Bilan et perspectives

Nous dressons ici un bilan des travaux de recherche que nous avons effectuer au cours de ce doctorat et des principales perspectives qui s'en dégagent.

Bilan

Les données volumineuses ont des caractéristiques propres qui posent problème à toutes les étapes du processus usuel de l'ECD. Après avoir dressé un bilan de ces problèmes, nous avons étudié les perspectives de l'apprentissage supervisé et semi-supervisé fondé sur la représentation, ainsi que la validation des découvertes lors de l'extraction de règles d'association.

En apprentissage supervisé, la représentation est, à notre sens, un modèle des données orienté vers la visualisation, à la fois des exemples et de leur relation de voisinage mais également du lien entre la position de ces exemples dans l'espace et leur étiquette de classe. Obtenir une telle représentation à partir de données volumineuses pose à la fois la question de sa qualité – liée aux espaces de grandes dimensions – et des temps de calcul – liés à la complexité des algorithmes. Après un état de l'art des différentes méthodes de représention, nous avons sélectionné les cartes auto-organisatrices qui répondent à l'exigence d'une complexité linéaire. Nous avons d'abord proposé une adaptation de ces cartes à l'apprentissage supervisé puis nous avons montré qu'il était possible de valider statistiquement la représentation des données qu'elles effectuent. Reste que la qualité de la prédiction des cartes souffre de la dimensionnalité des données. Les méthodes ensemblistes, d'un autre coté, sont réputées pour améliorer la qualité d'un classifieur seul en agrégeant les prédictions. Nous avons donc exploité les principes sous-jacents des méthodes ensemblistes, en particulier les notions de diversité et d'ambiguïté, pour construire des comités de cartes dont chacune réalise son apprentissage sur un sous-espace de l'espace d'apprentissage. En pratique, nous proposons deux solutions, l'une non-supervisée qui repose sur la classification des attributs et qui s'intéresse à la notion de diversité, l'autre supervisée, qui repose sur l'optimisation d'une mesure d'erreur prenant en compte l'ambiguïté des classifieurs. Cette dernière stratégie donne de très bon résultats en prédiction et montre

qu'il est possible d'obtenir un ensemble de cartes de qualité. Néanmoins, elle requiert de connaître l'étiquette de classe pour sélectionner les sous-espaces adéquats. Ce n'est pas le cas de la stratégie non-supervisée. Grâce à cette propriété, la stratégie non-supervisée résiste à des données bruitées et s'applique facilement dans un contexte semi-supervisé. C'est un avantage important face à des données volumineuses où le processus d'étiquetage par un expert trouve rapidement ses limites. Enfin, nous avons proposé une stratégie pour construire une représentation unique à partir des multiples représentations de l'ensemble, afin que l'apparente complexité émergeant de la multiplicité des cartes soit transparente à un utilisateur.

Finalement, l'apprentissage supervisé fondé sur la représentation des données, par des cartes ou des comités de cartes, offre à la fois un espace de visualisation et une structuration des exemples. Il est donc d'abord tourné vers la découverte de *connaissances orientées exemples*. Outre la prédiction d'une étiquette de classe pour un nouvel exemple, la représentation sert à visualiser la proximité entre exemples sur laquelle repose la prédiction mais aussi le voisinage d'un exemple donné, permettant ainsi la navigation à travers la base d'exemples. En apprentissage semi-supervisé, la représentation autorise la propagation des étiquettes de classes et, parce qu'elle fournit une visualisation des données, elle peut servir à choisir des exemples à étiqueter (notion d'apprentissage actif).

Dans certains cas, l'utilisateur est également intéressé par la découverte de *connaissances orientées attributs*. Les données génomiques nous semblent un bon exemple de cette problématique. Pour ces données, le biologiste ne veut pas seulement prédire, il veut également savoir quels sont les gènes et les réseaux de gènes qui mènent à cette prédiction. Pour découvrir de telles connaissances, nous nous sommes tournés vers l'extraction de règles d'association. Cette extraction suit une procédure déterministe dont le résultat est souvent un grand volume de règles difficilement exploitable par un utilisateur, particulièrement dans le cas de données en grandes dimensions où le nombre d'attributs augmente le nombre de règles possibles. Il existe des procédures qui mesurent l'intérêt des règles pour l'utilisateur afin de faire un tri parmi les règles admissibles. Nous proposons de compléter ces procédures en intégrant une étape supplémentaire de validation statistique des règles, en amont d'une mesure de leur intérêt.

