# Analyzing HCI Issues in Data Clustering Tools

Clodis Boscarioli[1], José Viterbo², Mateus Felipe Teixeira[1],
and Victor Hugo Röhsig[2]

[1] Western Paraná State University (UNIOESTE), Cascavel, Paraná, Brazil
{clodis.boscarioli,mateus.teixeira}@unioeste.br
[2] Federal Fluminense University (UFF), Niterói, Rio de Janeiro, Brazil
{viterbo,vrohsig}@ic.uff.br

**Abstract.** Due to the rapid growth in the volume of data stored in organizational databases and the human limitations in analyzing and interpreting data, appropriate technics are necessary to allow the identification of a large amount of information and knowledge in such databases. In this context, several techniques and tools have been proposed for enabling the end user to interpret his dataset. In this work we discuss the ways of interacting with cluster analysis tools, taking into account both the clustering and the interpretation stages. We investigate how usability and user experience aspects of such tools can improve the understanding of the discovered knowledge. Moreover, we evaluate the role of visualization methods in the comprehension of groups formed in cluster analysis using Knime, Orange Canvas, RapidMiner Studio and Weka data mining tools.

**Keywords:** Data Mining Tools, HCI, User Evaluation.

## 1 Introduction

Due to the rapid growth in the volume of data stored in organizational databases and the human limitations in analyzing and interpreting data, appropriate technics are necessary to allow the identification of a large amount of information and knowledge in such databases. The emerging analytical process called Knowledge Discovery in Databases (KDD), and for [1], is a non-trivial process for discovering valid, new, useful and accessible patterns in databases. KDD comprises three main steps: preprocessing for data preparation, data mining and post-processing, which includes the debugging and/or synthesis of the discovered patterns.

Data Mining is the core of the KDD process. It relies on a set of different algorithms to extract hidden patterns in databases. Such algorithms vary according with the purpose of the analysis, which may be identifying association rules or defining models for data regression, data classification or data clustering. Data clustering, in particular, may be defined as the identification of groups, i.e., subsets of data, in which there is a high internal cohesion among the objects that belong to a group, but also a large external insulation among groups.

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 22–33, 2014.

The cluster analysis process comprises two different stages. In the first stage, one or several algorithms, each using different ways for the identification and representation of its results, may be applied to identify the data clusters. The second stage consists in performing the interpretation of the results, what can be done applying methods based on data visualization. The main purpose of data visualization is to integrate the end user in the knowledge discovery process, providing a graphical representation of the database or the data clustering result. The end user will be able to interpret the results, for example, by identifying characteristics of a particular group, the relationship between patterns and distinctions between groups, spatial distribution of patterns, among others characteristics.

Human-Computer Interaction (HCI) and Data Mining researchers have long been working to develop methods to help end users to identify, extract, visualize and understand useful information extracted from huge masses of high dimensional databases. In this work we discuss the ways of interacting with cluster analysis tools, taking into account both the clustering and the interpretation stages. We investigate how usability and user experience aspects of such tools can improve the understanding of the discovered knowledge. Moreover, we evaluate the role of visualization methods in the comprehension of groups formed in cluster analysis. For this purpose, we selected a set of free and widely used tools: KNIME, Orange Canvas, RapidMiner and Weka.

This paper is organized as follows. Section 2 and Section 3 present some basic concepts about data clustering and data visualization, respectively. Section 4 describes the tools that were selected for our study. Section 5 discusses the HCI evaluations. Finally, in Section 6 we present our conclusions.

## 2 Data Clustering

The goal of data clustering, also known as cluster analysis, is to discover the natural grouping of a set of patterns, points, or objects [2]. At the end of the process, patterns belonging to the same group are more similar to each other and dissimilar to those patterns in other groups, based on a given measure of similarity. The basic idea of grouping data can be defined as the internal cohesion and external isolation of objects between groups [3].

