VISUALIZATION OF ASSOCIATION RULES

Bruno R. M. Santos, Deborah Ribeiro Carvalho
Pontifícia Universidade Católica do Paraná (PUC-PR)
E-mails: {bruno.muchau.santos@gmail.com, drdrcarvalho@gmail.com}

Abstract: In the past few years, many studies have been conducted about KDD, however its incorporation in in health management routines is far from the desired, and much meaningful information has been not used properly. The aim of this article is to present a tool that enables health managers to easily identify association of events that happen in health routines, by post processing the output of the Apriori KDD algorithm. The proposed tool facilitates the discovery of new patters, as it aims to highlight two key points in this process, which are the transitivity between events, and how intense an event association is. The proposed tool was evaluated by potential users, and 100% of them considered the tool helpful in the task of discovering association of events. It is expected that this tool serves as a boost for new researches and studies of KDD and data mining algorithms in the health management area, as it has a growing field of acting.

Keywords: Knowledge Discovery in Databases, Post-processing, Graph

1. INTRODUCTION

Since the increase in capacity and speed of data storage, information technology has been continuously looking for strategies which optimize the potential of this stored data in support of decision taking.

Data mining is the process of exploring big amounts of data, which aim is to discover anomalies, patterns and correlations, as well as descriptive, understandable and predictive models, so that decision-making processes can be easier and wiser [Zaki and Meira Jr 2014].

The amount of data that has been produced in the past few years has increased considerably, given that it is possible to create data from a wide variety of sources, like cellphones, social media networks, health transactions, and so on. However, more data and more information does not necessarily mean more knowledge. Therefore, by using data mining techniques, it is possible to filter all unnecessary information, and understand what is relevant and significant to one's particular needs, and then, make good choices and take proper actions based on the discovered results [Zaki and Meira Jr 2014].

The Knowledge Discovery in Databases (KDD) is a general process, which aims to convert raw data into useful information. It is composed of three phases: preprocessing, in which the cleaning and selection procedures of data is done, processing, which is about relevant information discovery, and post processing, which refines the results discovered in the earlier stages [Zaki and Meira Jr 2014].

Despite of the fact that many studies and researches have been conducted recently about KDD, its use in health management routines is far from the desired, and many meaningful data has not been properly treated, and therefore, valuable information has not been taken into account in health management processes [Carvalho, Escobar and Tsunoda 2014].

The health management area has been facing different challenges about the implantation of data mining systems, such as, health professionals do not have a basic knowledge about data mining methodologies and advantages, and nowadays, the available systems provide simple relations about health routines, which could be found by any other way [Carvalho, Escobar and Tsunoda 2014].

Therefore, the aim of this project is to provide a tool that enables health managers to easily identify association of events that happen in health routines.

The data used by this application is the output of data mining algorithms, so this project is limited to post processing the results and patterns that were previously discovered by the two first phases of KDD algorithms, which are the preprocessing and the information discovery stages. Consequently, only the third phase of the KDD process will be part of this project's scope. Thus, the referred tool generates a visualization of the discovered patterns, which allows health managers to better understand the data they are dealing with, and get relevant information from it.

The platform has some inputs from the user, and intuitiveness and a user-friendly interface are some of the requisites.

The tool has two key points: the first one is the ability to understand transitivity between the event group set, that is, when an event is associated with another one, which is linked to a third one, the first event will be consequently associated with the third event. This allows the user to discover events associations with a higher level of relationships, which would not be easily identified by reading the output of data mining algorithms, which generally are huge files that would need a big effort to be understood.

Besides that, the user is also able to recognize how intense an event association is, given that the tool highlights events with higher confidence. One more time, this task would consume a lot of time and would be hard to be executed without the help of computing software.

2. Literature Review

The data mining study and research, as well as ways of extracting knowledge from large-scale data, have been gaining much attention and interest by researchers and scholars of computing over the last years. In general, the purpose of visualization of association rules is the display of data in an easy and intuitive way to users. So many authors have of the scientific community have focused their studies on developing applications whose goal is to simplify the analysis of association rules.

Bruzesse and Davino [2003] proposed two different strategies to help the user analyze and interpret association rules. Both strategies are based on visualization concepts and pruning. The first method, called "Show and Prune", gives priority to the viewing stage that becomes an interactive tool to prune (delete) redundant association rules and with little sense of the purpose of the user. In the second strategy, called "Prune and Show", synthetic methods and multi-view patterns are applied to association rules remaining in the study group after passing through an initial filter, which was delimited by statistical tests.

Abdullah et al. [2014] developed a model to display least critical association rules. This model is composed of five main stages. The first one is exploration of the data set, given that all datasets used are in a flat file format. Each record is written in a line in the file, and stored separately from the others. The second component is the construction of Least Pattern Tree (LP-Tree) structure, which is based on the support descending orders of items. In the next stage, the Critical Relative Support (CRS) for each association rule is computed, wherein the support of the antecedent and of the consequent are used in this calculation. The definition of the Critical Least Association Rules, which are the rules that have a CRS value equal or more than the minimum CRS, is defined in the fourth component. The last one is the visualization itself, in which the critical least association rules are presented in a 3D chart.

