A Review of the Research on the Prediction of Learning Outcomes in the Field of Learning Analytics

Yuang Wei
North China University of Technology, China
2644595124@qq.com

Jining Xu[*]
North China University of Technology, China
jxu0422@ncut.edu.cn

Zehua Zhang
North China University of Technology, China
1124608664@qq.com

Zhijun Li
North China University of Technology China
lzj78@ncut.edu.cn

ABSTRACT

Learning analytics is a new research field that seeks beneficial results from various education data. Recently, the prediction of students' learning outcomes is becoming one of the hottest research issues. Despite large amounts of multilevel and multidimensional theoretical and practical research that has been performed by domestic and foreign scholars, comprehensive summaries are still lacking. Through systematic literature retrieval, this paper sorts out the current research hotspots, research limitations and main directions in the field of predictive analysis of learning results. In order to clearly describe and analyze the literature results, this paper uses the teaching goal classification theory, the individualized learning theory and the social cognitive theory as tools to classify the research, and conducts a systematic analysis of the empirical research results at different levels. The results show that in the current study of learning outcome prediction, most of them are based on knowledge space and made through students' learning behaviors, while there are few evaluations and predictions of students' potential traits from the level of ability and thinking. However, in recent years, the number of research results in this direction has been on the rise. Finally, this paper raises questions for future research and practice in the field may appear and potential research directions.

CCS CONCEPTS

- **Applied computing** → Education; E-learning.

KEYWORDS

Learning analytics, Learning outcome prediction, Predictive index, Learning behavior analysis, Literature review

ACM Reference Format:

Yuang Wei, Jining Xu, Zehua Zhang, and Zhijun Li. 2021. A Review of the Research on the Prediction of Learning Outcomes in the Field of Learning Analytics. In *2021 5th International Conference on Education and Multimedia Technology (ICEMT) (ICEMT 2021), July 23–25, 2021, Kyoto, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3481056.3481077

1 OVERVIEW OF STUDY ANALYSIS AND RESEARCH

1.1 The Source of Learning Analytics

With the advent of the era of big data, the application of data mining and analysis technologies in various traditional fields has become a norm. In recent years, with the global development of MOOCs, massive amounts of learning process data have been obtained. Applying data mining and data analysis to the field of education and optimizing teaching effects are the common needs of educators. At the same time, networked teaching, diversified interactive learning, and diversified teaching also provide new scenarios and development opportunities for applications in the field of learning analytics.

Learning analytics technology is a specific manifestation of education and teaching data mining. It mainly focuses on the use of advanced analysis technology and analysis tools for learning problem exploration, student learning ability evaluation, student learning result prediction, and optimal learning route research. At the end of February 2011, the First International Conference on Learning Analytics & Knowledge (LAKE) held in Banff, Alberta, Canada [1] defined Learning Analytics as: measurement, collect, analyze and report on student data of learning behavior and the learning environment, and it is the technology which used to understand and optimize learning and the environment it creates. Professor Siemens [2] of Athabasca University, Canada, in his "Learning Analytics: An Emerging Discipline" respectively discussed the roots of learning analytics technology, key technologies affecting learning analytics, popular learning analytics tools, techniques and applications procedures, and knowledge modeling are systematically explained, and a learning analytics cycle model is proposed, which includes seven parts: collection, storage, data cleaning, data integration, analysis, visualization, and action.

[*] Yuang Wei: primary author, China, master degree candidate. The major research directions are control theory, pattern recognition, and intelligent education.
Jining Xu: corresponding author& second author, China, associate professor, master's supervisor, Ph.D. The major research directions are control theory application, signal test and analysis, and Fieldbus application technology.
Zehua Zhang: third author, China, master degree candidate. The major research directions are control theory, intelligent education, and intelligent recommendation.
Zhijun Li: fourth author, China, full professor, master's supervisor. The major research directions are networked control system, fault diagnosis and fault-tolerant control, and advanced control algorithm.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICEMT 2021, July 23–25, 2021, Kyoto, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9022-4/21/07…$15.00
https://doi.org/10.1145/3481056.3481077

1.2 Research in The Field of Learning Analytics at Home and Abroad

1.2.1 Current Status of Foreign Research. In 2011, Brown [3] and Siemens proposed a model of learning analytics and learning process [4] based on its definition, and then made a continuous improvement loop model combined cycle characteristics [5]. Following the development of learning analytics model theory, many scholars have conducted empirical research. The learning analytics experiment carried out by Canadian scholar Leah P. Macfadyen [6] et al. used 3 semester learning behavior data of 118 students for regression analysis. The experiment showed that the accuracy of the model was 81%. From the theoretical perspective, it proves the possibility of developing an "early warning system" based on LMS (Learning Management System, LMS) data mining. Through the system, teachers can identify the danger signs of students' learning process early and give corresponding learning assistance in time. In 2011, the Society for Learning Analystics Research (SOLAR) proposed an integrated learning analytics system to construct a more comprehensive learning analytics system [7].

With the continuous enrichment of data, scholars have successively discovered and improved some problems in the learning model when the amount of data is sufficient. Elias [5] analyzed the learning process and its related stakeholders, and proposed the improvement of learning analytics. Model, the model includes three parts: data collection, information processing, and knowledge application. The data collection module involves data selection and capture; the information processing module involves integration and prediction; the knowledge application module involves extraction and application. These three parts are supported by four parts: organization, computer, related personnel and theory. Verbert [8] and other researchers proposed that the process of learning analytics should include four levels: perception, reflection, meaning construction, and influence. The perception phase represents data visualization; the reflection phase represents the adoption of opinions and functional evaluation; the meaning construction phase represents the suggestion of opinions after reflection and the innovation of new ideas; the influence phase is to implement ideas that users find useful and change their behavior.

