

Gene Expression Classification using Multi-Objective Ensembles

Ed Keedwell¹ and Ajit Narayanan²

Abstract. The classification of microarray data is an important application of computational intelligence techniques. It is important for the techniques chosen to construct models that are accurate, parsimonious and explicable. This paper investigates the use of multi-objective genetic algorithms to create a number of rules for classifying gene expression data and the combination of these rules into an ensemble classifier. The results demonstrate that the method of creating the ensemble is an intrinsic component of the classification process and that the use of ensembles can increase the accuracy of the classification.

1 INTRODUCTION

The advent of microarray technology has allowed biologists an unprecedented view of the genetic and biomolecular mechanisms underlying the behaviour of organisms. A typical microarray is capable of capturing the expression level of up to 30,000 genes and their transcripts at any particular time. The data generated by microarrays, however, has proved somewhat of a challenge for traditional analytical techniques, due largely to the size of the arrays themselves and the inherent combinatorial problems that arise from analysing a dataset with tens of thousands of variables. This paper describes a novel computational technique using multi-objective genetic algorithms for discovering genes that classify samples in large gene expression datasets.

Gene Expression Data Analysis

Generally speaking, microarray experiments fall into three categories: temporal, static and classificatory. Temporal microarray data is created by measuring gene expression values over a number of timesteps, often whilst subjecting the organism to some stimulus, so that the behaviour of genes can be measured to determine possible cause-effect relationships. The goal of analysing such data is to determine the pattern of excitations and inhibitions between genes, gene products and proteins that make up a gene network for that organism

through observing the interactions between genes between one timestep and the next. Genetic algorithms [1] and multi-objective genetic algorithms [2] have previously been successfully used for this purpose.

Static microarray data measures just once the gene expression values of a population of individuals. The aim is then to apply various techniques to separate the individuals into mutually exclusive subsets based on dividing the data into gene groupings that share patterns of co-expression. Typically, cluster analysis (hierarchical, k-means) sort samples into groups so that the degree of association between two samples is maximal if they belong to the same group and minimal otherwise. Cluster analysis is widely used to discover structures in data that may then require subsequent analysis and explanation.

Classification microarray data is created by taking an individual sample of a number of individual organisms or tissue samples which differ in some known respect from each other. Classification studies are most often seen in cancer research, where individuals are pre-sampled and separated into classes (typically case and control) according to an independent diagnosis, with membership of each class determined by the pathology of the individuals involved. The task for the analytical technique here is typically to differentiate between those individuals diagnosed with cancer and those without and also to identify a reduced set of genes for doing so, using the gene expression values of each of those individuals and the class to which they belong. Such 'gene reduction' methods are a form of feature selection due to their use of 'original' measurement values. In contrast, typical feature selection methods, such as support vector machines and recursive feature selection methods generally require an explicit search procedure, an evaluation criterion and a stopping criterion for terminating the search procedure [3]. One of the challenges for such techniques is how to identify optimal combinations (subsets) of features that together lead to successful feature selection, given the strategy of examining each feature in turn.

A common problem encountered with these studies is that the measurements gained from these studies often span thousands of genes (fields, attributes) and only a records (time-steps, samples), which makes it relatively unusual in structure and not as amenable to traditional analysis as static biological databases. The difficulty for all methods is 'under-determinism', i.e., there are very many variables in comparison to the number of records available. Another common problem is that, often, combinations of gene are required to correctly determine the class of the sample or the activation of the regulated gene. That is, one gene is unlikely by itself to classify all samples. In such circumstances, the search space grows from thousands of possibilities to a much

¹ School of Engineering, Computing and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF
Email: E.C.Keedwell@ex.ac.uk.

² School of Computing and Mathematical Sciences, Auckland University of Technology, Auckland 1142, New Zealand. Email: Ajit.Narayanan@aut.ac.nz.

larger potential search space of billions of combinations. It is this ‘curse of dimensionality’ that causes problems for traditional statistical and algorithmic approaches.

