RERANKING MEDLINE CITATIONS BY RELEVANCE TO A DIFFICULT BIOLOGICAL QUERY

Bo Han[1,2] Slobodan Vucetic[1] Zoran Obradovic[1]
[1]Center for Information Science and Technology
Temple University, Philadelphia, PA 19122, USA
[2]State Key Laboratory of Software
Wuhan University, Hubei Province 430072, P.R.China

ABSTRACT

We have initialized research aimed at 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. Our proposed approach to this problem is to train support vector machines as classifiers able to distinguish relevant citations from the rest of retrieved citations. We used their predictions to re-rank citations retrieved from PubMed and represented as vectors of term frequencies. 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 from 48% to 16%. On average only 13% and 28% of citations relevant to our target topic were recalled in the top 5 and top 10 citations retrieved by queering PubMed with disordered protein names. By our reranking method, this was improved to about 58% and 78%, respectively, suggesting that the proposed method might provide a cost-effective tool for identifying articles that are difficult to express as specific PubMed queries.

KEY WORDS

Information Retrieval, Text Mining, Support Vector Machine, Bioinformatics

1. INTRODUCTION

The richest source of knowledge in bioscience is free text contained in the vast corpus of published research. One of the challenges in bioinformatics is developing automatic procedures to extract relevant information from the literature [1,2]. Natural language is context sensitive and there are currently no successful general solutions for automated knowledge extraction from free text. This problem is further complicated by peculiarities of bioscience literature such as complex and ambiguous terminology and ample use of abbreviations.

Some specific information retrieval problems from bioscience data such as extracting known protein-protein interactions or gene names are, conceptually, relatively simple, and some promising solutions to these problems have been proposed [3,4,5,6,7]. However, a more general problem of constructing automated systems for understanding the results and implications of published bioscience research is far from being solved. Perhaps the best approach with respect to the current state-of-the-art is construction of hybrid systems aimed at facilitating information retrieval by human experts. In this paper we illustrate such an approach on the problem of locating publications that are likely to contain information about a specific protein property called protein disorder.

Disordered proteins are characterized by long regions of amino acids that do not have a stable three-dimensional conformation under normal physiological conditions. Recent results indicate that, despite the traditional view, disordered proteins are common in nature and are responsible for a spectrum of important biological functions [8,9]. However, due to the lack of systematic studies, the knowledge about protein disorder is still scattered across literature and described using non-unified terminology. For example, term “protein disorder” has been widely accepted only recently [10], and, historically, it has been referred to using terms such as “unfolded”, “denatured”, “unstructured”, “unstable”, etc.

In an initial effort to systematize knowledge about protein disorder, Dunker and collaborators invested more than a man-year effort in literature overview and they gathered valuable evidence about functions of 90 disordered proteins [8]. However, this approach cannot be feasible in future studies that should examine the evidence about protein disorder in thousands of proteins. Clearly, this process should be facilitated by use of an appropriate data mining technology.

Medline is a major repository of biomedical literature that stores citations for about 12 million biomedical papers and is incremented by additional 400,000 new citations each year. PubMed [11] is the most popular online interface to Medline documents that allows Boolean queries based on combinations of keywords and returns all citations satisfying the query. Such keyword-based search is very sensitive to the choice of a query, which is

related to complexity and ambiguity of terminology used in the citations.

For example, assume a user is interested in finding the papers useful to identifying functions of disordered protein called Apurinic/apyrimidinic endonuclease. Given a query "Apurinic/ apyrimidinic endonuclease AND disorder," the PubMed search system returns only one record unrelated to the posed question. If the query is reduced to "Apurinic/apyrimidinic endonuclease," the PubMed returns over 600 citations. Reading such a large number of articles could be overwhelming to the user of the system. Hence, automatically reranking the resulting citations according to their relevance to the question of interest could be extremely valuable.

In this paper we describe a procedure for an automatic extraction of PubMed citations on specific proteins and reranking according to their relevance to determining functional properties of disordered protein regions. Our approach to reranking is based on classification. The data set was composed of positive class examples representing citations with relevant information about protein disorder property, and of contrasting class examples representing arbitrary citations from PubMed. A classifier with soft output constructed on such data is used to rank citations based on their similarity to positive class citations.

