

Teaching Evaluation System by use of Machine Learning and Artificial Intelligence Methods

<https://doi.org/10.3991/ijet.v16i05.20299>

Jingjing Hu

Zhejiang Business College, Hangzhou, China
9061613@qq.com

Abstract—To explore the adoption of artificial intelligence (AI) technology in the field of teacher teaching evaluation, the machine learning algorithm is proposed to construct a teaching evaluation model, which is suitable for the current educational model, and can help colleges and universities to improve the existing problems in teaching. Firstly, the existing problems in the current teaching evaluation system are put forward and a novel teaching evaluation model is designed. Then, the relevant theories and techniques required to build the model are introduced. Finally, the experiment methods and process are carried out to find out the appropriate machine learning algorithm and optimize the obtained weighted naive Bayes (WNB) algorithm, which is compared with traditional naive Bayes (NB) algorithm and back propagation (BP) algorithm. The results reveal that compared with NB algorithm, the average classification accuracy of WNB algorithm is 0.817, while that of NB algorithm is 0.751. Compared with BP algorithm, WNB algorithm has a classification accuracy of 0.800, while that of BP algorithm is 0.680. Therefore, it is proved that WNB algorithm has favorable effect in teaching evaluation model.

Keywords—AI; machine learning; regression analysis; naive Bayes

1 Introduction

With the continuous update of emerging technologies such as AI and the progression of educational information infrastructure, the integration of information technology and education is also deepening. Many new technologies, such as online course, simulation teaching, and online education platform, have been widely adopted in college teaching. The popularization of these technologies has realized the integration and sharing of teaching resources and promoted the communication between teachers and students. Moreover, it has made up for the deficiency of traditional classroom teaching, and improved the teaching quality of teachers and enhanced the learning interest of students, thus network experimental teaching platform has attracted more attention from universities [1].

At present, AI technology has been able to improve teaching efficiency in learning tutoring, teaching evaluation, and teaching space optimization, and help students realize personalized learning [2]. Therefore, the close connection of education and AI

technology would promote the reform and innovation of college teaching and establish an education and teaching system suitable for students' lifelong development, which helps education to transform to high-level and precise type [3].

The teaching evaluation system can help schools judge teachers' teaching effect and students' learning outcomes, which is the evident for many universities to judge the teaching process of teachers. However, the current education evaluation system fails to reflect the teaching situation of teachers under the new technology, including the following aspects [4]. I. The evaluation method is out dated and inefficient, and the data credibility is low. II. The evaluation index is not perfect, and there is no evaluation content for multimedia teaching mode. III. The weight distribution of evaluation indexes is unreasonable and lacks objectivity and fairness. IV. There are deficiencies in the analysis and processing of the data at the later stage, and the practical information can't be extracted. Moreover, the implementation of teaching evaluation process is very complex, and massive data calculation is required. Therefore, it is urgent to establish an objective, efficient, and feasible teaching evaluation system and evaluation optimization process.

In this work, a teaching evaluation system based on machine learning is designed. Firstly, correlation analysis is performed on the acquired evaluation data. Secondly, association rules are followed to determine the relationship between indexes in teaching evaluation. Finally, the machine learning algorithm is adopted to optimize the data processing and build the teaching evaluation model, so as to realize the automation of teaching evaluation.

2 Methods

2.1 Design of teaching evaluation system

The traditional teaching evaluation includes student evaluation, teacher mutual evaluation, teacher self-evaluation, and expert evaluation. Due to the different evaluators, the content of the evaluation is also different. However, the current evaluation content is generally aimed at the content and attitude of teachers' teaching efforts. Such evaluation is too formal to reflect the merits and demerits of the actual teaching effect. Most of the current evaluation forms are statistical statements, which is not only require a large task of data analysis, but also can't describe the information presented in the data perfectly [5].

Combined with the characteristics of big data background, data mining technology is adopted to solve this problem [6]. Firstly, web-based evaluation process is adopted instead of paper evaluation forms. Meanwhile, the relationship between teaching effect and teaching level is explored after data processing, and the evaluation index system is then optimized. Finally, aided by machine learning classification algorithm, an appropriate teaching evaluation model is established to obtain rapid and objective teaching evaluation and help improve teaching management. The specific evaluation process of the teaching evaluation system designed in this study is shown in Fig. 1.