In order to evaluate association rules that are generated by data mining techniques, Blanchard, Giullet and Briand [2003] developed a user-driven and quality-oriented method for the visualization of association. This approach based on a specific model relies on rules interestingness measures and on interactive rule subset focusing and mining.

Chakravarty and Zhang [2003] defined a method that stores all association rules in a relational database. By filtering and rules order, the proposed system is considered an efficient and flexible way to manage and view a big set of association rules.

Junior [2005] talks about two different issues that are present in the data mining field: the first one is the high volume of discovered rules, which makes it difficult to identify more interesting and valuable rules, while the second one is related to the complexity of the concepts the data mining has. In most cases, the end user struggles to understand all of the techniques used, and then it ceases to use the system satisfactorily. Thus, in order to solve these issues, the author describes two different approaches to display a set of association rules and details which of the two methods is the most effective one.

Melanda [2005] says that there are many gaps concerning an intuitive method to prioritize and select the most appropriate rules in a large set. Thus, a model for post processing a particular group of association rules was presented, so that small groups of association rules, which were considered interesting, are presented to users.

Due to the existing gap in the easiness of discovering valuable and interesting knowledge from non-Information Technology professionals, Milani and Carvalho (2013) proposed a tool that aimed to integrate all existing steps in the KDD process. Whereas pre and post-processing, with data mining algorithms, in order to create an environment which could allow professionals and managers from different areas of expertise to run and discover interesting information from specific sets of data. Such proposal could be justified for many reasons: The pre-processing phase generally involves procedures, which can be performed by computing professionals, like preparing and organizing files in a very specific format. Furthermore, the post-processing alternatives have a very restricted availability, and it is difficult to obtain portable patterns, due to the variety of data formats. Therefore, the proposed tool puts together different algorithms, in a user-friendly interface, in order to generate valuable data without the need of computing experts' knowledge.

Despite of the fact that many researchers have spent a big effort on developing different applications, none of the articles mentioned above are focused on the relationship between different events, that is, it is not possible to check higher levels of events connectivity. The articles cited above aim to build a tool, which allows the user to see the entire whole group set as in a unique view, and it is difficult to recognize different relationships within the group of association rules.

Furthermore, some of the proposed projects do not show how intense those relationships are, since they are concerned about other ways of showing data.

3. METHODOLOGY

The tool was developed by using Java language, along with some external libraries. In order to create the user interface, we chose Swing, which is a graphic user interface toolkit for Java. In addition to that, Java JUNG [2003] was also used. JUNG stands for Java Universal Network/Graph. It is a framework, which provides a common and extendible language for modeling, analysis and visualization of data that can be represented as a network or as a graph. The proposed tool was developed in Java, it can be executed in any operational system.

The figure 1 shows the all interactions the user has with the proposed tool, that is, all input data and commands to generate a particular visualization, and everything the system gives the user during the processing.

The user should enter the proposed tool the file he wishes to be processed. Once the file has been entered, the tool does an initial processing of it, that is, it reads all rules and store them in an easy way so they can be processed. Then, the proposed tool gives the user the basic file information, which includes the number of rules, the number of unique antecedents, the number of unique consequents, the confidence mean value and the support mean value.

Once the user has this basic information, he needs to provide the application the data item he wishes to process, as well as the desired interval for the visualization. It is important to note that establishing this interval is not required, and the user is able to generate the visualization without doing do so.

Then, after the user defines all parameters, the tool process the rules contained in the input file, and then, the visualization is generated and shown in the screen. The user is able to interact with it, by moving and organizing the data the way he best needs.

If desired, the user is also able to export the visualization generated, and save it as an image in his computer, for later reviews of it. Consequently, the user does not need to reprocess a whole set of data, if he wants to check something that was previously generated.