To [4], data clustering is a general name for computational methods that analyze data regarding the discovery of sets of homogeneous observations. Given a database with n patterns, each measured by p variables, the goal of cluster analysis is to find a relationship that separates those patterns in g groups. The final purpose is to find out implicit relationships among data instances that were previously unknown.

There are several data clustering algorithms, each applying different techniques to identify and represent their results. The choice of which algorithm to use depends on the type of data to be analyzed and the purpose of the analysis.

In this work, we chose to use only the k-means algorithm, which was available in all analyzed tools. Even though it was first proposed over 50 years ago [5][6][7], k-means is one of the simplest and still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity [2].

**Algorithm 1.** - *K-means* algorithm

```
Step 1: Select K initial centroids
Step 2: Repeat
Step 3:    Assign each standard to the closest centroid
Step 4:    Recalculate centroids
Step 5: Until the groups remain stable
```

Algorithm 1 illustrates the process. Initially, we define the number k of groups to be formed. After that, we determine the initial k centroids, which can be calculated based on the available patterns, applying one of several different cluster initialization heuristics. For the next step each pattern in the database is associated to the closest centroid, based on a distance measure. After that, the k centroids are recalculated by finding the point that minimizes the average distance for each pattern in the group. This process is repeated until there is no more change in the formed groups.

Typically, k-means is run independently for different values of k and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clustering because k-means only converges to local minima. One way to overcome the local minima is to run the k-means algorithm, for a given k, with multiple different initial partitions and choose the partition with the smallest squared error [2].

## 3 Data Visualization

In a scenario where large amounts of data have to analyzed, the availability of data visualization techniques is particularly important to allow the end user to efficiently interpret the results of a clustering method, for example, by identifying common characteristics in a particular group, the relationship between patterns and the distinction between groups, the spatial distribution of patterns, among others.

The main idea of data visualization is to integrate the end user into the data analysis process, by providing a graphical representation of the database and the resulting data clusters. Besides that, the user can interact with the graphical representation and thus interpret the data clustering results in a more effective way [8].

Visualization is the process of representing data and information in a graphical way, based on visual representations and an interactive mechanism. The purpose of visualization is to give to the user some perception of what is being represented, not only creating a figure. In data clustering operations, the main interest is the visualization of clusters, so that their quality may be assessed, or the spatial distribution of patterns in a cluster can be understood.

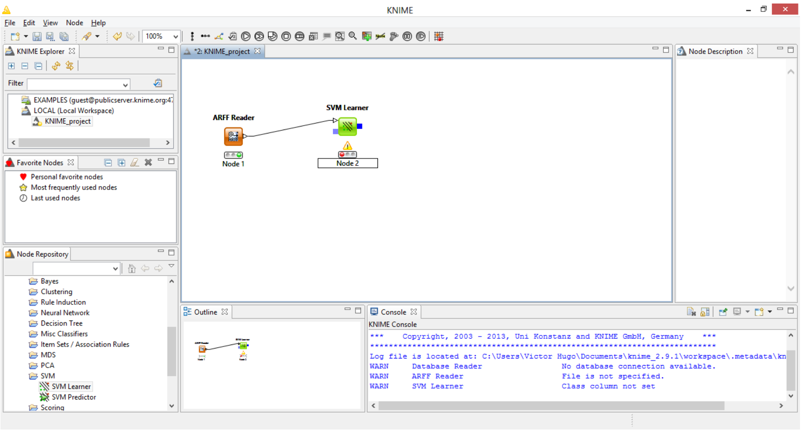
There are different techniques of data visualization, but most are limited by the dimensionality of the database to be explored, along with the dimensionality that can be represented by computational methods. Over time, various techniques have been developed for different types of data and also for different dimensionalities.

According to [9], the techniques of data visualization can be divided into:

- Two dimensions (2D) and three dimensions (3D) visualization techniques, such as pie charts, scatter and line charts, bar charts and cityscapes charts;
- Visualization techniques based on geometric projections such as parallel coordinates and star graph;
- Visualization techniques based on icons or iconographic, such as Chernoff faces;
- Pixel-oriented visualization techniques, such as Circle Segments;
- Hierarchical techniques, such as dendrogram, cone trees and cam trees.

