Investigation into Biomedical Literature Classification using Support Vector Machines

Nalini Polavarapu*
School of Biology
Georgia Institute of Technology
Atlanta GA 30332
gtg769e@mail.gatech.edu

Shamkant B.Navathe
College of Computing
Georgia Institute of Technology
Atlanta GA 30332
sham@cc.gatech.edu

Ramprasad Ramnarayanan, Abrar ul Haque, Saurav Sahay, Ying Liu
(ramji, abrar, ssahay, yingliu@cc.gatech.edu)
College of Computing
Georgia Institute of Technology
Atlanta GA 30332.

*Corresponding Author

Abstract

Specific topic search in the PubMed Database, one of the most important information resources for scientific community, presents a big challenge to the users. The researcher typically formulates boolean queries followed by scanning the retrieved records for relevance, which is very time consuming and error prone. We applied Support Vector Machines (SVM) for automatic retrieval of PubMed articles related to Human genome epidemiological research at CDC (Center for disease Control and Prevention). In this paper, we discuss various investigations into biomedical literature classification and analyze the effect of various issues related to the choice of keywords, training sets, kernel functions and parameters for the SVM technique. We report on the various factors above to show that SVM is a viable technique for automatic classification of biomedical literature into topics of interest such as epidemiology, cancer, birth defects etc. In all our experiments, we achieved high values of PPV, sensitivity and specificity.

1. Introduction

PubMed (Medline) is a huge repository of publicly available scientific literature. Currently, new data is being added to it at the rate of over 1500 abstracts per week. Most biomedical researchers want to access PubMed with specific goals based on the areas of interest. The ability to efficiently review the available literature is essential for rapid progress of research in scientific community.

The traditional literature database search involves the use of simple boolean queries, formulated using certain frequently used functionally important keywords the researcher is familiar with, followed by manual scanning of the retrieved records for relevance, which is time consuming, incomplete and error prone. Even with the formulation of complex queries, by a researcher over several years by continually adding new keywords encountered to the original query, the increase in the sensitivity of the searches is only marginal. Therefore, there is a pressing need for the development of automated literature mining techniques that can help the researchers to effectively harvest the heap of the knowledge available in the scientific literature.

Supervised algorithms such as Support Vector Machines (SVM) can be used for classification of biomedical literature into user defined categories. SVM is a machine learning algorithm that performs binary and multiway classification (pattern recognition) of the data into user defined categories [1]. Support Vector Machines maps non-linearly separable training vectors in input space to linearly separable higher dimensional feature space and finds a separating hyper plane with maximal margin in that higher dimensional space.

SVM has been widely used in text classification. The SVM method has been introduced in text classification by Joachims [2] and subsequently used in [3-10]. Joachims [2] applied SVM to text classification and reported that SVM yielded lower error than many other classification techniques. Yang and Liu [9] compared different classifiers, Naive Bayes (NB), kNN, and SVM and found that SVM performed at least as well as all other classifiers they tried. Dumais et al. [5] tested a novel algorithm for training SVM text classifiers and showed that this brings about training speeds comparable to computationally easy methods such as Rocchio. Han et al. (2003) [11] applied SVM for automatically extracting Medline citations of biomedical articles and reranking them according to their relevance to a certain biomedical property difficult to express as PubMed query. They reported that major improvements were achieved in reranking citations with respect to protein disorder-function relationships where the average relative ranking of a relevant citation was improved significantly.

In this paper, we report the results of application of SVM for incorporation of Human Genome Epidemiology (HuGE) relevant articles from PubMed database into CDC's HuGENet™ (http://www.cdc.gov/genomics/hugenet/) published literature database. Although the present study is limited to classifying the epidemiology related articles, the method described here has a wider applicability and can be used for classifying the articles by disease, by topic or even by domain of expertise. We also report the results of some preliminary analysis for multi-way classification using SVM with different types of cancer abstracts extracted from PubMed.

