A Novel Feature Selection for Gene Expression Data

Li-Yeh Chuang[1]*, Cheng-Hong Yang[2], Chung-Jui Tu[2], and Cheng-Huei Yang[3]

[1]Department of Chemical Engineering, I-Shou Univeristy.
[2]Department of Electronic Engineering, National Kaohsiung Univeristy of Applied Sciences.
[3]Department of Electronic Communication Engineering, National Kaohsiung Marine University, Kaohsiung, Taiwan

Abstract

The feature selection process can be considered a problem of global combinatorial optimization in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, and results in an acceptable classification accuracy. Therefore, a good feature selection method based on the number of features investigated for sample classification is needed in order to speed up the processing rate, predictive accuracy, and to avoid incomprehensibility. In this paper, particle swarm optimization (PSO) is used to implement a feature selection, and the K-nearest neighbor (K-NN) method with leave-one-out cross-validation (LOOCV) serves as an evaluator of PSO. The support vector machines (SVMs) with the one-versus-rest method serve as a classifier for the classification problem. Experimental results show that our method simplifies features effectively and obtains a higher classification accuracy compared to the other classification methods from the literature.

Keywords: Gene Expression Data, Particle Swarm Optimization, Support Vector Machines, Kernel-Adatron, One-Versus-Rest.

1. Introduction

DNA microarray is a biotechnological method that can rapidly measure expressions of several thousands of genes in a single experiment. The application of microarray data on classification of cancer types has become popular recently. Coupled with statistical techniques, gene expression patterns have been used in screening for potential tumor markers.

Gene expression data characteristically have a high dimension and few specimens, which makes it difficult for a general classification method to be trained and tested. Therefore, obtaining correct classification is difficult. In order to analyze gene expression profiles correctly, feature (gene) selection is most crucial for the classification process.

Several methods were used to perform feature selection on the training and testing data, for example genetic algorithms [1], branch and bound algorithms [2], mutual information [3], and tabu search [4]. In this paper, particle swarm optimization (PSO) is used to implement a feature selection, and the K-nearest neighbor (K-NN) method with leave-one-out cross-validation (LOOCV) serves as an evaluator of PSO. The support vector machines (SVMs) with the one-versus-rest method serve as a classifier for the classification problem. Experimental results show that our method simplifies features effectively and obtains a higher classification accuracy compared to the other classification methods from the literature.

2. Methods

2.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique, which was developed by Kennedy and Eberhart in 1995 [5]. PSO simulates the social behavior of organisms, such as bird flocking and fish schooling, to describe an automatically evolving system. In PSO, each single candidate solution can be considered "an individual bird of the flock", that is, a particle in the search space. Each particle makes use of its own memory and knowledge gained by the swarm as a whole to find the best solution. All of the particles have fitness values, which are evaluated by a fitness function to be optimized; they also have velocities which direct the movement of the particles. During movement, each particle adjusts its position according to its own experience and according to the experience of a neighboring particle, thus making use of the best position encountered by itself and its neighbor. The particles move through the problem space by following a current of optimum particles.

In this paper, a binary version of a PSO algorithm is used [6]. The position of each particle is given in a binary string form, which represents the feature selection situation.

2.2 Support Vector Machines

Support Vector Machines (SVMs) were originally introduced by Vapnik and co-workers [7] for classification tasks, and were subsequently extended to regression problems [8]. The idea behind SVMs is the following: input points are mapped to a high dimensional feature space, where a separating hyper-plane can be found. The algorithm is chosen in such a way as to maximize the distance from the closest patterns, a quantity which is called the margin. SVMs are learning systems designed to automatically trade-off accuracy and complexity by minimizing an upper bound on the generalization error provided by the Vapnik-Chervonenkis (VC) theory [9]. In a variety of classification problems, SVMs have shown a performance which can reduce training and testing errors, thereby obtaining a higher recognition accuracy. SVMs can be applied to very high dimensional data without changing their formulation.

In this study, Kernel-Adatron (KA) algorithms [7], are used to emulate SVM training procedures. By introducing Kernels into the algorithm it is possible to find a maximal margin hyper-plane in a high feature space, which is equivalent to nonlinear decision boundaries in the input space. In this study, the kernel function is used for the Radial Basis Function (RBF). C and r are used to control the trade-off between training error and generalization ability. The decomposition techniques used for KA are one-versus-rest.

2.3 One-Versus-Rest

The one-versus-rest method assembles classifiers that distinguish one from all the other classes. For each i, $1 \le i \le k$, a binary classifier separating class i from the rest is built. To predict a class label of a given data point, the output of each of the k classifiers is obtained. If there is a unique class label, say j, which is consistent with all the k predictions, the data point is assigned to class j. Otherwise, one of the k classes is selected randomly. Very often though, a situation arises in which consistent class assignment does not exist, which could potentially lead to problems [10].

2.4 PSO-SVM

Based on the rules of particle swarm optimization, we set the required particle number first, and then the initial coding alphabetic string for each particle is randomly produced. In our case we coded each particle to imitate a chromosome in a genetic algorithm; each particle was coded to a binary alphabetic string $S = F_1 F_2 \ldots F_n$, $n = 1,\ 2,\ \ldots,\ m$; the bit value {1} represents a selected feature, whereas the bit value {0} represents a non-selected feature.

