TP2 Fouille de Données
Préparation des données - Associations

29 janvier 2017
Lina Soualmia

1 Préambule

Les réponses aux questions sont à envoyer à fdd.m1.rouen@gmail.com avec comme objet : [FdD] TP2 GR(2, 3 ou 4) NOM sous la forme d'un fichier (.pdf, .txt, ...). N'oubliez pas d'indiquer le numéro de la question et répondez-y de manière claire.

2 Préparation des données

Environnement de travail :
Système WEKA
Fichier `weather.numeric.arff` du TP1.

2.1 Question 1

Weka fournit des méthodes, appelées *filtres* (package `weka.filters`) qui permettent de transformer les données pour une analyse plus efficace. A partir de l'onglet `Preprocess`, si un jeu de données a été chargé, la boîte `Filter` permet de choisir un algorithme de filtrage (bouton `Choose`).
Les filtres permettent de supprimer ou d'ajouter un attribut, de discrétiser un attribut, de remplacer des valeurs manquantes, de supprimer des exemples, de ré-échantillonner les données . . .etc.
Le bouton `Choose` ouvre une fenêtre qui permet de :

- Choisir un filtre non supervisé agissant sur les attributs ou sur les données,
- Définir un certain nombre de paramètres pour ce filtre.

Une fois le filtre choisi et paramétré, on doit l'appliquer au jeu de données (bouton `Apply`) ; la base courante est remplacée par le jeu de données modifié. On peut sauvegarder le jeu de données modifié (bouton `Save`) dans un fichier `.arff`.
Parmi les filtres sur les attributs :

- `Add` : permet d'ajouter un attribut nominal.
- `AddCluster` : permet d'ajouter un attribut nominal qui représente l'attribut de classe après une classification.
- `Discretize` : permet de discrétiser des attributs numériques continus.
- `MakeIndicator` : permet de construire un nouveau jeu de données avec un attribut booléen remplaçant un attribut nominal.
- `Normalize` : permet de normaliser tous les attributs numériques dans l'intervalle [0 ;1].
- `NumericToBinary` : permet de transformer un attribut numérique en binaire.

— `PKIDiscretize` : permet de discrétiser des attributs numériques continus.
— `Remove` : permet de supprimer des attributs.
— `RemoveType` : permet de supprimer tous les attributs d'un type donné (nominal, numérique, ..).
— `RemoveUseless` : permet de supprimer les attributs dont les valeurs varient peu.
— `ReplaceMissingValues` : permet de remplacer les valeurs manquantes des attributs nominaux ou numériques par leur mode ou leur moyenne.

Afin de simplifier l'interprétation des résultats, vous allez discrétiser l'attribut `temperature`. Cela aura pour effet de remplacer les valeurs numériques dans les instances par le nom de l'intervalle correspondant.
Cliquez sur le bouton `Choose` et sélectionnez le filtre `Discretize` dans la catégorie `Filter/Unsupervised/Attribute`. Cliquez sur le nom de l'algorithme afin de faire apparaître la fenêtre de choix des paramètres. Indiquez le numéro de l'attribut `temperature` dans la zone `attributeIndices` et laissez les autres paramètres à leur valeur par défaut. Cliquez sur le bouton `Apply` pour exécuter la discrétisation et sélectionnez l'attribut `temperature` en cliquant dessus. La description de l'attribut a changé et sont maintenant affichés la liste des intervalles générés et le nombre d'occurrences de chacun. Cette discrétisation est faite en largeur et les intervalles sont de même taille. Quels sont les intervalles pour lesquels il n'existe aucune instance dans le jeu de données ?

2.2 Question 2

Annulez la discrétisation précédente en cliquant sur le bouton `Undo`. Cliquez sur le nom de l'algorithme afin de faire apparaître la fenêtre de choix des paramètres et fixez le nombre d'intervalles à 4 dans la zone `bins`. Ne modifiez pas les autres paramètres. Quels sont les différents intervalles générés et leurs effectifs ?