The proposed solution belongs to one-class classification approaches where one class of data, called the target class, has to be distinguished from the rest of the attribute space. Previous solutions include kernel density estimation and support vector data description (SVDD) [12], and they are not robust to high dimension of attribute space. The novelty of our solution is in use of unlabeled examples that enforces learning relevant properties of positive examples with respect to the overall data distribution.

Using a bag-of-words approach, each citation is first represented as a vector of keyword frequencies in the document. To reduce the dimensionality, words are ranked by their ability to discriminate the citations related to the property of our interest (function of protein disorder) and only the most discriminative ones are used as attributes.

Support vector machines (SVM) with linear and nonlinear kernels are finally used to re-rank the citations returned by PubMed based on a given query. Use of SVM is motivated by their successful application to text classification [13]. The experiments indicate that SVMs often outperform other popular algorithms for text mining such as Naive Bayes or KNN, and that are robust even when documents are represented as high dimensional vectors. The reported experimental evidence also suggests that SVM with linear kernels are often almost as accurate as SVM with nonlinear kernels and therefore both will be considered in this study.

It is worth to note that our approach can be applied on a large class of similar problems in biological text mining. For a given biological property of interest, biologists can select a set of relevant citations and apply our approach to re-rank all citations discussing a certain protein returned by PubMed.

2. METHODOLOGY

The proposed approach consists of four steps. In the collection step citations on scientific papers are retrieved from PubMed. The most informative frequent words in positive and the contrasting citations are identified next in the keywords selection step. A feature vector representing frequencies of selected keywords is computed over all training citations and this representation is used for learning a classification model in the ranking step. The learned model is applied to re-rank all papers in a test set retrieved from PubMed by a protein name query. More detailed description of the proposed methodology is described in this section.

2.1 Data Retrieval

An automated protocol that enables communication with the PubMed web server was developed in order to retrieve Medline citations via the Hypertext Transfer Protocol. In this procedure, queries are submitted to PubMed and web pages with a list of the corresponding papers are received. These HTML files are parsed to extract URLs to the corresponding papers. The citations of the papers are downloaded and the title, authors, abstracts and Medical Subject Headings (MeSH) terms are automatically extracted and stored in a local database.

Using this protocol we retrieved from Medline a set of *positive examples* consisting of citations relevant to the topic of our interest (in our application it is learning about functions of known disordered proteins). As a *contrasting class* we collect a set of unlabeled citations for which their relevance to the topic of our interest is unknown. In our application contrasting citations correspond to the same set of proteins covered by the positive citations.

2.2 Knowledge Representation

In a bag of words representation each text is represented as a vector of frequencies of words that appear in the text. For categorizing the texts, low frequency words are often statistically inadequate while some very frequent words like stop words (e.g.. "and", "the", "a") are irrelevant. So, instead of representing each citation d as a bag of all its words, our objective is to represent it as a vector of features $\mathbf{x}_d = (x_{d,1}, x_{d,2}, x_{d,3}, \ldots, x_{d,K})$ corresponding to the selected K informative keywords.

Some uninformative words such as "the", "at" and "in" were first eliminated from citations by using a short stop list. To reduce redundancy, words in all citations were preprocessed into terms such that suffixes were eliminated from related words using Porter stemmer [14]. Since some biological abbreviations are very short, words shorter than three characters were kept unprocessed. In the training set, term frequencies at positive citations were compared to their frequencies at contrasting citations. The objective of this comparison was to identify discriminative high frequency words while eliminating stop words as much as possible. More precisely, for each term in the training set we measured the difference of its average frequency in the positive vs. contrasting citations.

The K most discriminative terms based on this measure were selected as the keywords for representing each citation as an K-dimensional vector of features. The so-called TF-IDF weight was used to construct the features where relevance of a keyword k to citation d was measured as the frequency of this keyword in citation d (called term frequency or TF) and the inverse document frequency (IDF) of the term in the whole training set. Using TF-IDF, the value of feature $x_{d,k}$ corresponding to keyword k in document d was computed as [15]

$$x_{d,k} = \frac{f(k,d)}{\max(f(k,d))} \log D / d_k$$

where f(k,d) is the frequency of keyword k at citation d, max(f(k,d)) is frequency of the most frequent keyword in document d, D is the total number of training citations, and d_k is the number of training citations containing keyword k. A high TF-IDF weight is achievable by a high term frequency in the given citation and a low frequency of the term in the whole training set.