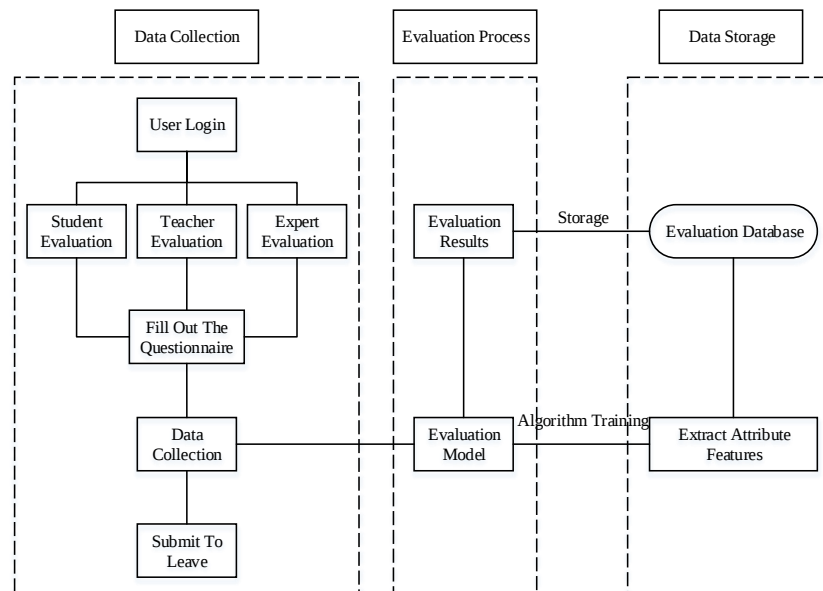


Fig. 1. Specific evaluation process of teaching evaluation system

2.2 Data collecting

The construction of the proposed teaching evaluation model involves data acquisition, preprocessing, and correlation analysis between data. Therefore, the design process includes the evaluation questionnaire, standardization of data, and correlation analysis. During data processing, regression analysis is adopted to detect abnormal data. Questionnaire survey is a common data collection method, whose sampling is extensive and representative, and it is efficient and easy to conduct quantitative analysis. Combined with the interaction and transmission of the network, it makes data collection more convenient and data processing more convenient [7]. This work evaluates the teaching effect of teachers from the perspective of evaluation in Fig. 2.

