DOI: 10.14529/ctcr190414

OPTIMIZATION OF CLASSIFIERS ENSEMBLE CONSTRUCTION: CASE STUDY OF EDUCATIONAL DATA MINING

Y.K. Salal, *Yasskhudheirsalal@gmail.com,*
S.M. Abdullaev, *abdullaevsm@susu.ru*
South Ural State University, Chelyabinsk, Russian Federation

The choosing the best prediction method of education results is major challenge of Educational Data Mining (EDM). This EDM paper compares the results of student's performance forecast produced by the individual binary classifiers (Naïve Bayes, Decision Tree, Multi-Layer Perceptron, Nearest Neighbors, Support Vector Machine algorithms) and their ensembles, which are trained (tested) on dataset containing up to 38 input attributes (weekly attendance in mathematics, the intensity of study, interim assessment) of 84 (36) secondary school students from Nasiriyah, Iraq. The two-class school performance was predicted – passing or not passing on final exam. Three following stages of comparison were completed. At the first stage of the experiment, the dependence of classifiers from the input attributes was investigated. It was shown that the forecast accuracy rises from 61.1–77.7% when all 38 attributes were used, to 75.0–80.5%, if base classifier trained with five attributes pre-selected by Ranker Search method. Then, in second stage, to each of the base classifier the AdaBoost M1 procedure has been applied and five homogenous ensembles were created. And only two of these ensembles demonstrated small rise of 3% in accuracy comparing to corresponding stand-alone classifier, but the overall maximal prediction accuracy of 80.5% stayed the same. Finally, comparing the accuracies of 77.7% and 83.3% achieved by the heterogeneous ensemble consisted of five simple voting base classifiers and by the heterogeneous meta-ensemble of five simple voting AdaBoost homogenous ensembles correspondingly, we conclude that improvement of the quality of the individual classifier or homogeneous ensembles allows to construct more powerful EDM prediction methods.

Keywords: base classifiers, educational data mining, ranker search method, adaptive boosting, heterogeneous ensembles.

Introduction

Educational Data Mining and Learning Analytics (EDM/LM) are promising scientific field to enhance of teaching and learning technologies of traditional and e-learning education [1–5] and to manage of various forms of constructivist education [6]. The wide availability of data mining tools such as R, scikit-learn for Pyton, and Weka [7] allows us to solve one of the main tasks of EDM/LA: to forecast of student's performance and to help the needy [8, 9]. Most commonly, this task is resolved by using of individual classifiers with learning following algorithms [10]: Naïve Bayes (NB), Decision Tree (J48), Multi-Layer Perceptron (MLP), Nearest Neighbors (1NN) and Support Vector Machine (SVM) and other algorithms from the top-10 list [11].

On the other hand, in pedagogical practice to identify problematic students and to solve their fate are used collective expert decisions. In this sense, in the EDM/LA we should use one of the meta-learning approaches consists of "learning from base learners" [12, 13]. The general purpose of this paper is to compare capacity of two types of heterogenous ensembles. First type of ensemble was created by base classifiers used NB, J48, MLP, 1NN and SVM and attribute structure improved by Ranker Search method application. The second ensemble consists of five homogeneous ensembles created by AdaBoost.M1 procedure from each of five base classifiers.

1. Data set

This particular student's performance dataset are collected from secondary school of Nasiriyah, Iraq for first semester 2018–2019 of Mathematic subject. Detail for educational dataset containing up to 38 input attributes is shown in Table 1. The dataset includes student's attributes like Name, Age, Gender, internal assignment attributes (Quizzes), number times of absence, in additionally two monthly exams,

finally the final exam of first semester for mathematic subject, collected by school reports and questionnaires are used for collecting data from the archives 120 students.