## 4 Data Mining Tools Analyzed

The tools selected for evaluation in this study are briefly described ahead. For each one, we first identified the available clustering techniques and cluster visualization methods. Knime [10] is a tool proposed for use in data mining, statistics and other areas. It has several methods that enable full knowledge extraction of a particular database. All the operation of Knime is based on the idea of adding method nodes to an execution workflow. Figure 1 shows Knime screen. In the upper left corner, the user can see the available methods. In the center, there is a window where the execution workflow is defined. In the bottom, the results are shown in textual format in a window. Some features are not explicitly presented, which can hinder their use. For k-means, Knime presents only the description of the centroids as textual output, with values that each feature and how many standards this centroid comprises.



```
***    Copyright, 2003 - 2013, Uni Konstanz and KNIME GmbH, Germany    ***
***************************************************************************
Log file is located at: C:\Users\Victor Hugo\Documents\knime_2.9.1\workspace\.metadata\kn
WARN     Database Reader                         No database connection available.
WARN     ARFF Reader                             File is not specified.
WARN     SVM Learner                             Class column not set
```

**Fig. 1.** KNIME Clustering Interface Example

Orange Canvas [11] is an open source tool for data mining with a focus on data classification, data regression and visual data mining. It also has data evaluation and data binding methods. The execution workflow of the tool is simple. Figure 2 shows Orange Canvas screen. As knime, it presents the structure of nodes, where each node added to the workflow canvas will perform a certain task. It also features a feedback system for each method, showing the input and output of each method.

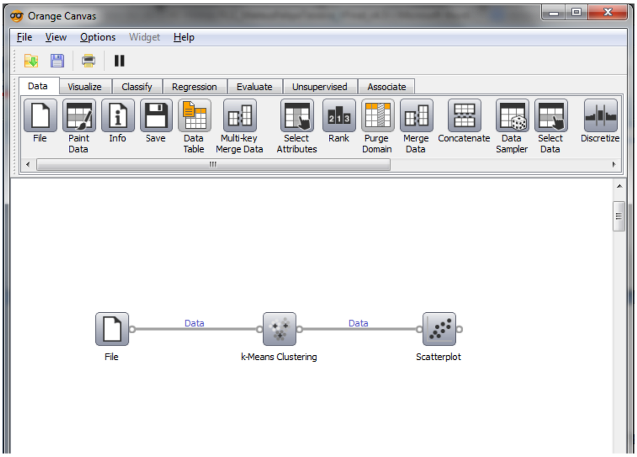
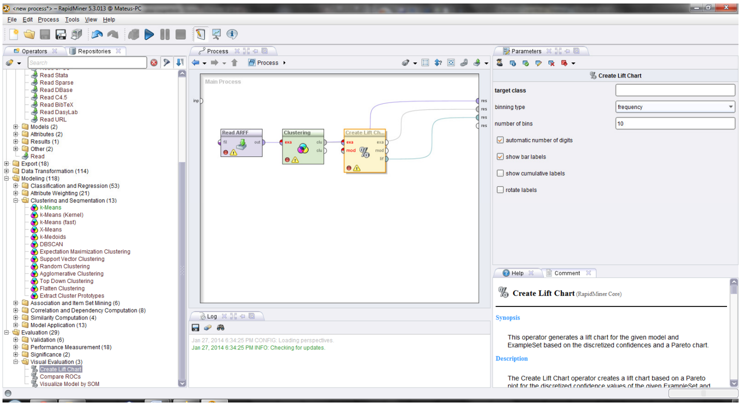


**Fig. 2.** ORANGE CANVAS Clustering Interface Example

RapidMiner Studio [12] is a data mining tool also focused on statistical, database and data analysis processes. It is a proprietary tool but there is also a trial version. Its main purpose is to provide a fully graphical desktop environment, with graphic elements that represent an operation of interest, for example, a data mining method. Figure 3 shows RapidMiner screen. The tool has a very intuitive execution workflow, where nodes represent a particular process. The output is textual, showing the number of groups formed and how these patterns are formed, and also a data table.



