Efficient and Intelligent Information Retrieval using Support Vector machine (SVM)

Monika Arora, Uma Kanjilal, Dinesh Varshney

Abstract—The information access is the rich data available for information retrieval, evolved to provide principal approaches or strategies for searching and browsing. The search has become the leading paradigm to find the information on World Wide Web. For building the successful information retrieval, there are a number of prospects that arise at the different levels where techniques can be considered. The present investigations explore the Support vector machine identified its level and classifies the documents on web. This paper attempts to develop a model for the efficient and intelligent retrieval. This paper attempts to propose the implement model for efficient and intelligent retrieval. In model it attempted to figure out the important factors for the successful efficient and intelligent retrieval. The proposed model is designed to collate all the differing views on information retrieval so as to construct a holistic theoretical which is considered to be the source of a system. This paper considers the application of Support Vector Machine for designing the model for efficient and intelligent retrieval. This will also consider a proposed model for developing successful retrieval.

Index Terms— Information Retrieval, Web Information Retrieval, Support vector machine

I. INTRODUCTION

SVMs (Support Vector Machines) are a technique for data classification [1]. SVM is considered easier to implement in the area of than neural networks too. This technique is used to achieve the higher accuracy in the process of retrieval. It also takes the challenges to handle a typical problem. The SVM working is as follows:This task firstly involves separating data/document into training from the testing tests. Each instance or records of the training set contains a target value and several attributes (feature or observed variable). The target of SVM is helping to produce a model or goal, which is based on training data and also predicts the target value of the test data. This training set enables the dataset the label pairs and classify it into the group for categorization. The support vector machine (SVM) is applied to achieve the optimized solution for the training set of data [2, 3]. The training vectors are mapped into a higher dimensional space by the use of function.

Manuscript received January 10, 2012.

Monika Arora, Department of IT/IS, Apeejay School of Management, Dwarka Institutional Area, New Delhi, India

Uma Kanjilal, Department of Library and Information Science, Indira Gandhi Open University ,Maidan Garhi, New Delhi-110 068,India

Dinesh Varshney, School of Physics, Devi Ahilya University, Khandwa road Campus, Indore, M. P. India and Multimedia regional Centre, Madhya Pradesh Bhoj (open) University, Khandwa road Campus, Indore- 452001, M. P. India

II. SVM AND ITS ATTRIBUTES

SVM uses a linear separating hyper plane. This function uses the maximal margin in this higher dimensional space. There is the penalty parameter which uses for the error term. Because of using the kernel values, it usually depends on the product of the inner feature vectors. Furthermore; it is called the kernel function. The generally used kernels are as follows: For 1, the training data is said to be linearly separable, when data can be separated at two hyper planes of the margins in a way that there are no points between them and then try to maximize their distance. This is the simplest kernel and shows good performance for linearly separable data. For 2, the polynomial kernel $K(x,y) = (\gamma x^T y + r)^{\partial}, \gamma > 0$ is a non linear kernel, used for large set of attributes values and polynomial kernels ,where the kernel values may go to infinity. That are linearly dependent on n dimensions, the kernel function of order n can be used to transform them into linearly independent vectors on those n dimensions. Once they are transformed into the dimension space, they become linearly separable. For 3, the radial basis function (RBF): $K(x,y) = \exp(-\gamma\|x\text{-}y\|^2), \gamma > 0$ "RBF kernel get better results than the linear kernels", the data set is linear separable (fully, i.e. with 100 % accuracy). The RBF is most popular in choosing of kernel types in Support Vector Machines (SVM). This is mainly because it is localized and has finite responses across the entire range of the real x-axis. This kernel is basically suited best to deal with data which have a class-conditional probability distribution function approaching the Gaussian distribution. It maps such data into a different space where the data becomes linearly separable [4].

