Advanced Science and Technology Letters
Vol.78 (MulGrab 2014), pp.49-54
http://dx.doi.org/10.14257/astl.2014.78.09

A kernel density estimation based text classification algorithm

Liu Yang[1,2], Han Liguo[1*1], Cai Xuesen[2]

1 College of GeoExploration Science and Technology, Jilin University, Changchun 130026, China
2 College of Computer Science and Technology, Changchun Normal University, Changchun, 130032, China
E-mail: han_lg@126.com

Abstract: This paper proposes a small sample text classification algorithm based on kernel density estimation. Firstly, the probability density of the text classification problem is estimated. Then we can construct auxiliary training samples by the estimated probability. Finally, the classification model is obtained with the help of auxiliary training samples. As the introduction of auxiliary training samples avoids over-fitting caused by small training samples, the proposed algorithm can effectively improve the performance of small sample text classification problems. The simulation experiments on the news text datasets fully verify the effectiveness of the proposed algorithm.

Keywords*:* text classification; probability density estimation; small sample; news text dataset

1 Introduction

Text classification [1] is an extremely important research direction in pattern recognition. And with the development of Internet technology, the role of Text classification is becoming more and more important. For example, Good public opinion analyses can be made through text recognition, thus make the government timely understand of the people's demands, and meanwhile conducive to make timely adjustment measures for the government. Shopping site can grasp consumers' attitude very well through Text recognition, and timely improve their service quality. Currently, there have been a lot of technologies applied to text categorization, such as Bayesian analysis method, KNN method[2], Support vector machine (SVM) method[3], Neural network method, the decision tree method and so on. As mature classification methods, these methods have achieved good learning results on text classification problems. However, in some cases, the number of training samples in text classification problem is very small, the traditional text classification are usually

[1] Corresponding author: Liguo Han

ISSN: 2287-1233 ASTL

unable to obtain better classification effect, thus a new text classification algorithm which is suitable for small sample text classification problem is required to be developed. For small sample problem, a small sample text classification algorithm based on kernel density estimation was proposed. The simulation experiments on the news text datasets fully verify the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. Section 2 introduces the problem of text classification and the algorithm of kernel density estimation. Section 3 presents the small sample text classification algorithm based on kernel density estimation. In Section 4, we apply the proposed algorithm on news text dataset, give the main results, and make a full analysis. Section 5 summaries the main contribution of this paper.

2 kernel density estimation

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable which was proposed by Rosenblatt and Parzen. Therefore it is also termed the Parzen-window method.

As the probability of a vector *X* lies in field *R* is:

$$P = \int_R p(X)dX \tag{1}$$

Let $\chi = \{x_1, \cdots, x_N\}$ be an independent and identically distributed sample drawn from some distribution with an unknown density *p*(*x*). Because each data point has a probability *P* of falling within *R*, the total number *k* of points that lie inside *R* will be distributed according to the binomial distribution:

$$P_k = \begin{pmatrix} N \\ k \end{pmatrix} P^k (1 - P)^{N-k} \tag{2}$$

the probability P can be estimated by following

$$\hat{P} = \frac{k}{N} \tag{3}$$

If, however, we also assume that the region *R* is sufficiently small that the probability density *p*(*x*) is roughly constant over the region, then we have

$$P = \int_R p(\mathbf{x})d\mathbf{x} = p(\mathbf{x})V \tag{4}$$

Where *V* is the volume of *R*.

Combining (3) and (4), we obtain the density estimate in the form

$$\hat{p}(\mathbf{x}) = \frac{k / N}{V} \tag{5}$$

Let the region *R* to be a small hypercube centred on the point *x* at which we wish to determine the probability density. In order to count the number K of points falling within this region, we define the following function

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \le \frac{1}{2},\ j = 1, \cdots, d \\ 0 & otherwise \end{cases} \tag{6}$$

Let h_n is the width of the region *R*, then its volume can be computed by

$$V_n = h_n^d \tag{7}$$

and the total number of data points lying inside this cube will therefore be

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \tag{8}$$

Substituting this expression into (5) then gives the following result for the estimated density at *x*

$$\hat{p}_n(\mathbf{x}) = \frac{k_n / n}{Vn} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \tag{9}$$

Clearly, $\hat{p}_n(\mathbf{x})$ subject to the following two conditions

$$\begin{aligned} &\hat{p}_n(\mathbf{x}) \ge 0 \\ &\int \hat{p}_n(\mathbf{x})\, d\mathbf{x} = 1 \end{aligned} \tag{10}$$

which ensure that the resulting probability distribution is nonnegative everywhere and integrates to one.

3 Algorithm design

With respect to small sample text classification problem, there are no sufficient samples for training. Thus the traditional classifiers tent to be over-fitting and cannot obtain a good performance. For small sample text classification problem, in this paper, we proposed a small sample text classification algorithm based on kernel density estimation, denoted by TCAKDE.

The main idea of this algorithm is that obtaining the probability density function of the original text data by kernel density estimation method. Thus the original training set can be expanded and the traditional classifiers can obtain a better classification performance. Table 1 below gives TCAKDE algorithm.

