Introduction générale

L'évolution récente des bases de données détenues par les institutions privées ou publiques est profondément marquée par l'association de différents événements. Le développement des moyens de stockage et des systèmes de transmission a rendu possible la numérisation massive de l'information et sa dématérialisation, ainsi que son acquisition automatique. Les données ainsi accumulées sont devenues complexes. Elles sont volumineuses (en nombre et en dimensions), de nature hétérogène (chiffres, texte, images, son, ...), de sources variées, souvent acquises automatiquement. Le DPI (dossier patient informatisé) est un excellent exemple de ces données complexes.

Ces données sont des sources d'information et de connaissance très utiles dans un environnement qui est souvent concurrentiel. Leurs caractéristiques font qu'elles exigent des méthodes automatiques de gestion, de traitement, de synthèse et d'interprétation. L'Extraction des Connaissances à partir des Données (ECD) ou *Knowledge Discovery in Databases* (KDD), puise ses outils au sein des bases de données, de l'intelligence artificielle et de la statistique. Le moteur de l'Extraction des Connaissances à partir des Données est la Fouille des Données ou *Data Mining* qui permet de rechercher dans les données des informations pertinentes, nouvelles ou cachées, à l'aide d'algorithmes d'apprentissage automatique. On peut distinguer deux grands types d'algorithmes, suivant que l'on veut structurer les données (apprentissage non supervisé) ou prédire un phénomène (apprentissage supervisé) à partir d'autres phénomènes plus facilement accessibles qui lui sont liés (prédire la classe ou l'étiquette des exemples à partir de leurs attributs descriptifs). C'est à ce dernier type d'apprentissage que nous nous sommes intéressée dans ce travail.

Les méthodes ensemblistes ont permis d'améliorer de façon spectaculaire les performances des classifieurs standard en apprentissage supervisé, en particulier le *boosting* qui effectue une relance adaptative du classifieur de base sur un échantillon *bootstrap* des individus, tout en mettant plus de poids au fil des itérations sur les individus difficiles à prédire. Malgré un grand nombre de résultats théoriques et pratiques dans divers domaines, l'apprentissage automatique doit faire face à certaines difficultés lorsqu'il est confronté aux particularités des données contenues dans des bases de données réelles.

Parmi les difficultés les plus connues, on citera les problèmes de complexité résultant du

grand nombre de cas traités, les paradoxes liés aux grandes dimensions, la sensibilité des algorithmes aux données bruitées et enfin le traitement des données lorsque la variable de classe est déséquilibrée, ce qui est souvent le cas pour des données réelles. Le dépassement de ces problèmes constitue un véritable enjeu pour améliorer l'efficacité du processus d'apprentissage face à des données réelles. Nous avons choisi dans cette thèse de réfléchir à des procédures adaptatives du type *boosting* qui soient efficaces en présence de bruit ou en présence de données déséquilibrées.

Notre premier chapitre est consacré au problème des données bruitées qui est devenu un problème majeur en apprentissage automatique. Nous nous sommes particulièrement intéressée au contrôle du bruit lorsque l'on utilise des procédures de *boosting* [Schapire and Singer, 1999]. En effet, les procédures de *boosting* ont beaucoup contribué à améliorer l'efficacité des procédures de prédiction en *data mining*, sauf en présence de données bruitées [Dietterich, 1999]. Dans ce dernier cas, le *boosting* a tendance à se spécialiser sur les exemples bruités, ce qui diminue ses performances en généralisation, et il voit sa vitesse de convergence diminuer. Pour venir à bout de ce double problème de surapprentissage et de vitesse de convergence, nous proposons AdaBoost Hybride [Bahri et al., 2009], une adaptation de l'algorithme standard Adaboost M1 [Freund and Schapire, 1996] qui est fondée sur un double lissage des résultats des hypothèses antérieures du *boosting*. Les résultats expérimentaux obtenus sur différents *benchmarks* de l'UCI, ainsi que sur le jeu de données réelles KDD-Cup 99 traité au chapitre 3 illustrent le bien-fondé de notre approche.

Dans le chapitre II, nous nous sommes intéressée à un autre problème posé par les données réelles, celui de la prédiction lorsque la distribution de la variable de classe est très déséquilibrée. Les méthodes à base de règles, par exemple C4.5 sont particulièrement pénalisées par une telle situation. Dans ce cas, les versions boostées de ces méthodes, telles que C4.5 ou les forêts aléatoires, ne suffisent pas à assurer de bonnes performances et en outre elles perdent l'intelligibilité des résultats qui font l'intérêt des méthodes à base de règles. Nous nous sommes fixée de répondre au problème du déséquilibre des classes en élaborant une méthode adaptative qui repose sur un classifieur faible à base de règles qui ne soit pas trop sensible au déséquilibre des classes.

Nous avons choisi, comme classifieur faible, la classification associative qui opère la prédiction à partir de règles d'association de classe et peut ainsi se focaliser sur des petits groupes de cas. Dans un premier temps, nous avons mis au point FCP-Growth-P [Bahri and Lallich, 2009b], une méthode de génération des règles d'association de classe qui soit plus efficace que les méthodes standard telles que Apriori ou FP-Growth en termes d'espace mémoire et de temps d'exécution. Nous avons ensuite élaboré le classifieur W-CARP [Bahri and Lallich, 2009a] qui construit une base de règles significatives ayant une confiance au moins égale à 50% à partir des règles fournies par FCP-Growth-P. Pour prédire la classe d'un nouvel exemple, W-CARP retient les règles de la base qui couvrent cet exemple

et construit le score de chaque classe en pondérant les prédictions des différentes règles par la valeur de leur mesure de Loevinger, qui a l'avantage de tenir compte de la fréquence du conséquent. Les résultats obtenus montrent que notre approche est plus économe en temps et en espace mémoire que les approches traditionnelles, ce qui permet de baisser le seuil de support, tout en assurant une précision au moins aussi bonne.

Nous proposons alors CARBoost [Bahri and Lallich, 2010], une méthode de classification associative adaptative qui utilise W-CARP comme classifieur faible. A chaque itération, CARBoost renforce le poids des exemples mal prédits lors de l'itération précédente et augmente le poids des règles qui ont bien prédit au moins un exemple mal prédit. CARBoost permet ainsi d'optimiser le système de poids appliqué aux différentes règles lors du calcul du score des différentes classes. Pour chaque exemple à prédire, CARBoost fournit la liste de règles significatives couvrant l'exemple, règles dont les poids ont été optimisés par la procédure itérative, ce qui permet de conserver l'intelligibilité des méthodes à base de règles.

Pour tester sur des données réelles, en vraie grandeur, le caractère opérationnel et la validité de nos propositions, ce qui est l'objet du chapitre III, nous avons choisi le domaine de la détection d'intrusions, un domaine d'actualité où le besoin de méthodes efficaces va croissant, face aux menaces intelligentes des attaquants (*traders*). Les données choisies sont les données KDD-Cup 1999. La difficulté de ces données réelles est que justement elles sont bruitées et très déséquilibrées. Dans un tel cas, les méthodes standard, en l'occurrence C4.5 et Adaboost M1, sont incapables d'étiqueter une connexion par l'une ou l'autre des classes les plus déséquilibrées. Les résultats obtenus attestent de la bonne performance des méthodes proposées, AdaBoost Hybride et surtout CARBoost, qui sont capables de retrouver avec une bonne précision la quasi-totalité des connexions relevant des classes les plus déséquilibrées, sans détériorer la prédiction des classes plus nombreuses.