Integrated Computational Learning Algorithm for Undergraduates' Academic Early Warning

Xiaoguang Sheng1,*, Qirui Yang1,2, Yu Han1 and Ying Wang1
1University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing, P.R.China 100049
2Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China

*Corresponding author email: shengxiaoguang@ucas.ac.cn

Abstract. With the development of education big data, it is helpful for education managers to use machine learning method to predict students' academic warning status, in order to ensure that students pass the course and graduate on time. Although significant progress on predicting academic warning status has been achieved in recent years, existing methods are often lack of generalization and are difficult to be applied in real scenarios. In this study, we divide students' scores into five coarse-grained dimensions: mathematics, foreign language, humanities, major and total score, and innovatively use machine learning ensemble model to predict college students' academic status. Experiments show that the coarse-grained dimension division is not only conducive to the generalization of the model, but also improves the accuracy of prediction academic warning status by 9.708%. Meanwhile, experimental results show that the multi machine learning model ensemble technique can effectively improve the prediction accuracy of college students' academic warning status. With only a small number of samples, the accuracy of predicting students' academic warning status by three months in advance reaches 97.52%, and that by the previous semester reaches above 97.93%.

1. Introduction
With the reform of higher education going deeper and deeper, it is more and more significant to construct a new student management system for colleges and universities. Improving students' academic performance and increasing their graduation rate is an important part of measuring the success of education reform. Therefore, it is urgent to develop an accurate, comprehensive and easily-generalizing method to predict students' academic status in advance and to provide targeted academic assistance timely. This can ensure the normal studying progress for student to the greatest extent, and also reduce the waste of educational resources.

In recent years, with the upsurge of data science, machine learning has gradually become an important method to analyze students' academic problems. For machine-learning-based prediction models, data is a paramount factor that affects the performance of the model. Most previous works adopted students' academic scores as main research target, and utilized the scores for training. Because the academic scores are the most intuitive representation of students' academic status. In addition, most of the existing methods treated all the scores equally without considering the difference between different disciplines. However, several psychological studies have shown that there are obvious discrepancies in different subjects. Therefore, based on existing methods, we take a step further and take into account the differences between different disciplines and the distribution of disciplines in different universities. Based on the relevant pedagogy and psychology literatures, we propose a new coarse-
grained taxonomy with five dimensions for multidisciplinary from the perspective of the characteristics of students' abilities. According to the research of pedagogy literature, we divide students' academic scores into five coarse-grained dimensions: mathematics, foreign language, humanities, major and total score. Moreover, effort on exploring various models is also a research hotspot. Several earlier innovative works have been put forward, but they usually require abundant data for training. As a sample-imbalance problem, academic warning is similar to financial fraud, cancer diagnosis and other issues. It is difficult to measure the performance of the model only depending on the accuracy. In predicting academic warning status, we should try the best to reduce the omission for positive samples. Therefore, this study considers six evaluation metrics: F1 score, Kappa, accuracy, omission rate, misjudgment rate and misjudgment correction rate. Our experiments show that a single machine learning model cannot balance these metrics well. Therefore, we propose a hybrid decision-making method of multiple machine learning ensemble model to maximize the prediction of students who should receive academic warning and improve other evaluation metrics. Thus, considering the above three aspects, the contributions of this paper are as follows:

1. We propose a new taxonomy to divide the students’ academic achievement into five dimensions, which is more effective and robust.
2. We propose a multi machine learning hybrid decision-making model to balance each evaluation metric and optimize the prediction accuracy.
3. With only a small sample of data, the accuracy of our method on predicting academic warning status reaches more than 98.5% by three months in advance.

