ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 141

International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2017)

Research on the Application of SVM Classification in Frontier Science

Mei Du* and Yue Zhu

Beijing Union University, College Management, Beijing, China Post 100101

*Corresponding author

Abstract—**The frontier field of science is very active in every subject, and it has a distinct mechanics. This law is highlighted in the number of research results in the rapid acceleration of growth, and has a strong academic radiation. Analysis of the frontier of scientific research results based on the characteristics, this paper selects the theme about Mars immigrants in space science, through the network to collect related theme papers, published time, keyword, cited information from the text, based on the statistical treatment of frequency, learning algorithms using support vector machine the field of quantitative research, tracking statement, development, in order to determine the identification of science frontier literature.**

Keywords-support vector machine (SVM); frontier science; text classification; machine learning

I. INTRODUCTION

As the most dynamic factor and the most powerful support of modern society, technology has an essential role on the process of economic construction. People have unlimited exploration of the unknown world while the resource of scientific exploration is limited. For this reason, whether the developed countries want to keep the dominant position on the area of research fronts or the developing countries want to achieve local penetration and leaping forward on the low starting point. All of these need the government to give key funding to the key science project which were selected cautiously. The most efficient and reasonable using of limited science resource is the most important way to have great development of country. In the strategic planning of making overall arrangements and promoting science process, more and more countries has realized that the selection of "priority areas" or "Key areas" is the key to accelerating the development of science and promoting the coordinate development of society economic.

Support Vector Machine (SVM) is a method of Machine Learning which is raised on the base of statistical theory, the method was put forth by AT&T Bell which was leaded by Vapnike [1], It has combined many theories such as Maximum Margin Hyperplane, Mercer core, convex quadratic programming and Slack variable, It could support an Function complexity Meaning depict which is independent of the dimensionality of the question. Using the Nonlinear Transformation function set ,it could also map vector to High dimensional feature space and produce Optimal hyperplane with the rule of support vector and the space-maximize of decision surface, then we could map the Linear decision boundary from High dimensional feature space to Nonlinear decision boundary in input space.[2]

SVM is commonly used to resolve binary classification problem and it doesn't has great performance to Multivariate classification problem. With the superiority in the working process of practical problems such as small sample，nonlinear，high dimensionality and Local minimum point, SVM could play an important role in the process of differentiation for "Research Front" literature. But in the process of solving support vector with SVM, it use the method of Quadratic programming and it related to the calculation of N-order matrix. N is the number of samples, it will take lots of machine resource and times to finish the storage computing of matrix. For this reason, it will be hard to finish while the amount of sample is very big.[3]

II. RESEARCH BACKGROUND

Foreign research has found that one notable aspect of research fronts is their potential to span traditional scientific disciplines. Potentially, for example, fronts that combine disciplines and challenge existing paradigms will have more difficulty being absorbed and may, in aggregate, presage paradigm shifts (Kuhn 1970). The progress of science is a result of the virtual and actual collaboration of thousands of scientists who, formally and informally, share their findings and build on one another's work. The research on explicit collaboration between scientists has emphasized the value of cross-company alliances, informal networks, and social capital (Gittelman 2003)[4]. The research area which was seen to have good potential or with great fund or maybe lead important business discover usually could attract most scientists interest.

A research front can be conceptualized as the convergence of scientific findings and social interests. New scientific findings may initiate the process of front formation by attracting the interest of more scientists who form social ties and generate more findings. The relevance and bearing of each new finding is continuously defined and evolved by the group. The foci of interests can be driven, of course, by the sources of funding as well as perceived scientific potential. This combined intellectual and social process is seen most vividly in the publications and citation patterns in science and technology. It is manifest in the emergence of clusters of highly cited papers representing the key scientific findings that are cited jointly. The authors of these cited and citing papers form what Derek Price has called an "invisible college".

In domestic research of research front area, the Chinese Academy of Sciences has mainly use the Incites database, the update speed of research front literature in this database is about 2-3 months. After the selection of research front literature and the related subject in Incites database, we can make batch download and lead them into excel table. High cited and circular references are the two important feature of research front literature.

