論 文
7-1-7

A Study on Cancer Diagnostic System Using a Fusion Method based on Genetic Algorithm and Support Vector Machine
(GA와 SVM에 근거한 Fusion Method을 이용한 암 진단시스템에 관한 연구)

하남(Ha-Nam Nguyen)[1] 최규석(Gyoo-Suck Choi)[2]

Abstract

Proteome patterns reflect the underlying pathological state of a human organ. It is believed that the anomalies or diseases of human organs are identified by the analysis of the pattern. There are many ways to analysis these patterns. <중략> (colon cancer and leukemia dataset) indicates that the proposed method shows better classification performance and more stable results than other single kernel functions.

Keywords Support Vector Machine, Machine learning, classification method, Genetic Algorithm.

요약

혈액에서 추출된 프로테옴 패턴(단백질 DNA 정보)는 인간 신체 기관의 병리학적 상태를 잠재적으로 반영하고 있다. 신체기관의 질병이나 이상은 이러한 프로테옴 패턴의 분석에 의해 식별될 수 있다고 알려져 있으며 프로테옴 패턴 정보를 분석하는 여러 가지 방법들이 현재 존재하고 있다. 본 논문에서는 SVM(Support Vector Machine)과 GA(Genetic Algoritm)의 융합에 근거하여 암 진단을 위한 디시전 모델의 효과적 학습(learning) 방법을 제안한다. <중략> 그 결과로서 개별적 kernel function 들보다 더 우수한 분류성능을 갖는 최적의 디시전 모델이 얻어졌다. 위암 데이터 셋 과 두 개의 일반 데이터 셋(대장암, 백혈병)을 사용한 컴퓨터 실험에서 제안된 방법이 다른 Kernel function 들에 비해 더 우수한 분류 성능을 보여주었다.

논문접수 : 2006. 1. 28
심사완료 : 2006. 2. 20

1) 정회원 : 한국 항공대학교 대학원 컴퓨터공학과 박사과정
2) 정회원 : 청운대학교 컴퓨터학과 교수

※ 본 논문은 2005년 청운대학교 학술연구비지원에 의해 연구되었음.

1. Introduction

Support vector machine [1-6] (SVM) is a learning method that uses a hypothesis space of linear functions in a high dimensional feature space. This learning strategy, introduced by Vapnik [2], is a principled and powerful method. In the simplest and linear form, a SVM is the hyperplane that separates a set of positive samples from a set of negative samples with the largest margin. The margin is defined by the distance between the hyperplanes supporting the nearest positive and negative samples. The output formula of a linear case is

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{x} - \mathbf{b}, \qquad (1)$$

where w is a normal vector to the hyperplane and x is an input vector. The separating hyperplane is the plane y = 0 and two supporting hyperplanes parallel to it with equal distances are

$$H_1 : \mathbf{y} = \mathbf{w} \cdot \mathbf{x} - \mathbf{b} = +1$$
$$H_2 : \mathbf{y} = \mathbf{w} \cdot \mathbf{x} - \mathbf{b} = -1 \quad (2)$$

Thus, the margin M is defined as

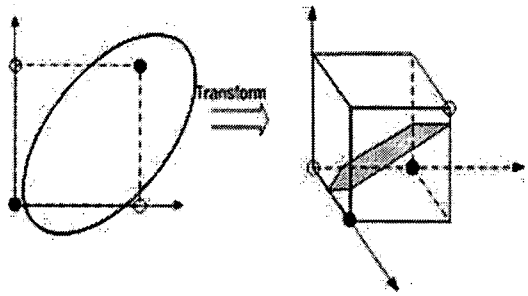
$$M=2/||w|| \qquad (3)$$

In order to find the optimal separating hyperplane having a maximal margin, a learning machine should minimize ||w|| subject to inequality constraints. This is a classic nonlinear optimization problem with inequality constraints. An optimization problem, which can be solved by the saddle point of the Lagrange function, is following

$$L(w,b,\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i y_i ([w^T x + b] - 1)$$

(4)

where i 0 are Lagrange multipliers.



[Fig. 1]An input space can be transformed into a linearly separable feature space by an appropriate kernel function

However, the limitation of computationalpower of linear learning machines was highlighted in the 1960s by Minsky and Papert [7]. It can be easily recognized that real-world applications require more extensive and flexible hypothesis space than linear functions. Such a limitation can be overcome by multilayer neural networks proposed by Rumelhart, Hinton and William [3]. Kernel function also offers an alternative solution by projecting the data into high dimensional feature space to increase the computational power of linear learning machines. Non-linear mapping from input space to high dimensional feature space can be implicitly performed by an appropriate kernel function (see [Fig. 1]). One of the advantages of the kernel method is that a learning algorithm can be exploited to obtain the specifics of application area, which simply can be encoded into the structure of an appropriate kernel function. One of the interesting characteristics on kernel functions. is that a new kernel function can be created by combining a set of kernel functions with the operators such as addition or multiplication operators [1].

