Ensemble prédicteur fondé sur les cartes auto-organisatrices adapté aux données volumineuses.

Elie Prudhomme*, Stéphane Lallich*

*Université Lumière Lyon 2, Laboratoire ERIC,
5 avenue Pierre Mendès-France
69676 Bron
eprudhomme@eric.univ-lyon2.fr, stephane.lallich@univ-lyon2.fr

Résumé. Le stockage massif des données noie l'information pertinente et engendre des problèmes théoriques liés à la volumétrie des données disponibles. Ces problèmes dégradent la capacité prédictive des algorithmes d'extraction des connaissances à partir des données. Dans cet article, nous proposons une méthodologie adaptée à la représentation et à la prédiction des données volumineuses. A cette fin, suite à un partitionnement des attributs, des groupes d'attributs non-corrélés sont créés qui permettent de contourner les problèmes liés aux espaces de grandes dimensions. Un Ensemble est alors mis en place, apprenant chaque groupe par une carte auto-organisatrice. Outre la prédiction, ces cartes ont pour objectif une représentation pertinente des données. Enfin, la prédiction est réalisée par un vote des différentes cartes. Une expérimentation est menée qui confirme le bien-fondé de cette approche.

1 Espaces de grandes dimensions

Les systèmes d'information tendent vers le stockage et l'analyse d'une quantité croissante d'information. En apprentissage, cette évolution a pour conséquence une augmentation massive du nombre d'individus et du nombre d'attributs. Elle est due à la fois à un faible coût de stockage et à une collecte plus facile des informations. Les individus d'intérêt sont ainsi devenus des objets de plus en plus complexes. C'est le cas par exemple des puces à ADN qui décrivent des individus à travers l'expression de milliers de gènes, des banques d'images dont chacune nécessite des centaines d'attributs ou encore de l'internet et des documents qu'il contient. Les outils classiques de l'apprentissage automatique, notamment ceux relatif à la prédiction, perdent une partie de leur efficacité pour traiter ces données (Verleysen, 2003). En effet, que ce soit du point de vue des individus ou des attributs, leur présence en un nombre important pose deux catégories de problème.

La première catégorie est relative à la qualité des données. Celle-ci dépend de la pertinence des informations récoltées. Elle est améliorée lors d'une phase de prétraitement. Du point de vue des individus, il s'agit de détecter ceux qui sont mal étiquetés comme ceux qui présentent des valeurs atypiques, ces deux types d'individus faussant fortement l'apprentissage et nuisant

à la généralisation (Beckman et Cooks (1983); Muñoz et Muruzábal (1998); Muhlenbach et al. (2003)). Du point de vue des attributs, il s'agit d'éliminer ceux qui apportent une information non pertinente ou redondante pour la prédiction, cachant ainsi l'information apportée par d'autres et augmentant inutilement la complexité du problème (Kira et Rendell, 1992). Or, plus le volume d'information est important, plus il est difficile de détecter les individus atypiques et de sélectionner les attributs pertinents.

La deuxième catégorie est relative aux algorithmes d'apprentissage mis en oeuvre. Tout d'adord, au travers de la complexité des algorithmes, la volumétrie des données a un impact direct sur le temps nécessaire à l'apprentissage. Ainsi, les algorithmes dont la complexité n'est pas linéaire en fonction des nombres d'individus et d'attributs ne sont-ils pas adaptés aux données volumineuses. Leur mise en place dans ce cadre est alors soumise à l'utilisation d'heuristiques ou de stratégies d'échantillonnage de la base d'exemples.

D'autre part, l'augmentation du nombre des prédicteurs pose certains problèmes théoriques qui affectent les algorithmes d'apprentissage. En particulier, cette augmentation dégrade la précision de l'apprentissage (malédiction de la dimension, Bellmann (1975)). Par ailleurs, elle affecte également la pertinence des mesures de distance entre les points (phenomène de concentration des distances, Demartines (1994)).

L'apprentissage supervisé de données volumineuses doit tenir compte de ces problèmes. Il s'agit donc de s'assurer de la qualité des données et de mettre en place des algorithmes adaptés aux problématiques particulières de ces espaces. Cet article étudie la capacité d'une méthode ensembliste fondée sur les cartes auto-organisatrices à répondre à ces attentes. Ici, les méthodes ensemblistes sont considérées au sens de Valentini et Masulli (2002) qui inclut le bagging (Breiman, 1996), le boosting (Freund, 1990) ou encore la combinaison de classifieurs sur des sous-ensembles d'attributs (Kuncheva et Whitaker, 2001).

La section suivante présente les cartes auto-organisatrises et leur adaptation à l'apprentissage supervisé (Sect. 2). Les Ensembles sont ensuite introduits (Sect. 3), puis leur utilisation avec les cartes auto-organisatrices (Sect. 4). Avant de conclure, nous présentons les résultats obtenus en appliquant notre méthodologie sur des données en grandes dimensions (Sect. 5).