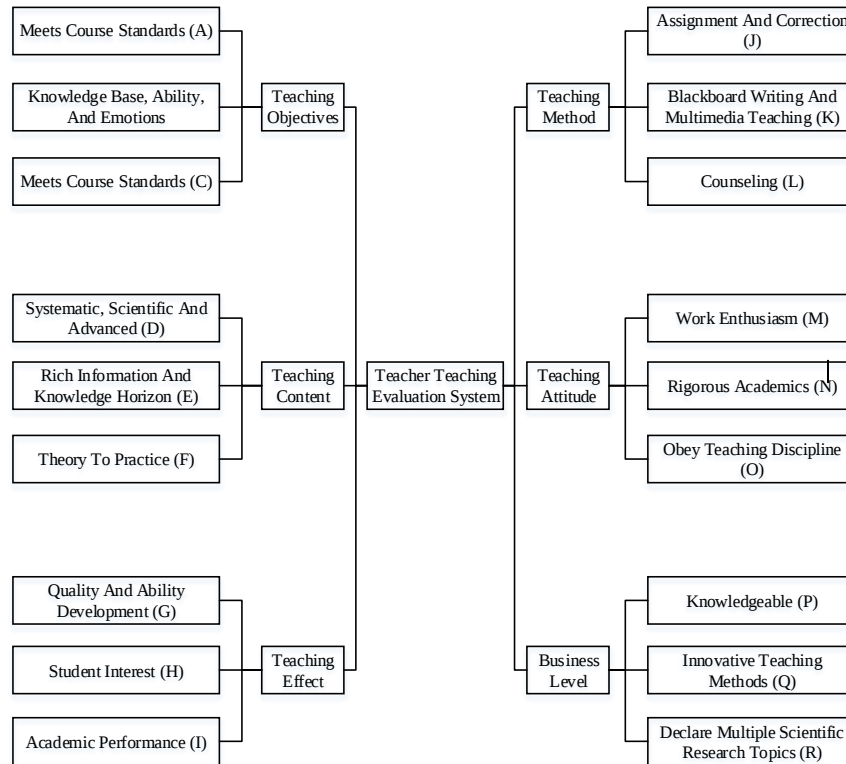


Fig. 2. Classification of teacher evaluation indexes

The index is a stipulation on one aspect of a specific target and reflects the characteristics of a certain aspect. Therefore, for the overall characteristics, it shouldn't start with a certain index, but deal with the relationship between each index, and a series of related index combinations should be set up to completely reflect the characteristics of things. Therefore, the evaluation of each teaching content in Fig. 2 is described as 1, 2, 3, 4, and 5, corresponding to unqualified, qualified, medium, good, and excellent, respectively. The Letters A, B, C, ••, Q, and R correspond to the 18 evaluation indexes in Fig. 2, and the letter T represents the comprehensive evaluation score. Each column of Table 1 contains a complete teaching evaluation record, and each letter corresponds to the corresponding value and the evaluation value of the evaluation index.

Table 1. Part of the data set after processing

	1	2	3	...	n-1	n
A	4	5	3	...	2	5
B	5	5	3	...	3	3
C	3	3	4	...	4	2
...	...	...	...	...	...	...
P	4	3	4	...	5	2
Q	4	4	5	...	3	4
R	5	2	3	...	4	3

2.3 Data processing

In teaching evaluation, evaluators can make extreme evaluations of teachers due to the influence of individual subjective emotions, resulting in the lack of credibility of evaluation data. Therefore, it is necessary to eliminate the teaching evaluation data deviating from the actual situation to ensure the objectivity and authenticity of the established teaching evaluation model.

Multiple regression analysis is widely adopted in actual data processing, which can analyze the data correlation of multiple independent variables and one dependent variable, and establish a prediction model. Multiple linear regression analysis refers to the linear relationship between independent variables and dependent variables [8]. Usually when a training data set $(Y_i, X_{i1}, X_{i2}, \dots, X_{ip}) \quad i = 1, \dots, n$ is given, a model is established through the training of a large number of data, which is $Y = f(X)$.

The deviation degree of the data is determined by the training data regression, and the normalized data of the 18 evaluation indexes are taken as the input variable during the training. The comprehensive evaluation results are taken as output variables to construct a prediction model. Then, the new input data are adopted to predict and evaluate the value comprehensively. By setting a threshold range, the evaluation value given by students with the predicted value is compared, the data deviating from the predetermined range is eliminated as abnormal data.

Scientific data analysis and screening of evaluation indexes can ensure the data independence between evaluation indexes, so it is necessary to analyze the data correlation of the evaluation indexes set above. The correlation coefficient indicates the degree of closeness between random variables, and the correlation coefficient equation can be taken to judge the degree of correlation and mutual influence between indexes, as shown in equation (1).

$$r = \frac{cov(X,Y)}{\partial x \partial y} \quad (1)$$

r represents the correlation coefficient between attributes X and Y , $cov(X,Y)$ is covariance, indicating the degree of coordination between indexes, ∂x and ∂y are both standard deviations, indicating the data volatility of indexes. When correlation coefficients between multiple attributes are calculated, the correlation coefficients can be formed into a matrix, as shown in equation (2).

$$R = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \quad (2)$$

The correlation analysis adopted relies on association rules to obtain the actual relationship between strong rule analysis and evaluation indexes. Association rules are adopted to express some correlation between the two sets of numbers X and Y , which is determined by the confidence c and support s [9].

The confidence level c measures the credibility of the rule, and is defined as the ratio of the number of events containing both X and Y to the number of events containing X in database A , that is, at least $c\%$ of the events contained in the event database also contains Y , the expression is shown in equation (3).

$$c(X \Rightarrow Y) = P(Y|X) = P(X \cap Y)/P(X) \quad (3)$$

The support degree s refers to the statistical importance of the rule in the entire data set, which represents the ratio of the number of events containing X in the event database A to the total number of events, indicating that at least $s\%$ of the events in the event database A contain X , as shown in equation (4).

$$s(X \Rightarrow Y) = P(X \cap Y) \quad (4)$$

The processed data set is calculated via the equation of the association rules to obtain strong rules between the evaluation indexes. The correlation analysis process of the association process is shown in Fig. 3.

