Computer and Information Science; Vol. 8, No. 3; 2015
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

A Novel Hyper-Active Algorithm to Estimate Missing Microarray Attributes

Baydaa Al-Hamadani[1] & Thikra Shubita[1]

[1] Faculty of Information Technology, Zarqa University, Jordan

Correspondence: Baydaa Al-Hamadani, Faculty of Information Technology, Zarqa University, P.O.Box 132222, Zarqa 13132, Jordan. Tel: 962-77-566-3578. E-mail: bhamadani@zu.edu.jo

Received: June 25, 2015 Accepted: July 14, 2015 Online Published: July 27, 2015

doi:10.5539/cis.v8n3p186 URL: http://dx.doi.org/10.5539/cis.v8n3p186

This research is funded by the Deanship of Research in Zarqa University /Jordan

Abstract

Classification the Microarray dataset is a powerful method used in clinical and biomedical studies, to estimate and diagnose some diseases like (cancer, non-cancer) depending on Gene expression. To be full beneficial, the gene expression dataset should be complete; i.e. with no missing data. Several approaches were proposed to deal with these missing values. In this paper, a robust algorithm is proposed based on the optimal fitting analysis to estimate the missing values in the microarray data. Then, the complete dataset is used to estimate the probability of lung cancer occurrence based on stochastic algorithm and support vector machine (SVM). The designed algorithm has been applied on different types of datasets varies from complete to different percent of missing data. Comparisons have been done with different other algorithms from the accuracy and error rates perspectives. The experimental results indicate that the proposed algorithm surpass other tested methods.

Keywords: Lung cancer, Microarrays, Gene expressions, Missing attributes

1. Introduction

Usually data from microarray experiments contain missing value either because of dust or scratches on the slide, error in experiments, image corruption and insufficient resolution (Mohammadi and Saraee, 2008). Many algorithms may require complete dataset to analyze gene expression, for example K-Mean clustering which loss effectiveness if deal with dataset having missing value (Ahmad and Miad, 2013). The proposed method depends on the microarray dataset analysis, which contains missing values as a common problem. To reduce the effect of analyzing the incomplete dataset we will estimate the missing value in gene microarray dataset.

Moreover, depending on Support Vector Machine (SVM), the system has the ability to learn from existing patient's results to use in diagnosing other cases. This property emphasizes the reusability of existing results, another classifier could be used is Ada-Boosting to classify and evaluate the final result of our model. The dataset used to test the system was the lung cancer Gene expression datasets.

Cancer is an important cause of death worldwide. There were an estimated 14.1 million new cases of cancer in the world: (53%) in males (47%) in females. Incidence rate shows that there are 205 new cancer cases for every 100,000 men in the world, and 165 for every 100,000 females (UK, 2013). The estimated number of people in Jordan is 155.4 cases of cancer per 100,000 and estimated 6,383 cases per year (Media, 2014). However, Lung cancer causes more deaths than the other three most common cancers together (Bowel, Breast and Prostate). Lung cancer special type of cancer represents the first place in worldwide and third place in Jordan (UK, 2013).

The first stage begins by making the dataset ready to be used by applying the preprocessing and production to determine the valuable data. The proposed system classifies the dataset and analyses it according to machine learning algorithms and techniques, for example (SVM), (GA) and Ada-Boosting. This will make our results meaningful to be predications and facilitating for implementation.

