The Intelligent Composition Evaluation System
Based on Anomaly Detection

Yue TONG* and Jian-She ZHOU

Capital Normal University, Research Center for Language Intelligence of China, Beijing 100048, China
*tongyue0816@126.com, *zhoujs@cnu.edu.cn

Keywords: Compositions evaluation; Language intelligence; Deep learning; Anomaly detection; Gaussian distribution.

Abstract. Compositions evaluation is one of the main tasks in Chinese teaching, how to evaluate compositions quickly and efficiently is an important assignment of language intelligence research. A method combining artificial intelligence and language intelligence is proposed to solve this problem, constructing a compositions evaluation system based on anomaly detection which belongs to unsupervised learning algorithm, applying Gaussian distribution theory to distinguish these abnormal characteristics from normal ones in a batch of compositions, according to which method could quickly identify anomalies in the compositions lead to low scores and ultimately achieve the goal of evaluating compositions accurately and efficiently.

Deep Integration of Artificial Intelligence and the Education

It is the deep integration of artificial intelligence and education that makes large-scale education digitalization and education network become a reality. Students can learn and practice on the Internet anytime and anywhere, and the job of exercise corrections are automatically completed by computer. Composition evaluating is the mainly and important problem of intelligent evaluation, it is essential of assessing compositions effectively to choose which algorithm to computing composition points and which features as the scoring dimension. Deep learning is the current mainstream artificial intelligence implementation method, which is mainly divided into supervised learning and unsupervised learning, both of which are used in the composition correction algorithm. The anomaly detection proposed here belongs to an unsupervised learning algorithm, it uses the Gaussian distribution to make normal data at the center position and abnormal data at the edge position, and determines the boundary of the abnormal data by setting the threshold. Compared with the supervised learning algorithm, what are the advantages of anomaly detection in compositions evaluation?

Firstly, supervised learning requires a sufficient number of positive examples from which to train features and make predictions for new examples. But the evaluating dimensions of student compositions is so complicated that it is difficult to collect enough positive example types. Anomaly detection which belongs to an unlabeled algorithm does not need to consider the characteristics of positive examples and collect them, Gaussian distribution can be used to detect positive examples. Therefore, even if there are no enough abnormal compositions examples, the abnormal data of compositions can be detected as well.

Secondly, supervised learning algorithm is suitable in the case of a positive example and a negative example with a balanced ratio, and the anomaly detection is run well in the case where the ratio of the positive example to the negative example is unbalanced. The students' writing ability is in a normal distribution trend (Fig. 1), medium-level composition accounts for the vast majority, and the poor compositions and excellent compositions are relatively small. Therefore, the evaluation of student compositions is more suitable for using anomaly detection algorithm.
The Gaussian distribution is a kind of probability distribution which looks very regular as Fig. 1, the probability of occurrence in the middle of a batch of samples is quite high, and the probability is getting smaller toward the two sides. The image appears as a bell, and the larger the number of examples, the smoother the data and the more pronounced the bell shape. The occurrence of moderate scores of compositions has a higher probability, and the one of abnormal composition and excellent composition is relatively small.

Thirdly, supervised learning can only deal with problems that have occurred in the training set. It is unpredictable for new data types, but the problems of student compositions are various so that the problems that have not appeared in the training set may occur at any time, however, the data set cannot contain all the problem examples. The anomaly detection does not consider how many kinds of abnormal problems exist in student compositions. It only calculates the data beyond the normal range. No matter which aspect of the writing caused the compositions looks not well. Even if there is a new abnormal data that is not in the training set, the status range will be detected as long as it is not normal.

Finally, the intelligent composition evaluation system is specially designed for K12 students. K12 students of different grades have different writing abilities. The writing ability of lower grade students is generally lower than that of upper grade students. Therefore, the evaluating standard of different grades are not the same. Supervised learning requires algorithmic model training for different grades of K12 composition data, which algorithm is very complicated and computationally intensive. The Gaussian distribution algorithm is used in the anomaly detection to avoid training for the scoring models of different stages, and only the overall distribution problem is calculated. Each evaluation essay takes the entire composition data of writing level into account, the medium compositions occupy the majority, and the excellent and the abnormal compositions account for a small proportion of the standard on the whole proportion, which guarantees the scientific and fairness on student compositions scores at different stages.

