First International Symposium on Information and Communications Technologies. October 7-8, 2004. Putrajaya, Malaysia.

FEATURE SELECTION METHOD USING GENETIC ALGORITHM FOR THE CLASSIFCATION OF SMALL AND HIGH DIMENSION DATA

Mohd Saberi Mohamad, Safaai Deris, Safie Mat Yatim, Muhammad Razib Othman
Artificial Intelligent and Bioinformatic Laboratory, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
Tel: 607-5537784, Fax: 607-5565044
Email: berie_ext2@lycos.com, safaai@fsksm.utm.my, safie@fsksm.utm.my, razib@fsksm.utm.my

Abstract:
Practical pattern classification problems require selection of a subset of attributes or features to represent the patterns to be classified. The feature selection process is very important which selects the informative features for used classification process. This is due to the fact that performance of the classifier is sensitive to the choice of the features used to construct the good classifier from small or high dimension data that are inherently noisy. In this paper, we propose an efficient feature selection method that finding and selecting informative features from small or high dimension data which maximum the classification accuracy. In this work, we apply genetic algorithm to search out and identify the potential informative features combinations for classification and then use the classification accuracy from the support vector machine classifier to determine the fitness in genetic algorithm. Experimental results with benchmark datasets show the usefulness of the proposed approach for small and high dimension data.
Keywords: Feature Selection, Genetic Algorithm, Classification, High Dimension Data

1. INTRODUCTION

Many classification tasks require learning of an appropriate classification function that assigns a given input to one of a finite set of classes. Feature selection is known to be a critical step in the design of pattern classifier for several reasons. Features selection methods aim at selecting a small or prespecified number of features leading to the best possible performance of the entire classifier. In many applications (e.g. medical diagnosis, speech recognition, text classification and image recognition) there is a practical need to reduce the number of measurements without significantly degrading the performance of the system [3]. Usually, feature selection methods easy to select subsets of feature from small dimension data (5-60 features) comparing then high dimension data (over 1000 features).

The feature selection refers the task of identifying and selection a useful subset of features to be used to represent patterns from a larger set of often mutually redundant, possibly irrelevant, features with different associated measurement risks [7]. Statistical model fitting or supervised learning systems generally do not have enough labelled training instances to fit accurate models over very large feature spaces, due to finite sample effects [1]. At the same time, in many cases it is difficult or impossible to know without training which features are relevant to a given task and which are effectively noise. As a result, the ability to select features from a huge feature set is critical for computer vision.

This work will propose a method using genetic algorithm to identify subset of features combinations from small or high dimension data for improve the classification accuracy using. The combinations of features are used for classification and then classification accuracy from SVM classifier used to determine the fitness function.

2. RELATED WORK

The idea of applying genetic algorithms as feature selectors is not novel. Of these, Yang and Hanovar[8] investigated combinations of genetic algorithm and neural network. Eads et al.,[2] and Sepulveda-Sanchis et al.,[6] combined genetic algorithm and SVM. Liu et al.,[5] combined the parallel genetic algorithm with classification method proposed by Golub et al.,[11] for gene expression classification. All researches above used the same model of chromosome representation in genetic algorithm values as shown in Figure 1 below.

10101000010000101…0111001011001

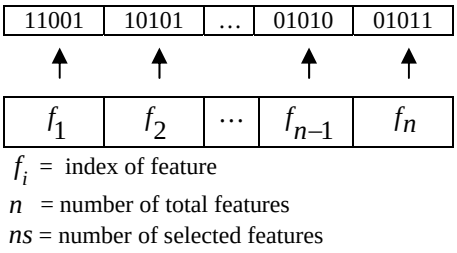
Figure 1. The representation of chromosome in GA used by previous research.

In which a bit value of 1 in the chromosome representation means that the corresponding feature is included in the specified subset, and a value of 0 indicates that the corresponding feature is not included in the subset. The advantage of this representation is that a standard and well understood GA could be used without any modification. Unfortunately, the model of chromosome is only appropriate for data that have small and medium features. It caused an exponential nature of subsets that exist as the number of features increases. So, if the number of features is too large, it is impossible to evaluate all possible combinations of features.

This work extends these and other author's work by modifying the chromosome representation in GA because the data used in this work have thousand of features. We also changed the fitness function, which motivate to get the best accuracy in training data. We replaced the neural network with the SVM to classify the subset of features.