Multi-Objective Evolutionary Computing in Gene Expression Analysis

Recently, ‘intelligent’ approaches based on machine learning techniques in artificial intelligence have been applied to the problems of microarray analysis (e.g. [4],[5]). The task of such techniques is to adopt traditional machine learning approaches, such as artificial neural networks and genetic algorithms (e.g. [6], [7]), to produce results that are primarily interpretable by experts in the domain rather than significant according to some statistical method. Multi-objective genetic algorithms are well-known for providing a number of solutions to select from and provide a more transparent result for interpretation by the domain expert therefore they represent a good choice of algorithm for use in this domain. Multi-objective genetic algorithms utilise similar genetic operators to their single-objective counterparts, namely a stochastic population generation process, random mutation and recombination, but they differ in the way that solutions are evaluated. Each solution in a MOEA will have multiple fitnesses and the superiority or otherwise of one solution over another is established via the dominance criterion. The algorithm produces a set of solutions which represent the optimal discovered trade-off between two or more objectives and is known as a pareto-front. For this to work effectively, objectives normally need to be antagonistic (e.g. a decrease in one objective incurs an increase in the other) and this is what is represented in the approaches below.

Multi-objective genetic algorithms have been used for deriving gene regulatory networks [2] whereby the connectivity of the network is one objective, and the RMSE of the derived network is the other. The resulting optimisations therefore create a pareto-front that consists of multiple networks that represent a trade-off between parsimony and accuracy. This trade-off is an important idea in science and embodies the principle of Occams’ razor, which in this case is the notion that given two regulatory networks with equal accuracy, the network with fewer connections should be selected. The work here uses a similar principle as applied to gene expression classification.

So far, only two systems have been proposed that use MOEAs for the classification of gene expression data. A system known as Memetic NSGA [8] was tested on a small 50 gene leukaemia dataset and was used to demonstrate the efficacy of the algorithm rather than to generate biological information. This is in contrast to the goal of this work which is designed to operate on gene expression datasets of any size. The most widely-cited system is that of Deb and Reddy [9] who proposed a gene expression classification system based on NSGA-II (the same algorithm that is used here) that classifies full gene expression datasets. This algorithm was able to discover small numbers of highly accurate genes. The objectives optimised by the MOEA were f1, the number of genes in the rule, f2, the number of errors in the training data and f3, the number of errors in the test data. These objectives gave rise to a number of rules which were able to classify

three biological datasets, namely the leukaemia (ALL-AML) (considered here), colon and lymphoma cancer datasets. The system was shown to be capable of discovering multiple rulesets with 100% accuracy on these datasets and often had less than five genes in the rule. However, the use of the test set in the optimisation procedure, firstly combined into one objective and latterly separated into two objectives removes the independence of that test set. The approach detailed here is significantly different, firstly from an algorithm perspective as this work uses a neural-genetic approach, secondly because the rules are then combined into ensembles for testing and finally because only the training dataset is considered in the computation of the fitness function, ensuring that the test set remains independent from the process of optimisation, a key factor in classification experiments.

Ensemble Classification

Ensemble classification is an established method of combining several independently developed classifiers into one classification mechanism. There is no established method for combining the classifiers into one classification approach but we investigate three popular choices here, using a winner-takes-all approach, a majority-vote approach and an averaged approach. A prerequisite for ensemble classification is that multiple classifiers must be created to be subsequently combined together.

A number of methods have been proposed for the creation of multiple classifiers, the most established of which are bagging and boosting. Bagging [10] involves randomly sampling from the dataset and then creating multiple classifiers from the sampled datasets. Boosting [11] is a sequential technique that creates multiple classifiers based on the errors of those classifiers that precede the current one. More recently, research has been conducted into the process of creating multiple classifiers with the view to them performing optimally as an ensemble. Significant work on the creation of neural network ensembles has been conducted [12] whereby diversity in the classifier is deliberately encouraged to increase performance when the networks are then combined into an ensemble. The same group has also considered the power of using a multi-objective algorithm to develop neural network ensembles [13]. The objectives used are the accuracy and diversity of the networks. It is this concept that is adopted in this paper, where a number of different objectives are considered.

