Cross Regional Resource Allocation Model and Solving Scheme Based on Different Consumer Groups

Yan Ma[1], Bin Ding[1], Qi Huang[1], Wei Xiao[1], Chengli Xu[2]

1 Heilongjiang Bayi Agricultural University
Heilongjiang Daqing, 163319 China
2 Huazhong Normal University
Hubei Wuhan, 430079 China

***Abstract* — In allusion to the unusable problem in the case of continuously generated data caused by operation of traditional integrated algorithm usually in batch-mode in distance education, an algorithm of classifier combination and increment- based integrated performance predictions for distance education students is proposed. Firstly, this paper has given a brief introduction of three concerned integrated classification algorithms: Naive Bayesian incremental version, 1-NN and WINNOW algorithm. Then, in the training data set, three algorithms are used to generate their own hypotheses. Finally, three hypotheses are integrated to predict the performance of students using the method of voting. Experiment of training set HOU provided by "information" courses of Greek distance learning universities has verified the validity and reliability of the proposed algorithm, and the experimental results show that compared to several other more advanced classifiers, the algorithm has achieved better precision of classification and less training time, which has provided strong prediction tools of students' performance for teachers.**

Keywords - integrated combination and increment; distance education; students' performance prediction; classifier; voting method

I. INTRODUCTION

The main feature of distance education is that teachers will not guide students in the classroom, but they complete teaching of learning materials based on the individual design, as well as the communication between students and teachers. In distance education, students often feel isolated, thus communication between students and teachers as well as other students is an important indicator of the successful program of distance education [1-2].

What distance education addresses is adults' special educational needs or discordance (age, profession and family responsibility, etc.) between students. Predictions of student performance [3], especially in environment of distance education, have gradually attracted people's attention, and the use of data mining and machine learning algorithm is a quite promising development direction able to achieve this purpose. In order to achieve true intellectualization and amenability, system model of teachers should learn from the data. However, the data set is so small that machine learning methods cannot be directly applied. Literature has solved this problem, giving the general method of creating precise classifier of educational data, and the success rate of the experiment that literature describes the prediction course is higher than 80% .

Literature has proposed a students' classification algorithm to predict students' final grades from the features extracted from web-based education system login data, and the analysis indicates that the Genetic Algorithm (GA) can improve and merge the accurate performance of classifier successfully, an increase than of about 10% -12% than non-GA classifiers. However, most of the integration algorithm is run in batch mode, that is, repeated reading and process of the entire training set. For instance, in the case of continuous generated data of educational environment, data storage of batch learning is impractical, so that these integrated learning algorithms are simply unavailable. The incremental learning ability is very important for machine learning methods of solving practical problems, and there are two main reasons: 1) it is impossible to collect all useful training examples before putting training system into use; 2) it takes a little time to modify the training system than to build a new system, which is particularly useful for real-time applications.

Based on the above analysis, it is proposed to use the voting method combined with Naive Bayesian incremental version, 1-NN and WINNOW algorithm, and the integrated algorithm proposed in this paper is used to predict the students' performance in distance learning system. Experimental results show that the proposed algorithm is useful in predicting students' performance, able to identify poor performance of students, but also enables the teachers to take preventive measures at an early stage. Even at the beginning of the school year, in order to provide additional help for at-risk groups, this paper more accurately diagnoses that the students' ability of performance can be enhanced with the adding of new curriculum data in semester, which has provided more effective results for teachers.

II. BACKGROUND

This chapter gives some basic theories, such as educational data mining, online learning algorithm and incremental integrated classifier conducted to predict students' performance.

Classification, clustering, visualization, association rule and statistical mining are usually to discover new, interesting and useful knowledge based on the students' use of data, mainly used in electronic learning problems or objects, able to handle and evaluate students' performance, to provide adaptive curriculum and learning suggestions based on students' learning performance, to handle and evaluate learning materials and web-based educational programs, to provide e-learning courses feedback for teachers and students, and to supervise atypical students' learning performance.

Classification (one of the most useful educational data mining tasks in e-learning) can be used for various educational objectives, such as [4]: the grouping of those students who prompt driving or error-driven, the identification of some usual misconception for these students, and the prediction in use of intelligent tutoring system / classification of students.