Jan 27, 2014 6:34:25 PM CONFIG: Loading perspectives.
Jan 27, 2014 6:34:25 PM INFO: Checking for updates.

target class
binning type: frequency
number of bins: 10
☑ automatic number of digits
☑ show bar labels
□ show cumulative labels
□ rotate labels

Create Lift Chart (RapidMiner Core)

Synopsis

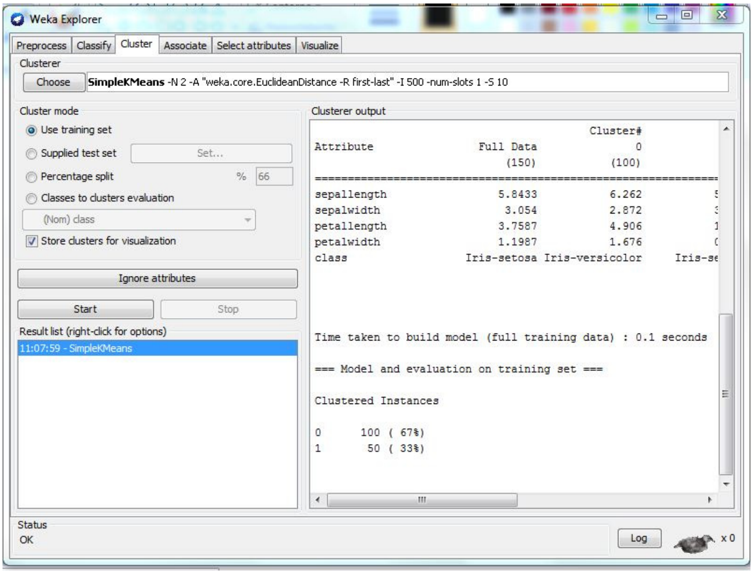
This operator generates a lift chart for the given model and ExampleSet based on the discretized confidences and a Pareto chart.

Description

The Create Lift Chart operator creates a lift chart based on a Pareto plot for the discretized confidence values of the given ExampleSet and

**Fig. 3.** RAPIDMINER Studio Clustering Interface Example

Weka [13] is an open source collection of machine learning algorithms for data mining tasks that contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. As a result of the execution of the k-means method, Weka shows the number of the iteration, the sum of the squared error within groups and the centroids formed. Weka technique of dispersion chart of does not show the data dispersion of a particular group, but the data dispersion according to one of its features depending on the group it belongs. Figure 4 shows Weka screen.



SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- ☑ Store clusters for visualization

11:07:59 - SimpleKMeans

```
                                   Cluster#
Attribute          Full Data              0
                       (150)          (100)
==============================================================
sepallength           5.8433          6.262           5
sepalwidth             3.054          2.872           3
petallength           3.7587          4.906           1
petalwidth            1.1987          1.676           0
class            Iris-setosa Iris-versicolor     Iris-se




Time taken to build model (full training data) : 0.1 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       100 ( 67%)
1        50 ( 33%)
```

Status
OK

**Fig. 4.** WEKA Clustering Interface Example

Table 1 shows the clustering methods and data visualization available in the evaluated tools. It is noteworthy that in the evaluation and accounting of the number of methods available, only basic methods were considered. For example, k-means variants present in some tools, such as fast k-means, were not counted as another data clustering method. For the hierarchical methods, each connection was not considered as a method, but the presence of the function was, since the hierarchical links were regarded as execution parameters in all tools.

