Improving Hospital Decision Making with Interpretable Associations over DataCubes

Carlos MOLINA a,1, Belén PRADOS-SUAREZ b, Miguel PRADOS DE REYES c and Carmen PEÑA YAÑEZ b

a Department of Computer Sciences, University of Jaen, Jaen, Spain
b Department of Software Engineering, University of Granada, Granada, Spain
c Department of Computer Sciences, San Cecilio Hospital, Granada, Spain

Abstract. In this paper we propose a new Classification based on Association Rules (CAR) algorithm that improves the interpretability of the results, works over real data from the electronic health records (EHRs), and allows the study of the patient as a whole. It enables tasks such as the discovery of relationships between diseases, or offering several alternative and reasoned diagnoses for the cases of patients with several diseases that analysed separately could lead to mistaken diagnosis. We aim to achieve several goals: to discover hidden relationships; to improve the interpretability and reduce the complexity of the result; to obtain more reliable diagnosis (getting alternative reasoned diagnoses and higher robustness to noisy rules), and to improve the quality of the classifier avoiding the usual overfitting problem. To this purpose, we define and exploit hierarchies defined over datacubes dimensions, and change the way the association rules are obtained, and their evaluation at the classification process. To prove the utility of our proposal we have used it in an example of cancer discrimination.

Keywords. CAR, Complexity Reduction, Interpretability Improvement, Classification, DataCube, Multidimensional Model.

Introduction

In the last few years important advances have been realized in the application of data mining methods over health care information with research purposes to improve the managerial decision support [1,2] or to support medical diagnosis [3,4].

Nevertheless as Jensen et al. indicate in [5], this is still an under-used source of information, due to three main reasons: the nature of the medical data itself; the use of statistical data instead of real data obtained directly from the electronic health records (EHR); and the difficulties for the user to understand and interpret the results of the data mining techniques. It all makes the medical doctors reluctant to the use of these advances, and especially the clinical decision support systems (CDSS) [6].

Here we propose a data mining method specifically designed to improve interpretability of the results that also works over real data. Therefore, although the proposal presented in this paper can be applied to any of these problems, here we focus on an example of use for CDSS.

1 Corresponding Author.
Among the wide variety of data mining methods [7, 8] the most widely used on CDSS are the neural networks [9], the decision trees [10], or combinations of them [11]. The main inconvenience of these proposals is that they are specifically designed for concrete pathologies or parts of the human body. As Isola et al. states in [9], it is necessary to count on systems capable of analysing the medical data set as a whole, in order to find out interactions or relationships between pathologies or different diseases with similar symptoms for patients with several diseases whose symptoms may lead to confusion in the diagnosis, second opinions or alternative diagnosis.

With CAR proposals, it is possible to achieve these goals since, as it is well known [12], they are designed to discover relations, and in addition, they are the most precise techniques due to their exhaustive exploration of the possibilities. Nevertheless, they still have three inconveniences. First, their results are difficult to understand. This is due to the rule set obtained with CAR proposals, which is used as an ordered list of rules. This set is checked in order, discarding the rules unsatisfied and stopping the process at the first one satisfied. In real systems, the antecedents of these rules are usually complex and not directly meaningful, so the user must take into account the previously discarded rules to understand the full meaning of the classification performed. Second, they are not robust to noise, since they are based on the selection of a single response, so if the rule selected or any of the rules discarded are wrong the final response of the system will also be wrong. Finally, another problem that classification methods (not only CAR based) must face is the over-fitting of the data.

It is necessary to find techniques capable of working over real data from the EHRs, of finding out hidden relations, that are also robust to noise, offering a good classification without over-fitting, and that provide several and reasoned responses which must also be easy to understand. To achieve these goals we present our proposal, based on three key points: the use of “datacubes” and the exploitation of the hierarchies defined over their dimensions, the change in the way the association rules are obtained, and a change in the evaluation of the rules at the classification process.

In section 1, we show the adaptation of the algorithm to reduce the complexity of the resultant set of rules. Examples about the performance of our proposal are presented in section 2, and finally we indicate our conclusions and final remarks in section 3.

1. Algorithm

In this section, we present a modification of the Complexity Guided algorithm for Association Rule Extraction (COGARE) described in [13]. The modification proposed here aims to extract association rules adapted for classification. Next, we comment the changes that must be introduced.

**Itemset generation.** When generating the frequent itemsets we change the candidates and only consider the itemsets containing an item of the class variable. If the itemset is not frequent, the method applies the generalization process but not on the class item (the granularity of this variable does not change in any moment).

**Rule generation.** At the rule generation stage, the method only takes into account the association rules of the form $A \rightarrow C$, where $C$ only has one item that belongs to the class variable. Therefore, for each frequent itemset found, the method only generates one rule.
Rule generalization. In the rule generalization phase the scheme is similar to the itemset generation, because the method tries to generalize the rule set considering all the variables but the class.

