Knowledge-Based Systems 56 (2014) 79–85

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An effective ensemble pruning algorithm based on frequent patterns

Hongfang Zhou *, Xuehan Zhao, Xiao Wang

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form 29 October 2013
Accepted 30 October 2013
Available online 9 November 2013

Keywords:
Ensemble pruning
Frequent pattern
Large-scale dataset
Transactional database
Boolean matrix

ABSTRACT

Ensemble pruning is crucial for the consideration of both predictive accuracy and predictive efficiency. Previous ensemble methods demand vast memory spaces and heavy computational burdens in dealing with large-scale datasets, which leads to the inefficiency for the problem of classification. To address the issue, this paper proposes a novel ensemble pruning algorithm based on the mining of frequent patterns called EP-FP. The method maps the dataset and pruned ensemble to a transactional database in which each transaction corresponds to an instance and each item corresponds to a base classifier. Moreover, a Boolean matrix called as the classification matrix is used to compress the classification resulted by pruned ensemble on the dataset. Henceforth, we transform the problem of ensemble pruning to the mining of frequent base classifiers on the classification matrix. Several candidate ensembles are obtained through extracting base classifiers with better performance iteratively and incrementally. Finally, we determine the final ensemble according to a designed evaluation function. The comparative experiments have demonstrated the effectiveness and validity of EP-FP algorithm for the classification of large-scale datasets.

1. Introduction

Ensemble learning has been a very popular research topic during the last decade. It has attracted scientists from several research fields including Statistics, Machine Learning, Pattern Recognition and Knowledge Discovery in Databases [1,2]. A number of excellent algorithms have been proposed to solve real-world problems, such as Bagging [3], Boosting [4], AdaBoost [5], WAVE [6] and RFW [7]. Ensemble learning trains a collection of base classifiers and then combines some of them with a strategy to solve given classification or regression tasks [8,9]. Remarkable improvements of the ensemble in predictive accuracy and generalization performance compared to a single classification have been observed in a broad scope of application fields, such as face recognition [10], science image analysis [11,12], medical diagnosis [13,14], intrusion detection [15]. However, with the rapid growth in the size of data, a large number of base classifiers are required to construct the ensemble, which would brings heavy computational burdens for the ensemble learning. So, traditional ensemble methods show inefficient classification performance in deal with large size datasets due to massive memory spaces and more executive time.

Generally speaking, typical ensemble learning is built in two phases: the production of multiple base classifiers and the combination of them to get an ensemble. Thanks to the widely accepted knowledge of that ensemble with a small size but carefully selected subset could work better than ensemble of all [16], the later could be expanded in another perspective which selects base classifiers with better performance instead of complete ensemble. This phase is commonly called ensemble pruning, selective ensemble thinning or ensemble selection, of which we use the first one in the paper. The major difference between ensemble pruning and other ensemble methods is that its main focus is to reduce memory and computational cost by choosing small sub-ensembles for classification, as well as maintaining or increasing predicative accuracy. It mainly has two aspects of superiority compared with ensemble methods: (1) the improvement of generalization ability and (2) the reduction of predictive overhead [17].

In order to solve the problem of classification for large size datasets, this paper proposes an ensemble pruning algorithm based on the mining of frequent patterns called EP-FP (Ensemble Pruning based on Frequent Patterns). The key ideas of EP-FP mainly consist of two points. One is mapping the dataset and base classifiers to a transactional database in which each transaction corresponds to an instance and each item corresponds to a base classifier. The other is deploying the mining of frequent base classifiers on a Boolean matrix which saves the classification resulted by the pruned ensemble. Combining with the transactional database and Boolean matrix, we transform the problem of ensemble pruning to frequent patterns mining. EF-FP algorithm extracts some base classifiers with better predictive performance to construct the final ensemble classifier.

The remaining of the paper is organized as follows: Section 2 briefly reviews the related work of ensemble pruning. Section 3 presents the transformation between the problem of ensemble

* Corresponding author. Tel.: +86 18966606688.
E-mail address: zhouhf@xaut.edu.cn (H. Zhou).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.10.024

pruning and the mining of frequent patterns. The proposed algorithm EP-FP is discussed in details in Section 4, including the related concepts and the framework of algorithm. Section 5 gives the experimental results. The final conclusion is drawn in Section 6.

2. Ensemble pruning

An ensemble is usually more accurate than a single learner, but existing ensemble methods often tend to construct unnecessarily large ensembles, which increases the memory consumption and computational costs. Ensemble pruning tackles this problem by selecting a subset of ensemble members to form sub-ensembles with less resource consumption and better performance than the original one.