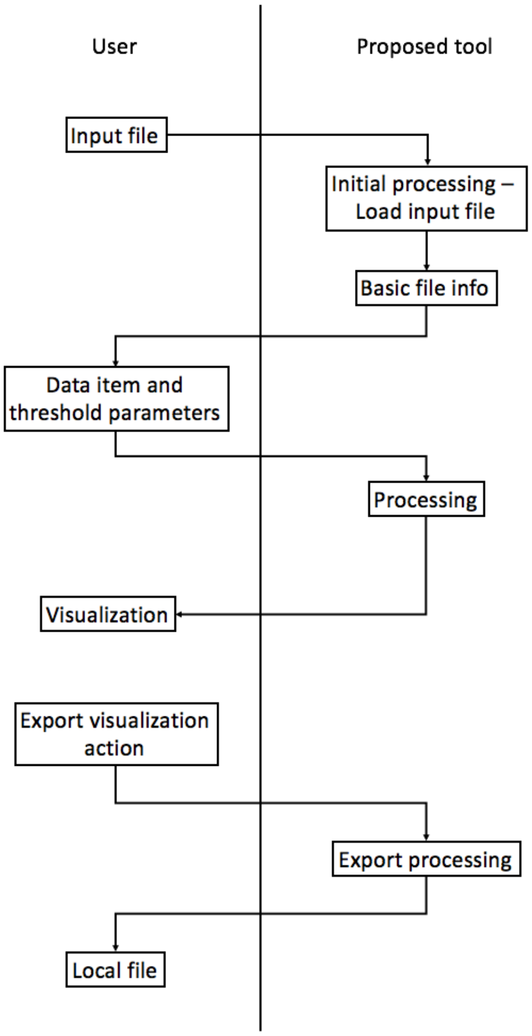

User
Proposed tool
Input file
Initial processing – Load input file
Basic file info
Data item and threshold parameters
Processing
Visualization
Export visualization action
Export processing
Local file

Figure 1: Input and output stream of the proposed tool

3.1. Data Input

Only the third stage of the KDD process is part of this project's scope, so the tool receives the output from a data mining algorithm, and all the processing happens from it.

The proposed tool receives data that is formatted as the output of the Apriori (Borgelt, 2004).

The association rule presents the structure X → Y, which means if X (antecedent) then Y (consequent) (Figure 2).

```
consequent <- antecedent (support_value%, confidence_value%)
```

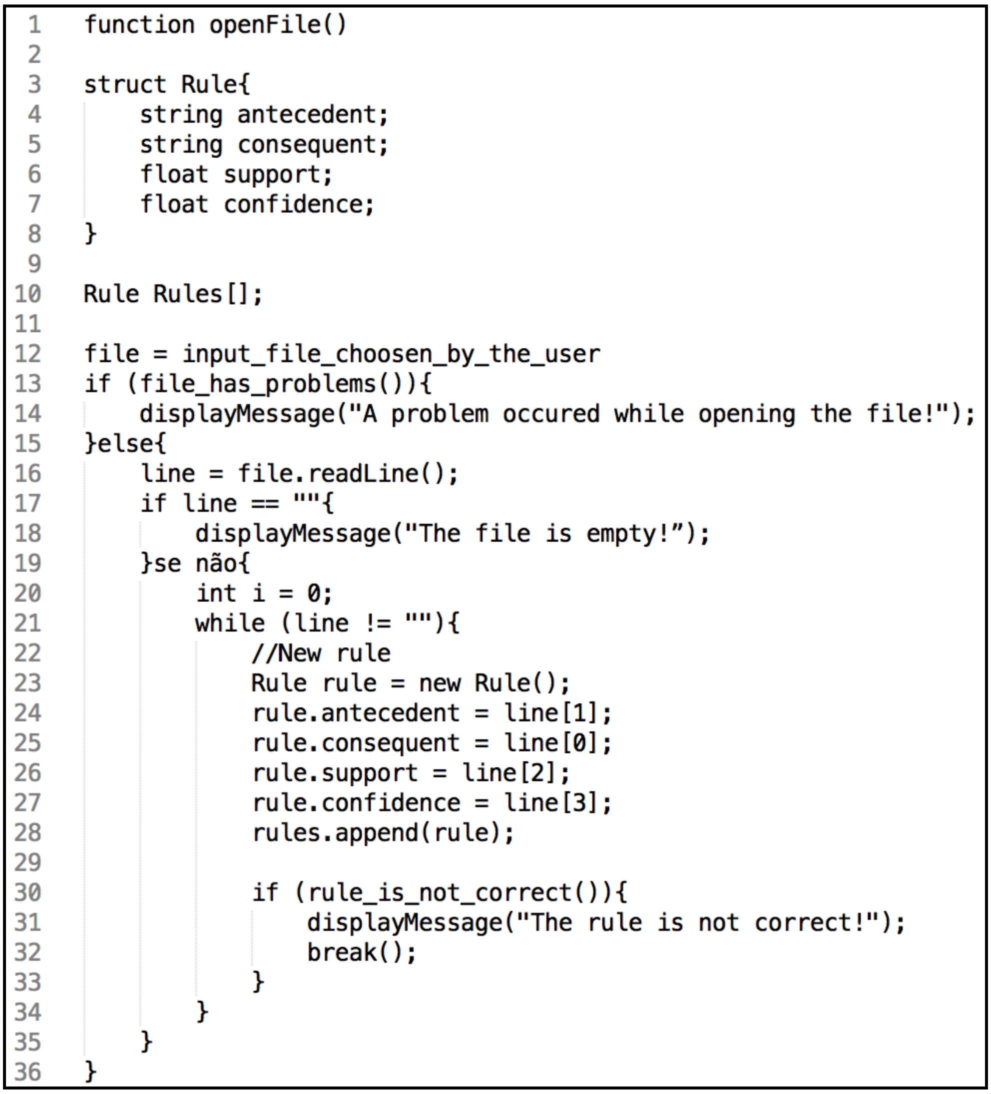
Figure 2: Association rules' general format

The restriction of this project is both the consequent and the antecedent are composed by only one data item.