2 Cartes auto-organisatrices, du non-supervisé au supervisé

Les cartes auto-organisatrices (ou *Self Organizing Map*, *SOM*) (Kohonen, 1982) permettent à la fois un apprentissage non-supervisé rapide des individus et leur représentation. Pour ce faire, elles se composent d'un réseau de neurones répartis uniformément dans un espace à 2 voire 3 dimensions. Chaque neurone est défini par un vecteur dans l'espace des individus, appelé vecteur de poids. Lors de l'apprentissage, les individus sont présentés successivement au réseau. Pour chaque individu, le neurone le plus proche (dit *Best Matching Unit, bmu*) et son voisinage dans le réseau sont modifiés afin qu'ensemble ils se rapprochent de l'individu.

L'algorithme classique de l'apprentissage du i^{eme} individu à l'instant t peut se résumer par la formule suivante de modification des poids w du neurone r :

$$w_r^{t+1} = w_r^t + h_r^t \times (x_i - w_r^t)$$

où $h_r^t = \alpha^t \times v_r^t$, avec α^t le pas d'apprentissage qui décroît linéairement avec le temps et v_r^t la fonction de voisinage qui définit l'étendue des neurones modifiés autour du bmu. En début d'apprentissage, tout à la fois les neurones à modifier se rapprochent fortement des individus (α^t grand) et le nombre de neurones à modifier autour du bmu (le voisinage) est large (v^t grand). Par la suite, l'ampleur de la modification et le nombre de neurones à modifier décroissent. Grâce à cet algorithme, on obtient une conservation de la topologie locale de l'espace des entrées. Pour deux individus, une proximité au sens des neurones correspond à une proximité dans l'espace de départ. La complexité des cartes auto-organisatrices est en $O(nd)$ (avec n le nombre d'individus et d le nombre d'attributs). Ceci en fait l'un des algorithmes d'apprentissage non-supervisé les plus rapides.

Du fait de ces propriétés (rapidité de l'apprentissage et préservation de la topologie), plusieurs auteurs ont adapté les cartes auto-organisatrices à l'apprentissage supervisé. Dans ce cadre l'approche la plus couramment utilisée est certainement la quantification vectorielle supervisée (notée *LVQ* pour *Learning Vector Quantization*) proposée par Kohonen (1988). Les neurones des cartes auto-organisatrices sont remplacés par des vecteurs de poids associés à l'une des classes à apprendre. L'apprentissage détermine itérativement le vecteur le plus proche de chaque individu présenté. A la différence des cartes auto-organisatrices, il sera le seul modifié. La règle d'apprentissage utilise alors la classe : si le vecteur et l'individu partagent la même classe le vecteur sera modifié de manière à se rapprocher de l'individu, dans le cas contraire il s'en éloignera.

Cette approche garde la simplicité et la faible complexité de l'algorithme des cartes auto-organisatrices face à la volumétrie des données. Cependant, elle n'effectue aucune projection de l'espace des entrées. Par ailleurs, l'étiquette aide à établir la position des vecteurs dans l'espace des entrées. Si cela conduit à une plus grande efficacité en phase de prédiction, cela pose aussi le problème de la conservation de la topologie. En effet, il n'est plus possible de représenter la topologie de l'espace des entrées puisque la position des prototypes ne consiste plus en une simple projection de cet espace sur celui d'une carte.

Afin d'obtenir une représentation topologique de l'espace des entrées indépendante de la classe, à la manière de Zighed et al. (2004) pour les graphes de voisinage, d'autres travaux ont séparé l'apprentissage en deux phases. Une première phase réalise la construction de la carte en se fondant uniquement sur les variables prédictives, la deuxième phase se charge d'étiqueter les neurones obtenus. L'étiquette correspond à la classe majoritaire des individus représentés par le neurone. Au terme de l'apprentissage, une fonction de prédiction permet d'attribuer son étiquette à un nouvel individu. Sur ce modèle, différentes méthodes ont été proposées. Elles divergent seulement sur la fonction de prédiction. C'est le cas de la méthode *Kohonen-KNN* (Zupan et al., 1994), qui sera améliorée par *Kohonen-WI* (Song et Hopke, 1996) puis par *Kohonen-Opt* (Prudhomme et Lallich, 2005b). L'avantage majeur de cette approche est de fournir, outre un modèle de prédiction, une carte qui constitue la projection de l'espace des prédicteurs. Du fait de la construction non supervisée de la carte, celle-ci est indépendante de l'étiquette des données.