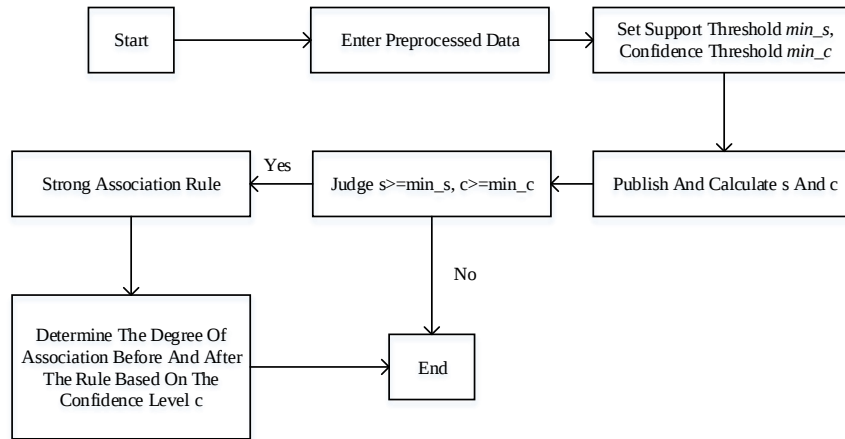


Fig. 3. Correlation analysis flow chart based on association rules

Through the data acquisition, processing, and correlation analysis of the teaching evaluation model, a scientific and effective evaluation index is designed. Moreover, the correlation between the indexes is analyzed to eliminate the large deviation of the data, and the teaching evaluation system is optimized.

2.4 Construction of teaching evaluation model based on machine learning

In machine learning, classification refers to summarize the special properties of the data in the training data set and find the appropriate description or model for each data set. Special descriptions are generated to classify the future data, and the classification of unknown data is inferred from the obtained model. The classification includes two processes: learning and classification. During the learning, a classifier can be trained with appropriate machine learning methods based on existing training data; while during the classification, the input unknown data are classified via a classifier.

Classification algorithms commonly adopted in machine learning include support vector machines, decision trees, neural networks, NB, etc. Among them, support vector machine is a binary classification model, which is a linear classifier with the maximum interval defined in the feature space. The model of decision tree usually has a tree structure, which represents the process of classifying instances based on features, and it includes three steps: feature selection, decision tree generation, and decision tree building. Artificial neural network is an effective method to solve nonlinear problems by reducing artificial factors. NB analysis is a classification method based on Bayes' theorem. Different classification algorithms have different effects in different situations, and there is no classification algorithm suitable for any situation and problem [10].

Due to the need of teaching evaluation, the classification algorithm is adopted to the construction of teaching evaluation model. The evaluation index is taken as the input value, the comprehensive evaluation result is taken as the class label, and the most suitable class label is given to the evaluation index via an appropriate classification algorithm. The performance of the classifier is evaluated by accuracy, which is defined as the ratio between the number of samples that the classifier can correctly classify and the total number of samples for a given data set. The equation is as follows (*pre_rate* represents the accuracy rate, *N_c* represents the number of correctly classified samples, and *N* represents the total number of samples).

$$pre_rate = \frac{N_c}{N} \quad (5)$$

According to the experimental results of Section 3.2, NB algorithm has a high accuracy rate and the shortest running duration in the classification of this data set. Therefore, the NB algorithm is selected to construct the teaching evaluation model.

Bayesian classification adopts the classification algorithm of Bayes' theorem to classify data. The principle of classification: after a large number of learning and training data sets, the prior probabilities of each category can be obtained; then, the calculation of the posterior probability belonging to different categories in an instance is performed; finally, the instance is judged to belong to the class with the maximum posterior probability. The NB classification algorithm is an efficient classification algorithm in the Bayes classification algorithm, which is simple, easy to explain, and is with fast computing speed and good stability.

NB classification model is based on the general Bayes classification model, which can remove the independence between attributes. Generally, $p(X)$ is a constant, and the calculation equation of NB algorithm is as follows.

$$p(C_j|X) \propto \argmax p(X|C_j)p(C_j) \quad (6)$$

In equation (6), $p(C_j)$ is the class prior probability, which is obtained by training a large number of data sets, and the calculation is as follows.

$$p(C_j) = \frac{s_j}{s} \quad (7)$$

In equation (7), s_j represents the number of C_j in the training samples, and s represents the total number of training samples.

