Fouille de données par programmation visuelle structurée avec KD-Ariane

Régis Clouard, François Rioult

Fouille de données par programmation visuelle structurée avec KD-Ariane

Régis Clouard*, François Rioult*

*GREYC CNRS UMR6072 - ENSICaen - Université de Caen Basse-Normandie
regis.clouard@ensicaen.fr, francois.rioult@unicaen.fr
https://forge.greyc.fr/projects/kdariane

1 Introduction

La programmation visuelle de composants de haut niveau permet d'améliorer le développement des chaînes de traitement. C'est pourquoi les plate-formes libres pour l'analyse de données (Weka [1], RapidMiner [2], Knime [3]) proposent des environnements de développement visuel. Pour autant, il y est difficile d'ajouter ses propres réalisations, par exemple des binaires compilés, car ces plate-formes requièrent l'utilisation d'une API dédiée, imposant par exemple un format spécifique pour la manipulation des données. Les possibilités de programmation structurée y sont également limitées : ces plate-formes sont conçues pour enrober leurs propres composants de base pour l'apprentissage automatique.

Nous présentons ici la plate-forme KD-Ariane, un déploiement d'outils pour la fouille de données dans l'environnement de programmation visuelle Ariane. Ce déploiement facilite la conception de chaînes structurées de traitements pour l'extraction de connaissance dans les données (Clouard (2009)).

KD-Ariane est simple à prendre en main et possède un fort potentiel pédagogique. Les composants graphiques exécutent simplement des commandes système, rendant aisé l'interfaçage avec les librairies disponibles. La puissance réside dans les possibilités de programmation structurée offertes par les boucles, les macro-composants et leur partage avec d'autres utilisateurs sous forme de routines. Enfin, les chaînes développées avec KD-Ariane sont exportables en scripts `shell` ou `perl`.

Ariane : c'est une plate-forme de programmation visuelle, initialement conçue pour valoriser les opérateurs de traitement d'image de la librairie Pandore (Pandore (2013)). Nous illustrons ici son potentiel dans la réalisation de chaînes de traitement pour l'extraction de connaissances dans les bases de données.

Le principe est celui de la programmation visuelle : des composants sont assemblés pour réaliser la chaîne de traitements. Ces traitements sont réalisés sur des fichiers de données, afin de produire d'autres données, ou des résultats sous forme de chaîne de caractère ou de valeur numérique.

1. http://www.cs.waikato.ac.nz/ml/weka/index.html
2. http://rapidminer.com/
3. http://www.knime.org/

2 Les composants d'Ariane

Un composant est représenté par une boîte colorée, constituée de plots de liaison, de paramètres et de résultats. Sur la figure 1, une chaîne élémentaire de traitement est représentée. Les fichiers d'entrée figurent sur la gauche et contiennent les données suivantes :

Fichier `sample.sup`

```
1 3
1 4
1 6
```

Fichier `sample.bin`

```
1 3 5
2 3 5
1 4 6
2 3 6
1 4 7
1 3 5
2 3 6
2 4 7
```

Cette chaîne élémentaire mesure le support (*i.e* le nombre d'occurrences) de chaque motif du fichier `sample.sup` dans les données de `sample.bin`. Elle ne conserve que les motifs fréquents, dont le support est supérieur à la constante `minsup`, ici fixée à 2.

