A Hybrid Feature Selection approach of ensemble multiple Filter methods and wrapper method for Improving the Classification Accuracy of Microarray Data Set

Prof. Dr. Ahmed Hassan
Math. Department
Faculty of Science, Zagazig University
Zagazig, Egypt

Prof. Dr. Ahmed Soufi Abou-Taleb
Biomedical Engineering Department
Faculty of Engineering, Cairo University
Cairo, Egypt

Asistant Prof. Osama Abdo Mohamed
Department Of Information Technology
Khulais Faculty of Computer and Information technology
, King Abd-ElAziz University
Khulais, Saudi arabia
Math. Department
Faculty of Science, Zagazig University
Osamaabdomail@yahoo.com

Mr. Amr Hassan
Faculty of Science, Zagazig University
Zagazig, Egypt

***Abstract*— Classification is one of the important tasks in bioinformatics. This can be obviously seen in cancer classification which is recently addressed by many researchers specially after emerging of microarrays. This technology opens the way for computer researchers to classify cancer samples without any previous biological knowledge. Microarrays data is facing a critical challenge because of the nature of high-dimensional and small sample size this problem forces scientists to design gene selection techniques as a preceding step to the implemented classifier. Feature selection is the assignment of selecting a small subset from original features that can perform maximum classification accuracy. This subset of features has some very important benefits such as; it reduces computational complexity of learning algorithms, improves accuracy and save time. In this paper we propose a hybrid feature selection techniques that performed in two phases. In the first phase we use ensemble filter method to ranking all genes and select top genes. This phase helps in improving the classification performance by removing redundant and unimportant features. In the second phase we used SVM-RFE (Recursive Feature Elimination) is a wrapper method. We evaluated our technique on three data sets of gene expression profiles; the Leukemia, the Lung and the Breast cancer; using six classifiers, which are Naïve Bayes (NB), Random Forest (RF), Decision Trees (C4.5), Support Vector Machines (SVM), K-Nearest Neighbor (KNN) and Logistic Regression (LR) and show the potentiality of the proposed method with the advantage of improving the classification performance.**

Keywords-: Feature Selection, Filter, Wrapper, SVM-RFE (Recursive Feature Elimination),SVM (Support Vector Machines), RF (Random Forest), NB (naïve Bayes), C4.5 (decision trees), KNN (k-nearest neighbor) , Microarrays, logistic regression (LR) and AUC (Area under Receiver Operating Characteristic Curve) , Bioinformatics, Classification and Data mining , Gene selection .

I. INTRODUCTION

Progress of information technologies has made the storage and distribution of data much easier in the past two decades. Huge amounts of data have been accumulated at a very fast pace. However, pure data are sometimes not-that-useful and meaningful because what people want is the knowledge/information hidden in the data.

Knowledge/information can be seen as the patterns or characteristics of the data [1]. Data Mining is the automated process involves the use of data analysis tools to discover previously unknown, valid patterns and relationships from large amounts of data stored in databases, data warehouses, or other information repositories using it to make crucial decisions. It is possible to put data-mining activities into predictive data mining, which produces the model of the system described by the given data set, or descriptive data mining, which produces new, nontrivial information based on the available data set [2].

Bioinformatics are data mining application those who specialize in use of computational tools and systems to answer problem in biology .Some of the grand areas of research in bioinformatics are analysis of gene expressions and mutations in cancer.

Cancer databases and gene expression values are the data used in this paper and is extracted from the emerging microarray technology.

High-throughput microarray technology is a hybridization procedure that enabled the simultaneous measurement of the abundance of tens of thousands of gene-expression levels from many different samples on a small chip. Microarray data takes the form of a huge M*N matrix, where M (rows) represents the genes, N (columns) represents the samples and each of its cells contains an expression value for a gene in a sample [3].

Data mining classification techniques aims to distinguish between two types of samples. Usually, these two types are positive, or case samples (i.e., taken from individuals that carry some illness) and negative, or control samples (i.e., healthy individuals). One first obtains a collection of samples with known type labels and uses it to build a classifier, which can later be used to classify unlabeled samples.