2.3 Support Vector Machines for Classification

Given a positive and a contrasting set of training citations represented as TF-IDF weights of K selected keywords, various classification models can be used for learning and ranking new citations. The base algorithm considered in this study is Support Vector Machines (SVM) [16] which is optimized to find a decision hyperplane with the maximal margin of separation between the two classes of high dimensional data.

In a case where classes are separable, the learning process consists of minimizing $\|\mathbf{w}\|^2$ under the constraint that $\mathbf{w}\cdot\mathbf{\Phi}(\mathbf{x}_d) + b \geq 1$ for citations from the positive class, and $\mathbf{w}\cdot\mathbf{\Phi}(\mathbf{x}_d) + b \leq -1$ for citations from the contrasting set, where $\mathbf{\Phi}$ is a mapping function, $\mathbf{w}$ is a corresponding weight, and $\mathbf{x}_d$ is a vector of TF-IDF attributes corresponding to citation *d*. In SVMs, $\mathbf{\Phi}$ is chosen to satisfy $\mathbf{\Phi}(\mathbf{x})\cdot\mathbf{\Phi}(\mathbf{y}) = K(\mathbf{x},\mathbf{y})$, where K is an appropriate kernel function.

In practice classes are often not separable, slack variables are introduced into optimization to relax the constraints. For a practitioner, this introduces a regularization constraint C>0 that should be set properly. In this paper we consider linear SVM with $\mathbf{\Phi}(\mathbf{x}) = \mathbf{x}$, and a nonlinear SVM with polynomial kernel $\mathbf{K}(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+\theta)^m$. Slack variables are used with a properly chosen C.

2.4 Ranking Citations through Classification

By SVM($\mathbf{x}$) we denote the output of a trained SVM when given a vector $\mathbf{x}$ of TF-IDF weights corresponding to a testing citation**.** We used this value as a confidence estimate that this citation belongs to a positive class such that large SVM($\mathbf{x}$) indicates large confidence. In the reranking application, SVM($\mathbf{x}$) confidence values were computed for all test citations obtained through PubMed search for literature about a certain protein. These scores were then used to re-rank the citations.

When the contrasting set is much larger than the positive set (as in our application), training set should be constructed carefully. In our approach, for purpose of achieving good generalization, we were constructing balanced training sets with the same number of positive and contrasting examples. This was achieved by taking a random sample from the contrasting set of the same size as the positive set.

Since this approach does not fully use the available information contained in the contrasting set, we also examined using an ensemble of such sets. There, multiple models were trained on multiple balanced samples with all positive examples and the same number of randomly drawn contrasting examples. The predictions of multiple models were then averaged to produce the final scores for ranking of citations. This approach is in its nature similar to popular ensemble methods such as bagging [17].

To provide an appropriate performance validation we performed k-fold cross validation experiments where the positive and a corresponding contrasting set were each partitioned randomly into k disjoint and nonredundant subsets such that in each of k experiments different k-1 subsets from each group were used as a *training set* for developing a reranking model while out-of-sample evaluation was performed on the remaining *test data* and the averaged results were reported.

2.5 Ranking Refinement

Following the recursive classification idea of Zhang and Wong [18] that was effectively used for gene selection, we also considered refining citations rankings through a two-stage SVM classification. In this approach at the initial phase all terms were used to develop an initial linear SVM. This model was used to weight terms according to their contribution to distinguishing positive

from contrasting data. The high weighted terms were selected to construct a feature vector used for developing a second stage SVM classifier as described in sections 2.3 and 2.4.

3. EXPERIMENTAL RESULTS

This section provides empirical evidence that our approach is efficient on reranking papers according to their relevance to disorder-function relationship.

3.1 Data Set

To evaluate the proposed approach to reranking we constructed a data set based on systematic literature search about cell functions of 90 disordered proteins [9]. This search resulted in 178 references confirmed to be relevant to the function of disordered proteins, out of which 5 were personal communication and another 7 were papers not listed in the Medline. We extracted citations for the remaining 166 references from PubMed to form a data set **P** of positive examples.