Les cartes auto-organisatrices sont relativement bien adaptées aux données volumineuses. D'abord en ce qui concerne leur passage à l'échelle, elles possèdent un algorithme d'une com-

plexité linéaire en fonction du nombre d'individus et de variables. Ensuite, du point de vue de la qualité des données, les cartes auto-organisatrices construisent une nouvelle représentation par le biais d'une projection non-linéaire. Elles s'intéressent ainsi à la pertinence des différents attributs en les pondérant implicitement en vue de la construction de la carte. Par ailleurs, la construction de cette représentation est robuste au bruit (Liu et al., 2006) et aux données manquantes. Une carte auto-organisatrice constitue en ce sens une bonne représentation des données. Qui plus est, il est possible de valider statistiquement cette représentation dans le cadre de l'apprentissage supervisé (Prudhomme et Lallich, 2005b). Cependant, les cartes auto-organisatrices restent sensibles aux problèmes théoriques rencontrés dans les espaces de grandes dimensions (voir Sect. 1). Afin de les adapter à ces espaces, nous avons choisi de les utiliser dans le cadre de méthodes ensemblistes.

3 Méthodes ensemblistes

3.1 Introduction.

Les Ensembles désignent une très vaste famille d'algorithmes d'apprentissage supervisé qui ont comme architecture fondatrice l'utilisation parallèle de plusieurs apprenants pour réaliser la prédiction. Plus précisément, (1) l'information – portée par les individus, les prédicteurs ou encore la variable de classe – est répartie entre différents apprenants, (2) chaque apprenant réalise l'apprentissage de l'information qui lui a été impartie et (3) pour prédire le résultat d'un nouvel exemple les résultats de chaque apprenant sont combinés. Cette architecture est utilisée avec succès par de nombreuses méthodes. Par exemple, le bagging (*Boostrap AGGregatING*) construit L apprenants à partir de L échantillons bootstrap de l'ensemble d'apprentissage (Breiman, 1996). Il joue ainsi sur la composante variance de l'erreur, principalement due à l'échantillon de départ. Le boosting construit, quant à lui, L classifieurs à partir de L échantillons pondérés de l'ensemble d'apprentissage. Au cours de ces échantillonnages, la pondération renforce la probabilité d'apparition des exemples les plus difficiles à apprendre. De cette manière, le boosting réduit la composante biais de l'erreur du classifieur (Meir et Rätsch, 2003). D'autres stratégies que l'échantillonnage des individus sont possibles pour la construction d'Ensembles. C'est le cas, par exemple, des mixtures d'experts, où chaque apprenant rentre en compétition avec les autres pour apprendre un individu (Jacobs et al., 1991). C'est aussi le cas des données issues de différents capteurs pour lesquels un individu est appris par tous les apprenants mais avec, à chaque fois, des attributs issus de capteurs différents (Duin et Tax (2000)). Une liste plus exhaustive peut être trouvée dans Valentini et Masulli (2002).

3.2 Diversité.

L'une des raisons majeure du succès de cette architecture est résumée par la notion de diversité (Brown et al. (2005) pour une revue de cette notion). Bien qu'il n'en existe pas de définition formelle, la diversité est la capacité des différents apprenants d'un Ensemble à ne pas commettre les mêmes erreurs lors de la prédiction. En effet, plus les apprenants se trompent sur des régions différentes de l'espace d'apprentissage, plus il est probable que la majorité d'entre

eux trouvent la réponse exacte pour une région particulière. Cette intuition sur la diversité est confirmée par plusieurs résultats théoriques.

Le premier d'entre eux a été obtenu par Geman et al. (1992) pour le cas d'Ensembles en régression. Ils montrent que l'erreur quadratique moyenne d'un Ensemble de M apprenants appliqué à la régression s'écrit :

$$E\left\{\left(\frac{\sum_i f_i}{M} - d\right)^2\right\} = \overline{bias}^2 + \frac{1}{M}\overline{var} + \left(1 - \frac{1}{M}\right)\overline{covar}\ . \tag{1}$$

avec $\overline{bias}$ le biais moyen, $\overline{var}$ la variance moyenne et $\overline{covar}$ la covariance moyenne des apprenants de l'Ensemble. Cette covariance quantifie la corrélation qui existe entre les erreurs des apprenants de l'Ensemble. En ce sens, elle constitue une expression de la diversité. De plus, à l'inverse de la variance et du biais, la covariance prend des valeurs négatives. Elle constitue donc un terme important de l'erreur quadratique moyenne d'un Ensemble : une corrélation négative entre les erreurs des classifieurs aura pour effet de réduire l'erreur totale.

