2006 IEEE Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

A Novel Support Vector Machine with Class-dependent Features for Biomedical Data

Nina Zhou and Lipo Wang

***Abstract*— In this paper we propose a novel support vector machine (SVM) with *class-dependent* features. According to an importance measure, e.g., the RELIEF weight measure or class separability measure, we rank the features importance for each class against the rest of classes. For each class we select an optimal feature subset using a classifier, e.g., the support vector machine (SVM). For the classification on these class-dependent feature subsets, we propose to construct a novel SVM using "one-against-all" in 2 processes: 1) construct one model for each class by training the classifier with the class's optimal feature subset; 2) during testing, each test pattern is tested on all models and the model with the maximum output decides the class of the test pattern. The method's performance is evaluated on two benchmark datasets. Our results indicate that our novel SVM classifier can effectively realize the classification of *class-dependent* feature subsets found by our wrapper approach which can remove irrelevant features for each class and at the same time maintain or even improve the classification accuracy in comparison with other feature selection methods.**

I. INTRODUCTION

SVM [25] [13] has already been used very successfully on many areas, such as pattern recognition, pattern classification, and regression problems etc., since it comes into our lives. Specially for its attractive features, regarding to the ability to effectively avoid overfitting, accomodate large feature spaces and condense the information of the data set [23], it now has already aroused the interest of many biomedical researchers, and has already been applied to the bioinformatics field, for example, cancer diagnosis [1] [6], protein secondary structure prediction [23][9][22] and gene analysis [2].

As a classifier, SVM has significantly better performance than or at least matches that of traditional machine learning approaches, including neural networks [23]. SVM was originally designed for binary classification, although most problems are multi-class. In order to solve those multi-class problems, some methods have been proposed to effectively extend the binary SVM to multi-class classification. The earliest used method is "one-against-all" [15], which is to construct D SVMs for the D-class problem. The i-th SVM is trained with all the examples in the i-th class with positive labels, and all the other examples with negative labels. Another method is to construct $D(D-1)/2$ classifiers and each classifier is trained on data only from two classes, which is called "one-against-one" method [24]. DAGSVM [10] is also a method based on solving binary classifications, which has the same training phase as "one-against-one" method and uses a rooted binary directed acyclic graph in the testing phase. Also several "all-together" methods are proposed, which are to consider all classes at once. With regard to those methods, Hsu and Lin [5] made a comparison ad pointed out that "all-together" methods have advantage on support vectors, "one-against-one" and DAGSVM methods are more suitable for practical use, and "one-against-all" is a good method whose performance is comparable to "one-against-one".

Since the given data sets we process are becoming increasingly larger in the number of patterns and the dimension of features or attributes, this may degrade the efficiency of most learning algorithms used on them, especially when those data exist some irrelevant or redundant features. For example, John [11] showed that a single irrelevant features in data set, such as credit-approval or diabetes, cut down the prediction accuracy of C4.5 by 5%. Langely and Sage [18] showed that the naive Bayes learner performs less optimally in the presence of irrelevant features. Fu and Wang [27] showed that deleting those irrelevant features can not only improve the classification accuracy, but also reduce the structural complexity of the radial basis function (RBF) neural network. In these cases, data dimensionality reduction (DDR) is urgent and necessary before a learning algorithm is used.

As to DDR, there are two commonly used techniques: feature extraction and feature selection. Feature extraction is the process of transforming the original set of features into a set of new features. Many current feature extraction techniques involve linear transformation of the original pattern vectors to new vectors of a lower dimensionality, e.g., principal component analysis (PCA) [16] and linear discriminant analysis (LDA)[26]. For PCA and LDA, although they reduce the dimension of features measured by the classifier, they do not actually reduce the number of features involved in the classification, because they linearly transform the original features into the new feature sets. Hence it will be possible for such techniques to hinder the target concept learning due to the irrelevant features. Feature selection [14] is the process of actually reducing the number of features in the classification, while maintaining acceptable classification accuracy. The technique for feature selection eliminate irrelevant features, leaving the best subset of features which retains sufficient information to discriminate well among classes. Thus how to decide which features are relevant or irrelevant and how to delete those features so as to form the optimal feature subset are the main objectives of feature selection.