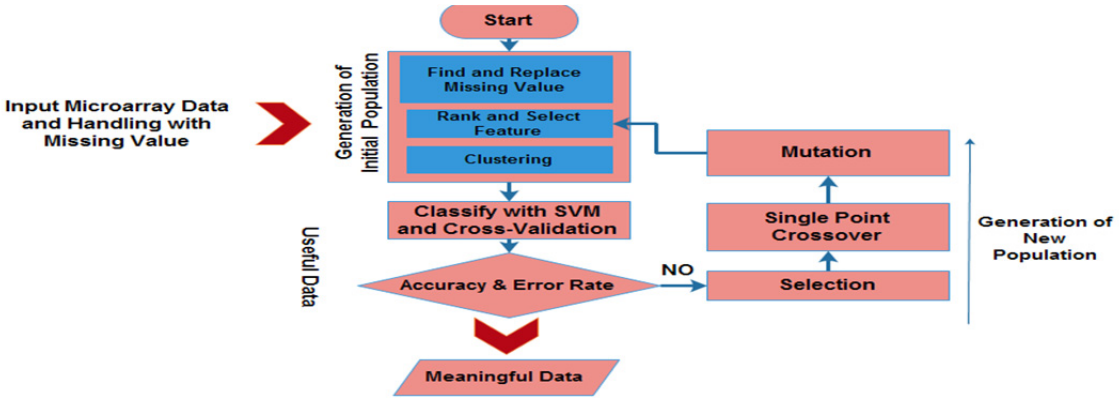

Figure 1. Flow Chart of Analysis and Classify Stage

Figure 1 illustrates the structure of the proposed system. The aim of the first step is to combine GA with other techniques to decrease the classification time and arrive to better accuracy. Then, Gene expression Microarrays dataset are going to be used to find the suitable method to solve the dataset problem with the missing attributes. Based on accuracy, the system can estimate how to enhance Microarray dataset. According to enhancement process, the system can decide if the patient has the lung cancer or not.

Fortunately, by the investigation to finding the perfect method that deal with this missing value, the proposed model suggested to use imputation method, to impute the missing value in lung cancer Gene expression dataset. This method selects the gene with expression similar to the gene that we interest to impute missing value, depending on correlation structure of microarray dataset (Ahmad and Miad, 2013).

However to deal with this missing attributes, the algorithm estimate it depending on its neighbor and using the ADAptive Linear Neuron (Adaline), which use the Least Mean Square (LMS) to complete the dataset. Because classification methods cannot deal with empty values, this missing value will reduce accuracy, and increase the error rate.

2. Literature Review

In 2005, Thanyalnk Jirppech and Stuart Aitken (Jirapech-Umpai and Aitken, 2005) proposed an evaluation method for identifying predictive Gene. They employed the Genetic Algorithm (GA) combination with K-Nearest Neighbor (KNN), and used the Rank Gene method, then applied Leave One-out Cross Validation (LOOCV) for error estimation. This model has a good accuracy on training dataset, but not replicate on testing dataset.

Another approach of classification Microarray dataset published in (Guan et al., 2009), which based on prior knowledge that used to improve the capacity, and could reduce the impact of noisy dataset. The (PAM) Predication Analysis for Microarray dataset started by clustering data and determine nearest centroid for unknown data, to identify the specific Genes. Also in this model used (SVM) Sport Vector Machine as a classification technique.

In (Twala and Phorah, 2010), the authors proposed a novel approaches that used for classification Microarray dataset. They used Decision Tree (DT) to classify Microarray dataset, and then estimate the probabilities by using the Logit model. The Logit model they proposed consists of (BLM) Binary Logit Model, and (MLM) Multinomial Logit Model that deal with three or more classes. This approach use GA and biological information from KEGG database, to achieve a robust subset of feature that has a high prediction capability. KEGG, authors used Microarray dataset as a guide for searching procedure according on ranking methods, and determine a threshold as early stopping criteria. Therefore, they have two procedure steps.

3. Hyper-Active Genetic Algorithm (HAGA)

This section characterizes theproposed algorithm and the technique that were used. Our method aims to classify Microarray dataset and generate the optimal Gene expression to achieve the highest accuracy. The design of HAGA based on the combination between different techniques to rank (score) the available data, then evaluate the accuracy for each group. By clustering the data, we can calculate the average rank for each cluster. The main goal of this combination is to accomplish the best accuracy, as well as minimize processing time.

As shown in Figure 2, the design of the algorithm has several stages. The detailed illustration of each stage listed in the following paragraphs.