Through the analysis of existing learning analytics models, Ifenthaler [9] pointed out that individual models still have shortcomings in guiding practice, and proposed a more specific learning analytics content framework on this basis. The framework includes content: (1) Personal characteristics (2) Social network (3) Body data (4) Courses (5) Online learning environment (6) Learning analytics engine (7) Reporting engine (8) Personalization and adaptive engine (9) Institutional strategy; (10) Management decision-making. Each part of the content discards the one-way linear relationship of the previous learning and analysis content, and instead forms a two-way multi-layer relationship, and the content of each part is further refined to make it more reliable and operable in guiding practice.

1.2.2 Current Status of Domestic Research. Compared with foreign countries, domestic research in the field of learning analytics started late, and most of them are introductions and reviews of some foreign studies.

Y. Y Li [10] and others proposed a conceptual model of learning analytics, which is an earlier research result in the field of learning analytics in China. Z. T Zhu [11] proposed a learning analytics model for smart education. X. Q Gu [12] and others in "learning analytics: emerging data mining technology" elaborated on the "past and present life" of learning analytics, the developing learning analytics technology and the application trend of learning analytics.

In recent years, domestic scholars have paid much attention to the field of learning. For example, Qing Li [13] and others introduced foreign learning analytics technology concepts, models, elements, sources, tools, and methods. F. Q Li [14] and others made a more in-depth interpretation of the connotation, process, tools, and methods of learning analytics, trying to promote the reform of learning analytics in schools and using it to promote better learning effects. S. P Wei [15] and others are analyzing the learning analytics technology at home and abroad, summarizing the key technology and analysis mode of learning analytics technology and showing the importance of learning analytics in online courses and the process of application through different perspectives.

X. H Yu, X. Q Gu, and others [16] established a behavioral model of learning analytics—the learning activity flow model. It complements the diversity of learning sources and the persistence of learning activities that were not considered in the previous analysis of learning behavior. And apply the model to the PLE-SRL (Personal Learning Environment Based Self-Regulated Learning, PLE-SRL) learning platform, and conduct empirical research on it. Z. T Zhu [11] and others regarded learning analytics technology as an important technical pillar of smart education, and respectively elaborated on the general design framework, resource process model and process dimension model of learning analytics.

1.3 Summary Method and Focus Points

In this review, we found 47 articles in 2010-2020 and 21 articles related to performance prediction on Knowledge Network, using learning, analysis, prediction and education as the common search theme. In order to ensure that this review can cover representative scholars and typical cases in the field, in addition to the keyword-based database search, the research literature of major scholars and representative research teams in the field is also searched separately. Through literature and literature citations, representative scholars' articles were cross-matched with keyword search articles, and literature was deleted and selected according to the degree of focus. Finally, a total of 60 articles were selected for review and analysis.

Screening conditions: The research should have a clear prediction object, that is, learners' academic achievements or academic risks or related cognitive skills improvement, etc.; Secondly, the research should clarify the study place and the participants of the research, the research method should be explicit, the predictive index should be sufficient, and the analysis process should be clear in the literature.

By summarizing the literature, it is found that although the problems currently studied by scholars seem to be in full bloom, the documents can be classified into several types of prediction, modeling, and correlation analysis. At present, there are more researches on correlation analysis of learning process data, and classification

Table 1: Summary of literature and algorithms involved in the review

Data sources and types	Research and analyze experimental results	Researcher	Percentage in the review (%)	Algorithm used
The Learning process and test data	Dropout rate	J Ma (2014), Balakrishnan (2013), Lara (2014), Z. X Jiang (2015), Bukralia (2015), Tsiakmaki M (2020)	16.8	Multiple regression, hidden Markov
	Course grades	Leah P. Macfadyen (2010), Brown (2012), Verbert (2014), C. K He and M Wu (2016), Z. J Mou and F. T Wu (2017), J. Y Wang (2020), X. L Song (2020), J Ma (2014), X. Z Wang and C. Q Huang (2018), H. T Sun (2012), F. T Wu and H Tian (2019), Boyer & Veeramachaneni (2015), Z. X Jiang (2015), Moreno-Marcos P. M. (2020) [54], Umer R. (2019) [55], Sokkhey P (2020) [56]	48.4	Random forest, linear regression, elastic network regression, support vector machine regression, gradient boosting tree, logistic regression
	The solution of ability mastery	Fujimoto (2020), Kaser (2017)	6.4	Naive Bayes model, dynamic Bayesian Network
	The solution of knowledge mastery	Pardos (2010), Chang (2006) [57]	6.4	Bayesian network
Student background data	Course grades	Lacave & Molina (2018) [58]	3.2	Bayesian network
Student social	Dropout rate	Lara (2014), Bukralia(2015)	6.4	Feedforward neural network, support vector machine, probability ensemble SFAM classifier
behavior data	Course grades	H. T Sun (2012), Ming N.C (2012) [59], Heise N (2020) [60]	9.6	Statistical Analysis
Basic problem solved				Algorithms suitable for solving
Predict the results of students' learning through the trend of students' learning process				Various regression algorithms
Prediction of learning outcomes of state variables in students' learning process over time				Hidden Markov
Correlation analysis of process data				Classification algorithms
Build network structure, causality model, predict learning results				Bayesian network
Summarize various data of students				Statistical Analysis

algorithm and regression algorithm at algorithm level are more popular, as shown in Table 1

2 MODEL AND METHOD OF PREDICTION AND ANALYSIS OF LEARNING RESULTS

A common question in learning outcome prediction is "How to predict learners' academic success or academic failure?" Researchers will choose different combinations of learning outcome indicators according to the needs of different scenarios, including student performance, dropout rate, classroom participation rate, Ability, knowledge mastery, homework completion, etc.

