Data Mining

Arbre de décisions

Arbres de décision et données binaires

Appliquez l'algorithme ID3 présenté en cours sur le tableau de données suivant.

Début de semaine	En groupe	Sexe Masculin	Amateur de théâtre	Apprécié
1	1	1	0	0
1	0	0	0	0
1	0	0	1	1
0	1	1	1	1
0	0	1	0	1
1	1	0	1	1
0	0	0	0	0

Arbre de décision et données nominales

Appliquez l'algorithme ID3 présenté en cours sur le tableau de données suivant.

Taille	Petit	Grand	Grand	Grand	Grand	Petit	Petit	Grand
Nationalité	Italien	Italien	Italien	Français	Allemand	Allemand	Allemand	Allemand
Situation F.	Cel.	Cel.	Marié	Cel.	Cel.	Cel.	Marié	Marié
Prend un crédit	Non	Non	Non	Oui	Oui	Oui	Non	Non

Arbre de décision et données mixtes

Appliquez l'algorithme ID3 présenté en cours sur les données suivantes. Pour cela, vous devez trouver une méthode pour vous ramener au problème précédent avec des données purement nominales. Les données sont fournies dans le formalisme du logiciel WEKA vu en TP, c'est-à-dire au format .arff.

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

***Mining* : le logiciel WEKA**

Prise en main et Etude de fleurs.

Nous allons analyser des données concernant de la botanique.
Lancer le logiciel WEKA : java –jar weka.jar
Ce logiciel est développé en Java, et fournit le code source et la documentation complète des classes. On peut ainsi réutiliser les différents outils dans une application personnelle en important les classes nécessaires. Sachant que le langage JAVA propose des paquetages spécifiques à la gestion de bases de données, c'est un environnement idéal pour combiner les connaissances acquises en base de données, fouille de données et en programmation objet dans le cadre d'un stage ou d'un projet par exemple.

Nous allons utiliser l'environnement Explorer.

L'interface est composée de plusieurs panneaux. Dans la suite, je fais confiance à votre perspicacité pour répondre à certaines questions grâce à l'interface proposée.

Le panneau Preprocess

Cette partie permet de retoucher les données pour les adapter à un format adéquate à la méthode de data mining que l'on sélectionnera plus tard.
Chargez le fichier de données iris.arff.
Combien y a-t-il d'attributs ? Quel est leur nom ?
Quelles sont les statistiques de répartition pour la largeur du pétale de chacune des fleurs analysés ?
De quel type sont chacune des données : binaires, nominal, numérique ?
Combien y a-t-il de données ?

On peut retoucher les données à ce niveau là. Par exemple, décider de ne pas utiliser un des attributs. Si l'on décoche petallength par exemple, pour rendre effectif le changement, il faut cliquer sur Apply Filter. Observez les modifications dans la zone Working Relation.

Revenez à l'usage de tous les attributs.

Quelles sont les trois classes d'iris répertoriées par les botanistes en fonction des 5 attributs métriques stockés ?

Nous reviendrons à ce panneau pour appliquer d'autres filtres avec d'autres jeux de données.

Le panneau Visualisation

Ce panneau permet de visualiser les données par projection 2D.

En modifiant les attributs projetés sur l'axe des X et l'axe des Y, déterminez un couple d'attributs visuellement optimal pour séparer les données dans les trois classes précisées par les botanistes.

A droite de la zone de projection, se trouve 6 bandes. Elles correspondent aux 6 attributs stockés pour chaque fleur dont le nom de l'espèce ou classe. En cliquant avec les boutons de la souris sur ces bandes, on modifie les sélections en X et Y pour la visualisation en 2D.

Le panneau Select Attributes

Ce panneau permet de faire de la sélection d'attributs pertinents dans un objectif de classification. C'est ici que vous pouvez trouver un moyen automatique de déterminer le meilleur couple d'attributs pour séparer les classes.

En laissant les modalités définies par défaut dans ce panneau, lancez la sélection d'attributs en cliquant sur Start. Quelle est la réponse fournie par l'algorithme ?
Est-ce en accord avec votre réponse ?
Visualiser la solution proposée pour vous en persuader.