2. Methods

2.1 Human screening of PubMed

New abstracts appearing in the PubMed database are currently being manually categorized as HuGE and populated into the CDC's HuGENet™ database by a human expert using a complex search query. The complex query CDC uses for screening the PubMed database was developed over four years by iteratively adding the new HuGE relevant keywords encountered that were absent in the original query. As of March, 2004 it consisted of 98 different keywords combined with boolean operators. It is important to note here that after manual processing by human expert, on average, only 5 - 10% of the articles retrieved from the PubMed database by the complex query are HuGE relevant and are being added to the HuGENet™ database (Figure 1).

1848 Total number of articles captured by the complex query

1544 Excluded based on reading titles

304 Selected for further reading based on reading titles

Manual Reading of full abstract of the above selected articles gives following:

174 HuGE articles – included in HuGENet database

130 NonHuGE articles – Not included in HuGENet database

Figure 1: Distribution of PubMed articles retrieved using the complex query: Weekly update of April 1, 2004.

2.2 Feature Selection

Different approaches were used for selecting the keywords to constitute the feature vectors for SVM. The keywords were generated using two different weighing schemes, Z-Score and TFIDF (Term Frequency x Inverse Document Frequency). The weighing schemes estimate the significance of words by comparing the frequency of words in a test set (HuGE) of abstracts with their frequency in a background set of abstracts. The background sets of abstracts were used to build a hash table of words and their respective statistics for comparison with the corresponding words in the training and test sets. The abstracts present in the PubMed database from 1969 till 2004 were used as the background set. Porter stemming algorithm [12] was used to truncate suffixes and trailing numerals so that words having the same root (e.g.,epidemic, epidemics, epidemiology, epidemiological etc.) are collapsed to the same word for frequency counting. The stop word list customized in the previous study [13-14] by adding biological methodology words to an online dictionary of 22,205 words (http://ftp.std.com/obi/Dictionary/dict), abbreviated PD+ was used to filter out non-scientific english words that carry low domain-specific information content The words with the high Z-Score and TFIDF values were selected as features for the SVM. The formulas used for the calculating the Z-Score and TFIDF values are given below.

2.3 Two Weighting Schemes for Keywords Extraction:

2.3.1 TFIDF method (Term Frequency x Inverse Document Frequency). The standard TFIDF function was used [15]. TFIDF combines term frequency (TF), which measures the number of times a word occurs in the HuGE's set of abstracts (reflecting the importance of the word to the HuGE), and inverse document frequency (IDF), which measures the information content of a word – its rarity across all the abstracts in the background set. The inverse document frequency (IDF) is calculated as:

$$idf^a = \log \frac{N}{df^a}$$

where idf^a denotes the inverse document frequency of word *a* in the background set; df^a denotes the number of documents (abstracts) in the background set in which word *a* occurs; and N is the total number of abstracts in the background set. TFIDF is defined as:

$$tfidf_H{}^a = tf_H{}^a \times idf^a$$

$tfidf_H{}^a$ denotes the weight of the word *a* to the HuGE abstracts *H*; $tf_H{}^a$ the number of times word *a* occurs in the set of HuGE abstracts *H*.

2.3.2 Z-Score method. The statistical formula used for calculating the Z-Score is given by

$$z = \frac{p_1 - p_2}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where *p1* is the probability with which a given word occurs in the HuGE abstracts, *p2* is the probability with which that particular word occurs in the background set, *n1* is the total word count in the HuGE abstracts and *n2* is the total word count in the background set. The formula used for the calculation of *p* and *q* are given below.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \qquad q = 1 - p$$

2.4 Our approach using SVM

After extracting keywords using Z-Score and TFIDF, eight different top ranked sets with varying number of keywords were used as features for SVM. They were:

1) Z-Score top 100 keywords

2) Z-Score top 500 keywords

3) Z-Score all 784 keywords

4) TFIDF top 100 keywords

5) TFIDF top 500 keywords

6) TFIDF top 750 keywords

7) TFIDF top 1010 keywords

8) TFIDF top 2010 keywords

The training and test sets were converted into an abstract vs keyword matrix, a format readable by SVM^{light} software [16]. In conversion of the abstracts in the training set into an abstract vs keyword matrix, +1 was used to denote the class label for positive (HuGE) abstracts and –l was used to denote the class label for negative abstracts (Non HuGE). The presence and absence of the keywords were represented by 1 and 0 respectively. The abstracts in the test sets were also converted to the similar format except for the class label, which is '0' for all the abstracts. Unless otherwise mentioned, the SVM was tested with linear kernel using default values of 'γ' and 'c' (cost) parameters.