Furthermore, a good number of approaches to ensemble creation have been applied to microarray classification. A new technique and a review of the current approaches can be seen in [19]. The approaches shown in the papers seen in [19] involve the ensemble of feature selection and classification techniques, often arranged in pairs. This differs significantly from the approach proposed here which simply uses the multi-objective algorithm to generate the rules required for ensemble creation. The use of feature selection methods has been criticised in the past as by necessity, they restrict the space available to the classifier and therefore may remove genes which look unpromising in isolation, but are influential when combined with other genes. Additionally, the

complexity of these techniques both in terms of implementation and computational complexity are likely to exceed that of a multi-objective GA run given that they comprise two separate stages and search the space of combined of classifier and feature-selection algorithms. In the case of [19] there are 6 classifiers and 7 feature-selection algorithms which must be searched combinatorially using a GA and its attendant computational cost before the feature-selection and classifier algorithms themselves are run.

2 METHOD

In 2003, a neural-network and genetic algorithm hybrid for the classification of microarray data was proposed [14]. In this paper, this technique is expanded to utilise a multi-objective algorithm from which a set of rules are created and then combined together to form an ensemble. The neural-genetic technique uses the evolutionary algorithm to select up to k genes from the set of possible genes in the data. The training data from these genes is then passed to the neural network along with the classification information. The neural network then adjusts its weights to reduce the error of the classification as much as possible. In this paper, a simple single layer is used to determine the weights, but more complex multi-layer arrangements are possible. The combined gene names and weights can then be printed in the form of rules which can then be easily interpreted by non-technical personnel. For a more in-depth description of the algorithm, readers are directed to [14].

The algorithm used here is significantly different in that it uses a multi-objective genetic algorithm. For this purpose the genetic algorithm has been replaced with NSGA-II [15], one of the most popular and successful elitist multi-objective genetic algorithms in current use. NSGA-II is popular because there are few additional parameters over the standard genetic algorithm and so it can replace the single objective GA without too much further parameter tuning. However, more objectives need to be identified for the problem and these are determined as follows:

2-Objective Example

1. Minimise error – over all samples in the training data
Where m = number of misclassifications and n is the size of the training set
2. Minimise number of genes in each rule

$$E = \frac{m}{n} \cdot 100$$

3-Objective Example

1. Minimise error
2. Minimise number of genes
3. Maximise diversity – *diversity is determined as the number of genes present in the current chromosome that are not present in the other individuals in the population*

4-Objective Example

1. Maximise sensitivity – $TP/(TP+FN)$
2. Maximise specificity – $FP/(FP+TN)$
Where TP = Number of true positives, FN = Number of false negatives, FP = Number of false positives, TN = Number of true negatives.
3. Maximise diversity
4. Minimise number of genes

The two objective example simply tries to maximise the conflicting requirements of accuracy and parsimony. The three objective example adds to this the dimension of diversity which is a key component of the work in [12] and [13] and allows a comparison to be drawn between the simple approach and one which incorporates a diversity component. There is evidence from [12] and [13] that the introduction of a diversity component improves the performance of the subsequent ensemble. The four objective example augments the three-objective with two established and conflicting accuracy measurements in sensitivity and specificity. This essentially means that the algorithm is maximising the area under the ROC (Receiving Operator Characteristics) curve and the diversity whilst simultaneously minimising the number of genes in each of the rules.

In each example, the algorithm is run for a set number of generations and the resulting pareto set of rules is used to construct the ensemble classifier. An additional further factor in the performance of an ensemble is how the classifiers are combined. Here, three methods are investigated:

Majority Vote – Each rule is given a single vote and the class with the most votes is then returned as the final decision. If there are equal votes for a class then the sum of fitnesses (accuracies) for that class is used to determine the classification.

Average Vote – Each rule is fired and the results averaged. If the average exceeds 0.5 then the classification corresponding to 1 is returned, otherwise the 0 classification is returned.

Weighted Vote – Similar to the average vote, but where the average is weighted according to the accuracy of each rule.

In the following experiments, the genetic algorithm is run 20 times for each set of objectives. For each run, the three methods of combining the rules to make an ensemble are tested against each other for accuracy on the training dataset and a completely separate testing dataset in terms of accuracy.