Table 1

Description of the dataset

#	Attributes	Description	Possible Values
1	Stu_N	Student name	String text
2	AG	Age of student	1: 14–16, 2: 16–18, 3: >18
3	Gender	Sex of Student	1: Male, 2: Female
4–7	Q1_M1, Q2_M1, Q3_M1, Q4_M1	Quiz 1, 2, 3, 4 for 1°, 2°, 3°, 4° week of first month	0 to 10
8–11	A1_M1, A2_M1, A3_M1, A4_M1	How many times absent 1°, 2°, 3°, 4° week of first month	0, 1, 2, 3, 4
12–15	Q1_M2, Q2_M2, Q3_M2, Q4_M2	As #4–7 of second month	0 to 10
16–19	A1_M2, A2_M2, A3_M2, A4_M2	As #8–11 of second month	0, 1, 2, 3, 4
20	Exam1	Exam of first attempt for semester	0 to 100
21–24	Q1_M3, Q2_M3, Q3_M3, Q4_M3	As #4–7 of third month	0 to 10
25–28	A1_M3, A2_M3, A3_M3, A4_M3	As #8–11 of third month	0, 1, 2, 3, 4
29–32	Q1_M4, Q2_M4, Q3_M4, Q4_M4	As #4–7 of fourth month	0 to 10
33–36	A1_M4, A2_M4, A3_M4, A4_M4	As #8–11 of fourth month	0, 1, 2, 3, 4
37	Exam2	Exam of second attempt for semester	0 to 100
38	AV_E1_E2	Average Exam1 & Exam2	0 to 100
39	FS_Exam	Exam First semester	0 to 100 (Class)

The output attribute is two class labeled as “Pass” if exam grade was ≥ 50 and as “Fail” if was not. Thus, two group of students with 80 students passed and 40 students that drop out on final exam were observed. The Weka version 3.8 (downloaded from https://sourceforge.net/projects/weka/) NB, J48, MLP, 1NN and SVM default algorithms were trained on randomly choosing data of 84 students (70% of data set) and then testing on the data of rested 36 students (30% of data set).

2. Methods of analysis and results

Three stages of the experiment can distinguish.

Stage of selection of attribuites. Initially, we find the base classifier’s accuracy applying then to all 37 input attributes. Then we use attribute selection methods to eliminate both irrelevant attributes and redundant ones. A simpler idea is to rank the effectiveness of each [14, 15]. We implement Ranker Search Method (RSM) on the dataset to compare between results models accuracy for prediction students performance. RSM [14] is combined by 3 feature selection techniques:

1) *Correlation Attribute Evaluation* which correlate each attribute of dataset and the output class evaluation, choosing the most relevant attributes by value of Pearson’s correlation;

2) *Information Gain Attribute Evaluation* is entropy measure introduced to machine learning by Quilan [16];

3) *Gain Ratio Attribute Evaluation* overcomes the bias of Information Gain across the features with the large number of values;

4) *Wrapper Subset Evaluation* introduced by Kohavi and John [17].

Table 2 show the best five attributes chosen by these attribute evaluators.

Table 2

Attributes Selected by Ranker Search Method

Attribute Evaluator	Attribute Arrange	Attribute Name	Ranked Values
Correlation Attribute Evaluation	37, 36, 19, 18, 29	Exam2, AV_E1_E2, Exam1, A4_M2, Q2_M4	0.572, 0.543, 0.524, 0.348, 0.311
Information Gain Attribute Evaluation	36, 37, 19, 29, 18	Exam2, AV_E1_E2, Exam1, Q2_M4, A4_M2	0.343, 0.295, 0.287, 0.094, 0.088
Gain Ratio Attribute Evaluation	37, 19, 36, 18, 29	AV_E1_E2, Exam1, Exam2, A4_M2, Q2_M4	0.295, 0.288, 0.226, 0.095, 0.094
Wrapper Subset Evaluation	37, 19, 36, 29, 18	AV_E1_E2, Exam1, Exam2, Q2_M4, A4_M2	0.326, 0.317, 0.295, 0.104, 0.101

It is easy to notice that besides the same attributes related intermediate examinations, with a significant lag two attributes related to attendance and quizzes were selected by RSM, also.

The second stage is boosting. Adaptive Boosting (AdaBoost) introduced by Freud and Shapire [18] to improve prediction ability of one single classifier (old) when we train new classifier based on the same algorithms but on the dataset updated by using the rules which increase the weight of examples misclassified by old one, and to decrease the weight of correctly classified examples. Thus, the weight tends to concentrate the weak classifier on "hard" exam. And at final of iteration procedures, ensemble of classifiers is produced, where all classifier are voted by their weight. We used AdaBoost.M1 to our purposes because of according to [5], AdaBoost.M1 is more adequate classifier for EDM/LA mining.

At the third stage, we compared the improvement of overall accuracy (A) and F-measure for minor class (F) of individual base classifiers after feature selection and boosting stages. As can be seen from Table 3, the major advance in forecasting capacity was observed after Ranker Search Method (RSM) application. In particular, we see that two algorithms J48 and NN, that took F less than 40%, after RSM increased up to 20% their predictability of minor class to do useful forecasts (> 50%) and permit their boosting.