2.5 Design of experiments to test SVM:

2.5.1 Training set. The 11000 abstracts present in the CDC's HuGENet™ database, as of March, 2004, were used as the positive training set. The Non HuGE abstracts were obtained by searching the PubMed database; using the complex query for the abstracts that appeared in it between 2000 and 2004; followed by removing the HuGE abstracts from them. A total of 11000 abstracts were then randomly selected from the Non HuGE abstracts and were used as the negative training set for the SVM. Two sets of training sets were compared, one consisting of equal number of positive and negative abstracts (11000 positives and 11000 negatives) and the other consisting of twice the number of positives over negative abstracts (11000 positives and 5600 negatives).

2.5.2 Test Set. The abstracts retrieved from the PubMed database using the complex query during four different weeks, Feb 12, 2004, Apr 1, 2004, Apr 8, 2004 and Jun 3, 2004 were used as the test sets for the SVM.

2.5.3 Sensitivity, Specificity and Positive Predictive Value. Three different metrics were used to evaluate the performance of SVM in classifying the abstracts. The classification of the abstracts by human expert was used as the "gold standard" against which the SVM classifications were evaluated by Sensitivity (also referred to as Recall), Specificity, and Positive Predictive Value (PPV) (also referred to as Precision).

Sensitivity (Sn) = TP/(TP + FN)

Specificity (Sp) = TN/(FP + TN)

PPV = TP/(TP + FP)

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

3. Results and Discussions

3.1 Performance of SVM with different sets of keywords

The performance of SVM with the eight different sets of keywords mentioned above was compared. Training set containing equal number of +ve and –ve abstracts was used. The keyword sets, TFIDF top 2010 and Z-Score all 784, performed better compared to the other sets. So these two sets were selected as features in the remaining of our comparisons. Best results were obtained (93.6 % average Sensitivity, 91.45% average specificity, 50% average PPV) using the top 2010 keywords obtained from TFIDF. (Table 1 and Figure 2).

Keyword sets	Feb12			Apr1			Apr8			Jun3		
	Sn	Sp	PPV	Sn	Sp	PPV	Sn	Sp	PPV	Sn	Sp	PPV
TFIDF top 100	88	90	36	88.5	88.6	44.7	89.5	91.3	53.6	92.6	89.7	44.4
TFIDF top 500	92.5	92.5	45	91.3	90.6	50.4	89	93	58.8	94	92	50.7
TFIDF top 750	94.5	92	44	92.5	90.5	50.5	90	93.7	61.7	93.3	92	50.7
TFIDF top 1010	94.5	92	44	92.5	90.5	50.4	90.1	93.4	60	93.3	92.2	51.5
TFIDF top 2010	97.2	91.8	43.8	92	91	50.3	89.5	93	59.3	96	90	47
Z-Score top 100	92.5	92	42.7	87.3	89	45	85.5	91	51.3	94	90.8	47.5
Z-Score top 500	95	74	19	93	69.7	24.2	94.2	73.2	28.2	94	73	23.6
Z-Score all 784	95	92.6	45	82.5	91.3	50	85.5	93.1	58.2	93.3	91.2	48.6

