Combinaison des cartes topologiques mixtes et des machines à vecteurs de support : Une application pour la prédiction de perte de poids chez les obèses

Mohamed Ramzi Temanni[*,**], Mustapha Lebbah[*], Christine Poitou-Bernert[**,***,****]
Karine Clement[**,***,****], Jean-Daniel Zucker[*,**]

[*] Université Paris 13, UFR de Santé,
Médecine et Biologie Humaine (SMBH) - Léonard de Vinci- LIM&BIO
74, rue Marcel Cachin 93017 Bobigny Cedex France
nom@limbio-paris13.org,
[**] Inserm, U755 Nutriomique, 75004 Paris, France;
[***] University Pierre and Marie Curie-Paris 6, Faculty of Medicine,
Les Cordeliers, 75004 Paris, France;
[****] AP-HP, Hôtel-Dieu Hospital, Nutrition department,
1 Place du parvis Notre-Dame, 75004 Paris, France
prénom.nom@htp.aphp.fr

Résumé. Cet article présente un modèle pour aborder les problèmes de classement difficiles, en particulier dans le domaine médical. Ces problèmes ont souvent la particularité d'avoir des taux d'erreurs en généralisations très élevés et ce quelles que soient les méthodes utilisées. Pour ce genre de problèmes, nous proposons d'utiliser un modèle de classement combinant le modèle de partitionnement des cartes topologiques mixtes et les machines à vecteurs de support (SVM). Le modèle non supervisé est dédié à la visualisation et au partitionnement des données composées de variables quantitatives et/ou qualitatives. Le deuxième modèle supervisé, est dédié au classement. La combinaison de ces deux modèles permet non seulement d'améliorer la visualisation des données mais aussi en les performances en généralisation. Ce modèle (CT-SVM) consiste à entraîner des cartes auto-organisatrices pour construire une partition organisée des données, constituée de plusieurs sous-ensembles qui vont servir à reformuler le problème de classement initial en sous-problème de classement. Pour chaque sous-ensemble, on entraîne un classeur SVM spécifique. Pour la validation expérimentale de notre modèle (CT-SVM), nous avons utilisé quatre jeux de données. La première base est un extrait d'une grande base médicale sur l'étude de l'obésité réalisée à l'Hôpital Hôtel-Dieu de Paris, et les trois dernières bases sont issues de la littérature.

1 Introduction

En apprentissage artificiel, on distingue deux grands thèmes, l'apprentissage supervisé et l'apprentissage non supervisé ; la plupart des problèmes d'apprentissage sont traités par l'une des deux approches. Selon les problèmes de classement, on a recours à de nombreuses méthodes telles que les machines à vecteurs de support (SVM) qui sont évaluées sur leur capacité à prédire correctement la classe des observations. Pour l'apprentissage non supervisé, on utilise souvent le modèle des cartes topologiques où les critères de qualité sont plus difficiles à définir ; ils s'articulent autour de l'interprétation des regroupements ou des partitions obtenues. Parmi les problèmes d'apprentissage, il existe une catégorie de problèmes qui sont appelés dans la littérature : difficiles, complexes, et plus particulièrement le problème traité dans ce papier concernant des données mixtes avec des variables quantitatives et qualitatives.

Une catégorie de modèles d'apprentissage spécifiques, combinant l'apprentissage non supervisé et supervisé, aussi bien dans le domaine de la classification hiérarchique et les arbres de décision que pour la recherche de partitions ont été développées, pour ce type de problème ; mais la majorité de ces modèles ne traitent que des données numériques, (Liu et al (2001); Rybnik et al (2003); Lebrun et al (2004); Sungmoon et al (2004); Shaoning et al (2005); Benabdeslem (2006)). Dans Wu et al (2004), les auteurs proposent d'utiliser les cartes topologiques de Kohonen (Kohonen (1995)) pour filtrer les données. Les observations non étiquetées héritent de l'étiquette de la classe du vote majoritaire de son sous-ensemble. A la fin de cette phase, un seul SVM est appris sur l'ensemble d'apprentissage initial réétiqueté, sans tenir en compte la partition des données. D'autre méthodes sont aussi inspirées des méthodes de partitionnement et de classement comme la définition de cartes topologiques dans l'espace de redescription (Sungmoon et al (2004)) ou l'utilisation des vecteurs supports pour définir une partition (Ben-Hur et al (2001)).

Notre approche est dédiée aux données numériques et/ou données mixtes, elle consiste à diviser le problème global de classement en sous-problème de classement guidé par la structure et l'organisation des données de la base en utilisant les cartes topologiques mixtes, (Lebbah et al (2005)). Le partitionnement des données avec les cartes, en une partition constituée de plusieurs sous-ensembles organisés vont servir à définir un classifieur pour chacun en utilisant les SVMs, Vapnik (1995). Les cartes topologiques mixtes sont utilisées dans notre modèle parce qu'elles sont de plus en plus utilisées comme outil de visualisation et de partitionnement non supervisé de différents types de données quantitatives et qualitatives codées en binaires. Elles permettent de projeter les données sur des espaces discrets qui sont généralement de dimensions deux et d'avoir des prototypes (représentants) du même type que les données initiales (quantitatives et qualitatives). Le modèle de base, proposé par Kohonen (Kohonen (1995)), est uniquement dédié aux données numériques. Les machines à vecteurs de support ont été développées dans les années 90 par Vapnik (1995). Ces méthodes ont été utilisées dans notre modèle parce qu'elles s'avèrent particulièrement efficaces. Elles peuvent traiter des problèmes mettant en jeu un grand nombre de variables tout en assurant une solution unique (pas de problèmes de minimum local comme pour les réseaux de neurones). L'algorithme sous sa forme initiale revient à chercher une frontière de décision linéaire entre deux classes, mais ce modèle peut considérablement être enrichi en se projetant dans un autre espace permettant ainsi d'augmenter la séparabilité des données.