Genetic algorithm [8-10] is an optimization

algorithms based on the mechanism of natural genetic procedure. Most of genetic algorithms share a common conceptual base of simulating the evolution of individual structures via the processes of selection, mutation, and reproduction. In each generation, a new population is selected based on the fitness values representing the performances of the individuals belonging to the generation, and some individuals of the population are given the chance to undergo alterations by means of crossover and mutation to form new individuals. In this way, GA performs a multi-directional search by maintaining a population of potential solutions and encourages the formation and the exchange of information among different directions. EAs are generally applied to the problems with a large search space. They are different from random algorithms since they combine the elements of directed and stochastic search. Furthermore, GA is also known to be more robust than directed search methods.

Recently, several researches have been working on GA/SVM to improve the performance of classification. Some of them use GA to optimize the number of selected features that are evaluated by classifiers [13, 14] and the best recognition rate of 80% was achieved in case of colon dataset. Other approach used GA to optimize the ensemble of multiple classifiers to improve the performance of classification [15].

In this paper, we propose a new kernel function combining a set of simple kernel functions for SVM and a method to train the combined kernel function. In the new learning method, GA is exploited to derive the optimal decision model for the classification of patterns, which consists of the optimal set of features and parameters of combined kernel function. The new method was applied to the classification of proteome patterns for the identification of breast cancer, which are extracted from actual clinical samples. Thecombined kernel function and the learning method showed faster convergence and better classification rate than individual kernel functions.

This paper is organized as follows. In section 2, our new combined kernel and its learning method are presented in detail. In section 3, we compare the performances of combined kernel functions and other individual kernel functions by the experiments with the classification of the datasets of colon cancer dataset, leukemia dataset and proteome pattern samples of stomach cancer. Finally, section 4 is our conclusion.

2.The Proposed Learning Method

2.1Overall Structure

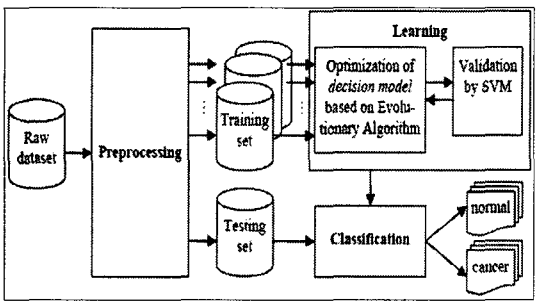
The proposed method is depicted in [Fig. 2] Our method consists of preprocessing, learning, and classification phase.

Firstly, in the preprocessing stage, training and testing sets consisting of a number of cancer and normal patterns is produced and passed to the learning phase.

Secondly, we applied a learning method exploiting GA and SVM techniques to figure out optimal decision model for the classification of proteome patterns in the learning phase. Here GA generates a set of chromosomes, each of which represents a decision model, by evolutionary procedures.

The fitness value of each chromosome is evaluated by measuring the hit ratio from the classification of samples with SVM classifier containing the decision model associated with the chromosome. n-fold validation method is used to evaluate the fitness of a chromosome to reduce overfitting [4]. Then only the chromosomes with a good fitness are selected and given the chance to survive and improve into further generations. Roulette wheel rule is used for the selection of chromosome [8]. Some of the selected chromosomes are given the chance to undergo alterations by means of crossover and mutation to form new individuals. One-point crossover is used, and the probabilities for crossover and mutation are 0.8 and 0.015 in turn. This process is repeated for a predefined number of times. At the end of GA procedure, the decision model with the highest hit ratios is chosen as the optimal decision model.

Finally, the optimal decision modelis used to build a SVM for the classification of novel samples and the performance of the model can be evaluated.



[Fig. 2] Overall framework of Proposed Method

2.2 Learning Combined Kernel Function and Feature Selection Using EA

A kernel function provides a flexible and effective learning mechanism in SVM, and the choice of a kernel function should reflect prior knowledge about the problem at hand. However, it is often difficult for us to exploit the prior knowledge on patterns to choose a kernel function, and it is an open question how to choose the best kernel function for a given data set. According to no free lunch theorem [4] on machine learning, there is no superior kernel function in general, and the performance of a kernel function rather depends on applications.