La validation statistique des règles pose le problème du contrôle du risque multiple, puisque à chaque mesure statistique d'une règle est associée à un risque de 1re espèce de sélectionner la règle à tort. Pour contrôler le risque issu d'une multiplicité de tests, nous avons proposé $BSFD$, une procédure fondée sur le *bootstrap*. Cette procédure contrôle que le nombre de fausses découvertes (nombre de règles sélectionnées à tort) ne dépasse pas un seuil donné à un risque fixé, critère que nous appelons $UAFWER$. Elle a l'avantage d'être non-paramétrique, ce qui évite d'étudier la loi de la statistique contrôlée sous l'hypothèse

nulle. Par ailleurs, nous avons participé à des travaux visant à améliorer la rapidité de convergence du *bootstrap*, donc celle de $BSFD$, par l'introduction de nombres pseudo-aléatoires. Les expérimentations menées sur $BSFD$ ont montré que cette procédure fournit des résultats comparables aux procédures de type FDR. La puissance du test de $BSFD$ est cependant légèrement supérieure puisque cette procédure contrôle que le nombre de fausses découvertes ne dépasse pas un seuil donné alors que les procédures de type FDR contrôle le nombre de fausses découvertes en moyenne.

Nous avons appliqué $BSFD$ à la sélection de règles d'association à la fois dans un contexte non-supervisé et supervisé. Dans le contexte supervisé, nous avons contrôlé que les règles augmentaient bien la probabilité du conséquent. Dans le contexte supervisé, nous avons introduit la notion de règles différentiellement exprimées. Ces règles sont valides sur les exemples d'une étiquette de classe, et non sur l'autre. Nous avons alors montré qu'en sélectionnant ces règles avec $BSFD$, nous pouvions obtenir un petit ensemble de règles qui soit exploitable. En particulier, dans le cas des données génomiques, les règles sélectionnées nous ont conduit à identifier un groupe de gènes potentiellement intéressants.

Perspectives

Les travaux sur les comités de cartes montrent qu'un comité de cartes peut conduire à la fois à une faible erreur en généralisation et à une visualisation des données, tout en étant applicable au contexte de l'apprentissage semi-supervisé. Certaines recherches supplémentaires nous semblent donc intéressantes à mener sur ces comités :

- Un des premiers aspect, pour rendre les comités de cartes totalement opérationnels, concerne la prise en compte des données déséquilibrées. On sait en effet qu'en pratique, les données présentent rarement une répartition équilibrée des modalités de l'étiquette de classe. Il est fréquent de se trouver dans des situations où une modalité de l'étiquette est sous-représentée par rapport aux autres (Thomas, 2009). Au niveau des comités, nos expérimentations ont montré que la classe majoritaire était logiquement favorisée lors de la prédiction. Pour rétablir une situation normale, une première solution consisterait à traiter le déséquilibre au niveau de la prédiction de chaque carte, comme nous l'avons suggéré lors des expérimentations. Comme nous sommes face à un comité qui réalise un vote pour déterminer la prédiction finale, il nous semble également intéressant d'étudier l'effet du déséquilibre sur ce vote. Une autre piste souvent exploitée pour gérer le déséquilibre consiste à jouer sur les données d'apprentissage à l'aide de stratégies d'échantillonnage pour favoriser la classe minoritaire, soit en augmentant le nombre d'exemples de cette classe, soit en diminuant les exemples des classes majoritaires. Dans le cas des cartes auto-organisatrices, cette solution pourrait être renversée : par exemple, au lieu de multiplier le nombre d'exemples de classe minoritaire, il semble envisageable

de biaiser la construction de la carte en faveur d'une sur-représentation des zones les moins denses en exemples. En effet, les cartes sont des méthodes de quantification vectorielle. Or, ces méthodes bénéficient d'un facteur de *magnification* qui leur permet de sur-représenter (ou sous-représenter) certaines régions de l'espace (Villmann et Claussen, 2006). Contrôler ce facteur pourrait être un moyen intéressant de traiter les données déséquilibrées dans le cadre d'un apprentissage supervisé fondé sur la représentation. Il permettrait d'améliorer, en même temps que la prédiction, la visualisation des régions les moins denses.