To actually visualize this, it is convenient to observe that the kernel (which is exponential in nature) can be expanded into an infinite series, thus giving rise to an infinite-dimension polynomial kernel: each of these polynomial kernels will be able to transform certain dimensions to make them linearly separable. Naturally, one would expect the RBF kernel to perform much better than either the Linear or the Polynomial kernel. However, this kernel is difficult to design, in the sense that it is difficult to arrive at an optimum σ and choose the corresponding C that works best for a given problem. The fact that certain combinations of σ and C make the SVM highly sensitive to training data also contributes to the error rate of the RBF-based SVM.

One of the advantages of the RBF kernel is that given the kernel, the weights, the number of support vectors Ns and the support vectors si are all automatically obtained as part of the training procedure, i.e, they need not be specified by the training mechanism. The RBF based kernel is more efficiently used for classification [5]. For 4, sigmoid kernel $K(x,y) = Tanh(\gamma x^T y + r)$ is not as efficient for classification as are the other three. Indeed, one of the fundamental requirements on a valid kernel is that it must satisfy Mercer's theorem, and that requires that the kernel be positive definite. However, the Sigmoid kernel is not necessarily positive definite, and the parameters κ and δ must be properly chosen. In cases where the kernel is not positive definite, the results may be drastically wrong, so much so that the SVM performs worse than chance.

A noteworthy point is that for a certain range of values of κ and δ, the kernel behaves as a linear kernel, while for a certain other range of values of the same parameters, it takes the form of a RBF kernel [4]. SVM case can handle the classification problem. Thus, in a way, it is an extension of the linear kernel, in that it gives the crucial transformation to "enable independence" among the training samples. The performance of this kernel is expected to be around the same as that of the linear kernel, since the principle behind the two is the same and the transformation is to just take them to different space. However, the performance does depend on the order p of the polynomial, since how well the data becomes separable depends on it.

III. ANALYZING DATA SET PREPARATION FOR SVM TESTING

In the experiment, the test data evaluates the retrieval performance of various relevance feedback methods on document based IR. A category and subcategories are assumed for the classification of the document of the individuals. The details of the test data are described in Figure 6.4.1. The dataset parameters are first picked from the database randomly, and this category is assumed to be the user's query target. The individual can be male or female based on the category IT for 1, HR for 2, Finance 3 and Marketing for 4.

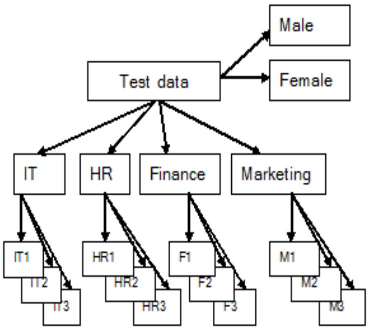


Figure 1: Data set formation

The training dataset is maintained for the class/category "IT" mostly concentrate on IT1, IT2, and IT3 and may have interest on any other subject area may be HR1 and so on. The dataset are maintained for all the different interest area persons. The system then improves retrieval results by relevance feedback. In each iteration of the relevance feedback process, given documents are picked from the database and labelled as either relevant or non-relevant based on the ground truth of the database. For the first iteration, two relevant documents and three non-relevant documents are randomly picked, and all methods are run based on the same set of initial data points. Table 1 presents a real-world example. These data structure are described in the table and also the sequence of the parameters. The dataset focuses on the 14 parameters as follows.

Table: 1 Dataset

No	Parameter	Description
1	Class / Category	1,2,3,4
2	Sex(M-1/F-2)	1 or 2
3	Database	0 or 1
4	Networking	0 or 1
5	Search engine	0 or 1
6	Compensation	0 or 1
7	HRM	0 or 1
8	Organisational Behaviour	0 or 1
9	Branding	0 or 1
10	Retail	0 or 1
11	Supply chain management	0 or 1
12	Mutual Funds	0 or 1
13	Derivatives	0 or 1
14	Census	0 or 1

The dataset consist of the mainly of four class/category such as IT, HR, Finance and Marketing. The parameter second classifies the data as male and the female. The parameters for 3-14 define the category of Yes/No i.e. 1/0. It specifies the area as 3-5 for IT category 6-8 for HR category, for 9-11 for Marketing category and 12-14 for the finance category. The model is prepared by considering the 100% accuracy using the SVM-Train and SVM-Predict functions. For the iterations afterward, each method selects give document based on their own display set selection algorithm. The accuracy for the set of training data used for model is 100%.