**Table 1.** Data mining tools and methods of data clustering and data mining

| Tool | Characteristics |
|---|---|
| **Knime** | **URL**: http://www.knime.org/ **Version**: 2.7.2 |
| | **Data clustering methods:** 5 |
| | Fuzzy c-Means, hierarchical methods (single-linkage, complete-linkage and average-linkage), SOTA – (Self Organizing Tree Algorithm), Learner e Predictor and k-means. |
| | **Data visualization methods:** 8 |
| | Box plot, histogram, lift chart, line plot, parallel coordinates, pie chart, scatter plot matrix and dispersion chart. |
| **Orange Canvas** | **URL**: http://orange.biolab.si/ **Version** : 2.6.1 |
| | **Data clustering methods:** 2 |
| | k-means, hierarchical methods (single-linkage, complete-linkage, average-linkage and ward) |
| | **Data visualization methods:** 12 |
| | frequencies distribution, box plot, general dispersion, linear projection, radviz, polyviz, parallel coordinates , survey plot, correlation analysis, mosaic display, sieve diagram, sieve multigram. |

**Table 1.** (*continued*)

| | |
|---|---|
| **RapidMiner** | **URL**: rapidminer.com/products/rapidminer-studio/ **Version** : 5.3.015 |
| | **Data clustering methods:** 9 |
| | k-means, k-medoids, DBSCAN, Expectation Maximization Clustering, Support Vector Clustering, Random Clustering, hierarchical methods (single-linkage, complete-linkage and average-linkage), Top Down Clustering and Flatten Clustering. |
| | **Data visualization methods:** 3 |
| | Lift chart, ROC curves and model visualization by self-organizing maps. |
| **Weka** | **URL**: http://www.cs.waikato.ac.nz/ml/weka/ **Version** : 3.7.8 |
| | **Data clustering methods:** 7 |
| | Cobweb, Expectation Maximization, Farthest First, Filtered Clusterer, hierarchical methods (single-linkage, complete-linkage, average-linkage, mean-linkage, centroid-linkage, ward, adjcomplete and neighbor-joining), Make Density Based Clusterer, k-means. |
| | **Data visualization methods:** 4 |
| | Histogram, dispersion chart, scatter matrix and tree visualization. |

## 5 HCI Evaluations

In order to evaluate how easy a user can learn to use each of these tools, we performed an evaluation by inspection using the cognitive walkthrough method [14], an IHC evaluation method whose primary purpose is to evaluate the ease of learning of an interactive system through the usage of its interface. As such, we developed a real use scenario, based on which usability tests were performed by a group of users.

The users' profile we defined for this inspection comprised IT professionals or students, who have interest in performing data mining tasks and already have at least a minimal prior contact with such area. Two different types of users performed the usability tests: (i) six graduate and undergraduate computer science students that attended data mining classes; (ii) five lecturers in the area of artificial intelligence, data mining or, statistics. For each tool, each user should execute the following steps:

1. Load a test file ("Iris.arff", obtained from [15])
2. Select, as the task to be performed, data clustering using k-means with $k = 3$;
3. Execute the data clustering operation;
4. Visualize the results.

After executing these tasks, the users answered questionnaires, providing data on their experience about using these tools. The questions, enumerated as follows, were formulated to assess the user profile (1 and 2), the organization of the tool's interfaces (3 to 6) and the usability and user's interaction with the tool (7 to 10).

1. How do you rate your knowledge on data mining?
2. How long have you used these systems?
3. The information available in the system's interface is well distributed, so as to contribute to the user's learning and memorizing?
4. The icons and control commands are well detached from other interface items?
5. Does the system's interface describes the task options offered, i.e., given a particular icon or menu, is there a brief specification of its functionality?
6. In the error messages, the help button is available?
7. This system is easy to use ?
8. Did you have some sort of difficulty in finding the information necessary to perform the requested tasks?
9. How do you evaluate the presentation of the data clustering results?
10. Did the available data visualization techniques help in the interpretation of the generated clusters?

### 5.1 General Overview

In order to execute the tasks included in the usability test, the user must identify the buttons or menu options for (a) specifying the path and name of the input file and loading the data; (b) selecting the k-means method, setting the necessary parameters and executing it; (c) selecting visualization methods and creating charts.