Table1: Results from SVM using different sets of keywords

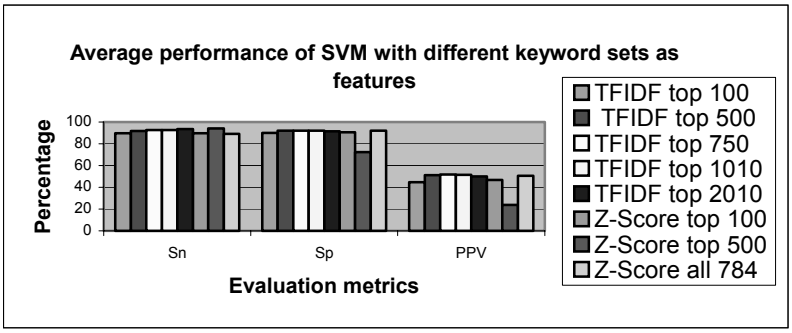

Figure 2: Average performance of SVM with different keyword sets as features

3.2 Performance of SVM with a bias for positive training sets

If a user is interested in sensitivity as the overriding criterion for classification, it is possible to influence the results by biasing towards the positive examples over negative ones. First, we can control the relative sizes of the two training sets. We compared the performance of SVM with two training sets using TFIDF top 2010 keywords as features. With twice the number of positives than the negatives in the training set, the sensitivity of the SVM increased consistently for each of the four sets, while reducing the specificity and Positive Predictive Value (PPV) (Table 2).

	Feb12			Apr1			Apr8			Jun3		
Training Set	Sn	Sp	PPV	Sn	Sp	PPV	Sn	Sp	PPV	Sn	Sp	PPV
Training set 11400 +ve and 11300 -ve	97	92	44	92	91	52	90	93	59	96	90	47
Training set 11400 +ve 5300 –ve	98	87	33	96	85	39	94	88	46	97	87	39

Table 2: SVM classifications with different training sets

Second, we can weigh the positives heavily over negatives in training the SVM. We tried weighing the positives over negatives by a factor of two, four and eight on a training set of equal positives and negatives and found that the sensitivity results consistently improved at the cost of Specificity and PPV(For Apr1, test set, the Sensitivity values are : 89.08, 97.13, 98.85 and 99.43). These results indicate that the outcome of the classification can be changed in response to the user's need by tweaking the training set or by assigning different weights to the training sets.

3.3 Union of results using keywords based on TFIDF and Z-Score methods

The performance of SVM was estimated by taking the union of the results obtained from using TFIDF top 2010 and Z-Score all 784 keyword sets. Briefly, the union of results is done as follows. If the SVM identified an article as false positive with both the keyword sets then it was considered as false positive. On the other hand if the SVM disagreed with the keyword sets i.e. classified the article as true positive with one keyword set and as false positive with the other set, then the article was considered as the true positive. The same rule applies to true negatives and false negatives (the article was considered as false negative if the SVM classified it as false negative with both the keyword sets. If there was a discrepancy in the classification with the two keyword sets then the article was considered as true negative) This was done to minimize the false positive and false negative error rates thereby increasing the sensitivity and specificity of the SVM. (Table 3 and Figure 3)

	Feb12			Apr1			Apr8			Jun3		
	Sn	Sp	PPV	Sn	Sp	PPV	Sn	Sp	PPV	Sn	Sp	PPV
TFIDF top 2010	97	91.8	43.8	92	91	50.3	89.5	93	59.3	96	90	47
Z-Score 784	95	92.6	45	82.5	91	50	85.5	93	58.2	93.3	91.2	48.6
Union of results	99.3	95.5	58.6	95.4	94.7	65	92.5	96.3	73.7	97.3	95.6	66.3

Table 3: Union of results using keywords based on TFIDF and Z-Score methods

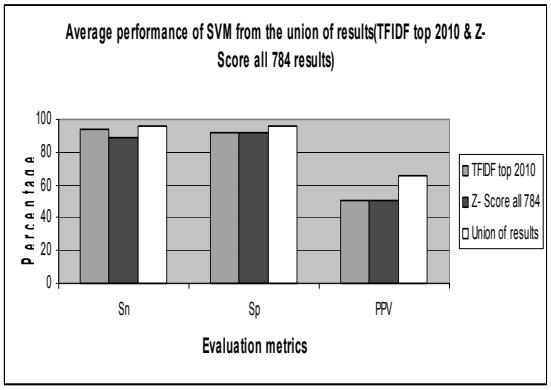

Figure 3: Average performance of SVM from the union of results

3.4 SVM classification outperformed Human expert classification

The false positives from the above result (i.e. union of results using keyword sets 3 & 8) were given to the CDC appointed expert, in charge of reviewing the literature for the HuGENet^{TM} database, for her scrutiny. In her inspection, she found that on average 50% of the false positives produced by the SVM were in fact true positives that were missed by her in her initial review process (Table 4 and Figure 4). Thus, our automated classification using SVM not only reduced the burden of manual processing, but also increased the sensitivity of the search.