Early Brown [17] sorted out the widely validated predictive indicators from the three aspects of learner's inherent indicators, behavioral performance indicators, and student works. And he discussed the predictive abilities and application cases of different indicators. Usamah [18] summarized 14 typical learning analytics systems and applications, and made predictions for learners. For the data needed for prediction, Verbert [19] and others focused on the data collection of test results such as learning behavior, social interaction, resource use, and time spent. Bukralia [20] chose academic ability, personal property, academic goals, and learning motivation as the input variables of the learning analytics system. Berry [21] summarized four types of factors that affect the continuity of learning (social factors, psychological factors, organizational factors, and economic factors) for the indicators that affect the prediction results, and three types of indicators that affect academic achievement (academic factors, demographics) Factors and cultural and social factors), expanding the scope of basic data sources for learning analytics.

Domestic scholars pay more attention to the predictive indicators related to learning behavior in the prediction research of learning analytics. In recent years, F. T Wu and Z. J Mou [22] proposed a learning outcome prediction framework based on learner behavior analysis, S Li [23] and others proposed six indicator dimensions based on learner online learning behavior input, C. K He and M Wu [24] By analyzing 16 MOOCs on the edX platform, it summarized the learning behavior characteristics of learners in multiple dimensions, and carried out data mining on some typical behavior characteristics and then made behavior predictions. Based on the original prediction framework, Z. J Mou and F. T Wu [25] tested the prediction ability of key indicators such as the number of video learning times, the number of text learning times, and the duration of evaluation participation on the MOOC platform. J. Y Jia and Y. Y Yu [26] proposed an online learning activity index based on the three dimensions of speed, quality, and quantity.

Based on the classification theory of teaching objectives, personalized learning theory and social cognitive theory, and the classification of learning results in the personalized behavior analysis model [22], this paper makes a classification analysis and summary of the existing literature. Specifically, the classification theory of teaching objectives is applied to the result analysis of learning characteristics and learning completion degree; using the theory of individualized learning analytics, we discuss the outcome analysis of learning activity participation and students' ability; The social cognition theory is used for the result analysis of the interactive level and the result analysis of the knowledge map. Based on these analysis ideas, this article divides the research results of learning outcome prediction into five categories, which are described as follows:

2.1 Analysis of Results Based on Learning Characteristics and Learning Completion

This type of research is based on learning completion and personal learning characteristics data to analyze the learning results. J. Y Wang [27] and others collected a data set of students in a certain area of Portugal through the UCI official machine learning database, and evaluated it three times in each academic year. In the test, set the three tests as G1, G2, G3, where G3 is the predicted target feature; the attributes with high correlation were selected by the correlation calculation of the original data set; the data set without feature selection is compared with the data set after chi-square test and random forest feature selection. It is found that the data set after feature selection based on the random forest method can make the prediction algorithm more accurate. Based on the same data set, X. L Song [28] and others conducted a feature selection study on the data. They chose linear regression, elastic network regression, support vector machine regression, and gradient boosting tree for data selection, compare algorithm performance indicators.

In an experiment to study the effect of completion of learning activities on learning outcomes, J Ma [29] and others used testing and discussion as learning activities, and performed multiple regression modeling on the degree of completion of the activities. The results show that the model will help teachers identify problems, a clear direction of the next training, played a role in early warning.

In the study of Z. X Jiang [30], the complete degree of "watching video + submitting homework" was used to analyze the learning of six courses (P, A, C, B, D, and I). The degree of completion is divided into five kinds of learners (Bystander, Anti-climax, Waver, Stalwart, Just-learner), the statistical number of learners who complete these activities.

At the same time, the statistics of the number of people who have obtained the network certificate (successful learning) and the number of times submitted to the test (the degree of completion of the activity). Then, according to the total score, the score statistics of the learners of different activity completion degrees are performed. Statistics show that the average score of those who are determined to complete is higher than the passing line, while the scores of other types of learners are almost 0 points. Obviously, it can be considered that those who are determined to complete are all potential certificate recipients. This empirical study shows an important correlation between activity completion and learning outcomes.

2.2 Analysis of Learning Results Based on Participation in Learning Activities

In the research on the degree of participation in learning activities, C. K He and M Wu [24] based on the operation of 16 courses of Harvard University and MIT on the edX platform from the fall of 2012 to the summer of 2013, first selects typical learning behaviors. Select the learning time, the number of learning events, the number of sampling statistics, the number of watching videos, the number of learning chapters, the number of posts, etc. as the key records of learning participation; the algorithm framework of logistic regression is constructed to predict whether learners can complete the learning task and obtain a certificate. It is confirmed that the learning result can be predicted through the analysis of learning behavior.