Although the pathways to accomplish the actions are different in each tool, we can conclude that the user will try to achieve the right outcome, since he knows the theoretical aspects of the implemented methods in order to perform the parameter settings. He must as well know the contents of the database to be analyzed, to be able to perform a semantic verification of the results obtained.

Moreover, among the tools there are particularities of interaction. In Knime or RapidMiner Studio, for example, the user needs to know in which category is classified each item he needs to perform the desired task. Each item/node (such as, loading database, clustering method and visualization method) is divided into subcategories, which should be known by the user. In such cases, the experience with any other tool or even some data mining theoretical knowledge will help the user to make an association between the interface elements and his objective.

In Weka, for action to open/load the database may get errors in identifying the tool component. For example, in some places, the component is called "ARFF Reader" and in others, it is available as "FILE". The idea that this component can also open/load a file with extension "ARFF" is implicit. Also on Weka, for choosing a data clustering or data visualization method, the user must know which methods to use and in which part of the tool interface he can find the desired items.

If the user does not know a priori the workflow necessary for carrying out his data mining tasks, he must learn while performing his activities. In Knime, RapidMiner Studio and Orange Canvas, the user is allowed to interact with the tool to build a workflow even when a node is being wrongly defined. These tools issue a warning to the user about the error only at the end, when he tries to execute the data mining process workflow. This behavior causes the user to lose time in his interaction with the tool, probably forcing the user to rebuild his workflow, entirely or in part, for not knowing exactly where the error occurred. A possible solution for this problem would be having the tool verifying the workflow node by node, while the user is building it, in order to identify possible errors and show alerts to the user as soon as possible.

### 5.2 Results Analysis

The usability test aims to assess how easily the user understands the usage of each selected interacting with its interface. The objectives of the usability test were defined and presented to the users, who, using the tools, should try to reach them. The users were divided in a group of lecturers, with a better understanding of data mining and greater experience with the tools, and a group of students with only basic knowledge, as depicted on Figure 5. The charts show that most users are well acquainted to Weka, while many never used Knime or Orange Canvas.

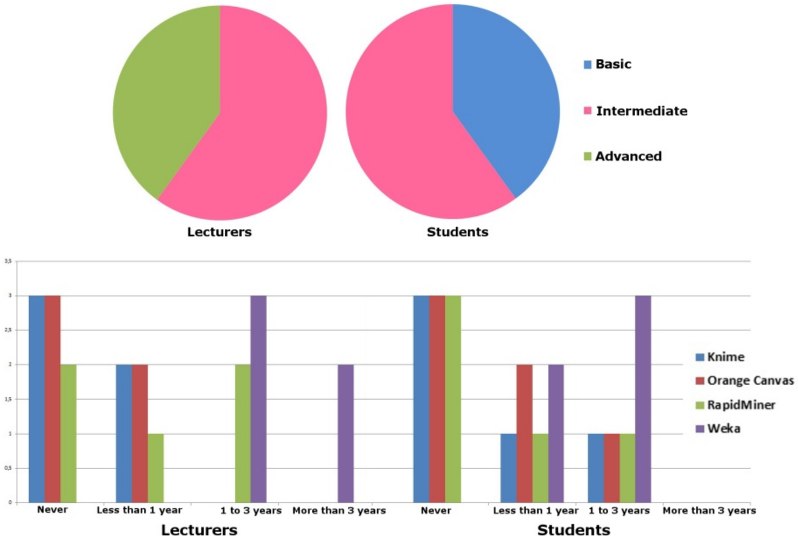


**Fig. 5.** Charts representing the level of expertise (top) and familiarity with the tools (bottom) of the two different user groups

Regarding the interface, both groups of users thought that the tools have the information well distributed, with the icons are well detached, what contributes to an easy understanding of the interaction process, as depicted in Figure 6 (a). Except for RapidMiner Studio, that has not been positively evaluated by the group of students with regard to this aspect. The reason may be that the set of objects needed to perform a task is large and complexly organized, requiring prior knowledge about these features, which was only found in the group of lecturers.