2. Related Work

Existing research shows that education administrators give early warning to students with academic risks, which can greatly reduce the rate of student academic failure. Data processing, visualization, educational data mining, technical learning enhancement, and predictive models are tools for learning analysis, designed to provide meaningful actions for educators. Existing student academic research usually explores academic warnings based on inter-disciplinary relationships under the credit system, which are called police afterwards. The existing research work mainly explores the dimensions that affect students’ academic performance, such as using parents’ education, family income and other factors to predict student performance. XU uses students’ online forums to predict individual subject scores. SUHAIL K uses the behavior records of the online learning system to predict the test scores of the subject. Zhou et al. uses the behavior records of the online learning system to predict the test scores of the subject. Zheng et al. predict the individual subject scores of students based on the completion of students' homework, class attendance, and quizzes. These tasks are often confined to the students' single subject scores, and it is difficult to provide educators with more effective suggestions.

As education informatization continues to deepen, student data is easier to collect. However, university data mainly presents three characteristics, "fragmentation", "sustainability" and "multi-dimensional". Student performance is the easiest to collect, which is also the data basis for most academic early warning predictions. However, there is a large amount of unprocessed data in the existing data, which makes the data value density low, and it is impossible to dig out the deep meaning behind it. On the other hand, although the performance of a course is the most important influencing characteristic of academic early warning, there are huge differences between different majors and different subjects. Previous methods vaguely treat student achievement as the same. Therefore, we need to consider the similarities and differences between disciplines, integrate the collected multi-dimensional data, and deeply analyze the living habits and learning situation of students. In this work, we systematically divide student performance into five dimensions according to subject attributes, namely mathematics, foreign language, humanities, professional courses, and total score, making our predictions more popular and practical.

Although the first step in increasing the likelihood of academic success is to identify at-risk students as early as possible. Researchers have also noticed the various analytical methods used to predict such students. Existing predictive early warning models for students’ academic studies include: (1) predicting the courses that will fail; (2) predicting students who are delayed or unable to
complete their studies on time\textsuperscript{[15,16]}, (3) predicting which students will receive academic warnings\textsuperscript{[17]}, (4) predicting which factors affect students' academic performance\textsuperscript{[15]}. This article will focus on (2) and (3) and explore (4) which courses are the main factors that affect students' failure to complete their studies on time in the process of research (2) and (3). For the prediction mechanism of the above research, we explored a variety of machine learning models, such as Random Forest\textsuperscript{[18]}, Logistic Regression\textsuperscript{[19]}, Support Vector Machine (SVM)\textsuperscript{[20]}, Decision Tree\textsuperscript{[21]}, Multiple Linear Regression\textsuperscript{[22]}, Multi-layer Perception\textsuperscript{[23]}, Naive Classifier\textsuperscript{[24]}, K-Nearest Neighbor\textsuperscript{[25]}, etc.

3. Method

In this thesis, in order to achieve a high-precision, strong generalization, and effective academic early warning method, the system we propose contains two strategies: a dimensional division method and a multi-machine learning integration method. In this work, unlike the previous fine-grained training vectors that lack generalization, we divide the students' performance in coarse-grained dimensions. According to the existing pedagogical research\textsuperscript{[21]} students' learning ability is mainly reflected in the four aspects of mathematics, foreign languages, humanities, and majors. Therefore, we need to classify students' grades according to course attributes. We first categorize the elective courses of students into four dimensions according to the starting school. In order to balance the influence of each dimension, we weight the student's four types of course results separately, and combine the total weighted results as the fifth coarse-grained. Finally, we normalize the students' scores in the above five rough dimensions in the major/discipline of this grade to alleviate the negative impact of subject differences.

In order to describe our model, we firstly define a set of training vectors \(X_{train} = \{x_1, x_2, ..., x_n\}\), the corresponding label \(Y_{train} = \{0,1\}\), the prediction label indicates whether the student is "received academic warning" ( If and only if the student receives an academic warning, the label is 1). For ease of description, we define the learner as \(h_i\) and discuss eight machine learning methods (KNeighborsClassifier\textsuperscript{[25]}, LogisticRegression\textsuperscript{[22]}, SGDClassifier\textsuperscript{[26]}, LinearSVC, GaussianNB\textsuperscript{[27]}, BernoulliNB, DecisionTreeClassifier\textsuperscript{[21]}, and MLPClassifier\textsuperscript{[23]}). We express the prediction output of \(h_i\) on the training vector \(X_{train}\) as an N-dimensional vector \(\{h_i^1(x); h_i^2(x); h_i^3(x);...; h_i^N(x)\}\), where \(h_i^j(x)\) is the output of \(h_i\) on the category label \(Y_{train}\). In order to describe the classification diversity of classifiers, we define the ensemble learner as \(f(x)\), so the square error of the individual learner and the ensemble learner is:

\[
E(h_i|x) = (f(x) - h_i(x))^2 \tag{1}
\]

\[
E(H|x) = (f(x) - H(x))^2 \tag{2}
\]

\[
\overline{E}(h|x) = \sum_{i=1}^{T} w_i \cdot E(h_i|x) \text{ represents the weighted average of individual learner errors, then:}
\overline{A}(h|x) = \sum_{i=1}^{T} w_i \cdot E(h_i|x) - E(H|x) = \overline{E}(h|x) - E(H|x) \tag{3}
\]

It can be seen from the above formula that when the accuracy of the individual learner is higher and the diversity is greater, the integrated learning method will greatly improve the performance of the model. At the same time, through experimental observation, we found that LinearSVC's classification prediction performance is excellent, but there are many early warning omissions. Unlike the LinearSVC method, the GaussianNB model has fewer early warning misses. Therefore, in this work, we will use the integrated learning method to predict the academic early warning of college students. In order to select a suitable individual learner, we will consider the diversity measure of individual classifiers. A represents the number of samples predicted by both \(h_i\) and \(h_j\) as positive; b represents the number of samples predicted by \(h_i\) as positive and \(h_j\) as negative; c represents the number of samples predicted by \(h_i\) as negative and \(h_j\) as predicted as positive; d represents \(h_i\) prediction number of samples predicted to be negative for the negative class \(h_j\), where \(a+b+c+d=\text{total sample m}\). Therefore, we will use the inconsistency measure \(dis_{ij}\), correlation coefficient \(\rho_{ij}\), Q-statistic \(Q_{ij}\), k-statistic \(k_{ij}\) to measure each body Diversity difference before learning.
\[ d_{i,j} = \frac{b+c}{m} \]  \hspace{1cm} (4)

\[ \rho_{i,j} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} \]  \hspace{1cm} (5)

\[ Q_{i,j} = \frac{ad-bc}{ad-bc} \]  \hspace{1cm} (6)

\[ k_{i,j} = \frac{p_1-p_2}{1-p_2} \]  \hspace{1cm} (7)

Where \( p_1 \) is the probability that the two classifiers agree; \( p_2 \) is the probability that the two classifiers agree by chance, which can be expressed as follows:

\[ p_1 = \frac{a+b}{m} \]  \hspace{1cm} (8)

\[ p_2 = \frac{(a+b)(a+c)+(c+d)(b+d)}{m^2} \]  \hspace{1cm} (9)

We integrate certain individual learner models, and we limit the output value types of individual learners. Different types of individual learners may produce different types of \( h^i_f(x) \). In this work, we use the class probability output method to vote. The class probability \( h^i_f(x) \in \{0,1\} \) is relative to an estimate of the posterior probability \( P(Y^i_f|x) \).

![Integrated learning example](image)

**Figure 1.** Integrated learning example.

### 4. Experiment

#### 4.1. Description of the Data

We collect 2369 students' data from some university from 2014 to 2019, including 119699 course grades records, which is sampled from the university educational administration. Raw data is consisted of grades data, class-picking time, submitting assignment data in course website and school precaution data. The grades data contains the information of class-picking time, final grades, whether do a make-up test and so on. Main dimension of the grades data is showed in Table 1.

| majority class class hour credit course's property grades is degree course have make-up test retake school term course's type |
|-------------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| major class class hour credit course's property grades is degree course have make-up test retake school term course's type |

For the convenience of making school precaution for students, we divide the students’ grades by time series. To be specific, choose the end of term as time frame.

#### 4.2. Data Preprocessing

In term of the grades data preprocessing, we focus on the distributed interval of student’s grades and their rank among the same major and grade. First, we digitize the descriptive evaluation given by some course for convenience. The transform rule is showed is table 2.
After digitizing all course grades, we process the data from two aspects, one is distributed interval of student’s grades, and the other is the students' grades and their rank among the same major and grade.

In the part of distributed interval, we divide data by terms, count the course number in every score interval, including the number of course that students get scores which is 0, 0 to 10 points, 10 to 20 points, 20 to 30 points, 30 to 40 points, 40 to 50 points, 50 to 60 points, 60 to 70 points, 70 to 80 points, 80 to 90 points and 90 to 100 points. In addition, we count the number of course that students take a make-up test and retake.

In the part of the rank among the same major and grade, we first divide course into mathematics & physics, foreign language, human culture and subject according to the college which provides the course. "human culture" mainly contains the general elective courses of liberal arts. "mathematics & physics " mainly contains the public elective courses of mathematics, physics which attend to be of science. "foreign language" contains all foreign language courses, and "subject" contains students' courses in their majority. Secondly, after dividing grades by terms, we calculate students’ credit weighted average grades in these four types of courses and all the courses. Finally, we rank students’ score in among the same major and grade, and normalization from 0 to 1. After processing, the dimension is showed in Table 3.

**Table 2.** Digitize descriptive evaluation.

|                | Unqualified: 0 | Failed: 0 | do not pass: 0 |
|----------------|----------------|-----------|----------------|
| Pass: 85       | Qualified: 60  | Meet a Minimum Standard: 60 |
| Medium: 66     | Good: 78       | Excellent: 90 |

4.3. Dimension Choices

To check the effectiveness for our method to choose the dimension. We will compare it with the method of dimension division in [5]. To compare them in experiment, we use three common machine learning classification algorithms to test, and evaluate the result by f1-score. It is known from figure 2, different dimension choices of training vector make the performance a significant difference. Our method which is "coarse-grained" have a better performance than "fine-grained" in different term and model in most instances.

| rank of first X terms of all courses | rank of first X terms of "Mathematics & physics " | rank of first X terms of "foreign language" | rank of first X terms of "human culture" | number of courses with 0 point | number of courses with 0 to 10 point | number of courses with 10 to 20 point | number of courses with 20 to 30 point |
|-------------------------------------|-----------------------------------------------|------------------------------------------|---------------------------------------|---------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
| number of courses with 30 to 40 point | number of courses with 40 to 50 point | number of courses with 50 to 60 point | number of courses with 60 to 70 point | number of courses with 70 to 80 point | number of courses with 80 to 90 point | number of courses with 90 to 100 point | number of courses with making up | number of courses with retaking |
| 0.33 | 0.33 | 0.36 | 0.40 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 15 | 21 | 8 | 0 | 0 |
| 0.64 | 0.64 | 0.98 | 0.98 | 0.48 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 4 | 25 | 33 | 1 | 0 |
| 0.16 | 0.10 | 0.36 | 0.22 | 0.42 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 17 | 17 | 22 | 14 | 4 | 2 |

It is noteworthy that for the convenience of generalization and promotion of our model, we use four types of courses’ rank and total rank, which can be referenced by university tended to either science or liberal arts.
4.4. Ensemble Learning

When using machine learning classification algorithms to predict the school precaution, we find the existing model cannot get good performance at all the evaluation index.