Online learning task is to obtain a set of concept descriptions from the time-stamped training data distribution, and such learning is crucial to many applications, such as computer security, intelligent user access and market basket analysis. Customers' preferences will change with the new products so as to make the service available. The response of conceptual shift algorithm must quickly and accurately converge to the new objective concept, but shall be valid in time and space [5].

Online learning algorithm processes each training instance once, and does not require storage and pretreatment, maintaining the current impact. For all hypotheses of training examples currently, such algorithm is also useful for large dataset, and the price of batch method is quite high for big data with requirements of multiple passes. Neural network learning batch algorithm will traverse dataset for multiple times, but neural network of online learning only needs to traverse data once. However, only once traversal of the data may have some related losses. All these algorithms have a known disadvantage, and namely, to carry out the study of a few examples is very difficult, and in order to solve this problem, some techniques rely on the window technique, which includes the final stored examples, and learning task is implemented only once when new examples are added. WM is the basis for many online algorithms, and it maintains weight vector of specialists set, and the prediction is output by weighted majority voting among experts. Vote perception can store more information during the training period, and then this elaborate information is used to make better prediction on the test set.

The concept of combined classifier is a new direction to improve the performance of classification. However, how to directly apply integrated approach is not very clear in line setting. One solution is to rely on users to specify the number of examples in of each basic learner's input stream, but the method assumes numerous cases about the structure of the data flow. There are also online upgrade algorithms of reweighted classifier, but these algorithms assume that the number of classification is fixed. Moreover, when the basic model is trained by a few cases, online upgrade may initially suffer a great loss, and the algorithm may never recover.

III. ALGORITHM PROPOSED

As we all know, the choice of optimized set of classifier is an important part of multiple classifier system, and the independence of classifier output is usually taken as an advantage that can get better multiple classifier system. In merge option of classifier, the voting method requires that the classifier has not any precondition [6].

When multiple classifiers are combined using the voting method, if most experts agree with them, they will make the right decisions, and based on the reliance of this point, this paper looks forward to better results [7].

Currently, there are three concerned integrated learning algorithms:

(1) Core of WINNOW algorithm is similar to perception. If $\sum_i x_i w_i > \theta$, it will classify a new instance x to class 2, or to class 1. However, if the predicted class is correct, WINNOW updates its weight as follows; if the predicted value is $y' = 0$ and the actual value is $y=1$, then the weight is too small. Therefore, as to such features for each $x_i = 1$, the weight is $w_i = w_i \cdot \beta$. Among it, the promotional parameter β is greater than 1. If $y' = 1$ and $y = 0$, then the weight is too large. Therefore, for each feature $x_i = 1$, $0 < \beta < 1$ in the weight $w_i = w_i \cdot \beta$ will be set to reduce the corresponding weight, which is called demotion parameters. WINNOW is an example of the index update algorithm, and weight of relevant features has grown exponentially, but the weight of uncorrelated features has reduced exponentially. For this reason, WINNOW can quickly adapt to changes in the objective function (concept drift).

(2) 1- Nearest Neighbor (1NN) is based on the principle that: there usually will be instances very close to the those within the data set, and these examples have similar properties. If the instances are marked with the class label, then unclassified instance tag value can be determined by observing its classes of NN. Absolute position of examples within this space is not so important as the relative position between instances, and relative distances are determined using the distance metric, which is ideal. Distance metric must be a minimum distance between the two instances of similar class, but the maximum distance between instances of different classes.

(3) Naive Bayesian classifier is the simplest form of Bayesian network, because it has identified that each feature and other features are independent assumptions, with the status of class characteristic given. Obviously, the assumption of independence is almost always wrong, but simple Naive Bayesian approach is still very competitive, even though it provides the estimate of poor probability of real basis. Naive Bayes algorithm is used for "batch mode", which means that the algorithm cannot execute most of its calculations after seeing its training examples but accumulates specific information on all training examples, then performing the final calculations throughout the group or "batch" of instances. However, it needs to be noted that

the there is not any fixed things in the algorithm to stop using it for incremental learning. For example, it is considered that the incremental Naive Bayes algorithm can be run, assuming that it only traverses a training data, and in the step 1, all initialized counts and sums is zero, and then the training examples are run through, one at a time. For each training instance, the feature vector X and its label are given, and the algorithm runs through the feature vector with the appropriate count incremented. In step 2, each count value is divided by the number of training examples of the same class and these counts and sums are converted into probability. Finally, previous probability $p(k)$ is computed as the score of all training examples of k class [8].