2.3 Question 3

Annulez la précédente discrétisation en cliquant sur le bouton `Undo`. Cliquez sur le nom de l'algorithme afin de faire apparaître le fenêtre de choix des paramètres et choisissez l'ajustement de la taille des intervalles par fréquences égales en mettant à `true` l'option `useEqualFrequency`. Laissez tous les autres paramètres booléens à `false`. Quels sont les différents intervalles générés et leurs effectifs ?

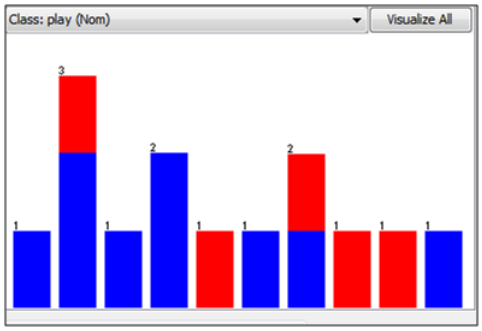
2.4 Question 4

Annulez la précédente discrétisation en cliquant sur le bouton `Undo`. Cliquez sur le bouton `Choose` et sélectionnez le filtre `PKIDiscretize` dans dans la catégorie `Filter/Unsupervised/Attribute`. Cliquez sur le nom de l'algorithme afin de faire apparaître la fenêtre de choix des paramètres et indiquez le numéro de l'attribut `temperature` dans la zone `attributeIndex`. Laissez les autres paramètres à leur valeur par défaut. Exécutez la discrétisation et indiquez les intervalles générés et leurs effectifs.
Enregistrez le jeu de données ainsi modifié dans un fichier `weather.nominal.arff`. Ouvrez ce fichier et remplacez les noms des intervalles par `cool`, `medium` et `hot` respectivement. Sauvegardez le fichier et chargez-le dans Weka.

2.5 Question 5

Cliquez sur le bouton `Choose` et sélectionnez le filtre `NumericToNominal` dans la catégorie `Filter/Unsupervised/Attribute`.

Cliquez sur le nom de l'algorithme afin de faire apparaître la fenêtre de choix des paramètres et indiquez le numéro de l'attribut `humidity` dans la zone `attributeIndices`. Laissez les autres paramètres à leur valeur par défaut.
Cliquez sur le bouton `Apply` pour exécuter la transformation et sélectionnez l'attribut `humidity` en cliquant dessus. La description de l'attribut a changé. Sont maintenant affichés la liste des valeurs nominales générées (le numéro de la valeur et son nom dans le jeu de données) et le nombre d'occurrences de chacune.
Dans le cadre en bas à droite, sélectionnez l'attribut `play` dans la zone `Class`. On peut observer que les 4 premiers histogrammes du graphique représentent en majorité des instances `play=yes`.



Quelles sont les valeurs numériques (`Label`) correspondant à ces 4 premiers histogrammes ?

2.6 Question 6

Afin de regrouper les 4 premières colonnes ensemble et les suivantes ensemble, cliquez sur le bouton `Choose` et sélectionnez le filtre `MakeIndicator` dans la catégorie `Filter/Unsupervised/Attribute`.
Cliquez sur le nom de l'algorithme afin de faire apparaître la fenêtre de choix des paramètres et :

- Indiquez le numéro de l'attribut `humidity` dans la zone `attributeIndices`
- Mettez à `false` l'option `numeric`
- Indiquez dans la zone `valueIndices` les numéros des valeurs qui correspondent aux 6 derniers histogrammes c-à-d les numéros `5-10`.