In feature selection, one usually chooses the same feature

The authors are with College of Information Engineering, Xiangtan University, Xiangtan, Hunan, China, and with Nanyang Technological University, Block S1, 50 Nanyang Avenue, Singapore 639798. ZHOU0034@ntu.edu.sg, lipowang@126.com

1-4244-0100-3/06/$20.00 ©2006 IEEE

subset for all classes in a given classification problem, which is called *class-independent* feature selection [27] [12] [21] [20]. In order to make advantage of the possibility that different groups of features may have different power in distinguishing different classes, we tend to choose a different feature subset for each class, which is called *class-dependent* feature selection [7] [8]. Class-dependent feature selection can have matched effect or be significantly better than that of class-independent feature selection. For example, Oh et al. proposed [7] [8] a *filter* approach to *class-dependent* feature selection to improve recognition performance for handwriting digits. They proposed an estimated class distribution and used it to calculate each feature's class separation for different classes. In [28], Fu and Wang used GA to select the optimal feature subset for each class based on an RBF classifier. This *class-dependent* feature selection approach made explicit use of the clustering property of the RBF neural network and can not work for a general classifier, i.e., SVM.

Considering many attractive features of SVM, we'd like to choose it as the classifier used in the process of selecting features and also in the final classification. When selecting class-dependent features, we first rank the importance of the features for class i ($i = 1, 2, ..., N$, N is the number of classes) using a feature importance measure, i.e., RELIEF weight measure [12] or the class separability measure [27]. We then select an optimal feature set from the ranking list for class i using SVM. This process will be repeated N times until N classes have their own feature subsets. Due to the fact of class-dependent feature selection, in this paper, we propose a novel SVM with *class-dependent* features. For each class, we construct one model by training the classifier with the class's optimal feature subset. Then during testing, each test pattern is tested on all models and the model with the maximum output decides the class of the test pattern.

The paper are organized as follows. In section 2 we first present our wrapper approach to *class-dependent* features selection, with two ranking measures. Then we propose our constructed SVM classifier with *class-dependent* features. In section 3, we present experimental results on two data sets from the UCI database [3], comparing *class-dependent* feature selection (CDFS) and *class-independent* feature selection (CIFS), together with published results on classification accuracy. In section 4 we provide a summary and discussion for future work.

II. Methodology

In this part, we first introduce our wrapper approach to selecting class-dependent features. Then we will simply review the original SVM and describe our constructed class-dependent SVM classifier. As to the RELIEF weight measure [12] and CSM [27][19], we won't give more details here.

A. Our Wrapper Approach to Selection of Class-dependent Features

Our wrapper approach to selection of class-dependent features consists of the following three steps (Fig. 1). In

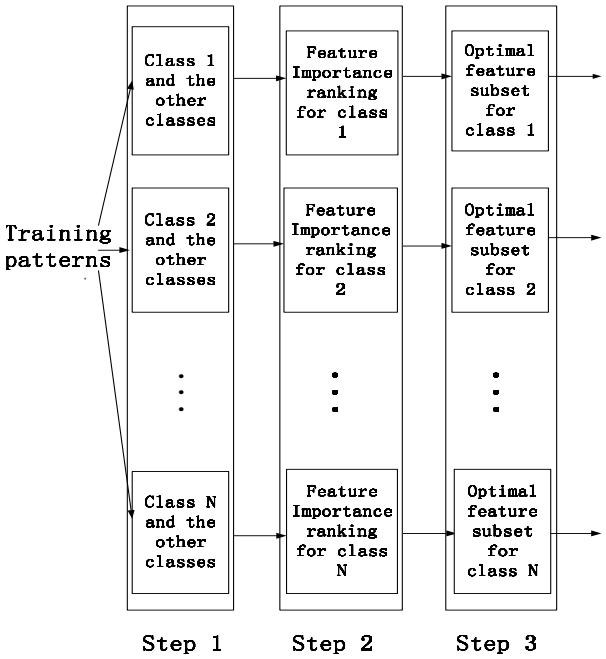

Fig. 1. Our wrapper approach to selection of class-dependent features.

step one, we convert a N-class classification problem to N 2-class classification problems, i.e., problem 1, problem 2, $\cdots$, and problem N. Problem i only has two classes: one is the original class, and the other includes all the other classes. It's goal is to correctly separate the two classes.

In step two, we need to rank the attribute importance using a feature importance ranking measure, i.e, probabilistic distance measure, RELIEF weight measure [12], class separability measure [27][19], information theoretic measure [17]. In this paper, we adopt the RELIEF weight measure and class separability measure to evaluate the importance of the features for each class and make a comparison about the final classifications. In this case, the attribute importance ranking thus obtained is called as *class-dependent attribute importance ranking*.