In the study of X. Z Wang and C. Q Huang [31], the theory that the Gini index is the uncertainty measure of random variables in information theory is used as the discriminant basis for the correlation between the characteristics of learning situation and results. The learning situation analysis experiment based on big data shows that in the comparison of the Gini gain results of each feature, the proportion of learning participation is 15.85%.

The analysis of learning participation can also predict the probability of students dropping out. Balakrishnan [32] uses the hidden Markov model to predict the probability of students dropping out according to the degree of learning participation based on the learning data of the MOOC of Duke University in 2012. The results show that students with high learning participation have a low probability of dropping out.

In the research of Maria Tsiakmaki [33], a custom Moodle plug-in was used to collect the usage of activities, resources, folders, forums, and the performance data of various quizzes. Collect data on a monthly basis, and evaluate the importance of features in prediction through extreme random trees. The prediction is carried out on the framework of AutoML compared with the classical data mining algorithm. Research shows that the Auto-WEKA algorithm under the AutoML framework can achieve higher prediction accuracy.

2.3 Analysis of Learning Results Based on Learning Ability

The calculation of learning ability is a unique path in the field of learning analytics. It includes research in the field of psychometrics. The item response theory proposed by the American measurement expert Lord in 1952 is now usually used to measure student ability.

Item response theory is based on the model of measurement, which believes that the estimation of the trait level of the measured ability depends on the individual's response behavior and the attribute of the measured item itself.

When item response theory is applied to the actual measurement process, it is an ideal state that one test corresponds to one ability. When each topic corresponds to multiple abilities, there will be a lot of problems that will lead to inaccuracy. Therefore, the multidimensional project response theory is the main research direction in recent years.

In the latest research of Fujimoto [34], based on the relevant data of 121 students in 29 projects, a hierarchical and dimensional project response theory model was constructed, which are Two-Tier Structure, Bifactor Structure, Six-Dimensional Structure, Two-Dimensional Structure, Unidimensional Structure. Afterwards, each model was solved posteriorly through the Naive Bayes method, and then the students' performance in each project was analyzed.

With the continuous development of machine learning technology, learning models and machine learning are also merging with each other. Kaser [35] and others introduced dynamic Bayesian networks to analyze the learning results of students' mathematical calculation abilities; through the measurement of different abilities, the dynamic Bayesian network was constructed with tests as the time axis. In the process of continuous updating and iteration, the probability of the student's learning result is obtained.

In the process of educational measurement, question bank construction is an important part. H. F Mu [36] constructed a computer adaptive English language test question bank based on item response theory. Y Wang [37] and others took 1633 students as the test target, chose the "modern educational technology" public course as the test course, used the project response theory analysis software BILOG to fit the data and the model, and used the model fitting index AIC (Akaike Information Criterion, referred to as AIC), BIC (Bayesian Information Criterion, referred to as BIC) and -2LL (-2Log-Likelihood, referred to as -2LL) for analysis. Delete the test questions whose significance level is less than or equal to 0.001, and then analyze the degree of discrimination and difficulty of the selected test questions, and find that they basically belong to normal distributions.

After analyzing the actual test data, the relevant conclusions about the amount of test information are obtained: The test information volume reaches the maximum I(θ)=44.76 at θ =-0.48, the standard error of the ability estimate is SE(θ)=15, and the test reliability rxx is as high as 98%. The test information content is ≥20 in the interval range where θ belongs to [-3.67, 2.50], and the minimum standard error SE (θ) is 0.22, which means that the test reliability RXX is ≥0.95 in the interval range of this ability.

The construction of a high-precision question bank can better understand the pros and cons of students, and even evaluate the lack of thinking level, and more accurately predict the learning effect.

2.4 Analysis of Learning Results Based on Interactive Hierarchical Analysis

Laurillard [38] proposed a conversation model in the learning process from the perspective of the interaction between teaching and learning. The interaction behavior in the model can be understood as follows: During the learning process, two levels of interaction will occur at the same time. The first level is the interaction between learner's behavior and the environment constructed by the teacher; This process is represented by the learning environment provides learners with tasks and feedbacks and learners take certain responses and adjust their behaviors according to the tasks and feedbacks, which is defined as adaptive interaction; another level of interaction is embedded in knowledge, which is the interaction between the concepts of students and teachers; The concept expressed as a teacher is described and presented in a certain way, and then acts on the student's concept construction process, leading to the change of the student's concept, and describing it in a certain way. This process is defined as conversational interaction.

In recent years, with the popularization and use of various online education platforms, human-computer interaction has become more common, generating massive amounts of interactive information data. H. T Sun [39] used the Moodle platform to record the learning behavior data of 106 students in basic computer network courses, and analyzed more than 62,000 behavior records. The results show that the positive interaction process will have a positive impact on the learning effect.

Boyer and Veeramachaneni [40] conducted data statistics of 14 interactive information variables and behavioral information variables on the three-quarters of MIT courses on the edX platform, including part of the time variables of students' learning, the number of learning achievements, the proportion of learning results, etc. This information contains good data requirements to meet the adaptive interaction and conversational interaction, as shown in Table 2

The paper cross-combines the data sources and targets of different quarters and then performs regression analysis, and compares and selects the model with the highest detection accuracy to predict the learning effect of students. In the teaching process, it also gave feedback to students' learning and conducted positive interventions, which achieved good results.

