Reducing SAGE Data Using Genetic Algorithms

Cheng-Hong Yang, Tsung-Mu Shih, and Li-Yeh Chuang

Abstract**—**Serial Analysis of Gene Expression is a powerful quantification technique for generating cell or tissue gene expression data. The profile of the gene expression of cell or tissue in several different states is difficult for biologists to analyze because of the large number of genes typically involved. However, feature selection in machine learning can successfully reduce this problem. The method allows reducing the features (genes) in specific SAGE data, and determines only relevant genes. In this study, we used a genetic algorithm to implement feature selection, and evaluate the classification accuracy of the selected features with the K-nearest neighbor method. In order to validate the proposed method, we used two SAGE data sets for testing. The results of this study conclusively prove that the number of features of the original SAGE data set can be significantly reduced and higher classification accuracy can be achieved.

Keywords**—**Serial Analysis of Gene Expression, Feature selection, Genetic Algorithm, K-nearest neighbor method.

I. Introduction

SERIAL analysis of gene expression (SAGE) is a technique proposed by Velculescu *et al.*, which allows global profiling of gene expression in a genome without a priori knowledge [1]. The SAGE technique enables biologists to identify a series of short sequences, as well as the count of each sequence (SAGE tag), out of the entire sequence of a specific cell or tissue type. Each short sequence is collected in a SAGE library, with the count of each short sequence representing the expression of the different genes. Hence, biologists can glean the differences of the gene expression of various cell or tissue types simply by browsing the library and comparing the data for each cell or tissue type. In recent years, the SAGE technique has been applied in cancer research [2], c-MYC identification [3], and the profiling of transcriptome differences [4], amongst others.

In general, SAGE generates large-scale gene expression data from specific cells or tissue. For example, the SAGE library GSM14731 [6] for medulloblastoma cerebellum, in GEO [5] contains 22,017 SAGE tags (genes). Obviously, researchers are faced with a huge challenge when trying to determine differences of gene expression between normal and abnormal samples, since the number of genes involved is so high. Recently, many methods have been developed to solve high dimensional (feature selection) problems in machine learning or data mining. Gamberoni and Storari classified SAGE data using C4.5 and Support vector machines (SVM) in a supervised learning technique, and clustered SAGE data using hierarchical clustering in an unsupervised learning technique [8]. Lin and Li used feature ranking combined with adaptive boosting (AdaBoost) and a SVM classification algorithm for analyzing SAGE data [9]. Alves *et al.* used the GRanular AdDel (GRAD) algorithm for feature selection with different classifiers, e.g. C4.5, SVM, the radial basis function (RBF) and neural network (NN) to predict SAGE data [10].

All the above mentioned feature selection methods obtain a reduced set of genes from the large-scale SAGE data and evaluate the classification accuracy of the selected gene subsets through classification algorithms. These feature selection methods facilitate the analysis of the SAGE data for biologist, since they avoid manual browsing and comparison of the original SAGE data. However, these methods still have not been maturated to a point where the reduced results produce classification accuracies of a high enough standard, so that future experiments could be negatively impacted by carrying over errors and deviations. Recent studies have indicated that superb feature selection performance can be achieved with evolutionary algorithms. We therefore used an evolutionary algorithm for feature selection in order to obtain to lower the number of genes involved and select only those gene subsets of the large-scale SAGE data that increase classification accuracy.

A genetic algorithm (GA) was used to implement the feature selection process, and the K-nearest neighbor (KNN) method was used to evaluate the classification accuracy of the selected gene data set. In the selection of SAGE data sets we focused on two visible data sets (74x822 and 90x27679) for testing. The results show that the proposed method achieved higher classification accuracy with fewer genes selected for the two SAGE data sets.

II. Experimental Methods

A. Feature Selection

The purpose of feature selection is to identify important features (representative features) in original data [14]. SAGE data sets are collected in SAGE libraries. These SAGE libraries can be divided into two classes and contain numerous SAGE tags (genes). As shown in Fig. 1, all of the SAGE libraries can be displayed as a gene expression matrix S. In the matrix S each row represents different SAGE libraries (samples) and each column represents different SAGE tags (features). The element exprij represents the gene expression of SAGE tag j in SAGE library i. The genetic algorithm in this paper is used for feature

Cheng-Hong Yang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan (e-mail: chyang@cc.kuas.edu.tw).

