Résumé

Le stockage n'étant plus soumis à des contraintes de coût importantes, les systèmes d'information collectent une quantité croissante de données, souvent via des processus d'acquisition automatique. En parallèle, les objets d'intérêt sont devenus plus complexes. Citons, par exemple, les dossiers médicaux de patients ou les données scientifiques comme les bases de données utilisées par les historiens, composés à la fois d'images et de texte. Pour leur analyse, les méthodes d'apprentissage doivent alors prendre en compte l'augmentation massive et conjointe du nombre d'exemples et d'attributs qui en résultent. Or, les outils classiques de l'apprentissage automatique ne sont pas toujours adaptés à ce changement de volumétrie tant au niveau de leur complexité algorithmique que pour appréhender la structure des données. Dans ce contexte de données volumineuses en apprentissage supervisé, nous nous sommes intéressés à l'extraction de deux catégories de connaissances, conjointement à la prédiction, la première relative à l'organisation des exemples entre eux et la seconde relative aux interactions qui existent entre les attributs.

Pour nous intéresser aux relations entre exemples, nous définissons le concept de représentation en apprentissage supervisé comme une modélisation et une visualisation des données à la fois du point de vue de la proximité entre les exemples et du lien entre la position des exemples et leur étiquette. Parmi les différents algorithmes recensés qui conduisent à l'obtention d'une telle représentation, nous retenons les cartes auto-organisatrices qui présentent la plus faible complexité algorithmique, ce qui les rend bien adaptées aux données volumineuses. L'algorithme des cartes auto-organisatrices étant non-supervisé, nous proposons son adaptation à l'apprentissage supervisé par le biais des cartes étiquetées (Prudhomme et Lallich, 2005b). Nous montrons également qu'il est possible de valider statistiquement la qualité de la représentation obtenue par une telle carte (Prudhomme et Lallich, 2005a). Les statistiques que nous proposons, sont corrélées avec le taux d'erreur en généralisation, ce qui permet de juger *a priori* de la qualité de la prédiction qui résulte de la carte.

Néanmoins, la prédiction des cartes auto-organisatrices n'est pas toujours satisfaisante face à des données en grandes dimensions. Dans ce cas, nous avons recours aux méthodes ensemblistes. Ces méthodes agrègent la prédiction de plusieurs classifieurs simples. En créant une certaine diversité entre les prédictions de ces classifieurs, les méthodes ensemblistes améliorent la prédiction qui aurait été obtenue par un seul classifieur. Pour créer cette diversité, nous apprenons chaque classifieur de l'ensemble (dans notre cas, des cartes auto-organisatrices) sur un sous-espace de l'espace d'apprentissage. La recherche de la diversité est ainsi l'occasion de diminuer la dimensionnalité du problème. Afin de choisir au mieux les sous-espaces, nous nous sommes inspirés des connaissances théoriques disponibles sur la répartition de l'erreur en généralisation d'un ensemble. Nous avons alors proposé deux

heuristiques. La première heuristique est non-supervisée. Elle repose sur l'interprétation des corrélations entre attributs pour déterminer les sous-espaces à apprendre (Prudhomme et Lallich, 2007). La seconde heuristique, au contraire, est supervisée. Elle optimise par un algorithme génétique une mesure de l'erreur d'un ensemble en fonction de l'erreur des classifieurs qui le composent (Prudhomme et Lallich, 2008b). Ces deux heuristiques conduisent à des ensembles de cartes (ou des comités de cartes) dont l'erreur en généralisation est plus faible que celle d'une carte seule apprise sur la totalité des attributs. Néanmoins, ils conduisent également à une multitude de représentations. Pour proposer une seule représentation à l'issue de l'apprentissage, nous introduisons la notion de *stacking* géographique. A l'inverse du *stacking* classique qui apprend la prédiction de plusieurs classifieurs, le *stacking* géographique apprend la représentation de plusieurs cartes auto-organisatrices par une nouvelle carte. Cette carte synthétique constitue la représentation issue d'un apprentissage par un comité de cartes.