Le panneau Classify

Pour prendre contact, nous allons réaliser une classification basique. L'algorithme ZeroR détermine la classe la plus présente dans l'ensemble d'apprentissage et génère une règle unique du type
: classe 1
Qui se lit, si n'importe quoi pour les attributs 1à 5 alors conclure que la fleur est de la classe 1.
Sélectionnez cet algorithme en cliquant sur la liste déroulante Classifier. Aucun paramètre n'est à préciser.
Lancer l'algorithme en cliquant Start.
Dans la zone Classifier Output, les résultats sont affichés. Après avoir rappelé quelques informations sur le nom de la relation (au sens des bases de données relationnelles) et sa structure, puis sur le type d'algorithme d'analyse utilisé, la partie Summary détaille les résultats en classification dont nous présentons l'essentiel dans ce qui suit.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          50               33.3333 %
Incorrectly Classified Instances       100               66.6667 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  Class
                 1         1          0.333      1      Iris-setosa
                 0         0          0          0      Iris-versicolor
                 0         0          0          0      Iris-virginica
```

```
=== Confusion Matrix ===

a   b  c   <-- classified as
50  0  0 |  a = Iris-setosa
50  0  0 |  b = Iris-versicolor
50  0  0 |  c = Iris-virginica
```

D'abord, il y a la façon dont on valide le résultat de l'algorithme : le plus souvent on utilise la technique de cross-validation. C'est-à-dire qu'on partage l'ensemble d'apprentissage en 10 ensembles par exemple. Puis, on entraîne avec 9 d'entre eux en testant sur celui qu'on a laissé de côté. On répète l'opération en permutant l'ensemble de test et on fait une moyenne. C'est ce qui assure l'un des meilleurs conditionnement d'un point de vue statistique lorsqu'on a peu d'exemples. La stratification permet d'assurer un équilibre supplémentaire entre les différentes classes dans chacun des sous-ensembles. Une autre possibilité si on a suffisamment d'exemples est de partager avec un certain pourcentage (66 % ici par défaut) l'ensemble d'apprentissage en conservant 34 % pour l'ensemble de test.

Ensuite, on présente des chiffres de bases : taux de classification correcte global par exemple. Ici 33%, ce qui est peu mais correspond au choix basique de l'algorithme. Et constitue une base de comparaison pour les autres algorithmes.

Puis, on présente des chiffres plus spécifiques. TP Rate pour taux de Vrais Positifs (True Positive) où taux de reconnaissance. FP Rate pour le taux de Faux Positifs (False Positive) ou taux de Fausses Alarmes.
La précision définie comme TP/(TP+FP) et le Recall définie comme TP/(TP+FN) où FN est le taux de Faux Négatifs.

A quoi correspondent les Faux Négatifs ?

Souvent, on présente une courbe Recall contre Précision car ces deux valeurs évoluent en fonction de certains paramètres des algorithmes. Selon vous, quelle est la meilleure région dans ce genre de courbe ?
Interprétez les chiffres proposés pour l'algorithme utilisé !

Enfin, on présente une matrice dite Matrice de Confusion. Elle se lit ainsi : 50 éléments ont été classés Iris-setosa et sont des Iris-setosa ; 50 autres ont été classés Iris setosa alors qu'ils sont des Iris-versicolor et de même pour les Iris-virginica.

Analysez les résultats obtenus si on utilise l'algorithme IB1 qui correspond à un algorithme du type 1-ppv ou 1-NN (1 plus proche voisin ou 1 Nearest Neighbour). Répétez avec un 2-NN.

Analysez les résultats obtenus si on utilise l'algorithme J48 qui est une version améliorée des arbres de décision type ID3. Représentez graphiquement sur papier l'arbre proposé et de fait écrivez les règles SI…ALORS qui en résultent.

Répétez avec l'algorithme J48-PART qui fournit directement une liste de règles.