This program should process input files that contain association rules related to health events, in order to discover how connected they are, so high-cost events can be determined easily.

3.2. Processing

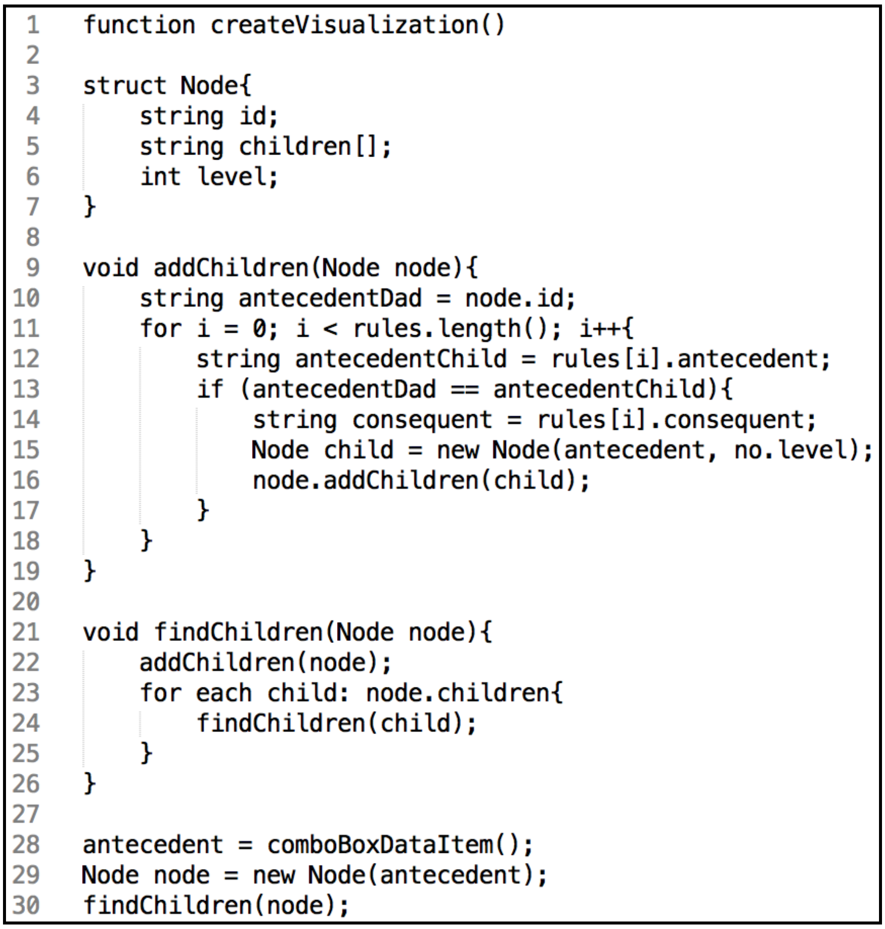
Once the user choses the input file, the proposed tool opens it in order to extract all association rules, and to store them in a structure that can be manipulated by the application. Several tests are performed to validate the input data, and these tests include the verification of the number of data items for the consequent and the antecedent, verification of values of support and confidence of each rule, and so on (Algorithm 1).



```
function openFile()

struct Rule{
    string antecedent;
    string consequent;
    float support;
    float confidence;
}

Rule Rules[];

file = input_file_choosen_by_the_user
if (file_has_problems()){
    displayMessage("A problem occured while opening the file!");
}else{
    line = file.readLine();
    if line == ""{
        displayMessage("The file is empty!”);
    }se não{
        int i = 0;
        while (line != ""){
            //New rule
            Rule rule = new Rule();
            rule.antecedent = line[1];
            rule.consequent = line[0];
            rule.support = line[2];
            rule.confidence = line[3];
            rules.append(rule);

            if (rule_is_not_correct()){
                displayMessage("The rule is not correct!");
                break();
            }
        }
    }
}
```

Algorithm 1: Pseudocode of the opening and loading the input file chosen by the user

In order to generate the visualization, the rules extracted from the input file have to be processed and manipulated according to the parameters given by the user through the graphic user interface. These parameters are the item data that is used to start the visualization, which is obligatory, and a support and confidence intervals, which are not required for the visualization to be generated (Algorithm 2).



```
function createVisualization()

struct Node{
    string id;
    string children[];
    int level;
}

void addChildren(Node node){
    string antecedentDad = node.id;
    for i = 0; i < rules.length(); i++{
        string antecedentChild = rules[i].antecedent;
        if (antecedentDad == antecedentChild){
            string consequent = rules[i].consequent;
            Node child = new Node(antecedent, no.level);
            node.addChildren(child);
        }
    }
}

void findChildren(Node node){
    addChildren(node);
    for each child: node.children{
        findChildren(child);
    }
}

antecedent = comboBoxDataItem();
Node node = new Node(antecedent);
findChildren(node);
```

Algorithm 2: Pseudocode of the definition of each rule's children from the initial data item

Recursive and iterative function calls were combined in order to define the sons each node has.

