Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics
& Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

A Novel Ensemble Selection Method for Cancer Diagnosis Using Microarray Datasets

Mohammed A. Gaafar
Computer and System Engineering
Alexandria University
Alexandria, Egypt
mohammed.a.gaafar@gmail.com

Noha A. Yousri
Computer and System Engineering
Alexandria University
Alexandria, Egypt
noha.yousri@alexu.edu.eg

Mohamed A. Ismail
Computer and System Engineering
Alexandria University
Alexandria, Egypt
drmaismail@gmail.com

***Abstract*—Ensembles of classifiers have recently proved their efficiency in cancer diagnosis based on microarray datasets. The main performance indicators, namely, accuracy and diversity, present the main focus of study when designing an ensemble. One other important performance indicator is classification robustness. In an attempt to improve the performance of an ensemble, the proposed algorithm presents a variation concerning the diversity method used. The proposed algorithm attempts to enhance the robustness of the classification by searching for an ensemble of diverse classifiers. Also, a comparison of the different diversity methods is presented in order to study their impact on the robustness of the classification. The experiments performed show that the diversity method used in the proposed algorithm outperforms the other diversity methods in terms of accuracy and robustness.**

***Keywords*-Gene Selection, Ensemble Selection, Microarray Classification, Cancer Classification.**

I. Introduction

Microarray technology gained great attention recently due to its ability to measure expression levels for thousands of genes simultaneously. In recent years, microarrays were employed in cancer diagnosis by classifying microarray samples [1] and clustering of gene expression profiles [2], [3]. The classification of microarray samples introduced many challenges due to the special nature of microarray data. One of the challenges it encountered is the high dimensionality problem. A microarray dataset consists of few tens or hundreds of samples and thousands of genes. Despite this large number of genes, only few of them are relevant to a specific cancer classification problem [4]. To overcome this problem many gene selection algorithms were proposed. The main goal of a gene selection algorithm is to enhance the classification accuracy by selecting the most informative genes. Informative genes are the ones whose expression values can distinguish a specific class of tumor or tumor subtype [5]. Gene selection algorithms can be classified into three main classes [6]; filters, wrappers and embedded methods. Filtering methods [7] select genes by measuring their relevance to the classification problem. Many measures were introduced like t-test [8], Pearson Correlation [9], Shanon's Entropy [8] and Mutual Information [10]. Wrapper methods [11] employ the classifier in the selection process by searching for the gene subset that gives the best classification accuracy. Embedded methods [6] are the methods that use the classifier itself as the feature selector such as *C4.5* and *ID3* [12]. One important issue with gene selection algorithms is the possibility of selecting redundant genes [13]. Many gene selection algorithms were introduced to avoid selecting redundant genes. Those methods aim at finding the most informative genes and at the same time minimizing the redundancy between selected genes.

Along with an efficient gene selection algorithm an accurate and efficient classifier should be employed to achieve high classification accuracy. Many classifiers were used to classify microarray samples such as K Nearest Neighbor (*KNN*) [14], Support Vector Machines (*SVM*) [15], Artificial Neural Networks (*ANN*) [16], and Decision Trees (*DT*) [17]. Although most of the mentioned classifiers may provide satisfiable accuracy, they encounter many issues such as over-fitting and low robustness. A classifier is considered robust if its output does not change with perturbations in the training data. The ensemble of classifiers approach was proposed in order to overcome the over-fitting problem and to enhance accuracy and robustness of single classifiers.

An ensemble of classifiers works by combining the output of its members using a voting or fusion scheme [18]. It can outperform its individual members if they are diverse [19]. Two classifiers are considered diverse if they produce different errors on the same set of samples [19]. Diversity of the ensemble members can be provided on one or more of the following three levels [18].

1) The samples level (Tr): Resampling of the training data is performed to train ensemble members by different subsets of the training data.
2) The features level (FS): Different feature subsets are used to train each ensemble member.
3) The classifiers level (Cls): Different classification algorithms are used for each ensemble member or the same classification algorithm is used but with different parameters.