PubMed citations contain both the paper abstract and its MeSH terms. A hierarchy of more than 17,000 MeSH terms is used by indexers at the National Library of Medicine to annotate Medline articles with up to 15 specific terms. Since MeSH terms provide very useful information about the subject of each biomedical article complete citations were used instead of abstracts only.

A baseline search method considered in this study was based on the protein name. We retrieved 18,499 citations from 90 PubMed queries based on the 90 disordered protein names reported at [8]. We denote this set of citations **C** as contrasting examples. With each contrasting example we also saved the name of its corresponding disordered protein, and the ranking after the PubMed query.

Sixty seven out of 90 proteins had more than 100 citations; 35 had more than 1,000 citations; and 12 had more than 10,000 citations. From the 166 citations from the positive set **P**, only 90 were also found in the control set and they corresponded to 61 disordered proteins. We denoted the set of these 90 papers as **P**90 (so, **P**90 $\subset$ **C**), while the remaining 76 papers were denoted as **P**76 (so, **P**76 $\not\subset$ **C**). In most cases, **P**90 citations were not listed as one of the top results after the protein-name based PubMed search. Only 12 were recalled among the top 5, and only 25 among the top 10 retrievals.

3.2 Feature Selection Results

Among all citations in positive and control sets, there were totally 20,127 unique terms. The 20 most frequent terms are listed in Table 1. We can see that some prepositions, such as "with", or some common words, such as "protein", appeared in the list. These terms could not make significant contribution to distinguishing the relevant papers.

To extract the most informative terms, we used our term-ranking procedure (section 2.2) to select the most informative K terms as keywords. Table 1 lists the top 20 keywords extracted by this method using the whole datasets **P** and **C**. Since in all experiments we performed 10-fold cross validation, the ranking was slightly different in different validations from Table 1. From visual inspection of the top 20 selected keywords it is evident that most of the selected keywords are very likely to be related to protein disorder property.

In Table 1 we also show the results of the refined feature selection (from section 2.5) using the whole datasets **P** and **C**. While the resulting selection is very similar to the results of the initial procedure from section 2.2, some differences could be observed. However, only from visual inspection it is difficult to conclude which selection is more related to protein disorder.

Table 1. Top 20 terms by 3 methods

Highest frequency	Initial selection	Refined selection
protein	structur	structur
structur	flgm	flgm
bind	conform	conform
metabol	disord	residu
chemistr	calcineurin	flagellin
that	helic	helic
for	loop	disord
dna	residu	sigma
sequenc	rai	reson
domain	unfold	nmr
with	flagellin	resolut
support	circular	rai
acid	dichroism	crystal
molecular	crystal	fragment
activ	calmodulin	solut
amino	reson	complex
gov	chemistri	secondari
peptid	domain	dichroism
cell	nmr	thermophilu
Factor	resolut	unfold

3.3 Performance Measure

To measure the quality of ranking, we defined an accuracy measure called the *relative position* (*rp*) as

$$rp = \frac{absolute\,rank\ of\ a\ citation}{number of\ retrieved\ citations}\ .$$

The *average rp* value among the queries on the test set (*arp*) is used as a measure of ranking quality.

The number of retrieved citations based on a protein-name query varies widely and so *rp* measure alone is insufficient. For example *rp*=0.2 is a very good result if the number of retrieved citations is ten (so, citation is ranked as second), but is less satisfactory if thousands of citations were . Therefore, we also report the proportion of relevant citations ranked among top 5 and top 10.

3.4 Evaluation of Reranking

We used 10-fold cross validation to estimate the reranking performance. We divided **P**90 into 10 subsets, trained and tested our model 10 times. Each time, 9 of 10 subsets from **P**90 were used in training while the omitted subset was retained for testing.

As a classifier we used SVM-Light [19] with C parameter set to a default value. Figure 1 shows relationship between the number of keywords K and *arp* ranking of linear SVM trained on keywords extracted with two different feature extraction methods. The *arp* of the citations retrieved by the PubMed protein name based query was 0.477 indicating near-random ranking with respect to protein disorder function. It could be seen that for any K our procedure resulted in a much better ranking, while the best performance was reached for K = 400. For K larger than 400 it seems that additional keywords were less related to protein disorder thus causing overfitting of SVM. Comparison of the two feature selection methods indicates that refinement resulted in a significant improvement over the initial reranking method (*arp* of 0.160 vs. 0.175 for K=400).