After obtaining the attribute importance ranking of each class, we then turn to step three: choose an optimal feature subset for each class through SVM. For each class we start with the most important feature as the first subset, and then each time add one attribute into the previous subset from the ranking list to form a new feature subset, until all the attributes in this class are added. Inputting each feature subset into the classifier, we can obtain different classification accuracy for different feature subsets. The optimal feature subset will be the one with the highest classification accuracy or lowest error rate.

For convenience, we use feature masks to represent feature subsets. Feature mask only has value "0" or "1", indicating the absence or presence of a particular feature. For example,

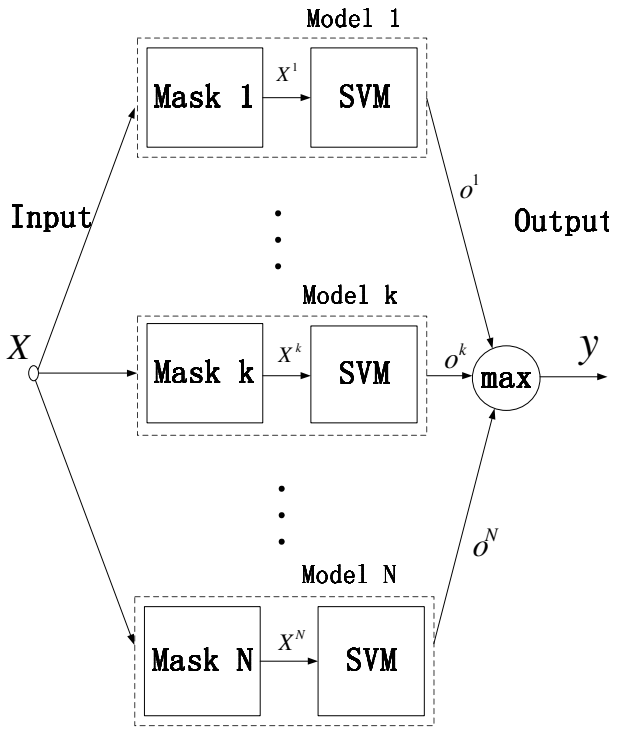

Fig. 2. Architecture of our novel SVM classifier with class-dependent features

if originally there are 5 features, i.e., $\{x1, x2, x3, x4, x5\}$, the optimal feature subset is $\{x_1, x_2, x_3\}$ with the forth and fifth features deleted, the feature mask will be $\{1, 1, 1, 0, 0\}$.

After introducing the wrapper approach to selection of class-dependent features, we will turn to the introduction of our novel SVM with class-dependent features.

B. Our Novel SVM with Class-dependent Features

In our experiment, because the data are with class-dependent feature subsets, we can not directly input the different feature subsets into any one of the SVM classifier mentioned above at one time. Here there are two choices. One is to recode the SVM classifier with "all-together" method for our purpose, which may be time-consuming and inconvenient. The other one is adopt the LIBSVM in our experiment, which is an efficient software for SVM classification, and make some modifications to form a class-dependent SVM classifier for our data. In our experiment, we will adopt the "one-against-all" method to construct a class-dependent SVM classifier, which is also consistent with our feature selection process. In Fig. 2, we present the classifier's architecture using LIBSVM to train and test the patterns with class-dependent feature subsets, which are described in the following 2 processes:

1) The training process:
In this process, we use training patterns to construct N SVM models, where N is the number of classes. The SVM k is trained using all of the training examples in class k with positive labels and all the other examples with negative labels. A feature mask is obtained for each SVM as described in the previous subsection. For instance, given a training pattern (X, y), where y is the class label of input X, the training pattern X to form the k-th model is presented as X^k with M_k features, M_k being the number of '1' in the feature mask of class k. $y = +1$ if X^k belongs to class k and $y = -1$ otherwise.
SVM model k solves the following problem (see [5]):

$$\begin{aligned} \min_{w^k, b^k, \xi^k} \quad & \frac{1}{2}(w^k)^T w^k + C^k \sum \xi^k \\ (w^k)^T \phi(X^k) + b^k \geq 1 - \xi^k, & \text{ if } y = +1, \\ (w^k)^T \phi(X^k) + b^k \leq -1 + \xi^k, & \text{ if } y = -1, \\ \xi^k \geq 0 & \end{aligned} \tag{1}$$

Here w^k denotes a set of linear weights connecting the feature space to the output space. Input X^k is mapped to a higher dimensional space by the transformation function ϕ. C^k is a tradeoff parameter between the error and margin, and ξ^k are "slack variables" in optimization theory. Minimizing $\frac{1}{2}(w^k)^T w^k$ is equivalent to maximizing the margin between two classes of data. When data are not linearly separable, $C^k \sum \xi^k$ is used as the penalty term to reduce the number of training errors.