4. RESULTS

To better exemplify the workings and the exit of the proposed algorithm, a database which contained data of employees who either were or weren't awarded a sick leave due to orthopedic issues was adopted. For these two groups, it was taken into account the total use of 82 orthopedic procedures, demanded between 2007 and 2013. These 82 procedures were indicated by five orthopedic doctors. In order to find the classifier, the statuses 'on sick leave' or 'not on sick leave' were adopted.

Figure 3 shows the first screen of the tool, in which the user may choose the input file containing the rules to be processed.

The system should be as more user-friendly as possible, so the less choices the user has to make, the better. That is why he has only one option available, which is to open an existing file.

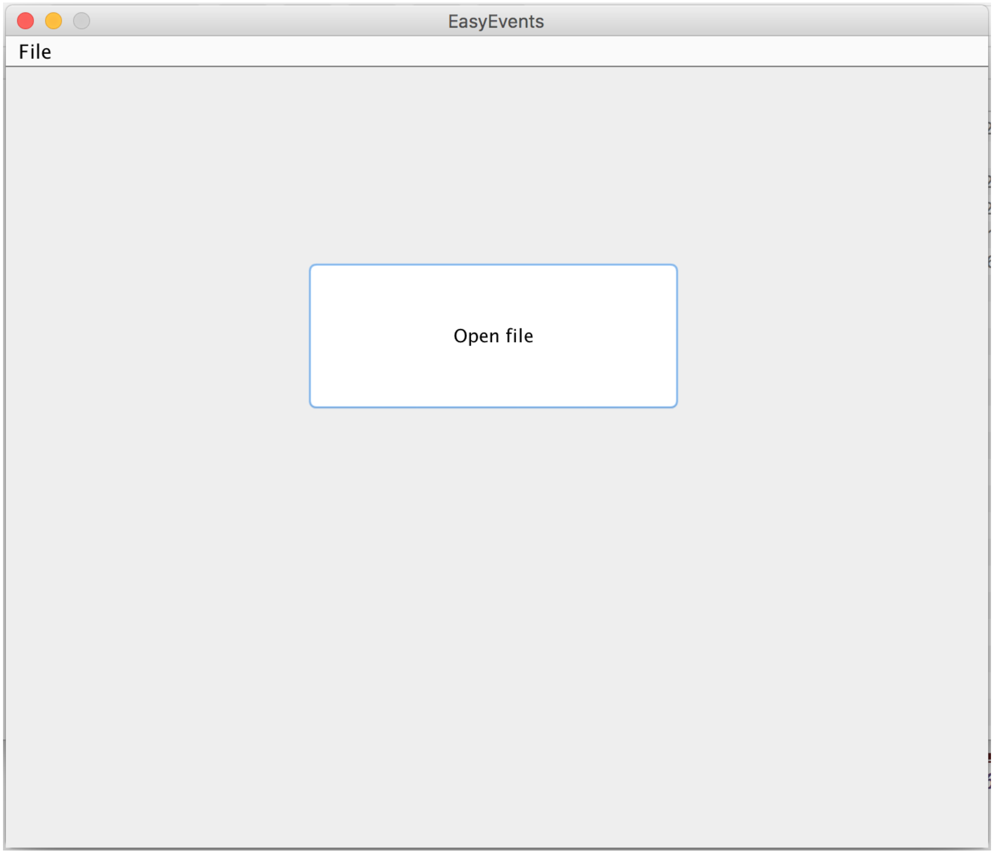


Figure 3: Screen which the user can select the input file to be processed

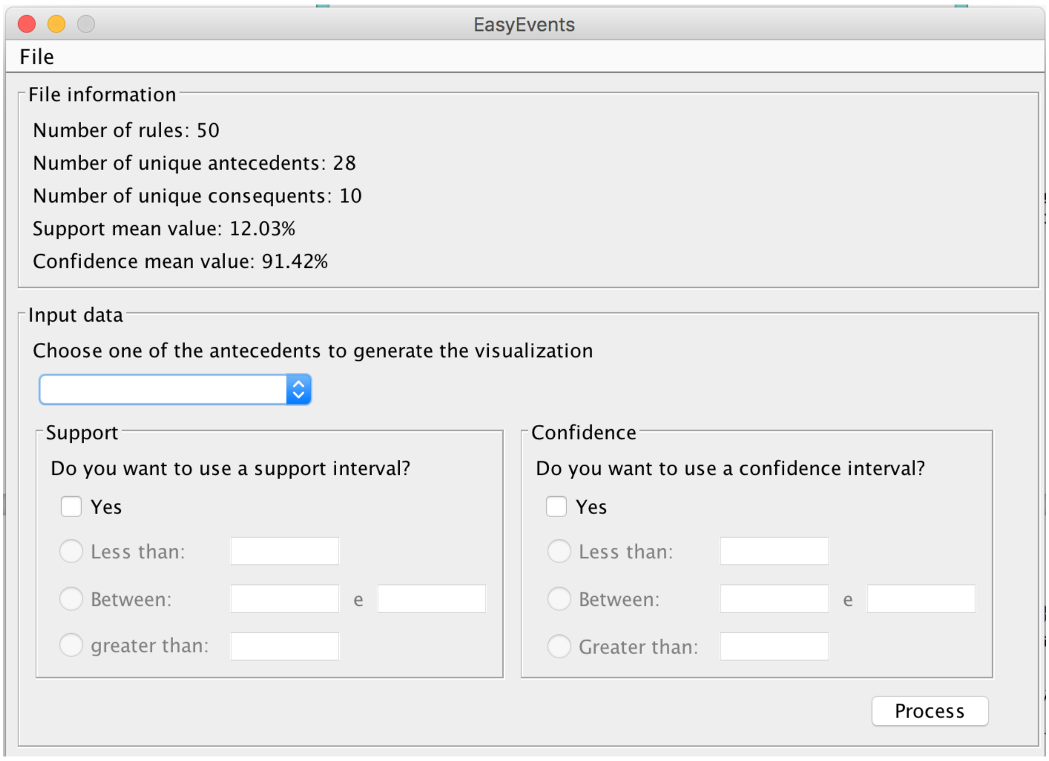
Figure 4 shows the screen after the initial file is loaded. The user is able to see the basic file information, and he should input the initial parameters for the processing to happen.