Cliquez sur le bouton `Apply` pour exécuter la transformation et sélectionnez l'attribut `humidity` en cliquant dessus.
La description de l'attribut a changé. Celui-ci ne possède plus que 2 valeurs nominales correspondant à des taux d'humidité faible ou moyen, et fort respectivement. Quel est le nom de ces valeurs nominales et le nombre d'instances pour chacune d'elles ?
Enregistrez le jeu de données ainsi modifié dans un fichier nommé `weather.discretized.arff`.

2.7 Question 7

Ouvrez le fichier `weather.discretized.arff` et remplacez les noms des intervalles par `low-medium` et `high` respectivement. Sauvegardez-le et chargez-le dans Weka.
Cliquez maintenant sur `Visualize all` afin d'afficher les histogrammes de répartition des instances `play=yes` et `play=no` pour chaque attribut.
Observez les histogrammes pour `temperature` et `humidity` et indiquez pour chacun si sa valeur doit être plutôt faible ou plutôt élevée pour que l'on puisse jouer.

2.8 Question 8

Cliquez sur l'onglet `Visualize` et double-cliquez sur le graphique correspondant aux attributs `temperature` et `humidity` afin d'afficher la répartition bi-dimensionnelle des valeurs. Quelles sont les combinaisons de valeurs de `temperaure` et `humidity` pour lesquelles toutes les instances correspondent à `play=yes` ?

2.9 Question 9

En fonction des connaissances acquises lors de cette analyse exploratoire, indiquez pour chaque attribut `outlook`, `temperature`, `humidity` et `windy` la valeur qui vous semble la plus représentative des classes `play=yes` et `play=no` ?

3 Règles d'association

L'extraction de règles d'association est accessible par l'onglet `Associate`. Les algorithmes implantés sont `Apriori, HotSpot, predictiveApriori` et `Tertius`.

3.1 Question 10

Chargez le jeu de données `weather.numeric.arff` dans Weka. Exécutez l'algorithme `Apriori` et avec ses paramètres par défaut. Que constatez-vous ? Quelle(s) conclusion(s) en tirez vous ?
Rappel : sur l'onglet `Preprocess`, le cadre `Filter` permet de définir des filtres de transformation des données. Ces filtres sont des opérations (suppression, normalisation, discrétisation, ...etc.) sur les données. Ils sont classés en deux catégories : filtres sur les attributs et filtres sur les instances. Les fonctions sont :

- `Apply` : appliquer le filtre.
- `Undo` : revenir en arrière (annuler les effets du dernier filtre).
- `Save` : sauvegarder le jeu filtré dans un fichier.

Allez sur l'onglet `Preprocess` et définissez un filtre `PKIDiscretize` pour discrétiser les attributs `temperature` et `humidity`. Ce filtre transforme les attributs numériques en attributs nominaux. Pour chaque attribut créé, les modalités correspondent à des intervalles de valeurs de même fréquence (même nombre d'instances pour chaque modalité). Visualisez les valeur de `temperature` qui doivent être discrétisées selon les modalités suivantes :

- `'(inf-70.5]'` : valeurs inférieures ou égales à 70,5.
- `'(70.5-77.5]'` : valeurs entre 70,5 et 77,5 incluse.
- `'(77.5-inf)'` : valeurs supérieures à 77,5.

et les valeurs de `humidity` doivent être discrétisées selon les modalités suivantes :

- `'(inf-77.5]'` : valeurs inférieurs ou égales à 77,5.
- `'(77.5-88]'` : valeurs entre 77,5 et 88 incluse.
- `'(88-inf]'` : valeurs supérieures à 88.

Sauvegardez le résultat dans un fichier `weather.discretized.arff`.
Afin d'augmenter la lisibilité des valeurs, remplacez dans ce fichier les noms des intervalles générés comme indiqué dans le tableau ci-dessous :

`temperature`		`humidity`	
Valeur	Remplacée par	Valeur	Remplacée par
`'(inf-70.5]'`	`cool`	`'(inf-77.5]'`	`low`
`'(70.5-77.5]'`	`temperate`	`'(77.5-88]'`	`medium`
`'(77.5-inf)'`	`hot`	`'(88-inf)'`	`high`

Sauvegardez le fichier et chargez-le dans Weka.