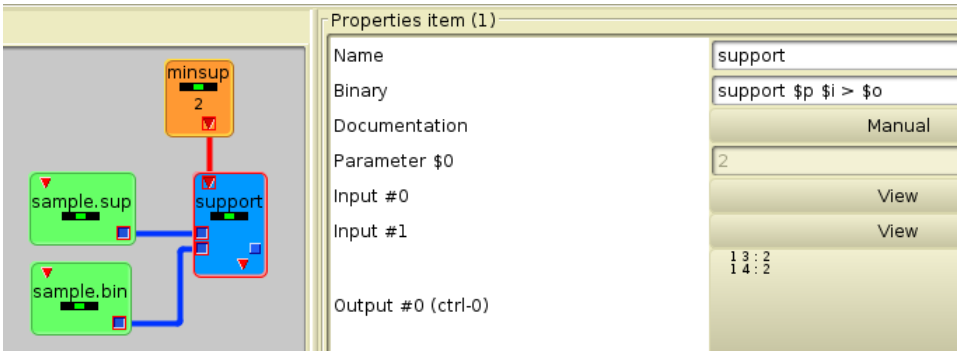

FIG. 1 – *Anatomie d'un opérateur.*

Le composant `support` effectue cette tâche en réalisant un appel au système :
- exécuter le binaire `support`
- avec les paramètres du composant (symbolisés par l'utilisation de `$p` dans la propriété `Binary`);
- indiquer au binaire les entrées (sympbolisées par `$i`);
- rediriger le résultat vers le fichier de sortie (symbolisé par `$o`).

3 Programmation visuelle structurée

Le point fort d'Ariane est de proposer des mécanismes puissants pour itérer les traitements. Ariane dispose ainsi d'opérateurs pour réaliser :
- un test d'alternative, qui oriente au choix vers deux traitements ;
- une boucle *for* ;

– une boucle *while*.

Ces opérateurs permettent de réaliser des chaînes complexes et structurées, en sortant du traditionnel cadre de la programmation par flots où les traitements sont simplement enchaînés sur les données, sans possibilité d'orientation alternative ni de répétition. On dispose ainsi d'une approche plus microscopique pour la programmation, ce qui permet d'envisager aussi bien la réalisation d'algorithmes complets, par exemple d'extraction de motifs, que le développement de chaînes de validation croisée, fondées sur des itérations.

Les développements effectués à l'aide d'Ariane sont facilement réutilisables grâce à :

– la conception de macro opérateurs ;
– l'enregistrement de ces macros dans une hiérarchie de routines ;
– l'exportation, le partage et la mutualisation de ces routines entre les utilisateurs.

4 Un exemple de réalisation

La figure 2 illustre ces concepts avec une chaîne mesurant la complexité de l'extraction des motifs fréquents. Pour cela, on souhaite chronométrer cette extraction dans une base de données classique de l'UCI, pour un seuil de support minimum allant de 1 à 50, et compter le nombre de motifs obtenu.

Nous utilisons donc une boucle *for* (*cf.* la partie supérieur de la figure 2), qui itère un traitement paramétré pour des valeurs de support minimum de 1 à 50. Cette boucle prend en entrée les données de benchmark (fichier `zoo.bin`) ainsi qu'un accumulateur destiné à recueillir les résultats de chaque itération. Cet accumulateur consiste en un fichier vide, créé à l'aide de la commande `echo -n`.

Le traitement à l'intérieur de la boucle *for*, sur la partie inférieure de la figure, consiste à recueillir la date d'exécution, extraire les motifs puis mesurer le temps écoulé depuis le recueil de date et enfin compter le nombre de lignes produites. Un simple `echo` assemble ces résultats et l'ajoute à l'accumulateur. Le résultat est visualisé à l'aide d'un script `GnuPlot`.

Références

Clouard, R. (2009). *Pour une ingénierie des connaissances pour le développement d'applications de traitement d'images*. Habilitation à diriger des recherches, Université Caen Basse-Normandie, Caen, France.

Pandore (2013). Pandore : A library of image processing operators (Version 6.6). [Software]. Greyc Laboratory. https ://clouard.users.greyc.fr/Pandore. [accessed July 2013].

Summary

This article is introducing the KD-Ariane platform, a deployment of tools for knowledge discovery processes in the visual environment Ariane. This deployment makes it easier to build structured streams of treatments for discovering knowledge in databases.