Table 1. TCAKDE algorithm

Input: original small sample training set $T = \{(x_i, 0), i = 1, \cdots, n_0\} \cup \{(x_i, 1), i = 1, \cdots, n_1\}$, classifier C, kernel function $\varphi(u)$
output：text classification model *TM*
process

```
    $\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}u^2\right\}$          //select Gaussian window as kernel function
$f$=$KDE$($T$, $\varphi(u)$);                    // estimate the density function $f$ on training set T by
                                          kernel density estimation method
 $AT$= Auxiliary_Sample_Generation($f$) ;     // construct auxiliary training set $AT$ by
                                          the density function $f$
$D = T \cup AT$ ;                  // obtain the new training set
$TM$=C($D$)    // train classifier C on the new training set D, and obtain the text
classification model $TM$
```

4 Experiments

4.1. Experimental data

To test the performance of SSTCAKDE algorithm, web page data has been selected to conduct text classification experiment in this paper. The selected data set comes from the news text which have been collected by sohu news site. To facilitate the experiment, merely four types of news topic including sci-tec, talk, education and economic have been extracted from the complex news content to conduct classification test. Meanwhile, in order to make the experimental data meet the demands of the algorithm proposed in this paper, the small number of text training data has been selected from each news topic in this experiment. In this experiment, for each type of news topics, 200 samples have been selected to conduct training, and 600 samples have been selected to conduct training. Using the method in literature [4] to conduct preprocessing of the collected new text data to get the training data and test data.

4.2. Indexes of classification performance

we chose precision (abbreviated as P) and recall (abbreviated as R). The calculation formula is as follows:

$$P = \frac{n_1}{n_2} \;,\; R = \frac{n_1}{n_3} \qquad (11)$$

where n_1 represents the number of pages to be classified correctly, n_2 represents the number of pages to be classified in this kind, n_3 represents the total number of pages belonging to this class.

4.3. Experimental method

In order to test the performance of TCAKDE algorithm, we compare KNN algorithm and SVM algorithm on the selected dataset. Detailed process is as follows: Set KNN as the classifier for text classification in SSTCAKDE algorithm and compare its result with that of KNN algorithm. Set SVM as the classifier in SSTCAKDE algorithm and compare its result with that of SVM. We select the C-SVM algorithm as the SVM algorithm, where C is a penalty factor. This experiment selects Gaussian kernel as kernel function:

$$K(x,y)=\exp(\frac{-\|x-y\|^{2}}{2\sigma^{2}}) \tag{12}$$

where σ is a width parameter, "x" and "y" are n-dimensional vectors.

4.4. Experimental result analysis

The average classification result on precision is reported in Figure 1, and the average classification result on recall is shown in Figure 2.

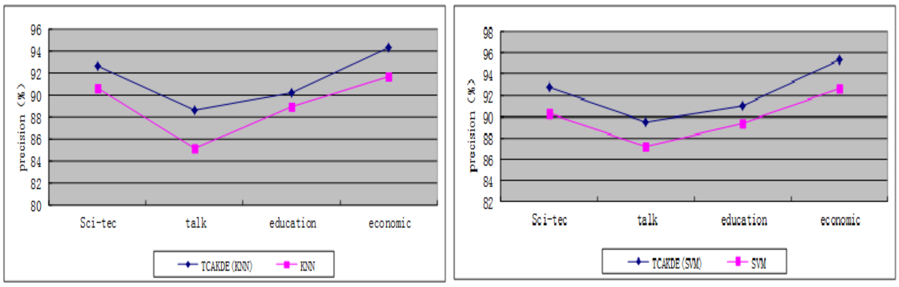

Fig. 1. Precision comparison on KNN and SVM models

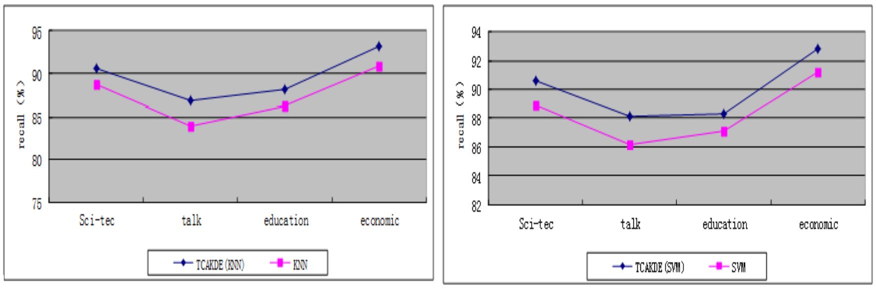

Fig. 2. Recall comparison on KNN and SVM models

As shown in the above two figures, the proposed TCAKDE algorithm performs better than KNN and SVM models on classification results. This is mainly due to a reasonable expansion on the original training set for TCAKDE algorithm. Thus the traditional classifiers can be trained with more training samples, avoiding over-fitting. Moreover the experiment also indicates that this method is an effective text classification algorithms which is independent of classifiers.

5 Conclusion

In this paper, we proposed a kernel density text classification algorithms. The algorithm can obtain a reasonable expansion on training set by using kernel density estimation method which has a good density function estimation performance. Therefor the traditional classifier can avoid over-fitting effectively and perform better on small sample text classification problems.

References

1. Bagdouri M, Webber W, Lewis D D, et al. Towards minimizing the annotation cost of certified text classification[C]//Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013: 989-998.
2. Cassotti M, Ballabio D, Consonni V, et al. Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method[J]. Alternatives to laboratory animals: ATLA, 2014, 42(1): 31-41.
3. Sharan R V, Moir T J. Comparison of multiclass SVM classification techniques in an audio surveillance application under mismatched conditions[C]//Digital Signal Processing (DSP), 2014 19th International Conference on. IEEE, 2014: 83-88.
4. Lan jun, Shi huaji, Li xingyi,etal. associative web document classification based on word mixed weight [J]. ComputerScience,2011,38(3):187-19