Table 4. Performance in different evaluation index.

| Algorithm           | Accuracy | Kappa | False Positive | False Negative |
|---------------------|----------|-------|----------------|----------------|
| LogisticRegression  | 0.937    | 0.582 | 17             | 8              |
| SGDClassifier       | 0.823    | 0.387 | 6              | 64             |
| LinearSVC           | 0.929    | 0.561 | 16             | 12             |
| GaussianNB          | 0.917    | 0.615 | 5              | 28             |
| BernoulliNB         | 0.927    | 0.634 | 7              | 22             |
| KNeighborsClassifier| 0.942    | 0.5546| 21             | 2              |
| DecisionTreeClassifier | 0.929 | 0.611 | 11             | 17             |
| MLPClassifier       | 0.939    | 0.604 | 16             | 8              |

From table 4. We can find that when we use eight common machine learning classification algorithms to predict the school precaution, although all the models perform well in accuracy, they make a significant difference in different evaluation index. In order to balance the recall, precision, kappa and accuracy, we use ensemble learning, which integrates some models have big different, to predict the school precaution.

By the experiment, we choose GaussianNB, BernoulliNB and KNeighborsClassifier to be our individual learner in the ensemble learning. It is showed in figure 3 that using soft voting which integrates individual learner perform better than individual learner and ensemble learning algorithm Adaboost. Finally, we decide to use soft voting to do ensemble learning, whose experiment result is showed in table 5.

Although our model performs imperfect in the accuracy, it does better in the f1-score and kappa. With more detailed observation, over 95% students we predict to get the school precaution incorrectly actually get the school precaution in the next time frame. In summary, our model is better than existing method when judged from all the evaluation index synthetically.
Figure 3. The performance of ensemble learner and individual learner. In the figure, "vote" means our soft voting algorithm, "gnb" means GaussianNB algorithm, "bnb" means BernoulliNB algorithm, "knn" means KNeighborsClassifier algorithm, "Ada3" means Adaboost algorithm with 3 learner, "Ada4" means Adaboost algorithm with 4 learner, "Ada5" means Adaboost algorithm with 5 learner, "Ada200" means Adaboost algorithm with 200 learner.

Table 5. Integrated learning experiment's result.

| Algorithm          | F1-score | Kappa | Accuracy | False Positive | False Negative |
|--------------------|----------|-------|----------|----------------|----------------|
| GaussianNB         | 0.4539   | 0.5051| 0.9477   | 0              | 18             |
| BernoulliNB        | 0.4811   | 0.5263| 0.9564   | 1              | 14             |
| KNeighborsClassifier | 0.4263   | 0.4542| 0.9765   | 7              | 0              |
| Adaboost3          | 0.1549   | 0.1775| 0.9738   | 9              | 0              |
| Adaboost4          | 0.1549   | 0.1775| 0.9738   | 9              | 0              |
| Ours               | 0.5911   | 0.6306| 0.9680   | 0              | 11             |

Table 6. Hyper-parameters for each learning methods during the experiments.

| Methods            | Parameters                                                                 |
|--------------------|-----------------------------------------------------------------------------|
| GaussianNB         | priors=None, var_smoothing=1e-09                                             |
| BernoulliNB        | alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True                  |
| KNeighborsClassifier | algorithm="auto", leaf_size=30, metric="minkowski", metric_params=None, n_jobs=None, n_neighbors, weights="uniform" |
| Ours               | estimators=[('mnb', GaussianNB(priors=None, var_smoothing=1e-09)), ('bnb', BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)), ('knn', KNeighborsClassifier(algorithm="auto", leaf_size=30, metric="minkowski", metric_params=None, n_jobs=None, n_neighbors=5, p=2, weights="uniform")), flatten_transform=True, n_jobs=None, voting="soft", weights=None] |

5. Conclusion
This article discusses the widespread use of machine learning methods in the context of educational big data, and proposes a generalized dimensionality division method. On this basis, we propose an early warning method for college students based on integrated learning. In this paper, we found that coarse-grained dimension partitioning can significantly improve model performance and at the same time improve model generalization ability. Our integrated learning method can effectively improve the prediction accuracy of college students' academic warning status. All in all, in this research, we propose a student academic early warning system, including the entire process of data collection, analysis and processing, to provide a more comprehensive reference for educators to develop student learning management.