2.5 Analysis of Learning Results Based on Knowledge Mapping

The knowledge mapping is transformed from the "knowledge engineering" first proposed by Professor Edward Feigenbaum of Stanford University. As a tool that can be displayed and evaluated visually, it can reflect learners' reserve of own knowledge more realistically and more intimately. The reserve situation of the learner, the construction of a more valuable association between the learner and the unknown target knowledge or entity, has the function of discovering the learner's potential learning problems and predicting their future learning performance. In the adaptive learning system, teaching managers can provide targeted guidance and intervention in teaching based on the predicted results of the knowledge map and learner performance, which can promote effective teaching and learning behaviors.

Educational knowledge graphs can be divided into two categories: static knowledge graphs and dynamic affair graphs [41]. The static knowledge graph uses the elements in the teaching process as nodes, and the logic between the elements is connected to

Table 2: List of interactive information [40]

Serial number	Interactive information
X1	Total time spent on all resources
X2	Number of different questions tried
X3	Number of submissions
X4	Number of right questions
X5	The average number of submissions per question
X6	The ratio of the time to correct the problem to the total time
X7	The ratio of the number of right questions to the number of tried questions
X8	Duration of the longest observed event
X9	Total time spent on lecture resources
X10	Time spent on book resources
X11	Total time spent on wiki resources
X12	Number of correct submissions
X13	The correct percentage of the total number of submissions
X14	The average time between issue submission and issue due date

form a network; the dynamic affair graph uses teaching events or activities as the representative object and uses the logical affair relationship as the connection to form a multi-relation graph.

Although the knowledge graph has broad application prospects in intelligent education, it is still in its infancy in terms of theoretical research, platform development, and application demonstration.

Regarding the general technical research on knowledge graphs, Q Liu et al. [42] proposed the technical framework of knowledge graphs, which was divided into three levels: information extraction layer, knowledge fusion layer, and knowledge processing layer; Z. L Xu [43] and others believed that the technologies involved in the knowledge graph include four aspects: knowledge extraction, knowledge representation, knowledge fusion, and knowledge reasoning; G. L Qi [44] proposed that the key technologies of knowledge graph construction include entity relationship recognition technology, knowledge fusion technology, entity link technology and so on.

In terms of empirical research, Pardos [45] et al. built a knowledge graph through knowledge points tracing, and used Bayesian networks to predict student learning effects. The model sets all priors of the PPS (Prior Per Student) model to the same value or assigns a prior value of a student as a shared value. In the model, the initial knowledge is equivalent to the prior knowledge, and the prior personalization is achieved by adding student nodes. The conditional probability table of the initial knowledge node is determined based on the student node value. The student node itself also has a conditional probability table, which determines the probability of a student with a specific ID. The parameter of this node is fixed at 1/N, where N is the number of students. The research builds a Bayesian network model based on initial knowledge + personalization + guess + slip probability, and predicts learning results.

In the model demonstration stage, they conducted tests and data collection on students in grades 7-12 on the mathematics tutoring system to simulate the different learning speeds of different students. The PPS model achieves 0.3515 on the average correlation coefficient of all problem sets, which is far better than the 0.1933 of the general knowledge tracing model (KT for short), and also achieves high accuracy in prediction.

3 THE ALGORITHM OF LEARNING RESULT PREDICTION

In the learning result prediction analysis model, scholars have used a variety of algorithms to predict the results. Under the same data, comparing the performance of different algorithms is also one of the key points of research

When Lara [46] and others studied the probability of students dropping out of school, they collected 100 students' courses for one semester, and selected four models to predict the probability of school dropout respectively: Educational Data Mining System (referred to as SEDM), Feedforward Neural Network (referred to as FFNN), support vector machine (referred to as SVM) and probabilistic integrated SFAM classifier (referred to as PESFAM). Each course consists of 10 units, 4 tests, 4 continuous assessment activities, and a final face-to-face examination. The predictive analysis results are shown in Figure 1

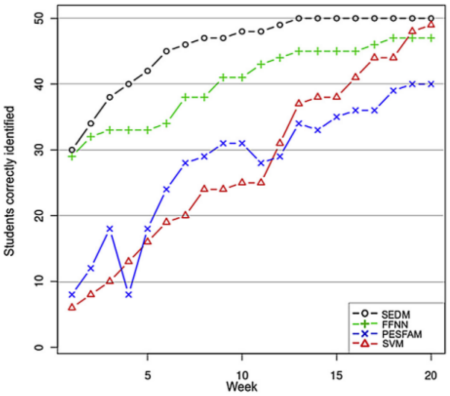

Figure 1: Prediction of student dropout probability under four models [46]

Table 3: Evaluation indicators of the four models in the tenth week [46]

Proposal	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)
SEDM	81.8	90	95	94
FFNN	55.6	50	90	82
PESFAM	23.5	40	67.5	62
SVM	10.5	20	57.5	50

Table 4: The integrated classifier evaluation indicators after bagging and upgrading a single classifier [47]

	Method of constructing a single classifier				Method of constructing an integrated classifier								
					Bagging				Boosting				Random Forest
Classifier type	BN	DT	ANN	SVM	BN	DT	ANN	SVM	BN	DT	ANN	SVM	DT
True Positive Rate	0.733	0.748	0.721	0.758	0.758	0.756	0.769	0.746	0.744	0.756	0.767	0.758	0.777
False Positive Rate	0.151	0.146	0.166	0.143	0.139	0.143	0.136	0.151	0.142	0.142	0.136	0.139	0.134
Accuracy	0.733	0.747	0.722	0.757	0.757	0.755	0.769	0.746	0.743	0.756	0.767	0.757	0.777
Recall Ratio	0.733	0.748	0.721	0.758	0.758	0.756	0.769	0.746	0.744	0.756	0.767	0.758	0.777