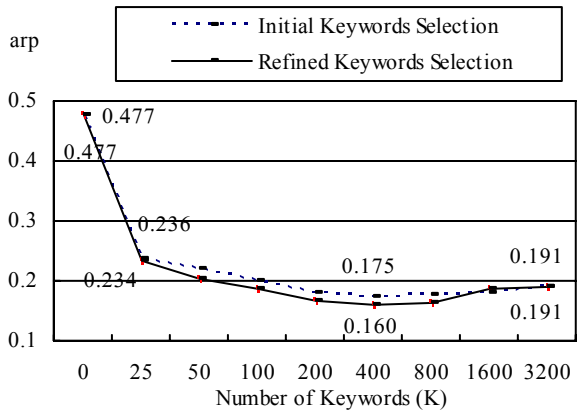

Figure 1. 10-fold cross validation testing of the average rank (*arp)* as a function of the number of keywords K.

To test robustness of the best performing reranking SVM (K=400) with respect to the sample choice, experiments were performed on balanced data consisting of an equal number of positive and unlabeled citations where training for each of 10-fold cross validation experiments was repeated for 100 random samples of unlabeled data and 100 SVM were constructed. The results of this experiment are summarized at Figure 2 where for each ensemble size *arp* results were shown over 10-fold cross validation experiments. The *arp* over ensembles of size up to 100 was consistently around 0.17 providing strong evidence that the proposed approach is stable.

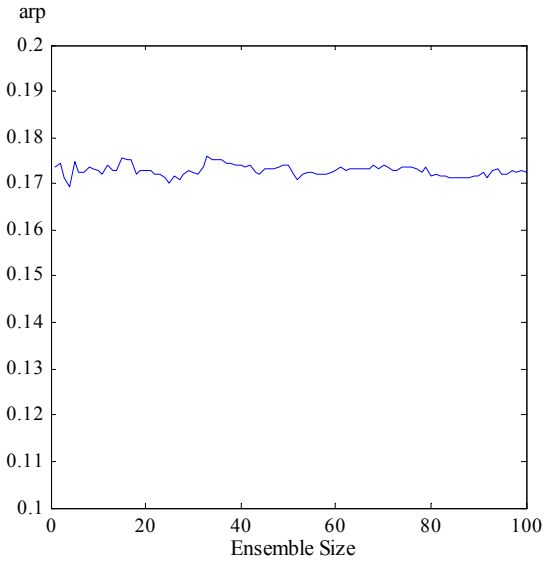

Figure 2. Dependence of *arp* to a size of an ensemble produced by combining outputs of linear SVM classifiers. Each SVM was trained on random samples of unlabeled and positive citations using a 400-keyword-representation of the initial selection method.

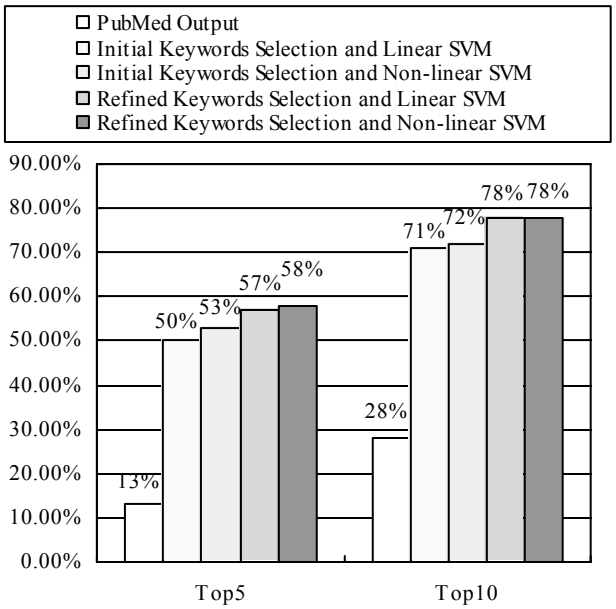

Figure 3. Fraction of relevant citations ranked as top 5 and top 10 before and after reranking (SVM were trained using representation with 400 keywords).

Comparison of ranking quality based on the top 5 and top 10 retrievals is summarized in Figure 3. In PubMed retrieval only 25 out of 90 citations (28%) were recalled in top 10, while after reranking with the initial or refined keywords selection and linear or non-linear SVM classifier, the number of citations ranked in top 10 has

been increased to 64 (71%), 65 (72%), 70 (78%), and 70 (78%) respectively.