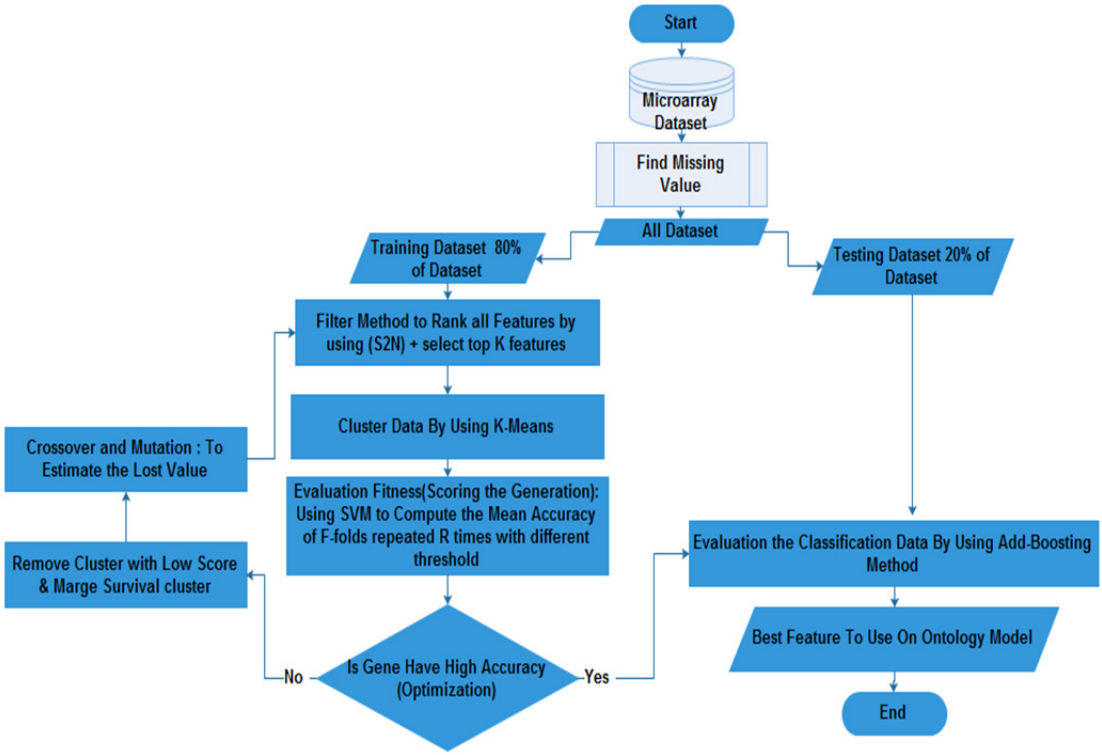

Figure 2. Hyper-Active Genetic Algorithm to diagnose lung cancer by using noisy microarray dataset

3.1 Find Missing Value

In the beginning, the algorithm starts with search operations to determine the missing value and remove it, then it estimates the value depending on its neighbor by using the Adaline technique that depend on LMS to complete the dataset.

$$n = f\left(Wp + b\right) \tag{1}$$

$$W^{new} = W^{old} + \eta e p^{T} \tag{2}$$

$$p^{new} = p^{old} + \eta e \tag{3}$$

Where (n) determine to be the result of multiple input neuron, (W) to be weight and will be initialized randomly, (b) is the bias that initialized to be 0.5, (P) is the selected gene, (f) the transfer function which selected to be Symmetrical Hard Limit, and (η) is learning rate to discover the best solution.

In this step of Adaline, firstly, determine the weight and bias randomly, then using set of iteration to find the optimal weight and bias depending on the Mean Square Error (MSE) as illustrated in equation (4). Secondly, we use the establishment weight and bias to estimate the rest of missing value in lung cancer dataset with minimum error and high accuracy.

$$MES = \frac{1}{n}\sum_{k=1}^{n}\left(t(k) - a(k)\right)^{2} \tag{4}$$

Where (n) is the number of gene, (t) is nearest gene, (a) is imputed gene, and MES is the stopping criteria of this step. After that, partitioning the dataset into two parts, the Training dataset that represent 80% from all dataset and Testing dataset for the rest of the dataset.

3.2 Filter Method

Filter method is type of feature selection, depending on rank all Training feature, and determine the top K features by using the Signal - To - Noise Ratio (SRN) ranking. This method represents the quality, just if there is a biological variability between samples.

Therefore, SRN is used as a measure for comparison:

$$\mu(X,Y)=\frac{|\mu_{+}-\mu_{-}|}{\sigma_{+}+\sigma_{-}} \tag{5}$$

where the data can be divided into two features, Y+ or Y- depending on the output. The (μ_{+}) and (σ_{+}) are the sample means and standard deviation respectively of vector *X* when *Y*=*Y*+, and (μ_{-}) and (σ_{-}) are mean and standard deviation respectively of vector *X* when *Y*=*Y*-.

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2} \tag{6}$$

Where (N) is Number of genes and (x_i) is the selected gene.