Table 5: Evaluation index values of eight classification algorithms before and after feature selection [48]

Algorithm	Before Feature Selection	After Feature Selection					
	Accuracy	Accuracy	Precision	Recall Ratio	F1	Mean Absolute Deviation	Root Mean Squared Error
Bayesian network	71.30%	71.30%	0.697	0.713	0.704	0.1581	0.3071
Naive Bayes model	64.81%	65.74%	0.638	0.657	0.645	0.1579	0.3416
Support Vector Machine	62.04%	62.04%	0.601	0.620	0.610	0.1519	0.3897
Sequential minimal optimization	59.26%	60.19%	0.584	0.602	0.592	0.2593	0.3452
logistic regression	62.96%	64.81%	0.666	0.648	0.653	0.1644	0.3494
Association rule	62.04%	65.74%	0.643	0.657	0.644	0.1709	0.3378
Decision Tree	62.04%	62.96%	0.623	0.630	0.625	0.1591	0.3628
Random forest	72.22%	73.15%	0.730	0.731	0.729	0.1560	0.2744

The performance evaluation data of the classifier can be intercepted for detailed observation and comparison. For example, the tenth week data table in the paper is shown in Table3.

It can be seen that SEDM has a better predictive effect on the data set selected in this paper than the other three methods.

Z. J Chen [47] studied the application effect of the latest ensemble classifier. The thesis selects 480 valid records and 16 attributes in the Kalboard360 learning management system. It is defined as low level (Low) with a score below 70, middle level (Middle) from 70 to 89, and high level (High) with a score above 90. Through two ensemble learning methods of Bagging and Boosting, different combination integration is carried out for a single classifier. The comparison table of prediction effects in the cited papers is shown in Table 4 below:

The results show that for the three algorithms of BN (Bayesian Network), DT (Decision Tree), and ANN (Artificial Neural Network), the classification performance can be improved to varying degrees by constructing an ensemble classifier; ensemble learning improved the performance of ANN type base classifier most significantly (increased by 6.5%); the DT type ensemble classifier obtained by the random forest method has the best performance. At the same time, it is shown that for the SVM algorithm, the construction of an integrated classifier cannot improve the classification performance. Compared with a single classifier, the performance is slightly reduced.

F. T Wu and H Tian [48] added learning behavior features into the algorithm to improve the prediction accuracy of the algorithm. They selected nearly 9,000 platform data records from 108 students, initially extracted 10 feature variables, and retained 8 feature variables after feature selection. The experiment uses eight classification algorithms at the same time, and uses the set of selected and unused feature variables as the basic data to predict the learning results. The test of the prediction algorithm uses the five-fold cross-check method. The test results in the cited papers are shown in Table 5

The results show that after deleting the two features, the prediction accuracy of the Bayesian network and the support vector machine algorithm remains unchanged. The accuracy of the other six algorithms has improved to varying degrees. Fewer "effective" feature variables were used to achieve better prediction results, indicating the effectiveness of feature selection.

4 CONCLUSION

The field of learning analytics is an emerging research field of education technology. At this stage, the analysis of learning effects based on different methods is still present as a relatively independent branch. Focusing on the prediction and analysis of learning results, this paper focuses on the model theory and empirical research of learning analytics methods, as well as the application effect of analysis and prediction algorithms, and sorts out the relevant literature results according to six different directions. It can be seen that the research in different directions has a strong degree of concentration at the problem level and the algorithm level, and the research conclusions show good support for each other. From this, we can see the research trend of learning outcome prediction in the field of learning analytics in the short term.

Among them, analysis based on learning interaction is a recent hot spot. Learning in the 21st century is understood as a participatory process [49], and the development of information interaction technology allows us to understand this process more clearly. For example, research on MOOCs and learning behavior analysis provides quantitative data in the learning process and improves the dimension of learning analytics. However, only learning interaction research is an incomplete analysis. The current MOOC-based predictive analysis has focused and in-depth on some issues, such as the specific discussion of the behavioral influence of video viewing, forum participation, etc. However, some very important tendency indicators are relatively less involved, such as past learning performance [50], initial knowledge [51], skill base [52], and learning drive [53].

Another research branch, project response theory and knowledge graphs can handle learning performance factors well, and there are also many empirical verifications in the quantification of skill base and solving predictions. The development of single-dimensional item response theory to multi-dimensional item response theory promotes the measurement and evaluation of students' ability. The progress based on the knowledge graph is strongly related to the research of initial knowledge and learning driving force.

On the one hand, future research and practice can comprehensively consider a variety of factors and algorithm models to enhance the comprehensive effect of learning analytics; on the other hand, it can uphold the concept of "building meaning from data" and combine learning basic theories to improve the level of model construction, and strive to achieve Accurate understanding and description of learner characteristics, the professional judgment of teaching process and effective intervention of academic risk. With the continuous advancement of technology, researchers should also be vigilant about the value judgment and meaningless mining of data to avoid misleading teachers and negatively impact students. Let learning analytics technology really help teaching and help achieve efficient and meaningful learning.

