Expert Systems with Applications 38 (2011) 13857–13862

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Pattern selection approaches for the logical analysis of data considering the outliers and the coverage of a pattern

Jeong Han [a], Norman Kim [b], Bong-Jin Yum [a,*], Myong K. Jeong [b]

[a] Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
[b] RUTCOR (Rutgers Center for Operations Research), Rutgers, The State University of New Jersey, Piscataway, NJ, USA

ARTICLE INFO

Keywords:
Classification
Logical analysis of data
Pattern selection
Set covering problem

ABSTRACT

The logical analysis of data (LAD) is one of the most promising data mining methods developed to date for extracting knowledge from data. The key feature of the LAD is the capability of detecting hidden patterns in the data. Because patterns are basically combinations of certain attributes, they can be used to build a decision boundary for classification in the LAD by providing important information to distinguish observations in one class from those in the other. The use of patterns may result in a more stable performance in terms of being able to classify both positive and negative classes due to their robustness to measurement errors.

The LAD technique, however, tends to choose too many patterns by solving a set covering problem to build a classifier; this is especially the case when outliers exist in the data set. In the set covering problem of the LAD, each observation should be covered by at least one pattern, even though the observation is an outlier. Thus, existing approaches tend to select too many patterns to cover these outliers, resulting in the problem of overfitting. Here, we propose new pattern selection approaches for LAD that take both outliers and the coverage of a pattern into account. The proposed approaches can avoid the problem of overfitting by building a sparse classifier. The performances of the proposed pattern selection approaches are compared with existing LAD approaches using several public data sets. The computational results show that the sparse classifiers built on the patterns selected by the proposed new approaches yield an improved classification performance compared to the existing approaches, especially when outliers exist in the data set.

1. Introduction

The *logical analysis of data* (LAD) is a new methodology aimed at discovering structural information hidden in data sets. Since the concept of the LAD was first introduced by Hammer (1986), many researchers have contributed to developing and clarifying various mathematical, theoretical, and applied aspects of the LAD (e.g., see Alexe & Hammer, 2007; Anthony, 1999; Bonates & Hammer, 2007; Boros, Hammer, Ibaraki, & Kogan, 1997; Boros et al., 2000; Bruni, 2007; Ekin, Hammer, & Kogan, 2000; Hertz, Dalang, Bengio, Goray, & Mayoraz, 2000; Mayoraz & Moreira, 1999; Ryoo & Jang, 2009). Various applications of the LAD have also been reported with successful results (Csizmadia & Vizvári, 2004; Kim & Ryoo, 2008; Lemaire et al., 2009; Ngai, Xiu, & Chau, 2009).

The hidden structural information (i.e., *patterns*) generated from the LAD distinguishes between observations in one class and those in the other. The LAD is a methodology for building a classification model based on patterns generated from two classes, which are called the positive and negative classes, respectively. The pattern for the positive (negative) class is a combination of certain attribute values occurring together only in the large number of the observations in the positive (negative) class. Therefore, the LAD patterns can be viewed as synthetic attributes which more closely reflect the positive or negative nature of an observation than the original attributes.

The basic approach of the LAD is to generate as many different patterns as possible to explain all of the observations in the original data set and then to choose the patterns by solving a set covering problem for a large set of patterns (Boros et al., 2000; Hertz et al., 2000). However, in the set covering problem, each positive (negative) observation, including an outlier, should be covered by at least one positive (negative) pattern. Consequently, too many patterns may be chosen, resulting in an overfitted classifier due to the selection of patterns that even cover outliers. The goal of this paper is to develop new pattern selection approaches that will avoid the problem of overfitting by accounting for the outliers and the coverage of a pattern.

This paper is organized as follows. The pattern selection process of the LAD is introduced in Section 2. Section 3 develops new approaches for selecting significant patterns for an improved classification performance of the LAD. These proposed pattern

* Corresponding author. Tel.: +82 42 350 3116; fax: +82 42 350 2420.
E-mail address: bjyum@kaist.ac.kr (B.-J. Yum).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.189

selection approaches are aimed at avoiding the problem of overfitting and improving the prediction accuracy of the LAD. In Section 4, we perform the computational experiments to compare the performance of the proposed approaches with the existing approaches. Finally, conclusions and recommendations for future research areas are given in Section 5.

2. The pattern selection process of the LAD

The LAD consists of five basic steps: data binarization, support set construction, pattern generation, pattern selection, and theory formulation (Boros et al., 2000; Han, Kim, Yum, & Jeong, 2010; Hertz et al., 2000). In this section, the pattern selection step of the LAD is described in detail. The aim of the pattern selection step is to find the minimal subset of the patterns that explain every observation. Boros et al. (2000) formulated the problem of pattern selection as a *set covering problem* such that every positive (negative) observation should be covered by at least one positive (negative) pattern.