3. PROPOSED APPROACH

Firstly the proposed approach called GASVM will combine GA with SVM without modification of chromosome representation. Secondly the proposed approach called New-GASVM will modify the model of chromosome representation. The overall classification strategy consists of two main components. The two main components are GA for features subset selection and SVM as classifier. Each individual of the current GA population represents a competing subset of features that must be evaluated to provide fitness feedback to the evolutionary process. In this work, we modify the original representation of chromosome suitable for huge-scale features as shown in Figure 2 below.



f_i = index of feature
n = number of total features
ns = number of selected features

Figure 2. The representation of chromosome for huge-scale data.

The new representation of chromosome support large features scale because it can fit the number of features selected. Hence, the New-GASVM is able to reduce the number of feature combinations (subsets).

It includes the real index f_i which indicates a selected feature of the i-th features among total features. Its length can vary according to the size of the number of total features n and the number of selected features ns. The chromosome length is equal size for each chromosome.

The fitness function of an individual is determined by evaluating the SVM. From early works [5], [10], [11], it is obvious that best accuracy for cross-validation from training set used as classification accuracy. So in this paper, we used 1-criteria fitness functions that containing just accuracy for testing data as mentioned in equation (1) below

$$fitness\ (x) = accuracy(x) \quad (1)$$

where $accuracy(x)$ is the cross validation accuracy of the SVM classifier trained using the feature subset of training data represented by x.

4. EXPERIMENTS

To investigate the performance and attributes of proposed approaches for small and high dimension data, some standardized benchmark datasets would be essential. The breast cancer and the leukaemia cancer datasets were selected to evaluate the performance of the approach. The two real world datasets have been used to evaluate some of current best classification algorithms such as Support Vector Machines, Kernel Fisher Discriminant, and AdaBoost [4], [5], [8], [10],[11]. In the experimental results section, the proposed approach will be compared against these powerful learning machines. Table 1 lists vital information pertaining to the two-benchmark datasets.

TABLE 1: BENCHMARK DATASETS DESCRIPTION

Description	**Breast Cancer**	**Leukaemia Cancer**
number of training samples	200	38
number of test samples	77	34
input dimension size	9	7129
number of classes	2	2
(*F*-fold cross validation)	100	38

4.1 Experiment Environment

Our experiments were run using Steady-State GA and roulette wheel selection strategy [12]. In these experiments, we assess two-point crossover and also applied Gaussian mutation operations at probability on each of the offspring strings produced from crossover [13]. The parameter setting in Table 2 is chose base on results of several preliminary runs.

TABLE 2. PARAMETERS OF EXPERIMENTAL ENVIRONMENT

GA Parameter	**Value**	**SVM Parameter**	**Value**
Size of population	100	Cost, C	100
Number of generation	10	Gamma, g	1/k,
Replacement rate	0.8		
Crossover rate	0.7		
Mutation rate	0.01		
Note: The k in the gamma option means the number of attributes in the input data.			

The mean *F*-fold cross validation accuracy in the breast cancer dataset is used to compare the classifier's performance against others. For leukaemia cancer dataset, the accuracy classifier is tested by leave one out cross validation (LOOCV) procedure because to make comparison with early works.

4.2 Result Analysis

From the Table 3 below, it is very interesting to note that New-GASVM performs better than all of them. Second good performs is GASVM. Only 8 features are selected from total features of the data using New-GASVM. The GASVM was selecting the features by randomly. But GASVM only required two features to achieve good (79%) accuracy. All previous works were using whole of features may have uninformative feature that caused the poor classification tasks

TABLE 3: BENCHMARK OF GASVM AND New-GASVM WITH CURRENT BEST CLASSIFIER ON BREAST CANCER DATASET

Classifier	**Cross Validation Accuracy (100 Fold)**	**Number of Features Selected**
GASVM	79	2
New-GASVM	82	8
DBNN	74.6	9
KFD	74.2	9
SVM	74.0	9
AB_R	73.5	9
MLP	72.4	9
RBF	72.4	9
AB	69.6	9

Note:
Classifiers in boldface were experimented in this research. Sources of other classifier results from [4], [8], [9]. Best result shown in shaded cells.

AB : AdaBoost
AB_R : Regularized AdaBoost
DBNN : Denoeux Belief Neural Network
KFD : Kernel Fisher Discriminant
MLP : Multilayer Perceptron
RBF : Single Radial Basis Function
SVM : Support Vector Machines
GASVM : Hybrid GA with SVM (proposed method)
New-GASVM: Hybrid GA with SVM (proposed method)

From the benchmark result in Table 4 it is evident that New-GASVM and LD classifier perform better than all of previous methods. But the GASVM has performed very poorly on the leukaemia cancer data. The GASVM was selecting the features by randomly. But GASVM only required 3568 features to achieve good (94.7368%) accuracy.

