expressions faisant intervenir rayons et marges ne constituent pas précisément une borne sur $\mathcal{L}_m$. La première de ces expressions, le terme de droite de la formule 2.33, est justifiée par la proposition suivante, conséquence directe du théorème 12.

Proposition 7 *Considérons la méthode de décomposition un contre un utilisant des SVM à marge dure comme classifieurs de base. Soient $h_{kl}(x) = \langle w_{kl}, \Phi(x) \rangle + b_{kl}$ la fonction calculée par le classifieur binaire entraîné à distinguer les catégories d'indices k et l, $\mathcal{D}_m(k,l)$ le diamètre de la plus petite boule de l'espace de représentation contenant ses vecteurs support et $\mathcal{L}_m(k,l)$ le nombre de ses erreurs résultant de la mise en œuvre sur l'ensemble d'apprentissage d'une procédure de validation croisée "leave-one-out". Alors,*

$$\sum_{1 \leq k < l \leq Q} \mathcal{L}_m(k,l) \leq \sum_{1 \leq k < l \leq Q} \mathcal{D}_m(k,l)^2 \left\| w_{kl} \right\|^2. \tag{2.33}$$

La seconde "borne" s'appuie sur des considérations différentes, fondées sur la notion de matrices de dispersion [84, 69]. Elle est égale à $\frac{\mathcal{D}_m^2}{\bar{\gamma}^2}$, où l'expression du carré de la "marge" $\bar{\gamma}$ est la suivante :

$$\bar{\gamma}^2 = \frac{1}{m^2} \sum_{k<l} \frac{m_k m_l}{\|w_{kl}\|^2}. \tag{2.34}$$

Nous présentons à présent des bornes exactes sur l'erreur empirique "leave-one-out" des différentes M-SVM (à marge dure).

2.7.1.3 Bornes "rayon-marge" pour les M-SVM

Notre première extension porte sur les M-SVM de Weston et Watkins et Crammer et Singer (dont les versions à marge dure sont indentiques). Si sa preuve suit les mêmes étapes que celle du théorème 12, elle présente des difficultés originales. Celles-ci se traduisent par l'obtention d'un majorant faisant intervenir, outre le rayon et les marges géométriques, un coefficient multiplicatif nettement plus difficile à évaluer. Sa formulation nécessite en effet l'introduction préalable d'une suite de problèmes de programmation quadratique $(\mathrm{PQ}\,(\alpha^*)_i)_{1\leq i\leq m}$ dépendant directement de la suite des observations d'apprentissage $((x_i, y_i))_{1\leq i\leq m}$. Ces problèmes sont paramétrés par le vecteur α^* des valeurs optimales des variables duales de la M-SVM (entraînée sur tous les exemples d'apprentissage). Ils ne sont en fait définis que pour les valeurs de i telles que $\sum_{k=1}^{Q} \alpha_{ik}^* > 0$, c'est-à-dire pour les vecteurs support. La fonction objectif commune, J_{VC}, définie sur $\mathbb{R}_+^{Qm}$, est donnée par :

$$\forall \lambda \in \mathbb{R}_+^{Qm},\ J_{\mathrm{VC}}(\lambda) = \sum_{k=1}^{Q} \left(\sum_{i=1}^{m} \lambda_{ik} \right)^2. \tag{2.35}$$

Ici encore, les variables indicées par le couple (i, y_i) sont en fait des pseudo-variables, toutes égales à 0. Ces éléments étant posés, l'expression des problèmes $\mathrm{PQ}\,(\alpha^*)_i$ est la suivante :

Problème 15 ($\mathrm{PQ}\,(\alpha^*)_i$)

$$\min_{\lambda} J_{VC}(\lambda)$$

$$s.c. \begin{cases} \lambda_{ik} = \frac{\alpha_{ik}^*}{\sum_{l=1}^{Q} \alpha_{il}^*}, & (1 \leq k \leq Q) \\ 0 \leq \lambda_{jk} \leq \frac{\alpha_{jk}^*}{\sum_{l=1}^{Q} \alpha_{il}^*}, & (1 \leq j \neq i \leq m), (1 \leq k \leq Q) \\ \sum_{x_j \in k} \sum_{l=1}^{Q} \lambda_{jl} - \sum_{j=1}^{m} \lambda_{jk} = 0, & (1 \leq k \leq Q-1) \end{cases}.$$

En notand K_i la valeur de J_{VC} associée à la solution optimale de $\mathrm{PQ}\,(\alpha^*)_i$ et

$$K_{\mathrm{VC}} = \sqrt{2 \max_{i\ :\sum_{k=1}^{Q} \alpha_{ik}^* > 0} K_i}, \tag{2.36}$$

on dispose alors de la borne suivante, correspondant au théorème 2 de [61].

Théorème 13 (Borne "rayon-marge" pour la M-SVM de Weston et Watkins) *Considérons une M-SVM de Weston et Watkins (ou Crammer et Singer) à Q catégories à marge dure. Alors, en reprenant les notations de la définition 7, on obtient pour majorant de $\mathcal{L}_m$:*

$$\mathcal{L}_m \leq \frac{K_{VC}}{Q} \mathcal{D}_m^2 \sum_{k<l} \frac{(1+d_{kl})^2}{\gamma_{kl}^2} \tag{2.37}$$

où $\mathcal{D}_m$ est le diamètre de la plus petite boule de l'espace de représentation contenant les vecteurs support.

Le raisonnement établissant que ce théorème se réduit bien au théorème 12 dans le cas où $Q = 2$ est instructif. Dans ce cas en effet, les contraintes du problème 15 se simplifient de la manière suivante :

$$\begin{cases} \lambda_i = 1 \\ 0 \leq \lambda_j \leq \frac{\alpha_j^*}{\alpha_i^*}, \qquad (1 \leq j \neq i \leq m) \\ \sum_{j=1}^m y_j \lambda_j = 0 \end{cases},$$

et l'on observe qu'une solution optimale est obtenue en posant $\lambda_j^* = 0$ pour tout indice j différent de i tel que $y_j = y_i$ et en distribuant une masse de 1 sur un certain nombre de paramètres λ_j^* tels que $y_j = -y_i$. Dans ces conditions, pour toutes les valeurs de i à considérer, $J_{\mathrm{VC}}(\lambda^*) = (\lambda_i^*)^2 + \left(\sum_{j\ :y_j \neq y_i} \lambda_j^*\right)^2 = 2$. En conséquence, $K_{\mathrm{VC}} = 2 = Q$. Pour conclure, il suffit alors d'observer que lorsque $Q = 2$, la somme apparaissant dans le terme de droite de la borne 2.37 se réduit à $\frac{1}{\gamma^2}$.

Le terme de marge $\sum_{k<l} \frac{(1+d_{kl})^2}{\gamma_{kl}^2}$ est très simple à calculer par application de l'équation 2.10. A l'inverse, la présence dans la borne multiclasse de la constante K_{VC} rend son calcul nettement plus coûteux que celui de la borne biclasse. Cependant, un simple changement de variables permet de réduire considérablement le nombre de variables intervenant dans les (au plus) m problèmes de programmation quadratique à résoudre. Le vecteur α^* et un indice i étant donnés, considérons le vecteur $\Lambda = (\Lambda_{kl})$ de $\mathbb{R}^{Q^2}$ défini par : $\Lambda_{kl} = \sum_{n=1}^Q \alpha_{in}^* \sum_{x_j \in k} \lambda_{jl}$ (il résulte de $\lambda_{jy_j} = 0$ que l'on a pour tout k $\Lambda_{kk} = 0$). On définit pour ce vecteur la fonction objectif J'_{VC}, de manière que $J'_{\mathrm{VC}}(\Lambda) = \left(\sum_{n=1}^Q \alpha_{in}^*\right)^2 J_{\mathrm{VC}}(\lambda)$. Celle-ci a donc pour expression : $J'_{\mathrm{VC}}(\Lambda) = \sum_{k=1}^Q \left(\sum_{l=1}^Q \Lambda_{lk}\right)^2$. On démontre alors que la valeur de K_i s'obtient en résolvant le problème suivant :

Problème 16 ($\mathbf{PQ}'(\alpha^*)_i$)

$$\min_{\Lambda} J'_{VC}(\Lambda)$$

$$s.c. \begin{cases} \alpha_{il}^* \leq \Lambda_{y_i l} \leq \sum_{x_j \in y_i} \alpha_{jl}^*, & (1 \leq l \leq Q) \\ 0 \leq \Lambda_{kl} \leq \sum_{x_j \in k} \alpha_{jl}^*, & (1 \leq k \neq y_i \leq Q) \\ \sum_{l=1}^Q (\Lambda_{kl} - \Lambda_{lk}) = 0, & (1 \leq k \leq Q-1) \end{cases}.$$

Cette reformulation du problème est rendue possible par le fait que nous ne sommes pas intéressés par le vecteur λ^* lui-même, mais simplement par la valeur de $J_{\mathrm{VC}}(\lambda^*)$.

L'extension de la borne "rayon-marge" dédiée à la M-SVM de Lee et co-auteurs possède naturellement un intérêt particulier du fait de l'équivalence des problèmes 10 et 11, qui permet de l'appliquer en utilisant une marge douce, et donc pour choisir la valeur de C. Sa preuve suit également celle du résultat de base. Pour introduire la formule correspondante, il convient de revenir sur la notion de marges géométriques multiclasses donnée par la définition 7. On a en effet $d_{\mathrm{LLW}} = \frac{Q}{Q-1}$ et par conséquent :

$$\forall (k,l) : 1 \leq k < l \leq Q,\ \gamma_{kl} = \frac{Q}{Q-1} \frac{1+d_{kl}}{\|w_k - w_l\|}.$$

Par suite, dans le cas d'une machine à marge dure, on dispose de l'équation :

$$\frac{1}{Q-1} 1_{Qm}^T \alpha^* = \sum_{k=1}^Q \|w_k^*\|^2 = \frac{1}{Q} \sum_{k<l} \|w_k^* - w_l^*\|^2 = \frac{Q}{(Q-1)^2} \sum_{k<l} \frac{(1+d_{kl})^2}{\gamma_{kl}^2},$$

soit encore :

$$1_{Qm}^T \alpha^* = \frac{Q}{Q-1} \sum_{k<l} \frac{(1+d_{kl})^2}{\gamma_{kl}^2}. \tag{2.38}$$

Ces précisions étant apportées, la borne, correspondant au théorème 2 de [161], est la suivante :

Théorème 14 (Borne "rayon-marge" pour la M-SVM de Lee et co-auteurs) *Considérons une M-SVM de Lee et co-auteurs à Q catégories à marge dure. Alors, on obtient pour majorant de $\mathcal{L}_m$:*

$$\mathcal{L}_m \leq \mathcal{D}_m^2 \sum_{k<l} \frac{(1+d_{kl})^2}{\gamma_{kl}^2} \tag{2.39}$$

où $\mathcal{D}_m$ est le diamètre de la plus petite boule de l'espace de représentation contenant les vecteurs support.

Cette borne est nettement plus simple à calculer que celle dédiée à la M-SVM de Weston et Watkins (compte tenu de l'équation 2.38, le seul problème à résoudre est la détermination de $\mathcal{D}_m$). Naturellement, elle se réduit également à la borne "rayon-marge" biclasse lorsque $Q = 2$.

