
PREPOS ENVIRONMENT: A SIMPLE TOOL FOR DISCOVERING INTERESTING KNOWLEDGE

Cristian Simioni Milani, Deborah Ribeiro Carvalho

Pontifícia Universidade Católica do Paraná (PUC-PR) Curitiba – PR – Brazil

cristian.milani@pucpr.br, ribeiro.carvalho@pucpr.br

Abstract: The major limiting factors in the use of the KDD process are the operational difficulties in the preparation of data and in the analysis of the discovered patterns, difficulties arising from the skills required for processing in different environments, the way data are available, time required for the process, among others. The KDD process could be better use if there were a greater number of tools available. This article is aimed to provide the community with an environment – PREPOS – for support in the pre and post-processing steps, integrated with data mining algorithms, that is, it integrates the three steps of the KDD process. Thus, the environment proposed and developed provides a contribution to the KDD area, because besides the new integrated features made available, the new environment was implemented in open source platform, i.e., new features can be added to the structure built.

Keywords: Data mining, KDD, Pre-processing, Post-processing.

1. INTRODUCTION

The search for strategies for KDD - Knowledge Discovery in Databases is an area of computer science that has been growing fast due to the constant need for tools to support decision-making. The KDD process facilitates the potentiation of use of the available databases, which increase the size of stored data, providing an opportunity for the discovery of patterns, trends and interesting information.

The use of knowledge extraction processes provides managers with new strategies that do not depend on the prior establishment of assumptions, such as the salesperson who achieved the highest number of sales, the amount sold, etc. So, they move to more complex questions such as: “Which users are more likely to become delinquent?”. Thus, the KDD process plays a key role by complementing the so-called traditional decision support systems based on the establishment of assumptions.

The KDD process is characterized as the “process of identifying valid, novel potentially useful and ultimately understandable patterns in data” (Fayyad et al., 1996). It is composed of three steps: pre-processing, data mining and post-processing.

The pre-processing step is usually the one that demands more efforts because besides the fact that databases are not formatted according to the specific requirements, several changes may be needed, such as (Rezende, 2005):

- Denormalization: consists in organizing all information of a given event in a unique record;
- Transformation: when there are limitations regarding the types of data requiring conversion to another specific type of data;
- Cleaning: elimination of noises or inconsistencies;

- Selection of attributes: properties or attributes that do not contribute to the knowledge discovery process are eliminated.

In the data mining step the task is selected according to the nature of the management problem posed, including: discoveries of association rules, classification and clustering.

The association rules are characterized by the structure $X \rightarrow Y$, if X then Y, where X and Y represent sets of data items. The classification task is aimed to discover and represent a predictive model. Such model can be represented in various ways, including, decision trees, which are easily transformed into IF-THEN rules. Finally, the clustering step identifies groups according to similarities/dissimilarities regarding the data items. Software selection is also comprised in this task, which, in turn, is related to the management problem.

Post-processing is aimed to facilitate the evaluation of patterns discovered by the manager that have greater potential to be interesting. This step is particularly necessary when a large set of patterns is discovered, and plays a key role by increasing the probability that KDD process discovers elements that add value to the managers' knowledge.

Among the several post-processing strategies, it is worth mentioning the allocation of interest measures to the patterns. These allocated measures can be of two natures: objective and subjective. The objective measures are usually statistical measures, calculated based on the data and structure of patterns. Prior knowledge and the objectives of the user are somehow necessary for the subjective measures. The interest measures can be allocated to association rules or even to those originated from the transformation of decision trees.

Because of all of these steps, as well as each of their respective peculiarities and requirements, the KDD process is a non-trivial process, and, so, the assistance of IT experts is necessary. Thus, the scientific community is doing research on new alternatives that facilitate the operation of knowledge extraction.

The KDD process could be better used if there were a greater number of facilitating tools available. This article is aimed to provide the community with an environment – PREPOS – for support in the pre and post-processing steps, which are integrated with data mining algorithms, that is, they integrate the three steps of the KDD process.

Such a proposal is justified for many reasons. The pre-processing step usually involves operations that can only be performed by IT experts, such as preparing files in standard text, with no separation between the fields (non-standard CSV- Comma-separated values). This is a very common situation in microdata provided by research institutions, including the IBGE (Brazilian Institute of Geography and Statistics).