In this paper we study small clusters of highly cited research, called ‘‘research fronts’’. We work to provide quantitative and qualitative support for continued, focused study of these areas as important for understanding the development of science and technology more broadly. These areas of intensive work are interesting to R&D laboratories looking for future innovation breakthroughs, venture capitalists looking to allocate investment, governments interested in promoting emerging science, and researchers hoping to work on promising topics.

III. Literature Collection and Data Cleaning

In this project, We collected literatures about “space science” which is published from 2005-2008 through the internet, all of these literatures was from mainstream journal or core journal. Google Scholar is the main resource this time. All of the literatures which were used in this experiment are already opened by the author and has no copyright issues. The original form of these literatures were pdf and we transferred them to text and stored in an csv document for the convenient of the follow-up working.

We need to extract a series of data such as Title, Time, Author, Nationality, Cited, Key Words, Website, Text Related, Articles and Reference.

The mission of extracting high-frequency words is finished in R environment with the package of TM and Snowball C. We made word frequency statistics, cleaning the results and finish normalization process for every literature. The main work include that replace the unnecessary sign with blank and delete the unnecessary words just like “a”, ”is” and “the” which is in a high frequency. With the calculation of the weight of high frequency words and combine with keywords which were given by the author we can product a document with 20 dimensional feature.

IV. SVM Classification Method

A. SVM Principle

Known train set T={(x1,y1),….,(xi,yi)},xi is the input quantity of sample, yi is the output quantity,

xi Є X={(Z1,Z2….Zn)},Z is the word frequency, yi Є Y={-1,1},i=1,…n, ”-1” indicate normal literature,”1” indicate the front literature.

Select proper core parameter, we select Gaussian radial basis function (RBF):

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2) \tag{1}$$

map the train set sample to Hibert space to get new relate train set: $\tilde{T}=\{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_i,y_i)\}$

Select proper error penalty parameter C to construct and solve Optimal problem

$$\begin{aligned}&\min_{w\in H,b\in R,\xi\in R^l}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i\\&\text{s.t.}\quad y_i((w\cdot\mathbf{x}_i)+b)\geq 1-\xi_i,\quad i=1,\dots,l,\\&\qquad\ \xi_i\geq 0,\quad i=1,\dots,l.\end{aligned} \tag{2}$$

ω in function(2) is the classification page of Hilbert space, ξi is the Slack variable, C is the penalty parameter, to construct Lagrange function, gain the dual problem of above problem is

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jy_iy_jK(\mathbf{x}_i,\mathbf{x}_j)-\sum_{i=1}^{l}\alpha_i\\&\text{s.t.}\quad \sum_{i=1}^{l}\alpha_iy_i=0,\\&\qquad 0\leq\alpha_i\leq C,\quad i=1,\dots,l.\end{aligned} \tag{3}$$

we can get the optimum solution $\boldsymbol{\alpha}^*=\left(\alpha_1^*,\dots\alpha_j^*\right)^T$

Select a Positive component $0<\alpha_j^*<C$ of $\boldsymbol{\alpha}^*$, to calculate threshold value

$$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*K(\mathbf{x}_i,\mathbf{x}_j) \tag{4}$$

Classification hyperplane decision function:

$$f(\mathbf{x})=\operatorname{sgn}(\sum_{i=1}^{l}y_i\alpha_i^*K(\mathbf{x},\mathbf{x}_i)+b^*) \tag{5}$$

B. Parameter Optimization of SVM Model

After the existence of train sample and the definition of RBF core function，the error penalty factor C parameter and core parameter must has to be confirmed. The research of Vapnike has shown that the performance of SVM have little related with the selected core function but the error penalty factor and core parameter, they are the main factor which effect the performance of SVM[1].

The function of error penalty factor C parameter is adjusting the portion of confidence range and empirical risk of SVM, it could get an compromise value to enhance the generalization ability. The core parameter is used to reflect the feature of data distribution to confirm the range of local field. Big value of core parameter means low variance. So the value of error penalty factor C parameter and core parameter heavily determined the learning ability and generalization ability [3].