REFERENCES

[1] LAKE. Learning Analytics and Knowledge. Retrieved 2012. http://lak12.sites.olt.ubc.ca.

[2] Siemens, G. 2013. Learning Analytics: The Emergence of a Discipline. American Behavioral Scientist, 57(10): 1380-1400.

[3] Brown M. 2011. Learning Analytics: The Coming Third Wave. EDU_CAUSE Learning Initiative Brief, (4): 1-4.

[4] X. F Wei and L. Q Song. 2013. Learning Analytics: A Better Understanding of the Individualized Learning Process of Students-Interview with Professor George Siemens, An Expert in Learning Analytics Research. China Educational Technology, 2013(9):1-4.

[5] Elias, T. Learning Analytics: The Definitions, the Processes, and the Potential. Retrieved 2011-12-26. https://learning analytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf.

[6] Leah P. Macfadyen and Shane Dawson. 2010. Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2):588-599.

[7] Siemens G, Gasevic D, Haythornthwaite C, Dawson S, Shum S B, & Ferguson R. 2011. Open Learning Analytics: an integrated & modularized platform. Society for Learning Analytics Research, 2011(8):2-18.

[8] Verbert K, Duval E, Klerkx J, *et al.* 2013. Learning Analytics Dashboard Applications. American Behavioral Scientist, 57(10):1500-1509.

[9] Ifenthaler D and Widanapathirana C. 2014. Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines. Technology Knowledge & Learning, 19(1-2):221-240.

[10] Y. Y Li, S. X Ma, and R. H Huang. 2012. Learning Analytics: Serving the Learning Process Design and Optimization. Open Education Research, 2012(05):20-26.

[11] Z. Z Zhu and M. D Shen. 2013. Learning analytics: the scientific power of smart education. E-education Research, 2013(5):5-19.

[12] X. Q Gu, J. L Zhang, and H. Y Cai. 2012. Learning Analytics: The emerging data technology. Journal of Distance Education, 2012(2):18-25.

[13] Q li and T Wang. 2012. A Review of the Status Quo of Research and Application of Learning analytics Technology. China Educational Technology, 2012(08):129-133.

[14] F. Q Li and W. Z Qian. 2012. Learning Analytics: a New Field of Research and Practice of Teaching Information. Modern Educational Technology, 22(07):5-10.

[15] S. P Wei. 2013. Learning Analytics: Mining the Value of Education Data under the Big Data Era. Modern Educational Technology, 23(02):5-11.

[16] X. H Yu and X. Q Gu. 2013. Learning Activity Streams: A Behavior Model for Learning Analytics. Journal of Distance Education, 31(04):20-28.

[17] M Brown. 2012. Learning Analytics Moving from Concept to Practice. EDUCAUSE Learning Initiative: EDUCAUSE. 2012:1-5.

[18] Buniyamin N. 2014. An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. Engineering Education. IEEE.

[19] Verbert K, Govaerts S, Duval E, *et al.* 2014. Learning dashboards: an overview and future research opportunities. Personal and Ubiquitous Computing, 18(6):1499-1514.

[20] Bukralia R, Deokar A. V, and Sarnikar S. 2015. Using Academic Analytics to Predict Dropout Risk in E-Learning Courses. Springer International Publishing.

[21] Berry L. J. 2017. Using Learning Analytics to Predict Academic Success in Online and Face-to-Face Learning Environments. US: Boise State University, 1-129.

[22] F. T Wu and Z. J Mou. 2016. The Design Research of Learning Outcomes Prediction Based on the Model of Personalized Behavior Analysis for Learners. China Educational Technology, 2016(01):41-48.

[23] S Li, Z. X Wang, C Yu, and Y Zong. 2016. Mining LMS Data for Behavioral Engagement Indicators in Online Learning Environments. Open Education Research, 22(02):77-88.

[24] C. K He and M Wu. 2016. Analysis and Prediction of Learning Behavior of Educational Big Data on edX Platform. Distance Education in China, 2016(06):54-59.

[25] Z. J Mou and F. T Wu. 2017. The Exploration of Learning Outcome Prediction Indicators and Analysis of Learning Group Characteristics for MOOC. Modern Distance Education Research, 2017(03):58-66+93.

[26] J. Y Jia and Y. Y Yu. 2017. Analyzing Learning Activity Index and Online Learning Activity Index. Distance Education in China, 2017(04):15-22+56+79.

[27] J. Y Wang, Y. F Zhang, and Z Xu. 2020. Student performance prediction based on feature selection optimization. Think Tank Era, 2020(01):124-125.

[28] X. L Song, X Qi, and B Wang. 2020. Research on Informationized Prediction of Student Performance Based on Machine Learning. Computer Programming Skills & Maintenance, 2020(04):110-112.

[29] J Ma, W Zhao, J Zhang, and Y Zhao. 2014. An Empirical Study of Learning and Analysis Techniques to Build Predictive Models. Modern Educational Technology, 24(11):30-38.

[30] Z. X Jiang, Y Zhang, and X. M Li. 2015. Learning Behavior Analysis and Prediction based in MOOC Data. Journal of Computer Research and Development, 52(03):614-628.

[31] X. Z Wang, C. Q Huang, J Zhu, and X. Q Xu. 2018. Study on Learning Condition Prediction Based on Big Data Analysis in Cloud Learning Space. E-education Research, 39(10):60-67.

[32] Balakrishnan & Eecs G. 2013. Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models.

