Universe Scientific Publishing

DOI:10.18686/ahe.v8i6.13545

Japanese Adverb Stylistic Feature Judgment Model Based on SVM Algorithm

Jingxu Cui

Taiyuan Normal University, Taiyuan Shanxi, 030000, China

Abstract: In this paper, a computer algorithm is used to classify the features of Japanese adverbs, and the professional degree and objective degree of adverbs are quantitatively analyzed, which can reveal their stylistic features more accurately. Traditional teaching evaluation methods are based on manual classification, which has many shortcomings in accuracy and objectivity. In this paper, a support vector machine (SVM) -based model is proposed for the feature classification research and evaluation of Japanese adverbs, which significantly improves the proportion of adverbs and the professionalism, objectivity, accuracy and reliability of the text. Thus, we can accurately grasp the distribution and specific stylistic features of Japanese adverbs in different texts. This paper proposes a model based on SVM algorithm to judge the stylistic features of Japanese adverbs, which provides an effective and innovative solution.

Keywords: Japanese adverbs; Support vector machine (SVM); Stylistic features; Machine learning; Characteristic analysis

1. Construction of stylistic feature judgment model for Japanese adverbs

In this paper, Support Vector Machine (SVM) is used as the basic model and multi-feature information is combined to construct a comprehensive evaluation index system. The specific steps are as follows[1]:

In order to accurately observe the overall distribution and individual characteristics of adverbs in various texts, the author extracts all the documents with marked information from the BCCWJ corpus and groups them according to the hierarchical data provided by the BCCWJ stylistic information database. A variable length corpus is usually in chapters and sections, with a maximum of 10,000 characters. In order to avoid the interference of different sample length on the adverb detection number, this paper chooses a fixed length corpus

Data preprocessing: The original data is cleaned, normalized, and feature selection is performed. The data analysis in this paper is divided into two steps. Firstly, the proportion of adverbs in these texts is calculated one by one, and correlation analysis and one-way analysis of variance are carried out to verify whether there are correlation and significant differences in the proportion of adverbs in different groups of texts. Then, according to the NINJAL stylistic index system, the detected numbers of all adverbs are grouped one by one to make a crosstabs and converted into one-dimensional lists. In model training, SVM model was used to learn the training data and optimize the SVM parameters. Through correspondence analysis, the distribution and individual characteristics of adverbs in various stylistic features are discussed.

2. Support vector machine algorithm principle

Support vector machine (SVM) is a supervised learning model based on statistical learning theory (SLT), which is mainly used for classification and regression problems. The core idea is to find the optimal classification hyperplane by maximizing the distance between the decision boundary and the nearest sample point, which can effectively improve the generalization ability of the model[2].

Its main features are global optimization, structural risk minimization and powerful generalization ability. In practical application, the advantages of SVM can be fully utilized to solve various complex classification and regression problems through reasonable selection of kernel function, optimization of parameters, effective training algorithm and model evaluation and selection. The specific methods to solve the problem using support vector machine are as follows:

Firstly, each sample is represented as an eigenvector, then a suitable kernel function or nonlinear transformation method is select-

ed, and then a support vector machine model is established based on the training sample set. Then, the parameters of the hyperplane are determined by solving the optimization problem, and the support vector is determined according to the solution of the optimization problem, that is, the sample located on the interval boundary. Finally, it is used to classify new unknown samples.

3. Feature association analysis

In terms of data processing, after preliminary data cleaning, standardization and normalization, no outliers, blank values and missing values were found, and correlation analysis was carried out on the original data set, and the following result graph was obtained:

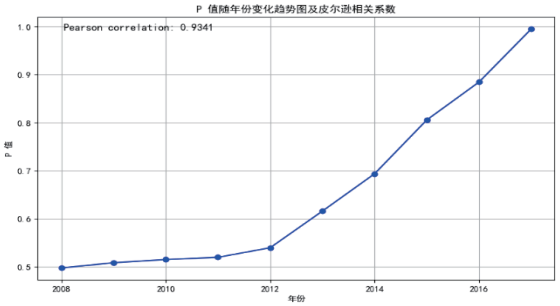

Fig. 1 Pearson correlation coefficient chart

From this figure, it can be clearly seen that the Pearson correlation coefficient is 0.9341, which is very close to 1, indicating that there is a moderate positive correlation between the evaluation indexes of stylistic features of Japanese adverbs and text features, including professionalism, objectivity and hardness. With the continuous growth of professional degree, the P-value of characteristic score is also increasing, and the overall trend is rising.