3.3 K-Mean Cluster

In this step, the algorithm identifies the correlated Gene cluster. Therefore, it computes the similarity between specific objects and the mean of all other Clusters. We suggest that the Gene clustering, and performing for estimation single Gene, is the main reason to developed grouping of Gene. The clustering step analyzing the Microarray dataset by using K-mean method that needed to determine the centroid for each cluster, and repeat that until no change on the average of centroid, until arrive to optimal cluster of feature.

However, the basic step of this stage is to determine the number of cluster (K), then determine the first objects as initial centroids, and then it determines the centroid coordinate. In addition, it specifies the distance between centroid and each object. In the end, it groups the object depending on minimum distance. Therefore, we can use the distance measurement to compute the cluster memberships, by using equation (7).

$$c(dataset,k)=\sum_{j=1}^{k}\sum_{i=1}^{t}\left\|g_i-c_j\right\|^2 \tag{7}$$

Where (t) is the Gene's number and (k) is cluster number.

3.4 Scoring Measure

We suggest using Support Vector Machine SVM as a linear classifier use in two-dimensional space to be used for two class of training dataset. Using SVM, the f-fold can be determined to compute the performance of Gene, and repeated this procedure R times with different threshold like.

Linear Kernel $$\mathrm{K}(\mathrm{X},\mathrm{Y})=\mathbf{X}^{\mathrm{T}}Y+C \tag{8}$$

Gaussian Kernel $$\mathrm{K}(\mathrm{X},\mathrm{Y})=\exp\frac{\left(-\left\|\mathrm{X}-\mathrm{Y}\right\|^2\right)}{2\sigma^2} \tag{9}$$

The performance of HAGA has been evaluated depending on 10-folds divided into 1-fold for *testing-dataset* and 9-folds for *training-dataset*. Thais process repeated 10 times with different Kernel function to get more efficient and effective classification. The entire previous step represents the first step in Genetic Algorithm (GA), which is the selection feature to select the first generation of our process. However, in this way, the proposed model merges the GA with other techniques to gain the optimal features of data and diagnose the lung cancer effectively.

3.5 Removing Low Cluster

Without finding, the cluster has optimal accuracy. The algorithm generates a new generation from the best cluster, by removing the clusters with minimum costs. This should increase the study accuracy. Then the remaining clusters combined together. The process of replacing iteration will help to complete the missing value and arrives to optimal data.

3.6 Crossover and Mutation

In this stage few members of the remaining data are used to be replaced and generate a new generation in order to reduce the noisy dataset. So we can use the 1-point crossover to exchange the data between parents to make new generation of genes. This process is done by selecting random index position in each gene and swapped the value in each one. After all these operations, the population size should be the same to go back and start the operation more times until reaching the optimal accuracy which will be near the 100%.

3.7 Evaluation the Classifications

After finding the optimal solution, HAGA proceed to the final evaluation stage to evaluate the classification dataset by using Ada-Boosting with the testing-dataset. The reason behind using this technique is to specify the sequence of weak learners on frequently improved versions of the data. All the predictions combined through summing the weights to produce the final prediction. However, the start is with training sample as a weak learner on the original data that came as the result of all previous stages. The weight is modified for the successive iteration, and then the learner algorithm is reweighted the data too. For this process, the weight of all the samples that are incorrectly predicted is increases, while the weights for the correctly predicted samples are decreased (scikit-learn, 2012). The data adaptations at each boosting iteration obtained by applying weights $w_1, w_2, ..., w_N$ to every training samples. Those weights are all set to $w_i = 1/N$ as a weak learner.

4. Experiments and Discussion

4.1 Dataset Samples

Gene expression microarray data on tumor pattern from 39 NSCLC sample, is the first dataset we used. This dataset is divided into two groups, "relapse" of 24 patient, and "non-relapse" for the remaining 15 patients. That classification depends on both clinical and radiological testing. This processed dataset described by 2880 Gene. Our dataset is lung cancer (University of Toronto, Ontario, Canada) (Ontario, 2015). From Figure 3 one can imagine the way we expressed and analyze the data from DNA Microarray.