Dans le cas de valeurs discrètes ordonnées, Tumer et Gosh (1995) formalisent le problème différemment. Pour une variable prédictive x et deux classes a et b (avec $a < b$), ils s'intéressent aux probabilités *a posteriori* réelles de prédire correctement a et b, notées $P(a|x)$ et $P(b|x)$. Lorsque les valeurs de x ne suffisent pas à séparer les classes a et b, il existe une erreur bayésienne irréductible. De plus, à cause des paramètres du modèle et du nombre fini d'individus, $P(a|x)$ et $P(b|x)$ ne sont qu'approximés. Outre l'erreur irréductible, il existe alors une erreur ajoutée. Tumer et Gosh (1995) ont montré que l'erreur ajoutée par un Ensemble s'écrit :

$$E_{add}^{ens} = \left(\frac{1 + \delta(M-1)}{M}\right) E_{add}\ . \tag{2}$$

E_{add} représente l'erreur ajoutée d'un apprenant et δ la corrélation des erreurs d'approximation (commises par les différents apprenants) des probabilités a posteriori. De même que pour la covariance, il s'agit là encore d'une mesure de la diversité. En effet, si $\delta = 0$, les apprenants sont indépendants, l'erreur de l'Ensemble est $\frac{1}{M}$ fois l'erreur d'un apprenant. A l'inverse, si $\delta = 1$, les apprenants sont dépendants, l'erreur de l'Ensemble est alors la même que l'erreur d'un apprenant.

Ces résultats montrent bien que la diversité est au coeur de la performance des Ensembles : elle assure que l'Ensemble est plus performant que l'algorithme d'apprentissage qu'il utilise. Cependant, aucun résultat théorique n'a pu être obtenu concernant cette diversité dans le cas de variables discrètes non-ordonnées. Autrement dit, aucune définition unanime ne se dégage. Pire, les différentes mesures existantes de la diversité ne semblent pas corrélées avec la diminution de l'erreur en prédiction d'un Ensemble par rapport à un apprenant seul (Kuncheva et Whitaker, 2003).

Cependant, la plupart des auteurs s'accordent sur son importance. En effet, outre l'extrapolation possible des résultats en régression ou pour des variables discrètes ordonnées, les résultats empiriques sont nombreux. La revue menée par Valentini et Masulli (2002) donne une idée de la fertilité de cette approche. Il apparaît qu'à chaque fois il est possible d'améliorer la prédiction d'un classifieur seul, que les différentes bases d'apprentissage soient construites à partir des individus, des attributs ou de la variable de classe.

3.3 Ensembles et espaces de grandes dimensions.

Dans le cadre des données en grandes dimensions, les Ensembles semblent donc particulièrement intéressants. Puisque le problème est le nombre d'attributs à apprendre par un classifieur, une solution adaptée consiste à profiter de l'approche ensembliste pour réduire ce nombre. En mettant en place plusieurs groupes d'attributs, chacun appris par un classifieur, nous visons plusieurs objectifs. Le premier est, bien sûr, de contourner les problèmes théoriques liés aux espaces de grandes dimensions. Cependant, en opposition à une sélection des attributs durant une phase de prétraitement, l'idée est ici de les conserver en totalité. L'avantage, comme le souligne Verleysen et al. (2003), est que la redondance de l'information prend part à une réduction du bruit sur chaque variable. Le deuxième objectif est l'amélioration de la prédiction. Elle résulte de la diversité des classifieurs qui sont construits sur chaque groupe. Ainsi, l'utilisation d'Ensembles dans le cadre de données en grandes dimensions permet de conserver la globalité de l'information tout en contournant les problèmes posés par des grandes dimensions.

4 Ensembles et cartes auto-organisatrices

La mise en place d'une approche ensembliste nécessite une réflexion sur trois points :
- Diversité : comment, à partir des données initiales, créer des jeux de données induisant des classifieurs divers ?
- Apprenants : quels algorithmes sont utilisés pour l'apprentissage de chaque jeu ?
- Aggrégation : comment les résultats des différents classifieurs sont-ils agrégés en vue d'obtenir la prédiction d'un nouvel individu ?

En ce qui concerne l'algorithme d'apprentissage, les cartes auto-organisatrices ont été choisies pour leur capacité de représentation et leur faible complexité algorithmique. Par la suite, nous nous intéressons donc plus précisément à la création de la diversité et à l'agrégation des différents résultats.

4.1 Diversité

Comme il a été souligné en Sect. 3.2, la diversité est un point crucial des Ensembles, puisqu'elle permet de rogner sur le taux d'erreur d'un apprenant seul. La Sect. 3.3 montre que l'utilisation des attributs pour créer cette diversité est une bonne idée dans le cadre des espaces de grandes dimensions. Les différents jeux de données seront donc instanciés à partir de différents groupes d'attributs. A savoir, si le jeu de données initial a n individus et d attributs, k nouveaux jeux de données sont constitués, chacun ayant n individus mais seulement d' attributs ($d' < d$). Il reste à déterminer comment ces partitions sont créées.