Finally, after the user has defined all the parameters, the visualization is generated, according to the figure 5.

As mentioned earlier, the user is capable of interacting with the graph generated. The instructions for doing so are provided in the screen. The user is also able to export the visualization into an image file, so he can check this visualization in future opportunities.

It is highly important to comment the two key points of the application: the first one is the color of the events. Events with the color, other than black, are events that appear more than once in the visualization pane. This helps the user identifies what events have more chances to happen by starting from a particular one.

The second one is the transitivity between the events, as now it is possible to determine events with a higher level of relationship.



File information

Number of rules: 50

Number of unique antecedents: 28

Number of unique consequents: 10

Support mean value: 12.03%

Confidence mean value: 91.42%

Input data

Choose one of the antecedents to generate the visualization

Support

Do you want to use a support interval?

□ Yes

○ Less than:

○ Between: e

○ greater than:

Confidence

Do you want to use a confidence interval?

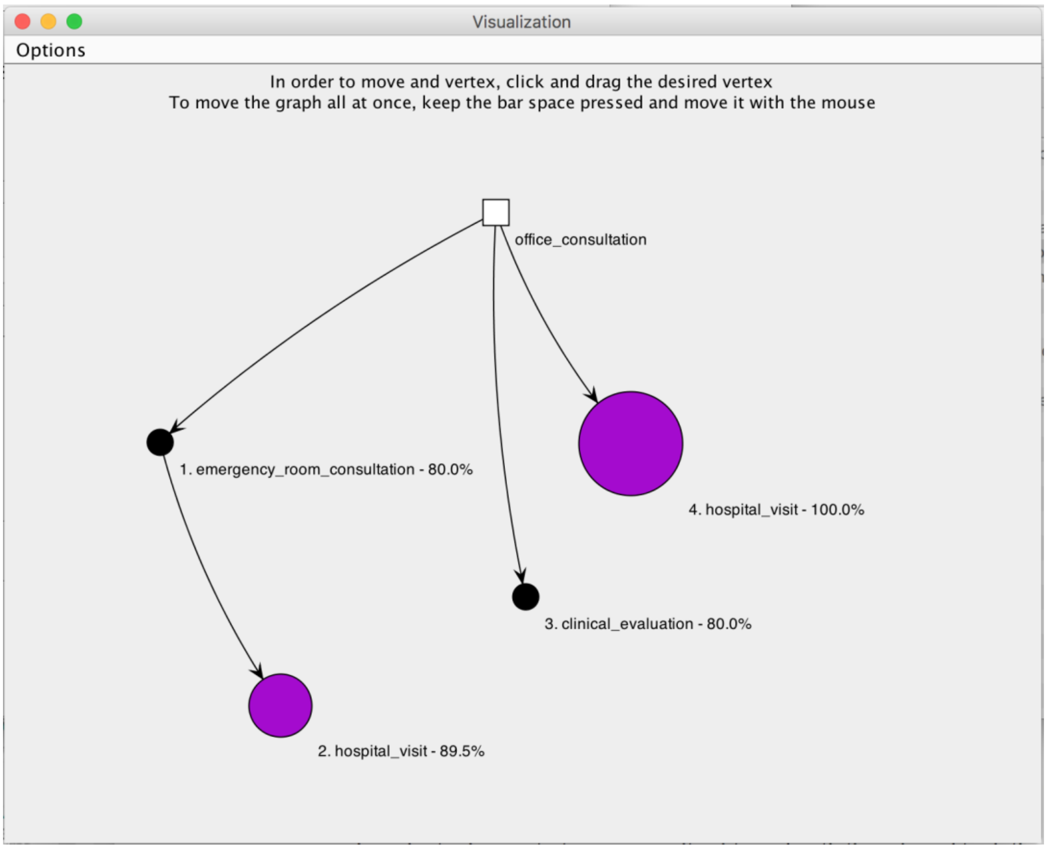
□ Yes

○ Less than:

○ Between: e

○ Greater than:

Figure 4: Screen which the user enters the necessary parameters to generate the visualization.


