Etat de l'art

Les premiers travaux introduisant le terme de fouille visuelle de données datent de la fin des années 1990 [Keim et Kriegel, 1996], [Brunk et al., 1997], [Cox et al, 1997] et [Inselberg, 1998]. Dans un premier temps, toute l'attention a été portée vers le développement de techniques performantes et innovantes d'accès, de représentation graphique et de traitement des données. La masse de données disponible dans le monde ne cesse d'augmenter. La FVD utilise la visualisation comme canal de communication pour la découverte de corrélations dans les données. L'espace disponible sur un écran pour la représentation de ces données est limitée. Une première préoccupation concerne l'amélioration des techniques existantes de stockage, d'accès et de représentation graphique des données. Dans cet ordre d'idées, [Keim, 1996] a utilisé des techniques orientées pixel pour la représentation des données multidimensionnelles (VisDB). L'outil segments de cercle de [Ankerst et al., 1999] utilise aussi le même principe. Un autre aspect abordé dans ce domaine a conduit à une modélisation de la tâche de FVD [Ankerst, 2000]. Le modèle de tâche de FVD présenté par Ankerst possède 3 variantes (figure 1.17) suivant le mode d'utilisation des représentations graphiques. Pour chacune de ces variantes, on dénote des phases de visualisation des données, d'application de méthodes d'analyse de données avant d'aboutir à la connaissance (modèle des données). En effet, lorsqu'on se sert de la visualisation comme support en fouille de données, après la sélection des données à exploiter (première étape des 3 variantes), une alternative se présente : soit l'utilisateur sélectionne et exécute un algorithme automatique de fouille de données, soit il procède à une visualisation (exploration) de l'ensemble de données. La visualisation peut être suivie de l'application d'une méthode automatique (ou interactive) de construction du modèle des données. L'étape suivante consiste en la visualisation des résultats. On assiste enfin à une évaluation puis à une exploitation de ces résultats considérés comme des connaissances nouvelles. Plus explicitement, les variantes de la figure 1(modèle de tâche de Ankerst) correspondent à :

- l'utilisation de la visualisation pour l'interprétation des résultats d'un algorithme automatique de fouille de données,
- l'utilisation de la visualisation pour une exploration des données qui permet à l'utilisateur d'avoir une idée générale des données, suivi d'une application d'un algorithme automatique de fouille de données à l'ensemble de données et de l'utilisation de la visualisation pour une interprétation des résultats finaux,
- dans le dernier cas, la représentation graphique sert de support aux traitements. Afin de construire le modèle des données, l'utilisateur interagit avec la représentation visuelle et procède à des traitements successifs qui lui permettent de construire un modèle des données.

