Apprentissage incrémental de règles de décision

Bassirou Seye, Clément Fournier,
Léandre Le Polles -- Potin, Pierre Testart

Encadreur : Laurence Rozé

Résumé

Notre projet a eu pour but de nous faire implémenter l'algorithme VFDR, en nous servant de l'API de Weka. Cet algorithme est décrit dans le document de recherche intitulé "Learning decision Rules from Data Streams" de João Gama et Petr Kosina.

1 Introduction

1.1 Notions d'apprentissage artificiel

Pour introduire les notions principales nécessaires à la compréhension du travail accompli, nous allons prendre un exemple qui nous servira de fil directeur : l'exemple classique de la différentiation entre les oies et les cygnes. La question posée est la suivante : "Comment un programme peut-il différencier ces deux oiseaux ?".

1.1.1 Règles de décisions

La classification est une tâche qui consiste à attribuer une étiquette à un individu donné en entrée. Notre exemple des oiseaux est un exemple de classification : à partir de données sur un oiseau, on doit déterminer si c'est une oie ou un cygne. Dans les algorithmes qui apprennent des règles, on base la décision d'attribuer une étiquette particulière à l'individu sur la validation de certaines "règles," c'est à dire la réalisation simultanée d'une ou plusieurs conditions sur les données qui décrivent l'individu à classifier.

Les données qui décrivent chaque individu sont les valeurs associées à des caractéristiques mesurables de l'individu, qu'on appelle des *attributs*. Par exemple, pour nos oiseaux, la taille en centimètre est un attribut numérique (sa valeur est un nombre) et la couleur du plumage est un attribut nominal (il prend ses valeurs dans un ensemble fini, par exemple "clair", "moyen" ou "sombre"). Les attributs nominaux et numériques sont les deux types de données que notre algorithme supporte.

Les règles qui nous permettent de prendre une décision de classification se présentent sous la forme de conjonctions de conditions sur les attributs de l'individu à classifier. On appelle ces conditions des *littéraux* (ou *antécédents*). Sur des attributs numériques, on compare la valeur avec des valeurs seuils, ainsi un antécédent numérique pourrait être `taille > 50cm`. Les valeurs des attributs nominaux ne sont pas ordonnées, les antécédents nominaux sont donc de la forme `plumage = clair`.

Une règle possible pour classifier les oiseaux est « Si plumage = "clair" et taille $>$ 80, alors classe = "cygne". » On voit donc que si les conditions de la règle sont réunies, on prend une décision de classification. Le but d'un algorithme de classification va être de consituer un ensemble de règles qui décrive au mieux les individus déjà observés pour avoir une estimation de la classe d'un nouvel individu à classifier. C'est le principe simplifié de l'apprentissage de règles.

1.1.2 Apprentissage

L'apprentissage de règles consiste à laisser l'algorithme définir seul les règles de décisions en lui fournissant un ensemble d'apprentissage, constitué d'individus dont la classe est déjà connue. En fournissant un jeu de données d'individus déjà étiquetés à l'algorithme, celui-ci pourrait par exemple répartir les individus sur un diagramme tel que celui-ci :

Et en déduire l'ensemble de règles suivant : « Si plumage = "clair" et taille $>$ 80, alors classe = "cygne" Si plumage = "foncé" et taille $>$ 97, alors classe = "cygne" Si plumage = "noir", alors classe = "oie" Sinon classe = "oie" »

1.1.3 Incrémentalité

Les algorithmes d'apprentissage incrémentaux sont capables de mettre à jour les règles de décisions à chaque individu catégorisé qu'on lui fournit, tout en laissant la possibilité à tout moment d'utiliser lesdites règles de décisions pour classifier un individu dont on ne connaît pas encore la classe.

Ces algorithmes sont conçus pour optimiser leur utilisation d'espace mémoire et le temps qu'ils mettent à classifier une instance, et sont donc adaptés à l'apprentissage sur des flux de données importants et continus.

1.2 Weka

Weka (Waikato Environment for Knowledge Analysis) est une plate-forme d'apprentissage artificiel, programmée en Java, permettant de réaliser de nombreuses tâches d'apprentissage et de classification. Elle rend accessible les différentes techniques de Data Mining et de Machine Learning et permet d'appliquer rapidement ces techniques sur des problèmes concrets.

1.3 Présentation de VFDR

L'algorithme VFDR, sigle de "Very Fast Decision Rules," est l'algorithme d'apprentissage incrémental de règles qui nous intéressera ici. Nous allons dans ce chapitre décrire le fonctionnement de cet algorithme.

1.3.1 Principe

L'algorithme commence avec un ensemble vide de règles et une règle par défaut. Cette règle, ne comprenant aucun littéral, pourrait être exprimé par “Sinon ...” comme dans l'exemple d'ensemble de règle ci-dessus. A chaque règle est associée une structure de donnée permettant de calculer les statistiques nécessaire pour le traitement de ladite règle.

Si tous les littéraux d'une règle sont vrai pour un individu donné, on dit qu'il est *couvert* par cette règle. Lorsque l'algorithme reçoit un individu étiqueté, il tente de le faire correspondre à chaque règle de l'ensemble de règles créés ou à la règle par

défaut. Si la règle qui le couvre contient suffisamment d'individus, elle est étendue pour améliorer sa précision. L'extension d'une règle consiste à lui rajouter un littéral. Lorsque c'est la règle vide qui devrait être étendue, une nouvelle règle est créée à la place. Ainsi, l'algorithme crée l'ensemble de règle qui servira à classifier les individus non étiquetés.

1.3.2 Description des statistiques

Pour ne pas saturer la mémoire, l'algorithme n'enregistre pas l'ensemble des individus traités, mais utilise une structure de donnée permettant de calculer les statistiques pertinentes liée à chaque règle. Cette structure de données est composée de :

- Un entier, correspondant au nombre d'individus couverts par la règle.
- Un vecteur d'entier, comprenant le nombre d'occurence de chaque classe parmis les individus couverts.
- Une matrice, représentant le nombre d'occurrence de chaque valeur des attributs nominaux pour chaque classe.
- Un estimateur gaussien, permettant de calculer pour chaque attribut numérique la probabilité de rencontrer une valeur supérieur à telle ou telle valeur déjà rencontrée, par classe.

1.3.3 Expansion d'une règle

1.3.4 Concept de règle non-décisionnelle

2 Second chapter

Travail réalisé : Organisation Comment fonctionne le programme ? Limites de l'implémentation

Comparer VFDR avec d'autres algorithmes comme VFDT Illustrer la comparaison avec un exemple : SPAM Illustrer le fait que VFDR soit incrémental

Limites de l'implémentation : nombre d'attributs ? Taille attributs ? Nombre

3 Third chapter

4 Conclusion

Références

[1] João Gama et Petr Kosina : Learning decision rules from data streams. *In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Two*, IJCAI'11, pages 1255–1260. AAAI Press, 2011.