

TD/TP n°4

Règles d'association (Motifs fréquents)

1. Description des outils d'extraction d'associations dans Weka

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

The screenshot shows the Weka Explorer interface. The 'Associate' tab is active. The 'Choose' button is highlighted, and the 'Apriori' algorithm is selected. The parameters for Apriori are: -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1. The 'Start' button is also visible. The 'Associator output' pane shows the following text:

```
Result list (right-click for c
14:03:23 - Apriori
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 47
Size of set of large itemsets L(3): 39
Size of set of large itemsets L(4): 6
Best rules found:
1. outlook=overcast 4 ==> play=yes 4   conf:(1)
2. temperature=cool 4 ==> humidity=normal 4   conf:(1)
3. humidity=normal windy=FALSE 4 ==> play=yes 4   conf:(1)
4. outlook=sunny play=no 3 ==> humidity=high 3   conf:(1)
5. outlook=sunny humidity=high 3 ==> play=no 3   conf:(1)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3   conf:(1)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3   conf:(1)
8. temperature=cool play=yes 3 ==> humidity=normal 3   conf:(1)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2   conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2   conf:(1)
```

Annotations in the image point to the 'Choose' button (labeled 'Choix de l'algorithme'), the parameter string (labeled 'Paramètre de l'algorithme'), the 'Result list' (labeled 'Liste des exécutions effectuées'), and the 'Associator output' (labeled 'Résultats de l'exécution sélectionnée').

Exercice 1 : Chargez le jeu de données "weather.arff" dans Weka. Essayez d'exécuter les algorithmes Apriori et Tertius avec leurs paramètres par défaut. Que constatez-vous ?

Application impossible de ces algorithmes car le jeu contient des valeurs numériques.

Les attributs numériques Temperature et Humidity doivent être discrétisés afin d'appliquer l'extraction de règles d'association. Les attributs du jeu de données doivent tous être qualitatifs uniquement (non numériques). Certains algorithmes de classification n'opèrent eux-aussi que sur des attributs qualitatifs.

1.1. Discrétisation des attributs numériques

Sur l'onglet `Preprocess`, le cadre `Filter` permet de définir des filtres de transformation des données. Ces filtres sont des opérations (suppression, normalisation, discrétisation, transformation, 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.

Exercice 2 : 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 qualitatifs. Pour chaque attribut créé, les modalités correspondent à des intervalles de valeurs de même fréquence (même nombre d'instances pour chaque modalité).

Visualisez les valeurs 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 inférieures 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.nominal.arff"`.

Exercice 3 : Afin d'augmenter la lisibilité des valeurs, ouvrez ce fichier dans un éditeur de texte et remplacez les noms des intervalles générés comme indiqué dans le tableau ci-dessous.

Temperature		Humidity	
Valeur	Remplacée par	Valeur	Remplacée par
<code>'[inf-70.5]'</code>	cool	<code>'[inf-77.5]'</code>	low
<code>']70.5-77.5]'</code>	temperate	<code>']77.5-88]'</code>	medium
<code>']77.5-inf[</code>	hot	<code>']88-inf[</code>	high

Sauvegardez ensuite le fichier et chargez-le dans Weka.

1.2. Algorithme Apriori

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 (confiance)	$\text{support}(A \wedge C) / \text{support}(A) = \text{Prob}(A \wedge C) / \text{Prob}(A)$
lift (intérêt)	$\text{Prob}(A \wedge C) / \text{Prob}(A) \times \text{Prob}(C)$
leverage	$\text{Prob}(A \wedge C) - \text{Prob}(A) \times \text{Prob}(C)$
conviction	$\text{Prob}(A) \times \text{Prob}(\neg C) / \text{Prob}(A \wedge \neg C)$

- `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 (motifs fréquents de longueur 1), et la liste ordonnées 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).

```

Associator output

Best rules found:

1. outlook=overcast 4 ==> play=yes 4    conf:(1)
2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
3. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
4. outlook=sunny play=no 3 ==> humidity=high 3    conf:(1)
5. outlook=sunny humidity=high 3 ==> play=no 3    conf:(1)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
8. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2    conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    conf:(1)
    
```

Support de l'antécédent Support de la règle Confiance de la règle

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 donc est vraie pour toutes les instances (confiance de 1, c'est-à-dire 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 la fenêtre `Result list`. Vous pourrez alors ouvrir le fichier enregistré avec un éditeur de texte ou un tableur. Ceci vous permettra par exemple de sélectionner les règles contenant un ou plusieurs items particuliers.

Exercice 4 : Appliquez l'algorithme d'extraction de règles d'association Apriori sur le fichier "weather.nominal.arff" avec les paramètres par défaut. Au besoin, vous augmenterez le nombre de règles extraites (paramètre `numRules`) pour répondre aux questions suivantes.

Identifiez dans le résultat les trois règles les plus fortes (confiance et support maximaux) permettant de prédire que l'on va jouer au tennis, c'est à dire les règles contenant `play=yes` dans la partie droite.

1.	==> play=yes	conf:()
2.	==> play=yes	conf:()
3.	==> play=yes	conf:()

Identifiez dans le résultat les trois règles les plus fortes permettant de prédire que l'on ne va pas jouer, c'est à dire les règles contenant `play=no` dans la partie droite.

1.	<code>==> play=no</code>	conf:()
2.	<code>==> play=no</code>	conf:()
3.	<code>==> play=no</code>	conf:()

Exercice 5 : Appliquez l'algorithme de classification supervisée Prism sur le fichier "weather.nominal.arff" et notez ci-dessous les règles de classification générées.

1.
2.
3.
4.
5.
6.
7.

Exercice 6 : Comparez les règles de classification générées par Prism et les règles d'association générées par Apriori dans l'exercice 4. Notez vos observations dans le cadre ci-dessous (identifiez les règles identiques ou les valeurs similaires par exemple).

--	--

Exercice 7 : Vous allez comparer les quatre mesures de précision en identifiez pour chacune les règles les plus fortes contenant `play=yes` puis `play=no` seul dans la partie 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.

Confidence :	<code>==> play=yes</code>	conf:()
	<code>==> play=no</code>	conf:()
Lift :	<code>==> play=yes</code>	lift:()
	<code>==> play=no</code>	lift:()
Leverage :	<code>==> play=yes</code>	lev:()
	<code>==> play=no</code>	lev:()
Conviction :	<code>==> play=yes</code>	conv:()
	<code>==> play=no</code>	conv:()

Les règles contenant seulement `play=yes` (respectivement `play=no`) 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 (respectivement où il est impossible de jouer).

Exercice 8 : Lancez l'extraction de règles avec la confiance comme mesure et identifiez la première règle contenant l'attribut `play=yes` dans la partie gauche. Faites de même pour `play=no`.

Si besoin, afin d'obtenir davantage de règles, augmentez le nombre de règles extraites (paramètre `numRules`) et diminuez le seuil de `minConfiance` (paramètre `minMetric`).

<code>play=yes</code>	<code>==></code>	conf:()
<code>play=no</code>	<code>==></code>	conf:()