Expert Systems with Applications 39 (2012) 5907–5919

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

# Adjusting and generalizing CBA algorithm to handling class imbalance

Wen-Chin Chen [a,b], Chiun-Chieh Hsu [a,*], Jing-Ning Hsu [c]

[a] Department of Information Management, National Taiwan University of Science and Technology, Taiwan
[b] Marketing Department, Chunghwa Telecom Co., Ltd., Taiwan
[c] Marketing Department, Data Communications Business Group, Chunghwa Telecom Co., Ltd., Taiwan

ARTICLE INFO

*Keywords:*
Associative classification
Direct marketing
Imbalance data
Class imbalance
Scoring
Probability classifiers

ABSTRACT

Associative classification has attracted substantial interest in recent years and been shown to yield good results. However, research in this field tends to focus on the development of class classifiers, but the required probability classifier of imbalance data has not been addressed comprehensively. This investigation presents a new associative classification method called Probabilistic Classification based on Association Rules (PCAR). PCAR is based on modifying the rule sorting index, the pruning method, and the scoring procedure in the CBA algorithm. CBA can be generalized to construct a probability classifier. Additionally, it can improve the efficiency of associative classification for predicting imbalance data. Experiments that use both benchmarking datasets and real-life application datasets reveal that the new method outperforms the previous associative classification algorithm and C5.0 for all datasets. Also, in some datasets, the predictive performance exceeds that achieved by logistic regression and the use of a neural network.



## 1. Introduction

Associative classification, one of the most important tasks in data mining and knowledge discovery, integrates two primary functions of data mining, which are classification and association-rule discovery. Focusing on a limited subset of association rules – those rules in which the consequent is restricted to the class variables – enables more accurate classifiers to be constructed. Several studies have established that associative classification is intuitive and effective in numerous cases (Dong, zhang, Wong, & Li, 1999; Jiang, Shang, & Liu, 2010; Liu, Hsu, & Ma, 1998; Liu, Han, & Pei, 2001; Wang & Zhou, 2000; Yin & Han, 2003; Yoon & Lee, 2007). The reasons for the favorable performance are obvious. Association rules search globally for all rules that satisfy minimum support and minimum confidence thresholds. They therefore contain the full set of rules, which may incorporate important information. The richness of the rules gives this technique the potential to capture the true classification structure of the data (Wang & Zhou, 2000). Associative classification is therefore gaining popularity. However, when class distributions vary significantly, it is not the most effective method (Janssens, Wets, Brijs, & Vanhoof, 2005).

In life, rare objects are commonly the most interesting. The same is true of data sets, which, after all, represent aspects of reality. Therefore, in data mining, rare objects are frequently of primary interest. Examples abound: they include predicting customer churn or customer purchases in marketing (Burez & Van den Poel, 2009; Kim & Street, 2004); identifying fraudulent credit card transactions (Chan & Stolfo, 2001); predicting pre-term births (Grzymala-Busse, Zheng, Goodwin, & Grzymala-Busse, 2000); and detecting oil spills from satellite images (Kubat, Holte, & Matwin, 1998). Studying rarity in the context of data mining is important because rare objects are typically much harder to identify than common objects, and only a few studies of associative classification have been published.

Liu, Ma, Wong, and Yu (2003) proposed an algorithm, called scoring based on associations (SBA), that exploits association rules to score the purchasing intentions of customers. This algorithm uses weighted scoring method to solve the issue that negative class rules are considered only because the proportion of positive examples in the training set is too low in the prediction of rare events when scoring is based on the best rule (SBR) (Liu et al., 2003), making the predictive performance of SBA better than that of C4.5 and the Naïve Bayesian method. However, this algorithm prunes rules by pessimistic error pruning (PEP) (Quinlan, 1992), and so many overlapping data cases that satisfy rules, raising the possibility of misevaluation in the scoring of a test set. Janssens et al. (2005) studied class imbalance classification and used a rule sorting index, "Intensity of implication", to mitigate the problem that the positive class rules are eliminated from a classifier. Chen, Hsu, and Hsu (2010) proved that the sorting index slightly improves the rank of the positive class rules. This study utilizes six groups of minimum support (minsup) and minimum confidence (minconf) to

* Corresponding author. Tel.: +886 2 27360617.
*E-mail address:* cchsu@cs.ntust.edu.tw (C.-C. Hsu).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.11.113

build a class classifier. This method can only generate six points on an ROC curve; it cannot control the number of potential customers. Finally, Chen et al. (2010) recommended a sorting index, confidence of undersampling, to solve the problem that positive class rules will not be chosen by the classifier because the original confidence in the prediction of rare events was too low. Moreover, this investigation solves the problem, that there was highly overlapped in both satisfied cases after pruning rules are applied with PEP, by applying Probabilistic Classification Based on Associations (PCBA), which is a rule pruning algorithm that revises CBA. With these modifications, SBA cannot only reduce the number of rules that are required but also increase the accuracy in predicting a positive class. However, the corrected SBA still uses the weighted scoring method; this not only underestimates high confidence values and the importance of rules that exceed minsup, but also eliminates the ability of associative classification is to interpret rare events easily.