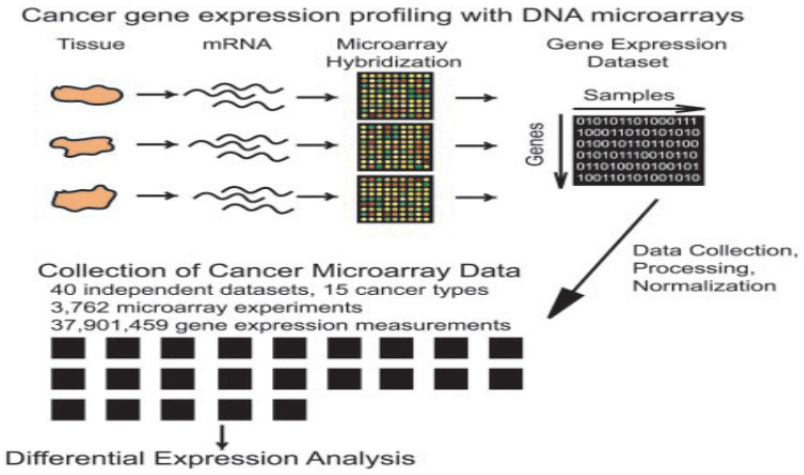

Figure 3. Dataset of Gene Expression Analysis DNA Data (Rhodes et al., 2004)

The second dataset used is Lung Cancer (University of Michigan) (Mechigan, 2015). This dataset has 7129 Genes for each sample, which include 86 primary lung neoplastic samples, and 10 of non-neoplastic lung samples.

The third dataset used is from Harvard Medical School, which consists of two datasets. The first dataset is Lung Cancer (Dana-Farber Cancer Institute, Harvard Medical School) (Harvard, 2015), which have 12600 Genes to describe each sample, and contain 203 Snap-Frozen lung tumor and normal lung sample, that divided for four groups . (1) "ADEN" for 139 samples of adenocarcinomas cancer, (2) "SQUA" represent by 20 samples of pulmonary carcinoids, (3) "SCLC" 6 samples of small-cell lung carcinomas, and (4) "NORMAL" for normal sample. This dataset was published in.

The last dataset is Lung Cancer (Brigham and Women's Hospital, Harvard Medical School) (Brigham, 2015), that use for classification the dataset for two group malignant pleural mesothelioma (MPM) and adenocarcinoma

(ADCA) for the lung dataset. This dataset consist 181 tissue samples (31 MPM and 150 ADCA) divided for 32 training sample (16 MPM and 16 ADCA), and 149 samples are used for testing. 12533 Genes describe each sample.

4.2 Experiments and Discussion

HAGA uses different techniques to improve the classification results. Table 1 lists the parameters that are used in this system to achieve the best accuracy, with low error rate and reduce the time of classification. We obtained the best fitness for all datasets, after some generation of making changes; to find the optimal Gene-Order, and to determine the illnesses for the patient depending on microarray dataset.

Table 1. HAGA Parameters

Dataset	Population Size	K-mean Clustering (k=10)	Generation Number
Michigan (complete)	95 **7130**	Iteration 6	20 selection feature 8 generation
Harvard 1 (30% missing)	203 12600	Iteration 11	186 selection feature 11 generation
Harvard 2 (30% missing)	181 12534	Iteration 12	150 selection feature **10 generation**
Ontario (80% missing)	39 2881	Iteration 2	24 selection feature **21 generation**

This system classifies around 12600 Gene with missing attributes. After dealing with these attributes, it uses feature selection to measure the performance of the best gene in each generation depending on K-mean clustering to improve the accuracy without reducing the performance. The classifying method chosen is SVM since it has the ability to avoid over fitting leading without affecting the accuracy results as we seen in Figure 4. The using of SVM prevents the overlapping between the information from microarray dataset after using the clustering.

