Advances in Social Science, Education and Humanities Research (ASSEHR), volume 189
2018 International Conference on Management and Education, Humanities and Social Sciences (MEHSS 2018)

Ensemble Learning on Scoring Student Essay

Haokun Liu [a, *], Yan Ye [b] and Min Wu [c]
Center of Modern Educational Technology, University of Science and Technology of China, China
[*, a] sa515004@mail.ustc.edu.cn, [b]zyyp@ustc.edu.cn, [c]minwu@ustc.edu.cn

Keywords: Ensemble Learning, Word2vec, Nature Language Processing, Score Essay.

Abstract. Automated essay scoring is becoming more and more concerned by the researchers. In this work, we develop a new way to extract Textual features, which is proved to be valid. First, we calculate the Distributed Representation from the WiKi corpus by the word2vec. Then we calculate the number of words, the number of dictionary,the diversity of words as the textual features by K-means and Distributed Representation. There will be 3*k textual features as the k represents the number of categories. Besides, we calculate the structure features including the length of essay,the number of paragraph, the length of sentence etc. We use several models such as XGBoost, Random Forest, GBDT to train the training set and predict the test set.Finally, We ensemble the prediction of those models as the final prediction.

1. Introduction

With the rise of online education, students' online writing is becoming more and more popular.While the manual scoring is time-consuming and laborious, an efficient automated scoring system is urgently needed. As we know, the evaluation of English composition is mainly based on three aspects: vocabulary, grammar and organizational structure.Vocabulary not only means that the word needs to be spelled correctly but also be required to apply properly.Besides, the diversity of vocabulary will be took into consideration. Grammar means that your sentences have no problem in grammatical requirements.Organizational structure means that the coherence of context and the clarity of the relations between sentences are took into consideration.

Obviously, it is easy to extract structure features such as the length of essay,the number of paragraph,the length of sentence etc. But those features are not good enough to evaluate the essay,as we know those are shallow features.What we need to do to describe the content quality of essay? We use the word2vec to generate distributed representation which represents the word's meaning. While all essay words' distributed representation have been calculated,we use K-means method to cluster the distributed representation and generate the number of words,the number of dictionary,the diversity of words as the textual features. We then calculate the structure features such as the length of essay,the number of paragraph,the length of sentence etc. We use several models such as XGBoost, Random Forest,GBDT to train the training set and predict the test set. And Finally, We ensemble the prediction of those models as the final prediction voted with different weight.

The following is a brief introduction to the related work. The third part describes the word2vec and k-means algorithm,which are used to generate textual features. The fourth part lists the structure features while the fifth part describes the ensemble of base models. The sixth part reports experimental results. At the end of the paper made a summary for this study and outline the future research work.

2. Related Work

As early as 1966, Page showed that an automated "rater" is indistinguishable from human raters (Page, 1966). In the 1990's more systems were developed; the most prominent systems are the Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998), Intellimetric (Elliot, 2001), a new version of the Project Essay Grade (PEG, Page, 1994), and e-rater (Burstein et al., 1998) [1].

Intelligent Essay Assessor(IEA) was developed by Foltz[2], which evaluates the score of the essay according to the semantic information of the essay. In 2002, Rudne and Liang had developed BETSY (Bayesian Essay Test Scoring system)based on Naive Bayes model,which is the only open source system[3].They combined PEG shallow features,LSA latent semantic features and E-rater features into the system.The advantage is that the system is simple and easy to be implemented. But the main default is that the building feature is difficult to meet the feature independence hypothesis of the naive Bayes model and cannot be built in the deep level. In addition to classification and regression methods, Yannakoudakis proposed a Learning to Rank method to learn a scoring model [4]. They ranked essays globally based on essay quality and it is the first time who used the ranking method on scoring essay. By extracting the features of the word, the part of speech tagging features and the syntactic features,they used Rank SVM model to predict essay score. Base on that,He developed a scoring essay model which was based on document list sorting(Listwise Learning to Rank)[5].