The adaptive functional values were data based on the particle features representing the feature dimension; this data was measured by leave-one-out cross-validation of a nearest neighbor (1-NN). The obtain feature subset by PSO was classified by a support vector machine (SVM) to obtain classification accuracy. SVM can decrease the training error and testing error, and increase the classification accuracy. The accuracy for the SVM evolves according to the K-fold Cross-Validation Method for small sample sizes, and according to the Holdout Method for big sample sizes [11].

Each particle renewal is based on its adaptive value. The best adaptive value for each particle renewal is *pbest*, and the best adaptive value within a group of *pbest* is *gbest*. Once *pbest* and *gbest* are obtained, we can keep track of the features of *pbest* and *gbest* particles with regard to their position and speed. Each particle is updated according to the following equations.

$$v_{pd}^{new} = w \times v_{pd}^{old} + c_1 \times rand_1 \times \left(pbest_{pd} - x_{pd}^{old}\right) + c_2 \times rand_2 \times \left(gbest_d - x_{pd}^{old}\right) \quad (1)$$

$$S\left(v_{pd}^{new}\right) = \frac{1}{1 + e^{-v_{pd}^{new}}} \quad (2)$$

$$\text{if } \left(rand < S\left(v_{pd}^{new}\right)\right) \text{ then } x_{pd}^{new} = 1 \text{; else } x_{pd}^{new} = 0 \quad (3)$$

The feature after renewal is calculated by the function $S(v_{pd}^{new})$ (Eq. 2), in which the speed value is v_{pd}^{new}. If $S(v_{pd}^{new})$ is larger than a randomly produced disorder number that is within (0, 1), then its position value F_n, $n = 1,\ 2,\ \ldots,\ m$ is represented as {1}. If $S(v_{pd}^{new})$ is smaller than a randomly produced disorder number that is within {0~1}, then its position value F_n, $n = 1,\ 2,\ \ldots,\ m$ is represented as {0}.

The inertia weight w was 0.9. The two factors $rand_1$, $rand_2$ and $rand$ are random numbers between (0, 1), whereas c_1 and c_2 are learning factors, usually $c_1 = c_2 = 2$.

3. Results and Discussion

In this study, the gene expression data were downloaded from http://www.gems-system.org. Presently, there is no standard for pre-processing of gene expression data in the microarray technique. Therefore, we used a standard normalization form to reduce all of the gene expression values to between 0 and 1, so as to effectively reduce the SVM training error, thereby improving accuracy for the classification problem. The data format was arranged as shown in Table 1 [12].

In this paper, binary PSO is used to serve as feature selection for gene expression data. It helps to improve the performance owing to its small number of simple parameter settings. A KA-SVM is used to classify the feature subset of the PSO, which can be obtained by comparing the characteristics of the general test data. The gene expression data have a fairly small sample size and high dimension. The SVM can be applied to very high-dimensional data by introducing a Kernel function to find a maximal margin hyperplane in a high feature space that is well suited to the gene expression data structure. At the same time, it reduces the amount of training and testing, thereby increasing the classification accuracy for gene expression data.

Table 1 shows that the number of necessarily selected features can be much reduced by the proposed method. This helps to increase the classification accuracy. Table 2 compares experimental results obtained by methods from the literature and the proposed method. The proposed method obtained the highest classification accuracy for the 9_Tumors, Brain_Tumor2, and Leukemia1 data sets. The classification accuracy of the 9_Tumors and Brain_Tumor2 data sets are 70.00% and 84.00%, respectively, an increase of 5% classification accuracy compared to methods using Non-SVMs and MC-SVMs. For the SRBCT data set, both the MC-SVM and the proposed method obtained 100% classification accuracy. However, the number of features selected is less in the proposed method. This means that not all features are needed to achieve total classification accuracy. For the data sets of 11_Tumors the classification accuracy is better than the classification accuracy of Non-SVMs and is comparable to the MC-SVM method. These results indicate that for gene expression data classification problems, the proposed method (binary particle swarm optimization) can severe as a pre-processing tool and help optimize the feature selection process, which leads to an increase in classification accuracy. A good feature selection process reduces feature dimensions and improves accuracy.

The parameters used in PSO are fewer. However, if the proper parameter values are set, the results can easily be optimized. Proper adjustment of the inertia weight w and the acceleration factors c_1, c_2 is very important. If the parameter adjustment is too small, the particle movement is too small. This will also result in useful data, but is a lot more time-consuming. If the adjustment is excessive, particle movement will also be excessive, causing the algorithm to weaken early, so that a useful feature set can not be obtained. Hence, suitable parameter adjustment enables particle swarm optimization to increase the efficiency of feature selection. For SVMs, correct parameter adjustment is crucial, since many parameters are involved. This can have a profound influence on the results. For different gene expression data, different parameters have to be set for SVMs. The two factors r and C are especially important. A suitable adjustment of these parameters results in a better classification hyperplane found by SVM, and thereby enhances the classification accuracy. Bad parameter settings affect the classification accuracy negatively. In this paper, we used the parameters $r = 2^{-10}$, $C = 2^{12}$, and $\eta = 0.1$ for all gene expression data. The parameters settings used in our study were optimized, and could be used as a reference for future studies.