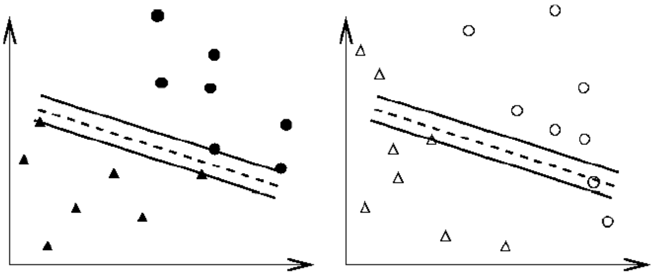
For the SVM-based techniques in the experiment, we implement the algorithms by modifying the codes in the libsvm library [6]. It notices that the experimental settings are important to impact on the evaluation results. To enable an objective measure of performance without bias, it chooses the same kernel and parameters for all SVM-based methods. In order to select the best kernel function for the current dataset, we performed an experiment to evaluate the performance of different kernels. The kernel functions involved in the experiment are listed on Table 1. The datasets are with four categories (4-Cat). Each category includes 4 documents belonging to a same semantic class. We evaluate the performance of different kernel functions by measuring their average precision on the top 16 retrieval result. This helps to create the model for the training dataset.

The best parameter might be affected by the size of data set as the data is in the large distribution it may spread and accuracy may effect. But as in practice it is obtained from cross-validation defining that data set should be suitable for the whole training set. The scaling values can be used for the avoiding the attributes in higher numeric ranges is dominating those in lower numeric ranges. Also, it is used to avoid numerical difficulties during the calculation. The kernel values depend on the usage of parameter by inner products , e.g. the feature vector such as linear kernel and the polynomial kernel, large attribute values might cause numerical problems. The same set if method with scale should be use for training data and also the testing data. For example, the scaling of the first attribute of training data also from the testing data.

IV. MODEL SELECTION

There are four common kernels available to use. The RBF kernel is more popular in usage because it uses the penalty parameter “C” for the parameter of the kernel chosen. Also the RBF kernel is rationally the first choice because it uses the nonlinearly maps samples and that to a higher dimensional space. But linear kernel provides up with the relationship between the class labels and attributes. The linear kernel is one of special case of RBF, which uses with a penalty parameter C .The performance as the RBF kernel with some parameters (C; γ) are same [7]. Also, the sigmoid kernel behaves like RBF for some of the parameters [4]. It is a polynomial kernel which considers more hyper-parameters than the RBF kernel.

Cross validation and Grid search are considered as two different parameters for an RBF kernel: C and γ for any problem. The model creation uses a parameter for selectionfor developing the best model. The main objective is to make out the best suited values for (C; γ) so that the classifier can accurately predict on the unknown data as testing data. The common approach applied is to divide the data set into two parts. The prediction accuracy is obtained from that unknown set to more precisely that reflect the performance on categorize an independent set of data. The cross-validation process called the improved version also uses a that is based on v-fold cross-validation. The data set in cross-validation process divide the training set into equal size v subsets. Also the one subset is tested by using the classifier qualified on the left over (v – 1) subsets. The whole training set is predicted or tested once so the cross-validation accuracy in terms of the percentage of data which are correctly classified. The circles and triangles which are filled are the training data set while hollow circles and triangles from the testing data. The testing accuracy of the classifier in Figures 2(a) and (b) is not good since it over fits the training data. The training and testing data used on the training and confirmation sets in cross-validation is not appropriate. It gives better testing accuracy as well as cross-validation using Grid search. It recommend a grid-search on C and using cross-validation to various pairs of (C; γ) values are applied and achieve the best cross-validation accuracy is picked. It found that trying exponentially rising sequences of C and γ in a practical method.s accuracy. Also, there are several advanced methods used similar to the cross-validation rate. Also, there are two motivations for the application (a) Training data and an over fitting classifier (b) Applying an over fitting classifier on testing data