Visualization
Options
office_consultation
1. emergency_room_consultation - 80.0%
4. hospital_visit - 100.0%
3. clinical_evaluation - 80.0%
2. hospital_visit - 89.5%

In order to move and vertex, click and drag the desired vertex
To move the graph all at once, keep the bar space pressed and move it with the mouse

Figure 5: Screen of the visualization of the data item chosen by the user from the input file

4.1. Evaluation

In order to evaluate the proposed tool, potential users of the tool filled in a form. The form is composed by thirteen question, and except by the first three, which have to be answered as "Yes" or "No", all of the other questions must be answered in a scale, given that their options vary from "Strong disagree" to "Strong agree".

The questions presented in the form are the following:

1. Had you ever worked with association rules before?
2. I already had previous data mining knowledge.
3. I consider myself as a data mining user.
4. I considered it easy to open a file that contained association rules.
5. The basic file information to be processed (number of rules, number of unique antecedents etc.) are interesting for me.
6. I consider it easy to select a data item to be processed.
7. I consider it easy to use a confidence interval to generate the visualization.
8. The error messages are useful to correct existing issues in the input file.
9. I consider it easy to generate the visualization of a particular input file.
10. I consider it intuitive the way the visualization was presented in the screen.
11. I consider it easy to move an event in the screen,
12. The system made it easy the task of identifying the relationship between data items.
13. I consider it easy to export the visualization into my computer.

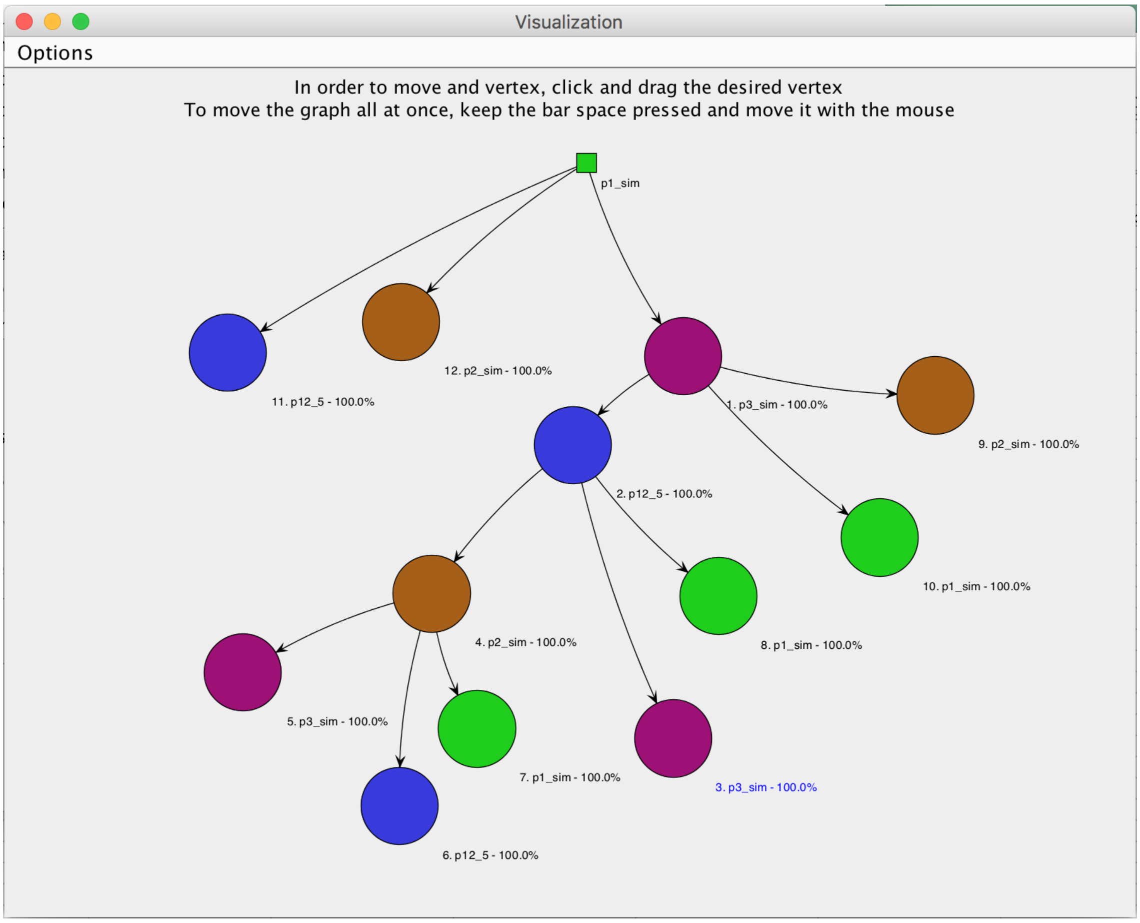
The potential users of the proposed tool were conducted to assess the quality of it, by following the instructions presented in a user guide. As the proposed tool is built to be very intuitive, no more formal guidance was provided, and the users should have been able to use the tool with no struggle. Some sample input files were made available to them, and once the visualization was generated, the users were asked to fill in the form above, so we could measure how user-friendly and easy-to-use the tool is.