	Feb 12	Apr 1	Apr 8	Jun 3
FP from union of results	100	89	57	74
TP after Human Expert Review	59	47	28	32
Percentage	59%	52.8%	49%	43.2%

Table 4: False Positive analysis

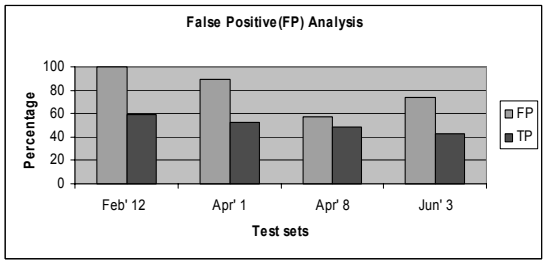

Figure 4 False Positive (FP) Analysis

3.5 SVM performance with the corrected training set

Based on the false positive analysis, we realized that the negative training set used was not good enough and it may have contained some of the positives that were missed by the human expert. Since its not feasible to manually check all the abstracts in the negative training set to remove the positives and also because our test set classifications were accurate as they were reclassified by human expert after SVM classification, the negatives from the three test sets Feb12, 2004, Apr 8, 2004 and Jun 3, 2004 were used as the negative training set while the positive training set was left unchanged. This training set, referred to as corrected training set was tested on the Apr 1, 2004 test set and used in our later comparisons. Also, the training set decreased in size because of the limited number of negative abstracts present in the three test sets. The results from SVM with the corrected training set using different sets of keywords, TFIDF top 2010 and Z-Score all 784, as features are shown in Table 5.

3.6 Performance of SVM with different training set sizes

We tested the accuracy of our models using a very small training corpus starting with just 50 positive abstracts. (In all our training sets, we use 50% positive and 50% negative examples).We varied the positive set size from 50 to 5362. Even with small training sets, SVM was able to pick up the right model parameters and gave reasonable results (Table 6).

3.7 SVM overall performance

With the corrected training set of equal number of positives and negatives and the union of results of SVM classifications obtained using TFIDF top 2010 keywords set and Z-Score all 784 keyword set on Apr 1, 2004 dataset, SVM achieved a Sensitivity of 96.3%, Specificity of 96.8% and Positive Predictive Value of 80.6% (Table 7).

3.8 SVM performance directly on the PubMed abstracts (without complex query)

Three random dates were chosen to check the performance of SVM directly on the articles contained in the PubMed even before applying the complex query. We were able to classify articles with 89.7% Sensitivity, 98.4% Specificity and 33.3% PPV. This shows that SVM can be used to classify the articles even without the complex queries. Also, the PPV of the PubMed search based on EDAT (Entrez date) increased 38 fold from 0.857% to 33.3% using SVM.

Training Set	Previous Training Set (5363 +ve's, 5316–ve's)	Corrected Training Set(5363 +ve, 5362 –ve)	Previous Training Set (5363+ve, 5362 –ve)	Corrected Training Set (5363 +ve, 5362 –ve)
Keywords	TFIDF top 2010	TFIDF top 2010	Z-Score 784	Z-Score 784
Sn	92.2	95.4	87.2	89
Sp	91	92.3	93	92
PPV	59.7	62.8	62.5	60.8

Table 5: Results from SVM after correcting the training set

Training Set Size	100	500	1000	2000	5000	10724
PPV	42.86	42.47	52.13	52.07	53.95	58.19
Sensitivity	84.48	90.8	91.38	86.78	90.23	95.98
Accuracy	87.93	87.55	91.29	91.23	91.83	93.13

Table 6: April 1 Results as a function of Training Set Size

Training Set	Corrected Training Set(5363 +ve, 5362 –ve)
Keywords	TFIDF top 2010 & Z-Score 784 (Union of results)
Sn	96.3%
Sp	96.8%
PPV	80.6%