Pour la compréhension de notre modèle, nous présentons dans la section 2, les différentes notations utilisées. Pour simplifier la présentation du papier, le modèle SVM et le modèle des

cartes topologiques ne seront pas présentés. Dans la section 2, nous présentons la combinaison des deux modèles que nous proposons d'utiliser pour le classement. Dans la section 3.1, une validation du modèle sur des données issues de la littérature ainsi que des données médicales réelles. Cette validation pemet de démontrer que notre modèle peut être utilisé pour augmenter les performances en classement du SVM sur ce type de bases de données.

2 Méthode hybride Cartes topologiques et SVM : CT-SVM

On suppose que l'on dispose d'une base d'apprentissage $\mathcal{A} = \{(\mathbf{z}_i, y_i); i = 1..N,\ \mathbf{z}_i \in \mathcal{D}\}$ où l'observation est $\mathbf{z}_i$, y_i l'étiquette de sa classe utilisé pour l'apprentissage du modèle SVM, et $\mathcal{D}$ représente l'espace des observations de dimension d. Les observations $\mathbf{z}_i$ sont composées de deux parties : la partie numérique $\mathbf{z}_i^r = (z_i^{1r}, z_i^{2r}, ..., z_i^{nr})$ ($\mathbf{z}_i^r \in \mathcal{R}^n$), et la partie qualitative codée en binaire $\mathbf{z}_i^b = (z_i^{1b}, z_i^{2b}, ..., z_i^{mb})$ ($\mathbf{z}_i^b \in \beta^m = \{0,1\}^m$). Avec ces notations une observation $\mathbf{z_i} = (\mathbf{z_i^r}, \mathbf{z_i^b})$ est de dimension $d = n + m$ (numérique et binaire). Comme tout modèle de cartes topologiques, nous supposons que l'on dispose d'une carte discrète $\mathcal{C}$ ayant N_{cell}. Cette structure de graphe permet de définir une partition de $\mathcal{D}$ en N_{cell} sous-ensembles qui sera notée $\mathcal{P} = \{P_1, ..., P_{N_{cell}}\}$. A chaque sous-ensemble P_c, on associe un vecteur référent $\mathbf{w}_c \in \mathcal{D}$ qui sera le représentant ou le "résumé" de l'ensemble des observations de P_c. Par la suite nous notons $\mathcal{W} = \{\mathbf{w_c} = (\mathbf{w_c^r}, \mathbf{w_c^b}); c = 1..N_{cell}\}$ l'ensemble des vecteurs référents constitués de la partie quantitative $\mathbf{w}_c^r \in R^n$ et de la partie qualitative $\mathbf{w}_c^b \in \{0,1\}^m$ suivant le même codage binaire que les données initiales, ce qui simplifie l'interprétation des référents. La partition $\mathcal{P}$ de $\mathcal{D}$ peut être définie d'une manière équivalente avec la fonction d'affectation de la carte ϕ qui est une application de $\mathcal{D}$ dans l'ensemble fini des indices $\mathcal{I} = \{1, 2, ..., N_{cell}\}$. Dans le cas où il y a eu regroupement des sous-ensembles, nous avons défini une application surjective χ de $\mathcal{I}$ dans l'ensemble des indices $\mathcal{J} = \{1, 2, ..., S\}$ où $1 \leq S \leq N_{cell}$. Si on utilise ces définitions, le sous-ensemble P_c est alors représenté par $P_c = \{\mathbf{z} \in \mathcal{D}/\phi(\mathbf{z}) = c, \chi(c) \in \mathcal{J}\}$, (si $\chi(c) = 1$ alors $\mathcal{P} = P_c = \mathcal{A}$). On notera par la suite, l'ensemble des indices $\mathcal{I}_p$ des sous-ensembles purs tel que $\mathcal{I}_p = \{c/\forall \mathbf{z} \in P_c, \chi(\phi(\mathbf{z})) = c, vote(P_c) = y_c\}$. y_c est l'étiquette du vote majoritaire à 100% du sous-ensemble P_c en utilisant la fonction $vote$. Par la suite, nous présentons un modèle de classement qui permet d'augmenter les performances en classement du SVM en utilisant le partitionnement des observations par les cartes topologiques.