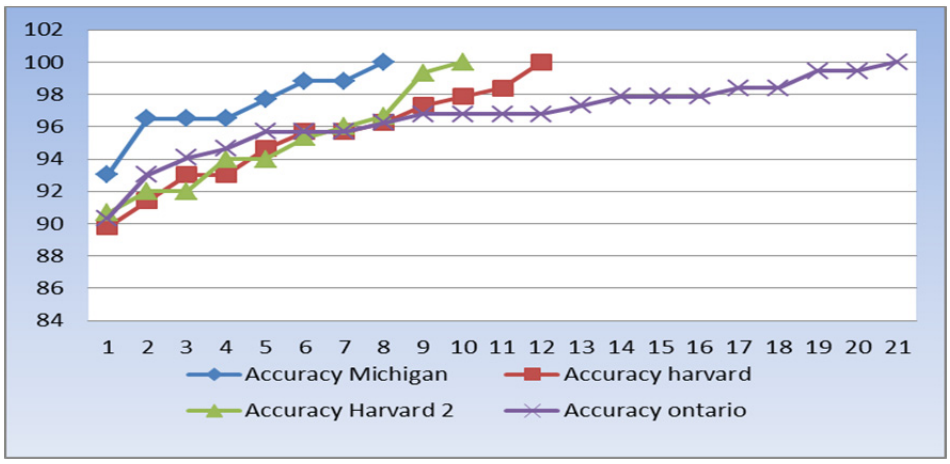

Figure 4. The Accuracy through Generation for all Dataset

As shown in Figure 4, the system achieves an accuracy ranged 90% to 100% for all the used training datasets through all the process generations. Then it reaches the optimal accuracy of 100% among 8 to 21 generations for all.

For evaluation the method and estimate small errors, the system depend the K-fold cross validation with K=5 as

a best result we obtained as shown in Figure 5 for all dataset that the system use and arrive to best error rate '0' after few generations.

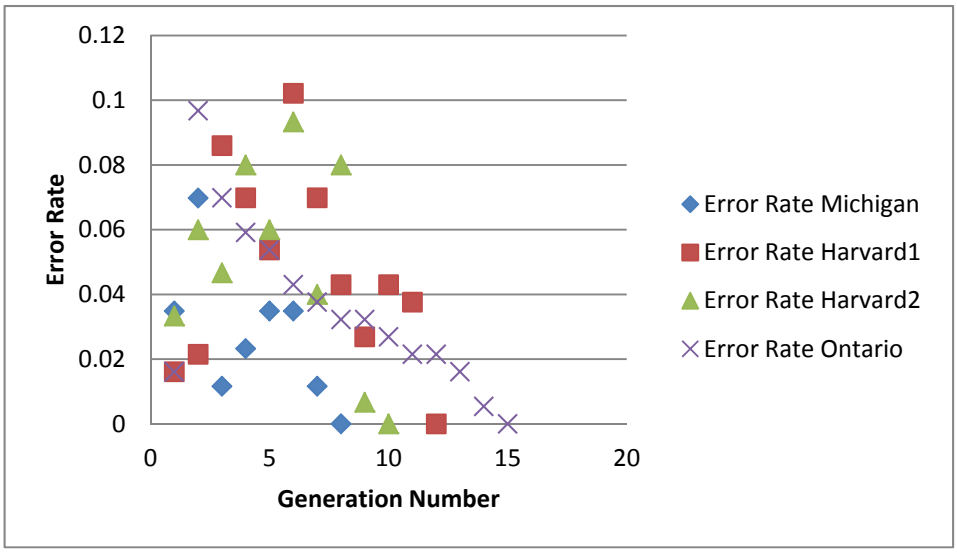

Figure 5. Error Rate through Process Generations

Table 2. Timing for Process System

Dataset	Process Time	Generation Number
Michigan Time	5.85307	8
Harvard Time	33.90013	12
Harvard2 Time	24.41596	10
Ontario Time	40.81986	21

Table 2 reports the time that the system process spends on training dataset to classify it and find the optimal result. The best timing is 5.85307 for Michigan dataset after 8 generations.

Figure 6 show as the result clustering after all process of the selection feature generated in the proposed algorithm.

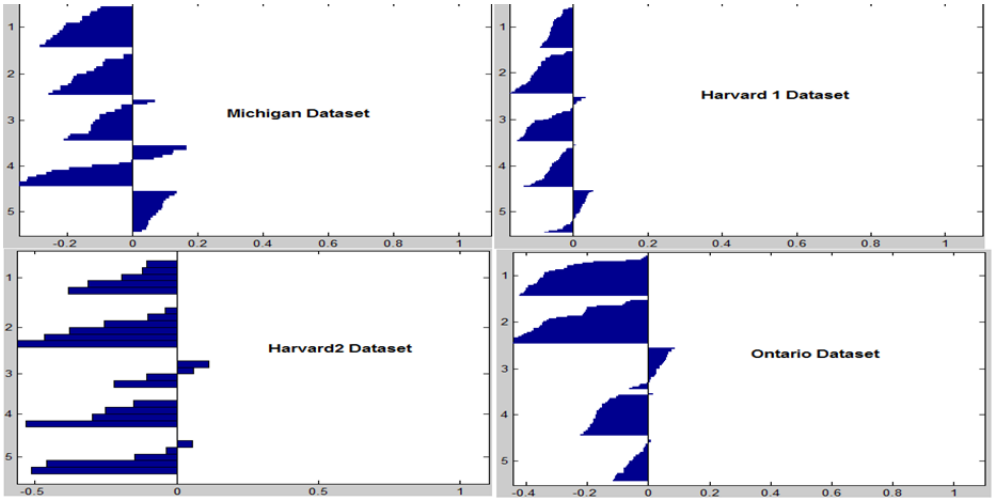

Figure 6. Clustring Ruselt after evaluation for all dataset of the Hyper-Active Genatic Algorithm