[33] Tsiakmaki M, Kostopoulos G, Kotsiantis S, *et al.* 2019. Implementing AutoML in Educational Data Mining for Prediction Tasks. Applied Sciences, 10(1):90.

[34] Ken A. Fujimoto & Sabina R Neugebauer. 2020. A General Bayesian Multidimensional Item Response Theory Model for Small and Large Samples. Educational and Psychological Measurement, 80(4).

[35] Kaser T, Klingler S, Schwing A G, & Gross M. 2017. Dynamic bayesian networks for student modeling. IEEE Transactions on Learning Technologies, 10(4): 1-1.

[36] Huifeng Mu. 2017. Research on Item Banking of International Academic English Proficiency Evaluating System. Technology Enhanced Foreign Language Education, 2017(03):9-14+35.

[37] Y Wang, S. J Chang, X. J Han, and H Lu. 2019. The Development and Validity Check of the Item Bank Based on Item Response Theory——Taking the Public Course of "Modern Educational Technology" as an Example. Modern Educational Technology, 29(10):41-47.

[38] Laurillard Diana. 2002. Rethinking University Teaching: A Conversational Framework for the Effective Use of Learning Technologies.

[39] H. T Sun. 2012. Case Study of Interactive Analysis of Distance Teaching from the Perspective of Learning analytics. China Educational Technology, 2012(11):40-46.

[40] Boyer S and Veeramachaneni K. 2015. Transfer Learning for Predictive Models in Massive Open Online Courses. Lecture Notes in Artificial Intelligence, 54-63.

[41] Z Li, D. D Zhou, and Y Wang. 2019. Research of Educational Knowledge Graph from the Perspective of "Artificial Intelligence Plus": Connotation, Technical Framework and Application. Journal of Distance Education, 37(04):42-53.

[42] Q Liu, Y Li, H Duan, Y Liu, and Z. G Qin. 2016. Knowledge Graph Construction Techniques. Journal of Computer Research and Development, 53(03):582-600.

[43] Z. L Xu, Y. P Sheng, L. R He, and Y. F Wang. 2016. Review on Knowledge Graph Techniques. Journal of University of Electronic Science and Technology of China, 45(04):589-606.

[44] G. L Qi, H Gao, and T. X Wu. 2017. The Research Advances of Knowledge Graph[J]. Technology Intelligence Engineering, 3(01):4-25.

[45] Pardos Z. A and Heffernan N. T. 2010. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. International Conference on User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg.

[46] Lara J A, Lizcano D, Martínez, María A, *et al.* 2014. A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. Computers & Education, 72, 23-36.

[47] Z. J Chen and X. L Zhu. 2017. Research on Prediction Model of Online Learners' Academic Achievement Based on Educational Data Mining. China Educational Technology, 2017(12):75-81+89.

[48] F. T Wu and H Tian. 2019. Mining Meaningful Features of Learning Behavior: Research on Prediction Framework of Learning Outcomes. Open Education Research, 25(06):75-82.

[49] Thomas D, Brown J. S. 2009. Learning for a World of Constant Change: Homo Sapiens, Homo Faber & Homo Ludens Revisited. In Proceedings of the University Research for Innovation: Proc. 7th Glion Colloquium (Montreux, Switzerland).

[50] Bainbridge J, Melitski J, Zahradnik A, *et al.* 2015. Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education. Journal of Public Affairs Education, 21(2):247-262.

[51] Alexander P. A and Judy J. E. 1988. The Interaction of Domain-Specific and Strategic Knowledge in Academic Performance. Review of Educational Research, 1988, 58(4):375-404.

[52] Snow C. E and Biancarosa G. 2003. Adolescent Literacy and the Achievement Gap: What Do We Know and Where Do We Go From Here?. NewYork, NY: Carnegie Corporation.

[53] Shum S. B and Crick R. D. 2012. Learning dispositions and transferable competencies: pedagogy, modelling and learning analytics. International Conferenceon Learning Analytics & Knowledge, 92-101.

[54] Moreno-Marcos P. M, Pong T. C, Munoz-Merino P. J, & Kloos C. D. 2020. Analysis of the factors influencing learners' performance prediction with learning analytics. IEEE Access, pp(99), 1-1.

[55] Umer R, Mathrani A, Susnjak T, & Lim S. 2019. Mining Activity Log Data to Predict Student's Outcome in a Course. In Proceedings of the 2019 International Conference on Big Data and Education, London, UK, 27–29 March 2019; pp. 52–58.

[56] Phauk Sokkhey and Takeo Okazaki. 2020. Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques. International Journal of Advanced Computer Science and Applications (IJACSA), 11(7):23-32.

[57] Chang K. M, Beck J. E, Mostow J, *et al.* 2006. A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. International Conference on Intelligent Tutoring Systems. Springer Berlin Heidelberg.

[58] Lacave C and Molina A. I. 2018. Using bayesian networks for learning analytics in engineering education: A case study on computer science dropout at UCLM. International Journal of Engineering Education, 34(3):879-894.

[59] Ming N. C, Ming V. L. 2012. Predicting Student Outcomes from Unstructured Data. In proceedings of the 2nd Workshop on Personalization Approaches for Learning Environments.

[60] Heise N, Meyer C. A, Garbe B. A, Hall H. A, & Clapp T. R. 2020. Table quizzes as an assessment tool in the gross anatomy laboratory. Journal of Medical Education & Curricular Development, 7, 238212052094182.