Tsung-Mu Shih is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan (e-mail: gmtsungmu@gmail.com).

Li-Yeh Chuang is with the Department of Chemical Engineering, I-Shou University, Kaohsiung, Taiwan (e-mail: chuang@isu.edu.tw).

selection; it allows us to identify fewer relevant genes in a specific SAGE data set (matrix form).

	tag$_1$	tag$_2$	tag$_3$	...	tag$_n$
library$_1$	expr$_{11}$	expr$_{21}$	expr$_{31}$	...	expr$_{n1}$
library$_2$	expr$_{12}$	expr$_{22}$	expr$_{32}$	...	expr$_{n2}$
library$_3$	expr$_{13}$	expr$_{23}$	expr$_{33}$	...	expr$_{n3}$
...	...	...	...	...	...
library$_m$	expr$_{1m}$	expr$_{2m}$	expr$_{3m}$	...	expr$_{nm}$

Fig. 1 Gene expression matrix S

B. Genetic Algorithm

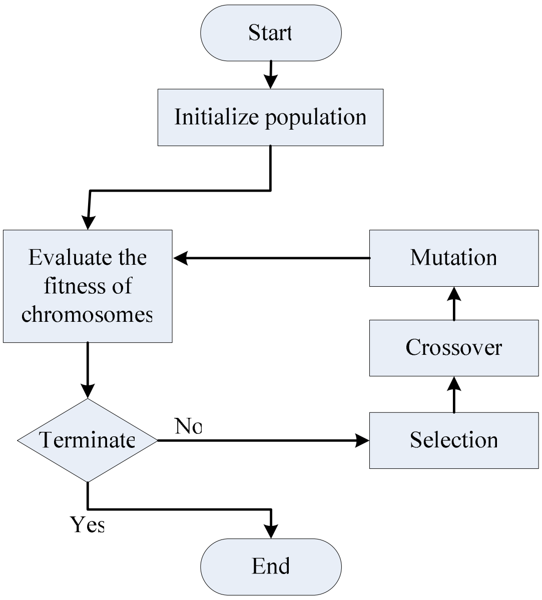

Fig. 2 The flowchart of genetic algorithm

A genetic algorithm (GA) is a kind of evolutionary algorithm [15]; it simulates the process of biological evolution to find near-optimum solutions in a complex problem and is based on the ‘survival–of-the-fittest’ principle by Darwin. Initially, a population of chromosomes is generated randomly. In the initial population, each chromosome represents a solution. Then, these chromosomes will be evolved (selection, crossover and mutation) repetitively until the best chromosomes are produced or a terminal condition is reached. Finally, only the best chromosomes survive. The flowchart of the GA process is shown in Fig. 2. Details of the GA information extraction process of from the specific SAGE data sets are described below.

1. Chromosome design. A chromosome is composed of a SAGE tag subset which was selected from a specific SAGE data set. The scope of the selected SAGE tags in a chromosome was restricted between tag$_1$ and tag$_n$. Different chromosomes contain different feature (SAGE tag) sets in the population.

2. Population initialization. A random selection function was used to generate a population of chromosomes. Each chromosome contained a random set of SAGE tags in the initial population, and the entire population was evolved by the GA in a simulated evolution process.

3. GA evolution. All chromosomes in the population are constantly subjected to the processes of selection, crossover and mutation in order to obtain a chromosome with the best fitness. Roulette wheel selection was applied to randomly choose chromosomes for crossover and mutation. Chromosomes with a higher fitness value have a higher probability of selection. In the crossover process, two-point crossover was used to generate offsprings by exchanging a random part of the SAGE tags from two randomly selected parents. During mutation, offsprings are generated by abandoning parts of the original SAGE tags and adding parts of non-original SAGE tags in a selected chromosome randomly. If, after the evolution process, a generated offspring is superior to its parent, the offspring will replace the parent in the population.

4. Chromosome evaluation. The fitness of each chromosome was evaluated using the K-nearest neighbor method. Chromosomes with high fitness values are regarded as having collected important SAGE tags from the original data set.

5. Termination. The entire process is stopped when an upper-bound of generations is reached. The best chromosomes are then saved.