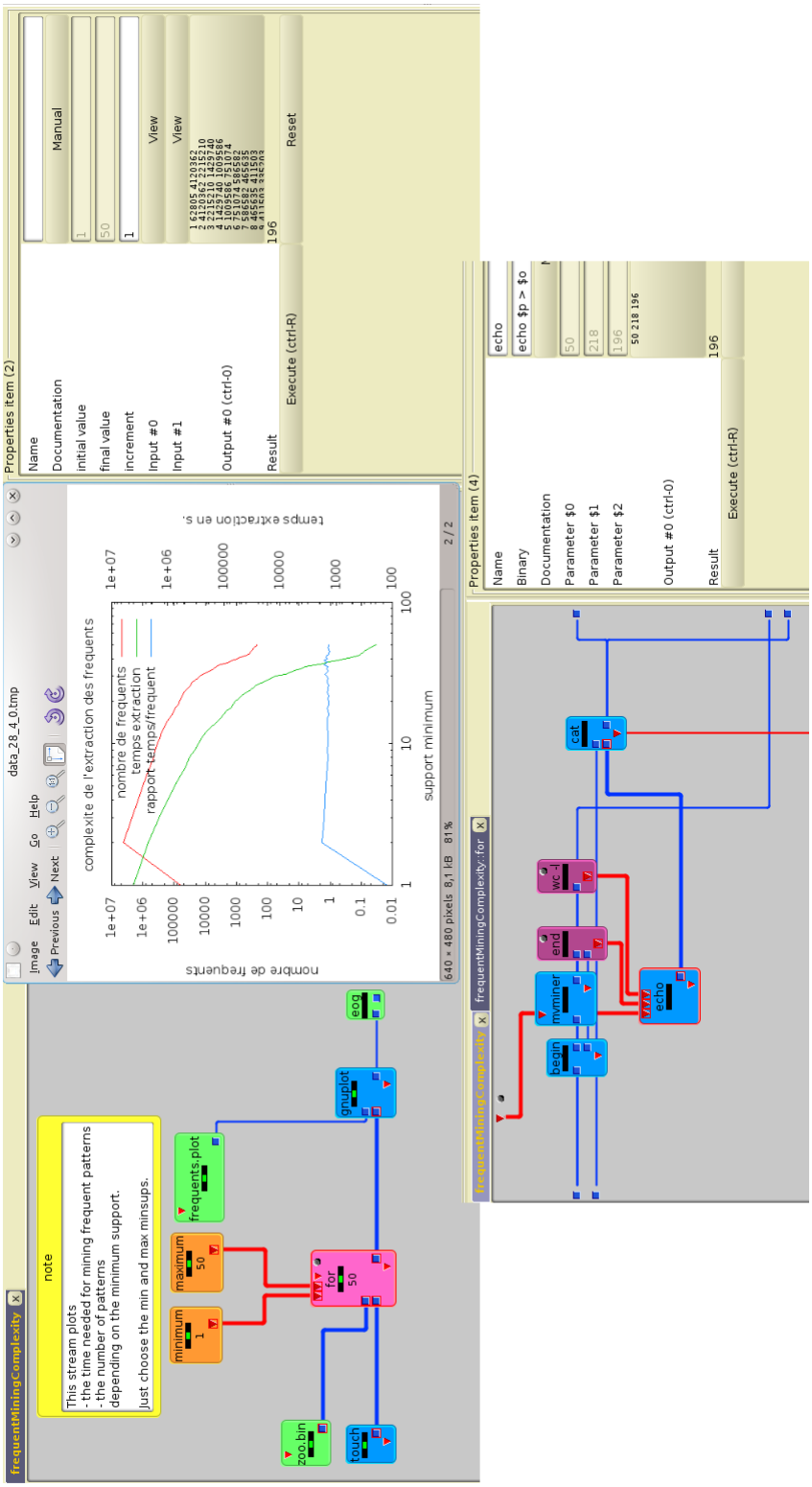

FIG. 2 – *Chaîne structurée pour mesurer la complexité de l'extraction des motifs fréquents*