International Journal of Control and Automation
Vol.7, No.8 (2014), pp. 419-426
http://dx.doi.org/10.14257/ijca.2014.7.8.38

Research on 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)*
[2]*School of managemen, Harbin University of Commerce, Harbin(China)*
[1]*sjm@hrbcu.edu.cn*, [2]*sunql@hrbcu.edu.cn*

Abstract

Research 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.

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

1. Introduction

How to choose the database for retrieval, complete the resource description and select, reduce the resource waste and optimize the resource distribution has been the key point to efficient information retrieval. Some traditional distributed retrieval technologies, such as Meta search engine, have resolved important problems like integrating single retrieval engine and merging the retrieval results. According to related experiments results, the efficiency and accuracy of Meta search engine are poor, wasting searching time and providing the customers with psychological burden.

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.

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

ISSN: 2005-4297 IJCA

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.

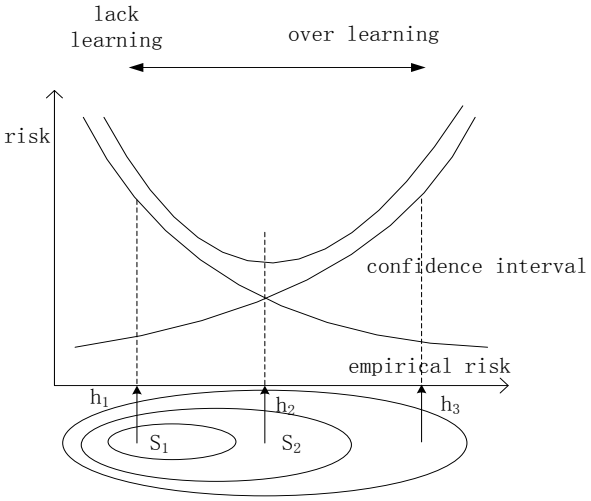

Figure 1. Structural Risk Minimization Principle

There are two ways to achieve the Structural Risk Minimization Principle. One way is to search the smallest empirical risk in each subset and then select the subset in which the sum of the smallest empirical risk and confidence interval is smallest. This method is time consuming so when there are too many subsets, it cannot work well. The other way is to design some structure for functions set to make each subset have access to a minimization empirical risk, and then choose the subset in which the confidence interval is the smallest. The function in this subset is optimal. SVM belongs to the second method.

2.1. Document Categorization Model of Sentence Collection Kernel Function

Sentence collection kernel function perfectly reflects the sentence level semantic relationships but ignores the order relationships between sentences. We use the word sequence kernel function to handle the sentence structure. In fact, sentence sequence kernel function is the expanded form of word sequence kernel function. Common word sequence kernel functions take the document as a sequence of words and ignore the sentence level. Sentence sequence kernel function takes the document structure level into consideration and proposed new interval calculate methods based on word sequence kernel functions.

The sequence composed by three Nouns has a stronger judgment than the sequence composed by preposition, article and noun. In order to alleviate the judgment imbalance,

different recession factors are introduced for different symbols. So the set mapping function changes to be:

$$\Phi u(s) = \sum_{i:u=s[i]} \prod_{1 \le j \le |u|} \lambda_{m,u_j} \prod_{il \prec k \prec i|u|, k \ne i} \lambda_{g,s_k} \qquad (1)$$

In formula 1, $\lambda_{g,x}$ is the recession factor when X is the interval , and $\lambda_{m,x}$ is recession factor when X is match. According to the definition, interval recession factor can be determined by morphology and match recession factor can be determined by inversed document frequency. The mapping of subsequence "u=we are friends" in sequence "we are good friends" is $\Phi u(s) = \lambda_{m,we}\lambda_{m,are}\lambda_{g,good}\lambda_{m,friends}$ (2)