Orange Canvas and RapidMiner were poorly evaluated by the lecturer with regard to the description of the tasks offered by the interface, i.e., such users considered that the interface of those tools tools did not provide enough explanation on the functionalities provided, as depicted in Figure 6 (c). The tools also had poor rating with respect to the descriptions of the elements contained in the interface and help options, as depicted in Figure 6 (d).

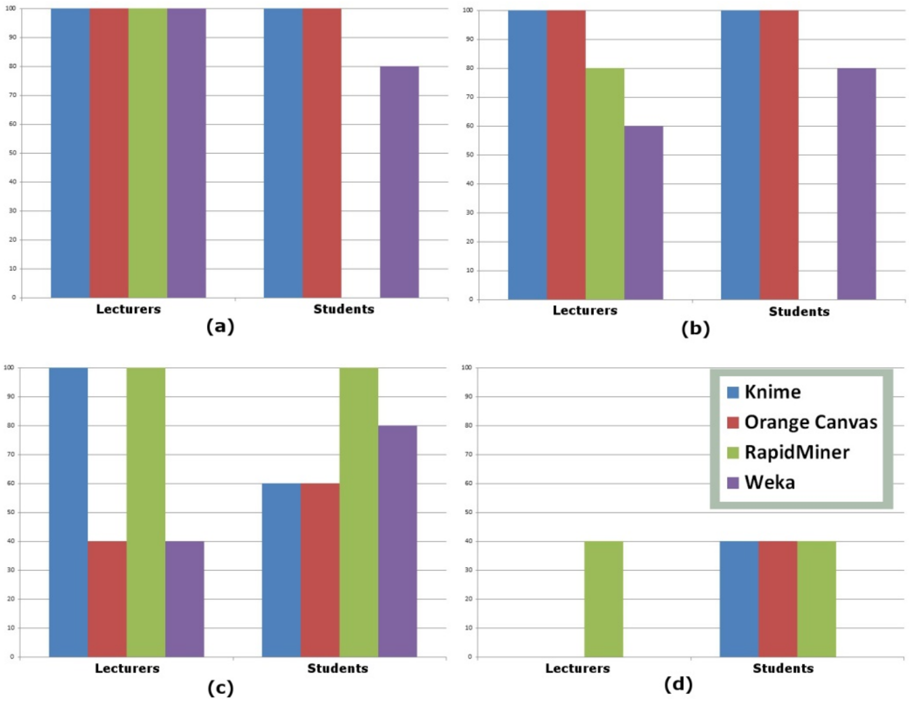


**Fig. 6.** Charts representing the proportion of users that evaluate positively the distribution of information in the interface (a), the emphasis in icons and commands in the interface (b), the description of the tasks offered by the interface (c), the availability of error messages (d)

According to the students, the Knime is the tool that presents the interface in which is easier to find information and the desired elements, while RapidMiner was the most negatively evaluated by this group, as depicted in Figure 7 (a). Among the group of lecturers, Weka was considered the best one, probably because in general all those users had great familiarity with the tool, as indicated in Figure 5.

As to the presentation of the data clustering results, Knime got the worst evaluation reported by the group of lecturers, but a good result among the students, as depicted in Figure 7. Using colors and providing the confusion matrix for presentation, were some of the improvement suggestions cited by the evaluators. On the other hand, Orange got a great result according to the students, while the lecturers found the performance the performance of the tool for visualizing the results is satisfactory. This is probably due to ease of visualization of the results provided by the tool, but that are somehow superficial. The other tools had a regular a performance.

