TP2

Exercice 1

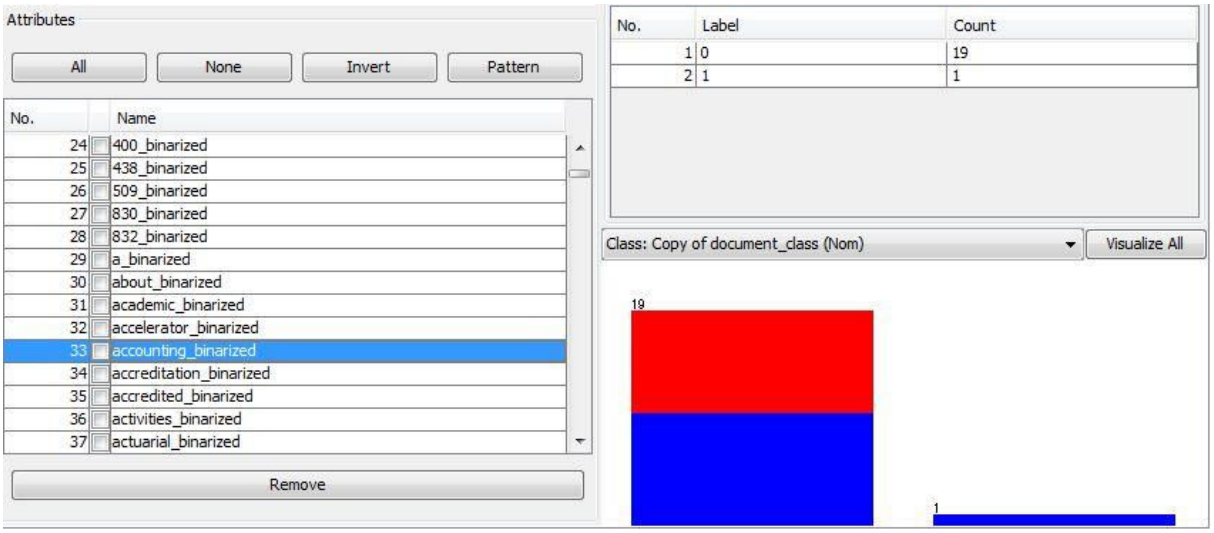
L'objectif de cet exercice est de trouver un sous-ensemble d'attributs permettant de mieux discriminer entre les classes : **attributs pertinents.** On appelle cette opération *sélection d'attributs*.

1. Reprendre le fichier doc_string.arff déjà prétraité en TP 1.
2. Examiner les trois attributs suivants : *accounting, science* et *sciences*. Remplir les tableaux de distribution suivants

		accounting	
		0	1
Classe	A		
	B		

		science	
		0	1
Classe	A		
	B		

		sciences	
		0	1
Classe	A		
	B		



3. Extraire des tableaux précédents des règles sous la forme :

   Si attribut = valeur_attribut Alors classe = valeur_classe
   Erreur = valeur_erreur
   Couvertur = valeur_couverture

4. Déduire quels sont les attributs pertinents.
5. Les attributs *science* et *sciences* sont-ils vraiment deux attributs distincts ? Les réunir en un seul attribut, et reconstruire le tableau de distribution. Le nouvel attribut est-il pertinent ?
6. Découvrir l'onglet *Select attributes*. Explorer les paramètres par défaut (*Attribute evaluator*, *Search method*)
7. En utilisant les paramètres par défaut, effectuer et comparer les résultats :
   - Une sélection d'attributs en gardant l'attribut *document_name*
   - Une sélection d'attributs en éliminant l'attribut *document_name*.

8. Classement d'attributs (ranking) : en utilisant *Attribute evaluator= GainRatioAttributeEval* et *Search method = Ranker*, effectuer un classement d'attributs. Chercher une explication dans les distribution des valeurs de classe par rapport aux valeurs d'attributs.
9. Effectuer la même opération en utilisant les filtres (*filters/supervised/attribute/AttributeSelection*).

Exercice 2

1. A partir du dossier data de weka, charger le fichier *weather.nominal.arff*. Remplir le tableau suivant

Attribut	Règles	Taux d'erreur	Taux d'erreur total
Outlook	sunny -> no overcast -> yes rainy -> yes		
Temperature	hot -> no mild -> yes cool -> yes		
Humidity	high -> no normal -> yes		
Windy	false -> yes true -> no		

2. Le classifieur OneR utilise un seul attribut (celui ayant le plus faible taux d'erreur) pour effectuer la classification. Quelles règles de classification aura-t-on si on utilise OneR ?
3. Vérifier la réponse de la question précédente. Onglet *Classify*: *Classifier/rules/OneR*.
4. Explorer le résultat de la classification
    a. *Detailed Accuracy By Class* (TP, FP, Precision, Recall,…)
    b. *Confusion Matrix*

5. Utiliser l'arbre de décision J48 pour le même et faites varier les techniques d'évaluation. Quelles remarques peut-on faire par rapport aux résultats obtenus ?
6. Visualiser l'arbre. Pour afficher l'arbre de décision, cliquer droit dans la partie "Result list (right-click for options)". Choisir l'option "Visualize tree".
7. Analysez l'arbre résultat.
8. Visualisez l'erreur de clssification "Visualize classifier errors".
9. Selectionnez "More options". Recréer l'arbre en verifiant l'effet des options.