Le comité obtenu à partir d'une heuristique non-supervisée a l'avantage de ne pas intégrer l'information relative à l'étiquette lors de son apprentissage. En premier lieu, cette particularité lui confère une certaine robustesse au bruit, robustesse que nous avons pu observer en ajoutant un bruit uniforme sur l'étiquette à prédire (Prudhomme et Lallich, 2009). Mais cette particularité est également intéressante dans le contexte de données semi-supervisées. Dans ce contexte, seulement une petite partie des exemples sont étiquetés et la question est de savoir comment tirer profit des exemples non-étiquetés pour améliorer l'apprentissage. Cette situation est fréquente dans le cas de données volumineuses où l'acquisition automatique des données rend difficile leur étiquetage exhaustif. Nous avons donc adapté les comités de cartes à l'apprentissage semi-supervisé en remplaçant les cartes étiquetées par leur version semi-supervisée (Prudhomme et Lallich, 2008a). Les résultats que nous avons obtenus sont compétitifs avec beaucoup de stratégies d'apprentissage semi-supervisé, particulièrement lorsque le nombre d'étiquettes est restreint. Ces résultats montrent aussi l'intérêt des ensembles en apprentissage semi-supervisé.

Les données volumineuses posent également un problème lors de la validation des connaissances extraites. En apprentissage, il est constant de devoir valider les découvertes effectuées, par exemple lors de la recherche de gènes discriminants en bioinformatique, de termes discriminants en fouille de textes ou lors de la sélection de règles d'association. Lorsque l'on traite des données volumineuses en grandes dimensions, on se retrouve souvent face à une multitude de découvertes à valider. Pour y parvenir, on utilise couramment des tests statistiques qui, chacun séparément, contrôlent le risque de 1^{e} espèce. La difficulté est que le nombre de fausses découvertes augmente conjointement avec le nombre de tests.

Nous proposons une méthode non paramétrique de contrôle du risque multiple ($BSFD$) qui utilise le bootstrap pour contrôler que le nombre de fausses découvertes ne dépasse

probablement pas une borne fixée au préalable (Lallich et al., 2004). Dans un premier temps, nous avons eu recours à cette méthode pour sélectionner un sous-ensemble pertinent de rêgles d'association (Lallich et al., 2006a). Nous nous sommes ensuite replacés dans le contexte supervisé où nous avons mis en oeuvre $BSFD$ dans le cadre de données d'expression de gènes (Lallich et al., 2006b) en l'intégrant dans un processus d'extraction des connaissances. Les données génomiques présentent la double particularité d'offrir un nombre important de prédicteurs (plusieurs milliers) conjointement à un nombre faible d'individus (quelques dizaines). Dans ce contexte, la prédiction est un exercice délicat, principalement à cause des difficultés de validation. Notre approche utilise des outils statistiques adaptés pour faire face à ce problème. A savoir, l'extraction de règles de prédiction (de type *Si A et B Alors C*) dont la pertinence est d'abord mesurée statistiquement (à l'aide de la confiance) avant d'être validée par notre méthode de contrôle du risque multiple fondée sur le bootstrap. Ainsi, nous avons pu mettre en évidence – en plus d'un ensemble de règles caractéristiques de chaque classe – un petit ensemble de gènes (donc d'attributs) particulièrement spécifiques de ces classes.

En outre, d'autres travaux ont visé l'amélioration théorique de $BSFD$. L'idée proposée a été d'utiliser des nombres quasi-aléatoires pour améliorer le processus de rééchantillonnage du boostrap (Teytaud et al., 2006). Ainsi, la variabilité de notre estimateur (dans le cas de $BSFD$, la variation de la mesure due à l'échantillonnage) est réduite. Ces travaux s'adaptent directement à d'autres utilisations du bootstrap et du bagging.