Integrated algorithm proposed in this paper starts from the establishment of set of three algorithms (NB, WINNOW, 1-NN), when a new instance arrives, algorithm running through it, and it receives each expert's forecast. Overall block diagram of the integrated approach proposed in this paper is shown in Figure 1, in which h_i is the hypothesis generated for each classification; x is classified instance, and $y*$ is the prediction of the proposed online integrated approach. The number of the model or run-time parameter is adjusted by the user, which is also a usability index of the algorithm. For non-professional data mining, none user will adjust parameters in the integrated approach proposed, which will be more attractive.

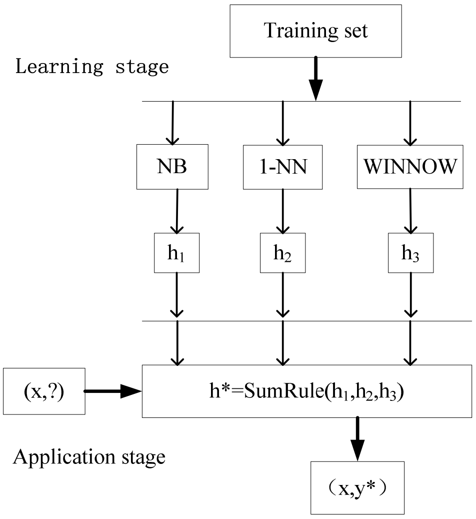

Figure 1. Integrated approach proposed in this paper

The reason why this paper uses the three specific algorithms is because they are easy to adapt to the online environment. Not only is the new mark of the same WRI available, but also the next mark of WRI will become available.

The reason why simple majority voting method is used in these three algorithms is because it is simple to use in online environment. The implementation incremental training learners' turning into upgrade package of new instances is available, but no new features become available.

Integrated approach proposed in this paper can easily achieve parallel learning algorithms in each machine. Parallel and distributed computing machine learning (ML) is very important to performers, and advantages of ML system parallel or distribution execution are: (i) increase in the speed; (ii) increase of the accessible applications (for example, it can handle multiple data).

IV. EXPERIMENT

In this paper, it takes the training set HOU as research objectives provided by "information" course in distance education of Greek universities. Basic education unit of HOU is module, and each student can register up to three modules per year, and "information" course is consisted of 12 modules, able to obtain bachelor's degree. Among HOU's INF10 modules, students must submit four copies of written assignments each year, participating in face-to-face four optional meetings with teachers, and sitting the final test 11 months later. Score system in Greek university is10-point system, and students' scores are greater than or equal to 5, and through the course or module, and it means fail if less than 5. There is a total of 1347 instances (students' records) who register INF10, as shown in Table 1.

Data are collected from two different sources, students' registration of HOU and teachers' records, and almost the relevant data for all students are collected. "Class properties" (the dependent variable) indicate two values that lead to test results in final exams; the "failure" means that students' performance is poor; "poor performance" means the students whose courses are suspended in school year (due to personal or professional reasons or the inability to turn in 2 copies of written work) and the students who do not take the final exam or attend the final exam but their scores are lower than five points; "pass" indicates the completion of the INF10 modules, and scores obtained in the final exam more than five points.

In the first stage (training phase), data collected from 2006-2007 school year are used to train each algorithm, and the training phase is divided into four consecutive steps. Step 1 includes the first written assignments and data obtained from the class; step 2 includes the data used in step 1 and the data of the second written assignment; step 3 includes data used in step 2 and the data of third written work, and step 4 includes data for use in step 3 and the data of the fourth written assignment.

Experiment 1: Comparison is made between the algorithm proposed in this paper and each online learning algorithm (Naive Bayes, 1-NN, WINNOW).

This paper minimizes the impact of prejudices of any expert system by adjusting the particular dataset of any algorithm, using the default-value of learning parameters as far as possible, which could result in lower estimated error rate, but this may affect bias of all learning algorithms. WRI-1 marked line in Table 2 indicates the prediction accuracy.

