Analyzing and Predicting the TEM-4 Performance of English Majors in China

Yao Meng[1], Xiangdong Gu[2], Qing Zhou[1] and Yu Zhong[3]
[1]College of Computer Science, Chongqing University, Chongqing, China
[2]Research Center of Language, Cognition & Language Application, Chongqing University, Chongqing, China
[3]School of Foreign Language and Cultures, Chongqing University, Chongqing, China

Keywords: Education Data Mining, TEM-4, Prediction, Naive Bayes Model, English Majors.

Abstract: Test for English Majors-Band 4 (TEM-4) is a national Test for Chinese English majors in the end of their second year at university. This paper focuses on analysis and prediction of the TEM-4 performance of 77 English majors in a Chinese key university. A rich amount of data was collected including students' demographics and family status, learning related achievement, motivation and learning journals they kept for a year with school's permission and students' willingness. The accuracy of three classification algorithms to predict students' TEM-4 performance were compared and Naive Bayes Classifier is verified to gain the highest accuracy. On predicting whether the students' TEM-4 scores might reach the excellent level, the accuracy of the model is above 90%. On predicting whether the students might pass the exam, the accuracy reaches 98%. One contributing finding of this study is that a richer set of data was collected, and we integrate the data. Another one is that students' written learning journals have been verified in the improvement of the accuracy of the prediction model which hasn't been explored in the previous researches about the test.

1 INTRODUCTION

Test for English Majors-Band 4 (TEM-4) has been carried out in China since 1991 by the National Advisory Commission on Foreign Language Teaching in Higher Education (NACFLT). It is a criterion-referenced test which can only be applied for by English major sophomores. TEM-4 aims to test the comprehensive ability of English majors in China, with very limited testing opportunities and highly authoritative testing result. Each student has only two times to sit for the test, and the test result is used as a prerequisite for graduation by some universities or for job-hunting by some companies. Therefor it is a high-stakes examination in China. If we can predict students' TEM-4 achievement level, teachers can provide timely intervention and guide the students who may fail in TEM-4 before the test as well as give further support to those excellent students to become more professional. The written examination and oral examination of TEM-4 are carried out and scored separately. In addition, TEM-4 score in this paper refers to the score of written examination.

Data mining technology can discover hidden information and patterns from a large number of data, and has been successfully applied in a wide range of fields such as education (Zhou, 2015). Education data mining is increasingly concerned by China and the world. For Chinese English education, there have been abundant research in studying the relationship between TEM-4 scores and students' characteristics. (Wen, 2009) conducted multiple-regression analysis to find out to what extent the performance on the three oral tasks can predict the scores on the written tasks of TEM-4. (Li, 2012) used TEM-4 scores as criteria to test whether self-assessment is a reliable and valid measure in evaluating students' language proficiency, and the result confirmed this hypothesis. (Xi, 2012) found that TEM-4 scores can be predicted by learning motivation and learning effort. (Li, 2013) applied stepwise multiple regression to analyze the differences in language aptitude components between high- and low-proficiency groups. However, previous studies pay more attention to analyzing the relation between TEM-4 performance and students' characteristics

Meng, Y., Gu, X., Zhou, Q. and Zhong, Y.
Analyzing and Predicting the TEM-4 Performance of English Majors in China.
DOI: 10.5220/0006263102560261
In *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017) - Volume 1*, pages 256-261
ISBN: 978-989-758-239-4

rather than the accuracy of prediction. Besides, the data of these research are always of certain types, such as gender, learning strategies and language abilities (grammar, pronunciation, listening, etc.), but lack of integrative consideration of these data.

This paper focuses on using multiple types of data to predict students who might fail the TEM-4 test and those who might get excellent scores. Our objective is to draw teachers' and relevant institution's attention to the students who might be on the risk with English learning and to discover what tools are more efficient to improve the learning experiences of the English students in China and beyond.

2 DATA DESCRIPTION

77 English majors of the same grade from a key university in China participated in this research. Students involved in this research took the TEM-4 test in their 4th semester. The score is conventionally transformed into one of four levels by the following classification: fail 0-59; pass 60-69; good 70-79; excellent 80-100. Among the students, 5 failed (6%) and 16 reached excellence (20%). In this paper, *normal* indicates the test participants except for those who got excellent scores. To explore as many features as possible which may be associated with TEM-4 scores and to establish an accurate prediction model, we collected four types of data from the students, which contained more than 24 features: (1) demographics and family socioeconomic status, (2) study-related records, including pre-university educational background and academic performance of their first four semesters in university, (3) results of achievement goal test (Elliot, 2001), (4) students' written learning journals they kept for a year. These data were collected with students' permission and willingness. Data of demographics, family socioeconomic status, and students' achievement goals were collected from questionnaires. With the teacher's advice and guidance, the students wrote learning journals in Chinese for one year, which are used as part of the features. Table 1 lists all variables except for the learning journals and presents a statistical description of the results. Section 3 will give a description of data pre-processing.

3 METHODOLOGY

This study aims at establishing a model to predict English majors' TEM-4 test result using four types of data mentioned above. Three datasets used to train the model were prepared. *Feature set I* consists of all features mentioned in Section 2. *Feature set II* was a subset selected by performing CfsSubsetEval feature selection algorithm (Hall, 1999), implemented in Weka, on feature set I. Students' written learning journals, as a new feature, was added in *feature set III* based on feature set II. We adopted three classification algorithms: Decision Tree, Logistic Regression and Naive Bayes. Each classification algorithm was then performed on the three feature sets respectively. The samples were divided into training set and test set, and leave-one-out cross-validation was used to verify the validity of predication models.

3.1 Text Feature Extraction

Quite different from English, sentences in Chinese consist of characters, which are written one after another in line with no spaces. Word in sentences consists of several adjacent characters. Such characteristic of Chinese language requires word segmentation to single out words from sentences. As different sentence division may lead to different valid words and meanings, we use an efficient Chinese text segment tool called Jieba to segment students' written journal texts. The Jieba tool is open-source uses a prefix dictionary and dynamic programming to find out the most probable combination base on word frequency. For words not in the dictionary, a HMM-based model is used with the Viterbi Algorithm (Sun, 2014).

By dropping words mentioned by less than 5 students, such as Thailand, 650 words remained from 66,495 word-segment results. The bag-of-words model as shown in Table 2, contains 650 words and their frequency in each student's journals.

3.2 Feature Selection

Feature selection refers to selecting N most effective features from M features ($N \leqslant M$), with the aim of reducing dimensions of the data strands. It is an important way to improve the performance of classification algorithms and a key step for data mining. The present study uses Weka to perform CfsSubsetEval, which can choose subsets of features that are highly correlated with the class but with low intercorrelation. The value of the subset of attributes

Table 1: Descriptive statistics of the variables.

Variables	Mean	SD	Min	Max
Dependent variables				
- TEM-4 Score	72.65	9.15	49.00	92.00
Demographics and family socioeconomic status				
- Age	18.14	0.616	17.00	20.00
- Gender(dummy: 1=male, 2=female)			1	2
- Mother's educational qualification[1]	3.34	1.80	1	6
- Father's educational qualification[2]	3.80	1.86	1	7
- Family Month Income[3]	2.35	0.79	1	5
- Development of hometown province[4]	2.25	0.60	1	3
learning related				
- The onset grades of English learning[5]	2.94	2.01	1.	6
- Type of highschool[6]	2.37	1.09	1.	4
- Choice of admission[7]	1.17	0.37	1	2
- English score of entrance examination	132.43	8.02	112.00	145.00
- Chinese score of entrance examination	117.50	8.02	104.00	145.00
- Math score of entrance examination	123.71	10.63	95.00	144.00
- Vocabulary size test score[8]	8405	1068	6500	11800
- Mean score of professional courses for the 1st semester	87.48	4.72	71.82	96.83
- Mean score of professional courses for the 2nd semester	86.55	6.12	66.98	97.45
- Mean score of professional courses for the 3rd semester	88.23	4.29	77.41	98.08
- Mean score of professional courses for the 4th semester	89.53	3.75	80.19	96.88
- Times of taking part in extracurricular activities	2.35	1.95	0	8
- Times of taking part in English contests	1.79	1.40	0	6
- Proportion of winning awards in English contests	0.46	0.842	0	1
Achievement goals[9]				
- mastery-approach	12.07	1.70	7	15
- performance-approach	10.69	2.13	3	15
- mastery-avoidance	11.57	1.53	8	15
- performance-avoidance	10.04	1.87	6	13

- Educational qualification [1, 2]:
  1=Primary School, 2=Secondary School, 3=Senior high School, 4=Technical Secondary School, 5=Junior College, 6=Bachelor, 7=Master, 8=PhD
- Family Month Income [3] (Yuan):
  1= [0, 2000), 2= [2000, 5000), 3= [5000, 10000), 4= [10000, 15000), 5= [15000, above]
- Development of hometown province [4]:
  1=Underdeveloped, 2= Developing, 3=Developed
- The onset grades of English learning [5]:
  1= before 3rd grade, 2= 3rd grade, 3= 4th grade, 4= 5th grade, 5= 6th grade, 6= 7th grade, 7= 10th grade
- Types of High school[6]:
  1= Foreign Language High School, 2= National or Provincial key high School, 3=Municipal or Country key high school, 4=Ordinary high school
- Choice of admission[7]:
  1=Voluntary choice of the major, 2= to be adjusted to the major
- Vocabulary size test score[8]:
  A test created by Paul Nation, containing 140 multiple-choice items, with 10 items from each 1000 word family level
- Achievement goals[9]:
  Elliot's and Murayama's Achievement Goal Questionnaire (Elliot, 2008). Every achievement goal includes three questions and all items are on a 5-point Likert scale.

Table 2: A sample of bag-of-words model extracted from students' written learning journals.

Student No.	English	Upset	Reading	Effort	Curriculum	Time	Final Examination	Insist on	…
001	6	0	9	3	17	12	6	5	…
002	17	5	0	0	5	8	19	0	…
…	…	…	…	…	…	…	…	…	…

is evaluated by considering the individual predictive ability of each feature and the degree of redundancy between them. More details can be found in (Hall, 1999).

3.3 Prediction Model

Common classification algorithms include Decision Tree, Artificial Neural Network, Support Vector Machine, Bayes and Association Rule. Among them, the Naive Bayes Model is simpler, more efficient and easier to interpret. It stems from classic mathematic theory with a sound mathematic foundation. Therefore, it is popular in the field of data mining. More details about it can be found in (Rish, 2001).

Naive Bayes Model uses a series of feature attributes to describe the sample, and then classifies the test sample based on the knowledge learnt from the training sample. For those to be classified, which class they fall in is decided by probability: they belong to the category where highest probability shows. In classification, for a given sample X, the probability of this sample belongs to category Y is calculated by the following formula (1).

$$P(Y_i|X) = \frac{P(X|Y_i)P(Y_i)}{P(X)} \quad (1)$$

In the present study, X presents students' features we concern about, such as age, family income, and P(X) presents the ratio of each feature's value in all samples. Y represents two classes of TEM-4 result (for scenario I they are successful or failed, while for scenario II they are excellent or normal). Since Naive Bayes assumes the X features are independent, formula (2) can be derived, and samples can be classified based on the conditional probability estimation of each feature in each class obtained from formula (2).

$$P(X|Y_i)P(Y_i) = P(Y_i)\prod_{j=1}^{m} P(X_j|Y_j) \quad (2)$$

4 RESULTS AND DISCUSSION

This study, we addressed two questions: whether students may pass TEM-4 and whether they may achieve excellent scores in TEM-4. To answer the two questions, we compared the classification effect of three common classification algorithms, Naive Bayes, Decision Tree and Logistic Regression, using feature set I, II and III respectively. For the two scenarios, predicting Success/Failure and Excellent/Normal, 18 models were built using three different algorithms and three feature sets. Table 3 only lists the feature set which provides the best prediction performance for a given achievement level and classification algorithm, and Figure 1 demonstrates the value of F-measure of three classification algorithms' best performance among three feature sets. As table 3 shows, the values of F-measure and AUC (area under ROC curve) indicate that the Naive Bayes algorithm should be a better binary classifiers in this research. In this section, three experiments of Naive Bayes Model verified to have the best prediction performance both on classification into Success/Failure and Excellent/Normal will be analyzed in detail.

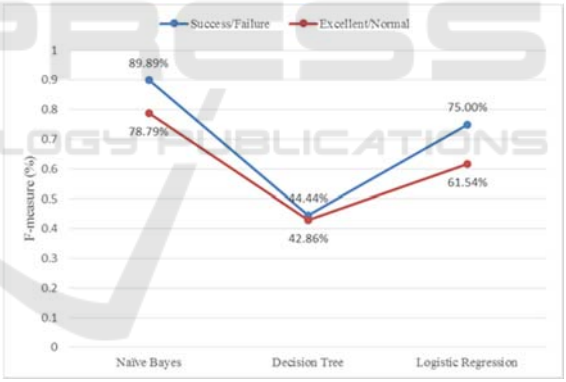

Figure 1: F-measure of three classification algorithms' best performance on three feature sets.

In Experiment 1, feature set I, including all attributes except the students' learning journals, was used to predict whether students may fail in TEM-4 and whether they can get excellent scores. As shown

Table 3: Classification into Success/Failure and Excellent/Normal.

Achievement Level	Classification Algorithm	Data	F-measure	AUC
Success/Failure	Naive Bayes	Feature set III	89.89%	0.80
	Decision Tree(J48)	Feature set II	44.44%	0.50
	Logistic Regression	Feature set III	75.00%	0.72
Excellent/Normal	Naive Bayes	Feature set III	78.79%	0.71
	Decision Tree(J48)	Feature set II	42.86%	0.48
	Logistic Regression	Feature set III	61.54%	0.50

Table 4: Naive Bayes model in predicting students' TEM-4 achievement level.

Experiment	Data	Achievement Level	Precision*	Recall*	Accuracy	F-measure	AUC
NO.1	Feature set I	Success/Failure	25.00%	40.00%	88.31%	30.77%	0.68
		Excellent/Normal	64.29%	56.25%	84.41%	60.00%	0.56
NO.2	Feature set II	Success/Failure	50.00%	60.00%	93.51%	55.55%	0.76
		Excellent/Normal	70.59%	75.00%	88.31%	72.73%	0.63
NO.3	Feature set III	Success/Failure	100.00%	80.00%	98.70%	89.89%	0.80
		Excellent/Normal	76.47%	81.25%	90.91%	78.79%	0.71

*Precision=Positive predictive value, Recall=Sensitive, Accuracy =Ratio of samples correctly classified, F-measure=a measure of a test's accuracy which considers both the precision p and the recall r, AUC=the area under a ROC curve

in Table 4, classifier trained is not satisfactory in prediction. It can only detect 40% and 56% of the students who failed and who obtained excellent scores in TEM-4 respectively.

In Experiment 2, feature set II was used for prediction which includes average scores of professional courses, high school type, onset grade of English learning and students' hometown provinces. As Table 4 shows, the recall of predicting students who might succeed/fail the test improves from 40% to 60%. In terms of prediction on excellent/normal students, the accuracy improves from 56% to 75%.

In Experiment 3, a new prediction model was established by adding some words extracted from the students' learning journals based on Experiment 2. *"Pressure"* was added to predict whether the students would pass TEM-4. *"Final exam"*, *"endeavor" and "modification"* were added to predict whether their TEM-4 scores would be excellent. The precision and recall improve again and reach over 75%, which gain the best predictive effect among the three experiments.

According to the three experiments, that feature set III owns the highest accuracy to predict whether the students may pass TEM-4 and whether their TEM-4 scores can reach the excellent level. The result of Experimental 1 is not ideal probably due to too many features which may lead to over-fitting. The prediction accuracy of the model was improved obviously by eliminating features with low relevance to the class e.g., students' TEM-4 achievement level, and by adding words extracted from the learning journals. Figure 2 and Figure 3 show the improvement of the value of F-measure and AUC during the process of model optimization.

According to the experiment results, the feature students' hometown provinces exists in all Naive Bayes prediction models. One possible explanation is that the students from economically-developed regions are more likely to access better English educational resources. The type of high schools and the onset grade of English learning, which represent the students' English educational background before entering the university, also contribute to the high prediction accuracy. It indicates that the students' English foundation may also be one important factor mediating students' English learning outcome at tertiary education. In predicting of whether the students will pass TEM-4 or whether their TEM-4 test scores will be excellent, professional course average score has not been screened out by the feature selection algorithm. This indicates that the professional course average score is a very contributing feature to predict the students' language

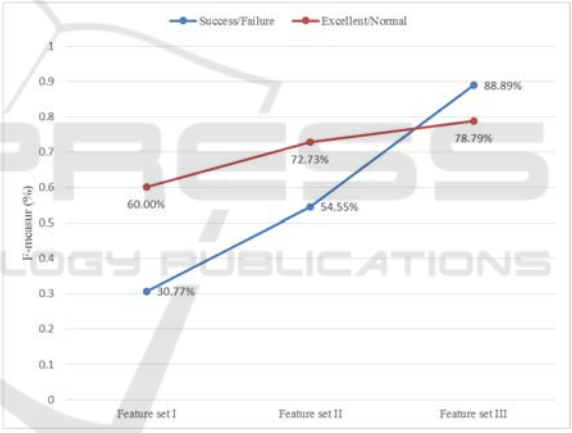

Figure 2: F-measure of Naive Bayes Model.

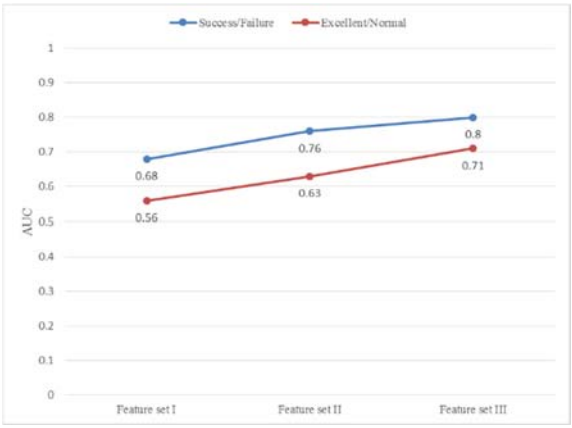

Figure 3: AUC of Naive Bayes Model.

proficiency. In particular, the learning journals also contributes to the prediction accuracy of the students' TEM-4 achievement level.

Based on the research results, English teachers might need to pay more attention to those students who come from ordinary high schools and from underdeveloped provinces. Due to their weak English foundation, they might face more pressure in language learning and are in need of more help, more encouragement and more flexible learning methods fit for them. In addition keeping regular learning journals is highly recommended, as it can help students to develop the habit of self-reflection and help them regulate their English learning. It can also help teacher to keep track on students' learning and emotional fluctuations at students' report in interviews.

5 CONCLUSIONS

In the present study, substantial data from English majors were collected to develop classification models to predict students' TEM-4 performance. Our methodology differs from priori ones in two aspects. Firstly, previous researches are focused on the relationship between students' characteristics and their TEM 4 scores rather than the accuracy of prediction. Our study employed Naive Bayes to predict whether students will pass the TEM-4 and whether they will obtain excellent scores. The accuracy of the established model is up to 98%. Secondly, a richer set of data was collected in the current study, including students' demographics and family socioeconomic status, learning related achievement, motivation and learning journals. What's more, we integrate the data. A contributing finding of this study is that students' written learning journals have been verified in the improvement of the accuracy of the prediction models. Our tentative suggestions for the English teaching are to try to understand the students' learning background and emotional fluctuations, and teaching in accordance with individual's aptitude. The English majors and the like are advised to keep English learning diaries, which help know one's own strengths and weaknesses, and have a positive attitude towards English learning and life.

Due to the imbalanced data in this study between those failed and those reached excellent in TEM-4, the findings of this study needs further validation by a much larger sampling, and should be generalized or used with caution.

ACKNOWLEDGEMENTS

We would very much like to thank Prof. Xiangdong Gu's and Prof. Qing Zhou's team and the external reviewers for their insightful feedback. The current study is supported by "The Short-term International Academic fund" of Chongqing University, Fundamental Research Funds for the Central Universities (Grant No. 106112015CDJSK04JD02) in Chongqing University, National Natural Science Foundation Project of CQ CSTC (Grant No. cstc2016jcyjA0276), Postgraduate Education and Teaching Reform Research Project in Chongqing Province (Grant No. yjg153023), Degree and Postgraduate Education Research (Grant No. C-2015Y0415-128).

REFERENCES

Chuanyi Li. (2012). The criterion-related validation of TEM4 based on test-takers' self-assessment, *2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI)*, Pages: 856 – 858.

Elliot, A. J., and McGregor, H. A. (2001). A 2× 2 achievement goal framework. *Journal of personality and social psychology*, *80*(3), 501.

Elliot, A. J., and Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology*, *100*(3), 613.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. The University of Waikato.

Li, L. (2013). Foreign Language Aptitude Components and Different Levels of Foreign Language Proficiency Among Chinese English Majors. In *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, pp. 179-196. Springer Berlin Heidelberg.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3, No. 22, pp. 41-46. IBM New York.

Sun, J. Y., (2014). "Jieba Chinese text segmentation." https://github.com/fxsjy/jieba, accessed July 21, 2014.

Wen, Q. F., and Wang, L. (2009). Validation of TEM 4-Oral. *Journal of PLA University of Foreign Languages*, 5, 009.

Xi, C. (2012). A Linear Regression Analysis on TEM4. *Journal of Civil Aviation Flight University of China*, *6*, 020.

Zhou Q, Mou C, Yang D. (2015) Research Progress on Educational Data Mining: A Survey. *Ruan Jian Xue Bao/ Journal of Software*, 26(11): 3026-3042(in Chinese).