Si l'on considère la corrélation entre attributs comme critère de regroupement, deux solutions opposées se dégagent. La première consiste à former des groupes d'attributs corrélés entre eux. Dans ce cas, chaque groupe ne contient qu'une petite partie de l'information globale. Un Ensemble de cartes auto-organisatrices construit sur ce modèle aura sans conteste une diversité importante puisque chaque jeu représentera une information différente. A l'inverse, la seconde solution consiste à former des groupes d'attributs non corrélés entre eux. Dans ce

cas, chaque groupe contiendra une partie importante de l'information globale. Cependant, au niveau de l'Ensemble, la question est de savoir si la diversité sera suffisamment importante. Plus précisément, le dilemme se résume au travers de deux grandeurs : la qualité d'un classifieur au sein de l'Ensemble et la diversité de l'Ensemble. Dans le premier cas, la qualité d'un classifieur est très faible mais la diversité très importante ; dans le deuxième cas la situation est inversée.

Ces deux solutions conduisent également à des qualités différentes en matière de représentation des données. Un groupe d'attributs non corrélés fournit, par le biais des cartes auto-organisatrices, une représentation possible des données initiales. Parce qu'une partie de l'information a été enlevée, cette représentation n'est pas optimale. Cependant, elle n'est pas unique non plus, l'Ensemble étant par nature construit à partir de plusieurs apprenants (ici plusieurs cartes donc plusieurs représentations). Ces représentations peuvent être directement utilisées pour la navigation à travers la base d'exemples ou pour la recherche d'information ; plusieurs représentations donnent plusieurs points de vue. A l'inverse, un groupe d'attributs corrélés n'est pas en mesure de constituer une représentation des données initiales. Tel que, ce type de groupe ne contient pas suffisamment d'information. En revanche, la projection d'attributs corrélés par la carte constitue une synthèse de l'information portée par un groupe, capable de limiter le bruit présent sur chaque attribut. En d'autres termes, ce ne sont pas les données initiales qui sont représentées, mais des sous-parties de ces dernières. Ces représentations ne peuvent pas être utilisées directement pour la navigation ou l'interrogation. Au contraire, il faut construire une représentation supplémentaire pour y parvenir.

Dans cet article, nous nous sommes intéressés à la formation de groupes d'attributs non corrélés entre eux. Ces groupes serviront à instancier les jeux de données appris par les cartes auto-organisatrices. La construction de ces groupes est réalisée en deux phases (voir Fig. 1) :

- Classification des attributs : cette phase a pour objectif de construire des grappes d'attributs par le biais d'une classification non supervisée. Parmi différentes méthodes possibles (classification hiérarchique (Ward, 1963) ou k-moyennes (McQueen, 1967) sur la matrice attributs/valeurs transposée, par exemple), nous avons choisi la méthode Varclus (SAS, 1989). Elle est conçue spécifiquement pour la classification d'attributs et présente l'avantage de déterminer automatiquement le nombre de grappes. Il s'agit d'une méthode divisive. A chaque itération, l'algorithme calcule les deux premiers vecteurs propres de la matrice attributs/valeur. L'idée est de construire les grappes autour de ces vecteurs en affectant chaque attribut au vecteur avec lequel il est le plus corrélé (r^2 maximum). Cependant, pour éviter que le premier vecteur propre ne soit corrélé avec la grande majorité des variables, une rotation orthogonale des vecteurs propres est effectuée par la méthode varimax (Kaiser, 1958). On itère pour chaque grappe ainsi formée jusqu'à ce que la deuxième valeur propre obtenue soit inférieure à 1. A l'issue de la procédure, on obtient une partition en k grappes d'attributs.
- Formation des groupes : à partir des k grappes initiales, k nouveaux groupes sont construits. L'ajout d'attributs à ces groupes s'est fait de 3 manières différentes :
  - Groupes complets : Pour chaque groupe, une variable est choisie au hasard à l'intérieur de chaque grappe initiale. On obtient donc k groupes de k variables, chacune représentant une grappe différente. Comme illustré par la Fig. 1, lorsqu'une grappe contient moins de k attributs, certains de ces attributs sont réutilisés. A l'inverse, lorsqu'une grappe contient plus de k attributs, certains sont omis.