Each attribute variable of the NB algorithm has independent conditions. When the number of attributes in the data set is large, the calculation overhead of $p(X|C_j)$ is large, and the introduction of the conditional independence assumption can reduce the overhead, but the calculation accuracy will be reduced, and the calculation of $p(X|C_j)$ can be simplified as follows.

$$p(X|C_j) = \prod_{i=1}^n p(x_i|C_j) \quad (8)$$

In this work, the WNB classification algorithm is adopted to assign reasonable weights to attributes according to their contribution to the classification results. While maintains the high speed of the NB algorithm, it also reduces the impact of the attribute condition independence assumption on the performance of the classifier [11], as shown in equation (9).

$$p(C_j|X) = \arg \max_{C_j} p(C_j) \prod_{i=1}^n p(A_i|C_j)^{w^i} \quad (9)$$

w^i represents the weight of attribute A_i , which shows the importance of different attributes in the classification process. The larger the value of w^i , the more important the corresponding attribute A_i is for classification.

From the correlation between the evaluation index and the comprehensive evaluation value for teaching evaluation data, it can be concluded that the value of each evaluation index has a different degree of influence on the evaluation result. Therefore, it proposes a method to determine the weight of each evaluation index by the relative probability of class attributes. Each attribute A_i may have K different values, with a_k representing its specific value, where $k \in K$. A specific instance X is assumed, when the attribute A_i of is a_k , the calculation of the correlation probability and uncorrelated probability of the attribute A_i , C_j for the category is as follows.

$$p(A_i|rel) = \frac{\text{count}(A_i=a_k \wedge C_j)}{\text{count}(A_i=a_k)} \quad (10)$$

$$p(A_i|norel) = 1 - p(A_i|rel) \quad (11)$$

In equation (10), count represents statistics, the value of the attribute A_i is a_k and belongs to the C_j category, and the equation for calculating the attribute weight is as follows.

$$w(A_i, a_k, j) = \frac{p(A_i|rel)}{p(A_i|norel)} \quad (12)$$

Therefore, the specific calculation of the weighted NB classification algorithm is as follows.

$$p(C_j|X) = \arg \max_{C_j} p(C_j) \prod_{i=1}^n p(A_i|C_j)^{w(A_i, a_k, j)} \quad (13)$$

In dataset A , if there are m class labels, n attributes, and k possible values for each attribute, the total weight of all attributes is $m \times n \times k$. The weights of the same attribute are different under different circumstances. According to the specific value of each attribute, the weight of the probability associated with the current class label is selected for calculation, and the result value of each category is compared. The category corresponding to the maximum value is the classification result.

3 Results and Discussion

3.1 Correlation analysis between evaluation indexes

Based on Python and pycharm platform, the correlation analysis of 440 pre-processed teaching evaluation data sets is implemented. The correlation coefficient between evaluation indexes is calculated based on the experimental data. Table 2 shows the maximum correlation coefficient between each evaluation index.

Table 2. Maximum correlation coefficient between indexes

Index	The Most Relevant Index	Correlation coefficient
A	O	0.333
B	M	0.677
C	O	0.685
D	N	0.217
E	P	0.652
F	Q	0.356
G	E	0.53
H	M	0.502
I	R	0.265
J	L	0.342
K	E	0.473
L	D	0.537
M	B	0.677
N	A	0.459
O	C	0.685
P	E	0.652
Q	N	0.453
R	E	0.563

From the data in Table 2, there is a strong correlation between the two evaluation indexes with large correlation coefficients. Therefore, an association rule experiment is conducted on this data set, and the minimum confidence level is set to 0.500 and the minimum support level is 0.300. After calculation, strong rules for some data in Table 3 are obtained.