The different diversity methods and their different combinations are summarized in *Table I* with their notations used in the remaining of the paper.

The main objective of the work presented in this paper is to enhance the robustness and the accuracy of ensembles

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

Cobmination	Classifiers Level	Features Level	Bagging
Cls	✔	-	-
FS	-	✔	-
Tr	-	-	✔
Cls_FS	✔	✔	-
FS_Tr	-	✔	✔
Cls_Tr	✔	-	✔
Cls_FS_Tr	✔	✔	✔

TABLE I: The Different Diversity Methods

of classifiers. In order to enhance the robustness, the algorithm increases the diversity between ensemble members by introducing diversity at two levels; features level, and classifiers level. To enhance the accuracy, the algorithm uses a gene selection algorithm that considers redundancy between selected genes. The proposed algorithm employs a genetic algorithm (*GA*) to search the space of possible ensembles for the most accurate and diverse ensemble of classifiers.

The remaining of this paper is organized as follows: a brief survey on current gene selection and ensemble selection algorithms is introduced in *section II*. In *section III*, the proposed algorithm is presented. In *section IV* and *section V* the experimental setup and the results of the performed experiments are introduced and discussed. Finally, *section VI* contains the conclusion and future work.

II. Related Work

In recent years, microarray samples classification and its challenges have acquired great attention. One of the challenges was the gene selection problem. As mentioned in the previous section, one problem with gene selection is the possibility of selecting redundant genes which may cause loss of information that affects the classification accuracy [13]. In order to solve this problem, many algorithms have been proposed. Ding et al. [13] have proposed the Maximum Relevance Minimum Reduncy (*MRMR*) algorithm. This algorithm selects genes iteratively. In each iteration, a gene is selected that has the maximum relevance to the studied classes and minimum redundancy with the genes selected in the previous iterations. Relevance of a gene is measured using any of the measures mentioned in *section I* and redundancy between genes is measured using *Pearson Correlation* or *Mutual Information*. Liu et al. [20] have proposed the use of *Conditional Mutual Information* to measure redundancy between genes along with Mutual Information to measure the relevance of the genes. They also proposed Information Correlation Coefficient (*ICC*) to measure the relevance of the genes [5]. And to reduce redundancy between selected genes, *Approximate Markov Blankets* were used. El Akadi et al. proposed *IGFS* [21] method to select genes using mutual information as a relevance measure and conditional mutual information to measure redundancy and interaction between genes.

Equally, ensemble selection problem became very important in the recent years. Kim and Cho [22] have proposed an ensemble selection algorithm that uses a *GA* to search for an ensemble of classifiers using a fitness function that considers accuracy and the number of ensemble members. Diversity between classifiers is introduced at the features and classifiers levels. The algorithm uses many different filtering gene selection algorithms to generate gene subsets that are used to train the ensemble members. Chen and Zhao [23] have proposed an ensemble selection algorithm that uses Estimation of Distribution Algorithm (*EDA*) [24] to search for an ensemble of classifiers with minimal error. The algorithm employs many measures to generate gene subsets using ideal marker genes. Those gene subsets are used to train *ANN* classifiers. Liu [25] has proposed an ensemble selection algorithm that uses *GA* to search for an ensemble of classifiers that has maximum accuracy and diversity. An ensemble of filtering techniques is used to generate a feature pool. The space of solutions consists of feature subsets used to train the ensemble members and weights for the ensemble members to be used for the voting to find the final answer. The fitness function for an ensemble is the sum of its accuracy and diversity.

The proposed algorithm tries to overcome the shortcomings of the aforementioned algorithms. This is discussed in details in the next section.