It is apparently seen from Table 2 that, in accordance with the - test in , the classification accuracy of the proposed integration algorithm is better than each of the other classifiers. Overall, the integration proposed algorithm is significantly more accurate than WINNOW algorithm in the four outputs of the four test steps. In addition, the proposed algorithm is significantly more accurate than 1-NN algorithm in the two outputs of the four test steps, and that the proposed method is significantly more accurate than NB algorithm in an output of four test procedures.

Experiment 2: Comparison is made between the integrated approach proposed and representative algorithm for each batch of advanced machine learning techniques. Batch algorithm is used for the upper algorithm for measuring the accuracy of learning algorithms, and most of the incremental versions batch algorithms are lossy. Lossless online learning algorithm is that when the same training set is given, the algorithm returns a hypothesis, which is the same with the one returned by its corresponding batch algorithm. C4.5 algorithm is a typical decision tree algorithm; RBF algorithm is a well-known learning algorithm of estimating neural network weight value, a representative neural network algorithm. In this study, 3-NN algorithm combines the noise of strong robustness; RIPPER is a representative learner of rule. Finally, Sequential Minimal Optimization (SMO) is a representative algorithm of SVMs. WRI-2, WRI -3 and WRI-4 marked lines in Table 3 indicate the prediction accuracy of each algorithm.

It is apparently seen from Table 3 that the integration proposed algorithm is significantly more accurate than RBF, BP and SMO algorithm in an output of the four test steps. In addition, the proposed algorithm is significantly more accurate than 3NN algorithm in the two outputs of the four test steps, and that the proposed method is significantly more accurate than RIPPER and C4.5 algorithm in an output of four test procedures. Finally, the proposed algorithm is significantly more accurate than the method of voting perception in the four outputs of the four test steps.

Experiment 3: Comparison is made between the proposed approach and several well-known integrated classifiers, and it must be mentioned that other integrated methods can only be used in batch mode, and this paper uses the batch integration algorithm as the upper algorithm to measure accuracy of integrated approach. The third experiment is used for comparison: (a) Adaboost decision stump algorithm and iterative 10 times algorithm, (b) random-forest integrated algorithm with 10 trees, (c) voting perception algorithm, (d) rotation forest algorithm with C4.5 and iterative 10 times algorithm. Table 4 significantly shows that the proposed method is significantly more accurate than batch integrated algorithms of other tests in an output of four test procedures.

As can be seen from Table 4, as previously mentioned, the main advantage of the integrated approach proposed in this paper is that it is easy to apply to the online environment. Not only is the new mark of the same WRI available, but also the next mark of WRI will become available. If the article has used another test learner and integrated approach, then when the next WRI marker becomes available, it shall be started from scratch to re-train the classifier.

Experiment 4: All algorithms of data concentration in the paper are used in batch learners, and Table 5 shows the training time.

As can be apparently seen from Table 5, the incremental update is faster than the currently seen return of a batch algorithm on all the data, and it might even be the only way, if all seen data currently cannot be stored or if online prediction and update needs to be performed in real time, but at least it has great rapidity.

TABLE I PRACTICAL PROPERTIES AND THEIR VALUES

Students' performance properties	
1st written assignments	Score<3, 3 ≤ score ≤ 6, score> 6
2nd written assignments	Score <3, 3 ≤ score ≤ 6, score> 6
3rd written assignments	Score<3, 3 ≤ score ≤ 6, score> 6
4th written assignments	Score <3, 3 ≤ score ≤ 6, score> 6

TABLE II CLASSIFICATION ACCURACY OF EACH ALGORITHM (%)

Algorithm	Algorithm in this paper	NB	1-NN	WINNOW
WRI-1	83.86	68.00	73.86	67.40
WRI-2	88.39	76.39	78.24	74.75
WRI-3	89.73	79.43	78.24	74.75
WRI-4	89.81	80.84	78.47	77.95
The average accuracy	88.95	76.17	77.20	73.71

TABLE III CLASSIFICATION ACCURACY COMPARISON OF EACH ALGORITHM (%)

	Algorithm in this paper	C4.5	3NN	RIPPER	SMO	BP	RBF
WRI-1	83.86	73.86	73.86	73.86	69.19	71.56	72.38
WRI-2	88.39	77.35	78.09	77.65	77.95	78.17	76.31
WRI-3	89.73	80.02	78.99	80.02	80.10	80.62	81.06
WRI-4	89.81	81.14	78.99	80.69	81.73	80.92	81.06
The average accuracy	88.95	78.09	77.48	78.06	77.24	77.81	77.70