<Table 1> Kernels are chosen to experiments in our study

Kernel function	Formula
Inverse Multi-Quadric	$\frac{1}{\sqrt{\|x-y\|^2+c^2}}$
Radial	$e^{(-\gamma\|x-y\|^2)}$
Neural	$\tanh(s\cdot\langle x,y\rangle - c)$

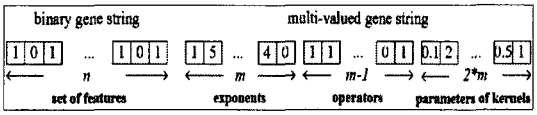
In our case, a new kernel function is created by combining the set of kernel functions (see <Table 1>). The combined kernel function has the form of

$$K_{Combined} = (K_1)^{e_1} \circ \cdots \circ (K_m)^{e_m} \quad (5)$$

where {Ki | i =1, ■, m} is the set of kernel functions to be combined, ei is the exponent of i-th kernel function, and ? denotes an operator between two kernel functions. In our case, three types of the kernel functions listed in <Table 1> are combined, and multiplication or addition operators are used to combine kernel functions.

The parameters in a kernel function play the important role of representing the structure of a sample space. The set of the

parameters of a combined kernel function consists of three part - i) the exponents of individual kernel functions, ii) the operators between kernel functions, iii) the coefficient in each kernel function. In thelearning phase, the structure of a sample space is learned by a kernel function, and the knowledge of a sample space is contained in the set of parameters. Furthermore, the optimal set of features should be chosen in the learning phase. In our case, GA technique is exploited to obtain the optimal set of features as well as the optimal combined kernel function.



[Fig. 3] Structure of a chromosome used in GA processing

The challenging issue of GA is how to map a real problem into a chromosome. In our learning method, we need to map feature space, the set of the parameters for kernels, and the set of operators combining kernels. Firstly, the set of features is encoded into a n-bit binary string to represent an active or non-active state of n features. Then the exponents of mindividual kernel functions, the operators between individual kernel functions, and the coefficients in each individual kernel function are encoded into a multi-valued gene string. The combination of the two gene string forms a chromosome in GA procedure which in turn serves as a decision model (see [Fig. 3]). In learning phase, simulating a genetic procedure, GA creates improved decision models containing a combined kernel function and a set of features by the iterative process of reproduction, evaluation, and selection process. At the end of learning stage, the optimal decision model consisting of a combined kernel functionand the set of features is obtained, and the optimal decision model is contained in a classifier to be used classify new pattern samples.

3. Experiment Results

In this section, we show the result from the classification based on the model trained by our learning method. Furthermore, the performance of the classification model with combined kernel function is compared to the performancesof the models with other kernel functions.

3.1 Dataset Descriptions

There are several microarray dataset from published cancer gene expression studies. In this paper, we have used two representative datasets (conlon cancer and leukemia datasets) among them and our own dataset (proteome patterns of stomach cancer dataset).

The stomach cancer dataset consists of 137 samples taken from stomach cancer patients. There are 67 normal and 70 cancer samples. The random 72 samples out of 137 samples were used as training set. The remaining 65 samples are used as test samples. The proteome pattern samples used for training and testing the models are provided by Cancer Research Center of Seoul National University in Seoul, Korea. The proteome patterns are extracted from the sera of the affected and the unaffected men. The proteome patterns of affected are extracted from the sera of the patients diagnosed with

stomachcancer at the gynecological clinic in the hospital. We chose 119 spots as the candidate features to be included in the optimal classification model. The normalized quantity values of 119 spots from each proteome image are extracted grouped into a feature vector representing the proteome pattern of a sample. <Table 2> shows the distribution of patients and normal individuals in our samples. <Table 3> shows the distribution of the ages of cancer patients and normal individuals.

<Table 2> Distribution of men and women in dataset

	Cancer group	Normal group	All
Man	34	32	66
Woman	36	35	71
All	70	67	137

<Table 3>Distribution of the ages of samples Age groupCancer groupNormal groupAgeless than 4086

40~60	33	34
more than 60	29	27

The colon cancer dataset [11] contains gene expression information extracted from DNA microarrays. The 62 samples dataset consists of 22 normal and 40 cancer tissue samples and each having 2000 features. 32 samples were chosen randomly as training set and the remaining samples were used as testing set. (Availble at: http://sdmc.lit.org.sg/GEDatasets/Data/ColonTumor.zip).