Acknowledgements
This work is supported by the University of Chinese Academy of Sciences (E0E58914).
References

[1] SALOMON G. Communication and Education: Social and Psychological Interactions [J]. 1981.

[2] SHIMIZU Y, UEDA K, CHIBA H, et al. Symmetries and dissociation dynamics of F1score excited states of BF3 [J]. Journal of Chemical Physics, 1997, 107(7): 2419-23.

[3] BALDWIN, ALBERT S. The NF-kappa B and I kappa B proteins: new discoveries and insights [J] Annual Review of Immunology, 1996, 14(1): 649.

[4] LIU, JUNXIU, YAO, et al. Data analytics enhanced component volatility model [J]. 2017.

[5] LYTRAS, MILTIADIS, D., et al. Big Data and Data Analytics Research: From Metaphors to Value Space for Collective Wisdom in Human Decision Making and Smart Machines [J]. 2017.

[6] TOETENEL L, RIENTIES B J B J O E T. Analysing 157 learning designs using learning analytic approaches as a means to evaluate the impact of pedagogical decision making [J]. 2016, n/a-n/a.

[7] GIUNCHIGLIA F, ZENI M, GOBBI E, et al. Mobile social media usage and academic performance [J]. 2018, 82(may): 177-85.

[8] TAI C K, CHUN L F D, LYTRAS M D, et al. Predicting At-risk University Students in a Virtual Learning Environment via a Machine Learning Algorithm [J]. 2018.

[9] XU X, WANG J, PENG H, et al. Prediction of academic performance associated with Internet usage behaviors using machine learning algorithms [J]. 2019.

[10] SUHAIL K, BEHAV Z B J C. Effects of excessive Internet use on undergraduate students in Pakistan [J]. 2006, 9(3): 297-307.

[11] QUAN W, ZHOU Q, ZHONG Y, et al. Predicting At-Risk Students using Campus Meal Consumption Records [J]. 2019, 35(2): 563-71.

[12] HELAL S, LI J, LIU L, et al. Predicting academic performance by considering student heterogeneity [J]. 2018, 161(DEC.): 134-46.

[13] GASEVIC D, DAWSON S, ROGERS T, et al. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success [J]. 2016, 28(JAN.): 68-84.

[14] HU Y H, LO C L, SHIH S P J C I H B. Developing early warning systems to predict students' online learning performance [J]. 2014, 36(469-78).

[15] ALHARBI Z, CORNFORD J, DOLDER L, et al. Using data mining techniques to predict students at risk of poor performance; proceedings of the 2016 SAI Computing Conference (SAI), F, 2016 [C].

[16] LAKKARAJU H, AGUIAR E, SHAN C, et al. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes; proceedings of the Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, F, 2015 [C].

[17] DUARTE R, RAMOS-PIRES A, GÔN?ALVES H J T Q M, et al. Identifying at-risk students in higher education [J]. 2014, 25(7-8).

[18] BREIMAN L J M L. Random forest [J]. 2001, 45(5-32).

[19] ALLISON P D. Logistic Regression Using the SAS System: Theory and Application [J]. 2001,

[20] MOHAMED M A, XIAO W. Q-METRIC BASED SUPPORT VECTOR MACHINE [M]. US. 2008.

[21] QUINLAN J R J M L. Induction on decision tree [J]. 1986.

[22] EBERLY L E. Multiple linear regression [J]. 2007.

[23] IAN, D., LONGSTAFF, et al. A pattern recognition approach to understanding the multi-layer perception [J]. 1987, 5(5): 315-9.

[24] KONONENKO I. Semi-naive bayesian classifier; proceedings of the European Working Session on Learning, F, 1991 [C].

[25] PETERSON L E J S. K-nearest neighbor [J]. 2009, 4(2): 1883.

[26] KABIR F, SIDDIQUE S, KOTWAL M R A, et al. Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier [J]. 2015.

[27] RAO V, TEH Y W. Gaussian process modulated renewal processes [J]. 2011.