– Groupes semi-complets : Lorsque k est grand devant $\sqrt{d}$, on peut se limiter à sélectionner $\sqrt{d}$ grappes pour représenter un groupe. Cela limite la taille des données à apprendre par chaque carte. Dans ce contexte, lorsqu'un attribut d'une grappe est selectionné pour un groupe, aucun attribut de cette grappe n'est re-sélectionné tant que les autres grappes n'ont pas fourni d'attributs. De la même manière, lorsqu'une grappe est à nouveau sélectionnée pour fournir un attribut, cet attribut est choisi parmi ceux qui n'ont pas encore été sélectionnés pour un groupe. Ces deux règles permettent de répartir au mieux l'information contenue dans les grappes. A l'issue de cette procédure, on obtient donc k groupes de $\sqrt{d}$ variables, chacune représentant une grappe différente.
– Hasard : Seuls $\sqrt{d}$ attributs sont sélectionnés pour représenter un groupe. Les attributs sont tirés sans remise parmi les d attributs de départ. On obtient donc k groupes de $\sqrt{d}$ variables, sans relation avec les grappes formées par Varclus.

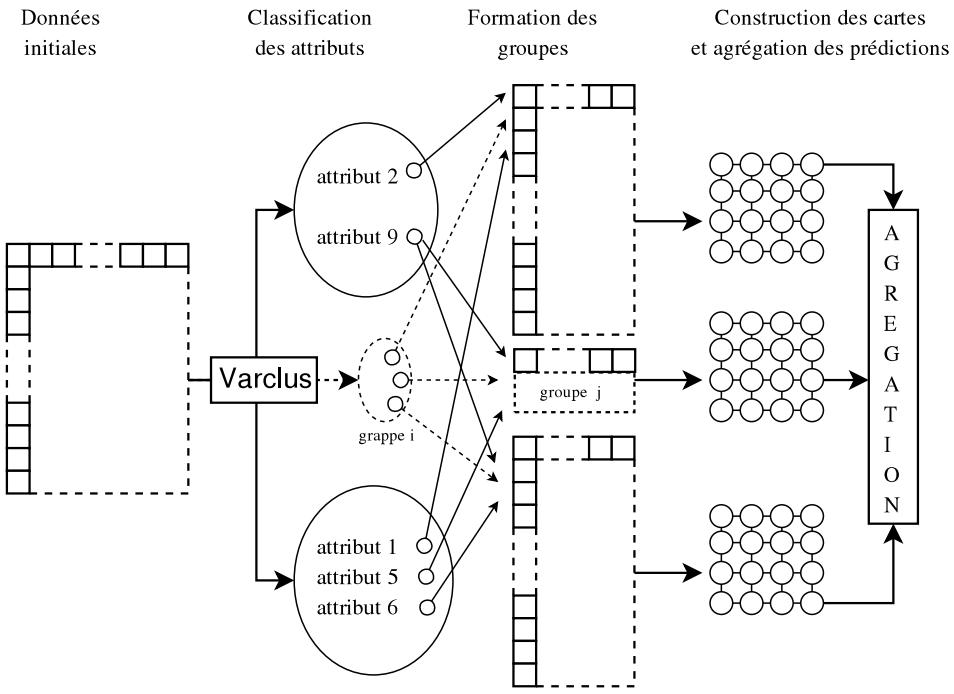

FIG. 1 – *Les différentes étapes de construction d'un Ensemble de cartes à partir de groupes d'attributs non corrélés.*

4.2 Agrégation

Une carte auto-organisatrice en prédiction est mise en place sur chacun des k groupes. Pour un nouvel individu, k prédictions doivent donc être agrégées pour donner la prédiction finale. Nous utilisons ici le vote à la majorité, qui à l'avantage d'être simple tout en donnant des résultats aussi satisfaisant que d'autres méthodes (Duin et Tax, 2000). Cependant, la prédiction

des cartes passant par la représentation, une agrégation fondée sur cette représentation est envisageable. En particulier, une statistique corrélée avec le taux d'erreur en généralisation d'une carte (Prudhomme et Lallich, 2005a) pourrait pondérer efficacement le vote.

5 Expérimentations

Avant de situer notre approche par rapport à d'autres méthodes ensemblistes, nous avons d'abord validé son bien-fondé. Pour ce faire, nous comparons ses résultats en 10-validation croisée aux résultats obtenus par Kohonen-Opt. De même, nous comparons les différentes méthodes de création des groupes définies en Sect. 4.1. Les données utilisées, présentées dans le tableau 1, proviennent de Newman et al. (1998). Pour chacun des jeux, les paramétrages des cartes sont identiques (dimension 20×20, 17 cycles d'apprentissage) donc non-optimisés. Les résultats sont reportés dans le tableau 2.