3 DATA

The two datasets used here are well studied in the gene expression literature. The first of these is known as the ALL-AML dataset and is designed to differentiate between two types of leukaemia, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) and was collected and studied in [16]. The data consists of 7129 genes, and 72 cases, of which 38 are of type B cell ALL, and 34 of type AML. This dataset is designed to discern between these cancers as they present very similar symptoms, but respond

very differently to treatments. This dataset is randomly separated (although preserving the class balance) into a training set of 52 and test set of 20 samples. The second dataset is taken from [17] and consists of 70 gene expression values for each of 104 individuals, of whom 73 suffer from myeloma and the remaining 31 are diagnosed as normal. The task is for the classification algorithm to correctly separate the individuals into the appropriate groups using minimal gene sets. A random test set of 40 individuals was used and the remaining 64 used for training.

4 RESULTS

The following experiments are designed to determine whether an ensemble approach can yield better performance on gene expression data. Furthermore, the goal is to determine the feasibility of an ensemble-based neural-genetic method for the classification of data in general. The experimentation is also expected to yield information as to which set of objectives yields rules that perform best as an ensemble, along with which combination method yields the best results.

ALL-AML Dataset

Table 1 – Comparison of the performance of each objective set on the ALL-AML dataset

	Mean Rule Number	Mean Train Error	Mean Test Error
2-Object.	5.38	8.14	4.98
3-Object.	6	8.41	5.08
4-Object.	11.57	8.98	4.84

Table 1 shows the performance of each set of objectives on the ALL-AML problem. As expected, the number of rules in the ensemble increases with an increase in the number of objectives. However this phenomenon is much more obvious with the move from 3 to 4 objectives than from 2 to 3, indicating that the diversity measure does not dramatically

Mean Performance on Training and Test Data for Each Ensemble Method

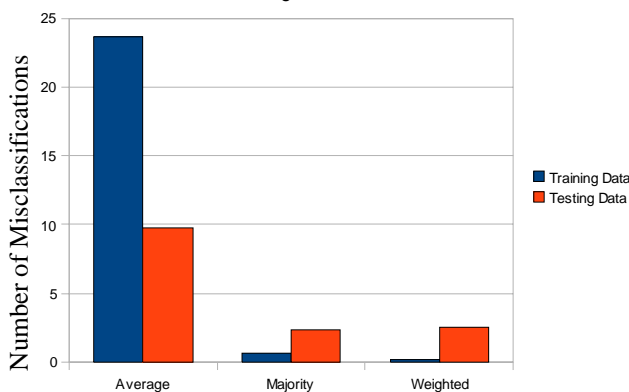


Figure 1 – Comparison of performance for each ensemble technique on the ALL-AML data

increase the number of rules in the ruleset, although it should increase the diversity of the rules. The training and test errors appear to be rather poor, but this is due to the performance of one of the ensemble creation techniques, as we can see in Figure 1.

As can be seen in Figure 1, the average method of combining the pareto-set of rules yields very poor performance. This indicates that the method of averaging the response of each of the rules is by far the least effective here and that the number of rules that fire and their accuracy are important factors in correct classification. Table 2 below shows the errors with the average method removed.

Table 2 – Revised Table with the average ensemble method removed

	Mean Rule Number	Mean Train Error	Mean Test Error
2-Object.	5.38	0.38	2.26
3-Object.	6	0.43	2.61
4-Object.	11.57	0.39	2.45

Table 2 shows the performance of just the weighted and majority vote ensembles on the same data and this indicates that performance particularly on the training data is much improved with less than half an error per ensemble on average. The results shown here therefore result in average testing classification accuracies of 87% (3 objective) to 89% (2 objective). However, individual rulesets achieved 100% accuracy on testing data (see Further Analysis).

Myeloma Dataset

The average ensemble method performed similarly poorly on the myeloma dataset and therefore the following results have had this method removed.

Table 3 - Comparison of the performance of each objective set on the Myeloma dataset with the Average ensemble method removed

	Mean Rule Number	Mean Train Error	Mean Test Error
2-Object.	6	0.86	3.59
3-Object.	6	0.95	3.29
4-Object.	7.6	0.95	3.68

Table 3 shows the relative performance of each set of objectives on the myeloma problem. An interesting property of this particular set of runs is that the 3-objective run did not yield larger rulesets than the 2-objective run in this case. This suggests that again, the diversity of the rules does not influence the generated pareto-front particularly. In fact, the number of rules in the ensemble does not change in the same way for any of the objectives as it did for the ALL-AML

dataset. This is likely to reflect the smaller number of genes in this dataset, which has been preprocessed and therefore only includes a small number of genes that are related to the classification. With relatively few genes to choose from, it is unlikely that rules with many genes will be present in the data and therefore this provides a natural limit to the size of the ruleset. The results shown here therefore result in average testing classification accuracies of 91% (2 objective) to 92% (3 objective). However, individual rulesets achieved 95% accuracy on testing data (see Further Analysis).