C. The K-nearest neighbor method

The K-nearest neighbor (KNN) method is a basic classification algorithm often employed in machine learning [16]. K-NN classifiers have attracted the interest of many researchers due to their theoretic simplicity and the comparatively high accuracy that can be achieved with it compared other for more complex methods. The evaluation is based on a specific distance (Euclidean distance was used in this paper). The test sample is of an unknown class, and the distance to other samples (training data) is calculated. The nearest samples are used to determine class membership. Test samples are classified in a certain class when the votes for this class are a simple majority.

Leave-one-out cross-validation (LOOCV) is typically used to evaluate the classification accuracy in classification studies [17]. LOOCV selects a single sample as a test sample and regards all other samples as training data for validation. The process is repeated so that each sample is set as the test sample once. We used LOOCV to evaluate the classification accuracy in this study.

III. Results and Discussion

The SAGE technique generates information of gene expression in various cell or tissue types. Several thousand genes are usually analyzed in a single study of biological samples. Therefore, it is necessary to filter SAGE gene expression data and obtain relevant genes by employing a feature selection process. The complex correlation between genes and diseases can be significantly reduced by using the filtered results. The experimental results of our study and a comparison with correlating articles are show below.

TABLE I
FORMAT OF SOURCE DATA SETS

Data set	Library amount	Gene amount	Normal tissue	Abnormal tissue
74x822	74	822	24	50
90x27679	90	27,679	30	60

A. Parameter setting

The population size and the number of generations was set to 1000, respectively, the crossover rate was set to 0.6 and the mutation rate was set to 0.01 in our study. The 1-nearest neighbor method (K=1, 1NN) was used to evaluate and calculate the classification accuracy for the SAGE data set.

B. SAGE data set

A typical SAGE data set consists of many different SAGE libraries, with all these libraries containing a large number of SAGE tags (genes). Table I shows the two SAGE data sets used in our experiment. The first data set is 74x822; it contains 74 SAGE libraries and 822 SAGE tags. The second data set is 90x27679, which contains 90 SAGE libraries and 27,679 SAGE tags. These two data sets were individually separated into two classes (normal and abnormal) for the brain, colon, ovary, etc., SAGE libraries of *Homo sapiens*. Both of the data sets were obtained from Tzanis *et al.* [7].

C. Results

The experimental results for the 74x822 test sample are depicted in Fig. 3 and Table II After employing the GA, 36 genes were selected from the original 722 genes (a reduction of 95.62%); the classification accuracy reached 89.19% at 100 generations. At 200 generations, 28 genes were selected (reduction of 96.59%) and 97.30% classification accuracy was achieved. The best gene subsets were obtained at 500~1,000 generations, with the number of genes selected at 24 (reduction of 97.08%) and 98.65% classification accuracy. The experimental results for the 90x27679 data set are shown in Fig. 4 and Table III The GA reduced the number of genes to 8,740 from an original number of 27,679 genes (reduction of 68.42%), with the classification accuracy at 91.11% after 100 generations. At 200 generations, 7,763 genes were selected (reduction of 71.95%) and 92.22% classification accuracy was achieved. At 500 generations, 7,921 genes were selected (reduction of 71.38%), and the classification accuracy was 95.56%. Finally, the best gene subsets were obtained at 1,000 generations, with the number of genes reduced to 4,914 (reduction of 82.25%) and classification accuracy at 97.78%.

For both the above test samples, the reduction of genes and the classification accuracy increased with a progressive number of GA generations. The results prove the effectiveness and swiftness of the proposed method for obtaining representative genes from high-dimensional SAGE data.

