

# Study of Ensemble Pruning Algorithms Based on the Idea of Pattern Mining

Yanping Li\*

Department of Computer and Information Engineering, Heze University, Heze, Shandong, 274015, China

**Abstract:** At first the thesis introduces a new concept: coverage-based pattern mining. Then based on this concept, two ensemble pruning algorithms, CPM-EP and PMEP are put forward. By the algorithms we get candidate sub-patterns, and applies coverage pattern mining and a greedy strategy to acquire candidate target ensemble. Experiments show that pattern mining is an efficient selective integration strategy, able to obtain predictive ability, small-scale target integrated classifier, the paper presents that the CPM-EP and PMEP algorithm is able to meet the demand for online learning.

**Keywords:** Chip images, filtering algorithm, noise.

## 1. INTRODUCTION

All along, machine-learning techniques have played a huge role in many fields. The ultimate goal of machine learning is to derive useful information and knowledge in order to guide the subsequent decision-making by analyzing the data. With the popularity of the Internet and the increasing data acquisition means, the amount of data people can get grows exponentially, which pose a challenge for traditional machine learning techniques. As for such transactions as online transactions over the Internet, online advertising, financial analysis, as well as search engine, it is of great significance to quickly and effectively study the large-scale, prolonged and persistent data [1-3]. Online machine learning is an important means for timely processing large amounts of data. Prediction performance and Prediction efficiency become the most important evaluation criteria for online learning methods.

In order to get the small-scale target integrated classifier with great generalized ability, this paper puts forward strategy of coverage-based pattern mining. Based on this strategy, two selective ensemble pruning algorithms, CPM-EP and PMEP are put forward. Experiments prove that pattern mining is an efficient selective integration strategy, able to obtain predictive ability, small-scale target integrated classifier, thus getting the small-scale target integrated classifier of great prediction ability. The two algorithms of CPM-EP and PMEP can meet the requirements of online learning.

## 2. SELECTIVE ENSEMBLE PRUNING

### 2.1. Selective Ensemble

Selective Ensemble is an added middle stage between the construction phase of base classifiers and the merger stage of results during the whole integrated learning process, that is, base classifier selection phase. The base classifier construction

and the prediction combination method used in the selective ensemble are similar to the ensemble learning, so the selective ensemble focuses on the study of the selection method of based classifiers. For a given set of training samples, the first step is to produce a number of base classifiers, and then use a validation set to cut the base classifiers. In the prediction period the prediction results of the new sample will be obtained through the combination of the cut integrated classifier.

A formalized description of the selective integration problems that will be discussed in this paper are shown as follows. A base classifier set  $BC = \{h_1, h_2, \dots, h_M\}$ , and a calibration sample set  $VS = \{X_1, X_2, \dots, X_N\}$  are given. Among them,  $M$  as the sum of base classifiers, and  $N$  is the sum of Validation Set. The goal of selective ensemble is to find a subset of  $BC$ , i.e. [4-6],  $PR$ , and to make the integrated classifier of  $PR$  the best metric of  $VS$ . One of the most commonly used metric is the Prediction accuracy, therefore it is:

$$PR = \arg \max \{Accuracy(ES, VS, \Psi) \mid ES \subseteq BC\} \quad (1)$$

Among them, Accuracy ( $ES, V_s, \Psi$ ) refers to the prediction accuracy acquired from the prediction of the samples in  $VS$  with the help of the combination strategy  $\Psi$ .  $PR$  is just the target integrated classifier of the selective ensemble.