2.7.1.4 Borne de Passerini et co-auteurs

Dans [169], Passerini et ses co-auteurs proposent une borne sur l'erreur "leave-one-out" des méthodes de décomposition fondées sur l'emploi d'ECOC. Cette borne repose sur une notion de marge multiclasse très proche de celle donnée par la définition 3. En reprenant les notations de la section 2.3.3, la marge en question est définie par :

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y},\ \mathcal{M}_{\text{PPF}}(x,y) = \min_{k \neq y} d_{\ell_{\text{dec}}}(M_{k.}, \tilde{g}(x)) - d_{\ell_{\text{dec}}}(M_{y.}, \tilde{g}(x)).$$

Les classifieurs de base utilisés sont les machines à noyau $h^{(l)}$ d'expression analytique :

$$\forall l \in \{1, \ldots, N\},\ h^{(l)}(.) = \sum_{i=1}^{m} m_{y_i l} \beta_{il} \kappa^{(l)}(x_i, .). \tag{2.40}$$

Ces éléments étant donnés, la borne de Passerini et co-auteurs est la suivante :

Théorème 15 (Théorème 4.1 dans [169]) *Considérons le classifieur à Q catégories obtenu par combinaison au moyen d'ECOC des machines à noyau définies par l'équation 2.40. Sous l'hypothèse que $d_{\ell_{dec}}(M_{k.}, \tilde{g}(x)) = -\langle M_{k.}, \tilde{g}(x) \rangle$, son erreur "leave-one-out" est majorée de la manière suivante :*

$$\mathcal{L}_m \leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left\{ -\mathcal{M}_{PPF}(x_i, y_i) + \max_{k \neq y_i} U_k(x_i) > 0 \right\}$$

où $U_k(x_i) = \langle (M_{k.} - M_{p.}), \tilde{g}(x_i) \rangle + \sum_{l=1}^{N} m_{y_i l}(m_{y_i l} - m_{kl}) \alpha_{il} \kappa^{(l)}(x_i, x_i)$ avec $p = \text{argmax}_{q \neq y_i} \langle M_{q.}, \tilde{g}(x_i) \rangle$.

Le théorème 15 utilise pour fonction de perte $\ell_{\text{dec}}\left(m_{kl} h^{(l)}(x)\right) = -m_{kl} h^{(l)}(x)$. Les auteurs proposent également un corollaire (corollaire 4.1) correspondant au cas où cette fonction n'est plus caractérisée que par le fait qu'elle est décroissante.

2.7.2 Discussion

Dans la section précédente, différentes bornes sur l'erreur "leave-one-out" de méthodes de décomposition et de M-SVM ont été présentées. A notre connaissance, la seule autre estimation de l'erreur "leave-one-out" d'une M-SVM, en l'occurence celle de Lee et co-auteurs, est décrite par ses concepteurs dans [137]. Elle repose sur le principe de "generalized approximate cross-validation" (GACV). Naturellement, le premier critère pour juger de l'utilité de ces formules est leur faculté à produire une valeur de la constante de marge douce de bonne qualité. Cela suppose qu'elles soient différentiables par rapport à cet hyperparamètre (c'était le souci premier de Wang et ses co-auteurs), ou au moins qu'elles puissent être évaluées simplement (dans un temps raisonnable) pour un nombre significatif de ses valeurs. Dans ces conditions, les travaux d'Hastie et ses co-auteurs rapportés dans [104] représentent une avancée importante pour la sélection de modèle. Ils introduisent en effet un algorithme permettant de parcourir toutes les solutions de l'apprentissage d'une SVM lorsque C varie, pour un coût proche de celui d'un apprentissage unique. En pratique, on passe d'une solution à l'autre par résolution d'un système linéaire.

Un point important, compte tenu du cadre de notre étude, est que cet algorithme s'applique également aux M-SVM. Ces travaux sont prolongés dans [136], où les auteurs proposent un algorithme dédié à leur propre modèle de M-SVM.

Les auteurs de [131] décrivent une méthode pour optimiser la constante de marge douce d'une SVM "de norme 2". Elle est fondée sur la résolution d'un problème de programmation convexe. Cet algorithme, étendu à notre version "à coût quadratique" de la M-SVM de Lee et co-auteurs, devrait fournir l'étalon pour juger de l'utilité pratique du théorème 14. Rappelons enfin que les méthodes fondées sur un critère de vraisemblance pénalisée sont actuellement celles qui font référence en sélection de modèle [152]. Leur application aux M-SVM constitue donc à nos yeux une perspective de recherche prioritaire.

2.8 Conclusions et perspectives

Dans ce chapitre, nous avons décrit les méthodes disponibles pour effectuer des tâches de discrimination à catégories multiples avec des SVM. Elles s'appuient soit sur une machine unique, soit sur un ensemble de machines binaires, utilisées dans le cadre d'une méthode de décomposition. Derrière leur grande variété de principes se cache une réalité pratique globalement admise par le plus grand nombre : des performances le plus souvent très similaires (difficiles à distinguer statistiquement), à quelques exceptions près. Les raisons qui font que telle solution se montrera plutôt à son avantage sur tel jeu de données, en fournissant des performances médiocres sur tel autre, demeurent le plus souvent obscures. Sans doute la voix de la sagesse peut-elle être empruntée à Friedman lorqu'il écrit dans [83], évoquant de manière plus générale la discrimination multiclasse : "La leçon la plus importante à tirer de l'exercice ci-dessus est que les performances relatives des différentes approches peuvent fortement dépendre du problème particulier auquel elles sont appliquées. Comme tous les autres aspects de la méthodologie de l'apprentissage, aucune approche (raisonnable) ne domine toutes les autres dans toutes les situations (raisonnables)". Dans ces conditions, à l'heure actuelle, le meilleur argument en faveur des SVM multiclasses est leur capacité à traiter les problèmes dans lesquels $\mathcal{Y}$ est un ensemble d'éléments structurés. Naturellement, il s'agit d'un argument important pour qui travaille à l'exploitation de données biologiques.

Toutes ces méthodes devraient s'appuyer sur des bornes de convergence uniforme du risque empirique. Cependant, historiquement, la pratique a précédé la théorie, et les résultats de consistance, ou les formules de complexité en échantillon, demeurent encore largement à établir. Aucun des résultats déjà disponibles n'apparaît trivial, ce qui laisse augurer de la difficulté de la tâche restant à accomplir. On remarquera en particulier que l'étude des mesures de capacité des systèmes discriminants multiclasses à vaste marge soulève d'importants problèmes techniques. Dans ce cadre, la prise en compte des propriétés d'un noyau est également plus délicate. Ces problèmes sont à notre avis fondamentaux, dans la mesure où ils sont révélateurs de la nature particulière des problèmes de discrimination à catégories multiples. Un autre enseignement qui se dégage de notre étude est donc le fait que la théorie de la discrimination multiclasse ne peut se réduire à celle de la discrimination biclasse.

Tant qu'une borne sur le risque suffisamment précise pour rendre compte de manière fine des différences de comportement entre les méthodes de décomposition, les M-SVM et les autres SVM multiclasses, ne sera pas disponible, le praticien sera réduit, pour faire un choix, à mesurer des performances empiriques en termes de taux de reconnaissance et de temps de calcul. De ce point de vue, les M-SVM doivent posséder encore un potentiel de progression important. Ni la programmation de ces machines ni la dérivation de méthodes de sélection de modèle dédiées n'ont fait l'objet d'études aussi poussées que dans le cas des autres modèles, alors même que ces problèmes appellent, comme toujours, des solutions spécifiques, allant bien au-delà d'extensions directes de solutions biclasses.

En l'absence de contribution plus novatrice dans la littérature, nous nous sommes concentrés ici sur une étude des capacités de généralisation des M-SVM reposant sur des schémas très classiques : calcul d'une borne sur le risque faisant intervenir une mesure de capacité globale, puis calcul d'une borne sur cette mesure, soit directement, soit en impliquant une dimension de Vapnik-Chervonenkis étendue, par le biais d'un lemme de Sauer-Shelah généralisé. L'utilisation de la complexité de Rademacher que nous avons présentée ne constitue encore qu'un travail initial. Les résultats les plus récents en théorie des bornes (voir par exemple [39]), reposent sur des inégalités de concentration [40] puissantes, ou des propriétés des moyennes de Rademacher. Ils font intervenir des mesures de capacité empiriques et permettent de

substituer à l'analyse dans le pire des cas une analyse locale. L'état de l'art concernant les SVM biclasses, que nous avons évoqué au long de ce chapitre, est complété par deux références majeures, [211, 33]. L'extension de ces résultats au cas des M-SVM constitue actuellement, avec la poursuite de nos travaux en sélection de modèle, la première de nos priorités. Ces deux axes de recherche sont naturellement intimement liés.

Chapitre 3

Application de SVM multiclasses en prédiction de la structure secondaire des protéines

3.1 Prédiction de la structure secondaire

Dans cette section, nous présentons le problème de traitement de séquences biologiques sur lequel ont porté l'essentiel de nos travaux appliqués concernant la mise en œuvre des M-SVM.

3.1.1 Présentation du problème

Connaître la structure d'une protéine est un prérequis pour comprendre précisément sa fonction. Les projets de séquençage à grande échelle qui se sont multipliés ces dernières années ont permis d'obtenir les séquences d'un très grand nombre de gènes et par suite celles d'un très grand nombre de protéines. Le phénomène a été accéléré par l'apparition de nouvelles techniques de séquençage à la fois rapides et à bas coût. Malheureusement, le nombre de structures connues n'a pas suivi la même progression. En effet, les méthodes expérimentales disponibles pour déterminer la structure tridimensionnelle (tertiaire), la cristallographie par rayons X ou radiocristallographie et la spectroscopie par résonance magnétique nucléaire (RMN) demandent beaucoup d'efforts et n'assurent pas l'obtention du résultat recherché (on peut remarquer par exemple que certaines protéines ne cristallisent pas). De ce fait, prédire la structure tertiaire des protéines *ab initio*, c'est-à-dire à partir de la seule séquence (structure primaire), est devenu l'un des problèmes centraux de la biologie structurale. Au début des années 60, Anfinsen proposa son "hypothèse thermodynamique" [77], impliquant que la séquence protéique contient suffisamment d'information pour garantir un repliement correct à partir d'un vaste ensemble d'états dépliés. Cette hypothèse s'appuyait en particulier sur des expériences de "dénaturation-renaturation" [7]. Si le problème considéré peut donc théoriquement être résolu, les difficultés pratiques, mises en évidence par exemple dans [121], sont telles qu'il est rarement abordé de manière directe, mais plutôt au travers d'une approche du type diviser pour régner. Dans ce contexte, une étape intermédiaire utile est la prédiction de la structure secondaire, qui représente un moyen de simplifier le problème en projetant la structure 3D très complexe sur une dimension, c'est-à-dire sur une succession d'états conformationnels associés à chaque résidu (acide aminé) de la séquence. La structure secondaire d'une protéine est constituée par les motifs réguliers (périodiques) et répétés du repliement de son épine dorsale. Les deux éléments structuraux les plus communs sont l'*hélice alpha* et le *brin bêta*. La figure 3.1 propose une représentation schématique de la structure secondaire de la protéine G [64], représentation obtenue avec le logiciel RasMol [192]. Cette structure est composée de deux parties principales : une hélice alpha, en rouge sur la figure, et un *feuillet bêta* constitué de quatre brins, en jaune. Du point de vue de la reconnaissance des formes, la prédiction de la structure secondaire peut être vue comme un problème de discrimination à trois catégories, consistant à affecter à chaque résidu de la séquence son état conformationnel, en hélice α, brin β ou structure apériodique (coil).

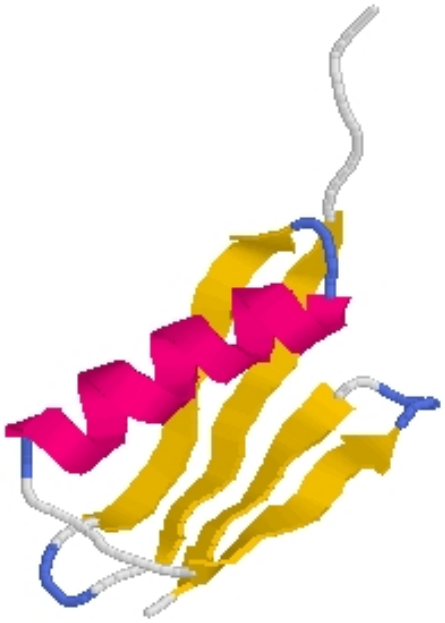

FIG. 3.1 – Représentation schématique des éléments structuraux de la protéine G.