III. The Proposed Algorithm

Most of the ensemble selection algorithms mentioned in the previous section do not consider redundancy between selected genes. Also, some of them do not consider diversity in the fitness or objective function they use. The proposed algorithm tries to overcome those problems by using *MRMR* algorithm for gene set reduction and gene subsets generation. Then, *GA* is used to search for an ensemble of *KNN* classifiers with maximum diversity and accuracy. *KNN* classifiers are used because of their low computational cost compared to other classifiers such *SVM* and *ANN*.

An important key in designing an efficient and accurate ensemble of classifiers is to increase diversity between its members. Increasing the diversity between ensemble members results in reducing the possibility that they produce errors on the same samples and increasing the robustness of the ensemble. In order to increase diversity between the ensemble members, the proposed algorithm introduces diversity at two levels:

1) Classifiers Level: Each ensemble consists of $\mathcal{C}$ *KNN* classifiers with different k values.
2) Features Level: Ensemble members are trained using disjoint gene subsets.

The problem of ensemble selection can be formalized as follows; given a microarray dataset $\mathcal{D}$ consisting of $\mathcal{N}$ samples and $\mathcal{G}$ genes, it is required to select a robust ensemble of classifiers consisting of n or less classifiers.

The proposed algorithm (demonstrated in *Figure 1*) consists of two phases; generation phase and search phase. The generation phase consists of two steps; gene set reduction and gene subsets generation. In the first step, reduction is carried out using *MRMR* algorithm [13]. *MRMR* algorithm works as follows; first, the most relevant gene is selected, and then for

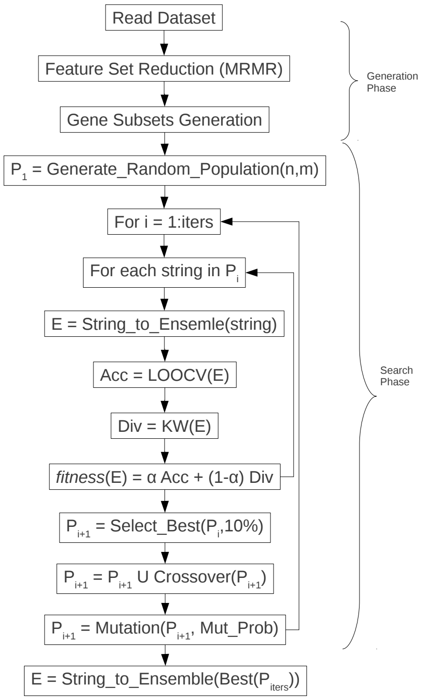
Read Dataset
Feature Set Reduction (MRMR)
Gene Subsets Generation
Generation Phase
P1 = Generate_Random_Population(n,m)
For i = 1:iters
For each string in Pi
E = String_to_Ensemle(string)
Acc = LOOCV(E)
Div = KW(E)
fitness(E) = α Acc + (1-α) Div
Search Phase
Pi+1 = Select_Best(Pi,10%)
Pi+1 = Pi+1 U Crossover(Pi+1)
Pi+1 = Mutation(Pi+1, Mut_Prob)
E = String_to_Ensemble(Best(Piters))

Fig. 1: Outline of The Proposed Algorithm

$\mathcal{M}$ iterations a gene is selected according to the objective function in *Equation 1*.

$$\underset{g\in\Omega}{m}ax\,[\,\mathcal{I}(g,h)-\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathcal{I}(g,i)\,] \tag{1}$$

Where g is the selected gene, h is the classification model, Ω is the set of the unselected genes, $\mathcal{S}$ is the set of previously selected genes, and $\mathcal{I}$ is the mutual information measure. The first term of the objective function measures the relevance of a gene to the classification model. The second term measures the redundancy between a gene and the previously selected ones. The use of *MRMR* algorithm helps in increasing diversity between ensemble members because the selected genes share less information [9].

In the second step, n disjoint gene subsets with cardinality m are generated from the reduced gene set. Gene subsets are also generated using *MRMR*. To generate the gene subsets; first, the most n relevant genes are selected, each one is placed in a different gene subset. Then, for each gene subset *MRMR* algorithm runs for $m-1$ iterations to select the other genes in each gene subset.