With the increasing number of base classifiers, the number of possible combination of base classifiers increases exponentially, so it is unrealistic to get a collection of the best base classifier in the brute-force method. Most of the existing selective ensemble algorithms choose the heuristic method to reduce the number of combination that needs to be checked [7-10]. Most of these algorithms have the following characteristics: (1) in the selection process, the base classifier is regarded as the basic unit of algorithms. There are too many base classifier combinations to be checked, so the speed of algorithms is slower. In addition, because of the base classifier as the basic unit, it is unavoidable there exists some sets of base classifiers of poor performances within the ensemble classifier, which may lead to the overall poor prediction performance of the ensemble classifier; (2) For a lot

\*Address correspondence to this author at the Department of Computer and Information Engineering, Heze University, Heze, Shandong, China; Tel: 13034551125; E-mail: [lyp5506@126.com](mailto:lyp5506@126.com)

of selective ensemble algorithms (such as GASEN, FS, and EPRL, etc.), it is very necessary to each need to look at the of, the algorithm need to recalculate the ensemble prediction accuracy (or other metrics) of every base classifier combination, which makes the computational expenses larger; (3) algorithms need to continually use the detailed prediction results of all base classifiers towards VS, in order to meet the demand of the prediction accuracy of constantly calculating the new ensemble classifier. If the problems are of large scale (the number of base classifiers and the check samples are very large), algorithms require a lot of memory space to save the prediction results of the base classifier, thus the space expenses will be large. In order to solve these problems, we propose the use of pattern mining techniques in selective ensemble.

## 2.2. The Strategy of Selective Ensemble Pruning Based on Pattern Mining

If a base classifier set, which has classified every check sample correctly, is regarded as a transaction, the corresponding transactions of these all samples can comprise a transaction database TD [11-14]. While by processing the transaction database, we can easily obtain the base classifier combination (pattern) which appear frequently. On the basis of the above idea, we make the following reasonable assumptions: in the frequently appearing patterns in TD, the prediction capabilities of their classifiers are likely to be positively related to each other, and if these patterns are combined, their prediction capabilities should also be strong; however, for those patterns which did not appear in TD, their prediction capability is likely to be negatively correlated, and the prediction capability of such modes should be rather poor. We can say that the larger the scale of TD is, the more reliable this assumption is. Obviously, the base classifier combinations of good predictive performance can be obtained by prediction combining these modes of high prediction capability. So the selective ensemble is the best mode combination.

The strategy of selective ensemble based on pattern mining has the following characteristics: (1) problems of the selective ensemble are turned into the mode mining problems in transactional database; (2) the target mode is regarded as a combination of several sub-modes; (3) the prediction ability of the target mode is the statistical function of each sub-mode involved in the combination, and can be calculated directly.

## 3. THE ALGORITHM OF SELECTIVE ENSEMBLE PRUNING BASED ON PATTERN MINING

### 3.1. The Framework Ensemble Pruning Algorithm Based on Pattern Mining

Based on the above ideas, we give a framework of ensemble pruning algorithm based on a Pattern Mining, just as shown in Algorithm 1 below. The framework describes the main steps of the selective ensemble strategy: The first step is to get the transaction database TD; the second step is to obtain the candidate sub-modes; the third step is to generate several candidate patterns. Each candidate pattern is a set of one or more candidate sub-modes, and in order to generate a

candidate mode, it needs to select some parts of the corresponding candidate sub-modes according to the evaluation; the fourth step is to select the best one from the candidate modes as the target mode according to the selected evaluation criteria (e.g. prediction accuracy, diversity, etc.). Algorithm 1 just describes the key idea of the framework in a general way; the other different algorithms based on this framework have their own distinct realization of these steps.

*Algorithm 1: Selective Ensemble algorithm framework based on pattern mining*

*Input:*

*BC: the set of base classifiers*

*VS: check sample set*

*Output:*

*PR: target ensemble classifier*

(1) *To show the results of each base classifier predicting the check samples as the transaction database TD;*

(2) *To obtain candidate sub-modes and their statistics;*

(3) *To determine the combination of the candidate sub-mode, to generate several candidate modes;*

(4) *To calculate a certain metric of the candidate modes according to the statistics of the candidate sub-modes and based on this metric value select the best candidate modes as the target pattern PR;*

(5) *Return PR.*

### 3.2. CPM-EP Algorithm

#### 3.2.1. CPM-EP Algorithm Ideology

The possible scale of target ensemble classifier (target mode) is the integers in the range  $[1, M]$ , so a candidate ensemble classifier (candidate mode) can be generated for each possible scale, and then according to a certain standard one classifier is selected from all the candidate ensemble classifiers as the ultimate goal. Suppose the candidate mode with a scale  $k$  is expressed by  $P_k$ . In the process of generating  $P_k$ , according to majority voting principle, only some parts of the candidate sub-mode with the length  $\lfloor k/2 \rfloor + 1$  can be combined. But for the base classifiers which don't appear in these candidate sub-modes, there is no opportunity to be selected into the candidate mode  $P_k$ .

According to the majority voting rules, when the ensemble classifier on a scale of less than or equal to  $k$  carries out a sample class prediction, if the number of base classifiers which have been correctly predicted is more than or equal to  $\lfloor k/2 \rfloor + 1$ , the ensemble classifier must be able to correctly classify the sample. According to this rule, firstly on the basis of the original transaction database TD we create a new transaction database  $TD_k$ , among which all transactions contain  $\lfloor k/2 \rfloor + 1$  items. So, if one transaction in  $TD_k$  can be covered by the candidate mode  $P_k$ , it means that at least  $\lfloor k/2 \rfloor + 1$  base classifiers in the mode  $P_k$  make a right classification about the corresponding samples. According to the majority voting rules, the ensemble classifier  $P_k$  makes a correct class prediction about this sample. While  $P_k$  is obtained in the coverage pattern mining method, that is:

$$Pk = CPM(TDk, k) \tag{2}$$

After the acquisition of all candidate mode  $Pk$  ( $1 \leq k \leq L$ ), from which to select the mode of the highest prediction accuracy of the check sample set will be selected as the target ensemble classifier, which is just the result of selective ensemble. In the following part we will describe the CPM-EP algorithm and its key steps in detail.

**3.2.2. PM-EP Algorithm Description**

Algorithm 2 shows the pseudo-code description of the algorithm of the CPM-EP. The algorithm input includes the base classifier set BC and check sample set VS, and the output is the target ensemble classifier PR. Firstly PR is initialized(the first line in Algorithm 2); then a transaction database TD is created according to the sample classification results of each base classifier; then all candidate patterns are created on the basis of the different mode scales (the third line in Algorithm 2), and the mode of the best prediction ability is selected as the target ensemble classifier.

Algorithm 2: CPM-EP(BC, VS)

Input: BC, the base classifier set, wherein the number of base classifiers is expressed by M.

VS, the check sample set, wherein the number of samples is represented by N

Output: PR, the target ensemble classifier

(1) PR =  $\Phi$ ; ncorrect=0; // Initialization

(2) TD= et\_predictions (BC, VS); // Acquire the description of the prediction results of the transaction database

(3) for each k in [1, M]

(3.1) TD '= refine\_transactions (TD, k, Ck); // delete the corresponding transactions which can easily classify samples

(3.2) TDk = generate\_new\_transactions (TD ', k); // create a new transaction database TDk for Scale k

(3.3) Pk = CPM (TDk, k); // use coverage mode mining techniques to obtain candidate modes Pk

(3.4) Sk = Coverage (Pk, TDk) + Ck; // to obtain the estimates of the prediction ability of the mode Pk

(3.5) if (Sk > ncorrect)

(3.5.1) PR = Pk; // record the best candidate mode

(3.5.1) ncorrect = Sk; // record the estimates of the prediction ability of the current best candidate mode

(4) return PR

**3.2.3. Simplification of Prediction results**

Sample  $X_i$  ( $1 \leq i \leq N$ ) corresponding affairs contains  $M_i$  base classifiers, if  $M_i$  meets:

That is to say, for given model length k, no matter what kind of base classifiers is chosen, in Pk there exist at least +1 base classifiers which predict sample  $X_i$  correctly. According to the majority voting principle, the ensemble results of Pk is surely right for  $X_i$ . If Sample  $X_i$  meets:

$$M_i < \lfloor k / 2 \rfloor + 1 \tag{3}$$

No matter what base classifier it is in Pk, the number of base classifiers which have been correctly predicted  $X_i$  is less than  $\lfloor k/2 \rfloor + 1$ . According to the majority voting rules, the ensemble classifier Pk makes a wrong ensemble prediction about  $X_i$ .

because the content of Pk will not affect their prediction result, in order to improve Pk's generalization ability and reduce the time needed for selection, we will delete the corresponding transactions of those samples which can meet the requirements of (2) and (3), getting a simplified transaction database. Assume that there are Ck samples meeting the requirement (2), Wk samples meeting the requirement (3), then the prediction accuracy of the candidate mode Pk to VS meet the requirement (4):

$$\frac{C_k}{n} \leq accuracy \leq 1 - \frac{W_k}{n} \tag{4}$$

Assume that k = 5, then:

$$\lfloor k / 2 \rfloor + 1 = \lfloor 5 / 2 \rfloor + 1 = 2 + 1 = 3$$

Because Sample  $X_1$ ,  $X_6$  and  $X_{16}$  meet the requirement (2), Sample  $X_4$ ,  $X_{13}$  and  $X_{14}$  meet the requirement (4), their corresponding transactions will be removed from TD. Table 1 gives the simplified results, in which the gray color is used to show the deleted transactions.

**Table 1. Concise results of transactions.**

Transactions Numbers	Item of transactions	Items No.
$X_1$	<i>h1, h2, h3, h6, h7, h8</i>	6
$X_2$	<i>h2, h3, h4, h5, h7</i>	5
$X_3$	<i>h2, h5, h6</i>	3
$X_4$		0
$X_5$	<i>h1, h2, h6, h8</i>	4
$X_6$	<i>h1, h2, h3, h4, h5, h6, h7, h8</i>	8
$X_7$	<i>h5, h6, h7</i>	3
$X_8$	<i>h2, h5, h7</i>	3
$X_9$	<i>h3, h4, h5, h7</i>	4
$X_{10}$	<i>h1, h2, h5, h6</i>	4
$X_{11}$	<i>h2, h5, h6</i>	3
$X_{12}$	<i>h1, h2, h4, h6, h7</i>	4
$X_{13}$	<i>h1</i>	1
$X_{14}$	<i>h1, h2</i>	2
$X_{15}$	<i>h4, h5, h6</i>	3
$X_{16}$	<i>h1, h2, h3, h4, h5, h6, h8</i>	7

**3.3. PMEP Algorithm**

PMEP algorithm is an optimization of the CPM-EP algorithm to reduce the selection time of the base classifiers. Al-

gorithm 3 gives a description of PMEP algorithm, its differences from CPM-EP algorithm are mainly reflected in the distinct acquisition methods of the simplified transaction database TD' and the transaction table TTK. PMEP directly obtains a streamlined transaction database from TD (Line 3 in Algorithm 3), then creates a FP-Tree (Line 4 in Algorithm 3). After that, for each possible length  $k$  of target modes, PMEP algorithm directly gets TTK from the FP-Tree, and then on this basis carries on the coverage pattern mining (the input equivalent to the coverage pattern mining is a transaction table and then it's no need to generate the transaction table. we use CPM\* to show the coverage pattern mining based on this modification), In this way, we can save the selection time to a certain extent. In PMEP algorithm, there is no explicit operation of acquiring TDk, but its content is directly implied in TTK. Because the operation of the coverage pattern is carried out directly in TTK (algorithm), thus avoiding the operation to obtain TTK for every value of  $k$ , thereby saving the amount of computation.