Répétez avec un classifieur Bayésien Naïf.

Le panneau Cluster

Maintenant, vous allez pouvoir tester par vous-même dans ce panneau l'algorithme des K-Means. Une fois lancé l'algorithme et les résultats obtenus, dans la zone Result List cliquez avec le bouton droit de la souris sur l'algorithme courant et demandez de visualiser le regroupement. En utilisant le bon couple d'attributs, vous verrez les croix correspondant aux éléments correctement classifiés et les rectangles aux éléments incorrectement classifiés.

Envol et Etude du temps.

Certains algorithmes ne fonctionnent qu'avec des données nominales.
Ouvrez le fichier weather.nominal.arff. Etudiez-le.

Appliquez successivement les algorithmes ID3 (dans le panneau Classify) et APRIORI (dans le panneau Associate) et analysez les résultats.

Explorez par vous-même d'autres solutions proposées (5 minutes). A votre avis, quelle est la meilleure stratégie pour interpréter des données ?

Ouvrez à présent le fichier weather.arff qui contient un mélange de données numériques et nominales. Essayez les algorithmes précédents.
La solution est d'appliquer des filtres pour transformer les données. Par exemple, pour passer de variables numériques à nominales, on utilise le filtre DiscretizeFilter. Il faut l'ajouter à la liste des filtres puis Apply Filter. Enfin, Replace les données pour agir sur les données modifiées. Observez attentivement les retours dans les zones d'informations.
Essayez à nouveau les algorithmes précédents. Puis supprimer le filtre de la liste d'attente.

Critique et Etude d'un spectacle

Récupérez le fichier resa.txt sur le site habituel.
Observez sa structure.
WEKA peut interpréter un tel fichier. Chargez-le.
Il vous est possible de le Sauver au format arff si vous le désirez, ce qui peut être une bonne idée. Attendez juste que le bouton soit accessible.

Commencez à appliquer certains algorithmes précédents pour analyser les résultats. Mais soyez malin… conditionnez bien le problème au départ : choix des attributs intéressants, quelle est la variable à expliquer …..

Pour aider certains, voici quelques étapes pour s'y retrouver :

Décochez la variable Nom et Apply Filter. Observez Working Relation. Sauver le fichier de données ainsi modifié si c'est possible.
Replace.
Combien y a-t-il d'hommes ?

Appliquez l'algo ZeroR. Demandez à ce que la variable expliquée soit Appreciation. Expliquez les résultats.
Visualisez les données. Positionnez Date en abscisse et Appreciation en Ordonnée. Quelle est l'appréciation globale du spectacle ? Même chose avec la variable Nombre de personnes ? Sexe ?

Sélectionnez les Hommes n'ayant pas du tout apprécier le spectacle (grâce à la zone Select Instance en choisissant Rectangle puis Submit).
A l'aide du Jitter (ajout d'une perturbation à chacun des points), dénombrez-les visuellement. Remettre le Jitter à 0.
En modifiant les axes X et Y, quel est le jour le moins favorable à ce spectacle (et en fait à tout spectacle en général) ? Aux garçons ?
Reset.

Enfin, on aimerait que tout ça soit plus synthétique et rassembler les appréciations A et B ensemble et les C et D ensemble. Qu'à cela ne tienne ! On va appliquer le filtre MergeTwoValuesFilter avec les paramètres 1, 2,0 puis 2,2,1. Attention les manips sont plus longues que ce que j'écris mais vous l'avez déjà fait. Normalement, l'attribut 2 (Appreciation) ne possède plus à la fin de la manip que deux valeurs : A-B avec 102 éléments et C-D avec 31 éléments. Continuez l'analyse précédente notamment en visualisant en X la Date et en Y le Nombre. Quel est le jour le plus mauvais ?

Enfin, transformez les variables numériques en nominal pour pouvoir appliquer les algos ID3 et APRIORI.

Bravo !

Pour les fanatiques : récupérer la classe Java sur le site. Et essayez de compiler dans l'environnement J-builder en n'oubliant pas d'indiquer le chemin de la librairie weka.jar.