Not to mention the post-processing alternatives, which besides the restricted availability, it is difficult to obtain portable patterns due to the large variety of formats. Thus, the PREPOS environment allows the user to perform all steps in a single environment, preventing the need for additional efforts to reconcile different environments and formats.

The PREPOS environment integrates two task algorithms. For classification, the algorithm J48 of the WEKA environment (Hall et al., 2009) and the C4.5 (Quinlan, 1993). The PREPOS offers the APRIORI algorithm for the discovery of the association rules, also from the WEKA environment, and the algorithm offered by Borgelt (2004). Four features are available for post-processing.

2. METHODS

JAVA language was chosen for the implementation of the tool and the use of API in WEKA environment (Hall et al, 2009) for the construction of some steps of the tool. It should be stressed that the strategy for construction of the tool allows the incorporation of new features or even other algorithms for data mining.

The general idea of the tool is to integrate all steps of the KDD process into a single environment. Thus, the features were grouped in two large sets. Data Mining and Utilities. The Data Mining group provides two task algorithms (discovery of association rules and classification), with two alternatives for each one, besides pre and post-processing features. In this first set, entry occurs from a single database. In Utilities, there are some procedures that can also be characterized as pre-processing, but which differ from other features because they need more than one database. The algorithms and features available in the PREPOS tool are listed in chart 1.

Chart 1. Algorithms and features available in the PREPOS tool.

Pre-processing	<ul style="list-style-type: none"> • Denormalize database; • Label data items for the APRIORI logarithm (Borgelt, 2004).
Data Mining	<ul style="list-style-type: none"> • APRIORI (Borgelt, 2004); • APRIORI of WEKA environment (Hall et al., 2009); • J48 of WEKA environment (Hall et al., 2009); • C4.5 (Quinlan, 1993);
Post-processing	<ul style="list-style-type: none"> • PAD (Simioni and Carvalho, 2013). • Filter for association rules; • Discovery of exception rules; • User-driven (the user specifies in advance in the environment what is interesting);

Two features were developed for the pre-processing step: denormalize and label data items for the APRIORI algorithm. Denormalize is essential, because the data often originate from the SGBDs – Management Database System, usually normalized. An example of denormalization can be seen in charts 2 and 3. Chart 2 shows an example of set of normalized data from one patient, and in chart 3, the same set is denormalized.

Chart 2. Normalized Data.

Patient Identifier	Procedure Performed
1	Consultation
1	Hospitalization
1	Scintigraphy
2	Emergency Consultation

Chart 3. Denormalized Data.

Patient Identifier	Procedures Performed
1	Consultation; Hospitalization; Scintigraphy
2	Emergency Consultation.

The pseudocode of the feature that performs database denormalization is shown in Chart 4. The parameters adopted are:

- Instances: instances of the database;
- i: index of the identifier attribute in the database;
- t: index of the item do be denormalized.

Chart 4. Pseudocode of the feature Denormalize Data.

```

function denormalize (instances, i, t)
begin
  k := 1;
  j := 1;
  R;
  v := distinctvalues(instances, i, t);      (* the result set *)
  while k ≤ v.lenght do begin
    R := append(vk, R);
    k := k + 1;
  end
  return R;
end
    
```

The APRIORI algorithm (Borgelt, 2004) does not work with structured databases or else, it requires that all individual data items in the set are identified. Thus, for situations in which the set of data to be mined contain similar domain values for different attributes, these must be individualized. One alternative is that the identifier is adopted as a label, that is, “label_domainvalue”. Chart 5 shows a variable representing climate conditions with their possible values and their respective transformed values.

Chart 5. Climate variable and its due transformation.

Climate	Labeled value
Sun	climate_sun
Rain	climate_rain
Cloudy	climate_cloudy

The pseudocode of the feature that labels data items is shown in chart 6. The parameters are:

- instances: instances of databases;
- attributes: attributes of databases;

Chart 6. Pseudocode of the Feature of Preparation of Database for the APRIORI algorithm

```

function preparator (instances, attributes)
begin
  k := 1;
  j := 1;
  R;                                     (* the result set *)
  while k ≤ instances.lenght do begin
    j := 1;
    while j ≤ attributes.lenght do begin
      R := R + append(instancesk, attributesj);

      j := j + 1;
    k := k + 1;
    end
  return R;
end

```

Although the algorithms for data mining are well disseminated, they were provided in the proposed tool. Such incorporation is aimed to ensure the integration of the three steps and to facilitate the use of post-processing features.