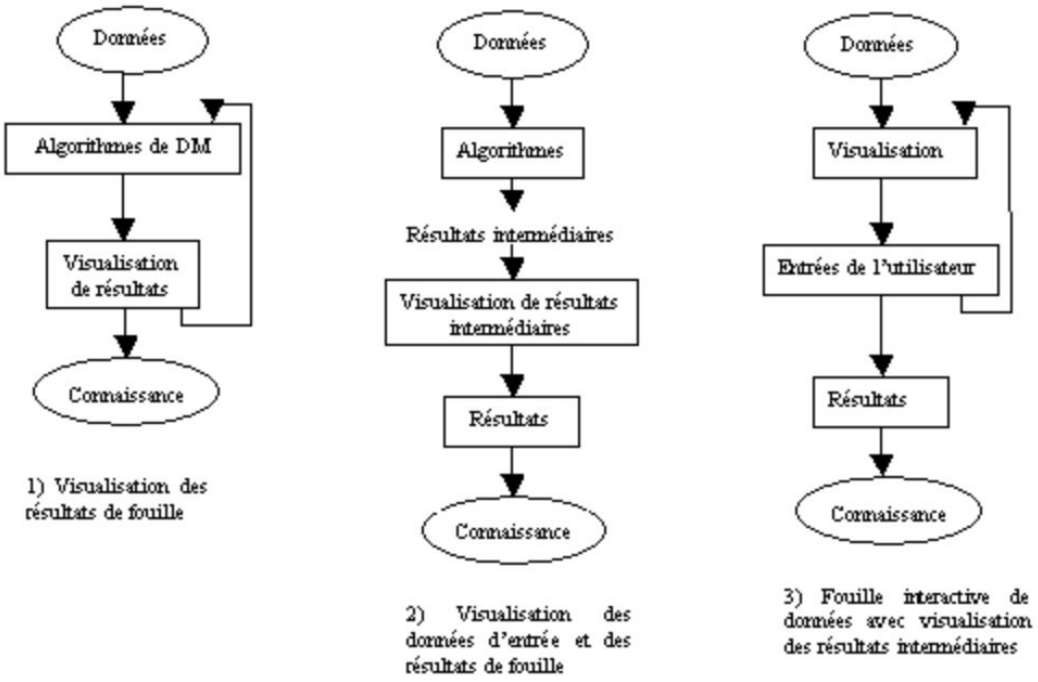


Figure 1 Fouille de données et visualisation

Pour les deux premières variantes du modèle de Ankerst, la plupart de techniques de représentations graphiques utilisées ont été présentées en début de ce chapitre. Nous allons à présent dresser un état de l'art des différentes techniques de visualisation qui permettent de découvrir des corrélations dans les données de façon interactive (variante 3 du modèle de Ankerst), de construire un modèle de données et de représenter les résultats issus de la construction du modèle des données.

Le modèle de tâche en FVD montre que la construction du modèle de données peut se faire de façon automatique (deux premières variantes du modèle de Ankerst) ou alors de manière interactive (troisième variante du modèle de Ankerst). Cet état de l'art est essentiellement basé sur la construction interactive du modèle de données qui possède de nombreux avantages par rapport aux algorithmes automatiques couplés ou non aux méthodes de représentation graphique. L'exécution des algorithmes automatiques d'analyse de données nécessite une étape préalable de paramétrage, ce qui n'est pas le cas en construction interactive du modèle de données. En effet, l'algorithme automatique se comporte comme une boîte noire recevant en entrée des données et fournissant en sortie un modèle de ces données. L'utilisateur ne participe pas à la construction de ce modèle, ce qui pourrait avoir une incidence sur le degré de confiance qu'il accordera au résultat. La vision humaine peut servir à capturer des corrélations complexes dans les ensembles de données au travers de représentations graphiques. Si l'utilisateur de l'outil de fouille interactive de données est un spécialiste du domaine des données, il peut utiliser ses connaissances du domaine de données durant le processus de fouille et non seulement au moment de l'interprétation des résultats (cas des algorithmes automatiques). La confiance au modèle de données ainsi construit est élevée car l'utilisateur a

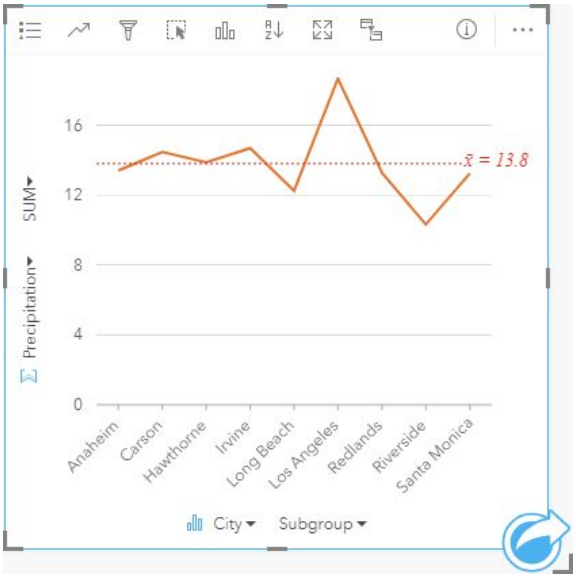
participé à sa construction. Le temps de traitement avec l'algorithme interactif peut s'avérer long, surtout pour de grands ensembles de données.
Les différentes variantes du modèle de Ankerst montrent aussi que l'étape de construction du modèle de données est suivie d'une étape de post traitement au cours de laquelle l'utilisateur peut avoir recours aux techniques de visualisation. A cet effet, il existe des techniques de visualisation telles que CUBEVIS [Poulet, 2001] ou Grand tour [Asimov, 1985]. Nous nous limitons ici à l'étape de construction du modèle de données.

1 Techniques standards 2D/3D

Visualiser de statistique sous la forme de graphiques est bien connus, ce qui est le plus courant méthode dans notre vie quotidien. 5 types le plus courantes graphiques de Data Visualisation est comme ci-dessous.

1.1 Graphique linéaire

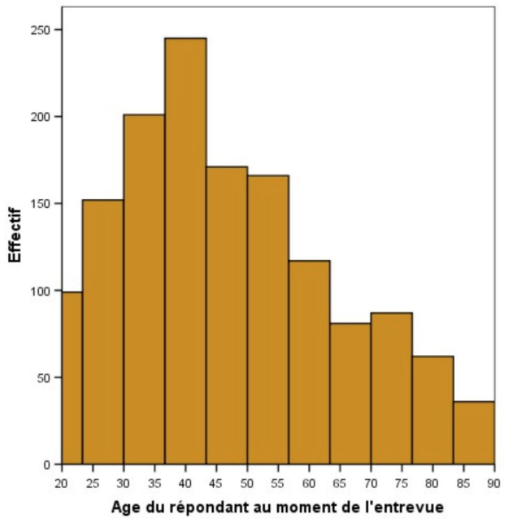
Les graphiques linéaires peuvent afficher des données continues qui changent avec le temps (en fonction des paramètres de mise à l'échelle courants). Ils sont donc parfaits pour afficher les tendances des données à intervalles égaux.



Le diagramme linéaire ci-dessus montre les fluctuations des précipitations d'une ville à une autre.

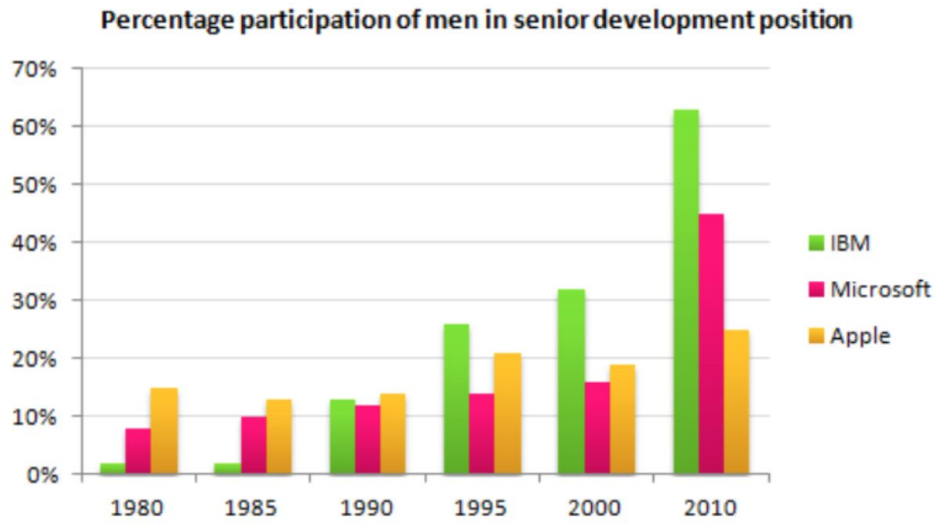
1.2 Histogramme

un **histogramme** est un graphique permettant de représenter la **répartition des valeurs d'une variable continue** (Continuous Data). Chaque colonne de l'histogramme représente un intervalle de valeurs. La hauteur des colonnes indique le nombre d'instances dans cet intervalle. L'examen de l'histogramme permet de se faire une idée claire sur **la distribution des valeurs** de la *feature* analysée.



1.3 Bar Plot

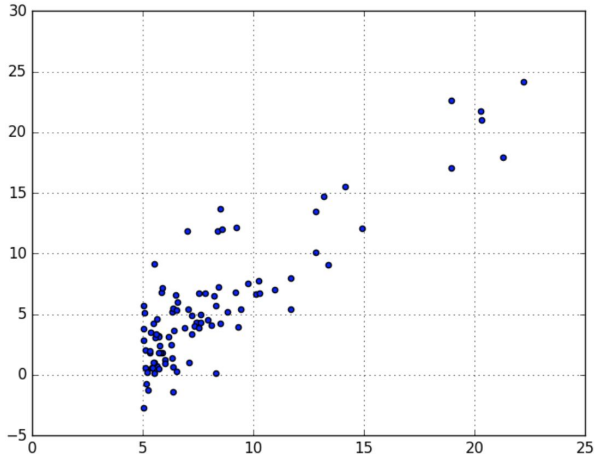
Les Bar Plots sont utilisés pour visualiser des données qualitatives (**Categorical Data**). Chaque "Barre" d'un bar plot représente une catégorie (modalité) et la hauteur de la barre indique la taille du groupe faisant partie de cette catégorie.



