Automated Essay Scoring Using Multi-classifier Fusion

Li Bin[1] and Yao Jian-Min[2]

[1] Anhui Radio and Television University, Hefei, China
Provincial Key Laboratory of Computer Information, Processing Technology, Soochow University, Suzhou, China
[2] Provincial Key Laboratory of Computer Information, Processing Technology Soochow University, Suzhou, China
liwangmy@163.com

Abstract. The method of multi-classifier fusion was applied to essay scoring. In this paper, each essay was represented by Vector Space Model (VSM). After removing the stopwords, we extracted the features of contents and linguistics from the essays, and each vector was expressed by corresponding weight. Three classical approaches including Document Frequency (DF), Information Gain (IG) and Chi-square Statistic (CHI) were used to select features by some predetermined thresholds. According to the experimental results, we classified the test essay to appropriate category using different classifiers, such as Naive Bayes (NB), K Nearest Neighbors (KNN) and Support Vector Machine (SVM). Finally the ensemble classifier was combined by those component classifiers. After training for multi-classifier fusion technique, the experiments on CET4 essays about same topic in Chinese Learner English Corpus (CLEC) show that precision over 73% was achieved.

Keywords: Automated essay scoring, feature selection, text categorization, multi-classifier fusion.

1 Introduction

While English writing is an essential part of the educational process, many raters find that assessing students' writing is one of the most expensive and time consuming activities for assessment programs. And one of the most difficulties is subjectivity in the grading process. Many researchers claimed that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness [1]. With different people evaluating different essays, reliability becomes a concern in the assessment process. Even with evaluating the same essay, differences in the background training and experience of the raters can lead to subtle but important differences in grading [2].

This issue may be faced through the adoption of automated assessment tools for essays. A system for automated assessment would at least be consistent in the way it scores essays, and enormous cost and time savings could be achieved. Using different methods, related works include PEG, the earliest AES system developed by Page et al [3]. The IEA system, developed in the late 1990s based on latent semantic analysis (LSA) [4]. Burstein et al. developed the E-Rater system [5] based on Microsoft parsing

Y. Wu (Ed.): ICCIC 2011, Part III, CCIS 233, pp. 151–157, 2011.

tools for the ETS. BETSY is developed by Lawrence M. Runder [6]. In China, research on automated essay scoring slightly delayed, the first person who researched on it was Professor Liang Mao-Cheng [7].

This paper presents an approach to essay scoring that builds on the text categorization literature and incorporates multi-classifier fusion technique. Section 2 describes the corpus organization, essay representation and feature selection. Section 3 describes the classifiers for essays scoring. Section 4 describes multi-classifier fusion technique method. Section 5 presents the experiments and the results and discusses the major findings. Section 6 summarizes the conclusions.

2 Courpus of Ganization and Feature Selection of Essays

A. Transform the Essay Scoring to Categorization

Several studies have reported favorably on computer grading of essays. The current systems have returned grades that correlated significantly with human raters. With our approach, we extracted the essays of the same theme (*Global shortage of Fresh Water*) from the *Chinese Learner English Corpus*. According to the different scores of essays, we observe that the sizes of the top and bottom groups were extremely small, so six categories from 230 essays between 7 to 12 points have been extended for our research. Forty-six essays from each group were randomly selected to be used as the trial sample. The remaining 221 essays were used as training corpus. The numbers of each score essays were distributed as follow.

Table 1. The Number Distribution of Essays

Score	7	8	9	10	11	12	Total
Training Set	24	28	40	35	31	26	184
Test Set	4	8	12	11	6	5	46
Total	28	36	52	46	37	31	230

B. Essay Representation and Feature Selection

The essays in their original form are not suitable to learn from. They must be transformed to match the learning algorithm's input format. Because most of the learning algorithms use the attribute-value representation, this means transforming each essay into a vector space.

First of all, essays need to be pre-processed. Then the transformation takes place. The vector space model can be represented as follow:

$$d_j = (w_{1j}, w_{2j}, w_{3j}, \ldots, w_{mj}) \tag{1}$$

Here, d_j denotes the jth essay, and w_{ij} denotes the weight of the ith feature in jth essay. w_{ij} can be deemed to the weight of features. We have extracted the features of contents and linguistics from the essays.

Content-based features include words and phrases and so on. Considering of the themes and contents, the more features which are helpful for assessment will be extracted if they are consistent with requirement of writing. Linguistics-based features include the superficial linguistic features (such as: the number of essays' words, the number of sentences, word lengths, etc. and complex linguistic features (such as: syntactic structure, part of speech, the number of spelling errors, syntax errors, etc), and 25 types of linguistic features have been collected in our research.

The dimensionality of the vector space mentioned above may be very high, which is disadvantageous in machine learning (complexity problem, over learning). And also in order to get more effective features, dimension reduction techniques are necessary. Document Frequency (DF), Information Gain (IG) and Chi-square Statistic (CHI) methods [8] are included in this study, each of which uses a feature-goodness criterion threshold to achieve a desired degree of feature elimination from the full features of an essay.

We have taken Naïve Bayes(NB), K-Nearest Neighbor (KNN)and Support Vector Machine (SVM) as classifications for our research.

3 Multi-classifier Fusion Method

Traditional text classification systems usually use one specific classifier, and it difficult to obtain good classification results if there are a large number of categories. Multi-classifier fusion method used to obtain better results. Its basic assumption is: given a task that requires expert knowledge to be performed, k experts may be better than one if their individual judgments are appropriately combined [9].

We have observed the different advantages and disadvantages of NB, KNN and SVM methods from the descriptions given by Section 3. And also, there are certain complementary relationships between them. So, in this study, we have combined those three classifications to hope them learn from each other.

A. Model of Multi-classifier Fusion

Multi-classifier fusion technique is based on the outputs of component classifiers. The overall probability of producing output y is then the sum over all the processes according to [10]:

$$P(y \mid x, \Theta^0) = \sum_{r=1}^{k} P(r \mid x, \theta_0^0) P(y \mid x, \theta_r^0) \tag{3}$$

Here $\Theta^0 = [\theta_0^0, \theta_1^0, \theta_2^0, ..., \theta_k^0]^t$ represents the vector of all relevant parameters. Figure I shows the basic architecture of an ensemble classifier whose task is to classify a test essay $\boldsymbol{x}$ into one of c_i (i=1, 2, … , M) categories.

The mixture of experts architecture consists of k component classifiers or "experts," each of which has trainable parameters θ_i, i=1,…, k. For each input essay $\boldsymbol{x}$, each component classifier i gives estimates of the category membership $g_{ir} = P(\omega_r \mid x, \theta_i)$.

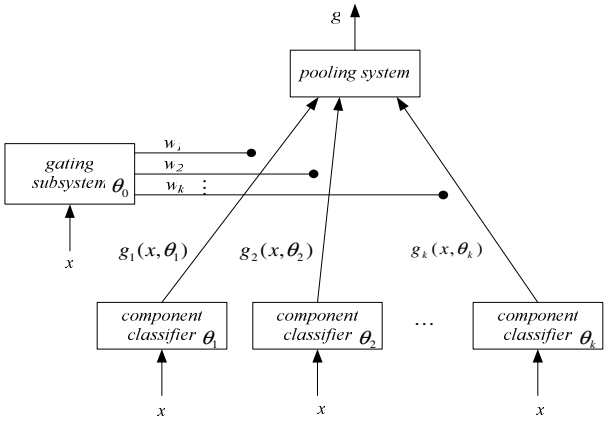

Fig. 1. The Figure of Multi-classifier Fusion

A test essay $\boldsymbol{x}$ is presented to each of the k component classifiers, each of which emits c scalar discriminant values, one for each category. The c discriminant values from component classifier r are grouped and marked $g(x,\theta_r)$ in the figure, with

$$\sum_{j=1}^{c} g_{rj} = 1 \text{ for all } r \tag{4}$$

All discriminant values from component classifier r are multiplied by a scalar weight gating w_r, governed by the *gating subsystem*. The mixture-of-experts architecture is trained so that each component classifier models a corresponding process in the mixture model, and the gating subsystem models the mixing parameters $P(r \mid x, \theta_0^0)$ in Eq.3. The outputs are weighted by the gating subsystem, and pooled for ultimate classification.