Figure 2 focuses on evaluating the score changes of professionalism, objectivity, and hardness with moderate positive correlation among text features.

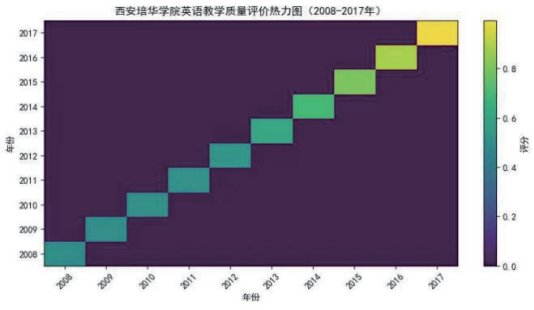

Fig. 2 heat map

First, the raw data is divided into independent variable X and dependent variable y. The data set is then split into a training set and a test set, where the test set accounts for 20% of the total data set. Next, the SVM model is initialized, the kernel function is set to radial basis kernel (rbf), the penalty parameter C is 3.0, and the gamma parameter is scale. Then, the SVM model is trained and the test set is predicted to get the predicted value. Finally, the mean square error (MSE) and root mean square error (RMSE) between the predicted value and the true value are calculated according to the algorithm evaluation index. MSE is the average of the sum of squares of the error between the predicted and actual values. RMSE is the square root of the mean sum of squares of the error between the predicted value and the actual value.

Tab.1 Mean square error (MSE) and root mean square error (RMSE) between predicted value and real value

Mean Squared Error:	0.3838040163960974
Root Mean Squared Error:	0.6195191816207932

As can be seen from the above table, the results obtained by using the SVM model alone are poor, and the accuracy of the results is not very accurate. So we need to find new ways to train the data.

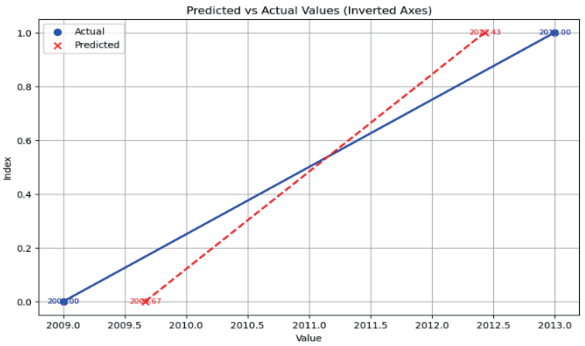

Fig. 3 SVM model prediction vs. actual comparison

Through the training, prediction and evaluation of the SVM model, the comparison between the predicted results and the actual results is visualized. The figure above shows that both lines show an increasing trend over time, but there are some differences in the growth rate of the two lines, the predicted value and the actual value do not completely coincide, and there is a certain deviation between the predicted value and the actual value at some points.

4. Stylistic characteristics of Japanese adverbs

The evaluation value of the future to the index is given by the SVM algorithm under the optimization algorithm.

As predicted in Figure 4, the p-value remains around 0.5. The P-value of the raw data increases significantly until it reaches a peak, then drops slightly, and then shows an upward trend again. The forecast results show that the P-value is around 1.2, indicating that the model

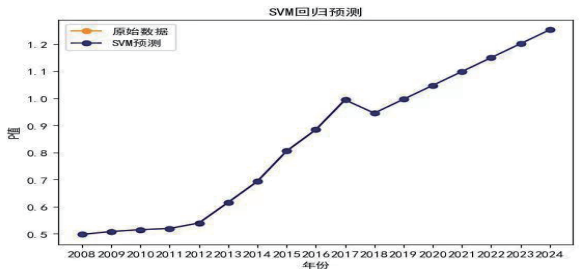

Fig.4 Prediction result graph

expects significant growth in the future. In the long run, the model predicts that the P-value will continue to rise, indicating that there may be potential for continued growth in the future.

Tab.2 Prediction accuracy

MSE:	0.004021313211270185
RSE:	0.06341382507994756

Using the function of Matplotlib library to draw line charts, the following figure is drawn to compare the performance of different models. On the whole, the data points of each line are marked, and it can be seen that all indicators show an upward trend over time. It also shows how the three measures change over time, and all measures improve over the time shown. After comparison, it is found that the overall trend of SVM under cross-validation is linear regression, which is relatively ideal and close to the original data, and the trend is close to the original data.

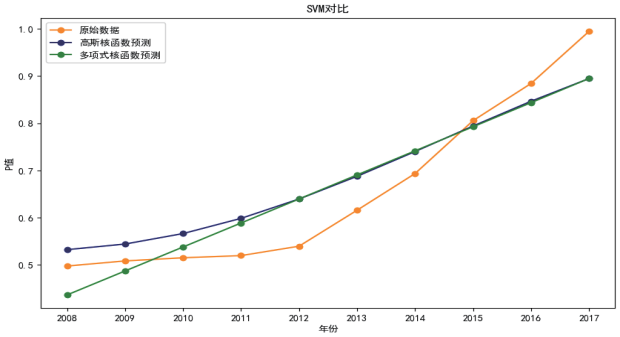

Fig. 5 Comparison of accuracy of each model

This graph shows how the raw data compares to two different support vector machine (SVM) kernel predictions, including Gaussian kernel predictions and polynomial kernel predictions. It is used to evaluate the performance of different SVM kernel functions in time series prediction. Among them, the very subjective adverbs include “ちょっと”, “とても”, “そう”, “やはり” and “ずっと”. Relatively subjective adverbs include “もちろん”, “すぐ”, “ただ”, “もし”, and “しばらく”. Relatively objective adverbs include “まず” “はっきり” “こう” “かなり” and “ついに”. Some non-objective adverbs include “たとえば”, “りゃく”, “とくに”, “ふたたび”, and “より”. Very soft adverbs include “きっと”, “まあ”, “とても”, “ちょっと”, “ずっと”, etc. Soft adverbs include “いったい”, “そう”, “ゆっくり”, “すこし”, and “あまり”. These adverbs are mostly used in spoken language. Stiff adverbs include “まず”, “つまり”, “むしろ”, “ふたたび” and “いかに”. Very blunt adverbs include “きわめて”, “より”, “たとえば” and so on. These adverbs are mostly used in written text. Non-conversational adverbs include “はじめて”, “すぐ”, “そう”, “もう”, and “はっきり”. These adverbs are mostly used in written language. Relatively conversational adverbs include “とくに”, “つまり”, “ちょっと”, “きっと” and “よく”. Examples of very conversational adverbs include “とても”, “いちばん”, “まあ”, “あまり” and “かならず”. More conversational adverbs are used in spoken language.

5. Conclusion

Based on SVM algorithm, this paper investigates the stylistic features of Japanese adverbs from multiple perspectives, relying on the classification information provided by large-scale native language corpus and BCCWJ stylistic information database. The correlation analysis shows that not only the proportion of adverbs and various stylistic features, but also the text features have a relatively significant correlation. Univariate analysis of variance shows that most texts with different types of stylistic features have significant differences in the proportion of adverbs. Correspondence analysis reveals the distribution of adverbs in different texts and their specific stylistic features.

References:

[1] Lu Chang. Characteristics and grammatical functions of adverbs in Chinese and Japanese [J]. Journal of Shenyang Normal University: Social Science Edition, 2005,29 (5): 3.

[2] Zhang Y Q. Evaluation Model of auxiliary teaching quality based on active learning support vector Machine [J]. Modern Electronic Technology,2019,42(07):112-114.