3.1.2 Etat de l'art

Les premiers travaux en prédiction de la structure secondaire datent de la fin des années 60. Depuis lors, ce problème a fait l'objet de recherches intensives. Il est possible de classer en trois grandes familles les méthodes qui ont été mises en œuvre pour le traiter. Historiquement, les méthodes les plus anciennes sont celles fondées sur l'exploitation de propriétés physico-chimiques [144, 143] et celles dites "statistiques" [52, 90], reposant principalement sur l'estimation des probabilités conditionnelles des états conformationnels à partir de statistiques d'ordres un ou deux calculées sur de petits peptides. Elles ont progressé lentement au cours des années 80-90 [86, 205, 89], jusqu'au moment où elles ont été pratiquement supplantées par des méthodes issues de l'apprentissage numérique. Les progrès les plus spectaculaires ont résulté de l'introduction dans le domaine de PMC. Dans un premier temps, une transposition directe du système NETtalk a permis à Qian et Sejnowski de faire passer le taux de reconnaissance d'environ 60% à plus de 64% [178]. Leur architecture doit être évoquée, car elle a été pratiquement systématiquement reprise dans les travaux ultérieurs. Deux PMC sont utilisés en cascade. Le premier, dit "séquence-structure", effectue la prédiction de la structure à partir de la séquence, le second, dit "structure-structure", lissant la prédiction initiale, en prenant en compte le fait que les états conformationnels des résidus consécutifs sont fortement corrélés. Le remplacement, en entrée du PMC "séquence-structure", de la simple structure primaire par un profil d'alignement, a eu pour conséquence un gain supplémentaire d'environ 6%, permettant d'atteindre pour la première fois la frontière des 70% de taux de reconnaissance. Ainsi, la méthode PHD [184], à travers ses développements successifs [185], a constitué l'état de l'art pendant la plus grande partie des années 90. Parallèlement, les systèmes mettant en œuvre des modèles de Markov cachés (HMM), introduits par Asai et ses collaborateurs [14], ont tiré profit, comme les autres, de l'accroissement de la taille de la "protein data bank" (PDB) [31, 30] pour prendre en compte de manière de plus en plus fine les règles syntaxiques régissant l'agencement des éléments structuraux. En la matière, la contribution la plus aboutie est probablement constituée par les travaux de Juliette Martin [151, 150]. Actuellement, les meilleurs systèmes de prédiction sont des modèles connexionnistes extrêmement complexes faisant intervenir de grands nombres de réseaux, à propagation avant ou récurrents [119, 18, 170, 177]. Leurs performances doivent plus à la qualité des alignements qu'ils exploitent qu'à la nature du modèle neuronal lui-même. De manière étonnante, celui-ci est souvent sur-paramétré, sans que l'on observe pour autant de

phénomène de sur-apprentissage [181, 91]. Les taux de reconnaissance les plus élevés rapportés saturent aux environs des 80% de résidus bien classés. De ce fait, le dernier grand article de synthèse que Rost a écrit sur la prédiction de la structure secondaire, [183], demeure aujourd'hui encore d'actualité.

La communauté de la biologie structurale prédictive est en attente d'un progrès du même ordre que celui ayant résulté de l'emploi de PMC ou d'alignements multiples. Un autre phénomène vient expliquer le ralentissement des progrès effectués. Il existe trois grandes familles de méthodes pour réaliser la prédiction de la structure tertiaire. La première est celle des méthodes de modélisation par homologie [53], la seconde étant celle des méthodes de "threading" [149]. La troisième, déjà évoquée, est celle des méthodes *ab initio*. Ce sont principalement ces dernières méthodes qui utilisent la prédiction de la structure secondaire. Or, on ne fait appel à la prédiction *ab initio* que lorsque la modélisation par homologie et le threading ont échoué. La conjugaison de deux facteurs positifs : les progrès méthodologiques et l'accroissement continuel des tailles des bases de référence, diminue sensiblement la fréquence de cette situation. De plus, la méthode de prédiction *ab initio* actuellement la plus performante, ROSETTA [203], n'est pas fondée sur la prédiction de la structure secondaire. Au contraire, elle a été utilisée dans [158] afin d'améliorer cette prédiction. La tâche qui nous intéresse a donc perdu, au cours des années, un peu de son importance. Elle pourrait revenir au devant de la scène si une avancée significative avait de nouveau lieu. Naturellement, il est plaisant d'imaginer qu'une telle avancée puisse résulter de l'emploi de méthodes à noyau, et plus particulièrement de M-SVM.

A notre connaissance, la première application d'une SVM, en l'occurence multiclasse, en prédiction de la structure secondaire, était une combinaison de modèles [92] (voir aussi [93]). Ces travaux, qui ont produit un gain de performance statistiquement significatif par rapport aux méthodes d'ensemble habituellement utilisées dans ce contexte, ont été poursuivis dans [99], avec le même succès. Parallèlement, les auteurs de [114] ont été les premiers à aborder directement le problème. Pour ce faire, ils ont employé des SVM biclasses à noyau gaussien sphérique, combinées par la méthode de décomposition un contre tous ainsi que d'autres méthodes élémentaires fondées sur un graphe de décision. Leur conclusion est que leur approche permet d'obtenir une très bonne valeur du coefficient Sov [186, 246] global : 76.2%. Des travaux similaires sont rapportés dans [237]. La principale différence par rapport aux précédents réside dans l'utilisation d'un codage physico-chimique des acides aminés. Le noyau, à l'inverse, demeure le même. Le taux de reconnaissance annoncé, 78.4%, est comparable à celui des meilleurs systèmes connexionnistes. Suivant l'usage commun, adopté par exemple par le challenge "critical assessment of techniques for protein structure prediction" (CASP), nous avons considéré jusqu'à présent que la prédiction de la structure secondaire était un problème de discrimination à trois catégories. En fait, les programmes d'assignation déterminant la structure secondaire à partir de la structure 3D considèrent plus d'états conformationnels. Ainsi, le programme DSSP [120], le plus utilisé en pratique, en considère huit, parmi lesquels celui correspondant aux coudes β, qu'une classification en seulement trois catégories associe à la structure apériodique. Ces coudes β sont prédits au moyen d'une SVM dans [247]. Les prédicteurs utilisés sont ceux correspondant au profil d'alignement obtenu par application de PSI-BLAST [5, 119] plus la structure secondaire prédite par PSIPRED [119]. Une fois de plus, les performances sont encourageantes, alors même que le noyau utilisé repose sur une simple gaussienne sphérique. On trouve plusieurs autres applications des SVM en prédiction de la structure secondaire utilisant des SVM biclasses et le noyau gaussien standard (voir par exemple [238, 101]). A l'inverse, à notre connaissance, une seule autre équipe a appliqué une M-SVM à ce problème. Dans [166, 167], les auteurs reprennent l'architecture introduite par Qian et Sejnowski, consistant à mettre en œuvre en cascade deux systèmes discriminants, l'un "séquence-structure" et l'autre "structure-structure". Les PMC de la méthode originale sont remplacés par des M-SVM de Crammer et Singer (voir la section 2.4.1.2). Ici encore, le noyau utilisé est le noyau gaussien standard. Nous présentons à présent le premier noyau dédié à la prédiction de la structure secondaire.

3.2 Noyau dédié à la prédiction de la structure secondaire

Nous avons vu dans la section précédente que la prédiction de la structure secondaire des protéines est un domaine qui est apparu il y a près de quarante ans et a fait depuis lors l'objet de recherches intensives. De nos jours, il n'est plus possible d'espérer dépasser l'état de l'art si ce n'est en mettant en œuvre un système discriminant spécialement conçu pour cette tâche, système intégrant autant que

possible l'importante expertise accumulée au cours des années aussi bien par le biologiste que par le bioinformaticien. Dans le cas où ce système est une méthode à noyau, cela passe naturellement par la spécification d'un nouveau noyau. Celui qui est présenté dans cette section repose sur des considérations biologiques très simples. Il s'agit d'un noyau gaussien entièrement caractérisé par le choix de la norme sur l'espace des exemples (espace des contenus de fenêtres d'analyse). Cette norme est paramétrée, si bien que le travail d'adaptation du noyau aux données relève directement de la sélection de modèle pour les M-SVM. Il vient donc compléter l'exposé effectué dans la section 2.7.

3.2.1 Choix des prédicteurs

La manière usuelle d'effectuer la prédiction de la structure secondaire au moyen de modèles de l'apprentissage statistique consiste à employer une approche locale. Plus précisément, les prédicteurs utilisés pour déterminer l'état conformationnel d'un résidu donné sont les acides aminés contenus dans une fenêtre d'analyse de taille fixe centrée sur ce résidu (dans certains cas, la fenêtre est asymétrique). Afin de coder le contenu de chaque position de la fenêtre, un vecteur de 22 composantes est utilisé. Chacune de ses 20 premières composantes est associée à un acide aminé donné (il y en a 20 différents), les deux composantes restantes étant utilisées respectivement pour représenter les acides aminés indéterminés, habituellement représentés par un 'X' dans les bases, et les positions vides de la fenêtre. Une fenêtre d'analyse peut contenir des positions vides lorsqu'elle déborde à l'une des extrémités (N ou C terminale) de la chaîne protéique traitée. En résumé, le codage utilisé pour représenter le contenu de la fenêtre est le codage orthonormal standard, qui n'induit aucune corrélation entre les symboles de l'alphabet. Ce qui apparaît a priori comme un avantage est ici un inconvénient, ainsi que nous le verrons dans la section suivante. Pour une taille de la fenêtre d'analyse $|W| = 2n + 1$, où la valeur de n est habituellement comprise entre 5 et 10, le nombre de prédicteurs est donc égal à $(2n + 1)22$. Seuls $(2n + 1)$ d'entre eux sont égaux à 1, les autres étant égaux à 0. Ceci produit un vecteur de grande taille très creux. La situation est différente lorsqu'un profil d'alignement multiple est utilisé à la place de la structure primaire. Une attention particulière doit être apportée à l'inclusion sous cette forme d'informations évolutives, dans la mesure où elle améliore significativement les performances des méthodes de prédiction comme nous avons eu l'occasion de l'observer dans la section 3.1.2. Dans un but de simplification de l'exposé, les détails concernant cette alternative sont reportés à la section 3.2.5.

3.2.2 Insuffisances des noyaux classiques

Considérons le vecteur $\mathbf{x}$ utilisé pour prédire l'état conformationnel d'un résidu donné. Alors, compte tenu du choix des prédicteurs décrit ci-dessus, $\mathbf{x} = (x_i)_{-n \leq i \leq n} \in \{0,1\}^{(2n+1)22}$, où x_i est le codage canonique de l'acide aminé occupant la position d'indice i dans la fenêtre d'analyse. En conséquence, la fonction calculée par un noyau gaussien standard appliqué à deux contenus de fenêtres $\mathbf{x}$ et $\mathbf{x}'$ peut être réécrite de la manière suivante :

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) = \exp\left(-\frac{(2n+1) - \sum_{i=-n}^{n} \delta_{x_i, x'_i}}{\sigma^2}\right) \tag{3.1}$$

où δ est de nouveau le symbole de Kronecker. Le terme de droite de (3.1) souligne le fait que le noyau dépend uniquement de la distance de Hamming entre les deux segments considérés. Cela constitue un bien pauvre résumé de l'information contenue dans les données. En effet, deux segments correspondant à des parties de protéines homologues superposables dans l'espace, et partageant donc la même structure secondaire, peuvent différer significativement en raison des phénomènes de l'évolution que sont les insertions/délétions ainsi que les substitutions. La distance de Hamming est très sensible aux premiers d'entre eux, tandis qu'elle ne prend pas en compte la nature des substitutions, mais simplement leur nombre. Elle ne prend pas non plus en compte l'influence relative des éléments du contexte en fonction de leur distance au centre de la fenêtre. En conséquence, on ne peut espérer qu'une telle combinaison noyau/codage des données puisse fournir des résultats satisfaisants sur le problème qui nous intéresse. Le constat serait similaire avec les autres noyaux standard. Cette simple observation est à l'origine du travail de développement de noyau décrit dans les sections suivantes, travail qui repose en particulier sur une extension au cas multiclasse de la notion d'alignement noyau-cible.