Pourcentage de la poste de développement senior de 3 sociétés

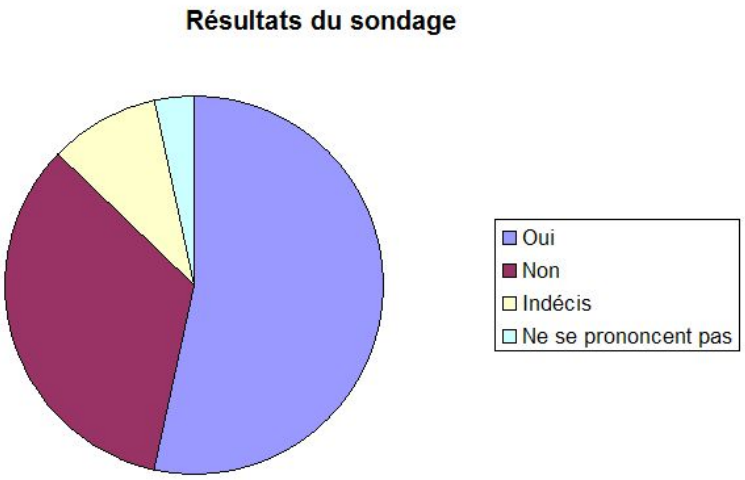
1.4 Scatterplot Matrices

Les Scatter Plot Matrices (Diagrammes de dispersion) permettent de visualiser la **corrélation entre deux variables continues.** On met la première feature sur l'axe des abscisses (X) et la deuxième sur les ordonnées (Y). La dispersion des points indique la relation entre les deux features.



1.5 Camembert

Les « camemberts » permettent de représenter des proportions et des pourcentages.



2 Techniques orientées pixel et construction d'un modèle de données

Les méthodes de visualisation qui utilisent cette technique sont : PBC [Ankerst, 1999], segments de cercle [Ankerst et Keim, 1996]) et DTViz [Han et Cercone, 2001].

2.1 Segments de cercle

Toute la base de données est représentée dans un cercle. Le cercle est divisé en segments, un segment pour chaque attribut. Dans le segment, chaque valeur d'attribut est représentée par une valeur de pixel. L'arrangement des pixels va du centre du cercle vers l'extérieur.

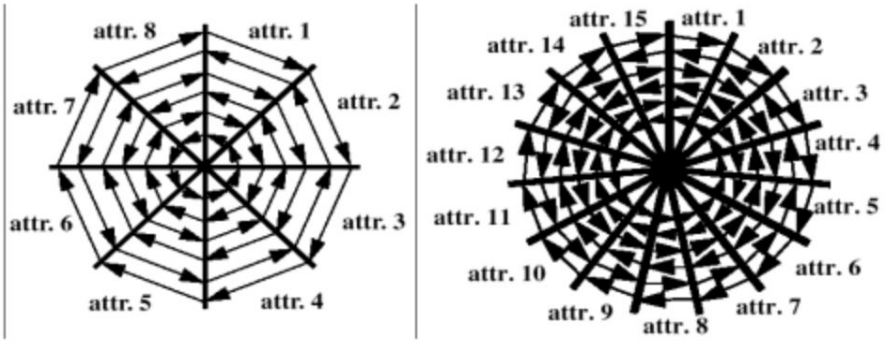


Figure 2 Représentation segments de cercles

Si on opère un changement dans le parcours utilisé pour remplir les pixels dans la représentation sous forme de segments de cercle, on obtient une représentation sous forme de barres rectangulaires.
Dans ces modes de visualisation, les données sont triées avant d'être affichées sous forme de pixels.

2.2 Barres rectangulaires

La figure 3 illustre la représentation en barres rectangulaires.

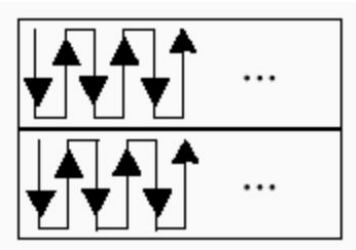


Figure 3 Représentation en barres rectangulaires

Segments de cercle et barres rectangulaires permettent de visualiser des ensembles de données relativement grands et d'appliquer l'algorithme PBC décrit dans le paragraphe 1.3. L'inconvénient de ce type d'approche tient au fait que suite au tri des données avant visualisation, l'information portant sur la distance des données est perdue. De plus, cette opération de tri constitue un coût supplémentaire dans les traitements par rapport aux techniques de représentation de type matrice de scatter plot.

