Advanced Science and Technology Letters
Vol.28 (NT 2013), pp.10-14
http://dx.doi.org/10.14257/astl.2013.28.03

The Optimal Information Retrieval Based on SVM

Sun Jianming[1] and Sun Qingli[2]

[1]school of computer and information engineering,
Harbin University of Commerce,Harbin(China)
Email :sjm@hrbcu.edu.cn
[2]school of management, Harbin University of Commerce, Harbin(China)
Email :sunql@hrbcu.edu.cn

Abstract. A new sentence level kernel function based on string sequence kernel function and word sequence function is proposed in this paper. Also two feasible algorithms-sentence collection kernel function and sentence sequence kernel function are put forward. These two kernel functions are based on sentence level and sequence kernel function. Pre-treatment text database is treated by practicing the SVM method. The classification and discrimination function of different text can be drawn. The resource description of every database is established in this paper. Then to classify the information with the sentence sequences function and optimize the distribution of time retrieval.

Keywords: SVM; selection strategy; text database; sentence sequences function

1 Introduction

There are several document classification methods popular at home and abroad, such as KNN, Decision Tree, simple Bayesian methods, SVM, Neural Networks, linear least squares fitting methods, maximum entropy model ,Genetic algorithm and so on. Many research results show that KNN and SVM are the best methods to do English document classification. Researching on the retrieval with time limitation in distributed retrieval, the optimal resource description and selection strategy is proposed with the SVM method combing the retrieval method based on topic clustering. Pre-treatment text database is treated by practicing the SVM method. The classification and discrimination function of different text can be drawn. The resource description of every database is established in this paper. Then to classify the information with the sentence sequences function and optimize the distribution of time retrieval. This system is proved to have excellent performance of classifying and retrieval.

ISSN: 2287-1233 ASTL

2 MODEL

SVM is a kind of machine learning method based on statistic learning theory. The number of samples should be infinity in traditional statistic, but in reality, the number of samples is limited. Therefore, some theoretical excellent learning methods cannot solve the practical problems perfectly. Statistic is the theory to study the small sample statistical estimation and forecast, including the following contents, statistical learning consistency conditions based on the Empirical Risk Minimization Principle, Guass distributing of learning methods, Empirical Risk Minimization Principle based on Guass distributing of learning methods and SVM methods to achieve these criteria. As shown in Figure 1, statistical learning theory proposed a new strategy to construct the set of functions as a sequence of subset of functions. Each subset is arranged according to VC dimension size and Empirical Risk Minimization is searched in every subset. The empirical risk and confidence interval are considered in subset and the minimization of real risk is obtained. This is the institutional risk minimization principle. Empirical risk depends on a specific function that depends on a set of functions. The selection of subset function corresponds to the traditional selection and the selection of specific function in subset corresponds to the traditional parameter estimation. Statistical learning theory provides a rigorous theoretical analysis of model selection based on Guass distributing and also proposed the conditions that a rational subset function structure should meet. The nature of the convergence of real risk is proposed based on Structural Risk Minimization Principle.

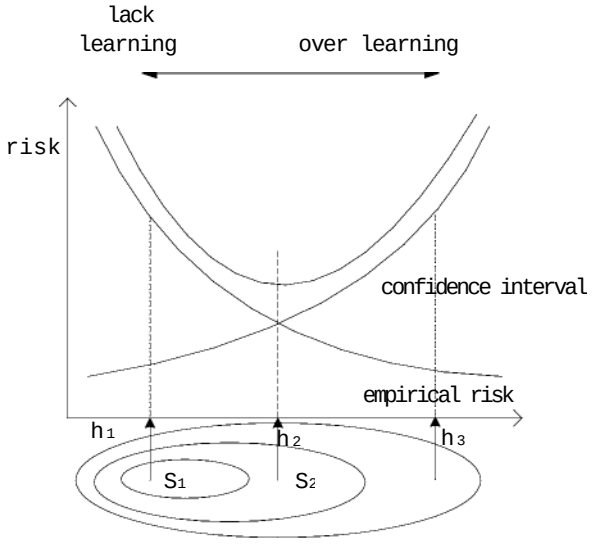

Fig. 1. Structural risk minimization principle

LIBSVM algorithm is a optimal method from Sequential Minimal Optimization and SVMlight and it has improvements for strategy selection in working set. SMO algorithm decomposes the optimal problems into series of QP problems. QP problems with two Lagrange multipliers are treated in iteration. Supposing the selected variables are α_1 and α_2, optimization problems can be described:

$$\min W(\alpha_1, \alpha_2) = \frac{1}{2} K_{11} \alpha_1^2 + \frac{1}{2} K_{22} \alpha_2^2 + s K_{12} \alpha_1 \alpha_2 + y_1 \alpha_1 v_1 + y_2 \alpha_2 v_2 + \Psi_{constant}$$