In order to post-process classifiers discovered by algorithms C4.5 (Quinlan, 1993) and J48 (Hall et al., 2009) the post-processing feature of Decision Trees – PAD was made available (Simioni e Carvalho, 2013), not only transforming decision trees in their corresponding set of rules. Also eliminating redundancies among the conditions that compose a given rule antecedent of the rules extracted, additionally assigning to each rule transformed the respective measure of interest based on multiple generalizations.

For post-processing patterns in the format of association rules discovered by the Apriori algorithm (Borgelt, 2004) and by the APRIORI of the WEKA environment (Hall et al., 2009), the following features were provided: Filter for association rules, extraction of patterns using the technique user-driven and Discovery of Exception Rules (DRE).

The main purpose of the feature Filter for Association Rules is to reduce the set of rules, facilitating not only the respective analysis, but also the processing time when this set of patterns is submitted to another post-processing feature. “Parameterization of the filtering process is made from the identifiers of data items that compose the rules”. (Simioni e Carvalho, 2013).

Discovery of Exception Rules (DRE) was provided because according to Simioni & Carvalho (2013) an exception rule is a specialization of a general rule, and tends to be more interesting to support the decision process than the general rule over which the exception is built.

In the “User-driven” procedure, the user carries out a thorough analysis of each rule, indicating which rules are considered interesting and which are not. Thus, the user guides the selection and elimination of rules. This module does not yet include the

calculation of subjective measures. However, as previously mentioned, new features can be added to the tool. Subjective measures such as: conformity, unexpected antecedent, unexpected consequent and unexpected antecedent and consequent (Sinoara, 2006), could be coupled to the environment.

In the second group – Utilities – a feature aimed to process files made available without any separator was implemented. Some examples include microdata from field research provided by the IBGE - Brazilian Institute of Geography and Statistics, e.g. PNAD 2004 and 2009 (National Household Sample Survey), POF 2008 (Household Budget Survey) and Census 2010. To this end, the dictionary of variables shown in chart 7 must be built and delivered. This dictionary is composed primarily of three elements: database name, name of the table to be created in the SGBD, variables and their respective initial and final positions, as well as typing.

Chart 7. Segment of the dictionary of variables for the PNAD 2004 database.

```
PNAD_2004
DOMICILIO
V0101;0;3;BIGINT
UF;4;5;BIGINT
V0102;4;11;BIGINT
V0103;12;14;BIGINT
V0104;15;16;BIGINT
V0105;17;18;BIGINT
(...)
```

After the development of the dictionary (chart7) transformation from TXT to CSV is performed, besides the scripts for creating the table and importing from SGBD into MySQL. After this import, the set of data will be available for several uses, including to be adopted as entry for the PREPOS tool, for KDD activities.

3. RESULTS AND DISCUSSIONS

Figure 1 shows the home screen of the tool, from which the user may select Data Mining or Utilities.



Figure 1. Home screen of the PREPOS tool.

In the Data Mining group, the user has the pre and post-processing features, as well as the algorithms that discover patterns. In figure 2, there is an example of navigation through the options of the Data Mining group, considering the database “*weather.nominal*”, provided by the WEKA environment (Hall et al., 2009).

The Pre-processing, Data Mining and Post-processing tabs become available when the database is loaded.

In order to demonstrate some results obtained with the referred tool, the database “*breast-cancer*” (Zwitter, 1998) was used. This database contains 286 instances and 10 attributes. This database was first prepared for the APRIORI algorithm (Figure 3).