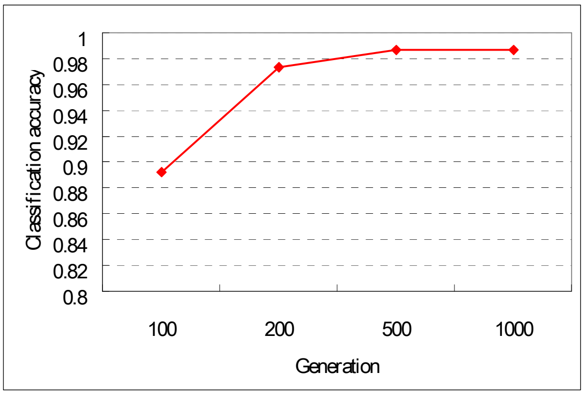

Fig. 3 Classification accuracy for 74x822

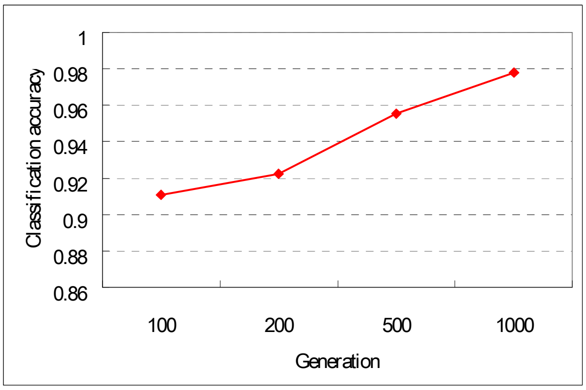

Fig. 4 Classification accuracy for 90x27679

TABLE II
REDUCTION OF GENES FOR 74X822

Generation	Reduction	Number of genes
100	95.62%	36
200	96.59%	28
500	97.08%	24
1,000	97.08%	24

TABLE III
REDUCTION OF GENES FOR 90X27679

Generation	Reduction	Number of genes
100	68.42%	8,740
200	71.95%	7,763
500	71.38%	7,921
1,000	82.25%	4,914

D. Discussion

Alves *et al*. used the GRAD algorithm to select genes in the 74x822 data set, with the different classifiers C4.5, SVM, RBF and NN used to evaluate the selected genes [10]. Their results show that C4.5 obtained the best classification accuracy (86.18%; reduction of 99.15%), far better than the one obtained with SVM, RBF and NN. A comparison of the results (Table IV) shows that our method achieved higher classification accuracy 12.47% higher than the one in the Alves *et al*. study, but that the reduction of genes was 2.07% lower. These numbers prove our method to be effective. Because the variations of gene expression in SAGE data are very small, the decision of where

TABLE IV
COMPARISON OF REDUCTION AND ACCURACY TO OTHER METHODS

Method	Data set			
	74x822		90x27679	
	Accuracy	Reduction	Accuracy	Reduction
Alves *et al*. [10]	86.18%	99.15%	n.a.	n.a.
Gamberoni and Storari [8]	n.a.	n.a.	82.20%	n.a.
Lin and Li [9]	n.a.	n.a.	85.60%	96%
Our method	98.65%	97.08%	97.78%	82.24%

to set the cut point for C4.5 may cause errors in calculating classification accuracy [13]. We therefore suggest using KNN rather than C4.5 when classifying selected genes for feature selection in SAGE data.

Gamberoni's and Lin's team both conducted studies of the 90x27679 data set [8, 9]. Gamberoni and Storari used a SVM to classify the data set and obtained classification accuracy of 82.2% [8]. Our classification accuracy is again 15.58% higher than the one produced in their study (Table IV). We believe that they did not use feature selection to reduce the number of irrelevant genes, and thus classification accuracy suffers as a result. Lin and Li used the F-score and linear weights for feature selection, and SVMs with various kernels for evaluating classification accuracy. The experimental results they produced, based on the same classifier, were superior to those published in Gamberoni's study (Table IV), a fact that proves the effectiveness of feature selection in obtaining representative genes from an original SAGE data set. A comparison of our results to the ones published in Lin's study [9] again shows that our method achieved classification accuracy 12.18% higher than the one in Lin's study. Many studies indicate that SVM is superior to KNN in term of classification [11, 12]. A GA is considered a SVM, and thus it can be expected that it outperforms the F-score and linear weights as classifiers for feature selection. A reduction of SAGE data with GAs or SVMs may yield better results (higher classification accuracy and more representative genes); however the cost in terms of computational time also increases substantially. When the number of samples is huge, the time complexity problem needs to be considered as well.

In this study, we used a GA to filter out irrelevant genes, and evaluated the results with 1-NN. The proposed method is efficient in identifying fewer and more representative genes in high-dimensional SAGE data. The obtained results indicate that the proposed method finds representative genes efficiently, and that substantially higher classification accuracy compared to other methods can be achieved.

IV. CONCLUSION

In this study, we successfully reduced the number of genes in SAGE data sets by implementing feature selection with a GA. KNN with LOOCV was used as a classifier to evaluate the reduction of genes. Two SAGE data sets, 74x822 and 90x27679 were used for testing. For 74x822 the number of genes was reduced to 97.08% of the original number of genes, and 98.65% classification accuracy was achieved. For the 90x27679 sample the reduction was 82.24% and classification accuracy 97.78%. The experimental results prove the method to be effective in discarding irrelevant genes and improving classification accuracy. Applied to SAGE libraries, the method can improve the efficiency of future studies.