Dans (Kuncheva , 2004, chapitre 6), l'auteur fournit une démonstration théorique pour ce type de modèles combinant partitionnement et classement. Si l'on considère que l'on dispose de S classeurs notés Cla associés à différents sous-ensembles P_i et si on note par $p(Cla_i/P_i)$ la probabilité du classement correct avec le classifieur Cla_i dans le sous-ensemble P_i, alors la densité de probabilité du classement correct de notre système de partitionnement et de classement s'écrit : $p(correct) = \sum_{i=1}^{S} p(P_i)p(Cla_i/P_i)$. $p(P_i)$ est la probabilité a priori que l'observation soit générée dans le sous-ensemble P_i. Pour maximiser ce mélange de probabilité, on choisit $p(Cla_i/P_i)$ tel que $p(Cla_i/P_i) \geq p(Cla_j/P_j), j = 1..S$.

Notre approche consiste à entrainer des SVMs ($Cla_i = SVM$) différents avec des sous-ensembles d'une partition $\mathcal{P}$ de la la base $\mathcal{A}$. Ceci permet de redéfinir des espaces de redescription différents (ou les mêmes) pour chaque sous-ensemble $P_c \in \mathcal{P}$, L'objectif de notre méthode CT-SVM est d'améliorer la discrimination en entraînant un SVM pour chaque sous-ensemble $P_c \in \mathcal{P}$ qui a plus d'une classe (les sous-ensembles non purs). Pour les sous-ensembles qui sont

composés d'observation de la même classe, aucun un SVM ne sera entrainé. Afin de réduire la partition et par conséquent le nombre de SVMs entrainés, nous avons utilisé la classification hiérarchique (CAH), sur l'ensemble des référents $\mathcal{W}$ de la carte pour réduire la partition ainsi le nombre de sous-ensembles,Yacoub et al (2001). Cette phase de réduction de la partition, qui consite à fusionner certains sous-ensembles, est optionnelle et elle peut être déterminée en interaction avec les experts et après visualisation des cartes topologiques ou avec un autre indice de regroupement, Vesanto et al (2000) .

L'algorithme de note modèle CT-SVM est le suivant : Pour un nombre de sous-ensembles S fixé faire :

- **Phase 1 :** Construction d'une partition $\mathcal{P} = \{P_1, ..., P_{N_{cell}}\}$ en utilisant les cartes topologiques mixtes constituées de N_{cell} cellules.
- **Phase 2 (optionnelle) :** Si $S < N_{cell}$ appliquer l'algorithme de regroupement pour construire la nouvelle partition $\mathcal{P} = \{P_1, ..., P_S / 1 \leq S \leq N_{cell}\}$
- **Phase 3 :** Détecter l'ensemble des indices $\mathcal{I}_p$ des sous-ensembles pures tel que $\mathcal{I}_p = \{c / \forall \mathbf{z} \in P_c, \chi(\phi(\mathbf{z})) = c, vote(P_c) = y_c\}$. y_c est l'étiquette du vote majoritaire à 100% du sous-ensemble P_c (toutes les observation de P_c portent la même étiquette y_c).
- **Phase 4 :** Apprentissage d'une SVM pour chaque sous-ensemble P_i tel que $i \notin \mathcal{I}_p$.

Remarque :
Pour l'apprentissage des cartes topologiques mixtes, nous avons utilisé notre programme développé en C/C++. Pour le regroupement des sous ensembles nous avons utiliser la classification hiérarchique. Nous avons aussi utilisé les programmes et l'heuristique développée par l'équipe de Kohonen, Vesanto et al (2000), pour estimer la dimension de la carte. Pour l'apprentissage du modèle SVM, nous avons utilisé la bibliothèque des programmes DAG-SVM (Directed Acyclic Graph SVM) dévellopé par Platt et al (2000); Cawley (2000).

Avec ce modèle CT-SVM, la topologie des observations est préservée grâce aux cartes topologiques. Lorsqu'on présente une nouvelle observation qui n'a pas participé à la phase d'apprentissage, elle sera projetée d'abord sur la carte topologique avec la fonction d'affectation associée ϕ. Puis, on utilisera la fonction d'affectation χ (voir §2), pour selectionner le sous-ensemble qui va déterminer le classifieur SVM associé. Cette methode d'affectation de notre classement permet de comprendre le comportement d'une observation à travers son référent $\mathbf{w}_c$ et/ou redéfinir une nouvelle partition en interaction avec l'expert. Si on note par svm_r la fonction de classement du modèle SVM du sous-ensemble P_r alors la fonction d'affectation globale de notre système s'écrit comme suite :

$$y_i = \begin{cases} svm_{\chi(\phi(\mathbf{z}_i))} & \text{si } \chi(\phi(\mathbf{z}_i)) \notin \mathcal{I}_p \\ vote(P_{\chi(\phi(\mathbf{z}_i))}) & \text{sinon} \end{cases} , \quad (1)$$

où $\mathcal{I}_p$ est l'ensemble des indice des sous-ensembles purs. $\chi(c) = c$ si $\mathcal{P} = \{P_1, ..., P_c, ..., P_{N_{cell}}\}$ et $\chi(c) = 1$ si $\mathcal{P} = \mathcal{A}$.