2) The testing process:
After all the models are formed, we will use them to test unlabeled patterns. Same as the training process, each testing pattern is filtered with one class's feature mask before input into the corresponding SVM model and the original attributes corresponding to '0' in the feature mask are removed. Assume that X is one original testing pattern, then N different feature subsets X^1, ..., X^i, ..., X^N will be separately input into the corresponding N models. We then classify X to the class which has the largest value of N decision functions:

$$\text{class of } X \equiv \texttt{argmax}_{j=1,\ldots,N}((w^j)^T \phi(X^j) + b^j) \tag{2}$$

In order to compare the N decision functions from LIBSVM on *class-dependent* feature subsets, we need specify the parameter '-b' as 1 in function 'svmpredict' [4] so that we can obtain a matrix output which consists of predicted labels and the corresponding probabilities. When one testing pattern is separately input into N models, we obtain N matrix outputs: O^1, ..., O^i, ..., O^N, all of which are vectors with dimension 3. The first element of the vector

TABLE I

THE NUMBER OF FEATURES ELIMINATED IN EACH OF THE 10 SIMULATIONS, THE AVERAGE NUMBER AND ITS STANDARD DEVIATION, USING RELIEF RANKING METHOD FOR THE BREAST CANCER DATA SET.

	The number of features deleted in each of the 10 simulations (RELIEF)	Average and Std (RELIEF)
Class-independent	3 3 4 3 3 0 0 4 1 1	2.2±1.5
Class-dependent	3 3 4 3 3 0 0 4 1 1	2.2±1.5

TABLE II

THE NUMBER OF FEATURES ELIMINATED IN EACH OF THE 10 SIMULATIONS, THE AVERAGE NUMBER AND ITS STANDARD DEVIATION, USING CSM RANKING METHOD FOR THE BREAST CANCER DATA SET.

	The number of features deleted in each of the 10 simulations (CSM)	Average and Std (CSM)
Class-independent	0 0 0 0 0 1 1 1 0 0	0.3±0.5
Class-dependent	0 0 0 0 0 1 1 1 0 0	0.3±0.5

is the predicted class label, the second and third element are probabilities. The probability that indicates the test data is in the class is what we need. Among the N probabilities obtained, the testing pattern belongs to the class with the largest probability, which is equivalent to (2).

III. EXPERIMENTAL RESULTS

We apply our feature selection method to the Breast cancer and Ecoli data sets from the UCI Machine Learning Repository databases [3] and then classify them with our novel SVM classifier. The Breast cancer data set has 699 samples, 9 attributes and 2 classes. The Ecoli data set describes the protein localization sites with 336 samples, 7 attributes and 8 classes.

From Table I to IV, we list the number of deleted features in each of the 10 simulations (10-fold) , the average number of deleted features and its standard deviation for two data sets. For Table I and Table II, we can see that the Breast Cancer data set have few irrelevant or redundant features. Using RELIEF measure, it has on average 2 features are removed. While using CSM, it has on average no features deleted. For Table III and Table IV, we can see the Ecoli data set has different feature subset for different classes with CDFS, e.g, for class 1 to class 5, the number of features eliminated is less, while for class 7 and 8, the number of deleted features is nearly 6. Compared with CIFS, the CDFS can show us that different features have different power in discriminating different classes. Then for these class-dependent features, our constructed novel SVM classifier will be able to effectively realize the classification. In table V, we list the classification accuracy with two ranking measures, two feature selection methods, and compare them with the classification accuracy of SOAP and Relief [21]. Our feature selection method has better accuracy with our constructed SVM. Besides these two biomedical data sets, we also try on several other data sets from UCI, such as Thyroid, Waveform, Pima, etc on., and also obtain very good results.

TABLE III

THE TOTAL NUMBER OF FEATURES ELIMINATED IN EACH OF THE 10 SIMULATIONS, THE AVERAGE NUMBER AND ITS STANDARD DEVIATION, USING RELIEF RANKING METHODS FOR THE ECOLI DATA SET.