ACKNOWLEDGMENT

This work was partly supported by the National Science Council in Taiwan under grants NSC96-2221-E-214-050-MY3 and NSC97-2622-E-151-008-CC2.

REFERENCES

[1] V.E. Velculescu, L. Zhang, B. Vogelstein and K.W. Kinzler, "Serial analysis of gene expression", *Science*, vol. 270, no. 5235, pp. 484-487, October 1995.

[2] L. Zhang, W. Zhou, V.E. Velculescu, S.E. Kern, R.H. Hruban, S.R. Hamilton, B. Vogelstein and K.W. Kinzler, "Gene Expression Profiles in Normal and Cancer Cells", *Science*, vol. 276, no. 5316, pp. 1268-1272, May 1997.

[3] T.C. He, A.B. Sparks, C. Rago, H. Hermeking, L. Zawel, L. T. da Costa, P.J. Morin, B. Vogelstein and K.W. Kinzler, "Identification of Myc as a target of the APC pathway", *Science*, vol. 281, no. 5382, pp. 1509-1512, September 1998.

[4] V.E. Velculescu, S.L. Madden, L. Zhang, A.E. Lash, J. Yu, C. Rago, A. Lal, C.J. Wang, G.A. Beaudry, K.M. Ciriello, B.P. Cook, M.R. Dufault, A.T. Ferguson, Y. Gao, T.C. He, H. Hermeking, S.K. Hiraldo, P.M. Hwang, M.A. Lopez, H.F. Luderer, B. Mathews, J.M. Petroziello, K. Polyak, L. Zawel, W. Zhang, X. Zhang, W. Zhou, F.G. Haluska, J. Jen, S. Sukumar, G.M. Landes, G.J. Riggins, B. Vogelstein and K.W. Kinzler, "Analysis of human transcriptomes", *Nature Genetics*, vol. 23, no. 4, pp. 387-388, December 1999.

[5] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky and R. Edgar, "NCBI GEO: mining tens of millions of expression profiles--database and tools update", *Nucleic acids research*, vol. 35, pp. 760-765, January 2007.

[6] GEO (Gene Expression Omnibus), "GSM14731", http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM14731.

[7] G. Tzanis and I. Vlahavas, "Accurate Classification of SAGE Data Based on Frequent Patterns of Gene Expression", *19th IEEE International Conference on Tools with Artificial Intelligence*, vol. 1, pp. 96-100, October 2007.

[8] G. Gamberoni and S. Storari, "Supervised and unsupervised learning techniques for profiling SAGE results", *In Proceedings of the ECML/PKDD Discovery Challenge Workshop*, pp. 121-126, September 2004.

[9] H.T. Lin and L. Li, "Analysis of SAGE Results with Combined Learning Techniques", *In Proceedings of the ECML/PKDD Discovery Challenge Workshop*, pp. 102-113, October 2005.

[10] A. Alves, N. Zagoruiko, O. Okun, O. Kutnenko, and I. Borisova, "Predictive Analysis of Gene Expression Data from Human SAGE Libraries", *In Proceedings of the ECML/PKDD Discovery Challenge Workshop*, pp. 60-71, October 2005.

[11] Y.F. Shi and Y.P. Zhao, "Comparison of Text Categorization Algorithms", *Wuhan University Journal of Natural Sciences*, vol. 9, no. 5, pp. 798-804, October 2004.

[12] L.Y. Chuang, C.H. Ke and C.H. Yang, "A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification", *International MultiConference of Engineers and Computer Scientists 2008*, vol. 1, pp. 146-150, March 2008.

[13] J.R. Quinlan, *C4.5: programs for machine learning*, Morgan Kaufmann, San Francisco, CA, USA, 1993.

[14] Wikipedia, "Feature Selection", http://en.wikipedia.org/wiki/Feature_selection.

[15] E. Elbeltagi, T. Hegazy and D. Grierson, "Comparison among five evolutionary-based optimization algorithms", *Advanced Engineering Informatics*, vol. 19, Issue 1, pp. 43-53, January 2005.

[16] Wikipedia, "k-nearest neighbor algorithm", http://en.wikipedia.org/wiki/K-nearest_neighbor.

[17] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", Bioinformatics, vol. 21, no. 5, pp. 631-643, March 2005.