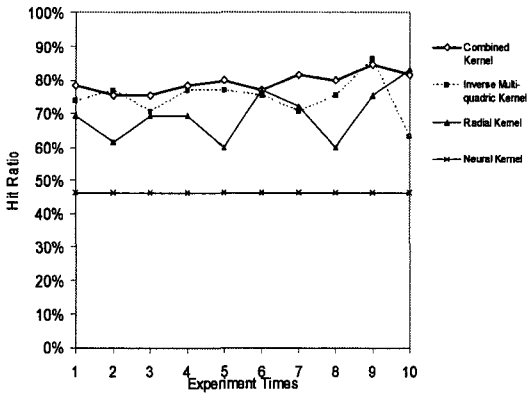
The leukemia dataset [12] consists of 72 samples that have to be discriminated into two classes ALL and AML. There are 47 ALL and 25 AML samples and each sample contains 7129 features. The dataset was divided into a training set with 38 samples (27 ALL and 11 AML) and a test set with 34 samples (20 ALL and 14 AML) (Availble at: http://sdmc.lit.org.sg/GEDatasets/Data/ALL-AML_Leukemia.zip).

3.2 Experimental Environments

All experiments are conducted on a Pentium IV 1.8GHz computer. The experiments are composed preprocessing, learning by GA to obtain optimal decision model, and classification (see Sec. 2). For preprocessing data, we normalize data and randomly build 10 pair of training/testing dataset. For EA, we have used tournament rule for selection method. Our proposed method was done with 100 of generations and 100 of populations. Our combined kernel function and three other kernel functions <Table 1> are trained by GA in learning phase with training set. Three kernel functions are chosen since they were known tohave good performances in bioinformatics field [4, 6, 13-15]. Also, 10-fold cross validation is used for the fitness estimating to reduce overfitting problem [4]. The optimal decision model obtained after 100 generations of GA is used toclassify the set of test samples. The experiments for each kernel function are repeated for 10 times to obtain generalized results. As the result of learning phase, an optimal decision model consisting of 15 most important features and optimal combined kernel function is obtained.

3.3 Experimental Results and Analysis

The classified results of stomach cancer dataset are shown in [Fig. 4] The graph indicated the proposed method with combined kernel function more stable than other cases. The average, highest, and lowest of hit ratios using four kernel functions are shown in <Table 4>. Here the combined kernel function case shows the best average performance with 79.33% correct classification. The standard deviation of it (see <Table 4>) also better than other cases.

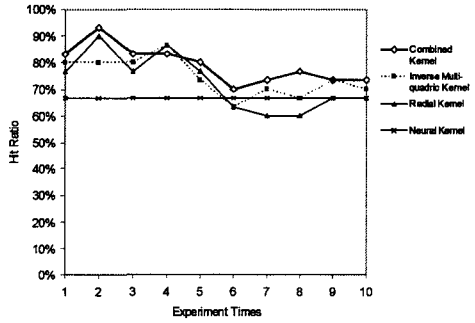


[Fig. 4] The comparison of hit rate in classification phase of combined kernel function case (bold line) with single kernel functions in stomach cancer dataset

<Table 4> The hit ratio of classification phase using the decision model obtained from 100 generations of GA in stomach cancer dataset

Kernel function name	Accuracy ± stdev
Combination	79.23±3.0%
Inverse Multi-quadric	74.62±6.0%
Radial	69.69±8.0%
Neural	46.15±0.0%

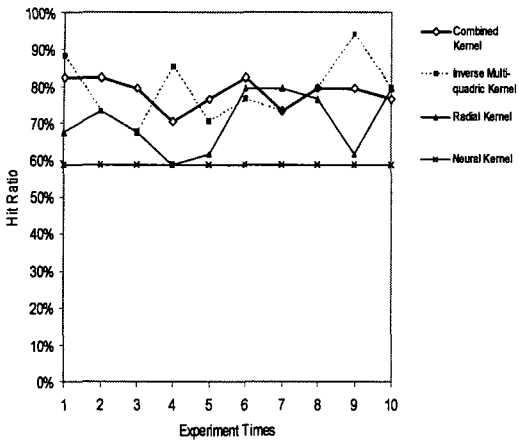
In the case of colon dataset, the experiments of proposed method with combined kernel function also show more stable and higher than other cases (see [Fig. 5]). According to <Table 5>, the result of combined kernel function case shows the best average performance with 79.0% of recognition rate with lower standard deviation values compared to single cases.