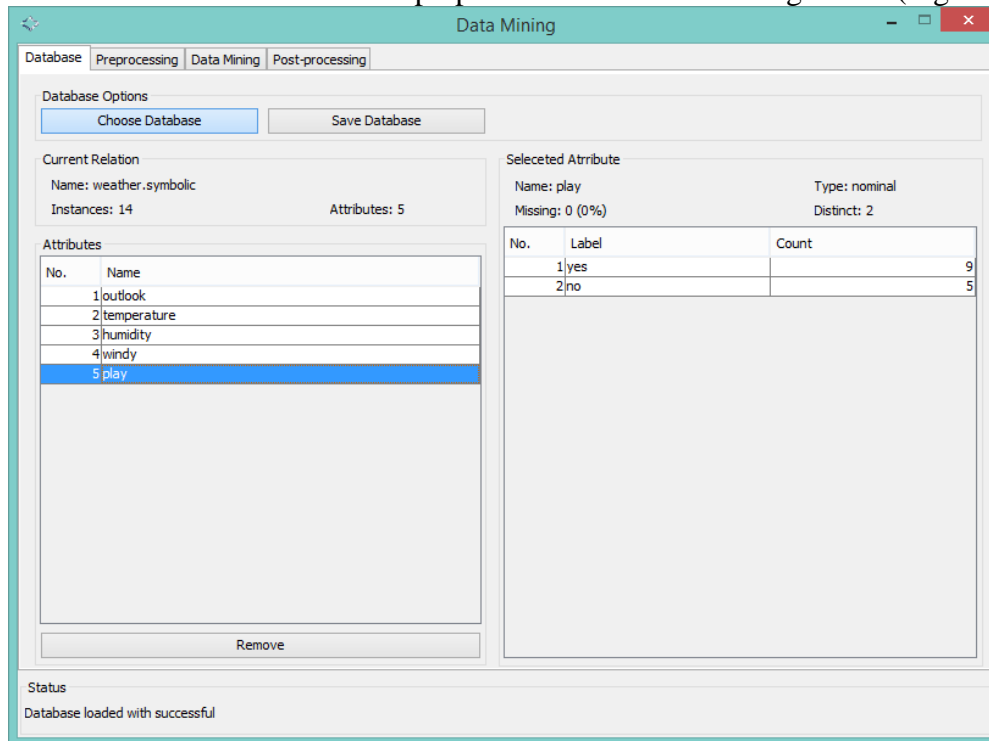


Figure 2. Database selection screen.

Another point to be stressed is that it is easy for users to extract knowledge from the various algorithms in a single interface. The environment in its initial version has four algorithms, two of them executable and two API calls of WEKA environment. That is, without the tool three different computer environments are required to perform this process: a different entry for each executable and one entry for the WEKA environment. Figure 4 shows examples of interface where algorithms for data mining are available.