We would take an example for describe the details for the procession of multi-classifier fusion technique. For the problem of classification about M classes C_t ($t = 1, 2, \ldots, M$) and K classifies θ_r ($r = 1, 2, \ldots, K$). We assume X_r is measurement vector for classifier θ_r and a test essay x, x will be put into the category whose posteriori probability is maximum. The formula has been expressed as follow:

$$x \to C_j, \text{when } P(C_j \mid X_1, \ldots, X_K) = \max_{t=1}^{K} P(C_t \mid X_1, X_2, \ldots, X_K) \tag{5}$$

We can deduce the formula (6) as follow using (3).

$$x \to C_j, \text{when} \sum_{r=1}^{K} (P(X_r) P(C_j \mid X_r)) = \max_{t=1}^{M} \sum_{r=1}^{K} (P(X_r) P(C_t \mid X_r)) \tag{6}$$

B. The Weight of Component Classifiers

The results of each component classifier will change with the features of the composites, so we need to combine a kind of classifies and adjust the weight of features

in consideration of different essays. Considering the weights of component classifiers increase with precision of them in direct proportion, besides, removing the classifiers which precision is too low [11], the weight of each component classifier can be constructed after the confusion matrix which is calculated as follows:

$$P(X_r) = Acc_i = \sum_{j=1}^{M} r_{jj}^{(i)} \Big/ \sum_{j=1}^{M} \sum_{l=1}^{M} r_{jl}^{(i)} \tag{7}$$

Here, $r_{ji}^{(i)}$ is one of elements of confusion matrix, it denotes the probability of which we divide the essays which in category C_j into category C_l using classifier i. We will get final result of classification if $P(X_r)$ in Eq.7.

4 Results and Analysis

A. Performance Measures

Categorization effectiveness is usually measured in features of the precision and recall. These probabilities may be estimated by features of the global contingency table [9].

Table 2. The Global Contingency table

Category set $C=\{c_1,c_2,c_3,\ldots,c_{\|C\|}\}$		Expert Judgments	
		YES	**NO**
Classifier Judgments	**YES**	$TP=\sum_{i=1}^{\|C\|} TP_i$	$FP=\sum_{i=1}^{\|C\|} FP_i$
	NO	$FN=\sum_{i=1}^{\|C\|} FN_i$	$TN=\sum_{i=1}^{\|C\|} TN_i$

Here, TP_i is the number of test essays correctly classified under c_i; FP_i is the number of test essays classified category c_i while their actual categories are not c_i; FN_i and TN_i are defined accordingly. Estimates of precision and recall may thus be obtained as [12]:

$$\text{Precision: } p=\frac{TP}{TP+FP}=\frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|}(TP_i+FP_i)} \tag{8}$$

$$\text{Recall: } r=\frac{TP}{TP+FN}=\frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|}(TP_i+FN_i)} \tag{9}$$

$$F_1 \text{ Measure: } F_1=2\times p\times r/(p+r) \tag{10}$$

Two methods are used to calculate the precision, recall and F_1 measure called micro-average and macro-average. Micro-average has been used to denote the standard of each point essays, and the macro-average for whole test set.

For evaluating the same essay, it's common that different raters can lead to subtle but important differences in grading. In this study, we hold that the results obtained with one point difference between the true values are also correct.

B. *Primary Results*

After experiment, the KNN classifier got the highest macro precision of 71.74% among other component classifiers, SVM classifier got the higher macro precision of 69.57%, and the macro precision of NB classifier is 63.04%. After in the proper fusion of these classifications, the macro precision increased to 73.91%. The detailed experimental Results of micro-average are given by Table 3. In the Table 3, "Def" denote the data can't be calculated because of none of essay be divided into a category.

C. *Discussion and Analysis*

Automated essay scoring works well. Normally a classifier is doing the job of inferring whether a document is about something or relevant to something. One expects the core of a category to be characterized by a few key concepts, and some larger number of highly associated concepts. The job of feature selection is to find these key concepts and so on. In contrast, in essay scoring, the classifier is trying to determine whether an essay is "good" or not.

Table 3. The Micro-average of Results

Score	Precision				Recall				F_1 Measure			
	NB	*KNN*	*SVM*	*Fusion*	*NB*	*KNN*	*SVM*	*Fusion*	*NB*	*KNN*	*SVM*	*Fusion*
7	*Def*	0.00	*Def*	*Def*	0.00	0.00	0.00	0.25	*Def*	0.00	*Def*	*Def*
8	1.00	1.00	*Def*	0.50	0.75	0.38	1.00	1.00	0.86	0.55	*Def*	*0.67*
9	0.63	0.57	0.67	0.71	0.75	0.92	1.00	1.00	0.68	0.70	0.80	0.80
10	0.73	0.77	*Def*	1.00	0.73	1.00	1.00	1.00	0.73	0.87	*Def*	*1.00*
11	0.50	0.14	*Def*	0.75	0.50	0.40	0.00	0.00	0.50	0.20	*Def*	*0.45*
12	0.20	*Def*	*Def*	1.00	0.20	0.00	0.00	0.00	0.20	*Def*	*Def*	*0.33*

Because we have exacted the essays of same theme; the characters of them are very similar. That causes the result of trial is not very well, although many methods about feature selection have been used for exacted the essays' features. And the feature selection is also the difficulty for essay scoring. Another reason for lower precision is the distributions of the essays are not symmetrical. In the corpus, the numbers of 7, 8, 11 and 12 points compositions are much less; the much more essays have been marked the 9 or 10 points. This led the inadequate training for high and low scores essays model, we found that the test essays with middle points got the much better effect than essays in marginal points.

According the output of experiment, the method of automated essay scoring using text categorization is feasible. But the performance of each component classifier is different, and they are not stable, such as when we take KNN classifier for test, the precision of 10 points and 11 points scores are 77% and 14%. So we combine the component classifiers by the techniques, and then get the final result by fused classifiers. The experimental results show that combination is an effective and stable method to enhance the performance of component classifiers.

5 Conclusion and Future Work

We presented the multi-classifier fusion technique to do essay scoring based on the well developed text categorization literature. Our evaluation of the approach based on words, phrases and lots of linguistic features and six categories is quite good. With fusing the different classifiers, we were able to achieve 73.91% precision.

We emphasize that this is a preliminary investigation. In next phase of our research, we will do our best to take measures to get more effective features with different methods to improve our accuracy. As the initial corpus of scored essays is very small, we will try to get larger corpus. And we will adjust and improve our fusion technique according to the characters of automated essay scoring.

References

[1] Valenti, S., Neri, F., Cucchiarelli, A.: An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education 2, 319–330 (2003)

[2] Blok, H., de Glopper, K.: Large Scale Writing Assessment. In: Verhoeven, L., De Jong, J.H.A.L. (eds.) The construct of language proficiency: Applications of psychological models to language assessment, pp. 101–111. John Benjamins, Amsterdam (1992)

[3] Burstein, J., Leacock, C., Swartz, R.: Automated Evaluation of Essays and Short Answers. In: Danson, M. (ed.) Proceedings of the Sixth International Computer Assisted Assessment Conference. Loughborough University, Loughborough (2001)

[4] Mitchell, T., Russel, T., Broomhead, P., Aldridge, N.: Towards Robust Computerized Marking of Free-text Responses. In: Danson, M. (ed.) Proceedings of the Sixth International Computer Assisted Assessment Conference. Loughboroug University, Loughborouh (2002)

[5] Rudner, L.M., Liang, T.: Automated essay scoring using Bayes' Theorem. The Journal of Technology, Learning and Assessment 1(2), 3–21 (2002)

[6] Valenti, S., Cucchiarelli, A., Panti, M.: Computer Based Assessment Systems Evaluation Via The ISO9126 Quality Model. Journal of Information Technology Education 1(3), 157–175 (2002)

[7] Liang, M-c.: A Model of Automated English Essay Scoring for Chinese Learner . Nan Jing: The PhD Thesis of Nan Jing university (2005)

[8] Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. In: Proceedings 22nd Annual International ACM SI2 GIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM Press, Berkeley (1999)

[9] Fabrizio, S.: Machine Learning in Automated Text Categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)

[10] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn., pp. 475–499. Wiley-Interscience Publication, Hoboken (2000)

[11] Tang, C.S., Jin, Y.: A Multiple Classifiers Integration Method Based on Full Information Matrix. Journal of Software (2003)

[12] Salton, G.: Automatic Text Processing: the Transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co, Boston (1989)