[Fig. 5] The comparison of hit rate in classification phase of combined kernel function case (bold line) with single kernel functions in colon cancer dataset

<Table 5> The hit ratio of classification phase using the decision model obtained from 100 generations of GA in colon cancer dataset

Kernel function name	Accuracy ± stdev
Combination	79.0±7.0%
Inverse Multi-quadric	74.33±7.0%
Radial	72.33±11.0%
Neural	66.67±0.0%



[Fig. 6] The comparison of hit rate in classification phase of combined kernel function case (bold line) with single kernel functions in leukemia dataset

In the case of leukemia dataset, the classified results of experiments with combined kernel function still seem more stable and higher than other single kernel function (see [Fig. 6]). The average, highest, and lowest of hit ratios using four kernel functions are shown in <Table 6>. The table shows us the best average is 78.82% of recognition rate in case of Inverse Multi-quadric kernel function. In this dataset, even though combined kernel function could not obtain the best average of corrected classification, but the results still stable than other cases (see <Table 6> - stdev=3%).

<Table 6> The hit ratio of classification phase using the decision model obtained from 100 generations of GA in leukemia dataset

Kernel function name	Accuracy ± stdev
Combination	78.24±4.0%
Inverse Multi-quadric	78.82±9.0%
Radial	70.57±8.0%
Neural	58.82±0.0%

4. Conclusion

In this paper, we proposed a new kernel function combining a set of kernel functions for SVM and its learning method exploiting GA technique to obtain the optimal decision model for classification. A kernel function plays the important role of mapping the problem feature space into a new feature space so that the performance of the SVMclassifier is improved. The combined kernel function and the learning method were applied to classify the clinical datasetsto identify cancer/normal groups. In the comparison of the classifications by combined kernel and other three kernel functions, the combined kernel function achieved the fastest convergence in learning phase and results in the optimal decision model with the highest hit rate in classification phase. Thus our combined kernel function has greater flexibility in representing a problem space than individual kernel functions.

Acknowledgment

The authors acknowledge that mySVM [16] was used for the implementation of the proposed method in this paper.

References

[1]N. Cristianini and J. Shawe-Taylor, "An introduction to Support Vector Machines and other kernel-based learning methods," Cambridge, 2000.

[2]V.N. Vapnik et. al., "Theory of Support Vector Machines," Technical Report CSD TR-96-17, Univ. of London, 1996.

[3]Vojislav Kecman, "Learning and Soft Computing: Support Vector Machines, Neural

Networks, and Fuzzy Logic Models (Complex Adaptive Systems)," The MIT press, 2001.

[4]Richard O. Duda, Peter E. Hart, David G. Stork, "Pattern Classification (2nd Edition)," John Wiley & Sons Inc., 2001.

[5]Joachims, Thorsten, "Making large-Scale SVM Learning Practical. In Advances in Kernel Methods - Support Vector Learning," chapter 11. MIT Press, 1999.

[6]Bernhard Schökopf , Alexander J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)," MIT press, 2002

[7]M.L. Minsky and S.A. Papert, "Perceptrons," MIT Press, 1969.

[8]Z.Michalewicz, "Genetic Algorithms + Data structures = Evolution Programs," Springer-Verlag, 3 re rev. and extended ed., 1996.

[9]D. E. Goldberg, "Genetic Algorithms in Search, Optimization & Machine Learning," Adison Wesley, 1989.

[10]Melanie Mitchell, "Introduction to genetic Algorithms," MIT press, fifth printing, 1999.

[11]U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proceedings of National Academy of Sciences of the United States of American, vol 96, pp. 6745-6750, 1999.

[12]T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, pp. 531537, 1999.

[13]H. Frohlich, O. Chapelle, B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm, Tools with Artificial Intelligence," Proceedings. 15th. IEEE International Conference, pp.142 148, 2003.

[14]Xue-wen Chen, "Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines, The Computational Systems," Bioinformatics Conference. Proceedings IEEE International Conference, pp. 505, 2003.

[15]Chanho Park and Sung-Bae Cho, "Genetic search for optimal ensemble of feature-classifier pairs in DNA gene expression profiles," Neural Networks, 2003. Proceedings of the International Joint Conference,pp. 1702 1707, 2003.

[16]Stefan Rüping, "mySVM-Manual," University of Dortmund, Lehrstuhl Informatik, 2000. [online]
(Available:
http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/)

응우옌하남

1998년 베트남 국립 하노이대학교 정보공학과 졸업 (공학사)
2003년 청운대학교 전자정보대학원 졸업 (공학석사)
2003년 ~ 현재 한국항공대학교 대학원 컴퓨터공학과 박사과정
관심분야 : AI 및 Data Mining, Machine learning, soft computing 등

최 규 석

1982년 연세대학교 전기공학과를 졸업 (공학사)
1987년 ~ 1997년 동 대학원 공학석사 및 공학박사
(주) 데이콤 연구소 연구원
(주) SK텔레콤 중앙연구소 책임연구원
1997년 ~ 현재 청운대학교 컴퓨터과학과 부교수
관심분야 : AI 및Mobile Computing, 이동통신, 인공생명 등