Information Sciences 320 (2015) 107–125

Contents lists available at ScienceDirect

Information Sciences

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

A novel method for constrained class association rule mining

Dang Nguyen [a,b], Loan T.T. Nguyen [c], Bay Vo [d,*], Tzung-Pei Hong [e,f]

[a] Division of Data Science, Ton Duc Thang University, Ho Chi Minh, Viet Nam
[b] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh, Viet Nam
[c] Faculty of Information Technology, VOV College, Ho Chi Minh, Viet Nam
[d] Faculty of Information Technology, Ho Chi Minh City University of Technology, Viet Nam
[e] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, ROC
[f] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, ROC

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form 26 April 2015
Accepted 3 May 2015
Available online 7 May 2015

Keywords:
Associative classification
Class association rule
Data mining
Useful rules

ABSTRACT

To create a classifier using an associative classification algorithm, a complete set of class association rules (CARs) is obtained from the training dataset. Most generated rules, however, are either redundant or insignificant. They not only confuse end users during decision-making but also decrease the performance of the classification process. Thus, it is necessary to eliminate redundant or unimportant rules as much as possible before they are used. A related problem is the discovery of interesting or useful rules. In existing classification systems, the set of such rules may not be discovered easily. However, in real world applications, end users often consider the rules with consequences that contain one of particular classes. For example, in cancer screening applications, researchers are very interested in rules that classify genes into the "cancer" class. This paper proposes a novel approach for mining relevant CARs that considers constraints on the rule consequent. A tree structure for storing frequent itemsets from the dataset is designed. Then, some theorems for pruning tree nodes that cannot generate rules satisfying the class constraints are provided and proved. Finally, an efficient algorithm for mining constrained CARs is presented. Experiments show that the proposed method is faster than existing methods.

1. Introduction

Association rule mining and classification are two important and common problems in the data mining field. Therefore, numerous approaches have been proposed for the integration of these models. Examples include classification based on association rules (CBA) [23], a classification model based on multiple association rules [21], a classification model based on predictive association rules [41], multi-class and multi-label associative classification [34], a classifier based on maximum entropy [35], the use of an equivalence class rule tree [39], a lattice-based approach for classification [29], the integration of taxonomy information into classifier construction [6], the integration of classification rules into a neural network [20], a condition-based classifier with a small number of rules [11], an efficient classification approach with a rule quality metric [10], and a combination of a Netconf measure and a rule ordering strategy based on rule size [15].

The implementation processes of these approaches are often similar. Firstly, a complete set of class association rules (CARs) is mined from the training dataset. Then, a subset of CARs is selected to form the classifier. However, the complete

* Corresponding author at: Faculty of Information Technology, Ho Chi Minh City University of Technology, Viet Nam.
E-mail addresses: nguyenphamhaidang@tdt.edu.vn, nguyenphamhaidang@outlook.com (D. Nguyen), nguyenthithuyloan@vov.org.vn (L.T.T. Nguyen), bayvodinh@gmail.com (B. Vo), tphong@nuk.edu.tw (T.-P. Hong).

http://dx.doi.org/10.1016/j.ins.2015.05.006
0020-0255/

set of CARs is often very large as it contains many redundant or insignificant rules. These useless rules not only waste storage space and decrease the performance of a classifier, but they also have a negative affect on decision-making.

To solve this problem, effort has been devoted to pruning redundant rules or ranking rules. The typical case of redundant rules is rule cover [36], where a rule is redundant if it is covered by others. The concept of sub-rules has been developed [23,29,39]. Another approach for pruning redundant rules to analyze the relations between rules with respect to objects was proposed by Liu et al. [24]. The authors used the Galois connections between objects and rules to determine the dependent relationships among rules. Their study showed that one rule is equivalent to another if they are supported by the same objects. Besides pruning strategies, some researchers have reported that ranking mechanisms are also important for CAR mining in associative classification since they directly affect classification accuracy. Therefore, several rule ranking approaches have been developed Li and Cercone [19] applied rough set theory to discover and rank significant rules Najeeb et al. [26] proposed a hybrid approach for rule ranking based on an evolutionary algorithm Cai et al. [8] ranked the interestingness of rules within an equivalent rule group for gene expression classification using two proposed interestingness measures called Max-Subrule-Conf and Min-Subrule-Conf Chen et al. [9] improved the performance of an associative classifier by rule prioritization. They proposed the MLRP algorithm, which re-ranks the execution order of CARs using rule priority to reduce the influence of rule dependence.

Although rule pruning and rule ranking approaches can help to eliminate redundant rules and obtain important rules, improving classifier performance, there has been little success with regard to discovering interesting or useful rules from an end user's point of view. In the real world, end users often consider the rules whose rule consequences contain one particular class. For example, while analyzing HIV/AIDS data, epidemiologists often concentrate on the rules whose rule consequences are HIV-Positive. Similarly, while mining banking data, bank loan officers often pay attention to the rules that classify a loan applicant into the "risky" class. Under this context, the present study proposes a novel and fast method for mining CARs with consideration of class constraints.

The contributions of this paper are stated as follows:

(1) A novel tree structure called the Novel Equivalence Class Rule tree (NECR-tree) is proposed for efficiently mining CARs with class constraints. Each node in NECR-tree contains attribute values and their information.
(2) Some theorems for quickly pruning nodes that are unable to generate rules satisfying the class constraints are developed.
(3) A novel and efficient algorithm for mining constrained CARs based on the provided theorems is presented.

The rest of this paper is organized as follows. In Section 2, some preliminary concepts of CAR mining are briefly given. Section 3 discusses work related to constrained association rule mining and CAR mining. The main contributions of the paper are presented in Section 4, in which the novel tree structure, called NECR-tree, is presented, and some theorems for eliminating unnecessary tree nodes are provided. The proposed algorithm for mining CARs with consideration of class constraints is also presented. Experimental results are discussed in Section 5. The conclusions and ideas for future work are given in Section 6.

2. Preliminary concepts

Let D be a dataset with d attributes $\{A_1, A_2, \ldots, A_d\}$ and n denote records (objects), where each record has an object identifier (*OID*). Let $C = \{c_1, c_2, \ldots, c_k\}$ be a list of class labels (k is the number of classes). Let *Constraint_Class* be a subset of C containing particular class labels considered by end users. A specific value of an attribute A_i and the mth record is denoted by a_{mi} ($m \in [1,n], i \in [1,d]$) and a specific value of class C is denoted by c_x ($x \in [1,k]$).

Definition 1. An *item* is described as an attribute and a specific value for that attribute, denoted by $\{(A_i, a_{mi})\}$ ($m \in [1,n], i \in [1,d]$).

Definition 2. An *itemset* is a set of *items*, denoted by $\{(A_i, a_{mi}), \ldots, (A_j, a_{mj})\}$ ($m \in [1,n], i,j \in [1,d]$, and $i \neq j$).

Definition 3. CAR R has the form $\{(A_i, a_{mi}), \ldots, (A_j, a_{mj})\} \rightarrow c_x$, where $\{(A_i, a_{mi}), \ldots, (A_j, a_{mj})\}$ is an *itemset* and $c_x \in C$ is a class label.

Definition 4. The actual occurrence $ActOcc(R)$ of rule R in D is the number of records of D that match R's antecedent.

Definition 5. The support of rule R, denoted by $Sup(R)$, is the number of records of D that match R's antecedent and R's consequent.

Definition 6. The confidence of rule R, denoted by $Conf(R)$, is defined as: