Incremental Naïve Bayesian Learning Algorithm based on Classification Contribution Degree

Shuxia Ren *, Yangyang Lian
School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin, China
t_rsx@126.com , sunshine_lianyy@qq.com

Xiaojian Zou
Military Transportation University, Tianjin, China
ala_99@126.com

***Abstract*—In order to improve the ability of gradual learning on the training set gotten in batches of Naive Bayesian classifier, an incremental Naïve Bayesian learning algorithm is improved with the research on the existing incremental Naïve Bayesian learning algorithms. Aiming at the problems with the existing incremental amending sample selection strategy, the paper introduced the concept of sample Classification Contribution Degree in the process of incremental learning, based on the comprehensive consideration about classification discrimination, noisy and redundancy of the new training data. The definition and theoretical analysis of sample Classification Contribution Degree is given in this paper. Then the paper proposed the incremental Naïve Bayesian classification method based on the Classification Contribution Degree. The experimental results show that the algorithm simplified the incremental learning process, improved the classification accuracy of incremental learning.**

***Index Terms*—incremental learning, Classification Contribution Degree, Naïve Bayesian**

I. INTRODUCTION

Data mining has been becoming more and more popular with all walks of life, from manufacturing to service industry, from medical diagnose to business decision, because its special advantage in discovering the potential decision information and business model [1]. Classification prediction is an important content in the research of data mining, which has been widely used in classification prediction and pattern recognition.

There are many classification methods such as decision tree, Naïve Bayesian classifier, support vector machine, neural network and so on. Naïve Bayesian classification method is widely used for its simple and efficiency. Its classification performance is comparative to neural network, decision tree and better than many other classifier in some cases [2]. The classification rules trained on a large scale of dataset with strong completeness can accurately response model of the practical problem. But in variety of applications such as medical decision making, stock analysis, industrial production, data are produced over time, and the training set are gained patched [3].

In order to ensure the classifier gotten by the Naïve Bayesian classification method more accuracy, the classifier must be retrained again and again whenever the training set is updated as the time goes by. The problems of wasting time and memory resources arise because of repeat training [4].

Incremental learning can amend the current classifier with knowledge learning from the new training set [5]. It can make full use of the previous trained classifier, making the previous work meaning. Meanwhile, incremental learning only needs new training set in memory, making the amending process fast and efficient. Compared with training a new system, amending a trained system will save lots of resources like time, space, manpower and material. Moreover, incremental learning can solve some problems of real-time application more accurately on the basis of small-scale study.

Incremental learning for a mining algorithm, especially the classification mining algorithms, is a very important ability. Many studies of incremental learning ability were down with many classification methods like RBF neural network, Support vector and k-Nearest Neighbor [6-9]. And the applications of incremental classification focus on predicting bugs in software, pattern classification, mining of frequent time sequence, analysis of medical data features, collaborative filtering and so on [10-12].

The research of incremental ability with naïve Bayesian classifier was first proposed by Gong Xiujun. He presented the learning formula of incremental Naïve Bayesian and selection strategy of amending samples [13]. Many improving methods on the select range of amending sample and sequence are made based on this paper. There are three kinds of incremental Naïve Bayesian classification algorithms.

The first kind of algorithm uses all the samples in the new training set for amending the current classifier, but prefer for samples whose sum loss of classification is lesser, which was first proposed by Gong. The second kind of algorithm predicts samples in the new training set with the current classifier first, and uses samples that are

doi:10.4304/jcp.9.8.1967-1974

correctly classified for amending the current classifier [14]. The third kind of algorithm also predicts samples in the new training set with the current classifier first, and uses samples that are wrongly classified to amend the current classifier. In the process of amending, these algorithms still prefer for samples whose sum loss of classification is lesser [15]. Correctly classified samples are samples whose predicted class labels are the same with their own class labels. And wrongly classified samples are samples whose predicted class labels are different from their own class labels.

The problems of three kinds of incremental Naïve Bayesian classification algorithms are as follows:

First, in the incremental learning process, the algorithms do not fully consider the factor of data representation, data redundancy and noise. Weak characterized data and redundant data will blur the category features. And the noisy data will decline the performance of the classifier. Obviously, it is not advisable to use all the new training samples to amend the current classifier [16]. In fact, whether a new instance is suit for amending the current classifier, we should consider the data qualities that affect the predict result, the representativeness, discrimination degree and redundancy of the classified information, comprehensively, instead of whether the sample is rightly or wrongly predicted.

Second, when the order of samples is different, the learning sequence obtained by the sum loss of classification is different with the same training set [17]. There are two extreme cases for example. When samples in new training set are descending according to the classification sum loss, all samples will be used to amend the current classifier. When samples in new training set are ascending according to the classification sum loss, only the first sample will be used to amend the current classifier.

In order to overcome the above problems, the concept of Classification Contribution Degree is presented with comprehensive consideration of classification discrimination, redundancy and noisy of the data, and then the Incremental Naïve Bayesian Learning Algorithm based on Classification Contribution Degree (INB-CCD) is proposed. The algorithm analyzes the effect on classification performance from the level of data quality, and avoids the dependence on sample order of learning result.

The rest of the paper is organized as follows. Section 2 presents the Naïve Bayesian classification method and further gives the learning formula of incremental Naïve Bayesian classification. In Section 3 the paper first gives the definition and theoretical analysis of Classification Contribution Degree, and then proposes the sample selection strategy of Naïve Bayesian incremental learning and at last gives the medical data mining algorithm based on Classification Contribution Degree. The effectiveness of the algorithm is studied in Section 4 through comparing its performance on five different medical datasets with the other three incremental learning algorithms. Finally, the paper concludes in Section 5.

II. Learning Formula of Incremental Naïve Bayesian

A. Naïve Bayesian Classification Mtheod

Naïve Bayesian classification method is mainly based on the Naïve Bayesian principle. So before the introduction of incremental Naïve Bayesian formula, we first present the Naïve Bayesian theory here. It gives the calculation of post probability according to the prior probability. Prior probability is the probability distribution of each class counted by the given data information. It contains two statistics: class prior probability and conditional probability. Post probability is the probability of each unclassified sample belonging to some certain class.

Suppose $X = \{A_1, A_2, ..., A_n\}$ is a data sample with n features, then post probability of sample X belonging to class C is $P(C \mid X)$. The Naïve Bayesian formula of calculating the post probability is as follows:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{1}$$

Where, P(C) is the class prior probability of a sample belonging to class C. P(X|C) is conditional probability of features same with sample X among samples whose class label is C.

The main idea of Naïve Bayesian classification method is as follows:

Suppose $C_1, C_2, ..., C_m$ represents the m different classes. For each test sample X, the classification method computes the posterior probability $P(C_j \mid X)$ through class prior probability $P(C = C_j)$ and class conditional probability $P(X \mid C_j)$. The value range of j is from 1 to m. And sample X belongs to the class whose posterior probability is the biggest of the all [18]. That is to say, Naïve Bayesian classification method classified the sample X as class C_i, when the sample X satisfied the following expression:

$$P(C_i \mid X) > P(C_j \mid X),\ 1 \le i \ne j \le m \tag{2}$$

We can learn from (1) that (2) equals to the following expression:

$$\frac{P(X \mid C_i)P(C_i)}{P(X)} > \frac{P(X \mid C_j)P(C_j)}{P(X)} \tag{3}$$

Where, $1 \le i \ne j \le m$

Moreover, (2) is equal to the following expression:

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j) \tag{4}$$

Where, $1 \le i \ne j \le m$

Therefore, the essence of training Naive Bayesian classifier is to calculate two parameters using statistical knowledge according to the training samples: prior probability and conditional probability.

B. Incremental Naïve Bayesian Formula

The process of classification forecasting with Naïve Bayesian classifier is as follows:

Firstly, calculating the posterior of test sample belonging to each class with the two parameters obtained from the new training information using Naïve Bayesian theory, and then dividing the test sample to the class with the largest posterior.

Therefore, the task of incremental learning of Naive Bayesian classifier is equal to how to determine the new prior probability and class conditional probability according to the prior information and new training information. The incremental amending process is shown in Fig. 1.

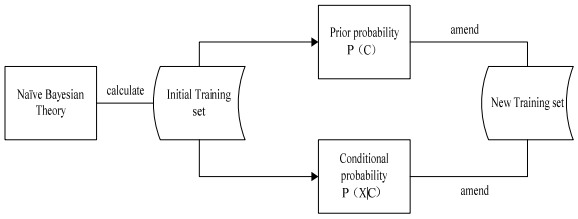

Figure 1. Incremental Amending of Naïve Bayesian classifier

It can be learnt from the Fig. 1 that, correction process of the incremental learning of Naïve Bayesian classifier is actually a recursive Bayesian estimation of parameters. Its advantage is that it can preserve the information in initial training data in the form of parameters. During the process of incremental learning, the system does not need to visit the raw data, but only needs to access the two saved statistic parameters. According to the information in new training set, the system amends and resaves the two statistic parameters. So, the incremental learning formula of Naïve Bayesian classifier is deduced by the statistical knowledge.

Suppose D is the current training set, T is the new testing set, $x_p = (A_1, A_2, ...A_i, ..., A_n) \in T$ is the new instance for amending, and c_p means its class label. $\theta_j = P(c = c_j)$ is class prior probability of class label c_j . Then the amending formula of class prior probability is:

$$\theta_j^{'} = \begin{cases} \frac{s}{s+1}\theta_j + \frac{1}{1+s} & \text{when } c_p = c_j \\ \frac{s}{s+1}\theta_j & \text{when } c_p \neq c_j \end{cases} \quad (5)$$

Where, $s = |D| + |T|$

$|D|$ represents number of samples in training set D, $|T|$ represents number of samples in new testing set T.

Suppose $\theta_{ik|j} = P(A_i = a_k \mid c = c_j)$ is class prior probability of attribute A_i of value a_k in class c_j . Then the amending formula of class conditional probability is:

$$\theta_{ik|j}^{'} = \begin{cases} \frac{m}{1+m}\theta_{ik|j} + \frac{1}{1+m} & \text{when } c_p = c_j \text{ and } A_i = a_k \\ \frac{m}{1+m}\theta_{ik|j} & \text{when } c_p = c_j \text{ and } A_i \neq a_k \\ \theta_{ik|j} & \text{when } c_p \neq c_j \end{cases} \quad (6)$$

$$m = |A_i| + count(c_j)$$

$|A_i|$ represents the numbers of values for attribute A_i , $count(c_j)$ represents number of samples whose class label is c_j .

Then the incremental amending of Naïve Bayesian classifier is completed after the modification of the two statistical parameters according to the (5) and (6). In order to illustrate the incremental learning process of Naive Bayesian classifier with two statistic parameters clearly, Fig. 2 shows incremental learning model of Naive Bayesian classifier with the three probability values of Naive Bayesian theory.

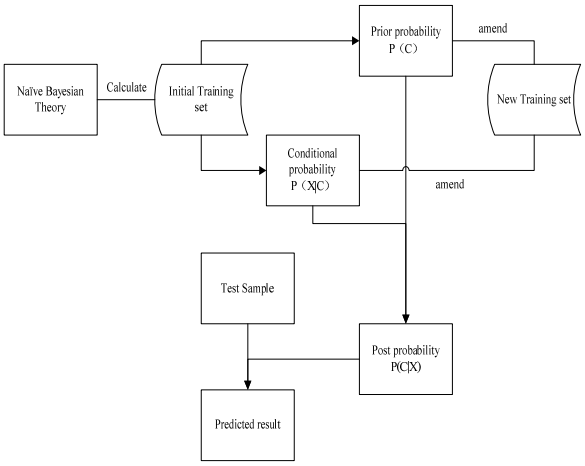

Figure 2. Incremental learning model of Naive Bayesian classifier

III. INCREMENTAL NAÏVE BAYESIAN CLASSIFICATION ALGORITHM

A. The Definition and Theory Analysis of Classification Contribution Degree

This paper defined the classification earnings obtained by the amending sample, namely the degree of reduction between classification results and sample observations, as classification contribution degree, which is denoted as con . Suppose $T(< x_1, c_1 >< x_i, c_i > ... < x_p, c_p >)$ is the new samples set. The testing class label of sample x_i is $c_i^{'}$ obtained by the current classifier, and $c_i^{''}$ obtained by classifier amended with sample $< x_k, c_k >$. Then the classification contribution degree of new testing sample x_k is:

$$con_k = \sum_{i \neq k} (\overline{c_i c_i^{'}} - \overline{c_i c_i^{''}}) \tag{7}$$

$\overline{cc^{'}}$ is distance between testing class label $c^{'}$ and observation class label c, and the value of distance $\overline{cc^{'}}$ is:

$$\overline{cc^{'}} = \begin{cases} 0, & c = c^{'} \\ 1, & c \neq c^{'} \end{cases} \tag{8}$$

Therefore,

$$\overline{c_i c_i^{'}} - \overline{c_i c_i^{''}} = \begin{cases} -1 & \text{when } \overline{c_i c_i^{'}} = 0 \text{ and } \overline{c_i c_i^{''}} = 1 \\ 1 & \text{when } \overline{c_i c_i^{'}} = 1 \text{ and } \overline{c_i c_i^{''}} = 0 \\ 0 & \text{when } \overline{c_i c_i^{'}} = \overline{c_i c_i^{''}} = 0 \text{ or } 1 \end{cases} \tag{9}$$

Through the analysis of (9), there are three cases about the new classifier amended by a sample. If the testing class label changes from 0 to 1, then the classification contribution degree of this sample is -1, and the sample has a negative influence to the improving of classification accuracy. If the testing class label turns from 1 to 0, then the classification contribution degree of this sample is 1, and the sample has a positive influence to the improving of classification accuracy. If the testing class label changes from 0 to 0 or from 1 to 1, then the classification contribution degree of this sample is 0, and the sample makes no difference on the improving of classification accuracy.

It can be learnt from (8) that the classification accuracy of current classifier is:

$$\frac{count(\overline{c_i c_i^{'}} = 0)}{|T|}, 1 \leq i \leq |T|$$

And the classification accuracy of classified amended by sample $< x_k, c_k >$ is:

$$\frac{count(\overline{c_i c_i^{''}} = 0)}{|T-1|}, 1 \leq i \neq k \leq |T|$$

Then the classification earnings or accuracy improvement of sample $< x_k, c_k >$ is:

$$\frac{count(\overline{c_i c_i^{''}} = 0) - count(\overline{c_i c_i^{'}} = 0)}{|T-1|}, 1 \leq i \neq k \leq |T|$$

Obviously, the sample with the largest classification earnings is most favorable for classification accuracy improvements.

The theoretical proof of classification contribution degree is as follows:

Expression①:

$$\max\{\frac{count(\overline{c_i c_i^{''}} = 0) - count(\overline{c_i c_i^{'}} = 0)}{|T-1|}\}$$

$$\Leftrightarrow \max\{(1 - \frac{count(\overline{c_i c_i^{''}} = 1)}{|T-1|}) - (1 - \frac{count(\overline{c_i c_i^{'}} = 1)}{|T-1|})\}$$

$$\Leftrightarrow \max\{\frac{count(\overline{c_i c_i^{'}} = 1) - count(\overline{c_i c_i^{''}} = 1)}{|T-1|}\}$$

Expression②:

$$\max\{\frac{count(\overline{c_i c_i^{''}} = 0) - count(\overline{c_i c_i^{'}} = 0)}{|T-1|}\}$$

$$\Leftrightarrow \max\{-(\frac{count(\overline{c_i c_i^{'}} = 0) - count(\overline{c_i c_i^{''}} = 0)}{|T-1|})\}$$

$$\Leftrightarrow \min\{\frac{count(\overline{c_i c_i^{'}} = 0) - count(\overline{c_i c_i^{''}} = 0)}{|T-1|}\}$$

Comprehensive expression ① and ②:

$$\max\{\frac{count(\overline{c_i c_i^{''}} = 0) - count(\overline{c_i c_i^{'}} = 0)}{|T-1|}\}$$

$$\Leftrightarrow \max \sum_{\leq i \neq k \leq |T|} \overline{c_i c_i^{'}} - \min \sum_{\leq i \neq k \leq |T|} \overline{c_i c_i^{''}}$$

$$\Leftrightarrow \max \sum_{1 \leq i \neq k \leq |T|} (\overline{c_i c_i^{'}} - \overline{c_i c_i^{''}}) \Leftrightarrow \max con_k$$

Visibly, the sample classification contribution degree can response the improved degree of classification by using the sample to amend the current classifier, namely the classification earnings gotten by the amending sample.

B. The Selection Strategy and Model of Incremental Learning

Because the amending sample sequence gotten by classification sum loss has a strong dependence on sample order in new training set, so the incremental learning result with the same training set may be different. This section comprehensively analyzes the factors affecting the prediction result in order to get a better selection strategy of amending samples. These factors are qualities of data, classification discrimination, representativeness, noisy and redundancy of training information, mentioned in section I.

There are two results using current classifier to test the new testing sample set: rightly classified and wrongly classified. The rightly classified samples demonstrate the current training set contains the classification information carried by the rightly classified samples in the new testing set. However, there are two reasons for wrongly classified samples. One result is that the data completeness of current training set is bad. The current training set doesn't contain the classification information carried by the wrongly classified samples. The other result is that the wrongly classified sample is noisy data.

In order to avoid the decline of classification accuracy and feature weakening brought by data noisy and redundancy, enriching the current training set with the strong classification representative samples, and avoiding the dependence on sample order of learning result, the paper amends the current classifier with two samples whose classification contribution degree is the largest among all rightly classified samples and wrongly classified samples separately.

As is known from (7), the sample with the largest classification contribution degree is satisfied with the following condition.

$$con_t \geq con_p \tag{10}$$

Where, $1 \le t, p \le |T|, t \ne p$

Obviously, sample with the largest classification contribution degree can improve the classification performance of the classifier best among all the samples. That is to say, sample with the largest classification contribution degree has strong representativeness, carries more classification information among all the samples in new training set. And it must not be noisy and redundancy data. Therefore, enriching the current training set with this sample can enrich the classification information of the training set effectively.

The main idea of Incremental learning based on classification contribution degree is that: testing the new training set with the current classifier first, putting instances whose class label are same with their tested label to the rightly classified sample set, and putting instances whose class label are different from their tested label to the wrongly classified sample set; then amending the current classifier with the samples whose classification contribution degree is the largest among all the rightly tested samples and wrongly tested samples separately.

Incremental learning model based on classification contribution degree is showed in Fig. 3.

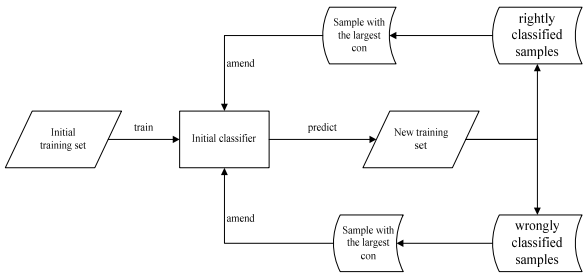

Figure 3. Incremental learning model based on classification contribution degree

C. Incremental Naïve Bayesian Classification Algorithm

The incremental Naive Bayesian classification algorithm based on the classification contribution degree first predicts samples in the new training set with the current classifier, and adds the rightly classified samples to set 'Tright', wrongly classified samples to set 'Terror'. Then the algorithm calculates classification contribution degree of each sample, and puts the samples with the largest classification contribution degree among sets 'Tright' and 'Terror' respectively to set 'Update'. Finally the algorithm modifies the current classifier with samples in set 'Update' according to the incremental Naive Bayesian formula. The process of algorithm is as follows:

Input: current training set D, current classifier nb, new training set T

Output: amended classifier nb'

Process:

Step1: predicting samples in the new training set with current classifier nb

Step2: putting samples that are rightly classified into set 'Tright', samples that are wrongly classified into set 'Terror'

Step3: computing classification contribution degree con_i of each sample $x_i \in Tright$, putting sample t_1 with the largest classification contribution degree into set 'Update'

Step4: computing classification contribution degree con_j of each sample $x_j \in Terror$, putting sample t_2 with the largest classification contribution degree into set 'Update'

Step5: amending the current classifier with samples in set 'update', and getting the amended classifier nb'.

The work flow of the incremental Naive Bayesian classification algorithm based on the classification contribution degree is shown in Fig. 4.

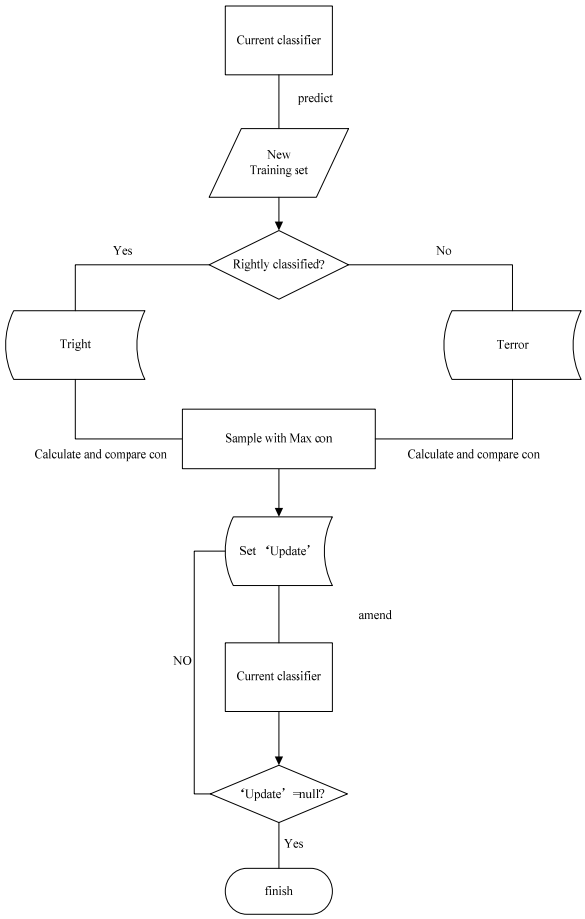

Figure 4. Flow chart of incremental Naïve Bayesian classification algorithm based on classification contribution degree

IV. EXPERIMENT

In order to verify the performance of the algorithm, the experiment is completed on five medical datasets. The five medical datasets are respectively Wisconsin Breast Cancer, Pima Diabetes, Bupa Liver Disorders, Hepatitis and Statlog heart disease. These datasets can be downloaded from website http://www.datatang.com/. And the detail information of the five medical datasets is shown in table 1.

TABLE I.

DETAIL INFORMATION FOR THE FIVE MEDICAL DATASETS

Name of datasets	size of dataset	Number of attributes	Number of classes
Wisconsin Breast Cancer	699	10	2
Pima Diabetes	768	9	2
Bupa Liver Disorders	345	7	2
Hepatitis	155	20	2
Statlog – heart	270	14	2

In the process of experiment, each medical dataset is divided into five parts in order to complete incremental learning. Four parts are used for initial training set, the other part is used for new training set. The usage of the five medical datasets is shown in table2. In order to ensure the objectivity of the experimental data, the experiment repeats the above process five times by using the idea of cross validation. And the experiment uses each part of five parts as the new training set in turn for each time, and the rest as initial training set. The final classification accuracy of each algorithm is taken as the average value of its five times experiments.

TABLE III.

THE USAGE OF THE FIVE MEDICAL DATASETS

Name of datasets	Number of dataset	Number of initial training set	Number of new training set
Wisconsin Breast Cancer	699	599	100
Pima Diabetes	768	614	154
Bupa Liver Disorders	345	276	69
Hepatitis	155	124	31
Statlog – heart	270	216	54

To demonstrate the performance of the modified sample's selection strategy of this algorithm, the experiment firstly trains the current classifier with the same Naive Bayesian classification method in the initial training sets of each medical dataset. The experiment uses Particle Swarm Optimization for feature selection in initial Naïve Bayesian classifier training process. Then the experiment uses algorithms presented in article [19-21] and INB-CCD proposed in this paper to complete incremental learning on those new training sets. At last the algorithm performance is illustrated by using the classification accuracy.

The main idea of incremental learning in article [19] is that: All new samples are used to amend the current classifier. But in the process of selecting the sample, priority selection is the sample with less sum loss of classification, until all the new samples are added to the current training set. For the convenience of description, incremental Naïve Bayesian classification algorithm of article [19] is marked to INB - L in this paper.

The main idea of incremental learning in article [20] is that: Firstly, predicting the samples in the new training set is completed. Secondly, amending the current classifier is completed by using the rightly classified samples. In the process of selecting the sample, priority selection is also the sample with less sum loss of classification, until all the new samples are added to the current training set. For the convenience of description, incremental Naïve Bayesian classification algorithm of article [20] is marked to INB - R in this paper.

The main idea of incremental learning in article [21] is that: Firstly, predicting the samples in the new training set is completed. Secondly, amending the current classifier is completed by using the wrongly classified samples. For the convenience of description, incremental Naïve Bayesian classification algorithm of article [21] is marked to INB-EL in this paper.

The final classification accuracy of each algorithm is shown in table3.

TABLE II.

THE FINAL CLASSIFICATION ACCURACY OF EACH ALGORITHM

Name of datasets	INB-CCD	INB-R	INB-EL	INB-L
Wisconsin Breast Cancer	98.92%	98.57%	98.57%	97.8%
Pima Diabetes	81.43%	80.13%	80.46%	77.08%
Bupa Liver Disorders	71.01%	60.14%	62.32%	62.03%
Hepatitis	98.57%	98.33%	89.84%	96.54%
Statlog - heart	87.96%	87.04%	87.96%	86.3%

The final classification accuracy line chart of each incremental algorithm is shown in Fig. 5 (in %).

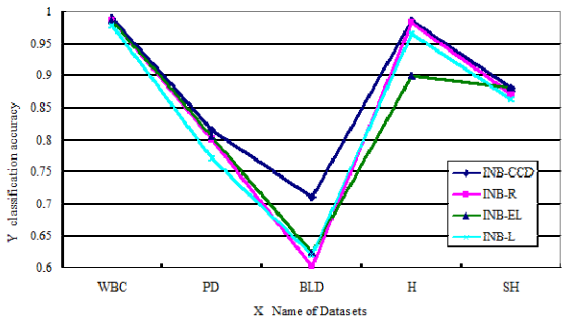

Figure 5. The final classification accuracy of each algorithm

For convenient, WBC is short for medical dataset Wisconsin Breast Cancer, PD is short for medical dataset Pima Diabetes, BLD is short for medical dataset Bupa Liver Disorders, H is short for medical dataset Hepatitis, and SH is short for medical dataset Statlog-heart in Fig. 5.

The analysis of Fig. 5 shows that:

(1) The classification accuracy of algorithm INB-EL is better than INB-L beside dataset Hepatitis. That is to say, amending the current classifier with rightly classified samples in the new training set is better than using all the samples in the new training set in most cases.

(2) The classification accuracy of algorithm INB-R is better than INB-L beside dataset Bupa Liver Disorders. It demonstrates that amending the current classifier with wrongly classified samples in the new training set is better than using all the samples in the new training set for most time.

From conclusion (1) and (2), we can learn that not all the samples are suit for incremental amending the current classifier, because data with less classification information will reduce the discrimination degree of classification, data's noisy and redundancy will decline the performance of the classifier.

(3) The classification accuracy of algorithm INB-R is better than INB-EL on dataset Hepatitis, but worse on dataset Pima Diabetes and Statlog-heart. It is clearly that algorithm INB-R and INB-EL are not always perform better for different dataset. That is to say, whether a sample is suit for amending the current classifier or not, we should consider the quality, information representative of each sample data, meanwhile the samples is rightly or wrongly predicted by the current classifier.

(4) The classification accuracy of algorithm INB-CCD is better than the other three algorithms on the five datasets. It demonstrates that the algorithm presented in this paper has stronger applicability for all the datasets, and improves the performance of the initial classifier effectively. The learning sample selection strategy based on classification contribution degree can avoid the data redundancy and noisy, meanwhile it can eliminate the dependence on the order of the new samples for the amending samples which are the same even if the order of the samples in new training set are different.

There is a simple analysis on the time complexity of the incremental Naïve Bayesian algorithm proposed in this paper and a comparison with the other three algorithms. The main operation of each incremental algorithm is comparison and amendment.

In the selection process of learning samples, the four algorithms all need N times compare, so the compare time complexity of each algorithm is O (N). And in the final amending process, the amending times of the three incremental learning algorithms based on sum loss of classification is between 1 and N, so their amend time complexity is O (N). However, because there are only two samples used for finial amending, so the amend time complexity of incremental learning algorithm proposed in this paper is O (2). Visibly, the incremental Naive Bayesian learning algorithm based on the classification contribution degree proposed in this paper not only improves the classification accuracy effectively, but also simplifies the amending process of incremental learning.

V. CONCLUSIONS

This paper discussed the importance of incremental learning ability of classification algorithms with the real-time, dynamic datasets, and launched the research of incremental learning ability with Naive Bayesian classification method. The paper first introduced Naïve Bayesian theory, Naïve Bayesian classification method and then presented the incremental learning formula of Naïve Bayesian classification method based on the analysis of the essence of incremental learning.

Through the analysis of the common problems on the existed amending sample selection process and the factors of classification performance from the level of data quality, INB-CCD algorithm is proposed. The paper gives the definition and theoretical analysis of classification contribution degree, which is earning of the classification gotten by an amending sample.

The amending sample selected by INB-CCD algorithm can avoid the negative influence of redundancy and noisy data and reduce the weakening of classification information brought by the samples whose data representation is not obvious. Only two samples are used to amend the initial classifier, which avoided the dependence on sample order, simplified the learning process and improved the speed of incremental learning.

The comparative analysis on the experimental result demonstrates that the new algorithm proposed in this paper has better data completeness and higher classification performance. The next research work will focus on these two problems. One is whether the algorithm can complete the current training set in greatest extent or not, the other is how to gain the classified information more effectively.

REFERENCES

[1] Xu Yingming, Wei Yongqing and Zhao Jing, "Incremental learning method of Bayesian classification combined with feedback information," Computer Application, vol. 31, Sep. 2011.

[2] HAN JOON KIM and JAEYOUNG CHANG, "Integrating Incremental Feature Weighting into Naïve Bayes Text Classifier," Proc. the sixth International Conference on Machine Learning and Cybernertics, 2007, pp. 1137-1143

[3] Shuling Di, Hui Li and Pilian He, "Incremental Bayesian Classification for Chinese Question Sentences Based on Fuzzy Feedback," Proc. 2010 2nd international conference on future computer and communication, vol. 1, 2010, pp. 401-404

[4] Phimphaka Taninpong and Sudsanguan Ngamsuriyaroj, "Incremental Adaptive Spam Mail Filtering Using Naïve Bayesian Classification," Proc. 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing, 2009, pp. 243-248

[5] Zi Yuan, Lili Yu, Chao Liu and Linghua Zhang, "Predicting Bugs in Source Code Changes with Incremental Learning Method," Journal of Software, Vol. 8, Jul 2013, pp. 1620-1633, doi:10.4304/jsw.8.7.1620-1633