Table 7: SVM overall performance

	Apr1 Test Set			Training Model Estimates	
Kernels	PPV	Sensitivity	Accuracy	Sensitivity	True Error
Linear	49.7	95.4	90.48	>79.47	<20.57
Polynomial	55.96	97.13	92.53	>74.08	<26.13
Polynomial1	58.19	95.98	93.13	>65.6	<34.98
RBF	51.54	95.98	91.13	>85.10	<56.02
Sigmoid	52.02	95.98	91.29	>95.11	<51.98

Table 8: April 1 Test Set trained on model with 5362 positives and 5362 Negatives

3.9 Parameter Tuning of SVM with different kernel functions

We conducted experiments for parameter tuning of our SVM algorithm with different kernel functions. The estimates reported in the table 8 refer to Apr 1, 2004 test set and include xi-alpha [17] estimates of true error (100%Accuracy) and sensitivity for five different kernels. Both the polynomial kernels used are of degree two but have different parameters. RBF and Sigmoid kernels behaved well only for a certain range of parameters whereas the linear and polynomial models generally performed well regardless of parameter choices on our discrete value datasets. In particular, the degree two polynomial1 function had best PPV and accuracy for our test set. (Table 8).

3.10 MultiWay-Classification

We have extended our binary classification work to explore multiway-classification of biomedical literature using SVM. We do so by incorporating training sets derived from known classes (e.g. types of cancers). In our current study we tried classifying literature related to 4 cancer types using the LIBSVM software [18]. We took 577 PubMed abstracts given by cancer experts at CDC with each abstract tagged with a cancer-type. They were randomly divided into two groups. The first group (training set) contained 200 abstracts with 50 abstracts of each cancer type whereas the second group (test set) contained the remaining 377 abstracts with non-uniform distribution of each cancer type. We tagged our training and test sets with individual tags for each class (e.g. 1 for colorectal, 2 for esophageal etc.). Then for each class, in comparison to the binary classification described in section 2.4, the decision to assign an abstract to that class is still regarded as a binary decision as follows. E.g., for colorectal cancer class, the label 1 corresponds to a positive assignment and labels 2, 3, 4 correspond to a negative assignment. For feature vectors, we used the top 1066 keywords that were generated based on their TFIDF. Using the Radial Basis Function (RBF) kernel with some parameter tuning, we were able to classify abstracts from these 4 cancer types with 93.9 % overall accuracy. Accuracy, sensitivity and PPV of each cancer type were also calculated (Table 9 and Figure 5). Given the complexity of the data, the initial results are very

promising and indicate that SVM can adequately classify data belonging to different classes.

Cancer Type	Accuracy	Sensitivity	PPV
Colorectal	95.2	91.1	98.1
Esophageal	98.7	100.0	66.7
Lung	96.8	96.4	96.4
Stomach	97.1	93.1	75.0

Table 9 Accuracy, Sensitivity and PPV for different cancer types using keywords based on TFIDF

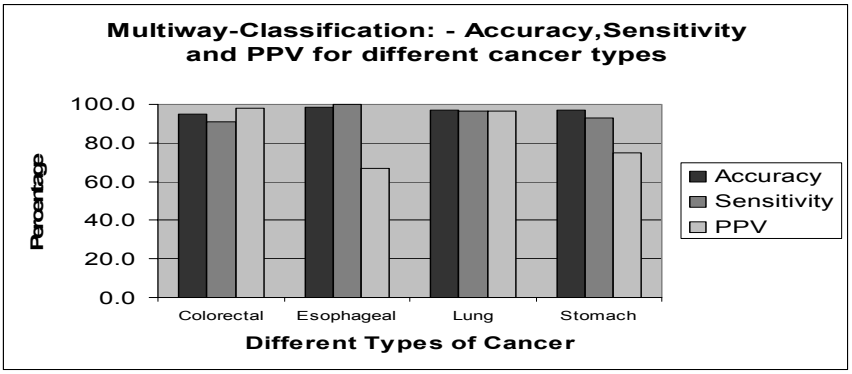

Figure 5 Multiway-Classification: Accuracy, Sensitivity and PPV for different cancer types from 377 abstracts using keywords based on TFIDF.