(a) Training data and an overfitting classifier (b) Applying an overfitting classifier on testing

Figure 2 (a) and (b) training data with over fitting classifier

The training data are applied by considering the default parameters for c and γ where s =2 and t=50. The above approach works well for problems with hundreds or more data points. For huge data sets a realistic approach is to at random choose a subset of the data set, by conducting the grid-search on them. The better-region-only grid-search ia applied to the complete data set. The testing data are prepared where instead of the class/category no, it is considered as serial no. The model, which was prepared for the 16 records are applied for the testing data for 50 records. The data according to the specification the all the 50 participants are mapped under 4 categories based on their parameter values. This technique indicated the classification at the large scale as well. Also this dataset is specifically for 4 classes it may be consider for thousands of the classes based on the data available. These guide our practice data for the process works as well as for data which do not have many features. Assume that if there are thousands of attributes, there may be data is useful for the need to prefer a subset of them before giving the data to SVM.

V. AN EXPERIMENT RESULT ON SAMPLE DATA

In this were the accuracy by the proposed procedure is compared. The experiments take cares of three problems mentioned in Table below. The software LIBSVM is applied to achieve the accuracy [8]. This accuracy is by direct by training and testing data. Secondly, it should also show the difference in accuracy with and without scaling that is not considered for this test data. The other parameters of training set attributes are considered and tested for the creation of model are able to restore them while the testing set. Also Thirdly, the accuracy for the (perfect model selection is he proposed procedure presented for the accuracy. Finally, tool in LIBSVM is demonstrate the use in which the whole dataset is applied automatically. Note same parameter selection tool will be applied for the huge set of data for the classification before retrieval.

The commands applied is as follows:_ Original sets with default parameters

```
$ ./svm-train test  test.model
$ ./svm-predict test.txt testmodel test.predict
! Accuracy = 100 % ( perfect model)
_ Original sets with default parameters
$. /svm-predict test3.txt testmodel test3.predict
```

The test3.predict provides the decision where actually the serial data should lie upon. The predicted classification of the data is described as follows.

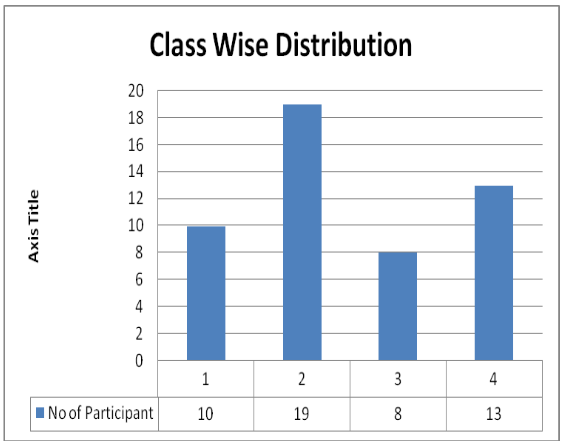


Figure 3: Class wise distribution

The most popular class is 2 with respect to the specialisation and also the linked people to gain the popularity. The training dataset used for the model creation defines a criteria where a set of the people into certain category. There overlapping in the stream such as an HR person also has interest in IT or vice versa. This model creates a understanding and categorise according to the training data. The model created based on the training data will be applied to the testing data for results.

The result in shown in figure 3 is homogeneous distributed across the 4 categories. It means that it is the best retrieval results produced by the same relevance feedback system for the four datasets, as data distributed in the figure. From the experiment, it draws the following Observation that the retrieval results of the included for relevance feedback techniques have been improved set after iteration. Thus, it shows that the retrieval performance of document based IR can be improved with relevance feedback techniques. Thus, these techniques are only able to retrieve the relevant images in a local area, and fail to retrieve other relevant images and improve the retrieval performance afterward.