Table 3

Accuracy and F-measure of different classifiers

Algorithm \ Classifier	Classifier with all attributes		Base classifier using RSM		AdaBoost.M1 classifiers	
	A, %	F1, %	A, %	F1, %	A, %	F1, %
Naïve Bayes, NB	77.7	71.4	80.5	74.1	80.5	74.1
Decision Tree, J48	61.1	36.4	75.0	52.6	75.0	57.1
Multi-Layer Perceptron, MLP	75.0	66.7	75.0	52.6	77.7	55.6
Nearest Neighbors, NN	72.2	37.5	75.0	57.1	75.0	57.1
Support Vector Machine, SVM	75.0	64.0	77.7	63.6	80.5	66.7

Evaluating the effectiveness of AdaBoost homogeneous ensembles show that accuracy rise only to 0–3% in comparison to of RSM classifier 0–14%. At the same time, the capacity of leading NB-based classifier remained unchanged.

Finally, we compare accuracy of 3 simple voting ensembles: one built from base classifiers (72.2%); second contain the classifiers which received after the first RSM stage (77.7%) and the ensemble obtained by combination of Adaboost ensembles (83.3%).

Conclusion

In this study, the main focus has been comparison of various models of machine learning algorithms based on NB, J48, MLP, 1NN and SVM algorithms and their ensembles. We observe that applying Ranker Search method to choose the best attributes have major effect on forecast evaluated by F-measure of less representative data class, and consequently permit us to use the improved weak classier as initial Adaboost resident. We can see also that accuracy of final heterogeneous ensemble, for the first time on the 3% maximum single classifier performance surpassed and homogenious combination of their ensembles.

References

1. Romero C., Ventura S. Educational Data Mining: A Review of the State of the Art. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 2010, vol. 40, no. 6, pp. 601–618. DOI: 10.1109/TSMCC.2010.2053532

2. U.S. Department of Education, Office of Educational Technology. *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief.* Washington, D.C., 2012, Available at: https://tech.ed.gov/learning-analytics/edm-la-brief.pdf (accessed: 03.07.2018).

3. Baker R.S., Inventado P.S. Educational Data Mining and Learning Analytics. *In: Larusson J., White B. (Eds.). Learning Analytics*. Springer, New York, NY, 2014, pp. 61–75. DOI: 10.1007/978-1-4614-3305-7_4

4. Calvet Liñán L., Juan Pérez Á.A. Educational Data Mining and Learning Analytics: Differences, Similarities, and Time Evolution. *RUSC. Universities and Knowledge Society Journal*, 2015, vol. 12, no. 3, pp. 98–112. DOI: 10.7238/rusc.v12i3.2515

5. Jovanovic M., Vukicevic M., Milovanovic M., Minovic M. Using Data Mining on Student Behavior and Cognitive Style Data for Improving E-Learning Systems: a Case Study. *I. Journal of Computational Intelligence Systems*, 2012, vol. 5, no. 3, pp. 597–610. DOI: 10.1080/18756891.2012.696923

6. Berland M., Baker R.S., Blikstein P. Educational Data Mining and Learning Analytics: Applications to Constructionist Research. *Tech Know Learn.*, 2014, vol. 19, pp. 205–220. DOI: 10.1007/s10758-014-9223-7

7. Slater S., Joksimovic S., Kovanovic V., et al. Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics,* 2017, vol. 42, no. 1, pp. 85–106. DOI: 10.3102/1076998616666808

8. Castro-Wunsch K., Ahadi A., Petersen A. Evaluating Neural Networks as a Method for Identifying Students in Need of Assistance. *SIGCSE' 17*, March 08–11, 2017, Seattle, WA, USA. DOI: 10.1145/3017680.3017792.

9. Hussain S., Fadhil M.Z., Salal Y.K., Theodoru P., Kurtoğlu F., Hazarika G.C. Prediction Model on Student Performance Based on Internal Assessment Using Deep Learning. *I. Journal of Emerging Technologies in Learning*, 2019, vol. 14, no. 8, pp. 4–22. DOI: 10.3991/ijet.v14i08.10001

10. Wu X., Kumar V., Quinlan R.J. et al. Top 10 Algorithms in Data Mining. *Knowl. Inf. Syst.*, 2008, vol. 14, pp. 1–37. DOI: 10.1007/s10115-007-0114-2

11. Kumar M., Salal Y.K. Systematic Review of Predicting Student's Performance in Academics *I. J. of Engineering and Advanced Tech.*, 2019, vol. 8, no. 3, pp. 54–61.