4. Conclusions

In this paper, we used PSO to perform feature selection and 1-NN serve as fitness function of PSO. The obtain feature subset by PSO was classified by a support vector machine (SVM) with the one-versus-rest method to obtain classification accuracy for five gene expression profiles. Experimental results show that our method simplified feature selection and the total number of parameters needed effectively, thereby obtaining a higher classification accuracy compared to other classification methods. The results obtained by the proposed method obtained the highest classification accuracy in four of five samples tested. The proposed method can serve as an ideal pre-processing tool in other areas to help optimize the feature selection process.

5. Acknowledgements

This work is partly supported by National Science Council in Taiwan under grant NSC94-2622-E-151-025-CC3, NSC94-2311-B037-001, NSC93-2213-E-214-037, NSC92-2213-E-214-036, and by CGMH fund CMRPG1006.

6. References

[1] Yang, J.H. and Honavar, V. (1998) Feature Subset Selection Using a Genetic Algorithm. IEEE Intelligent Systems, vol. 13, no. 2, pp. 44-49.

[2] Narendra, P.M. and Fukunage, K. (1997) A Branch and Bound Algorithm for Feature Subset Selection. IEEE Trans. Computers, vol.6, no. 9, pp. 917-922, Sept.

[3] Roberto B. (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5(4):537-550.

[4] Zhang, H. and Sun, G. (2002) Feature selection using tabu search method. *Pattern Recognition*, 35:

701-711.

[5] Kennedy, J. and Eberhart, R.C. (1995) Particle swarm optimization. in proceedings of the 1995 IEEE International Conference on Neural Networkds, volume 4, pages 1942-1948, Perth, Australia.

[6] Kennedy, J. and Eberhart, R.C. (1997) A discrete binary version of the particle swarm algorithm. Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'. 1997 IEEE International Conference on Volume 5, Oct. 12-15, pp. 4104 – 4108.

[7] Frieß, T., Cristianini, N., and Campbell, C. (1998) The Kernel-Adatron: a Fast and Simple Learning Procedure for Support Vector Machines. *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 188-196.

[8] Drucker, H., Burges, C., Kaufman, L., Smola, A. and Vapnik, V. (1997) Support Vector Regression Machines, In: Mozer, M., Jordan, M. and Petsche, T. (ed.), Neural Information Processing Systems, Vol. 9. MIT Press, Cambridge, MA, 155-161.

[9] Vapnik, V.N. (1995) The Nature of Statistical Learning Theory, Springer Verlag, New York.

[10] Scholkopf, B. and Smola, A. (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA.

[11] Stone, M. (1974) Cross-Validation choice and assessment of statistical predictions. Journal of the Royal Statistical Society B, 36, pp.111-147.

[12] Statnikov, A., Aligeris, C.F., Tsamardinos, L., Hardin, D., Levy, S. (2004) A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. in Bioinformatics, vol. 21(5), Sep., pp.631-643.

Table 1. Format of gene expression classification data

Dataset Name	Number of					Diagnostic task
	Samples	Categories	Variables (Samples)	Variables (genes)	Variables (genes) Selected	
9_Tumors	60	9	95	5726	2972	9 various human tumor types
11_Tumors	174	11	72	12533	6465	11 various human tumor types
Brain_Tumor2	50	4	207	10367	5516	4 malignant glioma types
Leukemia1	72	3	74	5327	2688	Acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B-cell and ALL T-cell
SRBCT	83	4	28	2308	1252	Small, round blue cell tumors (SRBCT) or childhood

Table 2. Accuracy of classification for gene expression data

Methods Datasets	*Non-SVM*			*MC-SVM*					*PSO-SVM*
	KNN	*NN*	*PNN*	*OVR*	*OVO*	*DAG*	*WW*	*CS*	*OVR*
9_Tumors	43.90	19.38	34.00	65.10	58.57	60.24	62.24	65.33	**70.00**
11_Tumors	78.51	54.14	77.21	94.68	90.36	90.36	94.68	**95.30**	95.04
Brain_Tumor2	68.67	60.33	62.83	77.00	77.83	77.83	73.33	72.83	**84.00**
Leukemia1	83.57	76.61	85.00	97.50	91.32	96.07	97.50	97.50	**98.57**
SRBCT	86.90	91.03	79.50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Average	72.31	60.30	67.71	86.86	83.62	84.90	85.55	86.19	**89.52**

Legends: (1) KNN: K-Nearest Neighbors. (2) NN: Backpropagation Neural Networks. (3) PNN: Probabilistic Neural Networks. (4) OVR: One-Versus-Rest. (5) OVO: One-Versus-One. (6) DAG: DAGSVM. (7) WW: Method by Weston and Watkins. (8) CS: Method by Crammer and Singer. Non-SVM and MC-SVM results taken from Statnikov *et al.* for comparison. PSO-SVM: the proposed method.