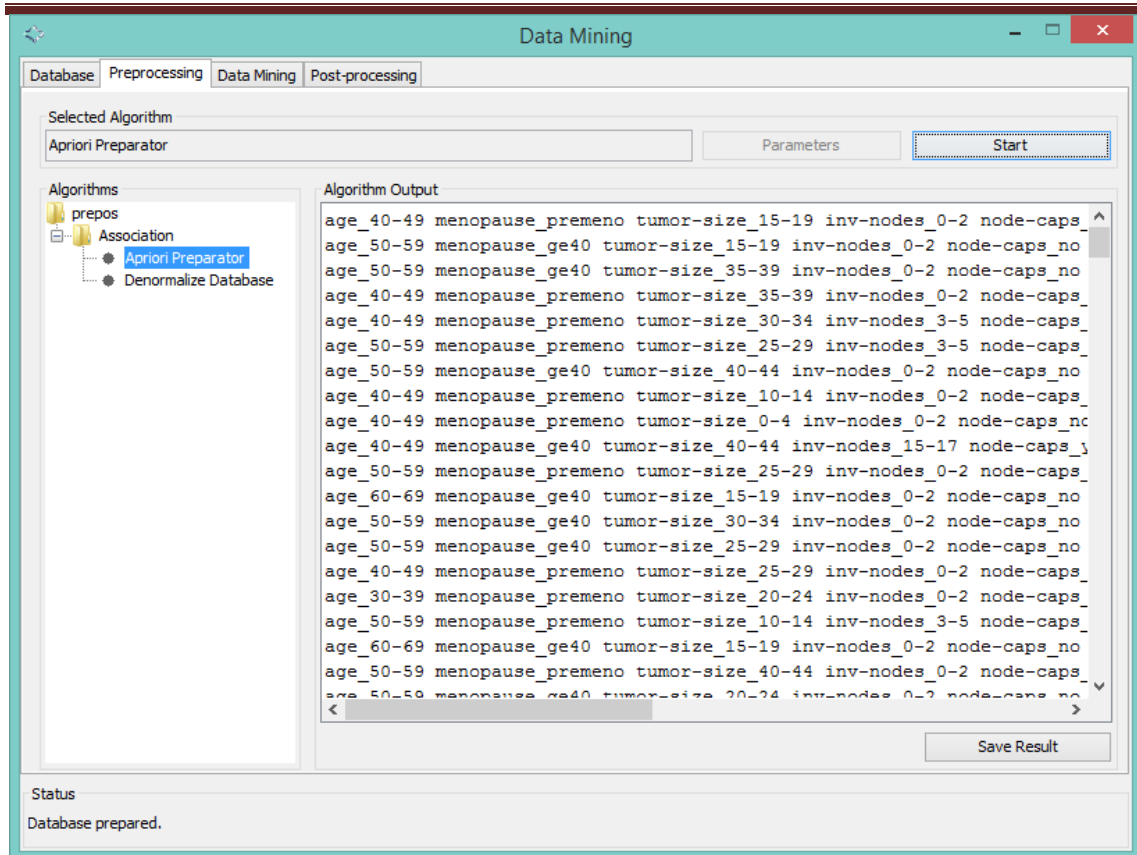


Figure 3. Example of feature for preparation of databases for the APRIORI algorithm.

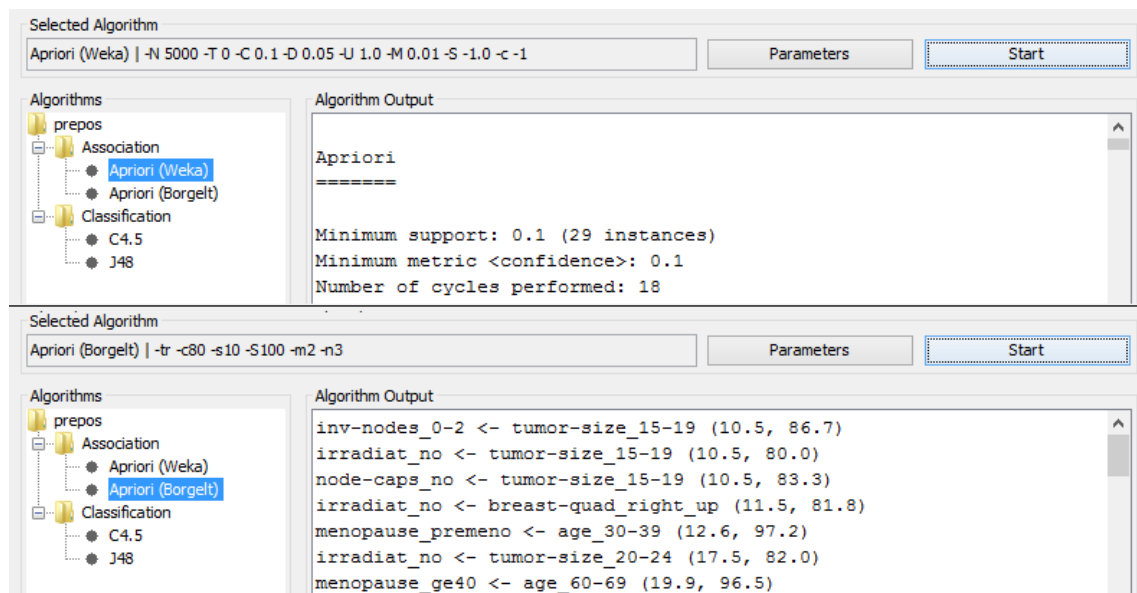


Figure 4. Examples of the choice of data mining algorithms.

As shown in Figure 4, by simply changing the selection of the algorithm tab, the user is able to interact with all available implementations in a single window, i.e., the environment becomes homogeneous bringing together the former three separate environments into only one environment.