3.2 Question 11

L'utilisateur définit le nombre de règles qu'il souhaite obtenir et la valeur de départ du seuil *minsupport*. Cette implantation de l'algorithme `Apriori` recherche les règles en diminuant successivement *minsupport* jusqu'à ce que :

— soit le nombre de règles demandé est atteint,
— soit la borne inférieure définie pour *minsupport* est atteinte.

Le seuil *minsupport* est défini comme un pourcentage de nombre d'instances du jeu de données.

Les paramètres de `Apriori` sont :

— `metric Type` : mesure de précision des règles de la forme $A \rightarrow C$. Les mesures de précision qu'il est possible d'utiliser sont :

Mesure	Calcul des valeurs de la mesure
`confidence`	$support(A \wedge C)/support(A) = Prob(A \wedge C)/Prob(A)$
`lift`	$Prob(A \wedge C)/Prob(A) \times Prob(C)$
`leverage`	$Prob(A \wedge C) - Prob(A) \times Prob(C)$
`conviction`	$Prob(A) \times Prob(\neg C)/Prob(A \wedge \neg C)$

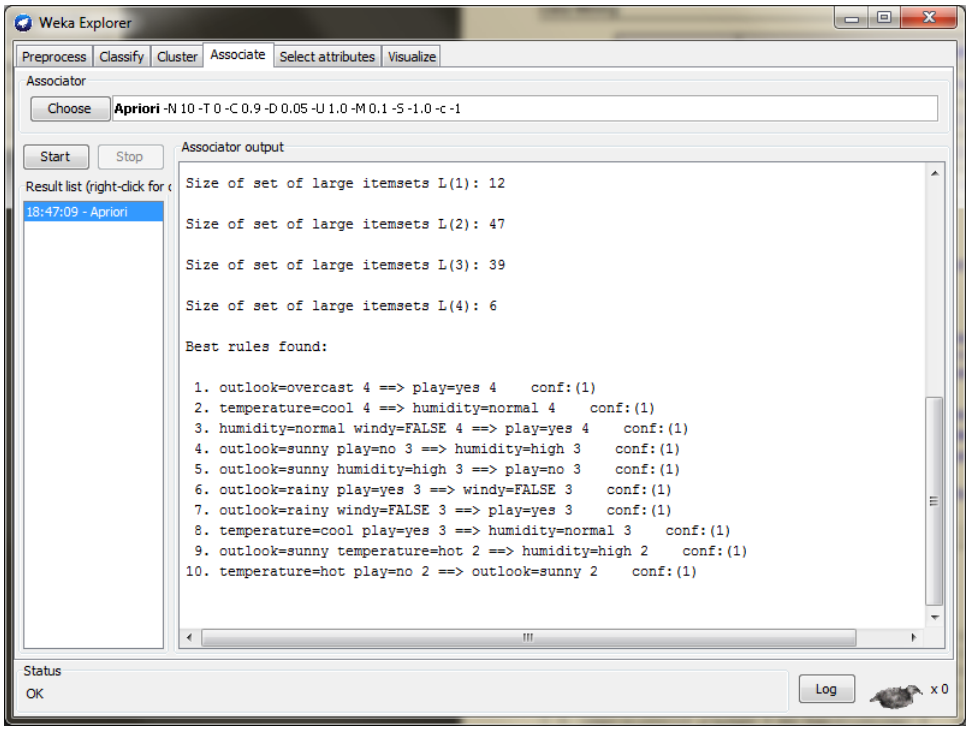
— `minMetric` : valeur minimale de précision des règles.
— `upperBoundMinSupport` : valeur initiale de *minsupport.*
— `lowerBoundMinSupport` : borne inférieure pour *minsupport.*
— `delta` : valeur de décrémentation de *minsupport.*
— `numRules` : nombre de règles à afficher.