2.3 Perception-Based Classification (PBC)

PBC est un algorithme interactif de construction d'arbres de décision qui utilise le principe de la pixellisation. La phase initiale de PBC permet de représenter graphiquement l'ensemble de données d'apprentissage (segments de cercle ou barres rectangulaires) et d'initialiser un arbre de décision au nœud racine, correspondant à l'ensemble de données d'apprentissage.
De façon concrète, la représentation graphique d'un ensemble de données aboutira à la figure 4 par exemple.

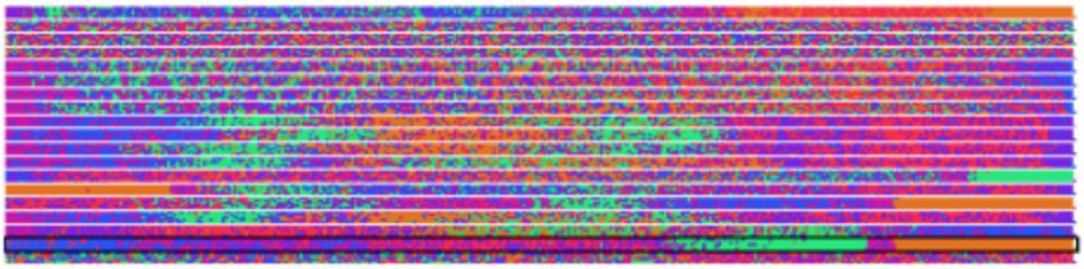


Figure 4 Exemple de représentation d'un ensemble de données

La visualisation permet de sélectionner de façon interactive des données et de procéder à des coupes. A partir de la représentation de la figure 5, pour la construction d'un arbre de décision par exemple, on peut procéder à des coupes interactives : binaires ou n aires mono variées. L'idée ici est de concevoir de façon interactive un modèle des données. A cet effet, on utilise la stratégie suivante : recherche de la meilleure partition pure, s'il n'en existe pas : recherche de la plus grande partition dominante, s'il n'en existe pas : recherche de l'ensemble de partitions dominantes.

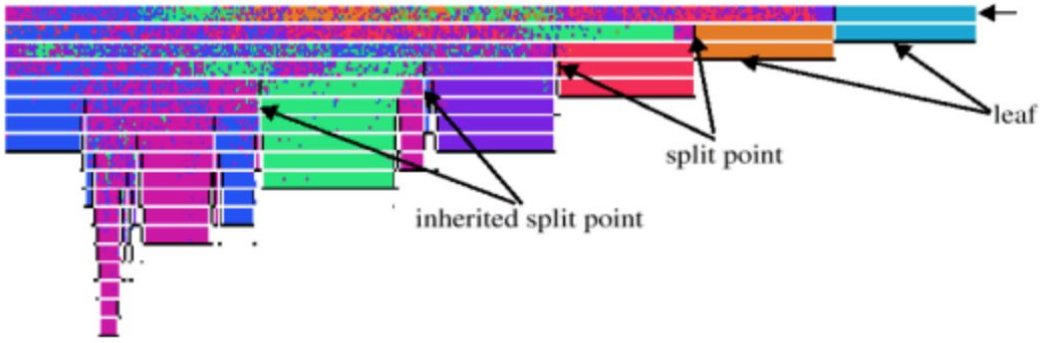


Figure 5 Représentation du processus de coupes successives

L'algorithme de construction interactive d'arbres de décision (CIAD) [Poulet, 2002b] et le module UserClassifier [Ware et al, 2001] de WEKA utilisent le même principe de construction d'arbres de décision mais sont basés sur des représentations matricielles. Les sections détaillent ces deux approches.

3 Techniques matricielles et construction interactive d'un modèle des données

3.1 CIAD (Construction Interactive d'Arbre de Décision)

CIAD est un outil permettant la construction interactive d'arbres de décision. Cette technique utilise des matrices de scatter plot comme technique de visualisation et permet pour des ensembles de données avec un nombre de dimensions (n inférieur à 20) une projection de n*(n-1)/2 matrices. Pour n>20, une représentation par défaut de l'ensemble de données est fournie avec une combinaison de 20 attributs au maximum. La première étape de traitement consiste à représenter graphiquement l'ensemble de données à traiter. La figure 6 représente une vue de l'ensemble de données segmentation de l'UCI [Blake et Merz, 1998] avec CIAD.