If the algorithm chosen is discovery of classifiers represented in the form of a decision tree (Figure 5), the user may transform this tree into production rules (Figure 6)

Simplified Decision Tree:

```
deg-malig in {1,2}:
| inv-nodes = 15-17: no-recurrence-events (1.0/0.8)
| inv-nodes in {12-14,18-20,21-23,24-26,27-29,30-32,33-35,36-39}: no-recurrence-events (0.0)
| inv-nodes in {0-2,3-5,6-8}:
| | inv-nodes = 6-8: no-recurrence-events (7.0/2.4)
| | inv-nodes in {0-2,3-5}:
| | | breast-quad in {left_up,left_low,central}: no-recurrence-events (151.0/33.0)
| | | breast-quad = right_low: no-recurrence-events (17.0/1.3)
| | | breast-quad = right_up:
| | | | tumor-size in {25-29,45-49}: recurrence-events (4.0/1.2)
| | | | tumor-size in {0-4,5-9,10-14,55-59}: no-recurrence-events (2.0/1.0)
| | | | tumor-size in {15-19,20-24,30-34,35-39,40-44,50-54}:
| | | | | node-caps = yes: recurrence-events (3.0/1.1)
| | | | | node-caps = no: no-recurrence-events (11.0/1.3)
```

Figure 5. Segment of the discovered decision tree.

Production Rules:

```
deg-malig>>{1,2} inv-nodes>>{0-2,3-5,6-8} inv-nodes=6-8 -> class=no-recurrence-events (7.0, 2.4)
deg-malig>>{1,2} inv-nodes=9-11 age>>{30-39,50-59,70-79} -> class=recurrence-events (3.0, 1.1)
deg-malig>>{1,2} inv-nodes=9-11 age=40-49 -> class=no-recurrence-events (2.0, 1.0)
deg-malig>>{1,2} inv-nodes=15-17 -> class=no-recurrence-events (1.0, 0.8)
deg-malig=3 tumor-size>>{0-4,45-49} -> class=no-recurrence-events (2.0, 1.0)
deg-malig=3 tumor-size=50-54 -> class=recurrence-events (1.0, 0.8)
```

Figure 6. Subset of production rules extracted from the decision tree.

Figure 5 shows a hierarchical structure of nodes in the two first lines that can be interpreted and transformed, such as IF (deg-malig= 1 OR deg-malig= 2) AND inv-nodes=15-17 THEN class=no-recurrence-events. This structure is represented in the fourth line of figure 6.

One of the most important contributions for the post-processing step is how the user can view the extracted knowledge. A statistical report is prepared for all features. Besides, it is possible to view the rules loaded in the environment. For example, when the user chooses to eliminate all redundancies of the production rules extracted from the database “breast-cancer” a report is generated (Figure 7). Moreover, besides the analysis of the extracted knowledge, users view statistics that may complement their understanding.

Based on the generated report (Figure 7) it is possible to observe that 21 rules were processed, of which, 14 (66.7%) had one or more redundant conditions in the antecedent of the rule. Considering the total set of rules, 20 redundant conditions were identified and eliminated, an average of 0.95 per rule. Among these attribute-operator-value conditions, the most frequent in the rules was “deg-malig=3”. The most frequently eliminate condition involved the attribute “tumor-size”. Or else, besides the set of non-redundancy rules, the user can view statistics for monitoring process traceability.

Number of rules: 21
 Number of rules with redundancy: 14
 Redundancy average: 66.67%
 Number of premises redundant: 20
 Redundancy premises per rule: 0.95
 Number of distincts attribute-value: 32

Statistics of attribute-value:

Item	Rules With	Redundancy Counter
deg-malign=3	11	0
tumor-size>>{10-14,15-19,20-24,25	9	8
inv-nodes>>{0-2,3-5,9-11,12-14,15	8	6
tumor-size>>{15-19,20-24,25-29,30	7	0
inv-nodes>>{0-2,9-11,12-14,15-17}	6	0
menopause=ge40	3	0
breast-quad>>{left_up,left_low}	1	0
breast-quad>>{right_low,central}	2	0
breast-quad=right_up	5	0
menopause=premeno	2	0
breast-quad>>{left_up,left_low,ri	1	0
deg-malign>>{1,2}	10	0
inv-nodes>>{0-2,3-5,6-8}	7	6
inv-nodes>>{0-2,3-5}	6	0
tumor-size>>{15-19,20-24,30-34,35	2	0
node-caps=yes	1	0
node-caps=no	1	0
menopause=lt40	1	0
tumor-size>>{25-29,45-49}	1	0
tumor-size>>{0-4,5-9,10-14,55-59}	1	0
inv-nodes=3-5	1	0
breast-quad>>{left_up,left_low,ce	1	0
breast-quad=right_low	1	0
tumor-size>>{10-14,40-44}	1	0
inv-nodes=6-8	1	0
inv-nodes=9-11	2	0
age>>{30-39,50-59,70-79}	1	0

Figure 7. Example of report generated by the PREPOS environment.

4. CONCLUSIONS AND FUTURE PERSPECTIVES

The major limiting factors in the use of the KDD process are the operational difficulties in the preparation of data and in the analysis of the discovered patterns, difficulties arising from the skills required for processing in different environments, the way data are available, time required for the process, among others. Thus, the environment proposed and developed – PREPOS - provides a contribution to the KDD area, because besides the new integrated features made available, the new environment was implemented in open source platform, i.e., new features can be added to the structure built. Thus, the user does not need to interact with different interfaces, which is the main problem encountered in this research field.

This tool is being used by researchers of the Postgraduate Program in Health Technology of the Pontifícia Universidade Católica do Paraná not only with the purpose of better exploring the available databases, but also to guide further research on the incorporation of data mining in routine activities related to health, both in management and in clinical practice.

Therefore, despite all the progress made, the KDD process remains the subject of research for new solutions that come close to the real interest of potential users. The explanation for this is that, despite the various experiments done, the KDD process is still little used in decision making in daily health care. (Mariscal et al. 2010), because of the lack of familiarity of experts with the involved methodology, advantages and disadvantages (Meyfroidt et al 2009), and the prevalence of statistical studies aimed to reveal simple linear relationships between health care factors (Cruz-Ramires, et al 2012).

REFERENCES

- Borgelt, C. APRIORI. Disponível em: <www.borgelt.net/apriori.html>. Acesso em: 16 dez. 2013.
- Carvalho D.R.; Moser A.D.; Silva V.A.; Dallagassa M.R. (2012) Mineração de Dados aplicada à fisioterapia. *Rev. Fisioterapia em Movimento*, Curitiba, v. 25, n.3, jul 2012.
- Cruz-Ramírez, M.; Hervás-Martínez, C.; Gutiérrez; P.A.; Pérez-Ortiz, M.; Briceño, J.; De La Mata, M. Memetic Pareto differential evolutionary neural network used to solve an unbalanced liver transplantation problem. *Soft Computing*, 2013. V17 n2 p 275-284.
- Fayyad, U; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, California, v. 17, n. 1, p. 37-54, 1996.
- Mariscal, G.; Marbán, O.; Fernandez, C. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 2010. V25:2, 137–166.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.
- Meyfroidt G.; Güiza, F.; Ramon, J. Bruynooghe, M. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, 2009. N 23 p 127–143.
- Milani C. S.; Carvalho D. R. Pós-Processamento em KDD. *Revista de Engenharia e Tecnologia*, v. 5, p. 151-162, 2013.
- Passos E.; Goldschmidt R. (2005) *Data Mining: um guia prático*. Rio de Janeiro: Elsevier, 2005.

- Rezende S.O. (2005) *Sistemas Inteligentes: fundamentos e aplicações*. Editora Manole, São Paulo, 2005.
- Rocha, M. R. M. *O Uso de Medidas de Desempenho e de Grau de Interesse para Análise de Regras Descobertas nos Classificadores*. 2008. 118 f. Dissertação (Pós-Graduação em Engenharia Elétrica) – Instituto de Engenharia Elétrica, Universidade Presbiteriana Mackenzie, São Paulo, 2008.
- Sinoara, R. A. *Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas*. 2006. 127 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2006.
- Zwitter, M. *Breast cancer data*. Institute of Oncology, Ljubljana, Yugoslavia, 1998.