	Classes	The number of features deleted in each of the 10 simulations (RELIEF)	Average and Std (RELIEF)
Class-independent	all classes	0 0 0 0 1 0 0 1 0 0	0.2±0.4
Class-dependent	class 1	2 2 2 2 3 2 2 2 3 2	2.2±0.4
	class 2	3 5 5 3 5 3 1 2 5 4	3.6±1.4
	class 3	0 1 1 1 0 2 0 0 1 1	0.7±0.7
	class 4	1 3 2 4 5 2 2 2 2 1	2.4±1.3
	class 5	3 4 3 2 1 4 0 4 4 1	2.6±1.5
	class 6	0 0 0 4 2 2 5 1 0 0	1.4±1.8
	class 7	6 6 6 6 6 5 6 6 6 6	5.9±0.3
	class 8	6 6 6 6 6 6 6 6 6 6	6.0±0.0

TABLE IV

THE TOTAL NUMBER OF FEATURES ELIMINATED IN EACH OF THE 10 SIMULATIONS, THE AVERAGE NUMBER AND ITS STANDARD DEVIATION, USING CSM RANKING METHODS FOR THE ECOLI DATA SET.

	Classes	The number of features deleted in each of the 10 simulations (CSM)	Average and Std (CSM)
Class-independent	all classes	0 0 0 0 0 1 0 0 0 0	0.1±0.3
Class-dependent	class 1	0 0 0 0 1 1 0 0 1 0	0.3±0.5
	class 2	0 0 0 0 0 0 0 0 0 0	0
	class 3	0 0 1 1 1 1 0 1 1 1	0.7±0.5
	class 4	2 4 1 4 2 4 4 5 4 4	3.4±1.3
	class 5	2 1 3 0 0 3 1 2 2 3	1.7±1.2
	class 6	4 4 4 4 4 2 4 4 4 4	3.8±0.6
	class 7	6 6 6 6 6 4 6 6 6 6	5.8±0.6
	class 8	6 6 6 6 6 6 6 6 6 6	6.0±0

TABLE V

CLASSIFICATION ACCURACY COMPARISON WITH OTHER METHODS FOR BREAST CANCER AND ECOLI DATA SET. OUR METHODS USE RELIEF AND CSM RANKING MEASURES (10-FOLD CROSS VALIDATION), IN WHICH '–' MEANS UNAVAILABLE. CDFS IS ABBREVIATION OF CLASS-DEPENDENT FEATURE SELECTION. CIFS IS ABBREVIATION OF CLASS-INDEPENDENT FEATURE SELECTION.

Method	Accuracy for Breast cancer data	Accuracy for Ecoli data
Without feature selection	96.49 %	86.61 %
CIFS with RELIEF	96.34 %	85.71%
CDFS with RELIEF	96.34%	86.31%
CIFS with CSM	96.49%	86.61%
CDFS with CSM	96.49%	86.91%
SOAP with C4.5[21]	94.84%	–
RELIEFF with C4.5[21]	95.02%	–

IV. SUMMARY

In this paper, we have proposed a *novel* SVM with *class-dependent* features selected from our wrapper approach. We try the RELIEF weight measure and class separability measure to rank the features importance and select the optimal feature subsets for each class with normal SVM. For the classification on the class-dependent feature subsets, we constructed a novel SVM classifier with different models to accommodate different feature subsets. The experimental results for UCI data sets show that each feature has different classification capability in discriminating different classes, and our method has improved or at least maintained the classification accuracy, while reducing the number of features. Hence both our *class-dependent* feature selection and SVM are very effective in deleting irrelevant or redundant features and realize the classification with class-dependent features.

Since class-dependent feature selection method requires determination of feature subsets of every class, it is likely to be more computationally expensive compared to other class-independent feature selection methods. However the extra computational cost may be worthwhile in certain applications where improvements of accuracy are very important and meaningful. When class-dependent feature selection is used on data, our SVM with class-dependent features can be used conveniently. Further research on feature selection and classification for biological data is currently in progress, such as choosing 'tagging' single nucleotide polymorphisms (SNP) [29] to classify the ethnic groups with individual genotype data.

REFERENCES

[1] A. Statnikov, CF Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", *Bioinformatics*, vol. 21, 2005, pp. 631-643.

[2] Brown et al., "Knowledge-based Analysis of Microarray Gene Expression Data Using Support Vector Machines", *Proceedings of the National Academy of Sciences*, vol. 97, 2000, pp 262-267.