Table 3. Part of strong rules between indexes

Strong rules	Strong rules	Explanation
B3=>M3	0.921	Remove B
M3=>B3	0.698	
C2=>O2	0.728	Remove C
O2=>C2	0.728	
R2=>C2	0.701	
Q2=>C2	0.698	
P2=>E2	0.753	Remove P
E2=>P2	0.823	
H2=>P2	0.689	
D2=>P2	0.705	

From the analysis data in Table 2, the correlation coefficients between indexes B and M, E and P, and C and O are relatively high, indicating a strong influence relationship between them. Combined with Table 3, B, C, P, and other indexes have high confidence, and these indexes are strongly correlated with other indexes. Other indexes can replace the existence of these indexes, so there is no need to retain these indexes in teaching evaluation and they were removed, and A, D, E, F, G, H, I, J, K, L, M, N, O, Q, and R is retained. Then the original 18 evaluation indexes are optimized into 15 independent evaluation index systems.

3.2 Comparison of accuracy of four machine learning algorithms

The accuracy and feasibility of each algorithm are judged by comparing the existing evaluation data with the four commonly used machine learning classification methods in Section 2.4. 440 experimental training sets and 140 test sets are adopted, and 20 cross-validation tests are conducted and two parallel tests are conducted to calculate the accuracy of each classification algorithm. Fig. 4 shows the actual comparison results of average classification accuracy of these algorithms.

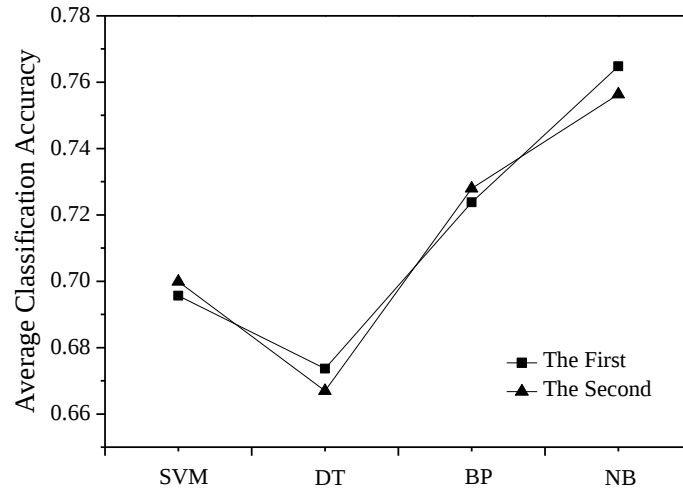


Fig. 4. Comparison of average classification accuracy of four commonly used classification algorithms

Fig. 4 presents that among the four common machine algorithms, the average classification accuracy of DT algorithm is the lowest, which is about 0.67; while NB algorithm has the highest classification accuracy, which is about 0.76. Therefore, it is proved that NB classification algorithm has good accuracy in the construction of teaching evaluation model. The average running duration of four classification algorithms is tested on the same data set. Fig. 5 shows the average running duration of various machine learning algorithms in processing the same number of data sets.

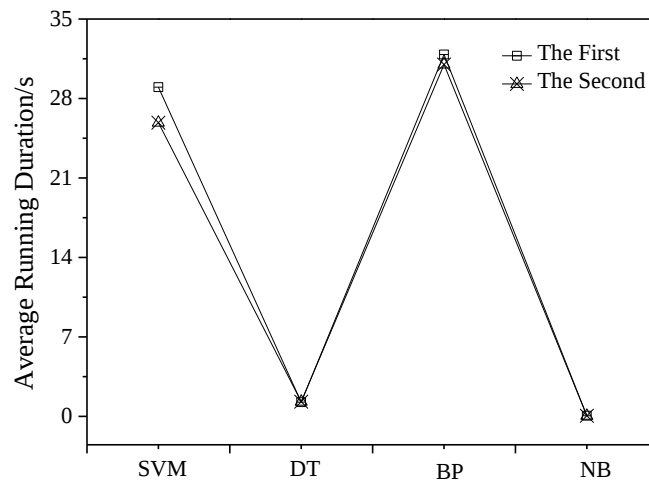


Fig. 5. Average running duration

Fig. 5 discloses that the running duration of the DT classification algorithm and NB classification algorithm in processing the average operation of the same data set are smaller than those of SVM and BP algorithms, while NB algorithm takes less time than DT algorithm. Combined with the experimental results, compared with other classification algorithms, the NB algorithm has a higher classification accuracy in the data set of the teaching evaluation system and has the shortest running duration. Therefore, the NB algorithm is chosen to construct the teaching evaluation model.