VI. CONCLUSION

The interactions between the parameters are widely considered. These parameter or documents create a relationship with a class. This class identifies the dataset and categorise it into a particular class for seeking the details. The interactions with documents are due to the information-seeking parameters which are studied. For the real-world the information retrieval, the data representation in categories may be an important step for evaluating the relevance feedback algorithms. We extract three different features to represent the data i.e. its attributes

It is investigated that SVM-based relevance feedback techniques can be used for solving the relevance feedback problems in document based IR. The imbalanced dataset problem in relevance feedback and proposed a novel relevance feedback technique with Support Vector Machine. The advantages of our proposed techniques are explained and demonstrated compared with traditional approaches. The experiments may be applied to the training data and also real-dataset. The experimental results demonstrate that SVM based relevance feedback algorithm is effective and promising for improving the retrieval performance in document based information retrieval.

REFERENCES

1. Quek C. Y. Classification of World Wide Web Documents.” 1997. Senior Honors Thesis, Carnegie Mellon University 1997.
2. Boser B. E., Guyon I., and Vapnik V. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144-152. ACM Press, 1992.
3. Cortes C. and Vapnik V. Support-vector network. Machine Learning, vol 20 pp 273-297,1995.
4. Lin H.-T. and Lin C.-J., “A study on sigmoid kernels for SVM and the training of non-PSD Kernels by SMO-type methods”, March 2003.
5. Chen S., Cowan C. F. N., and Grant P. M., "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks", IEEE Transactions on Neural Networks, Vol 2, No 2 (Mar) , 1991.
6. Cross, R. A bird’s-eye view: Using social network analysis to improve knowledge creation and sharing, IBM Corporation , 2002
7. Keerthi, S.S., Lin, C.-J., 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 15 (7), 1667–1689.
8. Chang C.-C. and Lin C.-J. LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

AUTHORS PROFILE

Ms Monika Arora, is Assistant Professor and Head of Department of I.T. She had over 11 years of experience in Industry, Teaching and Research. She is author of over 15 papers in national, International conference, national and International Journals. She played a role as reviewer in International conference of Data Management and International Journal of Interscience. She is also co-chair and chair of various conferences in Delhi Region. Her research Area is in the field of Intelligent and Efficient data retrieval in Semantic Web, Search engines, Social Network Analysis, Database System, Knowledge Management Systems.

Prof. Uma Kanjilal, is Professor of Library and Information Science and Director of School of Social Sciences at Indira Gandhi National Open University, New Delhi. She has done Ph. D. in Library and Information Science from Jiwaji University, Gwalior and PG Diploma in Distance Education from IGNOU. Prof. Kanjilal is at present coordinating major projects in the University and at national level like eGyanKosh, IGNOU FlexiLearn portal and Sakshat. The eGyanKosh Project won the Manthan South Asia 2008 award for best e-content for development in the e-education category. She has an experience of more than twenty years of working in Open and Distance Learning System. Prof. Kanjilal was a Fulbright scholar in 1999 – 2000 in University of Illinois, Urbana Champaign, USA where she worked on multimedia courseware development. Prof. Kanjilal has number of publications in national and international journals . She authored one book and co-edited two books.

Dr. Dinesh Varshney is Professor-School of Physcis, Devi Ahilya University, Indore (M.P.), India. He is a recipient of Dr. Kailash Nath Katju Award for outstanding contributions in Science for 2007 and Best research Scientist award of the University for 2009. He is author of over 350 research papers International journals and supervised 18 students for their Ph. D. He is engaged in application of information theory as quality improvement for audio, video and data transfer for Edusat establishment, Voice signal compression and spectrum analysis, efficient and intelligent data retrieval for Semantic web, Educational performance indicator as a data mining probe for higher education.