$$s.t. \quad y_1 \alpha_1 + y_2 \alpha_2 = -\sum_{i=3}^{l} y_i \alpha_i = constant$$

$$0 \le \alpha_i \le C, i = 1,2, \ldots, l$$

Supposing the initial feasible solutions are α_1^{old} and α_2^{old}, the optimal solutions are α_1^{opt} and α_2^{opt}. To meet the linear constraints $\sum_{i=1}^{l} y_i \alpha_i = 0$, the relationship between two variables should be:

$$y_1 \alpha_1 + y_2 \alpha_2 = constant = y_1 \alpha_1^{old} + y_2 \alpha_2^{old} = y_1 \alpha_1^{opt} + y_2 \alpha_2^{opt}$$

$$0 \le \alpha_1, \alpha_2 \le C, i = 1,2,\ldots,l$$

3 Algorithm Implementation

Generally, reading all the documents vectors and selecting some parameters (the type of SVM, the type of kernel function) can get the classification model documents, which are saved to be documents with ". Model" suffix. These documents can be used to forecast. First versions in this package cannot support the functions. So the source code of the packages must be modified to support all kinds of functions.
In old version, there are different calculating methods for different kernel functions.

For example, polynomial kernel function can be calculated by:

```
Case svm_parameter.POLY:
Return System.Math.Pow(gamma*dot(x[i],x[j])+coef0,degree);
```

We use custom kernel function and add a new kernel function category, PRECOMPUTED:

Easesvm_parameter.PRECOMPUTED:

Return x[i][Convert.ToInt32(x[j][0].value_Renamed)].value_Renamed;

Value_Renamed is can be distinguished from “value”.

4 Retrieval experiments based on optimal search strategy

There are 145 categories in all 7300 training samples. Ten categories with most samples such as earn, acq, money-fx, grain and so on are singled out. If the training samples belong to the selected categories, they are marked as positive classifications and other samples are marked as negative classifications. Using the sentence collection kernel function method can get the 10 discrimination functions for every category. We standardize two optional keywords and get the query vector that will be put into the 10 discrimination functions so the function value can be calculated. Only the categories with positive function value are considered. Adjusting the discrimination functions value can get the value of initial probability distribution:

$$p_i(x) = \frac{F_i(x)}{\sum_j F_j(x)}.$$

We will do retrieval in accordance with the order of $p_i(x)$. Graph 3 shows the comparison of the average search strategy and the optimal search strategy. The horizontal axis represents the number of all documents retrieved and the vertical axis represents the number of related documents retrieved.

5 Conclusion

An optimal search strategy based on new SVM method is proposed, resource constraint and accuracy taken into account. Experiment results shows this algorithm is better than the average search strategy. SVM of documents classification and kernel function are researched deeply in this paper. A new sentence level kernel function based on string sequence kernel function and word sequence function is proposed in this paper. Also two feasible algorithms-sentence collection kernel function and sentence sequence kernel function are put forward. These two kernel functions are based on sentence level and sequence kernel function. Experiment results shows that the new algorithm has better performance than traditional algorithm.

Acknowledgements. This work was supported by The Education Department of Heilongjiang province science and technology research projects (Grant Nos. 12531160)

References

1. Pal, M.and Foody, G.M.,Feature Selection for Classification of Hyperspectral Data by SVM, IEEE Geoscience and Remote Sensing.PP,99(2013)
2. M. Chi and L. Bruzzone, A semilabeled-sample-driven bagging technique for ill-posed classification problems , IEEE Geosci. Remote Sens. Lett. 2,1(2005).
3. I. Guyon , J. Weston , S. Barnhill and V. N. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46,13(2002)
4. H. Peng , F. Long and C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy , IEEE Trans. Pattern Anal. Mach. Intell. 27, 8(2005)
5. Yaoyong Li, Kalina Bontcheva and Hamish Cunningham., SVM Based Learning System for Information Extraction. Lecture Notes in Computer Science.3635(2005)
6. Edward Y. Chang.,PSVM: Parallelizing Support Vector Machines on Distributed Computers. Foundations of Large-Scale Multimedia Information Management and Retrieval.**978**(2011)
7. Giorgos Mountrakis, Jungho Im, Caesar Ogole.,Support vector machines in remote sensing: A review, Journal of Photogrammetry and Remote Sensing.66(2011)
8. M. Arun Kumar,M. Gopal., Least squares twin support vector machines for pattern classification. Expert Systems with Applications.36(2009)
9. R. Daz-Uriarte and S. A. de Andrs ,Gene selection and classification of microarray data using random forest, BMC Bioinf. 7,1(2006)
10. C.-W. Hsu and C.-J. Lin A comparison of methods for multi-class support vector machines, IEEE Trans. Neural Netw. 13,2(2002)
11. M. Pal, Margin-based feature selection for hyperspectral data, Int. J. Appl. Earth Observ. Geoinf. 11,3(2009)