[6] Hong Men, Lei Wang and Haiping Zhang, "Electronic Nose For The Vinegar Quality Evaluation By An Incremental RBF Network," Journal of Computers, Vol 7, Sep. 2012, pp. 2276-2282, doi:10.4304/jcp.7.9.2276-2282
[7] Xiaopeng Hua and Shifei Ding, "Incremental Learning Algorithm for Support Vector Data Description," Journal of Software, Vol 6,Jul 2011, pp. 1166-1173, doi:10.4304/jsw.6.7.1166-1173
[8] Shuli Han, Yujiu Yang and Wenhuang Liu, "Incremental Learning for Dynamic Collaborative Filtering," Journal of Software, Vol 6, Jun 2011, pp. 969-976, doi:10.4304/jsw.6.6.969-976
[9] Shaowei Wang, Hongyu Yang and Haiyun Li, "Facial Expression Recognition Based on Incremental Isomap with Expression Weighted Distance," ournal of Computers, Vol 8, Aug. 2013, pp. 2051-2058, doi:10.4304/jcp.8.8.2051-2058
[10] Jiyun Li, Junping Wang and Hongxing Pei, "Data Cleaning of Medical Data for Knowledge Mining," Journal of Networks, Vol. 8, Nov. 2013, pp. 2663-2670, doi:10.4304/jnw.8.11.2663-2670
[11] Shih-Yang Yang, Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, "Sun Incremental Mining of Across-streams Sequential Patterns in Multiple Data Streams," Journal of Computers, Vol. 6, Mar. 2011, pp. 449-457, doi:10.4304/jcp.6.3.449-457
[12] Hu Yin-E. and Ke Luo, "A selective naïve bayesian classification algorithm based on rough set," Proc. 2013 2nd International Conference on Manufacturing Engineering and Process, ICMEP 2013, vol. 325-326 2013, pp. 1593-1596
[13] Dewan Md. Farid and Chowdhury Mofizur Rahman, "Mining Complex Data Streams: Discretization, Attribute Selection and Classification," Journal of Advances in Information Technology, Vol 4, Aug 2013, pp. 129-135,doi:10.4304/jait.4.3.129-135
[14] Ozdikis Ozer, Senkul Pinar and Sinir Siyamed, "Confidence-based incremental classification for objects with limited attributes in vertical search," Proc. 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012, 2012, pp. 10-19.
[15] xiaopeng hua, Shifei Ding. "Incremental Learning Algorithm for Support Vector Data Description," Journal of Software, 2011, Vol. 6, No. 7, pp. 1166-1173.
[16] Yang, Fengqin, Zhou, Wanting and Wu, Di, etc., "An improved incremental Naïve Bayesian for Sentiment Classification," ICIC Express Letters, vol. 7, 2013, pp. 959-964
[17] Ouyang Zehua, Guo Huaping and Fan Ming, "Incremental learning of Naive Bayes parameter on gradual contracting space," Journal of Computer Applications, vol. 32, 2012, pp. 223-227
[18] Han Jiawei, "Data mining: concepts and techniques," Beijing: China Machine Press, 2012
[19] Gong XiuJun, Liu ShaoHui and Shi ZhongZhi, "An Incremental Bayes Classification Model," Journal of Computers, vol. 25, Jun. 2002, pp. 645-650
[20] Li Jinhua, Liang Yongquan and L Fangfang, "An Incremental Learning Model of Weighted Naive Bayesian Classification," Computer and Modernization, May 2010
[21] Li Xiaoyi and Xu Zhaodi, "Priciple and Algorithm of Incremental Bayes Classification," Journal of Shenyang Agricultural University, vol. 42, Jun. 2011, pp. 349-353

Shuxia Ren is born in Hegang, China, Sep. 1973. She received her bachelor's degree in computer application from Shenyang Polytechnic University, in 1997. She received her master's degree in software theory & application from Jinan University, in 2003. Currently she is an Associate Professor at Tianjin Polytechnic University. Her interests are in Data mining and special database.

Yangyang Lian is born in Handan, China, Jul. 1988. She received her bachelor's degree in school of mathematics and information science & technology from Hebei Normal University of Science & Technology, in 2011. And she is a research student of school of Computer Science and Software Engineering of Tianjin Polytechnic University Currently. Her major is Computer Technology and her interests are in data mining.

Xiaojian Zou is born in Dehui, China, May. 1978. He received his bachelor's degree in Information and Computing Science from the Jilin University, in 2003. He received his master's degree in operational research & cybernetics from the Nankai University, in 2011. Currently, he is a lecturer at Military Transportation University. His research interests include intelligent control and intelligent algorithm.