3.2.3 Alignement de noyaux

L'alignement de noyaux a été introduit dans [59], comme un moyen d'évaluer le degré d'adéquation d'un noyau à une tâche d'apprentissage donnée, et adapter en conséquence la matrice de Gram afin d'augmenter cette adéquation. Il s'agit donc fondamentalement d'une méthode conçue pour la transduction (voir en particulier le chapitre 8 de [229] ou [131]), dans la mesure où elle ne fournit pas d'expression analytique pour le noyau résultant.

Définition 26 (Alignement de noyaux [59]) *Soient κ et κ' deux fonctions noyau mesurables définies sur $\mathcal{T} \times \mathcal{T}$ où $\mathcal{T}$ est supposé être un espace probabilisé muni d'une mesure de probabilité $P_{\mathcal{T}}$. L'*alignement entre κ et κ', *$A(\kappa, \kappa')$, est défini comme suit :*

$$A(\kappa, \kappa') = \frac{\langle \kappa, \kappa' \rangle}{\|\kappa\| \|\kappa'\|} = \frac{\int_{\mathcal{T}^2} \kappa(t,t')\kappa'(t,t') dP_{\mathcal{T}}(t) dP_{\mathcal{T}}(t')}{\sqrt{\int_{\mathcal{T}^2} \kappa(t,t')^2 dP_{\mathcal{T}}(t) dP_{\mathcal{T}}(t')} \sqrt{\int_{\mathcal{T}^2} \kappa'(t,t')^2 dP_{\mathcal{T}}(t) dP_{\mathcal{T}}(t')}}.$$

Définition 27 (Alignement empirique de noyaux [59]) *$\mathcal{T}$, κ et κ' étant définis comme dans la définition 26, soit $T^n = (T_i)_{1 \leq i \leq n}$ un n-échantillon de variables aléatoires indépendantes distribuées suivant $P_{\mathcal{T}}$. L'alignement de κ et κ' par rapport à T^n est la quantité :*

$$\hat{A}_{T^n}(G, G') = \frac{\langle G, G' \rangle_F}{\|G\|_F \|G'\|_F} \tag{3.2}$$

où G et G' sont les matrices de Gram associées respectivement à κ et κ', calculées sur T^n, et $\langle ., . \rangle_F$ représente le produit scalaire de Frobenius entre matrices, si bien que $\langle G, G' \rangle_F = \sum_{i=1}^n \sum_{j=1}^n \kappa(T_i, T_j) \kappa'(T_i, T_j)$. $\|.\|_F$ représente la norme correspondante.

L'alignement de noyaux est donc un cosinus, et comme tel fournit une mesure de la similarité de ses arguments. Si κ est un noyau bien adapté au problème considéré et que κ' est bien aligné avec κ, alors κ' est également un bon noyau pour le même problème. En pratique, l'alignement n'étant pas calculable, puisque la distribution sous-jacente $P_{\mathcal{T}}$ est inconnue, il est estimé empiriquement au moyen de la formule 3.2. Cristianini et ses co-auteurs ont étudié les propriétés de concentration de la variable aléatoire $\hat{A}_{T^n}(G, G')$ autour de son espérance $A(\kappa, \kappa')$.

3.2.4 Alignement noyau-cible multiclasse : application au paramétrage d'un noyau

Considérons une famille de noyaux κ_θ de paramètre formel θ appartenant à l'ensemble Θ. L'alignement de noyaux étant défini, notre stratégie pour l'appliquer à la détermination partielle ou entière de θ peut être résumée de la manière suivante :

1. Choisir un noyau théoriquement idéal κ_t, que nous appellerons dans la suite le *noyau cible*, idéal au sens où il conduit à un classement parfait. En pratique, la matrice de Gram de κ_t doit pouvoir être calculée.
2. Etant donné un ensemble d'apprentissage $z^m = ((\mathbf{x}_i, y_i))_{1 \leq i \leq m}$, choisir θ^* en application du critère suivant :
$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \hat{A}_{z^m}(G_\theta, G_t),$$
où G_θ est la matrice de Gram associée au couple (κ_θ, z^m), G_t étant la matrice de Gram associée au couple (κ_t, z^m).

La conjecture qui est faite est que le noyau κ_{θ^*} ainsi choisi se comportera bien sur le problème considéré, pourvu que la famille $\{\kappa_\theta \ : \theta \in \Theta\}$ soit appropriée. Dans le cas biclasse (pour lequel $\mathcal{Y} = \{-1, 1\}$), le noyau idéal se définit de la manière suivante : $\kappa_t(\mathbf{x}, \mathbf{x}') = yy'$. L'extension multiclasse proposée par Régis Vert reprend le raisonnement géométrique sous-jacent, en utilisant pour représentants des catégories les sommets du simplexe décrit dans la section 2.4.1.3. On obtient ainsi dans le cas de Q catégories :

$$\begin{cases} \text{si } y = y', & \kappa_t(\mathbf{x}, \mathbf{x}') = 1 \\ \text{si } y \neq y', & \kappa_t(\mathbf{x}, \mathbf{x}') = -\frac{1}{Q-1} \end{cases}.$$

Notons que sous certaines hypothèses de régularité sur κ_θ, $\hat{A}_{z^m}(G_\theta, G_t)$ est différentiable par rapport à θ, et peut donc être optimisé en utilisant des techniques classiques, telles que les descentes en gradient évoquées dans la section 2.6.

3.2.5 Prise en compte d'informations évolutives dans un noyau de convolution

Dans ce qui suit, nous adoptons la terminologie de [244], qui nomme noyau de convolution tout noyau vérifiant $\kappa(t, t') = \kappa(t - t', 0)$. L'objectif est de prendre en compte pour ces noyaux deux des facteurs reconnus comme importants pour prédire la structure secondaire : la nature des substitutions entre deux segments et l'influence relative des acides aminés impliqués en fonction de leur position dans la fenêtre d'analyse.

3.2.5.1 Produits scalaires entre acides aminés

Dans la section 3.2.1, nous avons souligné le fait que le traitement standard des données de séquences en prédiction de la structure secondaire utilise un codage orthonormal des acides aminés. Cependant, il est connu que cette solution n'est pas satisfaisante. De fait, les biologistes ont produit un grand nombre de matrices de similarité (on parle également de matrices de substitution) pour les acides aminés, qui diffèrent toutes significativement de la matrice identité. C'est en particulier le cas des matrices "percent accepted mutations" (PAM) [62] et "blocks substitution matrix" (BLOSUM) [109]. Le problème soulevé par leur utilisation dans un noyau découle du fait qu'elles ne sont pas symétriques définies positives, et ne sont donc pas associées à un produit scalaire. Différentes solutions peuvent être envisagées pour surmonter cette difficulté. Les matrices étant symétriques, un moyen simple de les approximer par une matrice de Gram consiste à les diagonaliser et à annuler toutes les valeurs propres négatives. Une autre possibilité consiste à rechercher leur projection sur l'espace des matrices symétriques définies positives, l'opérateur correspondant étant associé à une norme matricielle, par exemple la norme de Frobenius déjà évoquée dans ce chapitre. Même s'il se peut que l'on ne dispose pas de l'expression analytique de l'opérateur de projection (le problème à résoudre n'est pas convexe), des estimations satisfaisantes peuvent être obtenues pas une simple descente en gradient [66]. Cette descente est alors réalisée par rapport aux composantes de la suite $(a_j)_{1 \leq j \leq 22}$ des vecteurs de $\mathbb{R}^{22}$ représentant les différents acides aminés, un contenu indéterminé ou une position vide.

Ce changement de produit scalaire entre acides aminés, qu'il soit issu ou non d'un changement explicite de leur codage, s'étend directement au cas où l'on utilise des alignements multiples. Dans ce cas, le profil présenté en entrée d'un classifieur neuronal (voir par exemple [184, 119, 177]) est simplement obtenu en calculant, pour chaque position de la fenêtre, une moyenne pondérée des vecteurs codant les acides aminés présents à la position correspondante de l'alignement. Le poids associé à un acide aminé particulier est sa fréquence d'occurence dans la position. Notons θ_{ij} la fréquence d'apparition de l'acide aminé d'indice j (de codage a_j) dans la position de l'alignement correspondant à la position d'indice i de la fenêtre glissante. Alors la fenêtre peut être représentée par le vecteur $\tilde{\mathbf{x}} = (\tilde{x}_i)_{-n \leq i \leq n}$ tel que $\tilde{x}_i = \sum_{j=1}^{22} \theta_{ij} a_j$. En conséquence, dans le calcul du noyau, le produit scalaire $\langle x_i, x_i' \rangle$ est simplement remplacé par :

$$\langle \tilde{x}_i, \tilde{x}_i' \rangle = \langle \sum_{j=1}^{22} \theta_{ij} a_j, \sum_{k=1}^{22} \theta_{ik}' a_k \rangle = \sum_{j=1}^{22} \sum_{k=1}^{22} \theta_{ij} \theta_{ik}' \langle a_j, a_k \rangle.$$

On tire ici profit de la linéarité du produit scalaire. Naturellement, comme dans le cas du "kernel trick", l'écriture $\langle a_j, a_k \rangle$ n'implique pas que l'on dispose de l'expression explicite de $(a_j)_{1 \leq j \leq 22}$.

3.2.5.2 Influence de la position dans la fenêtre

Nous avons vu que l'utilisation d'une fenêtre glissante était la norme en prédiction de la structure secondaire des protéines. De nombreuses études ont porté sur le choix de sa taille ou l'exploitation de son contenu. De bonnes illustrations sont fournies par [178, 249, 184]. En bref, une fenêtre trop petite ne contiendra pas assez d'information sur la conformation locale, tandis qu'une fenêtre trop grande

incorporera des données qui risquent de se comporter comme du bruit. Un moyen de surmonter cette difficulté consiste à choisir a priori une valeur élevée pour la taille de la fenêtre et à associer à chaque position un poids (indépendant de la nature de l'acide aminé), de manière à moduler son influence sur les calculs ultérieurs. La procédure peut être appliquée soit dans le cadre d'une prédiction directe des trois catégories, soit dans le cadre d'une décomposition un contre tous. Dans le second cas, il est instructif de comparer les pondérations obtenues pour les différents états conformationnels. Ceci avait déjà été réalisé avec succès par différentes équipes [88, 91]. Un point important est que ces études, bien qu'elles aient été fondées sur des approches très différentes, ont produit des distributions des poids très similaires. Ceci suggère qu'elles ont permis de mettre en évidence une propriété intrinsèque du problème considéré. Nous avons donc décidé d'incorporer une telle pondération dans notre noyau, les valeurs des poids étant obtenues au moyen de l'alignement noyau-cible multiclasse.