The high dimension of the data poses a real problem for standard classifiers Therefore feature selection is a common preprocessing step in many data analysis algorithms where it is used to find the smallest subset of features that maximally increases the performance of the model [4]. Feature selection techniques can be divided into three categories, depending on how they interact with the classifier. Filter methods directly operate on the dataset, and provide a feature weighting, ranking or subset as output. These methods have the advantage of being fast and independent of the classification model, but at the cost of inferior results. Wrapper methods perform a search in the space of feature subsets, guided by the outcome of the model (e.g. classification performance on a cross-validation of the training set). They often report better results than filter methods, but at the price of an increased computational cost [5]. Finally, embedded methods use internal information of the classification model to perform feature selection (e.g. use of the weight vector in support vector machines). They often provide good trade-off between performance and computational cost [6]. In this paper, we focus on this gene selection problem and propose hybrid ensemble approach of filter feature selection techniques and wrapper method for finding a very small subset of genes.

To evaluate and compare the proposed method to other feature selection methods, we used five classifiers – linear Support Vector Machine (SVM, Random Forest, K-nearest-neighbors (KNN), and Naïve Bayes to evaluate the selected features, and to establish the influence on classification accuracy.

The remainder of this paper is organized as follows. Section II, discusses the related work. In Section III describes the general scheme of our proposed approach. Experimental design and evaluation in Section IV. Section V analyzes these results. Finally, Section VI concludes the paper.

II. RELATED WORK

The goal of feature selection is to identify a more informative set of feature that are enhance the classification performance and in the context of microarray data analysis, can be divided into two major groups: filter methods [7] and wrapper methods [8].

the essential differences between the two methods are that a wrapper method makes use of the algorithm that will be used to build the final classifier while a filter method does not, and that a wrapper method uses cross validation to compare the performance of the final classifier and searches for an optimal subset while a filter method uses simple statistics computed from the empirical distribution to select attribute subset [9].

Wrapper methods could provide more accurate classification results than filter methods but would require much more computational costs.

In order to combine the strengths of filter and wrapper approaches while avoiding their drawbacks, some of the recent research works related to hybrid feature selection strategies are as follows:

Pengyi Yang and Zili Zhang [10] proposed hybrid algorithm, called GAEF (Genetic Algorithm with embedded filter), divides the feature selection process into two stages. In the first stage, Genetic Algorithm (GA) is employed to pre-select features while in the second stage a filter selector is used to further identify a small feature subset for accurate sample classification.

Pengyi Yang, Bing Bing Zhou, Zili Zhang, Albert Y. Zomaya [11] proposed multi-filter enhanced genetic ensemble (MF-GE) system is able to improve sample classification accuracy, generate more compact gene subset, and converge to the selection results more quickly.

SALAM SALAMEH, SALWANI ABDULLAH [12] proposed a three-stage selection algorithm by hybridizing the ReliefF, mRMR filter (as filters method) and GA (as a wrapper method) for addressing gene selection problem.

III. METHODOLOGY

System Description:

A flow chart of the proposed hybrid system is illustrated in fig.1. In this system the gene selection Process is sequentially divided into two phase, i.e., “filtering process” and “wrapper process”. In the filtering process we propose an ensemble approach for feature selection, where multiple filter feature selection techniques are combined to yield more robust and stable results. Ensemble of multiple filter feature ranking techniques is performed in two steps. It starts with bootstrap the training samples T=10 times (i.e., draw a sample of size N from the data with replacement T times) to get T rankings of all features by applying the ensemble approach on each sample to form a single feature ranking list as show in fig.2 .