Figure 7 illustrates the results of comparing HAGA to four other algorithms. The evaluations depend on training and testing accuracy to determine the best accuracy. HAGA achieved the best accuracy, with 100% for training-dataset and 99.462365% for testing-dataset.

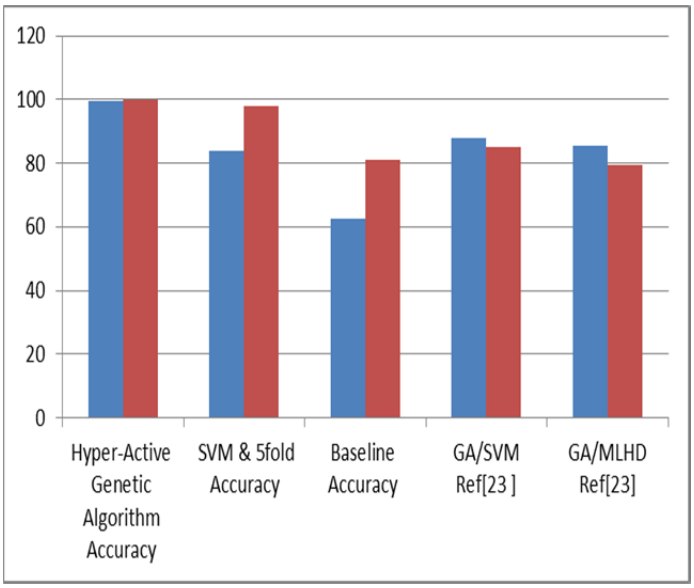

Figure 7. Results of comparing HAGA with four other algorithms

On the other hand, Table 3 listed the error rates of HAGA comparing with other algorithms. Again HAGA locates the best error rate of 0 with processing time equal to 5.853069.

Table 3. Comparing with other algorithm using the same dataset (Michigan dataset)

Method	Error Rate	Time
Hyper-Active Genetic	0	5.853069
SVM &5fold	0.020833	0.044
Baseline	0.07735	0.18750

5. Conclusion

The main contribution of this work is to build a novel genetic algorithm for the classification of incomplete microarray datasets. The design of the algorithm depends on merging several techniques to achieve the best performance comparing with other algorithms. Adaline, SRN, K-Mean, and SVM algorithm were used in the searching for missing attributes, filtering, clustering, and scoring stages. The performance of the algorithm has been analyzed on lung cancer dataset obtained from different biological datasets that are varied from complete sets to different ratio of missing attributes.

The experimental results of the designed algorithm show that the algorithm overwhelms other competitive ones in terms of accuracy and error rates, since it achieve 100% of accuracy and 0 error rate.

Reference

Ahmad, A., & Miad, F. (2013). Augmenting Breath Regulation Using a Mobile Driven Virtual Reality Therapy Framework". *IEEE Journal of Biomedical and Health Informatics, 18*, 746-752. Retrieved from http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html

Guan, P., Huang, D., He, M., & Zhou, B. (2009). Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. *Journal of Experimental & Clinical Cancer Research, 28*, 1-7. Retrieved from

http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard1.html

Jirapech-Umpai, T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics,* 6. Retrieved from http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Michigan.html

Mohammadi, A., & Saraee, M. H. (2008). Dealing with missing values in microarray data. 4th IEEE International Conference on Emerging Technologies ICET, Retrieved from http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Ontario.html

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., & Chinnaiyan, A. M. 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings of the National Academy of Sciences of the United States of America, 9309-9314. Retrieved from http://scikit-learn.org/stable/modules/ensemble.html

Twala, B., & Phorah, M. (2010). Predicting incomplete gene microarray data with the use of supervised learning algorithms. *Pattern Recognition Letters,* 31**,** 2061-2069. Retrieved from http://www.cancerresearchuk.org/cancer-info/cancerstats/world/incidence/