Eight users filled in the form, and all of them already had had some knowledge in data mining and association rules. For question 12, all users answered that they strongly agreed that the system made it easy to identify the relationship between data items.

To evaluate the answers, we exported them to a spreadsheet and the file was converted to the very specific format the proposed tool accepts, so we could find the relationships between the answers. Each question was assigned a particular label, which could identify this question from the others. For example, the question one, was labeled as "P1_" (P for "pergunta", that is question in Portuguese), the question 2, "P2_", and so on. The answers were denormalized according to their question plus their answer. Therefore, as the three first question could be answered as only "Yes" or "No", the answer for question number one would be "P1_Yes" or

"P1_No". The other questions could be answered as numbers in a scale, which the number one would represent "Strongly disagree", and five would be "Strongly agree". So for question number 4, which could be answered as "Strong agree", it would be represented as "P4_5".

The figure 6 shows the association between the answers, starting with question number one, which is "Had you ever worked with association rules before?", with a confidence of 100%:



In order to move and vertex, click and drag the desired vertex
To move the graph all at once, keep the bar space pressed and move it with the mouse

Figure 6: Screen of the visualization of the association of answers

From that particular question, all users that already have some knowledge in data mining concepts, considered it easy to identify association of events. Since from the transitivity of question number one, question 12 ("p12_5"), which stands for "The system made it easy the task of identifying the relationship between data items", was highly related to it, which confirms that this tool accomplished its proposal, which was to simplify the discovery of knowledge.

5. CONCLUSIONS

This paper proposed and tested two strategies to post-process patterns discovered in the format of classifiers represented from decision trees.

The health management area has been facing many challenges in the use of data mining tools, in order to allow health professionals to better understand all the data and information they deal with. This may be happening due to some very specific reasons: the systems and tools available today require IT professionals to treat the input data, so it can be set for processing, and the user-friendly tools available today are capable of providing simple's health routines relations. Thus, either the available activities are hard to be dealt with, and generally, health managers are not capable of manipulating it, or when the available tools are user-friendly and intuitive, they cannot provide relevant information.

Therefore, the tool proposed in this article aims to provide a contribution to the use of data mining and KDD algorithms to the health management area, since it provides an easy-to-use system that allow health managers to better understand and visualize the large-scale data they work with every day. Through the proposed tool, the user is able to identify higher levels of relationships between different events, and is capable of recognize how intense particular events are in a large group set.

It is expected that more researchers and computing scholars can put their effort in the implementation of data mining and KDD systems in the health management area, since this is a constant growing field of acting. It is also expected that decision-making processes can be easier due to the use of computing systems and its technology.

REFERENCES

ABDULLAH, Z.; HERAWA, T.; NORAZIAH, A.; DERIS, M. D. A Model for Visualizing Critical Least Association Rules. International Journal of Software Engineering and Its Applications, v. 8, n. 1, p. 167-180, 2014.

BLANCHARD, J.; GIULLET, F.; BRIAND, H. A User-Driven and Quality-Oriented Visualization for Mining Association Rules. Proceedings of the Third IEEE International Conference on Data Mining, 2003.

BORGELT, C. Apriori – Association Rule Induction / Frequent Item Set Mining. Disponível em: <http://www.borgelt.net/apriori.html>. Acesso em: nov 2016.

BRUZESSE, D.; DAVINO, C. Visual post-analysis of association rules. Journal of Visual Languages and Computing, v. 14, p. 621-635, 2003.

CARVALHO, D. R; ESCOBAR, L. F. A.; TSUNODA, D. Pontos de Atenção para o uso da Mineração de Dados na Saúde. Informação & Informação, v. 19, n. 1, p. 249-272, fev. 2014.

CHAKRANVARTHY, S.; ZHANG, H. Visualization of Association Rules over Relational DBMs. University of Texas at Arlington, 2003.

JAVA JUNG - "Java Universal Network/Graph Framework". Disponível em: <http://jung.sourceforge.net>. Acesso em: nov 2016.

JÚNIOR, F. A. N. Visualização de Regras de Associação. Tese de Doutorado, Belo Horizonte, MG: Universidade Federal de Minas Gerais, 2003.

MELANDA, E. A.. Pós-Processamento de Regras de Associação. Tese de Doutorado, São Paulo, SP: Universidade de São Paulo, 2004.

MILANI, C.; CARVALHO, D. R. Prepos Environment: A Simple Tool for Discovering Interesting Knowledge. Iberoamerican Journal of Applied Computing, v. 3, n. 2, p. 41-52, 2013.

ZAKI, M. J., MEIRA JR, W. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York, NY: Cambridge University Press, 2014.