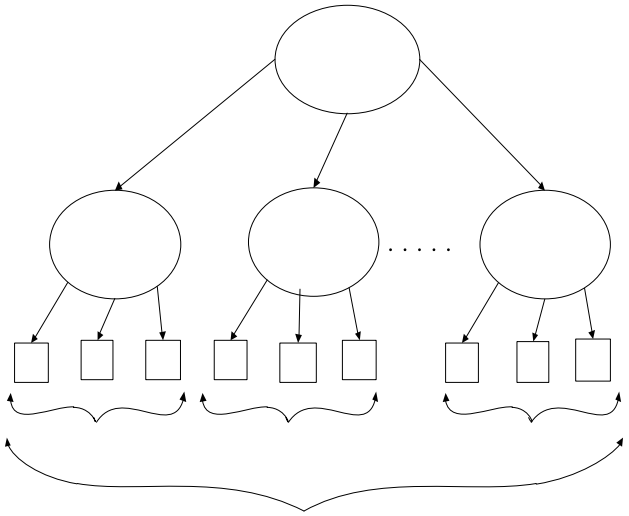

Fig.2 filtering process

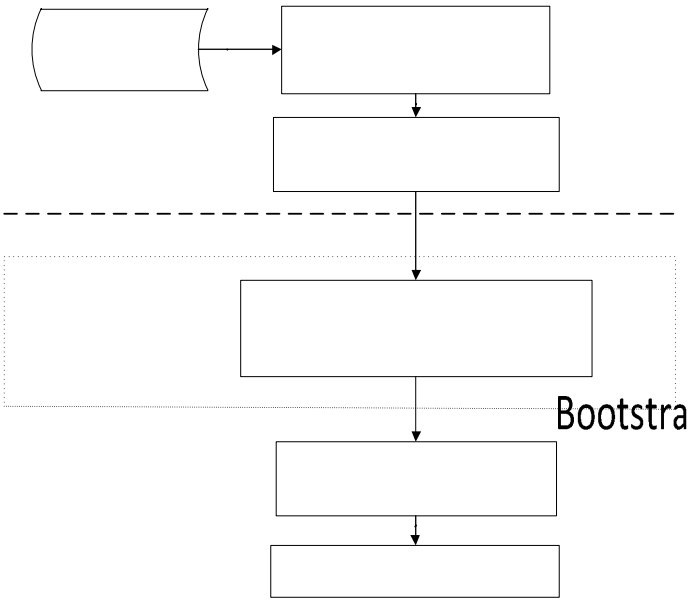

Fig.1 The Block diagram of the proposed method

In the wrapper process, SVM-RFE (Recursive Feature Elimination) is a wrapper method which performs backward feature elimination [13]. The idea is to find the m-features which lead to the largest margin of class separation, and uses the weight vector was a ranking criterion. The recursive elimination procedure of SVM-RFE is implemented as follows:

1. Start: ranked feature set R = []; selected feature subset S = [1,…, d];
2. Repeat until all features are ranked:
   a) Train a linear SVM with features in set S as input variables;
   b) Compute the weight vector;
   c) Compute the ranking scores for features in set

$$S : c_i = (w_i)^2 \quad (1)$$

   d) Find the feature with the smallest ranking

$$Score : e = \arg\min_i (c_i) \quad (2)$$

   e) Update: R= [e, R], S=S - [e];
3. Output: Ranked feature list R.

The algorithm can be generalized to remove more than one feature per step for speed up.

A. *Filter Feature selection methods in experiment*

In this subsection, we introduce three filter based feature ranking techniques in our approach system for experiments.

1) *Information gain:*

Given the entropy is a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases [14]. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (3)$$

IG is a symmetrical measure (refer to equation (3)). The information gained about X after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

2) *Gain Ratio:*

GainRatio incorporates “split information” of features into Information Gain statistic. The “split information” of a gene is obtained by measuring how broadly and uniformly it splits the data [15]. Let's consider again a microarray dataset has a set of classes denoted as ci, (i =1, ..., m), and each gene g has a set of possible values denoted as V .