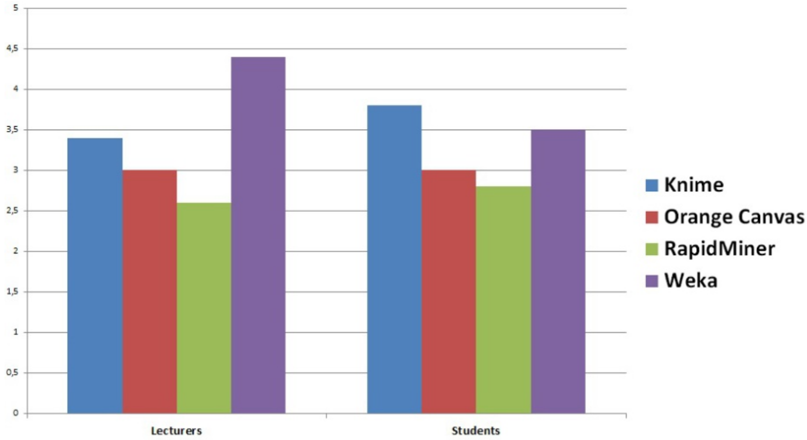


**Fig. 7.** Charts representing the user's average evaluation on how easy each tool

# 6 Conclusion

In general, we noticed that the analyzed tools have a very strong approach towards data visualization. Many of the tools discussed in this work enable the user to view the groups generated by the clustering methods, understanding, interpreting and extracting knowledge about the set of patterns and clusters.

Generally, the usability, ease of operation and the methods of each tool are good. In particular, the utilization of each tool, show a very clear workflow, having beginning, middle and end. As such, these tools may be used by people with any level of knowledge about them, or even about the methods provided by these tools.

We believe that the interdisciplinary integration of KDD and HCI may bring some significant contributions, such as improving the user's ability to gain useful knowledge from organizational databases and helping the end users to gain added values by making date useable and useful. However, the tools analyzed show that still several other HCI aspects need to be thoroughly approached.

# References

1. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthrusamy, R.: Advances in knowledge Discovery & Data Mining, California (1996)
2. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31(8), 651–666 (2010)
3. Cormack, R.M.: A Review of Classifications. JRSS, A 134, 321–367 (1971)
4. Everitt, B.S., Landau, S., Morven, L.: Cluster Analysis, 4th edn. Hodder Arnold Publishers, Londres (2001)
5. Macqueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, California (1967)

6. Wong, P.C.: Visual Data Mining. In: IEEE Computer Graphics and Applications, Los Alamitos (1999)
7. Ball, G., Hall, D.: ISODATA, a novel method of data anlysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA (1965)
8. Steinhaus, H.: Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci. IV(C1.III), 801–804 (1956)
9. Keim, D., Ward, M.: Visual Data Mining Techniques. Intelligent Data Analysis: An Introduction. University of Konstanz, Worcester Polytechnic Institute, USA (2002)
10. KNIME, Site Oficial da Ferramenta KNIME (2013), `http://www.knime.org/` (Acesso September 25, 2013)
11. CANVAS, Site Oficial da Ferramenta ORANGE CANVAS (September 09, 2013), `http://orange.biolab.si/` (acesso September 25, 2013)
12. RAPIDMINER, Site Oficial da Ferramenta RAPIDMINER STUDIO (2013), `http://rapidminer.com/products/rapidminer-studio/` (acesso November 17, 2013)
13. WEKA, Waikato Environment for Knowledge Analysis (2013), `http://www.cs.waikato.ac.nz/ml/weka/` (acesso September 25, 2013)
14. Whartonm, C., Rieman, J., Lewis, C., Poison, P.: The Cognitive Walkthrough Method: A Practitioner's Guide. Usability Inspection Methods, New York (1994)
15. Bache, K., Lichman, M.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2013)
16. Keim, D.A., Kriegel, H.: Visualization Techniques for Mining Large Databases: A Comparison. IEEE Trans. Knowledge & Data Engineering, 923–936 (1996)
17. Keim, D.A., Kriegel, H.P.: VisDB: Database Exploration using Multidimensional Visualization. IEEE Computer Graphics and Applications (1994)
18. Keim, D.A.: Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computers Graphics 8, 1–8 (2002)