- Nous avons insisté sur l'importance de la complexité des algorithmes. Une faible complexité permet le passage à l'échelle des algorithmes et leur adaptation à des données volumineuses. Sur ce thème, un autre point important concerne la parallélisation des algorithmes. Face à des données volumineuses et à des besoins d'exploitation en temps réel, il est intéressant de pouvoir profiter de toutes les ressources computationnelles à disposition. Les comités de cartes sont parallélisables en grande partie. L'apprentissage de chaque carte étant indépendant du reste du comité, il est possible de les apprendre parallèlement. L'optimisation du comité à l'aide d'un AG, en revanche, requiert de nouvelles recherches pour devenir parallélisable et ainsi améliorer à la fois la convergence et les temps de calcul.
- Pour obtenir un comité de cartes dont l'erreur en généralisation soit faible, nous avons proposé deux heuristiques, l'une non-supervisée et l'autre supervisée. En plus de faire intervenir l'erreur moyenne des cartes, la première repose sur le concept de diversité alors que la seconde utilise celui d'ambiguïté. Dans nos expérimentations, l'heuristique supervisée est plus efficace : bien sûr, elle utilise l'étiquette au cours de l'apprentissage. Cependant, d'autres résultats, comme ceux obtenus par Brown en corrélation négative, montrent que la notion d'ambiguïté est plus adaptée que celle de diversité quand elle est couplée à l'erreur moyenne pour expliquer l'erreur d'un ensemble. Partant de là, il nous semble que notre heuristique non-supervisée gagnerait à être fondée sur la notion d'ambiguïté à la place de celle de diversité. À l'idéal, bien évidemment, la transposition du principe de la corrélation négative à l'apprentissage non-supervisé nous semble une piste qui aurait l'avantage de l'élégance (et, sans doute, de la performance).

Nous avons insisté sur le rôle de la représentation pour un utilisateur du modèle. Cependant, nous n'avons pas développé au cours de cette thèse l'aspect utilisateur de cette stratégie d'apprentissage. Nous avons plutôt concentré nos efforts sur la recherche d'une représentation de qualité, applicable aux données volumineuses. L'autre versant des perspectives concerne donc l'aspect utilisateur d'un apprentissage fondé sur la représentation :

- L'application à l'apprentissage semi-supervisé est un des avantages d'un apprentissage supervisé fondé sur la représentation. Il nous semble que l'usage que nous faisons de la

représentation dans le contexte semi-supervisé pourrait être amélioré, notamment par la notion d'apprentissage actif. Si l'utilisateur est limité sur le nombre d'exemples qu'il peut étiqueter, il est nécessaire d'optimiser sa procédure d'étiquetage de sorte que pour un nombre d'exemples étiquetés donné, la généralisation aux exemples non-etiquetés soit la meilleure possible. La représentation, en tant que structure des exemples peut jouer un rôle dans ce processus. Cependant, même si la notion d'apprentissage actif a émergé récemment dans le domaine de l'apprentissage automatique, la quantité de travaux est déjà conséquente dans le domaine et un état de l'art est nécessaire avant d'envisager des solutions possibles.

- La représentation des données permet leur visualisation par l'utilisateur. Ici, nous considérons la représentation comme un espace de travail commun entre l'utilisateur et l'algorithme d'apprentissage. La visualisation de cette représentation constitue l'interprétation commune de ces deux acteurs du problème d'apprentissage. Sur cette base, l'utilisateur va pouvoir interagir. L'interaction d'un utilisateur avec la représentation peut avoir plusieurs buts :
  - Comme nous l'avons vu précédemment, elle peut s'intéresser à la phase d'étiquetage des données, pour choisir les exemples à étiqueter.
  - L'interaction peut concerner la phase de préparation des données, en permettant à l'utilisateur de détecter les incohérences (exemples atypiques ou mauvaise représentation au sens de l'utilisateur). Parmi les incohérences, citons également celles qui concernent la distance entre exemples, distance considérée au niveau de la représentation. En visualisant la représentation, l'utilisateur peut se rendre compte que certains exemples qu'il considère comme proches sont en fait éloignés. Mais, plus intéressant, il peut interagir pour signaler au système non seulement les proximités qui lui semblent contraires à son idée, mais apporter une correction à ce système.
  - La représentation peut également constituer une sorte d'index visuel permettant à l'utilisateur de trouver rapidement un exemple (cas de WebSOM, par exemple).