12. Smith-Miles K.A. Cross-Disciplinary Perspectives on Meta-Learning for Algorithm Selection. *ACM Comput. Surv.*, 2008, vol. 41, no. 1, Article 6, 25 p. DOI: 10.1145/1456650.1456656

13. Vilalta R., Giraud-Carrier C., Brazdil P. Meta-Learning – Concepts and Techniques. *In: Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 717–732. DOI: 10.1007/978-0-387-09823-4_36

14. Salal Y.K., Abdullaev S.M., Kumar M. Educational Data Mining: Student Performance Prediction in Academic. *I. J. of Engineering and Advanced Tech.*, 2019, vol. 8, no. 4C, pp. 54–59.

15. Trabelsi M., Meddouri N., Maddouri M. A New Feature Selection Method for Nominal Classifier Based on Formal Concept Analysis. *Procedia Computer Science*, 2017, vol. 112, pp. 186–194. DOI: 10.1016/j.procs.2017.08.227

16. Quinlan J.R. Induction of Decision Trees. *Machine Learning*, 1986, no. 1, pp. 81–106. DOI: 10.1007/BF00116251

17. Kohavi R., John G.H. Wrappers for Feature Subset Selection. *Artificial Intelligence (97)*, 1997, pp. 273–324 DOI: 10.1016/S0004-3702(97)00043-X

18. Freund Y., Schapire R.E. A Short Introduction to Boosting. *J. of Japanese Society for Artificial Intelligence*, 1999, vol. 14, no. 5, pp. 771–780.

Received 30 April 2019

УДК 004.8, 004.9 **DOI: 10.14529/ctcr190414**

ОПТИМИЗАЦИЯ КОНСТРУКЦИИ АНСАМБЛЯ КЛАССИФИКАТОРОВ: ПРИМЕР ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ОБРАЗОВАТЕЛЬНЫХ ДАННЫХ

Я.К. Салал, С.М. Абдуллаев

Южно-Уральский государственный университет, г. Челябинск, Россия

Сравниваются результаты прогнозирования итогов обучения бинарных классификаторов и их ансамблей с использованием пяти алгоритмов машинного обучения: Naïve Bayes, Decision Tree, Multi-Layer Perceptron, Nearest Neighbors, Support Vector Machine. Все классификаторы обучались (тестировались) на наборе данных, содержащих до 38 входных атрибутов, отражавших посещаемость уроков по математике, интенсивность обучения и промежуточные оценки 84 (36) учащихся средних школ из города Эн-Насирии, Ирак; прогнозировалось два класса их оценок на экзамене по математике. Эксперимент проводился в три этапа. Сначала было показано, что точность прогнозов классификаторов поднимается с 61,1–77,7 %, при использовании всего набора атрибутов, до 75,0–80,5 %, когда классификаторы обучались на данных из пяти атрибутов, выбранных методом ранжирования Ranker Search. Затем на втором этапе к каждому из этих слабых классификаторов была применена процедура бустинга AdaBoost M1 и были созданы пять однородных ансамблей. Некоторые из этих ансамблей демонстрировали 3%-ный рост точности, но их максимальная точность не превышала точности лучшего автономного классификатора (80,5 %). Тем не менее сравнение точности гетерогенного ансамбля, состоявшего из базовых классификаторов, обученных на ранжированных атрибутах (77,7 %), и мета-ансамбля, состоявшего из пяти однородных ансамблей AdaBoost (83,3 %), позволяет сделать вывод, что улучшение качества отдельных классификаторов и составление из них гетерогенных ансамблей позволяет построить более мощные методы анализа образовательных данных.

Ключевые слова: базовый классификатор, интеллектуальный анализ образовательных данных, метод селекции атрибутов Ranker Search, AdaBoost, гетерогенные ансамбли.

Салал Ясс Кхудейр, аспирант кафедры системного программирования, Южно-Уральский государственный университет, г. Челябинск; Yasskhudheirsalal@gmail.com.

Абдуллаев Санжар Муталович, д-р геогр. наук, профессор кафедры системного программирования, Южно-Уральский государственный университет, г. Челябинск; abdullaevsm@susu.ru.

Поступила в редакцию 30 апреля 2019 г.

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Salal, Y.K. Optimization of Classifiers Ensemble Construction: Case Study of Educational Data Mining / Y.K. Salal, S.M. Abdullaev // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2019. – Т. 19, № 4. – С. 139–143. DOI: 10.14529/ctcr190414

FOR CITATION

Salal Y.K., Abdullaev S.M. Optimization of Classifiers Ensemble Construction: Case Study of Educational Data Mining. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2019, vol. 19, no. 4, pp. 139–143. DOI: 10.14529/ctcr190414