The discriminative power of a gene g is given as:

$$\text{GainRatio}(g) = \frac{infoGain(g)}{split(g)} \quad (4)$$

in which:

$$split(g) = -\sum_{v \in V} \sum_{i=1}^{m} \frac{|s_v|}{|s|} \log \frac{|s_v|}{|s|} \quad (5)$$

Where S_v is the subset of S of which gene g has value v.

3) *Symmetrical Uncertainty :*

The Symmetrical Uncertainty method evaluates the worth of a gene by measuring the symmetrical uncertainty with respect to the sample class [16]. Each gene is evaluated as follows:

$$\text{symmU}(g) = \frac{2\,x\,((H(class)) - H(class|g))}{H(class) + H(g)} \quad (6)$$

Where H (.) is the information entropy function, H (class) and H(g) give the entropy values of the class and a given gene, while H(class|g) gives the entropy value of a gene with respect to the class.

IV. EXPERIMENTAL DESIGN AND EVALUATION

A. Datasets

We have conducted experiments on three microarray datasets of gene expression profiles. The datasets can be downloaded from http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. We have chosen Leukemia dataset, Breast Cancer dataset and Lung Cancer dataset from the repository. Leukemia dataset we analyzed was the well-known leukemia data studied by Golub et al. [17], which has been explored widely by many researchers. In this dataset, there are 72 observations, each of which is described by the gene expression levels of 7129 genes and a class attribute with two distinct labels: AML vs. ALL. The Lung Cancer dataset contains 186 tumor and 17 normal samples. Breast Cancer dataset this dataset is concerned with the prediction of patient outcome for breast cancer. The training set contains 78 patient samples, 34 of which are from patients who had developed distant metastases within 5 years ("relapse"); the remaining 44 samples are from patients who remained healthy from the disease for an interval of at least 5 years after initial diagnosis ("non-relapse"). There are 12 relapse and 7 non-relapse samples in the test set, and the number of genes is 24,481.

B. Experimental Design

Our experiments were conducted by using the following methods:

1) For each of the three gene ranking methods (Gain Ratio (GR), Information Gain (IG) and Symmetrical Uncertainty (SU)), we used it to select the top-20, top-10 and top-5 ranked genes and all of the top-ranked genes are directly used for classification.
2) Applied the hybrid system we have proposed that implemented using RapidMiner [18] (Version 5.2.008 for Windows).

3) Applied Filter Feature selection (Gain Ratio (GR)) and select top 100 ranked genes then apply SVM-RFE (Recursive Feature Elimination) is a wrapper method and select the top-20, top-10 and top-5 ranked genes to use for classification.

And we used WEKA [19] to measure the performance of each feature selection algorithm. WEKA is a well-known machine learning tool based on JAVA. And we evaluated selected feature subsets using six learning algorithms – naïve Bayes (NB), random forest (RF), decision trees (C4.5), support vector machines (SVM), k-nearest neighbor (KNN) and logistic regression (LR). We evaluated feature subsets using 10-fold Cross-Validation (CV) for all microarray data sets. The classification models are evaluated using the AUC performance metric that is an acronym for Area under Receiver Operating Characteristic Curve [20].

C. Experimental results

We have applied the three feature ranking techniques (GR, IG, SU), Hybrid system proposed and Gain Ratio filter ranker method Followed by SVM-RFE wrapper method to Lung Cancer, leukemia and Breast Cancer dataset. We have selected the top k (k is set to 20, 10, and 5) feature subsets for the experiments. After the feature selection, we used six learners, NB, RF, C4.5, SVM and KNN, to build classification models on the datasets with various selected subset of features. The classification models are evaluated in terms of the AUC performance metric. The results of the experiments are displayed in Table 1, 2, 3, 4, 5, 6, 7, 8 and 9. Each value in the table is determined by the row (ranker) and the column (learner) in which the value is loaded. It also depends on the value of k used for the table. The process of calculating AUC value for a table is performed in three steps:

1. Identify the row and column for which the AUC needs to be calculated. This helps in selecting a ranker and a learner.

2. Ranker is applied to the dataset to get the ranking list. The top k features are selected from the ranking list. The value of k can be determined by checking the table for which the AUC is calculated.