In the search phase, a genetic algorithm is used to search the space for the most accurate and diverse classifier using the fitness function proposed by in [17]. The fitness function is shown in *Equation 2*.

$$Fitness(\mathcal{E}) \;=\; [\,\alpha\, ACC(\mathcal{E}) + (1-\alpha)\, DIV(\mathcal{E})\,] \tag{2}$$

Where ACC is the accuracy of the ensemble $\mathcal{E}$ and DIV is its diversity, and α is a weighting parameter between accuracy and diversity. α's value is set prior to the algorithm execution and it can take a value between 0(searching for the most diverse ensemble regardless of accuracy) and 1(searching for the most accurate ensemble regardless of diversity).

C1	C1	C2	C2	C3	C3	C4	C4

Fig. 2: Chromosome for Ensemble of Four Classifiers

In order to enable the *GA* to apply genetic operations such as cross-over and mutation, the ensembles should be transformed into strings. In the proposed algorithm strings of integers are used instead of binary strings. A sample string is shown in *Figure 2*. Each ensemble member is represented by two digits in the string; the first digit specifies the index of the gene subset that is used to train the classifier, and the second digit specifies the number of genes to be extracted from that gene subset. For example, if a classifier is represented by the two digits $\mathcal{XY}$ then the first $\mathcal{Y}$ genes from gene subset $\mathcal{X}$ are used to train the classifier.

The genetic algorithm works as follows; initially a population of pop_size ensembles is generated randomly, the fittest 10% ensembles are selected. Then, cross-over is applied to the selected ensembles to generate a new population. Cross-over operation in the proposed algorithm does not differ from the operation used with binary chromosomes [22]. Mutation is applied to the new population with probability mut_prob. Mutation is applied to a string by randomly changing the value of a randomly selected one of its digit. The previous procedure is performed for a specific number of iterations $iters$. Finally the fittest ensemble is selected as the final solution.

In order to obtain the final output of an ensemble, the individual outputs of the ensemble members are merged using a weighted fusion scheme. Each ensemble member is assigned a weight based on its individual error rate. The error rates of all the ensemble members are used to generate a weighting vector $\mathcal{W}$. In this weighting vector, higher weights are given to the classifiers with less errors. Those weights are used to obtain the final output of the ensemble according to *Equation 3*.

$$\mathcal{H} = \sum_{i=1}^{n}\mathcal{W}_i.h_i \tag{3}$$

Where $\mathcal{H}$ is the ensemble output, n is the number of ensemble members, $\mathcal{W}_i$ is the weight of ensemble member i, and h_i is the output of ensemble member i.

IV. Experimental Setup

The set of experiments performed compare the robustness, accuracy and diversity of the different diversity methods. This comparison aims at proving that the diversity method used *Cls_FS* in the proposed algorithm outperforms other diversity methods in terms of accuracy and robustness.

The proposed algorithm has been tested against 6 microarray datasets shown on *Table II*. Due to the fact that there are many parameters to be studied, this study focuses mainly on

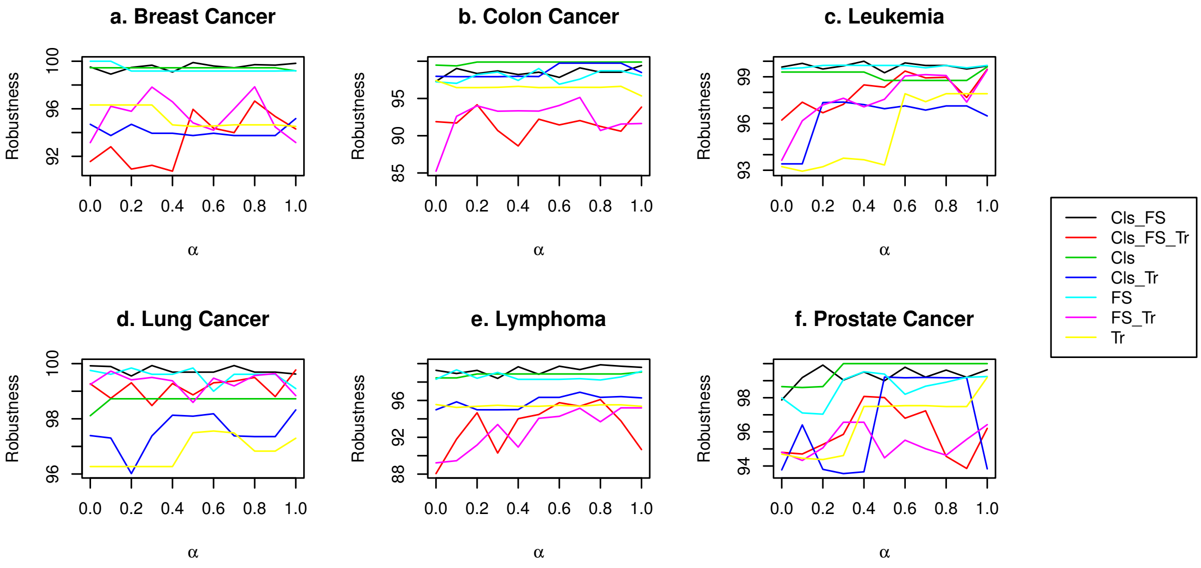

Fig. 3: Robustness of the 6 Datasets in the different combinations of diversity methods and different values for α

Datasets	# of Samples	# of Genes
Breast Cancer [26]	49	6817
Colon Cancer [27]	62	2000
Leukemia [28]	72	7129
Lung Cancer [29]	181	12600
Lymphoma [30]	77	6817
Prostate Cancer [31]	102	12600

TABLE II: Datasets Used in the Experiments

the weighting parameter in the objective function α and the combinations of diversity methods shown on *Table I*. The algorithm was implemented in R statistical tool [32].

The experiments were performed using the following configurations.

- $pop\_size = 500$, $iters = 100$, $mut\_pop = 0.05$
- Accuracy measure is Leave One Out Cross Validation.
- Diversity measure is Kohavi-Wolpert Variance (*KW*) [33].

After obtaining the final ensemble, perturbations were introduced in the datasets by removing randomly from 1 to 15% of the samples of each dataset. This procedure was performed 100 times. The robustness of the ensemble is calculated as the average similarity between the output of the ensemble trained using the original data and the ensembles trained using the perturbed data. The similarity between the output of two ensembles is calculated using the formula in *Equation 4*.

$$Sim(h_1, h_2) = \frac{\sum_{i=1}^{\mathcal{N}} (h_1(i) == h_2(i))}{\mathcal{N}} \qquad (4)$$

Where h_1 and h_2 are the outputs of two different ensembles and $\mathcal{N}$ is the number of samples. $Sim(h_1, h_2)$ ranges from 0 (the outputs of the two ensembles are completely different) to 1 (the outputs of the two ensembles are identical).

V. Results

Figure 3, *Figure 4* and *Figure 5* demonstrate the robustness, accuracy and diversity of the ensembles using the diversity methods shown in *Table I*. The plots of *Figure 3* demonstrate that there were three combinations which always gave better robustness with the six datasets using the different values of α. The three combinations are *Cls, FS,* and *Cls_FS* the diversity method used in the proposed algorithm. Some other diversity methods gave high robustness with a certain dataset such as *Tr* with Colon Cancer dataset in *Figure 3.b*, also *FS_Tr* and *Cls_FS_Tr* with Lung Cancer dataset in *Figure 3.d*. Those methods gave high robustness with only one dataset i. e. they are data dependent, but the other three methods (*Cls, FS,* and *Cls_FS*)gave high robustness regardless of the datasets.

It can be seen in the plots of *Figure 4* that four diversity methods gave the highest accuracy with the six datasets. Those combinations are the ones introduced diversity at the features level; *FS, Cls_FS_Tr, Tr_FS,* and *Cls_FS*. From *Figure 5*, it is noticed that the same combinations gave the highest diversity. In *Figure 5.d*, ensembles tested on Lung Cancer dataset have low diversity compared to other datasets because individual *KNN* classifiers used in the ensembles provide high accuracy on this dataset. This reduces the possibility of ensemble members to provide errors on different samples and reduces the ensemble's diversity. As discussed earlier in *section III*, increasing α's value increases accuracy and decreases diversity.

By comparing the robustness of the ensembles using the different diversity methods, it can be reported that introducing diversity on features level gives better accuracy and introducing it with diversity on classifiers level gives higher robustness. It can be concluded from those results that the diversity method

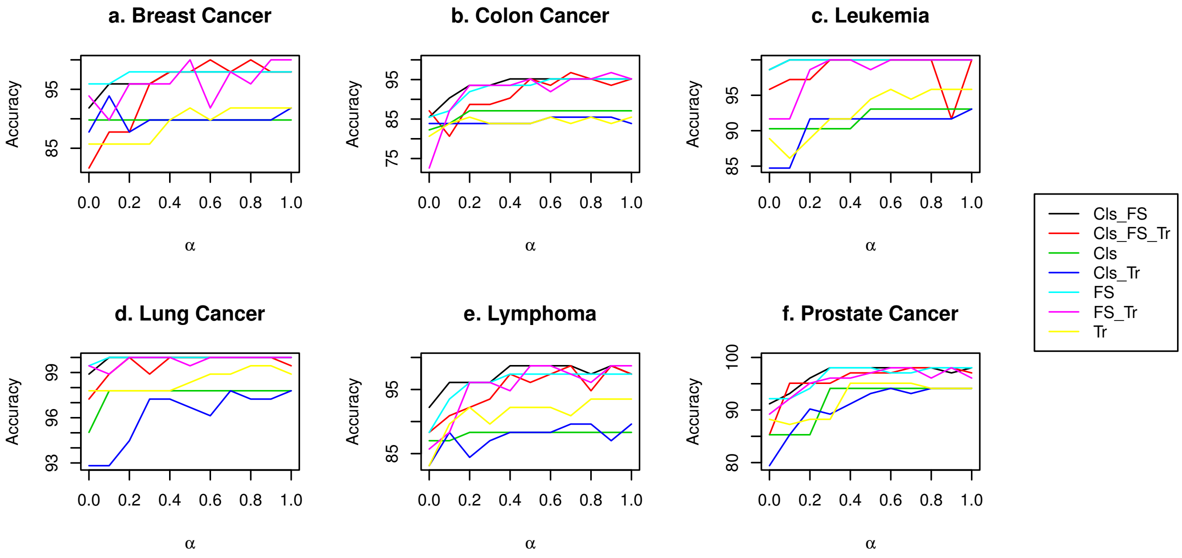

Fig. 4: Accuracy of the 6 Datasets in the different combinations of diversity methods and different values for α

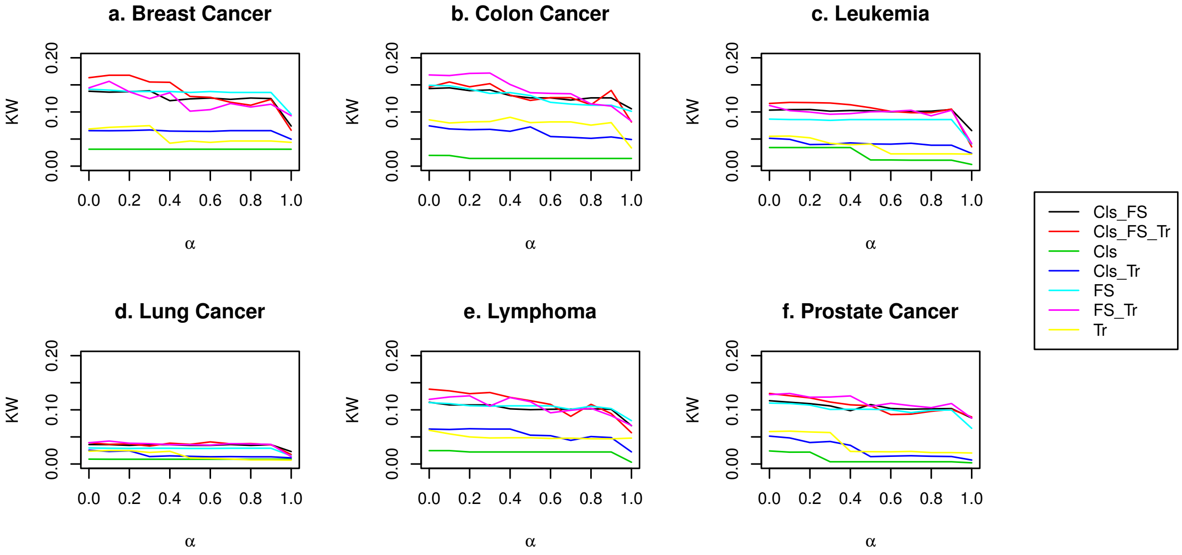

Fig. 5: Diversity of the 6 Datasets in the different combinations of diversity methods and different values for α

used by the proposed algorithm *Cls_FS* outperforms the other diversity methods in terms of accuracy and robustness.

VI. Conclusion

In this paper a novel ensemble selection method for microarray samples classification has been proposed. The algorithm aims at enhancing the robustness of the ensembles of classifiers. This was achieved by increasing diversity between ensembles members. The algorithm uses *GA* to search a space of ensembles using a fitness function that introduces a trade off between accuracy and diversity. The results showed that the chosen diversity method *Cls_FS* gives the highest robustness and accuracy for all the datasets and at the same time it gives satisfactory diversity values.

Future work includes studying the impact of changing the other parameters of the ensemble selection algorithm on the robustness of the ensembles such as the classifiers used and the diversity measure, etc. Also, It includes studying the impact of *GA* parameters. More testing for the algorithm with different type of datasets will be performed in order to widen the application domain of the proposed algorithm.

Acknowledgment

The authors would like to thank *Bibliotheca Alexandrina* for giving the permission to run the experiments on the High Performance Computer (*BA HPC*) there [34].

References

[1] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, mar 2002.

[2] N. A. Yousri, M. S. Kamel, and M. A. Ismail, "Pattern cores and connectedness in cancer gene expression," in *BIBE*. IEEE, 2007, pp. 100–107.

[3] N. A. Yousri, "Cluster-based characterization of gene over-expression in cancer sets," in *ISDA*. IEEE, 2010, pp. 74–79.

[4] X. Wang and O. Gotoh, "Inference of cancer-specific gene regulatory networks using soft computing rules," *Cancer Informatics*, vol. 9, pp. 15–30, Feb 2010.

[5] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 81–87, Feb. 2010.

[6] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *ARTIFICIAL INTELLIGENCE*, vol. 97, pp. 245–271, 1997.

[7] H. Mahmoodian, I. Saripan, M. Marhaban, R. Rahim, and R. Rosli, "New entropy-based method for gene selection," *IETE Journal of Research*, vol. 55, no. 4, pp. 162–168, 2009.

[8] I. n. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains." *Artificial intelligence in medicine*, vol. 31, no. 2, pp. 91–103, jun 2004.

[9] K.-J. Kim and S.-B. Cho, "Ensemble classifiers based on correlation analysis for dna microarray classification," *Neurocomputing*, vol. 70, no. 1-3, pp. 187–199, 2006.

[10] R. Cai, Z. Hao, X. Yang, and W. Wen, "An efficient gene selection algorithm based on mutual information," *Neurocomputing*, vol. 72, no. 46, pp. 991 – 999, 2009, brain Inspired Cognitive Systems (BICS 2006) / Interplay Between Natural and Artificial Computation (IWINAC 2007).

[11] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intellegence*, vol. 97, no. 1-2, pp. 273–324, Dec. 1997.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[13] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *J Bioinform Comput Biol*, 2003, pp. 523–529.

[14] L. Li, T. A. Darden, C. R. Weingberg, A. J. Levine, and L. G. Pedersen, "Gene assessment and sample classification for gene expression data using a genetic algorithm / k-nearest neighbor method," *Combinatorial Chemistry &amp;#38; High Throughput Screening*, pp. 727–739, dec 2001.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, mar 2002.

[16] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. D. Moor, "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection," *Artificial Intelligence in Medicine*, vol. 29, no. 12, pp. 39 – 60, 2003, artificial Intelligence in Medicine Europe AIME '01.

[17] D. Gacquer, V. Delcroix, F. Delmotte, and S. Piechowiak, "On the effectiveness of diversity when training multiple classifier systems," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ser. Lecture Notes in Computer Science, C. Sossai and G. Chemello, Eds. Springer Berlin / Heidelberg, 2009, vol. 5590, pp. 493–504.

[18] A. M. Canuto, K. M. Vale, A. Feitos, and A. Signoretti, "Reinsel: A class-based mechanism for feature selection in ensemble of classifiers," *Applied Soft Computing*, vol. 12, no. 8, pp. 2517 – 2529, 2012.

[19] T. G. Dietterich, "Ensemble methods in machine learning," in *INTERNATIONAL WORKSHOP ON MULTIPLE CLASSIFIER SYSTEMS*. Springer-Verlag, 2000, pp. 1–15.

[20] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recogn.*, vol. 43, no. 8, pp. 2763–2772, Aug. 2010.

[21] A. E. Akadi, A. E. Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, pp. 116–121, 2008.

[22] K.-J. Kim and S.-B. Cho, "An evolutionary algorithm approach to optimal ensemble classifiers for dna microarray data analysis," *Trans. Evol. Comp*, vol. 12, no. 3, pp. 377–388, Jun. 2008.

[23] Y. Chen and Y. Zhao, "A novel ensemble of classifiers for microarray data classification," *Appl. Soft Comput.*, vol. 8, no. 4, pp. 1664–1669, Sep. 2008.

[24] P. Larrañaga, *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Boston/Dordrecht/London: Kluwer Academic Publishers, 2002, ch. An introduction to probabilistic graphical models, pp. 25–54.

[25] K.-H. Liu, "The classification of microarray data using evolutionary classifier ensemble system," *IEIT Journal of Adaptive and Dynamic Computing*, vol. 2011, no. 4, pp. 34–39, 2011.

[26] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11 462–11 467, 2001.

[27] U. Alon, N. Barkai, D. A. Notterman, K. Gishdagger, S. Ybarradagger, D. Mackdagger, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, jun 1999.

[28] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct 1999.

[29] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[30] M. A. Shipp, K. N. Ross, P. Tamayom, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, A. T. Lister, and J. Mesirov, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 20, pp. 68–74, 2002.

[31] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203 – 209, 2002.

[32] "The r project for statistical computing," accessed: 10/6/2012. [Online]. Available: http://www.r-project.org

[33] R. Kohavi and D. H. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *MACHINE LEARNING: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL*. Morgan Kaufmann Publishers, 1996, pp. 275–283.

[34] "Bibliotheca alexandrina supercomputer project web page," accessed: 10/6/2012. [Online]. Available: http://www.bibalex.org/ISIS/Frontend/Projects/ProjectDetails.aspx?id=m8fC7jXMTFprEy98pIPBFw==