In order to evaluate the score of the essay delicately in all aspects, the researchers have put forward more points. Persing and Ng scored essay on the aspect of theme consistency [6] while Persing and other cooperators measured the quality of composition from the point of view of the organization [7]. Neural approaches have also been used for syntactic parsing. In (Vinyals et al., 2015), long short-term memory networks have been used to obtain parse trees by using a sequence-to-sequence model and formulating the parsing task as a sequence generation problem. In 2016, Kaveh Taghipour and Hwee Tou Ng developed A Neural Approach to Automated Essay Scoring [8].

3. Textual Features

As is known to us, the evaluation of English composition is mainly based on three aspects: vocabulary, grammar and organizational structure. So the essay content quality is one of the key factor in scoring essay. In 1986, Hinton proposed Distributed Representation,which designed to describe the similarity between words by measuring the distance(such as cos distance) between words.We use word2vec to generate the Distributed Representation with the WiKi corpus. The Distributed Representation can not only express grammatical and semantic information well, but also reveal a lot of potential language rules. However, the deficiency of the Distributed Representation is that each word only corresponds to a single word vector, which cannot solve the polysemy of a word. In view of these shortcomings,we use the subject-word-vector method, which combines the theme model and the Distributed Representation into together. The "theme" can be understood as a definite word for each word Theme. And the subject information of a word, as well as the context of a word, will also be used to train the Distributed Representation. The method based on the subject-word -vector can be understood that words can be expressed as different vectors under different topics. For example, apple is a fruit under the theme of “food" and is expressed as a technology company under the theme of information technology. Skip-Gram is one of the train model of word2vec. Its goal is to predict its context based on the current word. In this model, each word corresponds to a single distributed representation. Given the word sequences $D=\{w_1,...,w_M\}$, the goal of the Skip-Gram model is to maximize the average log probability:

$$l(D)=\frac{1}{M}\sum_{i=1}^{M}\sum_{-k\le c\le k,c\ne 0}\log^{\Pr(W_{i+c}|W_i)} \tag{1}$$

The k is the size of the sliding window, $\Pr(w_c \mid w_i)$ is calculated from softmax function:

$$\Pr(w_c \mid w_i)=\frac{\exp(w_c * w_i)}{\sum_{w_c \in w}\exp(w_c * w_i)} \tag{2}$$

And our model goal is to train distributed representation and the theme of word independently. Its goal is to maximize the following average log probability:

$$l(D)=\frac{1}{M}\sum_{i=1}^{M}\sum_{-k\le c\le k,c\ne 0}\log^{\Pr(w_{i+c}|w_i)}+\log^{\Pr(w_{i+c}|z_i)} \tag{3}$$

Comparing with the Skip-Gram model which use the current word to predict its context, our model takes the subject information of words and words into the consideration at the same time. The subject of a word is regarded as a pseudo word when it is realized, which can be regarded as a set of semantic information of all words under the subject. And finally, the word's distributed representation on certain theme is become the the connection between the word vector and the subject vector.

Considering that the method of Skip-Gram model can not solve the polysemous phenomenon of a word, we take the theme information into the distributed representation. And the words in each article can be divided into several categories according to their semantic and grammatical meanings. Each category represents a certain semantic information described by the author. Words and word distribution on certain category represents the strength and weakness of the semantic information. Here,we cluster word vectors by K-means[9] method. K-means is a highly efficient and simple clustering method. By changing the cluster center continuously, it will be clustered in the end. After the word vector is clustered according to the K-Means method, the words corresponding to each word vector will be classified to a certain category. Then we calculate the number of words,the number of dictionary,the diversity of words as the textual features. There will be 3*k textual features as the k represents the number of categories.

4. Structure Features

In addition to textual features, the structure features also do help to improve the precision of scoring essay system. We use Stanford Parser to produce the grammatical structure of sentences as well as part-of-speech tagging. There are 24 structure features show in Table 1.

Table 1. Feature set

Structure Feature	Feature Name
S1	The length of essay
S2	The number of the paragraphs
S3	The average length of the paragraphs
S4	The maximum length of the paragraphs
S5	The minimum length of the paragraphs
S6	The median length of the paragraphs
S7	The average length of the sentences
S8	The maximum length of the sentences
S9	The minimum length of the sentences
S10	The median length of the sentences
S11	The standard deviation length of the sentences
S12	The number of unique words in essay
S13	The number of periods
S14	The amount of comma
S15	The average length of words
S16	The median length of words
S17	The standard deviation length of words
S18	The number of nouns
S19	The number of verbs
S20	The number of adjectives
S21	The number of adverbs
S22	The number of prepositions
S23	The number of words whose character length greater than four
S24	The number of misspelled words

5. Ensemble Model

XGBoost is short for "Extreme Gradient Boosting", where the term "Gradient Boosting" is proposed in the paper Greedy Function Approximation: A Gradient Boosting Machine, by Friedman. XGBoost is based on this original model.Random forests or random decision forests[10][11] are an ensemble learning method for classification,regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set [12]. GBDT builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. As we know, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. We use three machine learning model including XGBoost, Random Forest, GBDT model to train and predict the essay score and then ensemble the prediction of those models as the final prediction.

6. Experimental Results

Table 2. model result

model	kappa
XGBoost	0.7937
Random Forest	0.7854
GBDT	0.7623
Ensemble	0.8047

Table 3 lists the data of the model prediction. We use Quadratic Weighted Kappa as the evaluation. The quadratic weighted kappa is calculated as follows. Firstly, and N-by-N histogram matrix O is constructed over the essay ratings, such that Ouija corresponds to the number of essays that received a rating I by Rater A and a rating j by Rater B. Besides, an N-by-N matrix of weights, w, is calculated based on the difference between raters' scores:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{4}$$

Farther more, an N-by-N histogram matrix of expected ratings, E, is calculated, assuming that there is no correlation between rating scores. This is calculated as the outer product between each rater's histogram vector of ratings, normalized such that E and O have the same sum. And from these three matrices, the quadratic weighted kappa is calculated:

$$k = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{5}$$

The Fisher Transformation is approximately a variance-stabilizing transformation and is defined:

$$z = \frac{1}{2} \ln \frac{1+k}{1-k} \tag{6}$$

Finally, the reverse transformation is applied to get the average kappa value:

$$k = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{7}$$

From the table 3, we could see that Boost performs better than the other two models and GBDT performs worst in those three models. What's more, with the help of ensemble method, which means that we let each model in the ensemble vote with certain weight, like what the bagging method act. The weight is calculated by training set and apply to the test set. The reason why the ensemble method works is that ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single model

would. So ensembles tend to yield better results especially when there is a significant diversity among the models.

7. Conclusion

In this paper, we have proposed an approach based on word2vec and ensemble method to tackle the task of automated essay scoring. We develop a new way to extract textual features, which proved to be valid and effective. Besides, by integrating three different machine learning model (Boost, Random Forest, GBDT), the prediction improve significantly. Our best system performs better than the strong open-source baseline and outperforms the baseline by 0.28% in terms of quadratic weighted Kappa. Furthermore, our approach doesn't use the deep neural network. Rather than RNN or CNN network, our approach save a lot of computing resource. So it is simple and efficient. What's more, by extracting features from the content of the essay, it can do well in avoiding the shortcomings of the non-text features.

References

[1]. Attala, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3).

[2]. Foltz P W, Latham D, Land Auer T K. The intelligent essay assessor: Applications to educational technology [J]. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1999, 1(2).

[3]. Rudner L M, Liang. Automated essay scoring using Bayes' theorem [J]. The Journal of Technology, Learning and Assessment, 2002, 1(2).

[4]. Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 180-189.

[5]. Chen H, He B. Automated Essay Scoring by Maximizing Human-Machine Agreement[C]// EMNLP. 2013:1741-1752.

[6]. Parsing I, Ng V. Modeling Thesis Clarity in Student Essays[C]// ACL. 2013: 260-269.

[7]. Parsing I, Davis A, Ng V. Modeling organization in student essays[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 229-239.

[8]. Kaveh Taghipour, Hewed Too Ngai. Neural Approach to Automated Essay Scoring[C]//The 2016 Conference on Empirical Methods in Natural Language Processing, 2016:1882-1891.

[9]. Kananga T, Mount D M, Netanyahu N S, et al. An efficient k-means clustering algorithm: Analysis and implementation [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(7): 881-892.

[10]. Ho, Tin Kim (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. p. 278–282.

[11]. Ho, Tin Kim (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. Doi:10.1109/34.709601.

[12]. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd Ed.). Springer. ISBN 0-387-95284-5.