3 Expérimentations

3.1 Problème réelle : prédiction de perte de poids après chirurgie chez les obèses

Dans la suite, nous avons illustré les performances obtenues par notre modèle en utilisant une base réelle. Il s'agit d'un extrait d'une base médicale regroupant des données clinico-biologiques portant sur l'étude de l'obésité (Hôpital Hôtel-Dieu, Paris). Ces données constituent une base difficile en classement. On retrouve un taux de bon classement faible quelque soit la méthode utilisée (46.8% avec "Random Forest", 50.9% avec l'arbre de décision et 55% avec SVM). Cette base permet d'illustrer le potentiel des différentes visualisations des cartes topologiques mixtes et à montrer l'intérêt de diviser le problème global de classement pour augmenter les performances en classement du SVM. Cet exemple porte sur des données réelles, issues d'une base de données caractérisant 101 patients, massivement obèses (BMI : Body Mass Index$> 40kg/m^2$), recrutés et suivis dans le service de Nutrition de l'Hôtel Dieu dans le cadre d'une chirurgie de l'obésité, (Crookes (2006)). La base de données comporte des variables cliniques et biologiques, recueillies avant l'intervention chirurgicale. Les patients sont classés en deux groupes (oui/non) suivant la médiane de perte de poids observée 3 mois après la chirurgie (gastroplastie par anneau ajustable ou bypass gastrique). Si la perte de poids est supérieure à la médiane, le patient est étiqueté "oui". Sinon il est étiqueté par "non". Chaque patient est caractérisé par 37 variables réelles (par exemple, le poids, le BMI, ALAT, ASAT, HDL, CRP...) et 13 variables qualitatives (exemple : diabète oui/non), caractérisant l'obésité et ses aspects cliniques et métaboliques, ainsi que ses complications multiples.

Pour étudier le comportement de notre modèle en classement, nous avons procédé à une validation croisée en variant le nombre de sous-ensembles de la partition et par conséquent, le nombre d'observations associées à chaque apprentissage d'un SVM. Ainsi, nous avons découpé la base complète en trois sous bases de même taille, B_1, B_2, B_3. On apprend sur deux bases parmi les trois et on teste les performances en classement sur la troisième en utilisant l'étiquette de perte de poids (oui/non) à trois mois. Ainsi, en utilisant le modèle CT-SVM (§2), trois cartes topologiques sont construites de dimension 3×4, ce qui fournit une partition de 12 sous-ensembles ($N_{cell} = 12$). Pour montrer l'importance de la taille de la partition, nous avons calculé les performances en classement en, variant le nombre de sous ensembles de 1 à 12. Dans le premier cas, l'application de notre modèle CT-SVM sur une partition avec un seul sous-ensemble est équivalente à entraîner un SVM binaire classique sur toute la base.

La figure 1 montre les trois variations du taux de bon classement des trois bases de test en fonction du nombre de sous- ensembles de la même partition. Dans le cas où la partition contiendrait un seul sous-ensemble, un seul SVM est entraîné sur toute la base. Ainsi, dans ce cas particulier, la fonction d'affectation des cartes topologiques ϕ n'influe pas sur la fonction d'affectation globale de notre modèle CT-SVM (formule 1). On observe aussi dans la figure 1, que l'augmentation du nombre de sous-ensembles de la partition permet d'augmenter les performances en classement sur les trois tests. En revanche, on constate que lorsque la taille de la partition est très grande, les performances diminuent. La partition contenant plusieurs sous-ensembles permet d'apprendre autant de SVMs que de sous-ensembles. Ainsi, la fonction d'affectation globale de notre modèle (formule 1) utilise d'abord la fonction d'affectation des cartes topologiques ϕ pour choisir le sous-ensemble, ainsi le SVM associé avec sa fonction d'affectation svm. Avec le premier test, on obtient au maximum, un taux de bon

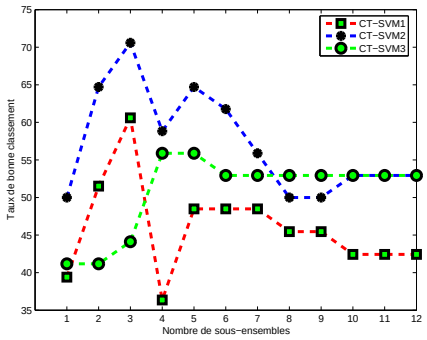


FIG. 1 – *Taux de bon classement avec CT-SVM en fonction du nombre de sous-ensembles.* □ *: base d'apprentissage :* B_1 *et* B_2*, base de test :* B_3 *;* • *: base d'apprentissage :* B_1 *et* B_3*, base de test :* B_2 *;* ○ *: base d'apprentissage :* B_2 *et* B_3*, base de test :* B_1*.*

classement de 60.6% avec trois sous-ensembles ; avec le deuxième test, on obtient un taux de bon classement 70.6% avec trois sous-ensembles. Finalement, avec le troisième test, on obtient 55.9 avec quatre sous-ensembles. Dans l'entraînement des SVMs avec notre modèle CT-SVM, nous avons utilisé la même fonction noyau linéaire. La validation croisée avec une variation du nombre de sous-ensembles montre l'intérêt et la difficulté de choisir la bonne partition pour une bonne discrimination. Cette partition est déterminée dans notre cas par expérimentation et visualisation des cartes topologiques. Cette validation croisée montre aussi l'intérêt de subdiviser le problème de classement global en sous-problèmes de classement pour améliorer les performances en classement.