TABLE IV CLASSIFICATION ACCURACY COMPARISON OF EACH ALGORITHM (%)

Algorithm	Algorithm in this paper	AdaBoost AdaBoost decision stump	Random Forest	Vote perception	Rotation forest
WRI-1	83.86	73.34	73.34	67.33	73.34
WRI-2	88.39	78.40	78.02	74.31	77.58
WRI-3	89.73	80.04	79.81	75.27	79.88
WRI-4	89.81	80.99	80.70	76.83	80.85
The average accuracy	88.95	78.19	77.96	73.44	77.91

TABLE V TRAINING TIME OF EACH ALGORITHM UNDER 2GHZ DUAL-CORE SYSTEM 3GB MEMORY ENVIRONMENT (SECONDS)

Algorithm	Algorithm in this paper	SMO	BP	RBF	AdaBoost decision stump	Random Forest	Rotation forest	Vote perception
WRI-1	0.01	0.07	3.87	0.05	0.02	0.16	0.01	0.07
WRI-2	0.01	0.13	5.54	0.05	0.03	0.27	0.02	0.09
WRI-3	0.01	0.13	7.3	0.06	0.03	0.33	0.02	0.1
WRI-4	0.01	0.18	7.68	0.06	0.03	0.59	0.03	0.1
The average accuracy	0.01	0.1275	6.0975	0.055	0.0275	0.3375	0.02	0.09

V. CONCLUSIONS CONFLICT OF INTEREST ACKNOWLEDGMENT

This paper proposes an integrated algorithm using method of voting and combined with three online classifiers: Naive Bayes, 1-NN and WINNOW algorithm. Through quite accurate prediction, the teacher has the ability to know which students can complete the module or course, and the initial accuracy of this prediction is 73%. Based on the students' demographic data, it reaches82% before the final exam. Data sets stem from module "information introduction", but the conclusion is replicable, which has currently caused extensive research interests in most modules of HOU for scholars. Experimental results show that compared to some more advanced classifier, classification tree integration algorithm proposed can more accurately predict the performance of students. In the future, the proposed algorithm will be applied to other remote data, and combined with other more advanced data mining techniques, a lot of experiments are conducted to better predict students' performance in distance education.

ACKNOWLEDGMENT

This work is supported by the The Philosophy and Social Science Planning Project of Heilongjiang Province in 2014 year, youth project, No.14C023.

REFERENCES

[1] J. He, Y. Geng and K. Pahlavan, “Modeling Indoor TOA Ranging Error for Body Mounted Sensors”, 2012 IEEE 23nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Sydney, Australia Sep. pp. 682-686, 2012.

[2] Li, Xiaoming, Zhihan Lv, Baoyun Zhang, Weixi Wang, Shengzhong Feng, Jinxing Hu. “WebVRGIS Based City Bigdata 3D Visualization and Analysis”. In Pacific Visualization Symposium (PacificVis), 2015 IEEE. IEEE, 2015.

[3] Li, Xiaoming, Zhihan Lv, Jinxing Hu, et al., “XEarth: A 3D GIS Platform for managing massive city information”. IEEE Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). IEEE, 2015.

[4] S. Li, Y. Geng, J. He, K. Pahlavan, “Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization”, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), Chongqing, China Oct. pp. 721-725, 2012.

[5] Su, Tianyun, Zhihan Lv, Shan Gao, Xiaolong Li, and Haibin Lv. “3D seabed: 3D modeling and visualization platform for the seabed”. In Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on, pp. 1-6. IEEE, 2014.

[6] Y. Geng, J. He, K. and Pahlavan, “Modeling the Effect of Human Body on TOA Based Indoor Human Tracking”, International Journal of Wireless Information Networks, vol. 20, No. 04, pp. 306-317, 2014.

[7] Zhang, Mengxin, Zhihan Lv, Xiaolei Zhang, et al.,. "Research and Application of the 3D Virtual Community Based on WEBVR and RIA." Computer and Information Science, vol. 2, No. 01, pp. 84, 2014.

[8] Zhong, Chen, Stefan Müller Arisona, Xianfeng Huang, et al., “Detecting the dynamics of urban structure through spatial network analysis”, International Journal of Geographical Information Science, vol. 28, No. 11, pp. 2178-2199, 2014.