Expert Systems with Applications 39 (2012) 12841–12850

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Increasing the effectiveness of associative classification in terms of class imbalance by using a novel pruning algorithm

Wen-Chin Chen [a,b], Chiun-Chieh Hsu [a,*], Yu-Chun Chu [a]

[a] Dept. of Information Management, National Taiwan University of Science and Technology, Taiwan, ROC
[b] Marketing Department, Chunghwa Telecom Co. Ltd., Taiwan, ROC

ARTICLE INFO

Keywords:
Associative classification
Direct marketing
Rare events
Class imbalance
Scoring
Probabilistic classifiers

ABSTRACT

Having received considerable interest in recent years, associative classification has focused on developing a class classifier, with lesser attention paid to the probability classifier used in direct marketing. While contributing to this integrated framework, this work attempts to increase the prediction accuracy of associative classification on class imbalance by adapting the scoring based on associations (SBA) algorithm. The SBA algorithm is modified by coupling it with the pruning strategy of association rules in the probabilistic classification based on associations (PCBA) algorithm, which is adjusted from the CBA for use in the structure of the probability classifier. PCBA is adjusted from CBA by increasing the confidence through under-sampling, setting different minimum supports (*minsups*) and minimum confidences (*minconfs*) for rules of different classes based on each distribution, and removing the pruning rules of the lowest error rate. Experimental results based on benchmark datasets and real-life application datasets indicate that the proposed method performs better than C5.0 and the original SBA do, and the number of rules required for scoring is significantly reduced.

1. Introduction

As data mining literature highly prioritizes classification and association-rule discovery, the feasibility of integrating both approaches has been extensively studied in recent years. Focusing on a limited subset of association rules, i.e. those rules where the consequent of the rule is restricted to class variables allows for the construction of more accurate classifiers (Liu, Chen, & Chen, 1999; Liu, Hsu, & Ma, 1998). Several works have demonstrated the intuitiveness and effectiveness of associative classification (Dong, zhang, Wong, & Li, 1999; Jiang, Shang, & Liu, 2010; Liu, Han, & Pei, 2001; Liu, Jiang, Liu, & Yang, 2008; Liu et al., 1998; Thabtah, Cowling, & Hammoud, 2006; Wang & Zhou, 2000; Yin & Han, 2003; Yoon & Lee, 2007). Association rules normally search globally for all rules that satisfy *minsup* and *minconf* thresholds. The richness of the rules makes this approach highly promising for accurately reflecting the classification structure in data.

Associative classification was first proposed in CBA (Liu et al., 1998), in which the conventionally adopted *Apriori* algorithm has been implemented to extract a limited number of association rules with their consequents limited to class labels. These rules are then sorted by descending confidence and are pruned to obtain a minimal number of rules deemed necessary to cover training data and achieve an acceptable accuracy. Confidence is a reliable measure when classes are equally distributed. However, direct marketing-related topics, including customer churn and customer purchases, are often rare objects (Gupta et al., 2006). When class distributions differ significantly, the response rate to a product promotion or retention is often extremely low, e.g., 1–2% (Gupta et al., 2006; Neslin, Gupta, Kamakura, Lu, & Mason, 2006; Piatetsky-Shapiro & Masand, 1999; Yang & Wu, 2006). Therefore, this is not the most adequate approach to follow (Janssens, Wets, Brijs, & Vanhoof, 2005). However, this class imbalance has seldom been addressed in associative classification. Liu, Ma, and Wong (2003) offers an algorithm SBA using association rules to produce a score for the data case in order to reflect the likelihood that the data case belongs to a rare class. Although SBA performs better than C4.5 and Naïve Bayesian, this algorithm conducts rule pruning by using the pessimistic error-rate-based method in C4.5 (Quinlan, 1992), subsequently failing to reduce the number of rules and ultimately resulting in an over-complication upon scoring. However, Janssens et al. (2005) developed a rule-ranking index, intensity of implication, to address issues involving a classifier, excluding positive class rules. However, conducting related tests (as described later in Section 5) reveals that this index can not significantly improve the ranking of positive class rules. Also, this work individually constructs classifiers with 6 *minsup* and *minconf* thresholds, a method that can only create 6 points on the ROC curve, yet can not control the entire potential customer list. This work presents a novel rule

* Corresponding author. Tel.: +886 2 27376766.
E-mail address: cchsu@cs.ntust.edu.tw (C.-C. Hsu).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.009

pruning algorithm to enhance the SBA in order to accurately predict a rare event with associative classification. The proposed algorithm is adopted from CBA, and is then adjusted to the probability classifier structure, explaining why it is referred to hereinafter as PCBA. PCBA is implemented in the following steps. After the previously flawed ranking rule from CBA is addressed, under-sampling is proposed to increase confidence of the positive class rules, subsequently allowing the classifier to produce a sufficient number of positive class rules. The conventional association rule mining procedure uses a single *minsup* and a single *minconf* during data mining. This is inappropriate for this work because the class distribution of our data is often extremely imbalanced. New algorithm sets distinguish *minsup* and *minconf* for positive and negative class rules based on the ratio of positive and negative examples in the training set. Finally, as for the inability of the CBA structure classifier to control the direct marketing name list, the rule pruning step with the lowest error rate on the CBA is removed. Experiments using benchmark datasets and real-life application data indicate that the proposed algorithm can decrease the number of rules for the classifier, as well as predict positive examples much more than the original SBA and C5.0.

