Prediction of Colon Cancer Using Evolutionary Multi-Agent System for Microarray Analysis

Gregor Stiglic
Laboratory for System Design
University of Maribor
Smetanova 17, SI-2000 Maribor
Slovenia
gregor.stiglic@uni-mb.si

Peter Kokol
Laboratory for System Design
University of Maribor
Smetanova 17, SI-2000 Maribor
Slovenia
kokol@uni-mb.si

Abstract – **This paper presents an application of Evolutionary Multi-Agent System (EMAS) to the analysis of gene expression data. Our goal is to find significant classification genes using simple classifiers that can be used by agents when searching through the gene expression database search space. This way we can get a small subset of significant features (genes) that can help us identifying the clinical state of the patient. The experiments show that agents improve their individual performance through evolution and collaboration with other agents. We present our results on the expression profiles of 2000 genes in 22 normal and 40 colon tumor tissues.**

I. INTRODUCTION

Microarrays have become important tool in profiling gene expression patterns. Their main advantages are: reproducibility and scalability of obtained data, short time of experiment and, of course, the large number of genes, the expression of which is measured. The technique of producing DNA microarrays is improving continuously. The results of improvement are better and more accurate gene expression databases. The problem in analysis of such databases is their multi-dimensionality, where we have large number of features (genes) and only a few instances (samples).

As a possible solution to the problem of classification in gene expression data, we propose a simple nearest neighbor classification method using only two features at a time. To narrow the large search space we employ the system of evolving agents searching for the best classifier. Agents in multi-agent systems are naturally led to building systems that adapt and learn through experience [1]. In our case agents can exchange information about the position of promising classification possibilities in two-dimensional feature space. Based on this fact two types of agents exist. The first type of agents are "static agents" which search in the proximity of the current best solution. The second type of agents are "dynamic agents" (we call them explorers) who are able to explore the search space without constraints. Using this technique we try to optimize current best solution and search for possible new promising points in the search space.

Tumor classification usually included diagnostic techniques like micro and macroscopic histology and tumor morphology. This way, however, we are unable to discriminate among tumors with similar histopathologic features. In the recent years we can notice the shift of interest from morphologic to molecular tumor classification.

Colon cancer database we are using in our paper was first presented and analyzed by Alon et al. in [2] and was later frequently used as a benchmark for testing the accuracy of gene expression classification methods on 62 human tissue samples. The results of the clustering method misclassified 8 samples. Three normal tissues were classified as tumor and five tumor tissues as normal. Another test on the mentioned database was performed using support vector machines (SVM) and was published in [3]. The results using SVM misclassified six samples (3 tumors and 3 normal tissues). Nguyen and Rocke [4] applied two feature selection (principal component analysis and partial least squares) and two classification methods (logistic discrimination and quadratic discriminant analysis) to the colon dataset. The best results were obtained using logistic discrimination which still missed four samples using leave-one-out cross-validation (LOOCV). From all investigated methods we found only one that managed to classify all samples correctly using LOOCV. It was performed by Fujarewicz and Wiench in [5]. They achieved these results using a combination of recursive feature selection (RFR) and SVM methods.

In the next section we describe the multi-agent environment and present basic agent properties and functions. After that a section with the results of predictive gene selection on tumor classification database is presented. In the final section we discuss about our method and possible further improvements of the multi-agent system for predictive gene discovery.

II. EVOLUTIONARY MULTI-AGENT SYSTEM

We know two types of hybrid systems combining evolutionary computation (EC) and multi-agent systems (MAS). In the first case we use evolution to help an agent solving problems using evolutionary techniques. In the other case we try to combine EC and MAS even more closely by using agents as a population of the evolutionary environment. The key idea of our system, which follows the second mentioned case, is that besides interaction mechanisms typical for MAS (such as communication) agents are able to reproduce (generate new agents) and may die (be eliminated from the system) [6]. A decisive factor of the agent's activity is fitness level, expressed by the ability of finding the best combination of genes for classification. Each agent presents a pair of features in the final ensemble of classifying agents and therefore we can keep a high level of diversity in the ensemble [7, 8, 9].

Ho introduced the idea of ensembles of k-Nearest Neighbor (k-NN) classifiers where the variety in the ensemble is generated by selection of different feature subsets for each ensemble in [10]. Since she generates

these feature subsets randomly she refers to these different subsets as random subspaces. She points to the ability of ensembles of k-NN classifiers based on different feature subsets to improve on the accuracy of individual k-NN classifiers because of the simplicity and accuracy of the k-NN approach. She shows that an ensemble of k-NN classifiers based on random subsets improves on the accuracies of individual classifiers on a hand-written character recognition problem.

A. Genotype and agent behavior

In the evolutionary system we have to follow the basic principles of the evolution theory like selection and inheritance. We can apply those principles to agents in form of:

- Death (elimination of agents from the system) and
- Reproduction (production of new agents)

Basic behavior parameters of agent are encoded in genotype and are inherited from its parent(s). Those parameters can be modified using mutation and recombination. Mutation in biology and also in computer science is the local search part of the evolution. Therefore the mutation operator should not significantly change the genotype parameters. This is accomplished allowing smaller changes with higher probabilities and larger ones with less. We use mutation modification distributed using a normal probability function as in research by Oechslein et al. [11].

In our system genotype of each agent consists of the following parameters:

- Exploring capability
- Speed
- Crowd factor
- Type of classifier

Exploring capability defines the level of agent's movement and can range from static to dynamic. Low exploring capability defines an agent as static, which means it will mainly search for the best solution near the best-known solution in the search space. In other case an agent has a freedom to explore in the unexplored search space.

Speed defines maximum movement capability when agent moves in the search space (i.e. maximum distance of move).

Crowd factor is a parameter that enables control over crowding effect. Our system allows dividing search space into n^2 equal sectors. Crowd factor defines a maximum number of agents in the sector before an agent moves to other sector because of overcrowding.

Type of Classifier represents parameters of used classifier. In our case this is the parameter of k-NN classifier that defines how many neighbors will contribute to the final vote of the test case.

Additional to agent's evolving parameters all agents have to follow some common rules that help them find better solutions in shorter time. Therefore each agent should follow these rules:

- Try to get an information in which areas of search space other agents found useful classification genes
- If the agent is "static" it should search near the best solutions
- Agents with the value of exploring parameter between "static" and "dynamic" can search far away from the best solution, but should try to follow the horizontal or vertical line from the best solution (this way an agent is using one of the genes selected by the best classifier so far)

B. Fitness Function

The k-nearest neighbors (k-NN) algorithm is a simple but effective classification algorithm. It is widely used in machine learning and has numerous variations [12, 13]. Given a test sample of unknown label, it finds the k nearest neighbors in the training set using Euclidean distance (d) and assigns the label of the test sample according to the labels of those neighbors. We use five different types of classifiers using 1 to 5 nearest neighbors. To ensure a majority in the voting process we use vote weighting where neighbor's vote is weighted by its distance to the test sample. The weight given to each vote is $1/d$.

When considering which accuracy estimation technique to use, we have an option to use a hold-out procedure, using a part of the samples to train a predictive model and using the rest of instances as a test set to estimate classifier accuracy. Another possibility is using n-fold cross-validation, which is typically implemented by running the same learning system n times, each time on a different training set of size $(n-1)/n$ times the size of the original data set. Because of this computational cost, cross-validation is sometimes avoided, even when it is agreed that the method would be useful. This is often the case in usual machine learning problems while in microarray classification the computational cost is usually not a problem because of a small number of samples. That is why a lot of authors use leave-one-out cross-validation method, in which one sample in the training set is withheld, the remaining samples of the training set are used to build a classifier to predict the class of withheld sample, and the cumulative error is calculated. LOOCV was often criticized, because of higher error variance in comparison to 5 or 10-fold cross-validation [14]. A recent study by Braga-Neto and Dougherty [15], stating they have not been able to verify the substantial difference in performance among mentioned methods, shows that LOOCV can still be considered as the most optimal classifier for microarray datasets.

In all our tests we use LOOCV for classifier accuracy estimation.

III. COLON CANCER DATASET

Colon cancer is second only to lung cancer as a cause of cancer-related mortality [16]. It is a genetic disease, propagated by the acquisition of somatic alterations that influence gene expression. Using DNA microarray

technology we are able to measure the expression level of thousands of genes simultaneously. The most exciting result of microarray technology research in the past has been the demonstration that patterns of gene expression can distinguish between tumors of different anatomical origin.

The first extensive study on detection of colon cancer using gene expression analysis was done by Alon et al. [2]. In this study, gene expression in 40 tumor and 22 normal tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6500 human genes. In our database we used 2000 statistically most significant genes that were chosen in the paper [2].

The paper of Alon et al. provides an analysis of gene expression data using top down hierarchical clustering. They show that most normal tissue samples cluster together and most cancer tissue samples cluster together. They explain that "outlier" samples that are classified in the wrong cluster differ in cell composition from typical samples. So called "muscle index" that measures the average gene expression of a number of smooth muscle genes is computed. Most normal samples have high muscle index and cancer samples low muscle index. The opposite is true for most outliers. Alon et al. also show that some genes are correlated with the cancer vs. normal separation but do not suggest a specific method of gene selection.

IV. EXPERIMENTAL RESULTS

Because we were using Nearest Neighbor classifier we had to do a simple preprocessing step before the actual testing began. We normalized all expression data using linear transformation to [0, 1] range. This way we prevent features with initially large ranges from outweighing features with initially smaller ranges. Using evolutionary method, we can expect lower time complexity of our method – usually we get very good solutions in less than ten minutes of running time on an average personal computer.

TABLE I

CLASSIFICATION ACCURACY OF THE BEST CLASSIFYING AGENTS

#	Classification Accuracy	Classifier Used	Predictive Gene Ids
1	95.2 %	2-NN	1967, 1449
2	93.5 %	4-NN	323, 625
3	93.5 %	2-NN	1411, 1047
4	91.9 %	3-NN	822, 763
5	91.9 %	4-NN	249, 769
6	91.9 %	1-NN	293, 1293
7	91.9 %	5-NN	1423, 1472
8	91.9 %	3-NN	765, 1715
9	91.9 %	5-NN	69, 493
10	91.9 %	4-NN	769, 249

In the first table (Table 1) we can see performance of the best ten agents with the smallest error rates on the Colon cancer database. The best performance was achieved using agent with 2-NN classifier, which reached classification accuracy of 95.2% in less than 3000 iterations (Fig. 1).

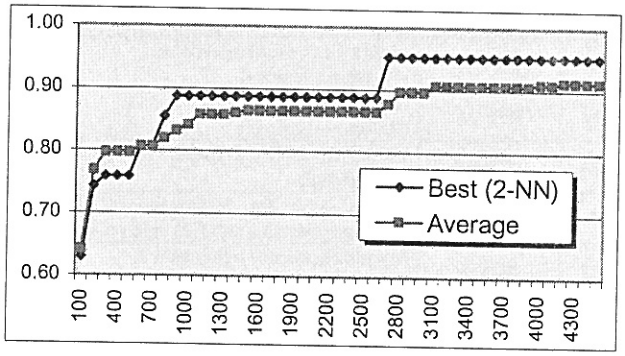

Fig. 1. Classification accuracy of the best against average agent in the first 4500 iterations

From the gene expression levels (Fig. 2) it is evident that the genes are discriminating jointly and not for themselves – we cannot spot any significant difference between the left and right group of samples (normal and colon cancer samples). We present only the 20 most important gene expressions here.

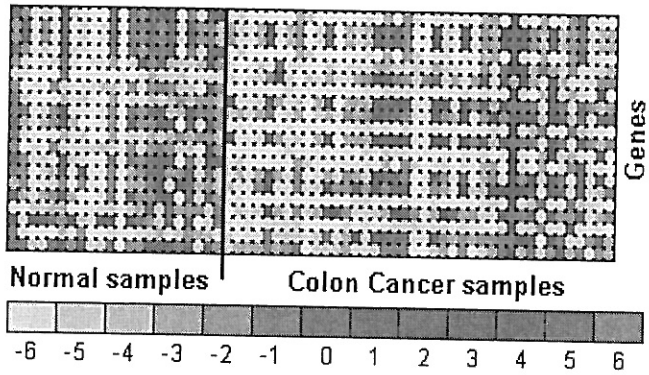

Fig. 2. Gene expressions for normal and colon cancer tissue samples

We also present the results of the accuracy estimation for the ensembles of 3, 5, 7 and 9 best performing agents. Odd number of agents in the ensemble is used, because each agent contributes a single vote for the questioned sample. Using LOOCV we get an overall accuracy for classification of all samples (Table 2).

TABLE II

CLASSIFICATION ACCURACY OF THE BEST CLASSIFYING ENSEMBLES OF AGENTS

#	Classification Accuracy	Agents Used	Ensemble
1	98.4 %	1 – 3	Top 3
2	96.8 %	1 – 5	Top 5
3	96.8 %	1 – 7	Top 7
4	98.4 %	1 – 9	Top 9
5	100.0 %	1, 2, 7	Combination of 3

After evaluation of accuracy for ensembles of top performing agents, we try to find the best combination of agents in the smallest possible ensemble. In the process of searching for the best combination of agents we try all combinations of three agents out of best ten agents. This way we find a combination of 6 genes (3 agents) that can classify all samples correctly. We have to know that it would take a lot more time to search the space using agents with 6 genes at a time. The most important genes for colon cancer detection are also presented in the Table 3.

TABLE III

LIST OF THE MOST SIGNIFICANT GENES

#	Name	Description
1967	T60778	Matrix GLA-protein precursor (Rattus norvegicus)
1449	R38758	Synaptic vesicle protein 2 (Rattus norvegicus)
323	U28963	Human Gps2 (GPS2) mRNA protein, complete cds
625	X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
1423	J02854	Myosin regulatory light chain 2, smooth muscle isoform (human)
1472	L41559	Homo sapiens pterin-4a-carbinolamine dehydratase (PCBD) mRNA, complete cds

Interesting conclusion from the list of highest ranked genes is that both genes in the first pair come from "Rattus norvegicus" group. The second gene (R38758) can also be found in the results of some rank-based gene selection models like [17], however we could not find the first gene in any rank-based method. This confirms that we can find promising solutions combining the genes that are linearly correlated to the decision class with those who are not.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel concept of searching simple classifiers in gene expression data classification problem. Our approach demonstrates its efficiency and effectiveness in dealing with high dimensional data for classification. The obtained results confirm that there is no need to reduce the initial set of genes using statistic gene ranking, as it is usually the case in microarray analysis. Using evolutionary multi-agent based system combined with simple nearest neighbor classifiers we can get very good results that are comparable to other much more complex methods. Our system is also very useful as a feature selection method and can effectively reduce the number of needed genes for discrimination between tissue classes. We prove this by comparing our result of no misclassified samples in Colon cancer dataset to the result by Fujarewicz and Wiench [5] who were using support vector machines. We matched their result using only three pairs of genes combined in an ensemble of nearest neighbor classifiers. Our further work will extend our concept to work with higher dimensionality (combination of three, four or five genes). We will also devote more attention to the comparison of statistically selected genes (rank-based methods) and genes selected using our method.

ACKNOWLEDGMENT

The authors would like to thank Alon et al. [2] for the original colon gene expression database, which is freely available at http://microarray.princeton.edu/oncology/.

REFERENCES

[1] P.J. Modi and W.M. Shen, "Collaborative Multiagent Learning for Classification Tasks," *Proceedings of the Fifth International Conference on Autonomous Agents*, 2001, pp. 37-38.

[2] U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci.*, Vol. 96, pp. 6745-6750.

[3] T.S. Furey et al., "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, Vol. 16, No. 10, pp. 906-914.

[4] D.V. Nguyen and D.M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, Vol. 18, No. 1, pp. 39-50.

[5] K. Fujarewicz and M. Wiench, "Selecting differentially expressed genes for colon tumor classification", *Int. J. Appl. Math. Comput. Sci.*, Vol. 13, No. 3, pp. 327-335.

[6] K. Socha and M. Kisiel-Dorohinicki, "Agent-based Evolutionary Multiobjective Optimisation", *Proceedings of CEC'02 - Congress on Evolutionary Computation*, Vol.1, 2002, pp. 109-114.

[7] G. Valentini and F. Masulli, "Ensembles of learning machines", *Neural Nets WIRN Vietr*, Vol. 2486, 2002, pp. 3-19.

[8] L. Kuncheva, "That Elusive Diversity in Classifier Ensembles", *Proceedings of Pattern Recognition and Image Analysis – IbPRIA*, Vol. 2652, 2003, pp. 1126-1138.

[9] P. Cunningham and J. Carney, "Diversity versus Quality in Classification Ensembles Based on Feature Selection", *Proceedings of 11th European Conference on Machine Learning*, Vol. 1810, 2000, pp. 109-116.

[10] T.K. Ho, "Nearest Neighbours in Random Subspaces", *Proceedings of 2nd International Workshop on Statistical Techniques in Pattern Recognition*, 1998, pp. 640-648.

[11] C. Oechslein, A. Hörnlein and F. Klügl, "Evolutionary Optimization of Societies in Simulated Multi-Agent Systems", *Modelling Artificial Societies and Hybrid Organizations*, 2000.

[12] R. Duda, P.E. Hart and D.G. Stork, "*Pattern classification*", Wiley, 2001.

[13] C. Yeang, "Molecular Classification of Multiple Tumor Types", *Bioinformatics*, Vol. 17, suppl. 1, 2001, pp. 316-322.

[14] T. Hastie, R. Tibshirani and J. Friedman, "*The Elements of Statistical Learning*", Springer, 2001.

[15] M. Braga-Neto and E.R. Dougherty, "Is cross-validation valid for small-sample microarray classification? ", *Bioinformatics*, Vol. 20, no. 3, pp. 374-380.

[16] G.A. Chung-Faye, D.J. Kerr, L.S. Young, P.F. Searle, "Gene therapy strategies for colon cancer", *Mol. Med. Today*, Vol. 6, no. 2, pp. 82–87.

[17] Y. Su, T.M. Murali, V. Pavlovic, M. Schaffer and S. Kasif, "RankGene: identification of diagnostic genes based on expression data", *Bioinformatics*, Vol. 19, no. 12, pp. 1578-1579.