3.2 Discussion

Puisque notre modèle utilise les cartes topologiques, on dispose d'un pouvoir de visualisation de la partition. L'application d'abord des cartes topologiques mixtes, va nous permettre d'analyser la répartition des observations et par conséquent les sous-ensembles qui ont servi au classement avec le SVM. Cette discussion va nous permettre de montrer l'intérêt de projeter les patients sur la carte pour comprendre le comportement et le profil de perte de poids du patient après chaque classement. L'apprentissage d'une carte de dimension 3×4 cellules effectué sur la base entière des patients, fournit pour chaque cellule un référent $\mathbf{w}_c$ composé de deux parties : la partie quantitative $\mathbf{w}_c^r$ et la partie qualitative $\mathbf{w}_c^b$ codée avec le codage disjonctif binaire. La figure 2.a présente la répartition des observations. On observe que la partition obtenue a permis de bien distribuer les observations sur 12 cellules de l'ensemble de la partition $\mathcal{P} = \{P_1, ..., P_{12}\}$, mais pour cet exemple, aucun des ces sous ensembles n'est pur. La figure 2.b présente la même répartition en distinguant ceux qui ont perdu ou non du poids à 3 mois par rapport à la médiane de l'ensemble des patients. On constate que les sous-ensembles sont mélangés. A l'aide de cette carte topologique 3×4, il est possible d'effectuer un certain nombre d'analyses de la base étudiée. Notre premier objectif est celui de partition-

ner les données, en tenant compte de leurs spécificités (données mixtes) pour augmenter les performances en classement. Pour visualiser la carte topologique, nous nous sommes limités à analyser les effets dûs à quelques variables pour lesquels l'exactitude des propriétés médicales retrouvées peuvent êtres vérifiées. En observant à la fois les deux figures 2.a et 2.b le médecin a détecté globalement trois grands groupes. Pour s'approcher de la partition du médecin, nous avons appliqué la CAH avec les référents de la carte pour avoir 4 sous-ensembles, $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$. La figure 7 présente la partition avec 4 sous-ensembles numéroté de 1 à 4. Cette répartition des données en quatre sous-ensembles et la répartition du médecin en trois sous-ensembles correspondent à la taille de la partition utilisée dans la phase de la validation croisée décrite ci-dessous.

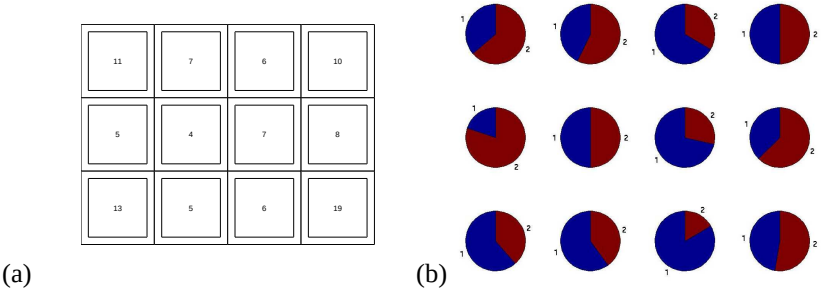
En visualisant à la fois les figures 2,3, 4, 5, 6 et la figure 7, il est possible de demander au médecin de définir les profils des patients. Ces profils vont servir à décrire les paramètres (variables) liés à la perte de poids et fournir des hypothèses de travail sur la résistance à la perte de poids fournies par le classeur SVM. Trois grands profils de patients sont définis selon la perte de poids à 3 mois. Le profil 1 est plutôt un bon profil par rapport aux pertes de poids à trois mois (figure 2.b) et correspond aux deux sous-ensembles P_1 et P_2 de la CAH. Le profil 2 est caractérisé par une perte de poids moyenne à 3 mois et correspond approximativement au sous ensemble P_4 de la CAH. Enfin le profil 3 est caractérisé par une perte de poids médiocre à 3 mois, ce qui aboutit à dénommer ce profil comme un "mauvais" profil en terme de perte de poids. Ce profil correspond au sous-ensemble P_3 de la CAH. Nous détaillons par la suite les deux profils 1 et 3 par rapport aux différentes variables clinico-biologiques.

Le profil 1 est caractérisé par un poids, un BMI (Body Mass Index) et une Dépense Energétique de Repos mesurée par calorimétrie (DERm) élevés. Les patients appartenants à ce profil ont une glycémie à jeûn et insulinémie élevées (figure 3) sans être diabétiques (figure 4). Il s'agit donc de patients insulinorésistants avant le stade de diabète. Le reste du profil métabolique est caractérisé par des HDL plutôt bas, des triglycérides (TG) et enzymes hépatiques (ASAT, ALAT et GGT) élevées,(figure 3). Dans les classes qualitatives "HTA" (hypertension) ou "SAS" (Syndrome d'apnées du sommeil) ces patients sont classés "oui" (figures 5 et 6). D'un point de vue inflammatoire, la CRP, la férritinémie (FERR), la SAA et l'orosomucoide (ORO), toutes des protéines de la phase aigue de l'inflammation, sont modérément élevées. Sur le plan nutritionnel, la TSH est basse, le profil protéique (albumine, préalbumine, RBP) et vitaminique est favorable, sans déficit. En conclusion pour ce profil, il s'agit de patients avec un poids très élevé mais dont le profil métabolique (figure 3) n'est pas trop évolué (sans diabète), sans inflammation importante et un bon profil nutritionnel.