3. Classification model is built using the dataset with selected features from the previous step.

Table 1

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.82	0.81	0.72	0.67	0.69	0.89	0.766667
IG	0.93	0.97	0.90	0.89	0.90	0.92	0.918333
SU	0.95	0.96	0.88	0.80	0.87	0.97	0.905
Hybrid system (Ensemble filter+ (SVM-RFE))	0.99	0.98	0.92	0.93	0.92	0.97	0.951667
GR+ (SVM-RFE)	0.98	0.97	0.92	0.89	0.92	0.96	0.94

AUC values for rankers with top 20 features for Lung Cancer dataset

Table 2

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.90	0.87	0.82	0.68	0.77	0.90	0.823333
IG	0.90	0.95	0.83	0.66	0.87	0.91	0.853333
SU	0.90	0.89	0.83	0.68	0.77	0.90	0.828333
Hybrid system((Ensemble filter+ (SVM-RFE))	0.99	0.98	0.92	0.91	0.93	0.94	0.945
GR+ (SVM-RFE)	0.98	0.97	0.91	0.86	0.93	0.94	0.931667

AUC values for rankers with top 10 features for Lung Cancer dataset

Table 3

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.76	0.72	0.67	0.66	0.63	0.71	0.691667
IG	0.93	0.94	0.85	0.66	0.84	0.93	0.858333
SU	0.90	0.88	0.85	0.69	0.80	0.91	0.838333
Hybrid system (Ensemble filter+ (SVM-RFE))	0.98	0.97	0.93	0.87	0.92	0.95	0.936667
GR+ (SVM-RFE)	0.96	0.96	0.88	0.84	0.89	0.94	0.91166

AUC values for rankers with top 5 features for Lung Cancer dataset

Table 4

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.99	0.99	0.87	0.93	0.93	0.98	0.948333
IG	0.99	1.00	0.87	0.94	0.98	1.00	0.963333
SU	0.99	1.00	0.87	0.96	0.96	1.00	0.963333
Hybrid system (Ensemble filter+ (SVM-RFE))	1.00	1.00	0.90	0.98	0.98	1.00	0.976667
GR+ (SVM-RFE)	1.00	1.00	0.86	0.98	0.98	1.00	0.9700

AUC values for rankers with top 20 features for leukemia Cancer dataset

Table 5

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.99	0.99	0.88	0.92	0.90	0.97	0.941667
IG	0.99	0.99	0.87	0.93	0.94	0.99	0.951667
SU	0.99	0.99	0.88	0.92	0.90	0.97	0.941667
Hybrid system (Ensemble filter+ (SVM-RFE))	1.00	1.00	0.89	0.98	0.98	1.00	0.975
GR+ (SVM-RFE)	1.00	0.99	0.87	0.98	0.98	1.00	0.97

AUC values for rankers with top 10 features for leukemia Cancer dataset

Table 6

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	1.00	0.99	0.89	0.90	0.92	0.95	0.941667
IG	0.98	0.99	0.88	0.92	0.91	0.96	0.94
SU	1.00	0.98	0.89	0.93	0.91	0.96	0.945
Hybrid system (Ensemble filter+ (SVM-RFE))	1.00	0.99	0.93	0.97	0.96	0.97	0.97
GR+ (SVM-RFE)	1.00	0.99	0.89	0.96	0.98	1.00	0.97

AUC values for rankers with top 5 features for leukemia Cancer dataset

Table 7

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.89	0.88	0.73	0.73	0.62	0.80	0.775
IG	0.86	0.88	0.70	0.75	0.76	0.79	0.79
SU	0.85	0.89	0.74	0.75	0.81	0.83	0.811667
Hybrid system (Ensemble filter+ (SVM-RFE))	0.92	0.88	0.77	0.86	0.79	0.88	0.85
GR+ (SVM-RFE)	0.94	0.90	0.76	0.88	0.69	0.88	0.8412