Les deux paramétrisations, la modification du "produit scalaire entre acides aminés" et la pondération des positions de la fenêtres d'analyse, peuvent être appliquées à tout type de noyau de convolution. Dans un but de simplification, leur utilisation est illustrée ici dans le cas d'un noyau gaussien utilisant des alignements multiples :

$$k_{\theta,D}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}') = \exp\left(-\frac{\sum_{i=-n}^{n}\theta_i^2\left(\|\tilde{x}_i\|^2+\|\tilde{x}_i'\|^2-2\langle\tilde{x}_i,\tilde{x}_i'\rangle\right)}{2\sigma^2}\right) \tag{3.3}$$

où $\theta = (\theta_i)_{-n\leq i\leq n}$ est le vecteur des poids et $D \in \mathcal{M}_{22,22}(\mathbb{R})$ la matrice des produits scalaires.

3.3 Evaluation des performances d'une M-SVM utilisant notre noyau

Nous avons souligné dans l'état de l'art que les meilleures méthodes de prédiction sont des systèmes très complexes, souvent constitués de centaines de modules. Les expériences relatées ci-dessous ont pour seule ambition d'établir qu'une M-SVM munie du noyau décrit dans la section précédente est significativement plus performante pour cette tâche que l'unité de base de nombre de méthodes de prédiction, le PMC.

3.3.1 Protocole expérimental

Pour évaluer notre classifieur, nous avons utilisé l'ensemble de 1096 protéines introduit dans [99] sous le nom P1096. Cet ensemble a été constitué de manière à satisfaire les exigences les plus fortes en termes de taux d'identité (voir [189] pour les détails). L'assignation de la structure secondaire a été effectuée au moyen du programme DSSP. La procédure utilisée pour réaliser le regroupement des huit états conformationnels initiaux dans les trois catégories de base est celle de CASP : H+G → H (hélice α), E+B → E (brin β), tous les autres états étant associés à la catégorie C (structure apériodique). Cette façon de procéder est connue comme étant celle conduisant au problème de prédiction le plus difficile [60]. Enfin, les alignements multiples, construits à partir des séquences identifiées par PSI-BLAST, ont été produits suivant le protocole décrit dans [177].

En présentant le problème de la prédiction de la structure secondaire, nous avons expliqué que le but premier du biologiste est de connaître la structure tridimensionnelle des protéines, afin de tenter d'inférer leur fonction. De ce fait, celui-ci souhaite principalement être en mesure d'identifier l'ensemble des éléments structuraux, avec leur ordre d'apparition sur la séquence. Un petit décalage entre les positions réelles et prédites des structures peut être toléré, mais la prédiction doit demeurer biologiquement plausible : aucune hélice ne doit être constituée de moins de quatre résidus (un tour d'hélice correspond à 3,6 résidus), deux structures périodiques ne peuvent pas être consécutives... Par suite, le seul taux de reconnaissance par résidu, noté Q_3 dans la littérature, n'est pas suffisant pour caractériser la qualité de la prédiction. De nombreuses mesures de qualité ont été introduites afin de rendre l'évaluation plus fine. Le lecteur intéressé trouvera dans [17] un exposé de synthèse sur la question. Dans ce qui suit, nous utilisons trois des mesures de qualité les plus usuelles : Q_3, les coefficients de correlation de Pearson/Matthews [154] et les coefficients Sov déjà évoqués plus haut. Ils fournissent des informations complémentaires.

Tandis que les coefficients de Matthews caractérisent la qualité de la prédiction pour un état conformationnel particulier (α/β/coil), ce qui permet, par exemple, de souligner une mauvaise reconnaissance des feuillets, la valeur des coefficients Sov donne une idée de la qualité de la prédiction au niveau des segments, satisfaisant ainsi l'une des principales exigences évoquées au début du paragraphe.

3.3.2 Estimation des paramètres du noyau

La matrice des produits scalaires entre acides aminés a été obtenue à partir de la matrice de similarité introduite dans [141]. La raison de ce choix tient au fait que cette matrice a été spécifiquement conçue pour effectuer la prédiction de la structure secondaire en s'appuyant sur la similarité de petits peptides, c'est-à-dire l'homologie de séquence locale (voir aussi [140, 89]). Dans ce contexte, elle s'est avérée supérieure à la matrice de Dayhoff. Parmi les deux options considérées à la section 3.2.5.1 pour engendrer la matrice de Gram, nous avons retenu celle utilisant la diagonalisation. Cependant, ce choix a en premier lieu été effectué dans un souci de reproductibilité des expériences, dans la mesure où la descente en gradient fournit des résultats très similaires [66]. Pour cet ensemble de produits scalaires, la valeur du vecteur de poids θ a été obtenue par application du principe d'alignement noyau-cible multiclasse, au moyen d'une descente en gradient (voir [232] pour le détail des calculs). La base d'apprentissage utilisée pour l'estimation de l'ensemble de ces paramètres était constituée des 1180 séquences employées pour entraîner SSpro1 et SSpro2. Cet ensemble, décrit dans [18, 177], est nommé dans la suite P1180. Ce choix était possible dans la mesure où aucune séquence de cette base ne présente avec une séquence de la base P1096 un taux d'identité supérieur à 25% (voir aussi [99]). En fait, la base P1096 a été assemblée en prenant en compte cette contrainte. La figure 3.2 illustre les valeurs obtenues pour les coefficients θ_i. Cette courbe

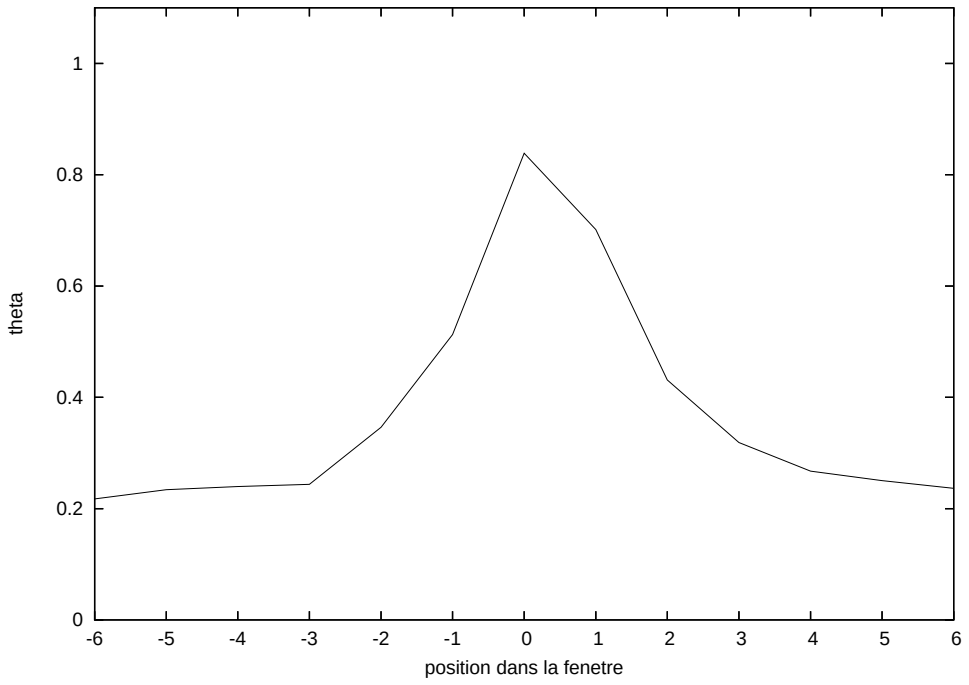

FIG. 3.2 – Vecteur θ du noyau défini par l'équation 3.3 optimisé par alignement noyau-cible multiclasse.

est très similaire à celles évoquées dans la section 3.2.5.2. L'une de leurs caractéristiques communes est

	séquence		alignement - profil		alignement - sortie	
	PMC	M-SVM	PMC	M-SVM	PMC	M-SVM
Q_3	61.6	62.0	72.0	72.3	68.9	69.6
C_α	0.46	0.47	0.63	0.64	0.55	0.59
C_β	0.33	0.35	0.53	0.54	0.42	0.48
C_c	0.38	0.38	0.53	0.54	0.47	0.46
Sov	53.9	54.2	65.1	65.3	64.0	64.8
Sov_α	57.8	57.9	66.5	66.7	64.4	65.0
Sov_β	44.7	46.1	61.5	62.3	58.4	61.6
Sov_c	57.3	57.3	66.7	66.8	64.2	66.1

TAB. 3.1 – Performances relatives d'un PMC et d'une M-SVM de Weston et Watkins mesurées sur la base P1096. La colonne "alignement - profil" correspond à l'utilisation d'un profil d'alignement, la colonne "alignement - sortie" correspondant à la combinaison, en sortie du classifieur, des prédictions associées à chacune des séquences d'un alignement.

leur asymétrie marquée en faveur du contexte droit (du côté de l'extrémité C terminale de la séquence). Ce phénomène n'a pas trouvé à ce jour d'explication biologique.

3.3.3 Expériences effectuées et résultats obtenus

Cette section décrit le protocole et les résultats de trois expériences inspirées respectivement par les travaux de référence de Qian et Sejnowski, Rost et Sander et Riis et Krogh (voir l'état de l'art). La M-SVM utilisée est celle de Weston et Watkins (voir la section 2.4.1.1). La première expérience compare la M-SVM à un PMC dans le cas où le vecteur de prédicteurs $\mathbf{x}$ correspond simplement au contenu d'une fenêtre d'analyse de taille 13 glissant sur la structure primaire. La seconde remplace la structure primaire par le profil d'un alignement multiple produit à partir du résultat de PSI-BLAST. Enfin, la troisième consiste à prédire séparément la structure des séquences d'un alignement multiple, et à combiner ensuite les prédictions obtenues au moyen d'une moyenne pondérée [226]. La pondération est celle proposée dans [110]. Dans les trois cas, nous avons utilisé la même procédure expérimentale, une validation croisée à cinq pas. Le PMC possédait une couche cachée de huit unités munies d'une sigmoïde, la fonction d'activation des unités de sortie étant la fonction softmax. La paramétrisation de la M-SVM est demeurée inchangée d'une expérience à l'autre, la constante de marge douce C étant fixée à 10.0 et la largeur de bande du noyau gaussien prenant la valeur $\sigma^2 = 10.0$. Les résultats obtenus sont résumés dans le tableau 3.1.

Dans chacune des configurations, l'accroissement du taux de reconnaissance résultant de l'utilisation d'une M-SVM est statistiquement significatif avec une confiance dépassant 0.95. La taille de la base (les 1096 protéines sont constituées de 255551 résidus) compense la faiblesse apparente de cet accroissement. Cependant, l'observation la plus encourageante est que toutes les mesures de qualité bénéficient de ce changement. Ceci est particulièrement net pour les brins β, qui sont habituellement les plus difficiles à prédire, puisqu'ils peuvent être impliqués dans des éléments structuraux non locaux : les feuillets β (voir la figure 3.1). Dans [114], les auteurs ont remarqué que les meilleurs indicateurs de la supériorité des SVM sur les PMC en prédiction de la structure secondaire étaient les coefficients Sov. Même si nous n'avons pas été en mesure de reproduire leurs expériences en obtenant des résultats similaires (nous avons observé des coefficients Sov plus faibles), la même conclusion peut être tirée ici.

Le nombre de variables duales de la M-SVM de Weston et Watkins étant égal à $(Q-1)m$, une idée de la complexité de la fonction calculée est fournie par le nombre d'exemples d'apprentissage pour lesquels au moins une variable duale appartient à l'intervalle ouvert $]0, C[$ (exemples situés sur l'une des C_Q^2 marges). Dans nos expériences (quinze apprentissages de la M-SVM), le pourcentage des points de ce type varie entre 25% et 30%.