Le profil 2 correspond à des patients ayant un BMI élevé et une leptine élevée (LEP, figure 3). Ils sont insulinorésistants mais pas diabétiques. Ils ont majoritairement une HTA et un SAS (figures 5 et 6). Les paramètres hépatiques et métaboliques sont normaux. L'adiponectinémie (ADIPO) est plutôt basse. En revanche, les paramètres inflammatoires (SAA et CRP) sont très élevés. Sur le plan nutritionnel, la TSH est normale, haute et les marqueurs nutritionnels sont bas (bilan protéique avec albumine, préalbumine et RBP, fer, vitamines A, E, B1, B12). Le profil 3 est un profil intermédiaire en terme de paramètres clinicobiologiques.

En conclusion, les deux profils de patients 1 et 2 sont caractérisés par des paramètres clinico-biologiques différents, notamment en terme de marqueurs d'inflammation et nutritionnels et sont aussi différents en termes de profil de perte de poids à 3 mois. Nous pouvons donc formuler l'hypothèse que le statut nutritionnel et l'état d'inflammation des patients avant

chirurgie pourraient être des éléments liés à la résistance à la perte de poids.



(a) (b)

FIG. 2 – *Carte topologique* 3×4 *(*$\mathcal{P} = \{P_1, P_2, ..., P_{12}\}$*). (a) Cardinalité des sous-ensembles (b) Répartition des pertes de poids à 3 mois. 1 : Pas de perte de poids ; 2 : perte de poids.*

3.3 Bases issues de la littératures

Dans cet exemple, trois bases d'apprentissage comportant un nombre variable d'observations ont été utilisées, (table 1) : Iris, Glass, Letter (Blake et al (1998)). Ces bases d'apprentissage et de test sont identiques à ceux pris dans l'article Benabdeslem (2006). Ceci va nous permettre de comparer nos résultats aux méthodes présentées dans l'article de Benabdeslem (2006) et montrer que notre méthode fonctionne aussi sur des bases de données classiques. Puisque toutes les variables sont quantitatives, l'utilisation des cartes topologiques mixtes se réduit pour ces bases à l'application de cet algorithme avec l'hyper-paramètre $F = 0$ qui correspond à la version batch des cartes topologiques classiques de Kohonen. Afin de mesurer la robustesse de notre système, l'apprentissage de notre modèle CT-SVM est réalisé sur les bases d'apprentissage présentées dans la table 1. L'affectation des observations de la base de test est réalisée à l'aide de la fonction d'affectation de notre modèle CT-SVM, présentée par la formule (1). La table 2 indique les performances atteintes avec notre modèle CT-SVM sur les bases de

nom/base	#Apprentissage	#Test	#classe	#variables
$Iris$	100	50	3	4
$Glass$	142	72	6	9
$Letter$	10000	5000	26	16

TAB. 1 – *Base d'apprentissage et de test.*

test des trois exemples en rappelant ceux du SVM classiques et l'algorithme DHSVM (Descendant Hierarchical Support Vector Machine). Dans la première base "Iris", le taux de bon classement est équivalent à celui du DHSVM et il est de l'ordre de 97.6%. Avec la deuxième base "Glass" une nette amélioration du taux de bon classement est constatée. On passe d'un taux de 71.5% avec le SVM "one againt one" à 81.9% avec notre modèle CT-SVM. Avec la troisième base, on constate que notre modèle CT-SVM arrive à un taux de 95.0% qui mieux que le MLP qui est de $\%85.2$, mais moins bon que le SVM classique et le DHSVM qui a le meilleur taux de 98.0%.

Base/modèle	one against one	one against all	MLP	DHSVM	**CT-SVM**
$Iris$	97.3	96.7	92.5	**97.6**	**97.6**
$Glass$	71.5	71.9	70.3	76.8	**81.9**
$Letter$	97.9	97.9	85.2	**98.0**	95.0

TAB. 2 – *Comparaison des performances en classement avec les algorithmes classique. SVM one against one, SVM one against all, MLP : Multi-Layer Perceptron, DHSVM :Descendant Hierarchical Support Vector Machine.*

4 Conclusion

Dans cet article, nous avons présenté un modèle de classement hybride, associant une méthode de partitionnement et une méthode de classement qui sont respectivement, les cartes topologiques et les SVMs. Ce modèle utilise l'organisation des données fournis par les cartes topologiques mixtes pour subdiviser l'espace des données afin d'apprendre un SVM spécifique pour chaque sous-espace des données. Notre modèle CT-SVM utilise la partition résultat des cartes topologiques, pour associer un SVM à chaque sous-ensemble de la partition avec des hyper-paramètres différents si cela est nécessaire. Les expériences effectuées montrent la robustesse de celui-ci à traiter des bases classiques avec uniquement des données réelles ou des données mixtes. D'autres parts, dans le cadre d'une application médicale réelle, nous avons vu que la quantité d'information fournie par ce modèle CT-SVM à travers les cartes topologiques mixtes est très importante et que le pouvoir de classement avec les SVMs est très performant. Nous avons aussi constaté, qu'il est important de choisir la taille de la partition. Ceci nous amène à réfléchir sur des indices qui permettent d'estimer la partition idéale et de faire une comparaison. Une comparaison avec d'autres méthodes classiques de classement est envisagée dans nos futurs travaux.