Données	**Attributs**	**Classes**	**Individus**
(1) Ionosphère	34	2	351
(2) Multi-features (Profile correlations)	76	10	2000
(3) Multi-features (Fourier coefficients)	216	10	2000

TAB. 1 – *Jeux de données*

Ils montrent d'abord l'avantage des méthodes ensemblistes sur une carte auto-organisatrice seule. Plus intéressant, cet avantage ne se retrouve qu'avec des groupes dont la création est guidée par la connaissance d'une classification des attributs (colonne Hasard, Tab. 2). Sans cette connaissance, la création d'un Ensemble de cartes n'améliore pas la capacité prédictive. On remarquera d'ailleurs que la création de diversité est l'élément majeur de la réussite de notre méthode. En effet, si l'on se limite à une ACP préalable des données, aux facteurs de laquelle on applique les cartes auto-organisatrices, les résultats sont décevants. Enfin, entre les groupes complets et semi-complets, le premier est plus performant (en moyenne). Il correspond à une affectation d'un attribut de chaque grappe dans chaque groupe. Cependant, les groupes semi-complets améliorent également les résultats d'une simple carte tout en limitant le temps d'apprentissage.

			Partitionnement		
Données	**Kohonen-Opt**	**ACP+Kohonen-Opt**	**Complet**	**Semi-complet**	**Hasard**
(1)	11.4	12.5	**8.4**	8.5	11.7
(2)	26.0	23.6	**19.8**	22.7	24.1
(3)	9.5	16.6	7.8	**7.6**	9.6

TAB. 2 – *Comparaison des différentes méthodes de partitionnement*

Dans un deuxième temps, nous avons comparé notre approche à d'autres méthodes ensemblistes. Les méthodes choisies sont le boosting d'arbre de décision et les forêts aléatoires

Données	**Ensemble de cartes**	ID_3	**Boosting** ID_3	**Forêts aléatoires**
(1)	8.4	10.1	8.5	**6.0**
(2)	**19.8**	33.3	37.6	21.6
(3)	7.8	23.6	8.1	**5.6**

TAB. 3 – *Comparaison de méthodes*

(Breiman, 2001), ces deux méthodes faisant référence. L'implémentation utilisée est celle de Tanagra (Rakotomalala, 2005). Les résultats sont reportés dans le tableau 3.

Les résultats de notre approche sont toujours supérieurs à ID3. Pour les jeux (2) et (3), elle se situe entre les forêts aléatoires et le boosting. Pour le troisième jeu, en revanche, les résultats sont différents. Sur ce jeu, le boosting donne des résultats plus mauvais que l'arbre de décision seul (différents paramétrages ont été testés sans fournir d'amélioration). C'est le cas lorsque le boosting sur-apprend les données, ce qui se produit généralement en présence de bruit. En outre, notre méthode donne alors de meilleurs résultats que les forêts aléatoires.

6 Conclusion

Cet article montre d'abord l'intérêt des méthodes ensemblistes pour le traitement des données en grandes dimensions. En divisant l'espace des prédicteurs pour construire plusieurs apprenants, les Ensembles contournent les problèmes théoriques liés aux espaces de grandes dimensions. Nous montrons que cette stratégie est payante : un Ensemble fondé sur les cartes auto-organisatrices améliore l'apprentissage d'une carte. Pour diviser l'espace, plusieurs stratégies sont possibles. Nous proposons une stratégie dirigée par la connaissance : un premier algorithme classifie les prédicteurs en grappe (construction de la connaissance) puis ces grappes sont utilisés pour former des groupes. Cette stratégie est payante car elle améliore les résultats donnés par un groupement aléatoire des attributs. Les forêts aléatoires, à l'inverse, utilisent une stratégie dirigée par le hasard : les arbres de décision sont construits sur des échantillons aléatoires des attributs et des individus. Les forêts aléatoires tirent parti des arbres de décision qui n'utilisent pas tous les attributs pour réaliser la prédiction, contrairement aux cartes auto-organisatrices. Enfin, cet article montre l'intérêt des cartes auto-organisatrices dans une approche ensembliste. Elles construisent une représentation sur laquelle se fonde la prédiction, représentation qui assure une certaine qualité des données. Grâce à cette représentation, les cartes sont alors plus robustes face à des données bruitées.

Associer les cartes auto-organisatrices et les Ensembles apparaît donc comme une piste de recherche prometteuse dans le cadre des données volumineuses. Ces travaux vont donc se poursuivre, d'abord en cherchant d'autres méthodes de partionnement, ensuite en essayant de mesurer la diversité créée entre les différentes cartes et enfin en étudiant la navigation à travers un Ensemble de cartes.

Références

Beckman, R. et R. Cooks (1983). Outliers. *Technometrics 25*, 119–149.

Bellmann, R. (1975). *Adaptive Control Processes : A Guided Tour*. Princeton University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Brown, G., J. Wyatt, R. Harris, et X. Yao (2005). Diversity creation methods : a survey and categorisation. *Information Fusion 6*(1), 5–20.

Demartines, P. (1994). *Analyse de données par réseaux de neurones auto-organisés*. Ph.d. dissertation, Institut National Polytechnique de Grenoble, Grenoble : France.