The fluctuation from being the best technique in the breast cancer data to the worst in leukaemia cancer data implies that there must be a certain drawback in GASVM that is severely affected when classifying the leukaemia cancer data. Comparing the breast cancer data to the leukaemia cancer data, the only obvious difference is in the dimension size of the data (Table 1). The model of chromosome representation in GASVM may not be able to support high dimension data. The New-GASVM was using new model of chromosome representation to support high dimension data.

TABLE 4: BENCHMARK OF GASVM AND New-GASVM WITH CURRENT BEST CLASSIFIER ON LEUKAEMIA CANCER DATASET

Classifier	**Cross Validation Accuracy (38 Fold)**	**Number of Features Selected**
GASVM	94.7368	3568
New-GASVM	100	50
LD	100	50
GAWV	94.7368	29
WV	94.7368	50

Note:
Classifiers in boldface were experimented in this research. Sources of other classifier results:[6],[10], [11]. Best result shown in shaded cells.

LD : Logistic Discriminant
WV : Weight Voting:
GAWV : Hybrid GA with Weight Voting
GASVM : Hybrid GA with SVM (proposed method)
New-GASVM: Hybrid GA with SVM (proposed method)

5. CONCLUSION

In this paper, we have investigated and solved the problem of conventional approaches for features selection of small or high dimension data and proposed a new approach for this problem. A major goal of this work is to propose an efficient feature selection method that finding and selecting informative features from small or high dimension data which maximum the classification accuracy.

To reflect the traits of feature combination, we used GA evaluated by SVM and selected features are good to get high classification accuracy for training data of small or high dimension data. Moreover, the model of chromosome representation in the proposed approach was reduces a combinations number of feature subsets with fitting of the chromosome length. Besides, the model is further to decrease the complexity searching on features space.

We are currently studying more principled design of fitness using domain knowledge as well as

mathematically well-founded tools of multi-attribute utility theory. Future work will experiment by designing of new domain related genetic crossover operator.

ACKNOWLEDGEMENTS

This work was supported by National Science Fellowship research program sponsored by Malaysian Ministry of Science, Technology and Environments (MOSTE).

REFERENCES

[1] A. K. Jain and B. Chandrasekaran, Dimensionality and Sample Size Considerations in Pattern Recognition Practice, Amsterdam: Handbook of Statistics, vol. 2, pp. 835 - 855, 1987.
[2] D. Eads, D. Hill, S. Davis, S. Perkins, J. Ma, R. Porter and J. Theiler, "Genetic algorithms and support vector machines for time series classification", 5th Conference on the Application and Science of Neural Networks, Fuzzy Systems and Evolutionary Computation, pp. 74-85, 2002.
[3] F. J. Ferri, V. Kadirkamanathan and J. Kittler, "Feature subset search using genetic algorithms", Proceedings of the IEEE Workshop on Natural Algorithms in Signal Processing, vol. 740, 1993.
[4] G. Rätsch, T. Onoda and K. R. Müller, "Soft margins for AdaBoost", Machine Learning, vol. 42, no. 3, pp. 287-320, 2001.
[5] J. Liu, H. Iba and M. Ishizuka, "Selecting informative genes with parallel genetic algorithms in tissue classification", Genome Informatics, vol. 12, pp. 14-23, 2001.
[6] J. Sepulveda-Sanchis, G. Camps-Valls, E. Soria-Olivas, S. Salcedo-Sanz, C. Bousono-Calzon, G. Sanz-Romero and J. Marrugat, "Support vector machines and genetic algorithms for detecting unstable angina", Computers in Cardiology, IEEE Computer Society Press, Menphis, USA, 2002.
[7] J. Yang and V. Hanovar, "Feature subset selection using genetic algorithm", Journal of IEEE Intelligent Systems, vol. 13, pp. 44-49, 1998.
[8] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. J. Smola and K. R. Müller, "Invariant feature extraction and classification in kernel spaces", Advances in Neural Information Processing Systems, Massachusetts, USA: MIT Press, vol. 12, pp. 526-532, 2000.
[9] S. N. Arjunan, Protein Secondary Structure Prediction from Acid Amino Sequences using a Neural Network Classifier based on Dempster-Shafer Theory, Master Thesis: Universiti Teknologi Malaysia, 2003.
[10] T. R. Golub, D. K. Slonim, P. Tomayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A Caligiuri, C. D. Bloofield and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, vol. 286, pp. 531-537, 1999.
[11] V. D. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data", Bioinformatics, vol. 8, no. 1, pp. 39-50, 2002.
[12] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, New York: Springer-Verlag, Third edition, 1996.
[13] Z. Michalewicz and M. Schoenauer, "Evolutionary algorithms for constrained parameter optimization problems", Journal of Evolutionary Computation, vol.4, no. 1, pp. 1-32, 1996.