ESIEA Datalab
Logiciel de Nettoyage et Préparation de Données

Christopher Corsia*

*ESIEA pôle ECD, 9 rue vésale, 75005 Paris
christopher.corsia@esiea.fr

1 Introduction

Il est communément admis que le temps de préparation des données peut occuper jusqu'à 80% du temps lors d'un projet industriel de fouille de données (Pyle, 1999). L'hétérogénéité des sources, la présence de valeurs manquantes, les erreurs de saisie ou de calcul, les pannes de capteurs, une mauvaise fusion de données sont autant de causes qui peuvent introduire erreurs et incohérences dans une table de données. ESIEA Datalab est une plateforme évolutive programmée en Java qui met à disposition de nombreux outils pour aider à la détection d'incohérences, la correction d'erreurs, la transformation ou la contrainte de variables, etc.

2 Le concept du logiciel

Le nettoyage et la préparation de données peuvent être vus sous la forme d'un processus représenté par la figure 1.

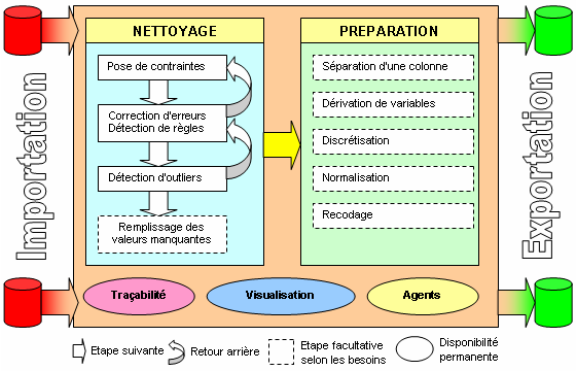

FIG. 1 – *Le nettoyage et la préparation de données vus comme un processus.*

Le logiciel n'impose pas ce processus à l'utilisateur, mais fournit tous les outils nécessaires à sa réalisation. En parallèle, le nettoyage et la préparation des données sont tracés dans

la console afin de pouvoir retrouver toutes les transformations et modifications effectuées sur les données et des agents fonctionnent en tâche de fond pour faire des suggestions et orienter l'utilisateur.

3 Les outils

Outre un vaste ensemble d'outils classiques, dans lesquels les algorithmes utilisés ont été adaptés à un contexte où toute valeur peut être manquante ou bien en erreur, ESIEA Datalab possède quelques outils originaux puissants qui permettent de traiter facilement des cas difficiles de nettoyage ou d'offrir des moyens de visualisation intéressants.

Type structuré. Grâce à la notion de type structuré, le logiciel est capable de détecter des erreurs dans des données symboliques possédant une structure. Une fois la structure d'une colonne spécifiée ou inférée, on peut contraindre les éléments de la structure à l'aide de formules et mettre ainsi en erreur les valeurs ne respectant pas l'une des contraintes.

Outils de visualisation. Parmi les outils de visualisation disponibles, ESIEA Datalab dispose de graphiques interactifs (matrice de nuages de points, coordonnées parallèles, etc.) qui permettent la sélection de valeurs et la réalisation d'actions sur celles-ci. On trouve aussi des outils originaux comme la carte « vue d'avion ». C'est un graphique qui représente dans une forme condensée toute une table, que l'on va utiliser avec des filtres qui vont colorer une sélection de valeurs. On a ainsi une vision totale de la table qui peut par exemple nous aider à estimer la densité des valeurs manquantes ou bien détecter des motifs.

4 Conclusion

ESIEA Datalab est un logiciel évolutif dont la simplicité d'utilisation des outils et les fonctionnalités adaptées permettent d'obtenir un gain de temps important sur le nettoyage et la préparation des données. Plusieurs améliorations sont en projet, notamment l'ajout d'une passerelle vers la librairie Java WEKA (Witten et Eibe, 2005).

Références

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Witten, I.H. et F. Eibe (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

Summary

ESIEA Datalab is an evolvable Java software program which goal is to clean and prepare data before an analysis. The software looks like a toolbox ready to use, including some interactive visualisation tools, suggestion agents and advanced functionalities implementing Data Mining algorithms.