Le résultat affiche le nombre d'itemsets (combinaisons de valeurs des attributs) fréquents classés par taille, par exemple `Size of set of large itemsets L(1)` indique le nombre de 1-itemsets fréquents, et la liste ordonnée des règles décrites sous forme textuelle.

Pour chaque règle $A \rightarrow C$ sont affichés :

— le support de l'antécédent : nombre d'instances contenant les valeurs de A.
— le support de la règle : nombre d'instances contenant les valeurs de A et C.
— la confiance de la règle : proportion d'instances contenant A et C parmi toutes celles qui contiennent A.

Les règles sont affichées par valeurs décroissantes de supports (nombre d'instances concernées) et de précision de la règle (la confiance ci-dessous).



La première règle indique que les 4 instances (support de l'antécédent) qui possèdent `outlook=overcast` possèdent aussi `play=yes`. Cette règle concerne 4 instances (support de la règle) et elle est donc vraie pour toutes les instances (confiance de 1, c-à-d 100%).
Afin d'explorer plus facilement les règles extraites, vous pouvez enregistrer le résultat de l'exécution dans un fichier en cliquant avec le bouton droit dans le fenêtre `Result list`.
Appliquez l'algorithme d'extraction de règles d'association `Apriori` sur le fichier `weather.discretized.arff` avec les paramètres par défaut.
Identifiez dans le résultat les trois règles les plus fortes (confiance et support maximaux) permettant de prédire que l'on va jouer au tennis.

3.3 Question 12

Identifiez dans le résultat les trois règles les plus fortes (confiance et support maximaux) permettant de prédire que l'on ne va pas jouer.

3.4 Question 13

Appliquez l'algorithme de classification supervisée `Prism` sur le fichier `weather.discretized.arff` et indiquez les règles de classification générées.

3.5 Question 14

Comparez les règles de classification générées par `Prism` et les règles d'association générées par `Apriori` à la Question 11. Notez vos observations (identifiez les règles identiques ou les valeurs similaires par exemple).

3.6 Question 15

Vous allez comparer les quatre mesures de précision et identifier pour chacune les règles les plus fortes contenant `play=yes` puis `play=no` seul dans la partie de droite. Si besoin vous augmenterez le nombre de règles (paramètre `numRules`) et diminuerez la valeur minimale de la mesure (paramètre `minMetric`) pour obtenir davantage de règles.

Les règles contenant seulement `play=yes` (resp. `play=false`) dans la partie gauche permettent d'identifier (dans la partie droite) les conditions climatiques les plus communes aux jours où il est possible de jouer (resp. où il est impossible de jouer).

3.7 Question 16

Lancez l'extraction de règles avec la confiance comme mesure et identifiez la première règle contenant l'attribut `play=yes` dans la partie de gauche.
Faites de même pour `play=no`. Si besoin, afin d'obtenir davantage de règles, augmentez le nombre de règles extraites et diminuez le seuil de *minconfiance*.

4 Comparaison de méthodes

On souhaite maintenant confirmer les observations faites concernant les attributs `temperature` et `humidity`.

4.1 Question 17

Pour cela vous allez évaluer l'influence de la valeur de `temperature`, seule, sur le fait que l'on jouera ou non. Identifiez la première règle qui permet de prédire que l'on jouera selon la température uniquement en utilisant la confiance comme mesure de précision.

4.2 Question 18

Identifiez la première la première règle qui permet de prédire que l'on ne jouera pas selon la température uniquement.

4.3 Question 19

Cela confirme-t-il les précédentes observations faites sans les règles d'association ?

4.4 Question 20

Identifiez de la même manière les règles permettant de prédire que l'on va jouer ou non en fonction de l'humidité (attribut `humidity`) uniquement. Cela confirme-t-il les précédentes observations faites sans les règles d'association ?