For the sake of brevity, only the procedure for the selection of positive patterns is described. The selection of negative patterns proceeds in a similar way. Consider $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ as the positive observations and $P_1^+, P_2^+, \ldots, P_R^+$ as the generated positive patterns. A Boolean vector $\mathbf{y}$ of size R indicates whether a pattern P_j^+ is included in the solution of the set covering problem ($y_j = 1$) or not ($y_j = 0$). A Boolean matrix $\mathbf{V}$ of size $N \times R$ is used to represent whether a positive observation $\mathbf{x}_i$ is covered by a $P_j^+ (V_{ij} = 1)$ or not covered ($V_{ij} = 0$). An observation $\mathbf{x}_i$ should be covered by at least one pattern, which is guaranteed by the following constraint:

$$\sum_{j=1}^{R} V_{ij} y_j \geqslant 1 \tag{1}$$

The goal is to find the minimal subset of positive patterns while satisfying the above inequality for all positive observations. This can be formulated as a set covering problem as follows.

$$\begin{aligned} Min \quad & \sum_{j=1}^{R} y_j \\ s.t. \quad & \sum_{j=1}^{R} V_{ij} y_j \geqslant 1, \quad i = 1, 2, \ldots, N \\ & y_j \in \{0, 1\}, \quad j = 1, 2, \ldots, R \end{aligned} \tag{2}$$

The set covering problem appears frequently in operations research, and it also has numerous applications in a variety of other fields. However, finding an optimal solution to the set covering problem is known to be NP-hard (Garey & Johnson, 1979; Kim & Ryoo, 2008). That is, if the problem size is large, the set covering problem may not be solved in polynomial time, thereby necessitating the development of efficient heuristic algorithms. A number of heuristic algorithms are currently in use, and the set covering problem can be solved satisfactorily using these algorithms in acceptable computation time.

The *simple greedy* heuristic algorithm is one of the popular heuristic approaches (Almuallim & Dietterich, 1991). At each step, the simple greedy algorithm selects a positive pattern that covers as many of the not-yet-covered positive observations as possible until all positive observations are covered. The selection of negative patterns proceeds in a similar way. The procedure for a positive pattern is outlined as follows:

1. Start with an empty solution set X and the set L of all positive observations.
2. Select a pattern which covers as many of the not-yet-covered positive observations as possible and add it to X. Then, delete the observations covered by the pattern from L.
3. If L is empty, stop. Otherwise, go to step 2.

3. Pattern selection approaches that consider outliers and the coverage of a pattern

Existing pattern generation approaches for the LAD may create an extensive set of patterns that contain many redundancies, ultimately resulting in overfitting, especially when outliers exist in the data set. The fundamental principle of the set covering problem for the LAD is that every positive (negative) observation should be covered by at least one positive (negative) pattern. In some cases, an observation may be covered by a single pattern and the coverage of this pattern is one or very small. This kind of observations may be considered as an outlier in the formulation of the LAD. Thus, when confronted with the presence of outliers in the data set, existing approaches tend to select more patterns to cover these outliers. If redundant patterns that cover only some outliers are contained in the solution, the LAD cannot maximize its predictive accuracy for the test data, with the consequence that the problem of overfitting arises in the patterns selected (Dietterich, 1995).

In addition, if a pattern covers only a few observations, the pattern may not be useful for formulating a theory – even though the observations are not outliers – due to the low explanatory power of these observations. In this case, the patterns selected by the existing pattern selection approaches can cause the problem of overfitting and subsequently the formulation of a theory becomes more complex than necessary. In order to overcome the overfitting issues, we propose a pattern selection approach that takes both outliers and the coverage of a pattern into account. We also propose a heuristic pattern selection approach based on the simple greedy algorithm to solve the problems in acceptable computational time.

3.1. Set-covering problem that considers outliers

As mentioned in Section 2, the set covering problem is frequently used to select patterns in the LAD. According to formulation (2), every positive (negative) observation must be covered by at least one positive (negative) pattern – even if the observation is an outlier. However, if an observation is an outlier, it does not need to be covered by any pattern, positive or negative. If the patterns which cover outliers are selected, the problem of overfitting occurs. Slack variables ξ_i will be used in the new pattern selection approach to deal with this problem. A Boolean vector ξ of size N is introduced to indicate whether an observation $\mathbf{x}_i$ is an outlier ($\xi_i = 1$) or not ($\xi_i = 0$). The requirement that each outlier does not need to be covered by any pattern can be achieved by modifying the right-hand side of the constraint in formulation (2) as follows:

$$\sum_{j=1}^{R} V_{ij} y_j \geqslant 1 - \xi_i$$

Note that if the ith observation is considered to be an outlier (i.e., $\xi_i = 1$), it needs not be covered by any pattern. Otherwise, each observation should be covered by at least one pattern. The objective function in formulation (2) should be modified adequately to select a reasonable number of patterns. The modified objective function can be written as:

$$Min \quad \sum_{j=1}^{R} y_j + C \sum_{i=1}^{N} \xi_i$$

where C, with an integer value, is a trade-off parameter between the number of selected patterns and outliers (i.e., trade-off between two conflicting goals: minimizing the number of selected patterns and minimizing the number of outliers). Therefore, the proposed set covering problem that considers outliers can be formulated as follows: