Analysis of massive unsupervised text sentiment based on rough set time series model

Du Baochen[1]

1. College of Software Engineering, Beijing University of Posts and Telecommunications, Beijing, 100091, China

Key Words: Text Sentiment Recognition, Sample Subspace, Dynamic Classification, Integrated Classification Model

Abstract: One text sentiment classifier constructed based on the mechanism of dynamic classification of sample space has been proposed to improve the accuracy of Chinese text sentiment recognition by starting from the perspective of integrated learning. This algorithm makes full use of the identification information within training sample space, makes adaptive classification for sample space by introducing kernel smoothing method, forms several multi-granularity subspaces with differences, and then constructs base classifier in each subspace respectively and finally integrates the output of all base classifiers to produce the final prediction results. Experimental results on Chinese data set have shown that this algorithm is superior to Bagging, Adaboost and other algorithms in precision ratio and recall ratio etc and is also with good application prospect in the sentiment recognition of large-scale sample set.

1 Introduction

With the development of Web2.0 technology as well as the increasing demand of people for information, more and more internet users express their views and opinions on one object through blog, micro-blog, forum, online reviews and other public communities. This information with subjective sentiments is mainly stored in the form of text. Under the trend of exponential growth of network text, public opinion analysts and users need to read a lot of comments every day and then attain valuable information from it. Viewpoints, attitudes, sentiments and other information in these texts can provide important research and investigation resources for customers, companies and government departments. Nevertheless, these increasing network comments also bring a lot of interference to common users and a great amount of noise data is not good for users to judge the objects accurately. It is difficult to explore potential important information quickly and steadily with adoption of manual mode. How to make use of automatic way to recognize and analyze these texts with sentimental signals has become a popular research topic in current network intelligent information processing. Therefore, effective sentiment recognition method is of great research value and practical significance for investigation and analysis of internet public opinion.

2 Integrated identification scheme based on sample subspace

2.1 Problem description

To construct vector space model, training and testing texts used for sentiment recognition can be expressed as vector composed of differnt sentimental words. Set $W=\{w_1, w_2, w_3..., w_d\}$ as the descending set of d sentimental words with the highest appearance frequency in the training set, f_{ij} is the freequency of No. j sentimental word (w_j) in No. i text, and then each sample s_i can be expressed as $(f_{i1}, f_{i2}, f_{i3}..., f_{id})$. In addition, define full sapce of training sample as $S=\{(s_i, c_i) \mid 1 \le i \le n\}$, for any sample $s_i \in \Re^d$ in the space, its sentiment category is $c_i \in C=\{-1,+1\}$, -1 is negative sentiment and +1 is positive sentiment.

Set $S_{k:n}^{r_k}$ is No. k sample subspace of s, in which r_k $(r_k < n)$ is the subspace granularity of No. k adaptive updating during iteration. Set $\Psi_k(S_{k:n}^{r_k})$ as the base classifier Ψ_k generated during training in subspace $S_{k:n}^{r_k}$, and then the integrated classification model based on L sample subspaces can be

expressed as:

$$Ens = RDS\{\Psi_k(S^{r_k}_{k:n}), R, fus, 1 \le k \le L\} \quad (1)$$

In which, fus is fusion strategy of base classifier. To confirm suitable granularity value of each subspace, it needs to construct an adaptive updating R distribution function. RDS is the R-based dynamic subspace method, which is the key ensemble classification algorithm adopted in this paper.

2.2 Overall Scheme

Based on above descriptions, problems need to be further confirmed in integrated identification scheme based on dynamic classification of sample space include: preparation of sentiment dictionary, selection method for sentimental characteristics, confirmation for the number of sample subspace (number of base classifier), selection method for sample subset within subspace, category of base classifier and confirmation for fusion strategy. For these problems, this paper plans to take following scheme:

(1) The vocabulary with emotional orientation information included in this paper is the best, which expresses the characteristics of overall sentiment; therefore, this paper adopts sentiment evaluation words set used in literature [1] and [2] as sentiment dictionary. The dictionary has integrated the standard sentiment words set published by Tsinghua University [3] and HowNet[4], in which there are 8015 commendatory terms and 6733 derogatory terms.

(2) For the selection of sentiment characteristics, firstly, it adopts word segmentation tool ICTCLAS2012[5] of Chinese Academy of Sciences to make batch word segmentation for text set, saves word segmentation results and then extracts sentiment words included in word segmentation results as characteristics based on sentiment dictionary made based on scheme (1).

(3) Take the number of sample subspace L as parameter for consideration and analyze the influence of different L values on sentiment recognition results.

(4) For the selection of sample subset, make random selection for sample setting and other probability distributions based on uniform distribution U. For subspace S_k, the selection steps are as following:

a) One random number u following distribution U will be produced within [0,1] interval;

b) If u meets $F_U(i-1) < u < F_U(i)\ (1 \le i \le n)$, and then select No. i sample to put into subspace, in which F_U is the density function following U distribution;

c) Return to (a) until r_k samples in subspace S_k finish the selection.

(5) Base classifier adopts support vector machine LSVM [1] (Linear Support Vector Machine) based on linear kernel function for its extensive application in text classification and good classification performance and processing speed in sparse feature space. LSVM is realized by adopting LIBLINEAR [2], which is suitable for large-scale sample processing.

(6) Fusion strategy adopting majority voting as base classifier

$$Senti(Ens, x) = \arg\max_{c\in\{-1,+1\}}\{\sum_{k=1}^{L}\Psi_k(S^{r_k}_{k:n})\} \quad (2)$$

In the formula, base classifier $\Psi_k \in \{0,1\}$, $Senti(Ens, x)$ is sentimental polarity attained by adopting integrated classification model Ens for input test sample x.

3 Dynamic subspace algorithm

In traditional integrated learning method, the granularity selection of subspace needs preset, which makes the sample number contained in each subspace as equigranular distribution and the difference between base classifiers can't be ensured. To get over this defect, dynamic subspace model has been introduced in this paper. This model adopts iterative method to construct each subspace; its specific method is to select suitable granule size adaptively according to classifier training performance constructed in different sub-regions. For this, it needs to build up R distribution to provide probability evidence for the granularity selection of subspace.

To construct R0 distribution, firstly, it needs to produce b training subsets and its production

mechanism obeys uniform distribution U, that is $\tilde{S}_t = RDS(S, r_t, U), t = 1,2,...,b$, in which the granularity of initial sample subspace r_t is confirmed based on formula (3) and floor means integral calculation:

$$r_t = b + floor\left(\frac{1}{2} \times \frac{n-1}{t-1} \sum_{s=1}^{t-1} h_s\right) \tag{3}$$

And then construct initial base classifier $h_t \ (t = 1,2,...,b)$ on b training subsets and estimate the training accuracy $\varphi(h_t)$ of each base classifier h_t in initial subspace $\tilde{S}_t$. The training accuracy represents the training effect when granularity is r_t. Attain discrete R0 distribution based on b training accuracy values and finally adopt KS method to make continuous treatments for it. KS smoothing process is as shown in formula (4):

$$f_R(r) = \frac{1}{\sum_{t=1}^{b} \varphi(h_t)\sigma} \left[\sum_{t=1}^{b} \varphi(h_t) K\left(\frac{r - r_t}{\sigma}\right)\right], r = 1,2,...,n \tag{4}$$

In which, K is kernel function and σ is smoothing factor.

In this research, it takes R distribution as basis and adopts inverse method produced by pseudo random number to determine the granularity of subspace. For this, set $F_R(r)$ as cumulative probability density and define as following:

$$F_R(r) = \sum_{t=1}^{r} f_R(t) \ (t = 1,2,...,r) \tag{5}$$

Selection steps for subspace are as following:

(1) One random number v following uniform distribution will be produced within [0,1] interval;

(2) In each iteration, cumulative probability density value is $F_R(r)$, if $F_R(r-1) < u < F_R(r) \ (1 \le r \le n)$, select r as the granularity of this subspace.

In the process of subspace formation, R distribution will be updated continuously, which will avoid the repeated selection of a lot of samples; in addition, the sentiment discrimination information with subspace of different granularities will be distributed in a more uniformed way, which provides classification complementary information for the construction of following base classifier and improves the effectiveness of integrated system.

4 Experimental research

4.1 Data set

The experimental data used in this paper is from Chinese comment corpus [5] about hotel and notebook provided by Tan Songbo. This corpus set is balanced data set with 2000 positive and 2000 negative respectively. Pretreatment of data set includes three parts: (1) Remove characters irrelevant to review topic, including html code and some guidance words, such as “free registration”, “hotel feedback” and “supplementary comment” etc; (2) Standardized treatment of data, including converting complex words to simplified words, converting half-width and full-width symbols; (3) Adopt sentiment dictionary prepared in scheme (1) in chapter 2.2 and then make statistics for the appearance frequency of all sentimental words in two corpus sets. After pretreatment, the related information of data set is as shown in table 1:

Table 1 Experimental Data Set Information

Number of words in corpse			
50.58	71.80	8.56	2099
31.61	33.61	4.92	1185

Number of words in corpse

It can be seen from table 1 that the length of comment text is generally short and each text includes few sentimental words, which accounts for less than 15% of the text. In addition, in two data sets, the proportions of sentimental words are 52.48% and 29.63% respectively. This

information granularity indicates that the dynamic classification method based on sample space is feasible.

4.2 Experimental design

To verify the recognition performance of RDS algorithm as well as its accuracy and robustness in Chinese sentiment recognition, the specific experimental design scheme is as following:

(1) Investigate the sampling number of subspace L (that is the number of base classifier) as well as the influence of number b produced by training subset in initial distribution R0 on the recognition performance of RDS algorithm. Select L at 20, 40, 80, 100, 120 and b at 5, 10, 20, 30, 40.

(2) Compare the recognition performances of RDS, integrated algorithm Bagging, Adaboost and single classifier LSVM and investigate the recognition effects of four methods in standard data set, in which LSVM classifies based on original sample space, Bagging algorithm [1] makes equigranular random classification for sample space, Adaboost algorithm [2] makes dynamic adjustment for sample weight and forms equigranular subspace in way of iteration. All the integration methods in the experiment adopt LSVM as bottom base classifier. Three kinds of integrated recognition algorithms use different base classifiers to make experiment (L=20, 40, 120). In addition, set the selection granularity of subspaces in Bagging and Adaboost algorithms at 50% of original training sample size and set initial training subset b at 30 in RDS.

(3) In bottom base classifier LSVM, set the loss parameter C of linear kernel function at 0.08 and cost parameter E at 0.01.

(4) In above two kinds of experiments, training set and testing set are segmented at the proportion of 4:1, which include 3200 and 800 samples respectively. Because the classification of subspace is with randomness, therefore, three kinds of integrated classification algorithms adopt the average value of ten independent operations. All the experiments are finished on MATLAB7.1 platform.

4.3 Experimental results and analysis

The selection of initial training subset is the key step of RDS algorithm, the bigger number B of initial training seed, the more feedback data will be provided for the generation of following subspace; on the contrary, the smaller b, the closer of subspace structure to the state of complete random selection and the lower robustness of integrated recognition model. But bigger b value will cause the decrease of algorithm performance and the calculation complexity of initial distribution R0 will be improved. The increasing of b value improves recognition effect to some extent. In view of the recognition conditions for two corpuses, b value is with better effect while increasing in the interval of [5, 30], but the improvement of performance is not significant if continue to increase b value, which explains that b=30 is a turning point for identifying the stability of performance. Moreover, generation number L of subspace has greater influence on recognizing the performance of model. In view of the differences in integrated learning, the bigger L, the difference of base classifier will decrease, and then several base classifiers will have the same classification ability; subspaces exceeding certain number do not provide effective classification complementary information for integrated model; under extreme conditions, phenomenon of equivalent effect between integrated classifier and single classifier will appear. It can be observed that when identify hotel data set, when the number of base classifier L=20, the value of F1 is only 84% due to less iteration quantity; when L=40, the recognition performance is good, achieving 85.74%; when L=120, the sample coverage rate selected in subspace achieves above 96%, but due to the decreasing of difference between base classifiers, F1 value will not increase any more, but presents a decreasing trend. In the processing of notebook corpus, the recognition effect of RDS is good; when b=30 and L=100, it attains the highest F1 value at 91%. In addition, while the number of base classifier L increases within the interval [20, 100], the improvement of recognition performance is significant and the highest improvement degree is 5%.

5 Conclusions

It can be seen from a series of experiments that when make Chinese text sentiment recognition with integrated learning method, its recognition performance is determined by the difference

between classification granularity of subspace and different subspaces. For the characteristics of big sample space and high sparse degree etc in Chinese sentimental recognition, one integrated classification method based on sample space classification has been proposed and adaptive selection for subspace has been made through kernel smoothing method. Experiments have shown that this kind of subspace selection method with multi-granularity can improve the performance of Chinese text sentiment recognition effectively and can realize sentiment recognition, viewpoint mining and other work for large scale comment set. In the following, it will make further research on the influence of sentimental characteristics distribution within sample space on algorithm performance in dynamic subspace. Moreover, a more optimized Chinese sentiment recognition algorithm will be designed by combining the integrated classification mechanism based on sentimental characteristics space.

References

[1] Chan C C, Liszka K J. Application of Rough Set Theory to Sentiment Analysis of Microblog Data[J]. Intelligent Systems Reference Library, 2013, 43:185-202.

[2] Yu J K, Zhao L. Sentiment Feature Selection Algorithm for Chinese Micro-blog[C]// International Conference on Management of E-Commerce and E-Government. IEEE, 2014:114-118.

[3] Lazaridou A, Titov I, Sporleder C. A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations[C]// Meeting of the Association for Computational Linguistics. 2013:1630-1639.

[4] Wang S, Li D, Wei Y. A method of text sentiment classification based on weighted rough membership[J]. Journal of Computer Research & Development, 2011, 48(5):855-861.

[5] Ahmad S R, Abu Bakar A, Yaakub M R. Metaheuristic algorithms for feature selection in sentiment analysis[C]// Science and Information Conference. IEEE, 2015:222-226.