Références

Benabdeslem, K. (2006). Descendant hierarchical support vector machine for multi-class problems. International joint conference on neural network (IJCNN 2006) Vancouver .

Ben-Hur, A., D. Horn, H.T. Siegelmann, and V. Vapnik (2001). Support vector clustering, Journal of Machine Learning Research, vol. 2, pp. 125.

Blake C.L and C.J. Merz (1998). "UCI repository of machine learning databases". Technical report. University of California, Department of information and Computer science, Irvine, CA. available at : *ftp ://ftp.ics.uci.edu/pub/machine-learning-databases*.

Cawley, G. C, (2000). MATLAB Support Vector Machine Toolbox ($v0.55\beta$) *http ://theoval.sys.uea.ac.uk/gcc/svm/toolbox*, University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ.

Crookes, P.F. (2006). Surgical treatment of morbid obesity. Vol. 57 : 243-264. annurev.med.56.062904.144928.

Kohonen, T. (1995). Self-Organizing Map. Springer, third edition Berlin.

Kuncheva, L. I. (2004) Combining Pattern Classifiers, Methods and Algorithms. A Wiley-Interscience publication. ISBN 0-471-21078-1.

Lebbah, M., A. Chazottes, S. Thiria and F. Badran. ESANN (2005) Mixed Topological Map, ESANN 2005, Bruges, April 26-27-28, Proceedings.

Lebrun, G., C. Charrier, O. Lezoray, H. Cardot (2004). Réduction du temps d'apprentissage des SVM par Quantification Vectorielle . CORESA (COmpression et REpréesentation des signaux Audiovisuels), pp 223-226.

Liu, R. and B. Yuan.(2001). Multiple classifier combination by clustering and selection. Information Fusion, 2 :163-168.

Platt, J., N. Cristianini, J. Shawe-Taylor (2000) "Large Margin DAGs for Multiclass Classification", in Advances in Neural Information Processing Systems 12, pp. 547-553, MIT Press.

Rybnik, M., A. Chebira, K. Madani (2003). Auto-adaptive Neural Network Tree Structure Based on Complexity Estimator. IWANN (1) : 558-565.

Shaoning, P., D. Kim, S.Y. Bang (2005). Face Membership Authentication Using SVM Classification Tree Generated by Membershipbased LLE Data Partition, IEEE Trans. on Neural Network, 16(2) 436-446.

Sungmoon, C., Sang Hoon Oh Soo-Young Lee (2004). Support Vector Machines with Binary Tree Architecture. Neural Information Processing. Letters and Reviews Vol. 2, No. 3.

Vapnik, V.N. (1995). "The Nature of Statistical Learning Theory", Springer-Verlag, New York, ISBN 0-387-94559-8.

Vesanto, J., J. Himberg, E. Alhoniemi and J. Parhankangas (2000). "SOM Toolbox-Team". Helsinki University of Technology. P.O.Box 5400, FIN-02015 HUT, FINLAND. http ://www.cis.hut.fi/projects/somtoolbox/.

Vesanto, J., Alhoniemi, E.(2000). "Clustering of the Self-Organizing Map", IEEE Transactions on Neural Networks. N 3 pp 586-600.

Wu, W., X. Liu, M. Xu, J. Peng, R. Setiono (2004) A Hybrid SOM-SVM Method for Analyzing Zebra Fish Gene Expression. 17th ICPR'04 - Volume 2 pp. 323-326.

Yacoub, M., F. Badran and S. Thiria (2001). A Topological Hierarchical Clustering : Application to Ocean Color Classification ICANN proceedings.

Summary

This paper introduces a classification model combining mixed topological map and support vector machines. The non supervised model is dedicated for clustering and visualizing mixed data. The supervised model is dedicated to classification task. In the present paper, we propose a combination of two models performing a data visualization and classification. The task of our model is to train topological map in order to cluster data set on organized subset. For each subset, we propose to train a SVM model. The global classification problem is devided into classification sub problem corresponding to the number of subset. The model is validated related to the obesity problem, which is provied by Nutrition team located in hospital Hôtel-Dieu in Paris.