The rest of this paper is arranged as follows. Section 2 introduces associative classification. Section 3 then describes the proposed rule pruning algorithm, PCBA. Next, Section 4 discusses the evaluation index for the SBA prediction efficiency, while Section 5 summarizes the results of the empirical evaluation. Conclusions are finally drawn in Section 6, along with recommendations for future research.

2. Classification based on associations

The conventional CBA and SBA algorithms must be thoroughly reviewed before the proposed algorithm is introduced. Class association rules are introduced first, followed by a description of the ranking and pruning mechanisms in CBA. Scoring the data using SBA is then outlined. Finally, exactly how association classifier recognizes the positive examples is reviewed.

2.1. Class association rules

Denote $I = \{i_1, i_2, \ldots, i_k\}$ as a set of literals, called items. Also, denote D as a set of transactions, where each transaction T represents a set of items such that $T \subseteq I$ As is well known, transaction T contains X, i.e. a set of items in I, if $X \subseteq T$ An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contains X also contains Y. The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contains $X \cup Y$ Given a set of transactions D, mining association rules involves generating all association rules that have support and confidence greater than a user-specified *minsup* and *minconf* (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994).

To make association rules appropriate for the classification task, the associative classification method focuses on a unique subset of association rules, i.e. those rules with a consequent limited to class variables only, i.e. the so-called class association rules (*CARs*). Thus, only rules of the form $A \Rightarrow c_i$, where c_i denotes a possible class, are generated. As for the class imbalance issue, these data cases can be classified into a positive or negative class, which is separately represented with c_p and c_n. Thus, if the consequent of *CARs* is the positive class, i.e. $A \Rightarrow c_p$, this rule is referred to hereinafter as the positive class rule. In contrast, it is called the negative class rule.

2.2. Ranking and pruning of CARs in CBA

A classifier in CBA is constructed (Fig. 1) based mainly on a database coverage pruning method, which is applied after all *CARs* are generated. As for the first step of pruning, the algorithm ranks all *CARs* and then sorts them in a descending sequence. As shown in the next section, this rank is subject to a modification implemented previously. *CARs* is ranked as follows: given two rules r_i and r_j, $r_i > r_j$ (or r_i is assumed to have a higher rank than that of r_j), if (1) $conf(r_i) > conf(r_j)$; or (2) $conf(r_i) = conf(r_j)$, but $sup(r_i) > sup(r_j)$; or (3) $conf(r_i) = conf(r_j)$ and $sup(r_i) = sup(r_j)$; however, r_i is generated before r_j. Each training sample is classified by the rule covering it and has the highest ranking. The pruning algorithm attempts to select a minimum number of rule sets, with each correctly classifying at least one training sample, to cover the training dataset as well as achieve the lowest error rate. The default class is set as the majority class among the remaining samples that are not covered by a rule in the final classifier.

From the above description, we can infer that the sorting in CBA is quite important because the rules for the final classifier are selected by following the sorted sequence. CBA sorts its rules based on the conditional probability (confidence). This index is a reliable measure when classes are equally distributed. However, when class distributions significantly differ from each other, and especially for classes whose frequency is low, this is not the most feasible approach to adopt. Therefore, Janssens et al. (2005) offers a ranking index, intensity of implication, for this particular problem, as shown below:

$$1 - F\left(support \times |D| \times \left(\frac{1}{confidence} - 1\right)\right)$$

where F refers to a cumulative density function of Poisson distribution of the parameter $\lambda = \frac{support}{confidence} \times (|D| - |C_i|)$, and $|C_i|$ denotes the number of transactions containing c_i in D. Since λ is a multiple of $|D| - |C_i|$, when the positive class is a rare event, $|D| - |C_p|$ becomes markedly larger than $|D| - |C_n|$. Therefore, this adjustment can increase the intensity of implication of the positive class rules. Consequently, it resolves the problem concerning CBA prediction inaccuracy with class imbalance data. However, following testing, applying this ranking technique appears to be limited in efficiency, as discussed in detail later in Section 3.1.

2.3. Scoring based on associations

Class classifiers are constructed using the CBA introduced in the above section. Restated, CBA is feasible for predicting customer class, yet ineffective in terms of selecting a specific number of customers willingness to purchase under direct marketing applications. As a solution to associative classification's lack of efficient prediction with class imbalance, Liu et al. (2003) developed a probability classifier based on SBA, an algorithm based on association rules. First, the algorithm produces *CARs* with *apriori* (as explained in Section 2.1). *CARs* are then pruned rules using the pessimistic error-rate based method in C4.5. This approach prunes a rule as follows: If the estimated error rate of rule r exceeds that of rule r^- (as obtained by deleting a condition from the conditions of r), then rule r is pruned. Finally, the algorithm obtains score S from the weighted average of confidence taken from all rules satisfying each data case.

$$S = \frac{\sum_{i \in POS} W_p^i \times conf^i + \sum_{j \in NEG} W_n^j \times conf_p^j}{\sum_{i \in POS} W_p^i + \sum_{j \in NEG} W_n^j}$$

where

- *POS* denotes the set of positive class rules that can cover the data case,