3.4 Discussion et perspectives de recherche

Dans ce chapitre, nous avons donné une illustration du potentiel que peut représenter pour le domaine du traitement de séquences biologiques l'utilisation de M-SVM munies de noyaux dédiés. Cette étude initiale a déjà été complétée par une seconde, mettant en œuvre la même architecture sur un autre problème de biologie structurale prédictive : la prédiction des peptides d'ancrage à l'interface membranaire [190]. Il convient de souligner que pour cette dernière étude, les meilleures performances ont été obtenues avec une matrice de substitution différente, ce qui souligne l'importance du choix de cette matrice.

Naturellement, de nombreuses améliorations peuvent être apportées aux modèles décrits dans la section 3.3.3. La plus simple consiste à reprendre l'architecture en cascade de Qian et Sejnowski, en post-traitant les sorties de la M-SVM au moyen d'un module de prédiction "structure-structure". L'incorporation de prédicteurs supplémentaires fournissant des informations de nature physico-chimique constitue un autre moyen sûr d'obtenir une amélioration significative des peformances. Nos travaux en combinaison de modèles [96, 99] nous incitent également à combiner des M-SVM paramétrées de manières différentes, ou à combiner des M-SVM avec d'autres modèles, par exemples les réseaux de neurones récurrents (BRNN) de la méthode de prédiction SSpro2. Enfin, la question du choix d'un autre noyau est évoquée au chapitre 4.

Reprenant l'idée déjà développée dans [93], nous travaillons actuellement au post-traitement des prédictions par un module de programmation dynamique inspiré du modèle de HMM inhomogène (IHMM) introduit dans [179]. Ceci nécessite de calculer des estimations des probabilités a posteriori des classes à partir des scores produits par la M-SVM. Pour ce faire, étendant au cas multiclasse l'idée de Platt [174] décrite dans la section 2.3.2, nous utilisons des exponentielles normalisées. L'objectif de cette étude est d'obtenir des éléments structuraux prédits dont les longueurs soient plus proches des longueurs des éléments structuraux réels. A l'évidence, les mesures de qualité qui nous permettront d'évaluer le succès de ce modèle hybride sont en premier lieu les coefficients Sov.

Chapitre 4

Programme scientifique

Les deux chapitres précédents ont fourni un exposé synthétique de nos travaux sur les SVM multiclasses. Ceux-ci ont principalement porté sur la conception, la mise en œuvre et l'étude des performances en généralisation de ces machines. Ils ont trouvé une application privilégiée : la prédiction de la structure secondaire des protéines. Dans les deux cas, nous avons dégagé des perspectives précises, à mettre en œuvre à court ou moyen terme. Le présent chapitre a pour objet nos autres perspectives de recherche, souvent à plus long terme, qui s'inscrivent dans le cadre plus large du projet scientifique de l'équipe ABC. Ceci appelle naturellement une structure constituée de deux sections principales, l'une dédiée aux contributions de portée générale en apprentissage, l'autre aux contributions relevant de la biologie computationnelle.

4.1 Apprentissage automatique

4.1.1 Bornes sur les performances en généralisation des systèmes discriminants multiclasses

La section 2.5.1.3 a exposé le cheminement permettant, lorsque le modèle considéré est une M-SVM, de réduire le calcul d'une borne sur le nombre de couverture apparaissant dans le terme de contrôle du théorème 1 à celui d'une borne sur les nombres d'entropie de l'opérateur d'évaluation correspondant. Ce cheminement s'appuie, si l'espace de représentation est de dimension infinie, sur l'application du théorème de Maurey-Carl. Après la publication de notre article sur la sélection de modèle [98], nous avons poursuivi l'exploration de cette idée, dont l'exploitation pratique nécessite la détermination de la valeur de la constante universelle c apparaissant dans la formule (2.18). Autant le cas où l'opérateur S appartient à $\mathfrak{L}(\ell_1^n, H)$ est aisé à traiter, autant le résultat dual pose de très grandes difficultés. Le nombre de tentatives infructueuses effectuées par Williamson et ses co-auteurs est de ce point de vue éloquent. De février à juillet 2008, l'équipe accueillera Myriam Bertrand, dans le cadre d'un CRCT. Nous profiterons de ce renfort pour relancer cette étude, en nous appuyant sur des références bibliographiques déjà identifiées, portant sur la théorie des opérateurs [118, 199, 48], et les variables aléatoires à valeurs dans des espaces de Banach [134].

La thèse d'Olivier Bousquet [42] a connu un fort retentissement dans la communauté apprentissage, en mettant en lumière les bénéfices que celle-ci peut tirer des récents développements dans le domaine des inégalités de concentration [147, 40], de la théorie des processus empiriques [225] et plus particulièrement de l'utilisation des moyennes de Rademacher. Nous souhaitons continuer à étendre ces recherches au cas multiclasse. De manière générale, le travail de synthèse sur la théorie de la discrimination exposé dans [39] demeurera pour nous une source d'inspiration privilégiée dans l'optique d'une extension au cas multiclasse des travaux de pointe sur les bornes, standard ou relevant de l'apprentissage PAC-bayésien [156, 132, 6].

4.1.2 Théorie et pratique des machines à noyau

En dehors des SVM, la famille des machines à noyau comprend plusieurs systèmes discriminants calculant des dichotomies, parmi lesquels certains se distinguent par leurs propriétés statistiques ou leurs performances expérimentales, comme les machines à point bayésien [111] et la "Kernel Projection Machine" (KPM) [34]. Nous souhaitons étendre ces machines (borne sur le risque et algorithme d'apprentissage) au cas multiclasse. L'extension de la KPM devrait en particulier faire l'objet d'une collaboration avec Laurent Zwald.

Ces travaux généralistes (indépendants de toute application particulière) seront complétés par la conception et la mise en œuvre de noyaux dédiés aux séquences. Pour ce faire, nous nous appuierons de manière privilégiée sur les références suivantes : [138, 139, 54]. Une collaboration prévue avec Gérard Biau en apprentissage des données fonctionnelles [32] devrait nous permettre d'aborder le traitement de séquences autres que biologiques.

4.1.3 Apprentissage non supervisé

La présence dans l'équipe de Fabienne Thomarat, spécialiste de phylogénie moléculaire, nous incite à travailler à la conception et l'amélioration de méthodes phylogénétiques [79]. Nos recherches se développeront prioritairement suivant deux axes. D'une part, nous étudierons la robustesse des méthodes de classification aux données lacunaires et bruitées [242, 172]. D'autre part, nous développerons des méthodes de classification fondées sur l'emploi d'un noyau reproduisant et adaptées aux grandes masses de données [216].

4.2 Biologie computationnelle

En biologie computationnelle, nous souhaitons continuer à travailler principalement sur des problèmes de traitement de séquences. Cette thématique de la reconnaissance des formes, encore largement ouverte, est celle où notre expertise se trouve être la plus grande, et pour laquelle nous cherchons de manière privilégiée à développer des solutions à la fois génériques et opérationnelles (aboutissant à la mise en ligne de logiciels librement utilisables par la communauté). La majorité des problèmes auxquels s'intéressent les équipes de biologistes et bioinformaticiens avec lesquelles nos liens, établis de longue date, sont les plus forts, relève de ce domaine.

4.2.1 Ingénierie du noyau

Les travaux que nous effectuerons sur ce thème sont très liés à ceux évoqués dans les sections 4.1.2 et 4.1.3. Un axe que nous privilégierons pour le traitement des séquences biologiques est celui des noyaux fondés sur des HMM [115, 239, 107]. En comparaison des autres types de "string kernels" (voir par exemple [146, 198]), ceux-ci apparaissent, au moins en principe, mieux adaptés à la prise en compte des phénomènes de l'évolution biologique que sont les substitutions ainsi que les insertions/délétions. Leur utilisation soulève cependant des problèmes techniques importants (faiblesse du lien existant entre qualité de la modélisation et pouvoir discriminant, mauvais conditionnement de la matrice de Gram...), que nous nous attacherons à résoudre.

4.2.2 Développement d'architectures hybrides, intégrant systèmes discriminants et génératifs

Nos recherches dans ce domaine s'appuieront essentiellement sur la poursuite du développement de la méthode de prédiction de la structure secondaire des protéines globulaires décrite au chapitre 3. A la fin des années 90, nous avons introduit en prédiction de la structure secondaire l'idée d'effectuer un post-traitement des sorties de SVM au moyen de HMM [92]. Cette idée, inspirée par les célèbres travaux en parole associant réseaux de neurones et HMM (voir en particulier [163]) a récemment prouvé son utilité dans d'autres domaines de la reconnaissance des formes, dont à nouveau le traitement automatique de la parole [208]. Nous entendons continuer à la développer, en intégrant les contributions qu'effectuera Julien

Vannesson jusqu'à la fin de son contrat d'ingénieur, en novembre 2007, et celles d'Emmanuel Didiot, qui a rejoint le projet dans le cadre de l'emploi d'ATER qu'il occupe depuis septembre. Cette étude, conçue comme une suite des travaux exposés dans [97], fera également l'objet d'une collaboration avec Alain Lifchitz et Régis Vert. Notre principal objectif, dans la continuité des travaux du projet GENOTO3D, sera de prendre en compte les "dépendances à long terme", c'est-à-dire les interactions entre éléments distants dans la séquence mais proches dans la structure. Nos travaux en prédiction de la structure secondaire s'effectueront principalement dans le cadre de collaborations avec Nadir Mrabet et Gianluca Pollastri. Ils ne constitueront pas notre seule contribution à la biologie structurale prédictive, puisque nous avons également prévu de poursuivre la collaboration avec Nicolas Sapay et Gilbert Deléage portant sur la prédiction des ancrages membranaires interfaciaux dans les protéines membranaires monotopiques [190].

Bibliographie

[1] F. Aiolli and A. Sperduti. An efficient SMO-like algorithm for multiclass SVM. In *Neural Networks for Signal Processing 2002*, pages 297–306, 2002.

[2] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837, 1964.

[3] E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary : A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1 :113–141, 2000.

[4] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4) :615–631, 1997.

[5] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Miller. Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Research*, 25(17) :3389–3402, 1997.

[6] A. Ambroladze, E. Parrado-Hernandez, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *NIPS 19*, 2007. (à paraître).

[7] C.B. Anfinsen, E. Haber, M. Sela, and F.H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, 47(9) :1309–1314, 1961.

[8] D. Anguita, S. Ridella, and D. Sterpi. A new method for multiclass support vector machines. In *IJCNN'04*, pages 407–412, 2004.

[9] D. Anguita, S. Ridella, and D. Sterpi. Testing the augmented binary multiclass SVM on microarray data. In *IJCNN'06*, pages 3924–3926, 2006.

[10] C. Angulo and A. Català. K-SVCR. A multi-class support vector machine. In *ECML'00*, pages 31–38, 2000.

[11] C. Angulo, X. Parra, and A. Català. K-SVCR. A support vector machine for multi-class classification. *Neurocomputing*, 55(1–2) :57–77, 2003.

[12] M. Anthony and P.L. Bartlett. *Neural Network Learning : Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[13] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–404, 1950.

[14] K. Asai, S. Hayamizu, and K. Handa. Prediction of protein secondary structure by the hidden Markov model. *CABIOS*, 9(2) :141–146, 1993.

[15] V. Badeva and V. Morozov. *Problèmes incorrectement posés, théorie et applications*. MASSON, 1991.

[16] G.H. Bakir, T. Hofmann, B. Schölkopf, A.J. Smola, B. Taskar, and S.V.N. Vishwanathan, editors. *Predicting Structured Data*. The MIT Press, Cambridge, MA, 2007.

[17] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics*, 16(5) :412–424, 2000.

[18] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11) :937–946, 1999.

[19] P.L. Bartlett. The sample complexity of pattern classification with neural networks : The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2) :525–536, 1998.

[20] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3 :463–482, 2002.

[21] P.L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C.J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 4, pages 43–54. The MIT Press, Cambridge, MA, 1999.

[22] P.L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities : Some asymptotic results. In *COLT'04*, pages 564–578, 2004.

[23] E.B. Baum and D. Haussler. What size net gives valid generalization ? *Neural Computation*, 1 :151–160, 1989.

[24] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1) :74–86, 1995.

[25] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V.N. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2 :125–137, 2001.

[26] K. Benabdeslem and Y. Bennani. Dendogram-based SVM for multi-class classification. *Journal of Computing and Information Technology - CIT*, 14(4) :283–289, 2006.

[27] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5 :1089–1105, 2004.

[28] K. Bennett, N. Cristianini, J. Shawe-Taylor, and D. Wu. Enlarging the margins in perceptron decision trees. *Machine Learning*, 41(3) :295–313, 2000.

[29] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004.

[30] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1) :235–242, 2000.

[31] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank : a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3) :535–542, 1977.

[32] G. Biau, F. Bunea, and M.H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51(6) :2163–2172, 2005.

[33] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 2007. (à paraître).

[34] G. Blanchard, P. Massart, R. Vert, and L. Zwald. Kernel projection machine : a new tool for pattern recognition. In *NIPS 17*, pages 1649–1656, 2005.

[35] A. Bordes, L. Bottou, P. Gallinari, and J. Weston. Solving multiclass support vector machines with LaRank. In *ICML'07*, pages 89–96, 2007.

[36] A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6 :1579–1619, 2005.

[37] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[38] L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors. *Large-Scale Kernel Machines*. The MIT Press, Cambridge, MA, 2007.

[39] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. *ESAIM : Probability and Statistics*, 9 :323–375, 2005.

[40] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

[41] J.-M. Bouroche and G. Saporta. *L'analyse des données.* Presses Universitaires de France, 1994.

[42] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms.* Thèse de doctorat, Ecole Polytechnique, 2002.

[43] R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs. i. the method of paired comparisons. *Biometrika*, 39(3/4) :324–345, 1952.

[44] E.J. Bredensteiner and K.P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1/3) :53–79, 1999.

[45] H. Brezis. *Analyse fonctionnelle, Théorie et applications.* MASSON, 1993.

[46] P. Burman. A comparative study of ordinary cross-validation, ν-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3) :503–514, 1989.

[47] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. SVM and kernel methods MATLAB toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[48] B. Carl and A. Pajor. Gelfand numbers of operators with values in a Hilbert space. *Inventiones mathematicae*, 94(3) :479–504, 1988.

[49] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators.* Cambridge University Press, Cambridge, 1990.

[50] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5) :1155–1178, 2007.

[51] O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1) :131–159, 2002.

[52] P.Y. Chou and G.D. Fasman. Empirical predictions of protein conformation. *Annual Review of Biochemistry*, 47 :251–276, 1978.

[53] C. Combet, M. Jambon, G. Deléage, and C. Geourjon. Geno3D an automated protein modelling Web server. *Bioinformatics*, 18 :213–214, 2002.

[54] C. Cortes, P. Haffner, and M. Mohri. Positive definite rational kernels. In *COLT'03*, pages 41–56, 2003.

[55] C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995.

[56] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2 :265–292, 2001.

[57] K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2) :201–233, 2002.

[58] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, 2000.

[59] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In *NIPS 14*, pages 367–373, 2002.

[60] J.A. Cuff and G.J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins : Structure, Function, and Genetics*, 34(4) :508–519, 1999.

[61] Y. Darcy, E. Monfrini, and Y. Guermeur. Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes. In *JdS'06*, 2006.

[62] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. In M.O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–358. National Biomedical Research Foundation, Silver Spring, Washington DC, 1978.

[63] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, B 39 :1–38, 1977.

[64] J.P. Derrick and D.B. Wigley. The third IgG-binding domain from streptococcal protein G : An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *Journal of Molecular Biology*, 243(5) :906–918, 1994.

[65] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[66] E. Didiot. Conception et mise en œuvre de M-SVM dédiées au traitement de séquences biologiques. Mémoire de DEA, DEA informatique de Lorraine, 2003.

[67] T.G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2 :263–286, 1995.

[68] K.-B. Duan and S.S. Keerthi. Which is the best multiclass SVM method ? An empirical study. In *MCS'05*, pages 278–285, 2005.

[69] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (Second Edition).* Wiley Interscience, 2001.

[70] R.O. Duda, J.W. Machanik, and R.C. Singleton. Function modeling experiments. Technical Report 3605, Stanford Research Institute, 1963.

[71] R.M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6) :899–929, 1978.

[72] R.M. Dudley. A course on empirical processes. In P.L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XII - 1982*, volume 1097 of *Lecture Notes in Mathematics*, pages 1–142. Springer-Verlag, 1984.

[73] R.M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4) :1306–1326, 1987.

[74] A. Elisseeff. *Etude de la complexité et contrôle de la capacité des systèmes d'apprentissage : SVM multi-classe, réseaux de régularisation et réseaux de neurones multicouches.* Thèse de doctorat, ENS Lyon, 2000.

[75] A. Elisseeff, Y. Guermeur, and H. Paugam-Moisy. Margin error and generalization capabilities of multi-class discriminant models. Technical Report NC-TR-99-051-R, NeuroCOLT2, 1999. (révisé en 2001).

[76] A. Elisseeff and M. Pontil. Leave-one-out error and stability of learning algorithms with applications. In *NATO-ASI Series on Learning Theory and Practice*, pages 111–130, 2002.

[77] C.J. Epstein, R.F. Goldberger, and C.B. Anfinsen. The genetic control of tertiary protein structure : Studies with model systems. In *Cold Spring Harbor Symposium on Quantitative Biology*, volume 28, pages 439–449, 1963.

[78] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1) :1–50, 2000.

[79] J. Felsenstein. *Inferring Phylogenies.* Sinauer Associates, Inc., Sunderland, MA, 2004.

[80] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, Chichester, second edition, 1987.

[81] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3 :95–110, 1956.

[82] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1) :119–139, 1997.

[83] J. Friedman. Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University, 1996.

[84] K. Fukunaga. *Introduction to Statistical Pattern Recognition. Second Edition.* Academic Press, New York, 1990.

[85] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2 :721–747, 2002.

[86] C. Gaboriaud, V. Bissery, T. Benchetrit, and J.-P. Mornon. Hydrophobic cluster analysis : an efficient new way to compare and analyse amino acid sequences. *FEBS letters*, 224(1) :149–155, 1987.

[87] O. Gascuel. La dimension de Vapnik-Chervonenkis - Application aux réseaux de neurones. In S. Thiria, Y. Lechevallier, O. Gascuel, and S. Canu, editors, *Statistique et méthodes neuronales*, chapter 15, pages 244–261. DUNOD, 1997.

[88] O. Gascuel and J.-L. Golmard. A simple method for predicting the secondary structure of globular proteins : implications and accuracy. *CABIOS*, 4(3) :357–365, 1988.

[89] C. Geourjon and G. Deléage. SOPMA : significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS*, 11(6) :681–684, 1995.

[90] J.-F. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory. *Journal of Molecular Biology*, 198 :425–443, 1987.

[91] Y. Guermeur. *Combinaison de Classifieurs Statistiques, Application à la Prédiction de la Structure Secondaire des Protéines*. Thèse de doctorat, Université Paris 6, 1997.

[92] Y. Guermeur. Combining discriminant models with new multi-class SVMs. Technical Report NC2-TR-2000-086, NeuroCOLT2, 2000.

[93] Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2) :168–179, 2002.

[94] Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics*, 2007. (soumis).

[95] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8 :2551–2594, 2007.

[96] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deléage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5) :413–421, 1999.

[97] Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, chapter 9, pages 193–206. The MIT Press, Cambridge, MA, 2004.

[98] Y. Guermeur, M. Maumy, and F. Sur. Model selection for multi-class SVMs. In *ASMDA'05*, pages 507–517, 2005.

[99] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, and P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, 56C :305–327, 2004.

[100] Y. Guermeur and O. Teytaud. Estimation et contrôle des performances en généralisation des réseaux de neurones. In Y. Bennani, editor, *Apprentissage Connexionniste*, chapter 10, pages 279–342. Hermès, 2006.

[101] J. Guo, H. Chen, Z. Sun, and Y. Lin. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins : Structure, Function, and Bioinformatics*, 54 :738–743, 2004.

[102] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson. Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48(1) :239–250, 2002.

[103] L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. *Theoretical Computer Science*, 261(1) :81–90, 2001.

[104] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5 :1391–1415, 2004.

[105] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2) :451–471, 1998.

[106] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[107] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSD-CRL-99-10, Departement of Computer Science, University of California at Santa Cruz, 1999.

[108] D. Haussler and P.M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2) :219–240, 1995.

[109] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22) :10915–10919, 1992.

[110] S. Henikoff and J.G. Henikoff. Position-based sequence weights. *Journal of Molecular Biology*, 243(4) :574–578, 1994.

[111] R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1 :245–279, 2001.

[112] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963.

[113] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2) :415–425, 2002.

[114] S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure : Support vector machine approach. *Journal of Molecular Biology*, 308 :397–407, 2001.

[115] T. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *ISMB'99*, pages 149–158, 1999.

[116] T.S. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *1999 Conference on AI and Statistics*, 1999.

[117] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C.J.C Burges, and A.J. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 11, pages 169–184. The MIT Press, Cambridge, MA, 1999.

[118] W.B. Johnson and G. Schechtman. Embedding ℓ_p^m into ℓ_1^n. *Acta Mathematica*, 149 :71–85, 1982.

[119] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292 :195–202, 1999.

[120] W. Kabsch and C. Sander. Dictionary of protein secondary structure : Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12) :2577–2637, 1983.

[121] M. Karplus and G.A. Petsko. Molecular dynamics simulations in biology. *Nature (London)*, 347 :631–639, 1990.

[122] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, volume 1, pages 382–391. IEEE Computer Society Press, 1990.

[123] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3) :464–497, 1994.

[124] S.-J. Kim, A. Magnani, and S. Boyd. Optimal kernel selection in kernel Fisher discriminant analysis. In *ICML'06*, pages 465–472, 2006.

[125] G. Kimeldorf and G. Wahba. Some results of Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1) :82–95, 1971.

[126] J. Kindermann, E. Leopold, and G. Paass. Multi-class text categorization with error correcting codes. In *PKDD'00*, pages 558–565, 2000.

[127] S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited : A stepwise procedure for building and training a neural network. In F. Fogelman-Soulié and J. Hérault, editors, *Neurocomputing : Algorithms, Architectures and Applications*, volume F68 of *NATO ASI Series*, pages 41–50. Springer-Verlag, 1990.

[128] P. Koiran and E.D. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54(1) :190–198, 1997.

[129] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations, series 2*, 17 :277–364, 1961.

[130] U. Kreßel. Pairwise classification and support vector machines. In B. Schölkopf, C.J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 15, pages 255–268. The MIT Press, Cambridge, MA, 1999.

[131] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5 :27–72, 2004.

[132] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *NIPS 15*, pages 423–430, 2003.

[133] Y. Lechevallier. Classification non supervisée. In S. Thiria, Y. Lechevallier, O. Gascuel, and S. Canu, editors, *Statistique et méthodes neuronales*, chapter 10, pages 171–189. DUNOD, 1997.

[134] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, Berlin, 1991.

[135] Y. Lee. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. Technical Report 1063, University of Wisconsin, Madison, Department of Statistics, 2002.

[136] Y. Lee and Z. Cui. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16 :391–409, 2006.

[137] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines : Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465) :67–81, 2004.

[138] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel : A string kernel for SVM protein classification. *Proceedings of the Pacific Symposium on Biocomputing*, 2002.

[139] C. Leslie and R. Kuang. Fast kernels for inexact string matching. In *COLT'03*, pages 114–128, 2003.

[140] J.M. Levin and J. Garnier. Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochimica et Biophysica Acta*, 955 :283–295, 1988.

[141] J.M. Levin, B. Robson, and J. Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS*, 205(2) :303–308, 1986.

[142] Z. Li, S. Tang, and S. Yan. Multi-class SVM classifier based on pairwise coupling. In *SVM'02*, pages 321–333, 2002.

[143] V.I. Lim. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *Journal of Molecular Biology*, 88(4) :873–894, 1974.

[144] V.I. Lim. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *Journal of Molecular Biology*, 88(4) :857–872, 1974.

[145] Y. Lin. Support vector machines and the Bayes rule classification. *Data Mining and Knowledge Discovery*, 6 :259–275, 2002.

[146] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine learning Research*, 2 :419–444, 2002.

[147] G. Lugosi. Concentration-of-measure inequalities. Lecture notes, Summer School on Machine Learning at the Australian National University, Canberra, 2004.

[148] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3, 1969. (en russe).

[149] A. Marin, J. Pothier, K. Zimmermann, and J.-F. Gibrat. FROST : a filter-based fold recognition method. *Proteins*, 49(4) :493–509, 2002.

[150] J. Martin. *Prédiction de la structure locale des protéines par des modèles de chaînes de Markov cachées*. Thèse de doctorat, Université Paris 7, 2005.

[151] J. Martin, J.-F. Gibrat, and F. Rodolphe. Choosing the optimal hidden Markov model for secondary-structure prediction. *IEEE Intelligent Systems*, 20 :19–25, 2005.

[152] P. Massart. Concentrations inequalities and model selection. In *Ecole d'Eté de Probabilités de Saint-Flour XXXIII*, LNM. Springer-Verlag, 2003.

[153] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5) :2326–2366, 2006.

[154] B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2) :442–451, 1975.

[155] E. Mayoraz and E. Alpaydin. Support vector machines for multi-class classification. Technical Report 98-06, IDIAP, 1998.

[156] D.A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3) :355–363, 1999.

[157] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141 :148–188, 1989. Cambridge University Press, Cambridge.

[158] J. Meiler and D. Baker. Coupled prediction of protein secondary and tertiary structure. *PNAS*, 100(21) :12105–12110, 2003.

[159] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London, Series A*, 209 :415–446, 1909.

[160] M. Minoux. *Programmation Mathématique - Théorie et algorithmes.* Dunod, 1983.

[161] E. Monfrini and Y. Guermeur. A quadratic loss multi-class SVM. Technical report, LORIA, 2007. (en préparation).

[162] M. Moreira and E. Mayoraz. Improved pairwise coupling classification with correcting classifiers. In *ECML'98*, pages 160–171, 1998.

[163] N. Morgan, H. Bourlard, S. Renals, M. Cohen, and H. Franco. Hybrid neural network/hidden Markov model systems for continuous speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4) :899–916, 1993.

[164] B.K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1) :67–97, 1989.

[165] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics, 1994.

[166] M.N. Nguyen and J.C. Rajapkse. Multi-class support vector machines for protein secondary structure prediction. *Genome Informatics*, 14 :218–227, 2003.

[167] M.N. Nguyen and J.C. Rajapkse. Two-stage multi-class support vector machines to protein secondary structure prediction. In *Pacific Symposium on Biocomputing 10*, pages 346–357, 2005.

[168] M. Opper and O. Winther. Gaussian processes and SVM : Mean field and leave-one-out. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, chapter 17, pages 311–326. The MIT Press, Cambridge, MA, 2000.

[169] A. Passerini, M. Pontil, and P. Frasconi. New results on error correcting output codes of kernel machines. *IEEE Transactions on Neural Networks*, 15(1) :45–54, 2004.

[170] T.N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G.P. Gippert, and O. Lund. Prediction of Protein Secondary Structure at 80% Accuracy. *Proteins : Structure, Function, and Genetics*, 41(1) :17–20, 2000.

[171] W.W. Peterson and E.J. Weldon, Jr. *Error-Correcting Codes.* The MIT Press, Cambridge, MA, 1972.

[172] H. Philippe, E.A. Snell, E. Bapteste, P. Lopez, P.W.H. Holland, and D. Casane. Phylogenomics of eukaryotes : Impact of missing data on large alignments. *Molecular Biology and Evolution*, 21(9) :1740–1752, 2004.

[173] J.C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 12, pages 185–208. The MIT Press, Cambridge, MA, 1999.

[174] J.C. Platt. Probabilities for SV Machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, chapter 5, pages 61–73. The MIT Press, Cambridge, MA, 2000.

[175] J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *NIPS 12*, pages 547–553, 2000.

[176] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[177] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47(2) :228–235, 2002.

[178] N. Qian and T.J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202 :865–884, 1988.

[179] P. Ramesh and J.G. Wilpon. Modeling state durations in hidden Markov models for automatic speech recognition. In *ICASSP-92*, volume I, pages 381–384, 1992.

[180] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5 :101–141, 2004.

[181] S.K. Riis and A. Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of Computational Biology*, 3 :163–183, 1996.

[182] F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65 :386–408, 1958.

[183] B. Rost. Review : Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134(2) :204–218, 2001.

[184] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232 :584–599, 1993.

[185] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins : Structure, Function, and Genetics*, 19(1) :55–72, 1994.

[186] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235 :13–26, 1994.

[187] G. Rätsch, A.J. Smola, and S. Mika. Adapting codes and embeddings for polychotomies. In *NIPS 15*, pages 513–520, 2003.

[188] A. Sakurai. Tighter bounds of the VC-dimension of three-layer networks. In *WCNN'93*, volume III, pages 540–543, 1993.

[189] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins : Structure, Function, and Genetics*, 9 :56–68, 1991.

[190] N. Sapay, Y. Guermeur, and G. Deléage. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics*, 7(255), 2006.

[191] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13 :145–147, 1972.

[192] R.A. Sayle and E.J. Milner-White. RasMol : Biomolecular graphics for all. *Trends in Biochemical Sciences*, 20(9) :374–376, 1995.

[193] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3) :297–336, 1999.

[194] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *KDD'95*, pages 252–257, 1995.

[195] B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors. *Advances in Kernel Methods, Support Vector Learning*. The MIT Press, 1999.

[196] B. Schölkopf, R. Herbrich, and A.J. Smola. A generalized representer theorem. In *COLT'01*, pages 416–426, 2001.

[197] B. Schölkopf and A.J. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

[198] B. Schölkopf, K. Tsuda, and J.-P. Vert, editors. *Kernel Methods in Computational Biology*. The MIT Press, Cambridge, MA, 2004.

[199] C. Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *Journal of Approximation Theory*, 40 :121–128, 1984.

[200] J. Shawe-Taylor and M. Anthony. Sample sizes for multiple-output threshold networks. *Network : Computation in Neural Systems*, 2 :107–117, 1991.

[201] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

[202] S. Shelah. A combinatorial problem : Stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1) :247–261, 1972.

[203] K.T. Simons, R. Bonneau, I. Ruczinski, and D. Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins : Structure, Function, and Genetics*, Suppl. 3 :171–176, 1999.

[204] A.J. Smola. Regression estimation with support vector learning machines. Diplomarbeit, Technische Universität München, 1996.

[205] V.V. Solovyev and A.A. Salamov. Predicting α-helix and β-strand segments of globular proteins. *CABIOS*, 10(6) :661–669, 1994.

[206] E.D. Sontag. Feedforward nets for interpolation and classification. *Journal of Computer and System Sciences*, 45(1) :20–48, 1992.

[207] E.D. Sontag. VC dimension of neural networks. In C.M. Bishop, editor, *Neural Networks and Machine Learning*, pages 69–95. Springer-Verlag, Berlin, 1998.

[208] J. Stadermann and G. Rigoli. A hybrid SVM/HMM acoustic modeling approach for automatic speech recognition. In *INTERSPEECH 2004 - ICSLP*, pages 661–664, 2004.

[209] I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18 :768–791, 2002.

[210] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1) :128–142, 2005.

[211] I. Steinwart and C. Scovel. Fast rates for support vector machines. In *COLT'05*, pages 279–294, 2005.

[212] M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1) :29–35, 1977.

[213] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3) :293–300, 1999.

[214] J.A.K. Suykens and J. Vandewalle. Multiclass least squares support vector machines. In *IJCNN'99*, pages 900–903, 1999.

[215] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8 :1007–1025, 2007.

[216] F. Thomarat, C.P. Vivares, and M. Gouy. Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. *Journal of Molecular Evolution*, 59(6) :780–791, 2004.

[217] A. Tikhonov and V. Arsenin. *Solutions of Ill-Posed Problems*. Winston & Sons, 1977.

[218] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML'04*, pages 823–830, 2004.

[219] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6 :1453–1484, 2005.

[220] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1) :135–166, 2004.

[221] W. Utschick and W. Weichselberger. Stochastic organization of output codes in multiclass learning problems. *Neural Computation*, 13(5) :1065–1102, 2001.

[222] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11) :1134–1142, 1984.

[223] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

[224] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[225] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes, With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[226] J. Vannesson. Contribution au développement d'une méthode hybride discriminante-générative de prédiction de la structure secondaire des protéines. Mémoire de Master 2R, Master informatique de Nancy, 2007.

[227] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[228] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[229] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.

[230] V.N. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9) :2013–2036, 2000.

[231] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2) :264–280, 1971.

[232] R. Vert. Conception et mise en œuvre de M-SVM dédiées au traitement de séquences biologiques. Mémoire de DEA, DEA informatique de Lorraine, 2002.

[233] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF, Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, 1990.

[234] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 6, pages 69–88. The MIT Press, Cambridge, MA, 1999.

[235] G. Wahba, Y. Lin, and H. Zhang. GACV for support vector machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, chapter 16, pages 297–309. The MIT Press, Cambridge, MA, 2000.

[236] L. Wang, P. Xue, and K.L. Chan. Generalized radius-margin bounds for model selection in multiclass SVMs. Technical report, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, 2005.

[237] L.-H. Wang, J. Liu, Y.-F. Li, and H.-B. Zhou. Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Informatics*, 15(2) :181–190, 2004.

[238] J.J. Ward, L.J. McGuffin, B.F. Buxton, and D.T. Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13) :1650–1655, 2003.

[239] C. Watkins. Dynamic alignment kernels. Technical Report CSD-TR-98-11, Department of Computer Science, Royal Holloway, University of London, 1999.

[240] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

[241] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *ESANN'99*, pages 219–224, 1999.

[242] J.J. Wiens. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, 52(4) :528–538, 2003.

[243] R.C. Williamson, A.J. Smola, and B. Schölkopf. Entropy numbers of linear function classes. In *COLT'00*, pages 309–319, 2000.

[244] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines *via* entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6) :2516–2532, 2001.

[245] R.C. Williamson, A.J. Smola, and B. Schölkopf. Entropy numbers, operators and support vector kernels. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 9, pages 127–144. The MIT Press, Cambridge, MA, 1999.

[246] A. Zemla, Č. Venclovas, K. Fidelis, and B. Rost. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins : Structure, Function, and Genetics*, 34(2) :220–223, 1999.

[247] Q. Zhang, S. Yoon, and W.J. Welsh. Improved method for predicting β-turn using support vector machine. *Bioinformatics*, 21(10) :2370–2374, 2005.

[248] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5 :1225–1251, 2004.

[249] X. Zhang, J.P. Mesirov, and D.L. Waltz. Hybrid system for protein secondary structure prediction. *Journal of Molecular Biology*, 225 :1049–1063, 1992.

[250] A. Zien, F. De Bona, and C.S. Ong. Training and approximation of a primal multiclass support vector machine. In *ASMDA'07*, 2007.

Annexe A

Principales publications