*Algorithm 3. PMEP (BC, VS)*

*Input: BC the base classifier set, wherein the number of base classifiers is expressed by M*

*VS, check sample set, wherein the number of samples is represented by N*

*Output: PR, the target ensemble classifier*

(1)  $PR = \Phi$ ;  $ncorrect = 0$ ; // Initialization

(2):  $TD = \text{get\_predictions}(BC, VS)$ ; // Get a description of the transaction database of predicted results

(3)  $TD' = \text{refine\_transactions}(TD, 1, C)$ ; // delete the corresponding transaction to the easily classified samples

(4) *Tree Build, FP-Tree (TD')*; // create an FP-Tree for TD'

(5) *for each k in [1, M]*

(5.1)  $TTk = \text{generate\_combined\_database}(Tree)$ ; // get the transaction table TTK from the FP-Tree

(5.2)  $Pk = CPM * (TTk, k)$ ; // obtain candidate patterns Pk with the help of the coverage mode mining techniques

(5.3)  $Sk = \text{Coverage}(Pk, TTK) + C$ ; // mode Pk predictive ability estimates

(5.4) *If (Sk > ncorrect) then*

(5.4.1)  $PR = Pk$ ; // record the current best candidate mode

(5.4.2)  $ncorrect = Sk$ ; // record the estimates of the prediction ability of the current best candidate mode

(6) *return PR*

Compared with CPM-EP algorithm, algorithm PMEP further optimizes the efficiency of the implementation. Their difference lies mainly in the following two points: (1) The prediction results are streamlined: As we have discussed above, for the given mode length  $k$ , should not affect the contents of PK. Specifically speaking, the items in the transactions of these samples should not be counted into the frequency information of the items, which will not affect the item order in transactions. Algorithm CPM-EP for each pos-

sible mode length  $k$ , will delete the corresponding transactions of the sample which can meet the above two conditions, so each  $k$  obtains the different simplified transaction database, which avoids the impact of the above samples. Algorithm PMEP uses the same simplified transaction database TD' for every possible mode length. in the process of simplification it just deletes corresponding transactions of the samples which meet the conditions  $M_i = 0$  and  $M_i = M$ . Therefore, there may still be redundant transactions, which have a certain influence on Pk's results. (2) create a transaction table: for a given mode length  $k$ , the transaction table in algorithm CPM-EP is obtained by combining the same transactions in the database TDk. The transaction table algorithm PMEP is directly obtained from FP-Tree, and it's unnecessary to generate transaction database TDk for each different  $k$ .

## 4. ANALYSIS OF EXPERIMENTAL RESULTS

### 4.1. Experimental Algorithm

In the experiment, in addition to the CPM-EP and PMEP algorithms proposed above, the other algorithms involved in the comparison include: Select Best, Bagging. And other typical selective ensemble algorithms in different categories include: GASEN, Forward Selection, Orientation Ordering, Margin Distance Minimization, MDSQ and CPF algorithm.

Both CPM-EP and PMEP adopt the coverage pattern mining method. In the selection ensemble experiments, we will analyze and compare the performance of the selection ensemble based on pattern mining and other selection ensemble methods from all aspects, and at the same time verify the effect of the optimization method adopted by the PMEP algorithm by experiments.

### 4.2. Experimental Methods

Experiments use the cross-validation method for ten times, that is, the data set is randomly divided into 10 parts. In every cross, one part is taken as a test set and the other 9 parts as the training sets (known as cross-training set). Each cross totally trains 100 base classifiers and the training set of every base classifier is obtained from the cross-training set by using bootstrap to resample. In order to improve the diversity of the base classifiers, we set the size of the base classifier training set as half of the cross-training set. The check sample set used by selective ensemble is just the cross-training set.

### 4.3. The Size of The Target Ensemble Classifier

Fig. (1) gives the sizes of target ensemble classifiers obtained in the selective ensemble method. From the Figure it can be seen that the target ensemble classifiers generated by CPM-EP are the smallest, with an average value of 7.36. The results of the algorithms PMEP and FS is equal to that of CPM-EP algorithm, with their average value 7.44 and 7.48 respectively. For CPM-EP and PMEP algorithms, their ensemble classifiers in the 12 data sets are of the same scale, and the results in other data sets are similar, which shows the results of CPM-EP and PMEP algorithms have the great similarity. GASEN uses the preset threshold  $\lambda$  to select base classifiers and the result is 9.57. Algorithm OO uses the vec-