3.3 Comparative analysis of classification accuracy between NB algorithm and WNB algorithm

440 data records are extracted from the evaluation and teaching database as the training set, and 140 data as the test set. 10 cross-validation experiments are implemented to test the classification accuracy of NB algorithm and WNB algorithm. Fig. 6 shows the comparison results of the accuracy of NB algorithm and WNB algorithm.

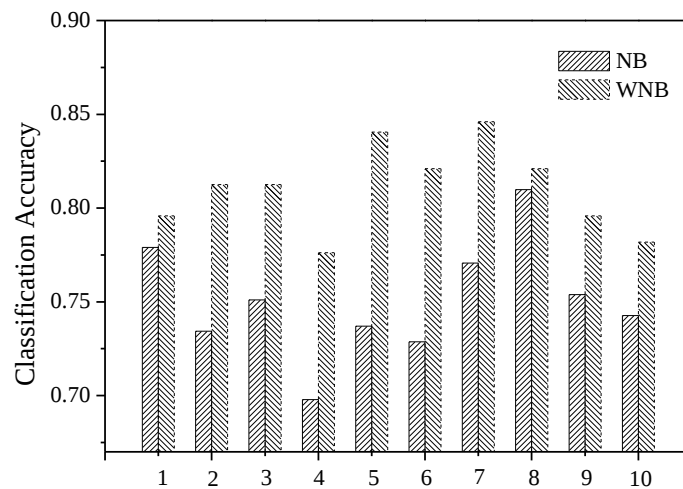


Fig. 6. Classification accuracy of NB algorithm and WNB algorithm

According to the experimental results, the classification accuracy of the WNB algorithm is higher than the classification accuracy of the NB algorithm generally. On this data set, the average classification accuracy of WNB is 0.817, while the average classification accuracy of NB algorithm is 0.751. Therefore, it is suggested that the weighted NB algorithm has a better classification accuracy on the teaching evaluation data than the traditional NB algorithm.

3.4 Comparison experiment of accuracy between BP algorithm and WNB algorithm

In the field of teaching evaluation, the most commonly used classification method is the BP neural network [12]. The weighted NB classification method is adopted in this work to construct the teaching evaluation model, and the following content is the comparison of data classification accuracy of WNB and BP algorithm.

BP classification algorithm and WNB algorithm are trained with 440 data, and 140 data are tested. When BP algorithm is trained, the number of input layer nodes is set as 15, hidden layer node as 6, output layer node as 1, activation function is “tanh”, the learning rate is 0.01, and the number of cycles is 20,000. Then, the BP algorithm and the WNB algorithm are compared in 10 cross experiments. Fig. 7 shows the comparison results of the classification accuracy between the BP algorithm and the WNB algorithm.

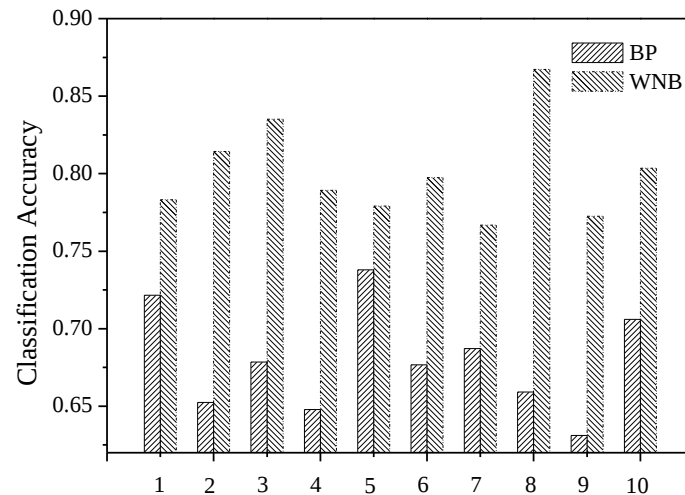


Fig. 7. Comparison of classification accuracy between BP algorithm and WNB algorithm

According to the experimental results, the classification accuracy of the WNB algorithm is higher than the classification accuracy of the BP algorithm generally. On this data set, the classification accuracy of WNB is 0.800, while the classification accuracy of BP algorithm is 0.680. Therefore, it can be obtained that the WNB algorithm can achieve better results in the classification processing on the teaching evaluation data set than the BP artificial neural network classification algorithm.