**Application of Anomaly Detection in Composition Evaluation**

Three steps of anomaly detection are required in the essay evaluation. First, choose the right features $x$. The selected features will change significantly when the student's composition is abnormal, which either far beyond or less than the normal range. Second, get the parameters $\mu$ and $\sigma$. $\mu$ is the average of the training examples and $\sigma$ is the variance of the examples and the mean. Third, give a new example and predict the examples. Above all, the features suitable for the anomaly detection algorithm are collected from the student composition. These features include the chapter features, word features, emotional features, and dynamic features in compositions.
Features Refer to Chapters in Compositions

The data set is composed of a large number of $X$ which is set as features of the composition.

$x_1 =$ the position of the composition heading in the word vectors

The prescriptive topic compositions are typical examples for anomaly detection. The composition theme has been clearly defined before writing, therefore, the students’ compositions topics should be similar and the essay topics are similar in word vector distribution. It will be deducted on the composition topic if the threshold epsilon $\varepsilon$ is exceeded. Try different values in the process of setting the threshold $\varepsilon$, or choose a threshold with the largest F-score value, or the threshold that performs well in other respects.

$x_2 =$ the position of the composition topic in the word vectors

The topic of propositional compositions is highly similar. If some compositions topic vectors are not closely to the intensive district composed of other compositions vector and exceed the threshold, they will be judged as off-topic compositions.

$x_3 =$ the amount of paragraphs in the composition

A complete essay has an appropriate proportion in the number of paragraphs. The number of paragraphs also imply the structure logic and hierarchical of the composition to a certain extent. If an 800-word essay has only one paragraph or as many as 20 paragraphs, the threshold will deduct scores about the paragraphs.

$x_4 =$ the amount of words in the composition

The amount of words in the composition is clearly defined, and most of the written texts are in compliance with the regulations. Compositions that exceed the word threshold will be deducted.

Features Refer to Words in Compositions

$x_5 =$ the degree of repeated words

The number of repetitions of the words of the composition after the word segmentation is judged as one of the criteria for judging the richness of the composition words, and if the number of repetitions of the words exceeds the set threshold, the score will be lowered.

$x_6 =$ the degree of different words

The number of different words appearing in the composition is also one of the criteria for judging the richness of the lyrics. The more diverse the words were used, the richer the composition’s expression and the higher the scores were given.

$x_7 =$ the level of the word rank

According to the student dictionary to determine the complexity of the words, using senior words indicates that the student has master the words extraordinary. If a composition exceeds most of other composition on words degree, it will be added scores.

Features Refer to Emotion in Compositions

The psychological development of most K12 students is normal, and only a small number of students have problems in their psychology. Therefore, it is suitable to use an abnormality detection algorithm to screen out the emotional abnormal composition. A number of emotional dimensions that play important roles in the development of students is used to screen students' psychological development.

$x_8 =$ the emotion of identity

$x_9 =$ the emotion of security

$x_{10} =$ the emotion of pleasure

$x_{11} =$ the emotion of caring

Dynamic Features during the Period of Writing

The intelligent composition evaluation can not only score the compositions which have already been written, but also make the performance of the students in the writing process as the evaluation dimensions of the evaluating.

$x_{12} =$ typing speed at the time of writing
The frequency of pause when writing
\( x_{13} \)

How long during of each pause
\( x_{14} \)

How long to finish the writing
\( x_{15} \)

\[ \text{dataset} \{ x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(n)} \} \in \mathbb{R}, m = 30,000,000, n = 10000 \]

Based on the massive data and deep learning algorithm to train the model, 30,000,000 compositions were collected as the data set of the evaluation, and the feature dimensions of the composition correction was set up to 10000 dimensions, feature \( x \) belongs to real number, \( m \) indicates the total number of examples and \( n \) indicates the dimensions of features. The Gaussian distribution was used to model the data of each dimension to find out the outliers. The anomalous characteristics will be much larger or far below than the average. The data set is distributed to the training set, cross-validation set and test set according to the ratio of 6:2:2. All of which are unmarked data. After the model training, the model is tested and evaluated through the cross-validation set and the test set, and the optimization is continuously optimized. After the training set is trained, the model is tested by cross-validation sets and test sets to continuously iterate and optimize it.

**Fit the Parameters \( \mu \) and \( \sigma \)**

Find the appropriate parameters \( \mu \) and \( \sigma \), try to fit the parameters \( \mu_1, \mu_2, \mu_3, \mu_n \), and \( \sigma^2_1, \sigma^2_2, \sigma^2_3, \ldots, \sigma^2_n \). Parameters can be calculated with the following formulas, \( \mu \) can be get from formula (1) and \( \sigma \) can be get from formula (2).

\[
\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \quad (1)
\]

\[
\sigma^2_j = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2 \quad (2)
\]

Formula (1) calculates the average of all examples, formula (2) calculates the variance between the example and the mean. Fig. 2 explains the parameters \( \mu \) and \( \sigma \) specifically, the closer to \( \mu \), the greater the probability, and the farther away from \( \mu \), the smaller the probability. The smaller the \( \sigma \), the closer the sample is to \( \mu \), and the larger \( \sigma \), the farther the sample is from \( \mu \).

**Evaluating the New Examples**

Given a new example and test each dimension of the example to see if it is abnormal. The evaluating formulas are as follows:

\[
p(x) = \prod_{j=1}^{n} p(x_j; \mu_j; \sigma^2_j) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left(- \frac{(x_j - \mu_j)^2}{2\sigma^2_j}\right) \quad (3)
\]

\[
\begin{align*}
y &= 0, \text{ if } p(x) \geq \varepsilon, \\
y &= 1, \text{ if } p(x) < \varepsilon
\end{align*}
\]

The probability \( p(x) \) of whether the example is abnormal is obtained from the formula (3), comparing \( p(x) \) with the threshold \( \varepsilon \) which has been verified previously. If \( p(x) \) is greater than \( \varepsilon \), there is no need to flag, indicating the example is not an abnormal example and conversely, a flag is required to indicate that the example is an abnormal example. A positive example is represented by \( y=1 \) and the negative one is represented by \( y=0 \).

**Conclusion**

Anomaly detection can pick out abnormal compositions in large-scale data without considering the positive examples of the data set when evaluating compositions. All examples in the data set are used as a reference to allocate the proportion of excellent, medium and abnormal compositions. It avoids the complicated operation of setting different scoring standards for K12 students in different
grades, ensures scoring quickly and efficiently and screened anomalous essays accurately, and reduces the complexity of the algorithm while improving the quality and efficiency of compositions evaluation. It is also necessary to point out that anomaly detection is only one of the algorithms in the compositions evaluation system which can't solve every problem, the anomaly detection is suitable for reviewing batch essays and is not suitable for judging a single essay. It can't be evaluated by normal distribution in terms of analyzing and calculating the concentration and uniqueness of the theme, the smooth flow of the text, the fluency of the statements, the use of punctuation, and the implicit rhetorical techniques.

References
[1] D. Erhan. Why does unsupervised pre-training help deep learning [J]. J Mach Learn Res, 2020, 11: 625-660.
[2] G. Hinton, O. Vinyals and J. Dean. Distilling the knowledge in a neural network [J]. arXiv preprint arXiv: 1503.02531, 2015.
[3] J.A. Comenius. Great Didactic [M]. Beijing: people's education Press, 1979.
[4] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods [C]. In Proc. BMVC, 2011.
[5] N. Chomsky. Rules and representations [M]. Oxford: Basil Blackwell, 1980.
[6] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. In Proc. CVPR, 2014.