Duin, R. et D. Tax (2000). Experiments with classifier combining rules. In J. Kittler et F. Roli (Eds.), *Multiple Classifier Systems*, Volume 1857 of *Lecture Notes in Computer Science*, Cagliari :Italy, pp. 16–29.

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers.

Geman, S., E. Bienenstock, et R. Doursat (1992). Neural networks and the bias/variance dilemma. *Neural Computing 4*(1), 1–58.

Jacobs, R., M. Jordan, et A. Barto (1991). Task decomposition through competition in a modular connectionist architecture : The what and where vision tasks. *Cognitive Science 15*, 219–250.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*, 187–200.

Kira, K. et L. A. Rendell (1992). The feature selection problem : Traditional methods and a new algorithm. In *Tenth National Conference on artificial intelligence*, pp. 129–134.

Kohonen, T. (1982). Self-organization of topogically correct feature maps. *Biological Cybernetics 43*, 59–69.

Kohonen, T. (1988). Learning vector quantization. *Neural Network 1*, 303.

Kuncheva, L. et C. Whitaker (2001). Feature subsets for classifier combination : An enumerative experiment. In *MCS '01 : Proceedings of the Second International Workshop on Multiple Classifier Systems*, London, UK, pp. 228–237. Springer-Verlag.

Kuncheva, L. et C. Whitaker (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning 51*(2), 181–207.

Liu, Y., R. H. Weisberg, et C. N. K. Mooers (2006). Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research (Oceans) 111*.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297.

Meir, R. et G. Rätsch (2003). An introduction to boosting and leveraging. *Advanced Lectures on Machine Learning*, 118–183.

Muhlenbach, F., S. Lallich, et D. Zighed (2003). Identifying and handling mislabelled instances. *Journal of Information Intelligent Systems 22*, 89–109.

Muñoz, A. et J. Muruzábal (1998). Self-organizing maps for outlier detection. *Neurocomputing 18*(1), 33–60.

Newman, D., S. Hettich, C. Blake, et C. Merz (1998). UCI repository of machine learning databases.

Prudhomme, E. et S. Lallich (2005a). Quality measure based on Kohonen maps for supervised learning of large high dimensional data. In *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, Brest : France, pp. 246–255.

Prudhomme, E. et S. Lallich (2005b). Validation statistique des cartes de Kohonen en apprentissage supervisé. In *Actes de EGC'2005*, Volume 1 of *RNTI-E-3*, pp. 79–90.

Rakotomalala, R. (2005). Tanagra : un logiciel gratuit pour l'enseignement et la recherche. In *Actes de EGC'2005*, Volume 2 of *RNTI-E-3*, pp. 697–702.

SAS (1989). *SAS/STAT user's guide, Version 6, Fourth Edition*, Volume 2. SAS Institute Inc.

Song, X.-H. et P. K. Hopke (1996). Kohonen neural network as a pattern recognition method based on the weight interpretation. *Analytica Chimica Acta 334*, 57–66.

Tumer, K. et J. Gosh (1995). Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. Technical report, Computer and Vision Research Center, University of Texas, Austin.

Valentini, G. et F. Masulli (2002). Ensembles of learning machines. In M. Marinaro et R. Tagliaferri (Eds.), *Neural Nets WIRN Vietri-02*, LNCS. Springer-Verlag.

Verleysen, M. (2003). *Limitations and future trends in neural computation*, Chapter Learning High-Dimensional Data, pp. 141–162. IOS Press.

Verleysen, M., D. François, G. Simon, et V. Wertz (2003). On the effects of dimensionality on data analysis with neural networks. In J. A. J. Mira (Ed.), *International Work-Conference on Artificial and Natural Neural Networks : Computational Methods in Neural Modeling*, Volume II, pp. 105–112. Springer-Verlag.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association 58*(301), 236–244.

Zighed, D. A., S. Lallich, et F. Muhlenbach (2004). A statistical approach of classes separability. In H. M. e. H. T. T. Elomaa (Ed.), *Revue Applied Stochastic Models in Business and Industry*, pp. 475–487. Springer-Verlag.

Zupan, J., M. Novic, X. Li, et J. Gasteiger (1994). Classification of multicomponent analytical data of olive oils using different neural networks. *Analytica Chimica Acta 292*, 219–234.

Summary

The Knowledge Discovery process encounters difficulties with the large amount of data to treat. Indeed, some theoretical problems related to high dimensional spaces then appear and degrade the predictive capacity of the algorithms. In this article, we propose a methodology adapted to both the representation and the prediction of large datasets. For that purpose, groups of non-correlated attributes are created in order to overcome problems related to high dimensional spaces. An Ensemble is then set up to learn each group with a self-organizing map. Besides the prediction, these maps also aim at providing a relevant representation of the data. Finally, the prediction is achieved by a vote of the different maps. Experimentation performed show the revelance of this approach.