4 Conclusion

To build a teacher teaching evaluation system that adapts to the education environment in the new era, the machine learning algorithm under the background of AI is adopted to build the teaching evaluation system, and the role of machine learning algorithm in teacher teaching evaluation system is discussed. Firstly, it poses the existing problems in the current teaching evaluation system, then puts forward corresponding solutions and introduces the relevant theories and technologies needed to solve the problems. Then, the actual construction of teaching evaluation system model is implemented. Finally, the performance of the teacher teaching evaluation model is tested and compared with that of the BP algorithm. The results reveal that the evaluation system model designed in this work has a good performance in both accuracy and speed. However, there are still some deficiencies, and there are deviations between the operational conclusion of the actual algorithm and the theoretical value, and more data is needed to train the model before being put into use. In addition, the artificially set confidence interval will affect the objectivity of the actual results, which should be solved in the follow-up research.

5 References

- [1] Zou S. (2017). Designing and practice of a college English teaching platform based on artificial intelligence. *Journal of Computational and Theoretical Nanoscience*,14(1): 104-108. <https://doi.org/10.1166/jctn.2017.6133>
- [2] Eaton E., Koenig S., Schulz C., et al. (2018). Blue sky ideas in artificial intelligence education from the EAAI 2017 new and future AI educator program. *AI Matters*,3(4): 23-31. <https://doi.org/10.1145/3175502.3175509>
- [3] Kim K., Park Y. (2017). A Development and Application of the Teaching and Learning Model of Artificial Intelligence Education for Elementary Students. *Journal of The Korean Association of Information Education*,21(1): 139-149
- [4] Zhao X., Guo H. T., Huang C. L., et al. (2017). Teaching evaluation system research based on structure entropy weight method. *Journal of Discrete Mathematical Sciences and Cryptography*,20(1): 179-191. <https://doi.org/10.1080/09720529.2016.1178915>
- [5] Wei Z., Baocun L. (2018). The undergraduate teaching evaluation system in China: Progress, problems and suggestions. *Chinese Education & Society*,51(4): 248-259. <https://doi.org/10.1080/10611932.2018.1493865>
- [6] Wang Z., Liu J. (2019). A teaching quality evaluation system of massive open online courses based on big data analysis. *International Journal of Emerging Technologies in Learning (iJET)*,14(14): 81-91. <https://doi.org/10.3991/ijet.v14i14.10818>
- [7] Rizzuto M. (2017). Design recommendations for self-paced online faculty development courses. *TechTrends*,61(1): 77-86. <https://doi.org/10.1007/s11528-016-0130-8>
- [8] Ho J., Plewa C., Lu V. N. (2016). Examining strategic orientation complementarity using multiple regression analysis and fuzzy set QCA. *Journal of Business Research*,69(6): 2199-2205. <https://doi.org/10.1016/j.jbusres.2015.12.030>
- [9] Osadchiy T., Poliakov I., Olivier P., et al. (2019). Recommender system based on pairwise association rules[J]. *Expert Systems with Applications*,115: 535-542. <https://doi.org/10.1016/j.eswa.2018.07.077>

- [10] Brewster L. R., Dale J. J., Guttridge T. L., et al. (2018). Development and application of a machine learning algorithm for classification of elasmobranch behaviour from accelerometry data. *Marine Biology*,165(4): 62. <https://doi.org/10.1007/s00227-018-3318-y>
- [11] Jiang L., Zhang L., Li C., et al. (2018). A correlation-based feature weighting filter for naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*,31(2): 201-213. <https://doi.org/10.1109/tkde.2018.2836440>
- [12] Zhou Y., Zheng S., Zhang G. (2020). Machine-learning based study on the on-site renewable electrical performance of an optimal hybrid PCMs integrated renewable system with high-level parameters' uncertainties. *Renewable Energy*,151: 403-418. <https://doi.org/10.1016/j.renene.2019.11.037>

6 Author

Jingjing Hu is a lecturer and master of Zhejiang Business College, mainly engaged in teaching management and student learning evaluation.

Article submitted 2020-12-05. Resubmitted 2021-01-13. Final acceptance 2021-01-16. Final version published as submitted by the authors.