Approche relationnelle de l'apprentissage de séquences

Clément Charnay*, Nicolas Lachiche*, Agnès Braud*

*ICube, Université de Strasbourg, CNRS
300 Bd Sébastien Brant - CS 10413
F-67412 Illkirch Cedex
{charnay,nicolas.lachiche,agnes.braud}@unistra.fr

Des flux de données aux sources de données ouvertes donnant accès à des informations en temps réel, de plus en plus de données peuvent être vues comme des séquences ordonnées. Les besoins en termes d'apprentissage automatique sur ces données séquentielles deviennent donc importants, comme pour des tâches de prévision du futur de la séquence. Dans ce contexte, nous nous intéressons à l'apprentissage supervisé hors ligne de la séquence : étant donnés des exemples ordonnés, nous voulons construire un modèle qui utilise dans ses hypothèses des propriétés des exemples passés.

Pour ce faire, nous introduisons une représentation relationnelle des données séquentielles, où chaque exemple, représenté dans une table, est associé avec tous ses prédécesseurs, relation représentée dans une autre table. Ces approches évitent une représentation attribut-valeur où à chaque exemple sont associés sur la même ligne les exemples précédents dans la limite d'une fenêtre choisie au préalable, approche plus lourde et ne permettant pas d'apprécier les tendances de la séquence.

Attribut-valeur	Relationnel
donnees(x_n, $\mathbf{y_n}$, x_{n-1}, y_{n-1}, x_{n-2}, y_{n-2}, …, x_{n-l}, y_{n-l})	donnees($\underline{id}$, x, $\mathbf{y}$) association($\underline{idPrincipal}$, $\underline{idAnterieur}$, $dist$)

Dans une représentation attribut-valeur, une seule table est utilisée pour représenter les données. Pour un exemple donné, le vecteur de variables prédictrices x_n associées à l'exemple ainsi que la valeur à prédire y_n sont stockées. De plus, on associe à l'exemple les valeurs des variables prédictrices et de la cible pour les exemples antérieurs à l'exemple courant, dans une limite de l exemples. Ainsi x_{n-1} et y_{n-1} représentent les informations associées à l'exemple précédant directement l'exemple courant. Plus généralement, x_{n-k} et y_{n-k} représentent les informations associées au k-ème exemple précédant l'exemple courant. C'est une approche par fenêtrage où $l + 1$ désigne la largeur de la fenêtre.

Cette approche présente un inconvénient : la taille de la fenêtre doit être définie à l'avance. Nous proposons de dépasser cette contrainte grâce à une autre représentation, relationnelle, des données séquentielles. Dans cette représentation, une table est utilisée pour représenter les attributs des exemples courants. Une deuxième table est utilisée comme table d'association pour mettre en correspondance deux lignes de la table des données.

Les colonnes *idPrincipal* et *idAnterieur* de la table *association* référencent toutes deux la colonne *id* de la tabledonnees. Ainsi, une ligne de la table *association* indique que l'exemple