Further Analysis

The performance of the ensemble tends to track that of the best performing rule in most instances. There are some occasions where the ensemble improves on the performance of the top performing rule which appear more often in the myeloma dataset than the ALL-AML. In general the ensemble performance on the myeloma dataset is at least as good as the best performing rule and often better than it.

The diversity preservation mechanism did not appear to have the desired effect of creating a number of different individuals with similar classification performance. The aim was to minimise the number of similar genes with other individuals in the whole population but perhaps this measure should be limited to those individuals in the pareto-set to further reduce the number of overlapping genes.

The best rule discovered by the approach classifies both the training and testing data with zero errors on the ALL-AML dataset. This rule states that AML is characterised by the absence of L05148, M89957 and X69111. L05148 codes for protein-tyrosine kinase and is involved with T-cell regulation [18]. M89957 codes for the cell surface glycoprotein, part of the B-cell receptor complex and therefore has involvement in the immune system. Finally, X69111 codes for the helix-loop-helix protein ID3 which is expressed in blood lymphocytes.

The best rule discovered on the myeloma dataset classifies the training data with zero errors and makes two errors on the testing set. This rule states that myeloma is characterised by the absence of three genes, M63928 a T-cell activation antigen and tumour necrosis factor receptor, X16832 codes for human cathepsin H. A number of cathepsins (G,L and K) have previously been associated with myeloma, but H does not appear to have such a strong biological association. L36033 codes for a pre-B cell stimulating factor which would again appear to have immune system influences.

5 CONCLUSIONS

The neural-genetic method of discovering classification models from gene expression data has been extended to include a multi-objective element. The use of a multi-objective GA allows the algorithm more flexibility in the rules that it creates in that it can balance the requirements for parsimony and accuracy. The pareto set of solutions can then be interrogated by experts in the field and an extra level of

confidence can be attributed to a set of similar rules with differing numbers of genes.

The creation of a set of results as opposed to a single rule begs the question of how best to combine the multiple results into a single classifier. This has been investigated here through the creation of an ensemble from the rules generated by the multi-objective GA. Three different methods of combining the rules were investigated and the conclusion can be drawn that the weighted and majority vote approaches performed very similarly in the trials on the data here and both are preferred to the averaging approach which yielded very poor results. An interesting point to note here is that the performance of the ensemble was limited to the performance of the best performing rule for the ALL-AML dataset, but this was not the case for the myeloma dataset where the ensemble performance was often better than the best performing rule. The tentative conclusion can be drawn then that the use of an ensemble is recommended to ensure robust classification over a number of datasets providing that the weighted or majority vote ensembles are used. If this is the case then the evidence suggests that the ensemble will deliver results at least as good as those of the best performing rule and occasionally better depending on the dataset.

Three different sets of objectives were considered to create the ensembles, and the genuinely surprising result is that the number and type of objective had little effect on the resulting accuracy of the ensemble. It was expected that the introduction of a diversity measure would increase the generality of the ensemble and reduce the potential for it to overfit, but there appeared to be little advantage in these areas for the three and four objective examples where the diversity measure was present. The increase in objectives did yield an increase in the number of rules in most cases, but it is not at all evident that the increased number of rules results in more robust classification or any increased protection against overfitting. This is a surprising result and perhaps suggests that certain of the rules enjoy a dominance over the others in the ruleset. This is certainly possible in the weighted voting strategy as rules with good fitnesses will prevail, but the majority vote strategy provides an unbiased view providing that tie-breaks are not routinely required.

Overall, the ensemble system has not yielded dramatic improvements in accuracy over the original multi-objective system. However, the rules generated by the system have been shown to be both accurate (zero errors for ALL-AML and two errors for myeloma datasets) and to have some biological plausibility according to the current known function of these genes.

REFERENCES

- [1] Keedwell, E., Narayanan, A., (2005) "Discovering Gene Regulatory Networks with a Neural-Genetic Hybrid" [IEEE/ACM Transactions on Computational Biology and Bioinformatics](#), July-September 2005, Vol 2., No.3, pp 231-243, IEEE Computer Society
- [2] Spieth C, Streichert F., Speer N., and Zell, A., (2005) "Multi-Objective Model Optimization for Inferring Gene Regulatory

- Networks” in the proceedings of Conference on Evolutionary Multi-Criterion Optimization (EMO) 2005 Guanajuato, Mexico, 2005. LNCS 3410, pp. 607-620, Springer-Verlag
- [3] Dash, M and Liu, H (1997) Feature Selection for Classification Intelligent Data Analysis Vol 1., pp131-156
- [4] Keedwell, E.C. and Narayanan, A. (2005). *Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems*. Wiley.
- [5] Fogel, G.B., Corne, D.W. and Pan, Y. (2007) *Computational Intelligence in Bioinformatics*. Wiley-IEEE.
- [6] Ando, S., Iba H.. (2001a) "Inference of Gene Regulatory Model by Genetic Algorithms", Proceedings of Conference on Evolutionary Computation 2001 pp712-719
- [7] Ando, S., Iba H., (2001b) "The Matrix Modeling of Gene Regulatory Networks -Reverse Engineering by Genetic Algorithms-", Proceedings of Atlantic Symposium on Computational Biology, and Genome Information Systems & Technology 2001.
- [8] Praveen K., Sharath S. Rio D'Souza G, K. Chandra Sekaran (2007) "Memetic NSGA A Multi-Objective Genetic Algorithm for Classification of Microarray Data" in the Proceedings of the 15th International Conference on Advanced Computing and Communications, 2007, pp75-80
- [9] Deb, K. and Reddy, A.R., (2003) "Classification of two-class cancer data reliably using evolutionary algorithms", BioSystems **72** (2003), p. 111.
- [10] Breiman, L.: Bagging predictors. Machine Learning, 24, 123-140 (1996)
- [11] Freund, Y. (1995): Boosting a weak learning algorithm by majority. Information and Computation 121, 256-285 (1995).
- [12] Brown, G. Xin Yao Wyatt, J. Wersing, H. Sendhoff, B. (2002) Exploiting ensemble diversity for automatic feature extraction Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02. Vol. 4, pp 1786- 1790 vol.4
- [13] Chandra, A and Xin Yao, "Multi-objective Ensemble Construction, Learning and Evolution," Proc. of the PPSN Workshop on Multi-objective Problem Solving from Nature (part of the 9th International Conference on Parallel Problem Solving from Nature: PPSN-IX), 9 - 13 September 2006, Reykjavik, Iceland.
- [14] Keedwell, E.C. and Narayanan, A. (2003) "Genetic algorithms for gene expression analysis" in Applications of Evolutionary Computing LNCS 2611 Gunther Raidl et al (Eds.), proceedings of EvoBIO2003 1st European Workshop on Evolutionary Bioinformatics pp 76-86
- [15] Deb, K, Amrit Pratap, Sameer Agarwal and T. Meyarivan (2000). A Fast and Elitist Multi-Objective Genetic Algorithm-NSGA-II. KanGAL Report Number 2000001
- [16] Golub T.R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M. L. Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S.. (1999) "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" Science Vol 286 pp 531-536
- [17] Page, D., Zhan, F., Cussens, J., Waddell, W., Hardin, J., Barlogie, B., Shaughnessy, J. "Comparative Data Mining for Microarrays: A Case Study Based on Multiple Myeloma." Poster presentation at International Conference on Intelligent Systems for Molecular Biology August 3-7, Edmonton, Canada. Technical report available from mwaddell@biostat.wisc.edu
- [18] Chan AC, Iwashima M, Turck CW, Weiss ^a ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain. Cell. 1992 Nov 13;71(4):649-62
- [19] Kim, K-J and Cho, S-B (2008) "An Evolutionary Algorithm Approach to Optimal Ensemble Classifiers for DNA Microarray Data Analysis" in IEEE Transactions on Evolutionary Computation Vol. 12 No. 3, June 2008.