Iterative approach. One of the main changes of the algorithm is the use of an iterative approach. Most of the proposals first generate all the association rules and later select the classification rules from this set. Here we propose the following iterative approach:

- First the method generates frequent itemsets of size 2 (one item from the class variable and another one from other variable),
- Then it generates the association rules of size two, pruning those with a certainty factor (CF) under the threshold given as parameter, and selects the rules from this set, so the method generates a classification model using simple rules (one antecedent and the class),
- Next, the method calculates the frequent itemset of size 3 using frequent itemsets of size 2. In this stage, the algorithm tries to reduce the number of rules, deleting unnecessary frequent itemsets as explained in the next subsection,
- After it the method generates the rules for the frequent itemsets of size 3, and selects the rules to classify,
- The method stops if the result is not better than using rules of size 2. If we obtain a better classification, the process continues with itemsets of size 4, and so on.

Space reduction. The aim of this process is to reduce the candidate frequent itemsets to be generated. The main idea is to avoid building more specific rules i.e. bigger itemsets in areas of the search space where the classification was made correctly with simpler rules. The method takes the current classification rule set, and looks for areas correctly classified with the current rule set. Those areas are delimited by the rules with CF equal to one (all the elements which satisfy the antecedent of the rule have the same class). Next, the process keeps out from the frequent itemsets those including the antecedent of any of these rules, and in the next iteration, no itemset in those areas will be considered. Avoiding the generation of larger and more complex rules in these areas, the use of unnecessary inserted items in the antecedent of the rules is also avoided, and hence the over-fitting they could introduce.

Rules selection. On each iteration, the algorithm selects a set of rules to build the classifier. First, the method sorts the rules according to the quality of each one in decreasing order (better rules first). Then it uses the first rule as initial classifier a computes its quality. The rest of the process checks iteratively whether the next rule improves the quality or not. If there is an improvement, the technique selects the rule and proceeds to the next iteration. Otherwise, the rule is dismissed.

Rule evaluation. As mentioned above in most of the CAR proposals the rule set obtained is used as an ordered list of rules. In our proposal, we do not establish an order between rules since we use a classic inference schema, considering that a rule is satisfied with a degree $\otimes(\mu(A_1), \ldots, \mu(A_n))$, where $A_i$s are the items in the antecedent and $\mu(A_i)$ the degree at which the record satisfies the item. As t-norm we use the minimum. We combine the result of each fired rule with the same consequent (class) with a t-conorm to obtain one unique value for each class (we use the maximum). The class with higher degree is the one selected to classify the new record.
Figure 1. COGARE structure with iterative process for rule generation

2. Example

To test the proposal we use the Wisconsin Diagnostic Breast Cancer (WDBC) at UCI ([14]). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. We consider 31 attributes and a total of 569 instances. We build over each attribute a hierarchy with five values (Very low, low, medium, high and very high) and another with only three values (low, medium and high). To obtain the partition of the values into the categories we have used the method presented in [15]. With this datacube we want to know the estimated type of the cancer: malignant or benign, so we consider as class the dimension Diagnosis. The parameters for the process are the following: Support=0.005, CF=0.1, Max complexity=0.2 (value that sets a high reduction of the number of rules in the association rule extraction), AcceptRatio=1.2 (a worse rule set may be accepted if the complexity reduction is high), β=10 (penalty, in the generalization process, for using values which produce a rule set that was not accepted in the past), α=0.7 (having less rules is preferred to improving the abstraction). To test the quality of the method we apply 10-fold cross-validation getting an average correct classification of 95.129% with 47.7 rules on average. Therefore, the method gets a good result.

3. Conclusions

In this paper, we have presented a modification of the COGARE algorithm for the extraction of association rules. With the methodology proposed four goals, finding out
hidden relations with the use of association rules may be achieved firstly. Secondly, we achieve robustness to noise, since we consider all the possible responses of the system, and provide the user with a combination of them too. Thirdly, an improvement of the interpretability of the results, since we have reduce the complexity of the result (less and simpler rules), and each rule is a result per se that can be interpreted independently. It means that each rule helps to understand not only the criteria under which the classification is being performed, but also the characteristics that determine each class. This research is still in progress and the algorithm is being tested under more data sets to be able to quantify the improvements in the above mentioned goals. Finally, we would like to remark that our proposal, which could also be applied for research purposes or managerial decision support, works over real data from the EHRs. It means that it can analyse the medical data set (and therefore the patient) as a whole, which is especially useful for tasks such as obtaining alternative diagnosis or discover relationships between diseases, offering also reasoning. It all may reduce the reservations of the medical doctors to use the CDSS.

References