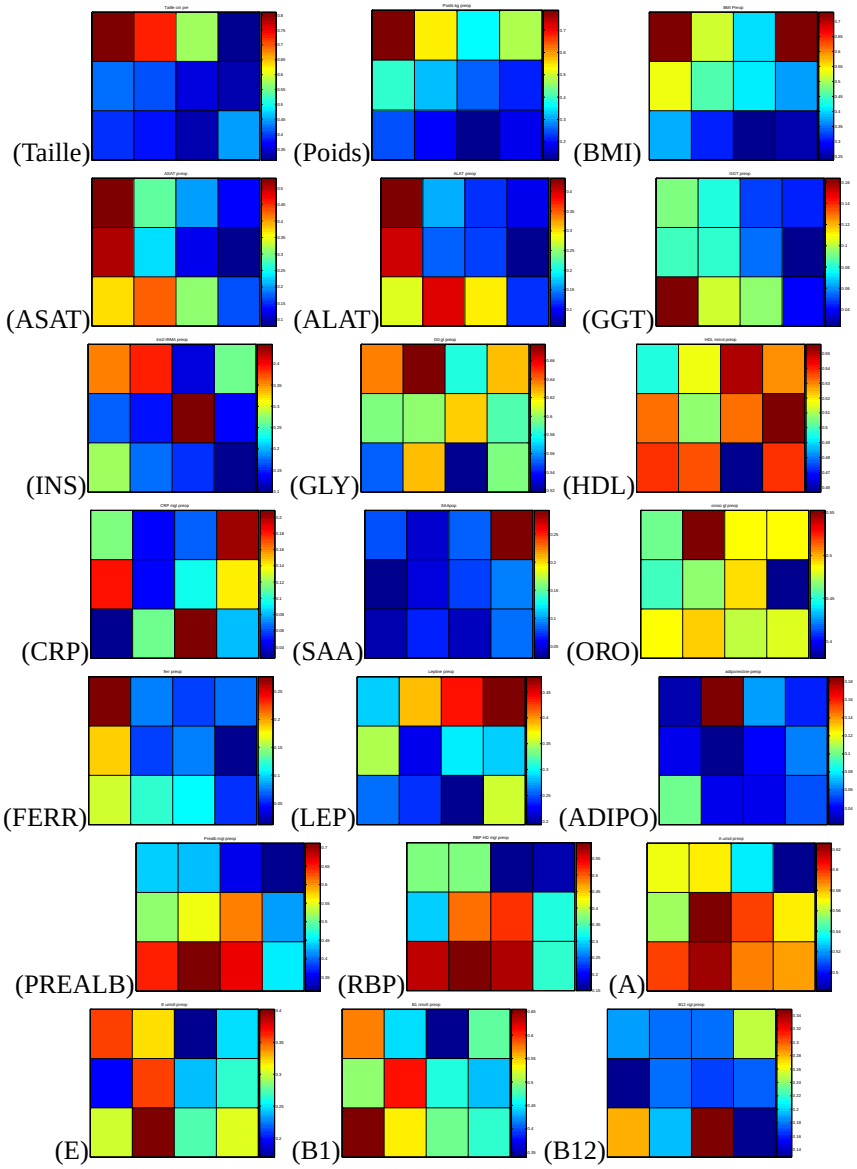


FIG. 3 – *Carte topologiques décrivant la variation sur les variables Taille,poids,BMI (Body Mass Index), ALAT, ASAT,GGT, INS(insuline), GLY (glycémie),HDL, CRP,SAA,ORO (orosomucoide),FERR (férritinémie),LEP (leptine),ADIPO (adiponectinémie),PREALB (préalbumine),RBP, A, E, B1, B12.*

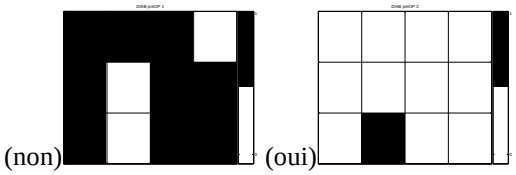


FIG. 4 – *Carte topologiques représentant les deux modalités non et oui de la variable qualitative Diabète.1 et 0 représente respectivement la présence ou l'absence de la modalité.*

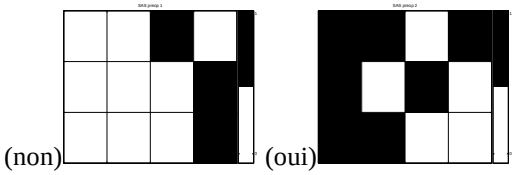


FIG. 5 – *Carte topologiques représentant les deux modalités non et oui de la variable qualitative SAS. 1 et 0 représente respectivement la présence ou l'absence de la modalité.*

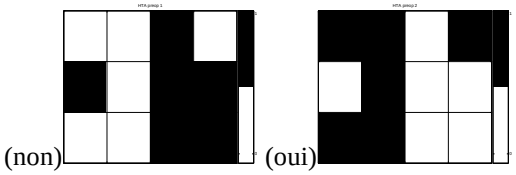


FIG. 6 – *Carte topologiques représentant les deux modalités non et oui de la variable qualitative HTA. 1 et 0 représente respectivement la présence ou l'absence de la modalité.*

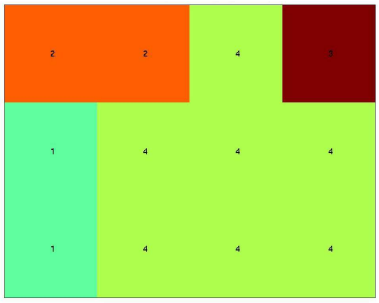


FIG. 7 – *Carte topologiques* 3×4 *après le partitionnement de la CAH.* $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$.