An IG-RS-SVM classifier for analyzing reviews of E-commerce product

Jiajun Ye[a], Huan Ren[b] and Hangxia Zhou[c] *

College of Information Engineering, China Jiliang University, Hangzhou 310018, China

[a]254525738@qq.com, [b]864450177@qq.com, [c]zhx@cjlu.edu.cn

* Corresponding author

Keywords: e-commerce; feature selection; ensemble learning; support vector machine

Abstract. Analyzing reviews of E-commerce product is a kind of text classification which belongs to supervised learning. Due to the huge number of words, high dimensional feature space is a serious problem in text classification. In order to solve it, a new algorithm, IG-RS-SVM, is proposed. Information Gain (IG) is a feature selection algorithm which can reduce the dimension of feature subspace. Random subspace, a kind of ensemble learning algorithm, can divide the feature space to smaller ones each submitted to a base classifier such as Support Vector Machine (SVM). After experiments, it shows that IG-RS-SVM algorithm can effectively improve the text classification accuracy.

Introduction

The reviews of E-commerce product belong to a kind of text sentiment analysis. Though collecting the consumers' reviews after they purchased E-commerce products, analyzing their emotions, moods and attitudes can help other consumers decide whether to buy and also help quality supervision departments find quality problems as soon as possible which is propitious to the implementation of the supervision and spot check. Currently, there are two ways to analyze the reviews of E-commerce products[1]. One is based on emotional knowledge, and the other is based on data mining. With natural language and some existing dictionaries, the first method makes a decision to the comments directly. This method not only need to establish a huge emotional dictionary, but also can't judge the emotional tendency accurately because of the complexity of Chinese. The second method uses data mining algorithms for text classification.

Text Classification

Original text is unstructured data which computer can't understand, so that it must be converted into structured data. Text segmentation is a major link in the pretreatment. It can transform text information to structured data and delete a large number of redundant contents (including punctuation marks, stop words, repeated contents and so on). English text segmentation is relatively simple because it only need operate according to the space and punctuation. But Chinese text need to be done by relevant algorithm segmentation. For example, SCWS Chinese segmentation system [2] and ICTCLAS Chinese segmentation system are often used at present.

The text after pretreatment is not totally structured that it need a mathematical model to represent itself. The most commonly used model of text feature representation is vector space model (VSM). In that model, text is considered as a vector space consists of a set of orthogonal vectors [3].

Data mining algorithm is used to classify structured data. Common methods of data mining for text classification are Bayes classifier, support vectors machine(SVM), decision-tree and so on. Ensemble learning can effectively improve the classification efficiency of the algorithm which uses some simple classification algorithms to get a number of different learning machines and then combines them into integrated learning machine. The ensemble learning algorithm is widely used in image processing, biomedical and control engineering and other related fields.

There are some researches about text classification applied by ensemble learning algorithm. Literature [4] used a Bagging algorithm with attribute selection. This algorithm can only evaluate the contribution to the classification of a part of attributes, but can't evaluate the contribution to the classification of single attribute. In order to improve the accuracy of text classification on high dimension, literature [5] put forward RS-SVM algorithm, but without considering the dimension problem itself when choosing the feature subspaces, it was unable to filter out the features which were redundant or no contribution.

Aiming at disadvantages of the above algorithms, considering the single feature contributions to text classification and text and text dimensional reduction using on high dimension, this paper puts forward IG-RS-SVM algorithm.

IG-RS-SVM Algorithm

Information Gain. The VSM model usual has a high dimension which can reach tens of thousands or even more and most of them are redundant or irrelevant. Redundant features may cause a decline in the classifier performance and affect efficiency of data mining by analysts. Feature selection is a good way to reduce dimensions of VSM so that it can achieve the goal to improve the classification accuracy and reduce computational complexity. Common methods of feature selection are document frequency(DF), information gain(IG), mutual information(MI), chi-square(CHI) and so on [6]. IG is proved as a better method compared with five feature selection algorithms [7].

The amount by which the entropy of the class decreases after observing a certain feature reflects the additional information about the class that feature provides[8], is called Information Gain. Formula is as follows:

$$IG(W)=H(C)\text{-}H(C|W)=\text{-}\sum_{i=1}^{n}P(C_i)\log P(C_i)+\begin{pmatrix} P(w)\sum_{i=1}^{n}P(C_i|w)logP(C_i|w)+ \\ P(v)\sum_{i=1}^{n}P(C_i|v)logP(C_i|v) \end{pmatrix} \quad (1)$$

C represents text category. W represents text feature, $W\in\{w,v\}$. $P(C_i)$ represents the probability that the text belongs to C_i. $P(w)$ represents the probability that W appears and $P(v)$ represents the probability that W doesn't appear. $P(C_i|w)$ represents the probability that the text belongs to C_i with W, $P(C_i|v)$ represents the probability that the text belongs to C_i without W.

Random Subspace. However, the dimensionality of feature can be few thousands even after feature selection. Fig. 1 shows the structure of Random Subspace. In Random Subspace, after dividing the original feature space to feature subspaces, each subset is submitted to a base classifier in the ensemble [9]. Combined with the result of each base classifier, final result is obtained by a majority vote.

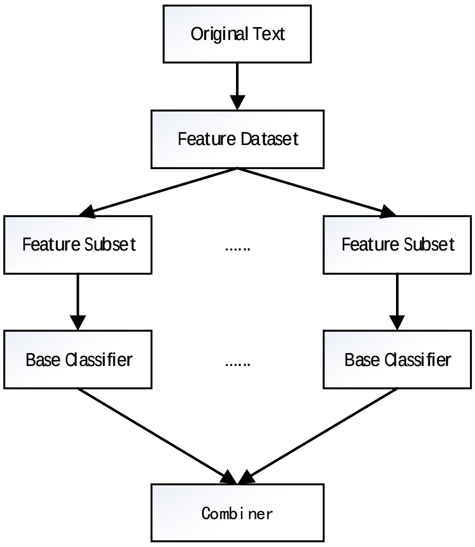

Fig. 1. The structure of random subspace

Support Vector Machines. Support Vector Machines (SVM) is a kind of machine learning methods which is proposed by Vapnik et al [10]. It has become the hotspot of machine learning because of its excellent learning performance on solving linear, nonlinear and high dimensional pattern recognition, and it also can be applied to the function fitting other machine learning problems [11-12]. The principle of SVM is to get a hyperplane to make sure that the distances of two kind points closing to the hyperplane are the farthest.

SVM has the advantage of dealing with nonlinear problems by introducing the feature transform the nonlinear problem in the original space into the linear problem in the new space, such as $(x_i \cdot x_j) \rightarrow (j(x_i) \cdot j(x_j))$ which can be remembered as a kernel function as $K(x_i, x_j) = (j(x_i) \cdot j(x_j))$ .So the final decision function is:

$$f(x) = \operatorname{sgn}\{\sum_{i=1}^{n} l_i y_i K(x_i \cdot x_j) + b\} \tag{2}$$

Using different kernel functions will have different forms of the nonlinear support vector machine. Now more commonly used kernel function mainly has three types:

Linear Kernel: $K(x, x') = (x \cdot x')$;

Polynomial Kernel: $K(x, x') = [(x \cdot x') + 1]^q$;

RBF Kernel: $K(x, x') = \exp(-\frac{\|x - x'\|^2}{s^2})$.

IG-RS-SVM Algorithm. For a text, the text after pretreatment can be expressed as $d_i = \{t_{i1}, t_{i2}, t_{i3}, \ldots, t_{in}\}$ (t represents feature, i represents the number of feature) and the category c_i which it belongs to. So the text dataset can be represented by $D = \{(d_1, c_1), (d_2, c_2), (d_3, c_3), \ldots, (d_m, c_m)\}$ and m which means the number of text in dataset.

IG-RS-SVM algorithm is described as follows:

Input: text dataset $D = \{(d_1, c_1), (d_2, c_2), (d_3, c_3), \ldots, (d_m, c_m)\}$ and m.

Output: classification result $F(d_i), F(d_i) \in C$

Setp1: transforming text dataset into VSM;

Setp2: calculating the text entropy of each feature in VSM, then putting them into a feature set;

Step3: sorting the feature set and delete the feature whose value is 0;

Step4: rebuilding a new VSM according to the new feature set;

Step5: choosing the number of SVM classifier;

Step6: to each SVM classifier, random generating a feature subspace samples from the new feature set;

Step7: classifying subspace samples with SVM classifier;

Step8: combined with the result of each SVM classifier, outputting the result by a majority vote or through after the combination.

IG-RS-SVM Algorithm

Precision, Recall and F-measure are commonly used to be evaluation indicators in the field of text classification. Computation formula is as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad F\text{-}Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3}$$

For binary classification, TP refers to the true positive which means forecast result and actual result are both true; FP refers to the false positive which means forecast result is true but the actual result is false; FN refers to the False Negative which means forecast result is false but the actual result is true. Precision is the ratio about the number of actual classification of the total samples. Recall is the ratio about the number of actual classification of the total actual samples. F-measure is harmonic mean between Precision and Recall.

Precision measures the ability of classification to refuse to the irrelevant information. Recall measures the ability of classification to classify the relevant information. F-measure measures the comprehensive ability about Precision and Recall.

Experiment Results Analysis. In order to verify the effectiveness of IG-RS-SVM algorithm in e-commerce product reviews analysis, this article selects the classic MovieReviews data set including 1000 positive evaluations and 1000 negative evaluations. This experiment cross validation method where dataset is divided into 10 portions, take nine as train data and the other one as the test data. At last, experiment uses the average value of each experiment. After text pretreatment, we get a feature dataset including 1165 features, then we use IG algorithm to keep features whose values are greater than 0. This new feature dataset includes 311 features which mean dimension of the feature fell sharply. Four classifiers are used for classification of two feature datasets in the experiments, and the experimental results are shown in Table 1.

Table 1. Classification results with different algorithms

Classifier	Correct/%	SD	Precision/%	Recall/%	F-measure/%
NB	79.1	3.81	79.5	78.5	79.0
SVM	80.7	2.99	80.4	81.0	80.7
RS-SVM	83.9	2.57	82.2	86.6	84.3
IG-NB	83.3	3.25	84.3	81.7	83.0
IG-SVM	86.8	3.65	86.7	86.9	86.8
IG-RS-SVM	86.2	2.62	84.1	89.3	86.6

Through the analysis of Table 1, we can obtain the following results:

(1) Without IG algorithm and RS algorithm, SVM got a better result than NB;

(2) Without IG algorithm, SVM with RS algorithm got a better result;

(3) Without RS algorithm, using IG algorithm have a certain upgrade of two algorithms.

(4) With IG algorithm and RS algorithm, considering the factors of standard deviation, SVM got a best result.

ROC Area of these classifications is the range of [0, 1] which is usually more than 0.5. The closer to 1 of the values is, the better the performance of the classifier is. If the value is equal to 0.5, it means that classifier is completely ineffective.

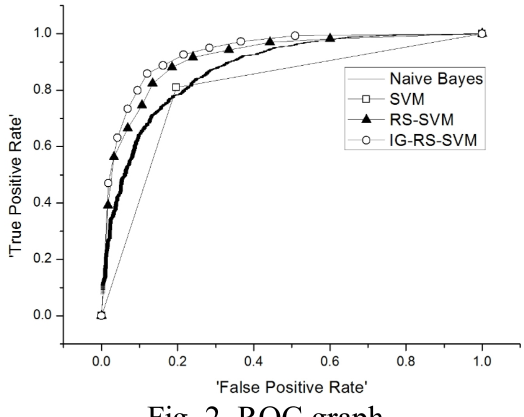

Fig. 2. ROC graph

Fig. 2 shows the ROC graph where the ROC Areas of RS-SVM and IG-RS-SVM are both the highest. The value of RS-SVM is 0.915 and the value of IG-RS-SVM is 0.927. It seems that IG-RS-SVM classifier is the best among them.

Parameter Analysis. This experiment adopts three kinds of commonly used SVM kernels such as Linear Kernel, Polynomial Kernel and RBF Kernel.

The Random Subspace Rate is also an important parameter in this algorithm which means the ratio of the feature subset. This experiment selects 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 of the proportion to evaluate the classification results under different proportions.

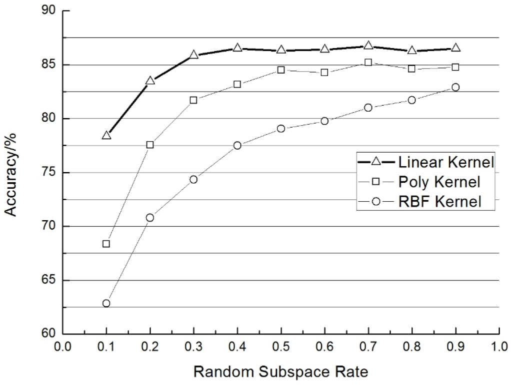

Fig. 3. Results of different ratios and kernels

Fig. 3 shows the classification result of Linear Kernel is the best, polynomial kernel is second, and the classification of RBF kernel is the worst. When the ratio is 0.7, Linear Kernel and Polynomial Kernel get a best result, and the classification results of the RBF kernel is increased with the Random Subspace Rate.

In summary, SVM classifier using the linear kernel is recommended in practical application.

Conclusions

Because of the problem about the unsatisfactory RS-SVM classification result under the high dimension of feature space, this paper introduces IG feature selection algorithm to reduce the dimensions of the feature subspace with Random Subspace algorithm. The experimental results show that, compared with other classification algorithms, IG-RS-SVM has greatly improved in classification accuracy and stability. And considering the influence of SVM kernel and Random Subspace Rate for the classification results, comparing the experimental results, show that the SVM linear kernel and 0.7 Random Subspace Rate obtains good results.

The reviews of E-commerce product quality not only provide great reference value to consumers, but also contribute to E-commerce product quality monitoring by the quality supervision departments. Through the analysis of the review of E-commerce products, governments can issue an alert to avoid the problems caused by the quality of the product, and provide a guarantee for the healthy development of E-commerce.

Acknowledgements

We are grateful to Dr. Songguo Liu (National Risk Monitoring Center for Quality of E-commerce Products, Hangzhou, China) for providing the data for analyzing, and Dr. Jun Feng (Hangzhou Institute of Calibration and Testing for Quality and Technical Supervision, Hangzhou, China) for improving algorithm of data mining. This work was supported by the General Administration of Quality Supervision, Inspection and Quarantine of the People’s Republic of China (No. 20130K001-2) and by the National Natural Science Foundation (No. 61027005).

References

[1] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and trends in information retrieval, 2008, 2(1-2): 129-135.

[2] X. Fang, S. Wang, S. Cao, A Chinese Search Approach Based on SCWS, Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Berlin Heidelberg: Springer, 2014: 665-671.

[3] Q.L. Guo, Y.M. Li, Q. Tang, The similarity computing of documents based on VSM, Application Research of Computers, 2008, 11: 3256-3258. (In Chinese)

[4] B. Robert, G. Ricardo, Q. Francis, Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern recognition: The Journal of the Pattern Recognition Society, 2003, 36(6): 1291-1302.
[5] G. Wang, S.L. Yang, Study of Sentiment Analysis of Product Reviews in Internet Based on RS-SVM, Computer Science, 2013, 40(11A): 274-277. (In Chinese)
[6] H.T. Ng, W.B. Goh, K.L. Low, Feature selection, perceptron learning, and a usability case study for text categorization, ACM SIGIR Forum, 1997, 31(SI): 67-73.
[7] Y. Yang and O.J. Pedersen, A comparative study on feature selection in text categorization, Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 412–420.
[8] C.J. Shang, D. Barnes, Combining support vector machines and information gain ranking for classification of mars McMurdo panorama images, IEEE International Conference on Image Processing, 2010: 1061-1064.
[9] M.J. Gangeh, M.S. Kamel, R.P.W. Duin, Random Subspace Method in Text Categorization, International Conference on Pattern Recognition, 2010: 2049-2052.
[10] V. Vapnik, The nature of statistical learning theory, Springer Science & Business Media, 2000.
[11] Q. Liu, C. Cui, H.X. Zhou, Application of a kind of modified SVM multi-class classification algorithm in wireless sensor networks, Journal of China Jiliang University, 2013 (3): 298-303. (In Chinese)
[12] S.L. Wang, Intrusion detection system for WSNs based on SVM, Transducer and Microsystem Technologies, 2012, 31(7): 73-76. (In Chinese)