[3] C.L. Blake and C.J. Merz. (1998) *UCI repository of machine learning databases*, Technical report, Department of Information and Computer Science, University of California, Irvine, CA. [Online]. Available at: `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[4] C.-C. Chang and C.-J. Lin. (2001) *LIBSVM: a library for support vector machines*. [Online]. Available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] C.-W. Hsu and C.-J. Lin, "A Comparison of methods for multiclass support vector machines", *IEEE Trans. Neural Network*, vol. 13, 2002, pp 415-425.

[6] H.L. Walker,Jr., W. Lut, M. Dan, E. Mark, S. Rizly and A. Frances, "Applying support vector machines to breast cancer diagnosis using screen film mammogram data", *17th IEEE Symposium on Computer-Based Medical Systems*, 2004, pp. 224.

[7] I.S. Oh, J.S. Lee and C.Y. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, 1999, pp 1089-1094, .

[8] I.S. Oh, J.S. Lee and C.Y. Suen, "Using class separation for feature analysis and combination of class-dependent features", *in Fourteenth International Conference on Pattern Recognition*, vol. 1, 1998, pp 453-455.

[9] J. Casbon, "Protein secondary structure prediction with support vector machines", *M.Sc. thesis, Univ. Sussex*, Brighton, U.K., 2002.

[10] J.C. Platt, N. Cristianini and J. Shawe-Taylor, "Large margin DAGs for multiclass classification", *Advances in Neural information Processing Systems*, MIT press, vol.12,2000, pp. 547-553.

[11] J. H. Geroge, "Enhancements to the data mining process", *Computer science department, School of Engineering, Standford University*, 1997.

[12] K. Kira and L.A. Rendell, "The feature selection problem: Traditional methods and a new algorithm", *in: Proc. of AAAI'92*, 1992, pp 129-134.

[13] L.P. Wang, Editor, *Support Vector Machines: Theory and Applications*, Springer, 2005.

[14] L.P. Wang and X.J. Fu, *Data Mining with Computational Intelligence*, Springer, Berlin; 2005.

[15] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. ackinger, P. Simard and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition", *Int. Conf. Pattern Recognition*, 1994, pp 77-87.

[16] L.H. Chen and S. Chang, "An adoptive learning algorithm for principle component analysis", *IEEE Trans. Neural Networks*, vol. 6, 1995, pp 1255-1263.

[17] M. Last, A. Kandel and O. Maimon, "Information-theoretic algorithm for feature selection", *Pattern Recognition Letters*, vol. 22, 2001, pp 799-811.

[18] P. Langely and S. Sage, "Induction of Selective Bayesian Classifiers", *Proc. of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1999, pp 399-406.

[19] P. Devijver and J. Kittler, *Pattern recognition: a statistical approach*. Prentice-Hall Int., 1982.

[20] Ran Gilad-Bachrach, Amir Navot and Naftali Tishby, "Margin based feature seleciton- theory and algorithms", *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada; 2004.

[21] Roberto Ruiz, Jesus S. Aguilar-Ruiz and Jose C. Riquelme, "SOAP: efficient feature selection of numeric attributes", *IBERAMIA*, 2002, pp 233-242.

[22] S. Oliver, S. Ingolf and L. Thomas, "Local protein structure prediction using discriminative models", *BMC Bioinformatics*, vol.7, 2006.

[23] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach", *Journal of Molecular Biology*, vol. 308, 2001, pp 397-407.

[24] U. Kressel., "Pairwise classification and support vector machines", *Advances in Kernel Methods:Support Vector Learning*, MIT Press. Cambridge, MA, 1999, pp 255-268.

[25] V. Vapnik, *Statistical learning theory*, Wiley, NewYork; 1998.

[26] W. Malina, "Two-parameter fisher criteria", *IEEE Trans. Syst. Man, and Cyber.-B:Cybern.*, vol.31, 2001, pp.629-636.

[27] X.J. Fu and L.P. Wang, "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance", *IEEE Trans. Syst., Man, Cybern.–B: Cybern.*, vol. 33, 2003, pp 399-400.

[28] X.J. Fu and L.P. Wang, "A GA-based novel RBF classifier with class-dependent features", *in: Proc. 2002 Congress on Evolutionary Computation*, vol. 2, 2002, pp 1890-1894.

[29] The International HapMap Consortium, "Integrating ethics and science in the International HapMap Project", *Nature Reviews Genetics*, vol. 5, 2004, pp 467-475.