In order to express the learning ability and generalization ability of SVM quantitatively, in the evaluation of algorithm we select n-fold cross-validation. First of all, we random put L number of sample points into N number of subset with no cross. We select on of them as the train set and other of them as the sample test set, then we could make n time of different train test and get N number of different Pi value. Through the calculation of the average value of classification accuracy we can judge the performance of classifiers quantitatively.

V. Experiment Data Analysis

In this experiment, we tried to use 50 literatures which was mixed by the proportion of 5:5 of "Research Fronts" literature and normal. Nevertheless, we insert a new field "Type" in the feature table to identify whether the literature belongs to research fronts. The Type will be "1" if the literature belongs, it will be "-1" if it is not. With the proportion of 8:2, we let the machine to make random screening and separate the train set(80%) and test set(20%).

Using the "E1071" package in R software, we could finish the work of SVM which include parameter optimization, training the data sample set, date test and getting the average accuracy rate. The parameter "gamma" is the "y" in formula 1.With the using of tune.svm function, we could make 10-fold cross validation and get the gamma and cost parameter with best classification performance. The effect of classification of gamma and cost has been listed in table 1. Among the table 1, the value of No. 1-4 is based on the sample set with the proportion of 5:5 of research front literature and normal literature. The value of No. 5 in table 1 is based on the proportion of 3:7.

TABLE I. SAMPLE OF SVM PARAMETER

No.	Gamma	Cost	Error
1	0.0625	2	0.3912
2	0.0625	1	0.3583
3	0.0625	4	0.3333
4	0.00390625	8	0.3416
5	0.00390625	0.003906	0.275

According to the value of error we could select the parameter, the option value with low error rate is（c=4，g=0.0625）.We use this parameter to make many times of sampling training and make prediction of test set. Through the comparison of the value of Type we could gain and calculate the accuracy rate, then get the average value in the final. With the comparison of data which is gained by the empirical parameter (c=10, g=0.1) shown in table 2. We can find out that the classification performance of classifier has gotten distinct development after the optimization of parameters.

TABLE II. ACCURACY COMPARISION

	With Empirical parameters	With Optimal parameters
average	55%	66%

Chaomei Chen tested the ETD results, the average precision and recall are 0.317 and 0.359 respectively[5]. The average precision and recall of our projects are 0.5 and 0.8, respectively. The purpose of this method is to develop a learning model to get a better accuracy and recall rate, and finally the results of the system is judged by the domain experts to do the final judgment.

VI. Conclusion

Using the method of SVM classification, this paper puts forward a feasible method to judge the frontier of science. In the process of dealing with literature of research fronts, With the using of the TF-IDF method to extracting the feature of these literatures，We select the high-mark feature words to combine with the key words in the literature to product an high-dimensional classification sample data set which was consist of eigenvectors. Using the Gauss radial basis function algorithm in SVM and combine with N-fold cross test method, we made classification experiment for the selection of "Research fronts" literature. The result of experiment has shown that supervised Learning Support Vector Machine (SVM) has better performance in the aspect of literature selection. In the following study, we will try to test the larger sample set and the higher dimensional data to evaluate the role of SVM in the selection of the frontier science.

Reference

[1] Corinna Cortes,Vladimir Vapnike.Support-vector networks, Proc.Of the 25th Machine Learning,1995, 20:273-297

[2] Ma erlun,Zheng Yannan. On the Parametr Optimization of SVM Classification Method: taking the application in Yellow River Estuary Wetland for Example. Value Engineering .2014

[3] Zhang yi,liu yijian,luo yuan. A Parameter Optimized C-SVM approach for EEG classification and its application Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition).vol.26 No.1.2014.1

[4] S. Phineas Upham, Henry Small. Emerging research fronts in science and technology: patterns of new knowledge development. Scientometrics (2010) 83:15–38. DOI 10.1007/s11192-009-0051-9

[5] Chaomei Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 2006, (3)

[6] Challenges and opportunities with big data. Alexandros Labrinidis,H. V. Jagadish. Proceedings of the VLDB Endowment . 2012

[7] Local-learning-based feature selection for high-dimensional data analysis. Sun, Yijun, Todorovic, Sinisa,Goodison, Steve. IEEE Transactions on Pattern Analysis and Machine Intelligence . 2010