The following formula is used to make the inversed document frequency of the key words embedded into the word sequence kernel function:

$$idf_c = \log\left|\left(\frac{N}{N_c}\right)\right| \qquad (3)$$

In formula 3, N represents the number of all documents in set and Nc is the number of the documents including the key words "c". The result of this formula is between 0 and log(N). The recession factor must be standardized:

$$\lambda_c = \frac{\log\left|\left(\frac{N}{N_x}\right)\right|}{\log(N)} \qquad (4)$$

2.2. Optimal Search Model based on SVM

Supposing S is the set of all the document databases and there are n documents in S. Every document in S corresponds to a vector in feature space. After treating the users' query conditions, query vector q can be acquired and q is seen as a document in feature space. It is document implementation that depends how to represent the document vector. Retrieving a specific task is defined as a process in which we search in the feature space and find all the results conforming to the specified requirements. Documents included in results are closest to the query vector. The similarity of documents vectors is calculated by kernel function or the inner product in the feature space. The feature space is divided into m subspaces and there is a document element used by independent subject database in every subspace:

$$\bigcup_{i=1}^{m} S_i = S \qquad (5)$$

According to the definition of optimal search theory, two functions are needed to treat every specified query vector q. One is initial probability distribution function corresponding to every independent subject database and the other is detection function. How to determine the initial probability distribution function of feature query vector is vital to the model. In the previous method, query is seen as to classify the documents. So it is natural to use the documents classification method to determine the relevance between query vector and every subject. Because of the performance of SVM in documents classification, SVM is selected as classifier. Some documents are training documents for all the subject databases. Documents belonging to the database are positive examples and others are negative examples. The classifier model is acquired by SVM, which can be used to determine other samples category. Every database corresponds to a certain classifier. There are m classifiers corresponding to m databases, which provide description information for these subject databases. When it is necessary to get the search results by query vector, classifying q should be used to get the decision function as well as

similarity. After making choices among these decision functions, approximate estimates of initial probability distribution function can be calculated. In some case, in order to facilitate computation, linear detection function can be used. After determining initial probability distribution and resource restrictions, search order and resource allocation scheme for every subject database can be calculated. IR strategy of standard vector space model is used. When selecting a subject database, query vector q is taken as a document. We need to calculate the similarity between q and other documents in database and select the document whose similarity is bigger than the threshold. If the allocated resource reaches the maximum, we stop search in this database. When the allocated resource runs out, the results from databases are sorted by similarity and results information can be shown for users.

2.3. LIBSVM Algorithm

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:

$$\min W(\alpha_1,\alpha_2) = \frac{1}{2}K(x_1,x_2)\alpha_1^2 + \frac{1}{2}K(x_2,x_2)\alpha_2^2 + y_1y_2K(x_1,x_2)\alpha_1\alpha_2$$
$$-(\alpha_1+\alpha_2) + y_1\alpha_1\sum_{i=3}^{l} y_i\alpha_i K(x_i,x_1) + y_2\alpha_2\sum_{i=3}^{l} y_i\alpha_i K(x_1,x_2)$$

$$s.t. \quad \alpha_1y_2 + \alpha_2y_2 = -\sum_{i=3}^{l} y_i\alpha_i = \text{constant}$$

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

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 $\sum_{i=1}^{l}\alpha_i y_i = 0$, the relationship between two variables should be: $\alpha_1y_1 + \alpha_2y_2 = \text{constant} = \alpha_1^{old} + \alpha_2^{old} = \alpha_1^{opt} + \alpha_2^{opt}$

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

The independent variables of the objective function are limited in a segment on flat (a1, a2). It turns into a one-dimensional problem and the solution procedure is as follows:

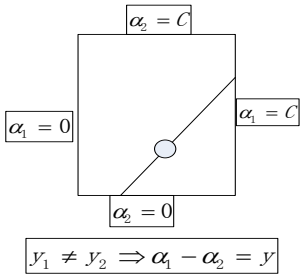

Figure 2. LIBSVM Algorithm

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.