SBA, which also applies PEP pruning ulres, is better than SBR because it adopts the weighted scoring method to solve the inefficiency problem of predicting rare events which are caused by the lack of under-sampling of SBR. Also, Chen et al. (2010) proved that it can improve SBA efficiency by pruning with PCBA, which coordinates with the confidence of undersampling sorting rules. Accordingly, it can replace the weighted scoring method of SBA with under-sampling. Based on the above considerations, this study develops a new algorithm, PCAR, with increased efficiency of associative classification for predicting rare events. This algorithm utilizes confidence of undersampling to sort class association rules. Furthermore, it is resulted from collocating PCBA to prune rules and modifying the SBR scoring method. The modifications of SBR are focused on replacing confidence with the rank sorting in order of descending the possibility that rules be applied to predict a positive example. Zero, the original scoring standard, is replaced with the score whose rank falls in between positive class rules and negative class rules in a test case that does not satisfy with any rule. This investigation proves that the customers in the positive class that are identified using the two algorithms are completely the same when the number of positive classes that is predicted by CBA equals that predicted using PCAR. Restated, PCAR generalizes CBA in that the number of predictive class, which is determined by class association rules, is two. Finally, experiments in which both benchmarking datasets and real-life application data are used demonstrate that the predictive performance of the new method in all datasets is better than that of the previous associative classification algorithm and C5.0; in some datasets, it is also better than that of logistic regression and the use of a neural network.

The rest of this paper is organized as follows. Section 2 introduces the algorithm of associative classification to predict rare events in the past. Section 3 describes the proposed classification algorithm, PCAR. Section 4 considers experimental data sets and discusses the index of PCAR prediction efficiency. Section 5 summarizes the results of the empirical evaluation. Section 6 finally draws conclusions and presents recommendations for future research.

## 2. Predicting rare events using associative classification

The major steps in using associative classification to predict rare events must be thoroughly reviewed before the proposed algorithm is introduced. Class association rules (CARs) are introduced first, and then pruning methods associated with CAR are described. CARs' sorting index is then elucidated. Finally, the exact way in which an association classifier recognizes positive examples is reviewed.

### 2.1. Class association rules

Let $I = \{i_1, i_2, i_k\}$ denote a set of literals, called items. Also, let D be 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 $X$, which is 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 $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in the transaction set $D$ with confidence $c$ if $c$ of the 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 the transactions in $D$ contain $X \cup Y$. Given a set of transactions $D$, mining association rules involves generating all association rules that have support and confidence greater than a user-specified and single minsup $s_{\text{minsupp}}$ and minconf $c_{minconf}$, respectively (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994). To make association rules suitable for the classification task, the associative classification method focuses on a unique subset of association rules – those rules with a consequent limited to class variables only; these are called the class association rules. Hence, only rules of the form $A \Rightarrow c_i$, where $c_i$ denotes a possible class, are generated.

With respect to the issue of class imbalance, these data cases can be classified into a positive or a negative class, represented by $c_p$ and $c_n$, respectively. Let $S(c_j)$ be the ratio of the number of data in class $j$ to the total number of data in the training set. Since $S(c_p)$ is obviously smaller than $S(c_n)$, for minimum support, the positive class rule will not be able to generate if the value of $s_{\text{minsupp}}$ is too high. If the value is too low, then overfitting will occur. For minimum confidence, inadequate positive class rules can be easily obtained if the value of $c_{minconf}$ exceeds the proportion of positive examples in the training set. In contrast, the positive class rules will be repeated in negative class rules. Liu et al. (2003) solved the above problems by setting different minsup and minconf values for positive and negative class rules (as in the following functions) based on the proportions of positive and negative examples in the training set.

$$\text{minsup}(c_i) = s_{\text{minsupp}} \times S(c_j)$$
$$\text{minconf}(c_i) = S(c_j)$$

Finally, if the consequent of CARs is the positive class, $A \Rightarrow c_p$, then the rule is referred to hereinafter as the positive class rule. Otherwise, it is called the negative class rule.

### 2.2. Rule pruning algorithms

Many association rules are well known to be redundant and minor variations of others. Such insignificant rules should be pruned. Pruning can remove a huge number of rules with no loss of accuracy. It also improves the prediction efficiency of classifier (Liu et al., 2003). In the past, the pruning of algorithms has been described as follows.

(1) Pessimistic Error Pruning (Quinlan, 1992): The pruning function uses the pessimistic error rate based on the method in C4.5. Notably, the error rate of a rule is 1 – 'the confidence of the rule'. The technique prunes a rule as follows. If a rule's estimated error rate is higher than the estimated error rate of the rule $r^-$ (obtained by deleting one condition from the conditions of $r$), then rule $r$ is pruned.
(2) Highest ranking pruning (HRP) (Liu et al., 1998) : In the first step of pruning, the algorithm ranks all CARs and then sorts them in order of descending sorting index. The sorting index will be discussed in the following section. Each training sample is classified by the rule that covers it and has the highest ranking. The pruning algorithm attempts to keep those rule sets, each of which correctly classifies at least one training