In all experiments with non-linear SVM classifiers, we applied a second degree polynomial kernel. From the results summarized at Figure 2 it is evident that linear kernels were fairly similar to non-linear kernels in reranking quality. This result is consistent with some previously reported results in text mining [13].

4. CONCLUSIONS

Automatically reorganizing documents according to their relevance to a topic that is difficult to express as a query is an important objective for biological knowledge discovery. We described and tested an approach for an automatic extraction of Medline citations on specific proteins and reranking them according to their relevance to determining functional properties of disordered protein regions. Our approach to reranking is to select the keywords by analyzing relevant vs. contrasting citations of unknown relevance, followed by learning linear and non-linear kernel SVM classifiers, and using them to sort citations according to the strength of the prediction. Experimental results provide evidence in support of effectiveness of our approach. The next objective of our research is to characterize other problems in biological text mining where the proposed approach with appropriate modifications can be beneficial.

REFERENCES

[1] M. Craven, and J. Kumlien, Constructing biological knowledge bases by extracting information from text sources, *Proc.7th Intl'l. Conf. Intelligent System for Molecular Biology*, Heidelberg, Germany, 1999, 77–86.

[2] U. Hahn, M. Romacker, and S. Schulz, Creating knowledge repositories from biomedical reports: The MEDSYNDIKATE text mining system, *Proc. Pacific Symp.Biocomputing*, Hawaii, USA, 2002, 338-349.

[3] N. Collier, C. Nobata, and J. Tsujii, Extracting the names of genes and gene products with a hidden Markov model, *Proc. 18th Int'l Conf. Computational Linguistics*, Saarbrucken, Germany, 2000, 201–207.

[4] E.M. Marcotte, I. Xenarios, D. Eisenberg, Mining literature for protein-protein interactions, *Bioinformatics. 17*(4), 2001, 359-363.

[5] D., Proux, F. Rechenmann, L. Julliard, V. Pillet, and B. Jacq, Detecting gene symbols and names in biological texts: A first step toward pertinent Information extraction. *Genome Informatics 9*, 1998, 72-80.

[6] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran, Robust relational parsing over biomedical literature: extracting inhibit relations, *Proc. Pacific Sym. Biocomputing,* Hawaii, USA, 2002, 362-373.

[7] T., Rindflesch, J. Rajan, and L. Hunter, Extracting molecular binding relationships from biomedical text, *Proc. 6th Conf. Applied Natural Language Processing,* Seattle, WA, USA, 2000, 188–195.

[8] A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva, and Z. Obradovic, Intrinsic disorder and protein function, *Biochemistry*, *41*(21), 2002, 6573 - 6582.

[9] A.K. Dunker, C.J. Brown, and Z. Obradovic, Identification and functions of usefully disordered proteins, *Advances in Protein Chemistry*, *62*, 2002, 25-49.

[10] P.E. Wright, and H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, *293*(2), 1999, 321-331.

[11] Entrez PubMed, www.ncbi.nlm.nih.gov/PubMed.

[12] D.M.J. Tax and R.P.W. Duin, Uniform object generation for optimizing one-class classifiers, *Journal of Machine Learning Research, 2*(Dec), 2001, 155-173

[13] T. Joachims, Text categorization with support vector machines: learning with many relevant features, *Proc. 10th European Conf. Machine Learning*, 1998, 137-142.

[14] Porter, An algorithm for suffix stripping, *Program*, *14*(3), 1980, 130-137

[15] G. Salton, and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, *24*, 1988, 513-523.

[16] V. Vapnik, *The nature of statistical learning theory* (New York : Springer-Verlag, 1995).

[17] L. Breiman, Bagging predictors, *Machine Learning, 24*(2)*,* 1996, 123-140.

[18] X. Zhang, and W.H Wong, Recursive sample classification and gene selection based on SVM: Method and software description, Biostatistics Dpt.Tech Report, *Harvard School of Public Health*, 2001.

[19] T. Joachims, Making large-scale SVM learning practical, in B. Schölkopf, C. Burges, and A. Smola (Ed.) *Advances in kernel methods - support vector learning,* 11 (MIT Press, 1999).