Zhou et al. first puts forward the theory of ensemble pruning and confirm that ensembling a subset of classifiers could be better than using all of the classifiers. They propose the algorithm GASEN based on the genetic algorithm (GA) by evolving the voting weights of individual classifier in an neural networks and selecting the classifiers with weights above a certain threshold to form a sub-ensemble [16]. Although obtaining the optimal ensemble pruning has been proved to be a NP-complete problem [18,19], Zhang et al. has formulated the problem as a quadratic integer programming problem with the sub-ensemble in the optimal accuracy–diversity balance. It obtains approximate solutions in polynomial time by applying semi-definite programming techniques [20]. Another method relies on ordered aggregation, where original classifiers are reordered according to some criteria, and sub-ensemble is constructed according to the order. A representative of this approach can be referred in Ref. [21], in which base classifiers are ordered by the increasing values of angles between the signature vectors and the reference vectors. Moreover, Greedy algorithm is also one of the extensively investigated topics related to ensemble pruning. It processes quickly only considering a small subspace among all the possible combinations. However, this would cause a strong possibility of suboptimal solution of the ensemble pruning [22].

The proposed method in this paper belongs to a family of ensemble pruning methods based on the mining of frequent pattern, which is a vital topic in the relative research fields of data mining and machine learning. Its emphasis is to find meaningful relationships between item and item for the transactional database or relation database [23]. Pattern-based selection is a surprised strategy for ensemble pruning, which combines the technology of pattern mining and ensemble pruning. Currently, the research in this field has just begun and a lot of knowledge is required to explore such as the generation of transactional database and the mining of useful patterns [24]. Pattern-based ensemble pruning applies the technology of transactions processing to pick up better base models and then combines them to construct the final ensemble classifier. The most important is organizing the classifications resulted by pruned ensemble on test datasets as the transactional database and finding a certain collection of base models with superior predictive performance. PMEP [25], the most representative method, utilizes the structure of FP-tree to save classification results and extract frequent base classifiers. It adopts the strategy of majority voting to complete the selection of ensemble. Experiments performed demonstrate that pattern-based ensemble pruning shows good predictive performance.

3. Problem transformation

Inspired by the theory of pattern-based ensemble pruning, this paper proposes a novel method to solve the problem of classification for large-scale datasets with the aid of transactional database and Boolean matrix. We map the dataset and base classifiers to a transactional database, in which each transaction stands for an instance and each item stands for a right-classified base classifier for the instance. Boolean matrix is used to compress classifications resulted by the set of base classifiers, in which each row corresponds to an instance, and each column corresponds to a base classifier.

Let $D = \{d_1, d_2, \ldots, d_n\}$ be a dataset with n instances, the attributes of instance d_p $(1 \leqslant p \leqslant n)$ is composed of the pruned ensemble $C = \{c_1, c_2, \ldots, c_m\}$ with m base classifiers, i.e. $d_{pq} = c_q$ $(1 \leqslant p \leqslant n,\ 1 \leqslant q \leqslant m)$. Suppose $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items for the transactional database $T = \{t_1, t_2, \ldots, t_n\}$. Dataset D is mapped to the transactional database T, where transaction t_p $(1 \leqslant p \leqslant n)$ is corresponding to instance d_p and item i_q $(1 \leqslant q \leqslant m)$ is corresponding to base classifier c_q, Additionally, Boolean matrix $M = \{(d_p, c_q) = \{0, 1\} | 1 \leqslant p \leqslant n, 1 \leqslant q \leqslant m\}$, instances as rows and base classifiers as columns, is used to save the classification resulted by the ensemble C on dataset D. Element as 1 means that instance d_p could be predicted rightly by base classifier c_q, but 0 means the opposite. An example of the transactional database and Boolean matrix for a certain database with eight instances is shown in Table 1 and Fig. 1 respectively.

4. Proposed algorithm

4.1. Related concepts

In order to describe the proposed approach clearly, we define several key concepts and equations used in this paper as follows.

Definition 1 (*Support*). Given a set of base classifiers $X \subseteq C$, the support of X denoted by $Supp(X)$ is defined as the proportion of the maximal set $R(X) \subseteq D$ classified rightly by X. It is calculated as:

$$Supp(X) = \frac{|R(X)|}{|D|} \tag{1}$$

where $|D|$ is the number of instances in dataset D.

Definition 2 (*Classification matrix*). Classification matrix is a special Boolean matrix which has removed the rows with all the elements being total 1 or total 0.

Definition 3 (*Row array*). An array which length is equal to the size of dataset is used to save the decimal representations corresponding to the binary results of classification for the dataset. An example is illustrated in Fig. 1.

Definition 4 (*Classification difficulty*). Classification difficulty is used to measure the degree of being classified rightly for an instance. It is calculated as:

Table 1
Transactional database.

Tid	Item set
t_1	i_3
t_2	i_1, i_2, i_3, i_4
t_3	i_1, i_3
t_4	i_1, i_3
t_5	i_1, i_3
t_6	i_2, i_3
t_7	
t_8	i_1, i_3, i_4