3.1. Rewriting the Prediction Module

This paper only predicts custom kernel function and ignores other support such as oneclaSSSVM, SVR. Decision function can be implemented:

```
Int I;

Int len=model.len;

Double Sum=0.0;

For (I=0;I<len;I++)

{

Sum+=model.sv_soef[0][I]*x[I+len].value_Renamed;

}

Sum-=model.rho[0];

Renturn sum
```

After simply judging the decision function, we can get the final classification decision results. If the judgment value is positive, the result is positive classification. If the judgment value is negative, the result is negative classification.

3.2. Kernel Function Calculation

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. Experimental Results

4.1. Classification Performance Experiments of Sentence Collection Kernel Function

The number of documents is limited. 100 positive samples and 100 negative samples are selected for every classification. The number of correct classified positive examples (TP), the number of wrong classified positive examples (FP) and wrong classified negative examples (FN) are calculated by the classification results. Using these results can calculate the experiment performance evaluation indexes:

Accuracy: the proportion of correctly classified positive examples in all documents, $p = \frac{TP}{TP + FP}$.

Recall: the proportion of correctly classified positive examples in all positive examples, $r = \frac{TP}{TP + FN}$.

F: harmonious function of accuracy and recall $F1 = \frac{2pr}{p + r}$

Tow kernel functions introduced in the paper are used in the experiment. One is sentence collection kernel function proposed in the paper and the other is word sequence kernel function. The parameters in two functions are the same. The maximum sub sequence length is 2 and the recession factor is 0.9. Table 1 is the classification results of word sequence kernel function and Table 2 is the classification results of sentence sequence kernel function.

Table 1. The Classification Results of Word Sequence Kernel Function

categories	precision	recall	F-score
Earn	1	0.72	0.8308
Corn	0.7646	0.93	0.8435
Acp	0.6667	0.95	0.7940
ship	0.8876	0.88	0.8731
wheat	0.7639	0.84	0.80
average	0.81618	0.868	0.82797

Table 2. The Classification Results of Sentence Sequence Kernel Function

categories	precision	recall	F-score
Earn	0.9417	0.72	0.8308
Corn	0.929	0.929	0.929
Acp	0.9151	0.971	0.9418
ship	0.9418	0.971	0.9558
wheat	0.9577	0.913	0.933
average	0.93727	0.919	0.92626

Seen from Table 1 and Table 2, performance of sentence collected is superior to that of word sequence kernel function. It proves there are some advantages existing in sentence level kernel function proposed in the paper. Better classification results and shorter classification time should be reached through improvements.

4.2. 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:

$$pi(x) = \frac{Fi(x)}{\sum_{\forall j, Fi(x) \geq 0} Fj(x)}.$$

We will do retrieval in accordance with the order of $p(x)/n_i$ .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.

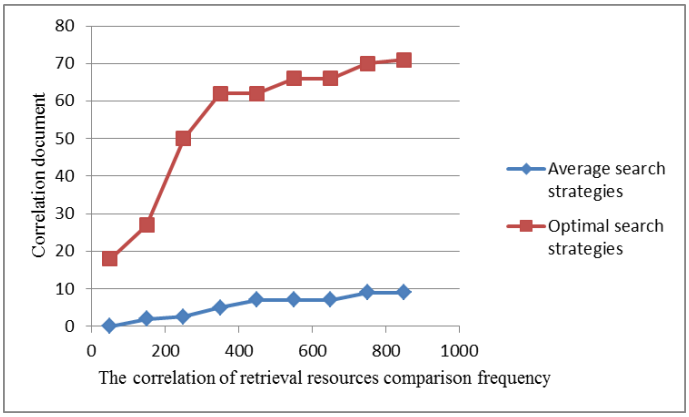

Figure 3. Optimization Search Strategy and Average Retrieval Strategy Comparison Chart

Seen from the Figure 3, the growth rate of average search is linear, which means it is necessary to search all the documents. While by optimal search only 15% of documents need to be searched. The advantage of optimal search strategy is to determine the position of relevant document in the database using the shortest time.

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)