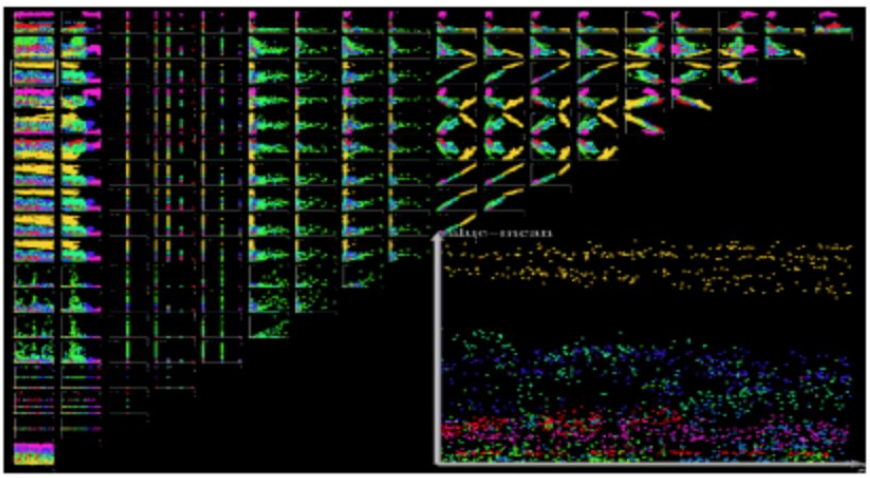


Figure 6 Représentation de l'ensemble de données segment avec CIAD

La couleur représente la classe. Les coupes effectuées pour la construction du modèle de données sont de type oblique en 2 dimensions donc sur deux variables. Ces différentes coupes sont effectuées grâce aux capacités humaines en reconnaissance de formes. Les étapes successives de ce traitement sont illustrées par la figure 7.

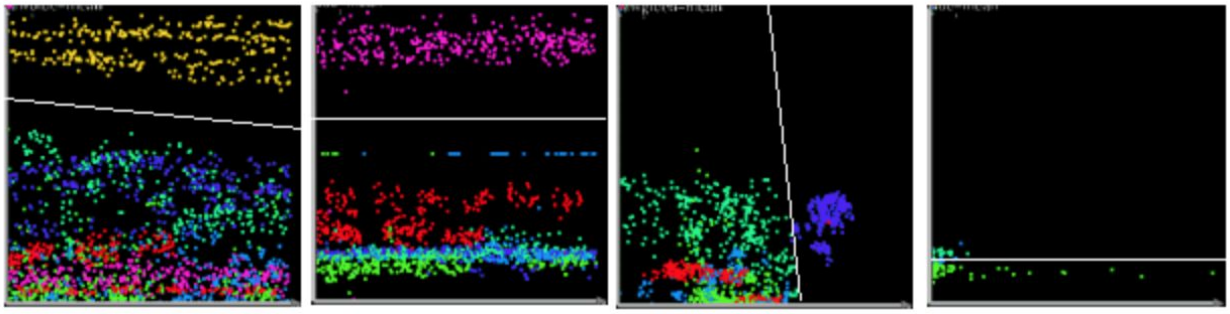


Figure 7 Construction interactive du modèle de données avec CIAD

Les 4 premières coupes représentent les classes 2, 7, 6, 3. Ces classes représentent 57% des individus de l'ensemble de données. CIAD peut être exécuté en modes 100% interactif, mixte ou alors 100 % automatique. Par rapport aux méthodes automatiques, CIAD permet d'obtenir une précision équivalente avec des tailles d'arbres inférieures.

3.2 UserClassifier de WEKA

Le module UserClassifier de WEKA est une implémentation de PBC qui utilise aussi des matrices 2D pour la construction interactive d'arbres de décision. UserClassifier à l'étape initiale de présentation de données ne permet pas d'avoir une vue globale de l'ensemble de données à traiter. Une seule matrice 2D est présentée l'écran. Pour le traitement de grands ensembles de données, la notion de contexte est perdue, il n'est pas possible de visualiser toutes les paires possibles d'attributs en même temps à l'écran. Moyennant des efforts, il est possible d'accéder

à toutes les paires de combinaison possibles d'attributs une par une, ce qui n'est pas le cas avec CIAD qui ne peut aller au-delà d'un certain nombre de dimensions (limite dûe à l'utilisation de la représentation sous forme de matrice de matrices de scatter plot). Les coupes opérées avec le module UserClassifier sont rectangulaires, polygonales ou alors sous forme de polylignes (voir la figure 8).

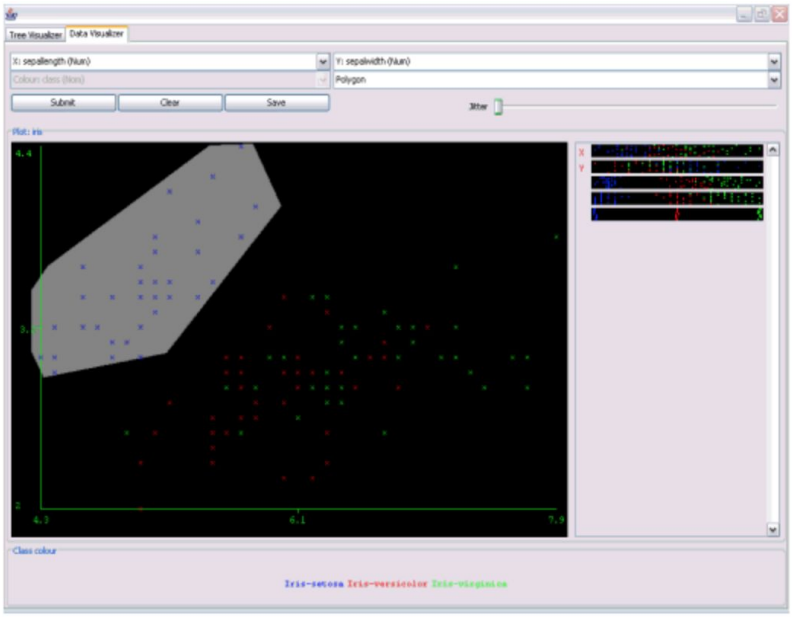


Figure 8 Représentation et construction interactive du modèle de données avec UserClassifer de WEKA

UserClassifier hérite de tous les inconvénients de la représentation graphique sous forme de matrice de scatter plot, notamment, l'impossibilité de traiter des ensembles de données pourvus d'un nombre élevé d'individus. De plus, il n'existe pas de mécanisme d'aide aux utilisateurs durant la construction du modèle des données.

Reference

1. [Blake et Merz, 1998] Blake C., Merz C.: UCI Repository of machine learning databases, [www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, University of California, Department of Information and Computer Science, 1998.
2. [Inselberg, 1998] Inselberg A.: Visual Data Mining with Parallel Coordinates, Computational Statistics Vol. 13(1), pp.47-63, 1998.
3. [Keim et Kriegel, 1994] Keim D.A., Kriegel H.-P.: VisDB: Database Exploration using Multidimensional Visualization, Computer Graphics & Applications Journal, pp.40-49, 1994.
4. [Brunk et al., 1997] Brunk C., Kelly J. Kohavi R.: MineSet : an integrated system for data mining, International Conference on Knowledge Discovery and Data Mining (KDD'97), AAAI Press, pp 135-138, 1997.
5. [Cox et al, 1997] Cox K.C., Eick S.G., Wills G.J., Brachman R.J.: Visual Data Mining: Recognizing Telephone Calling Fraud, Fraud, Data Mining and Knowledge Discovery Vol. 1, pp.225-231, 1997.
6. [Ankerst, 2000] Ankerst M.: Visual Data Mining. PhD Thesis, Ludwig Maximilians University of Munich, 2000.
7. [Han et Cercone, 2001] Han J., Cercone N.: "Interactive Construction of Decision Trees" in proc. of PAKDD'2001, LNAI 2035, pp.575-580, 2001.
8. [Poulet, 2002b] Poulet F.: Cooperation between automatic algorithms, interactive algorithms and visualization tools for visual data mining. In Proc. of Visual Data Mining workshop, PKDD2002, pp. 67-79, 2002.
9. [Edwige P. Fangseu Badjio, 2005] Edwige P : Evaluation qualitative et guidage des utilisateurs en fouille visuelle de données. 2005