4. Discussion

Automated and standardized categorization and classification of the biomedical literature is an important challenge facing the scientific community. Due to the vast amount of data produced by emerging biomedical research, manual classification is not feasible. We tried classifying medical literature using supervised learning techniques such as decision trees (69.12 accuracy), k-Nearest Neighbor (82.90 accuracy), Neural Networks (85.12 accuracy) and found that SVM performed better than these on average. In our investigation into the use of SVM for efficiently classifying HuGE medical abstracts, a high degree of accuracy (96.8%), sensitivity (96.3%) and positive predictive value (80.6%) was achieved. Furthermore, SVM outperformed a human expert working on the problem fulltime for 4 years, by identifying 20% additional HuGE abstracts that were missed in human inspection.

Our initial results using SVM for multiway classification of cancer abstracts are also very promising (overall accuracy 85.7%). These results indicate that with careful parameter and feature selection SVM can be used for efficient and accurate classification of biomedical literature. In our preliminary analysis, slight improvement was achieved using leave-one-out cross validation technique which needs to be further explored. Use of different keyword weighting schemes and effect of using different kernels is an interesting area to explore. In future we wish to develop our technique into a tool useful for the average biomedical researcher and intend to develop good benchmarks (parameters, kernels) and incorporate them into this personalized tool to be used by the scientific community.

Acknowledgements

We thank Abhishek Dabral from Georgia Institute of Technology for editorial assistance and constructive criticism. We also thank Wei Yu, Bruce Lin, Mindy Clyne, Muin Koury from the Office of Genomics and Disease Prevention, Centre for Disease Control and Prevention (CDC) and Venu Dasigi from Southern Polytechnique State University for their invaluable comments and suggestions. Also, we appreciate CDC for providing the data from their Human Genome Epidemiology Network (HuGENet™) database.

References

[1] Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[2] Joachims, T. (1998) Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 137-142.

[3] Drucker, H., Vapnik, V., and Wu, D. 1999. Automatic text categorization and its applications to text retrieval. IEEE Transactions on Neural Networks 10, 5, 1048–1054.

[4] Dumais, S. T. and Chen, H. 2000. Hierarchical classification of Web content. In Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval (Athens, GR, 2000), pp. 256–263.

[5] Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of CIKM 98, 7th ACM International Conference on Information and Knowledge Management (Bethesda, US, 1998), pp. 148–155.

[6] Klinkenberg, R. and Joachims, T. 2000. Detecting concept drift with support vector machines. In Proceedings of ICML-00, 17th International Conference on Machine Learning (Stanford, US, 2000).

[7] Taira, H. and Haruno, M. 1999. Feature selection in SVM text categorization. In Proceedings of AAAI-99, 16th Conference of the American Association for Artificial Intelligence (Orlando, US, 1999), pp. 480–486.

[8] Joachims T (2002). Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publishers.

[9] Yang, Y. and Liu X. (1999) A re-examination of text categorizationmethods. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999.

[10] Simon Tong, Daphne Koller (2000). Support Vector Machine Active Learning with Applications to Text Classification. Proceedings of ICML-00, 17th International Conference on Machine Learning.

[11] Han, B., Vucetic, S., and Obradovic, Z., Reranking Medline Citations By Relevance to a Difficult Biological Query, IASTED Int'l Conf. Neural Networks and Computational Intelligence, Cancun, Mexico, 2003.

[12] Porter, M. (1980) An algorithm for suffix stripping, Program, 14:130-137.

[13] Liu, Y. et al. (2004) Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering, Proceeding of IEEE Computational Systems Bioinformatics (CSB 2004), p394-404, Stanford University, Stanford, Aug. 16th – Aug. 19th, 2004.

[14] Liu, Y. et al. (2005) Text mining biomedical literature for discovering gene-to-gene relationships: a comparative study of algorithms. IEEE/ACM Transaction on Computational Biology and Bioinformatics, Vol.2, No.1, Jan-March 2005.

[15] Salton, G., and Buckley, C. (1988) Text –weighting approaches in automatic text retrieval. Information Processing and Management, 24:513-523.

[16] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods- Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[17] Joachims, T (2000) Estimating the Generalization performance of an SVM Efficiently. International Conference on Machine Learning (ICML).

[18] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.