AUC values for rankers with top 20 features for Breast Cancer dataset

Table 8

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.87	0.87	0.76	0.74	0.67	0.84	0.791667
IG	0.81	0.85	0.74	0.74	0.65	0.80	0.765
SU	0.84	0.88	0.76	0.75	0.75	0.82	0.8
Hybrid system (Ensemble filter+ (SVM-RFE))	0.92	0.88	0.80	0.81	0.75	0.90	0.843333
GR+ (SVM-RFE)	0.92	0.87	0.64	0.86	0.77	0.93	0.831666

AUC values for rankers with top 10 features for Breast Cancer dataset

Table 9

Ranker	NB	RF	C4.5	SVM	KNN	LR	Average
GR	0.82	0.78	0.73	0.73	0.62	0.84	0.753333
IG	0.72	0.86	0.73	0.70	0.70	0.69	0.733333
SU	0.79	0.86	0.80	0.68	0.77	0.75	0.775
Hybrid system (Ensemble filter+ (SVM-RFE))	0.93	0.91	0.83	0.83	0.81	0.92	0.871667
GR+ (SVM-RFE)	0.88	0.83	0.72	0.82	0.74	0.89	0.81333

AUC values for rankers with top 5 features for Breast Cancer dataset

V. ANALYSIS OF RESULTS

In this study, we compared the classification accuracy of the selected gene sets from Three feature ranking techniques (GR, IG, SU) individually, Hybrid system proposed and Gain Ratio filter ranker method Followed by SVM-RFE wrapper method . Instead of trying to acquire the highest classification performance, we focus on differentiating the classification power of different gene selection algorithms. The ranking and classification of each dataset are repeated 6 times and each time the top 20, 10 and 5 genes are used for sample classification. We report the average of the results of the classification.

The evaluation results obtained using three microarray dataset Lung Cancer, leukemia Cancer and Breast Cancer for

top 20,10 and 5 ranked genes are depicted in Table 1, 2, 3, 4, 5, 6, 7, 8 and 9 respectively. It is easy to see that the Hybrid system proposed has higher average AUC classification performance for all datasets and the second best in average AUC classification performance over all datasets show in GR/SVM-RFE Gain Ratio filter ranker method Followed by SVM-RFE wrapper method. Given the fact that the SVM-RFE part of these two algorithms are the same, the natural explanation of the improvement is attributed to the fusion of ensemble multiple filter selection techniques.

VI. CONCLUSION

In this paper, we proposed hybrid system consist of two phases the first phase we use ensemble multiple filtering algorithms and the second phase we used wrapper method .this system used for select a small number of informative genes from gene expression data which increases the classification accuracy, and then used Naïve Bayes (NB), Random Forest (RF), Decision Trees (C4.5), Support Vector Machines (SVM), K-Nearest Neighbor (KNN) and Logistic Regression (LR) to evaluate the classification performance by using three microarray gene expression datasets namely the Leukemia Cancer, Lung cancer and breast Cancer datasets. Experimental results showed that the proposed method simplified gene selection and the total number of parameters needed effectively, thereby obtaining a higher classification performance compared to other feature selection methods. The classification accuracy obtained by the proposed method was comparatively higher than other methods. In the future, the proposed method can assist in further research where Feature selection needs to be implemented. It can potentially be applied to problems in other areas as well.

.

REFERENCES

[1] Hui-Huang , editors." Advanced data mining technologies in bioinformatics". Idea Group Inc., 2006.

[2] Mehmed ,Kantardzic " Data Mining: concepts , Models, Methods and algorithms" Wiley IEEE Aug.2011 .

[3] C.S. Kong,J. Yu, F.C. Minion, K. Rajan, "Identification of Biologically Significant Genes from Combinatorial Microarray Data," ACS Combinatorial Science, 2011.

[4] Ivan Kojadinovic, Thomas Wottka, I Remia, and UniversitÃl' De RÃl'u-nion. Comparison between a filter and a wrapper approach to variable subset selection in regression problems. Moon, pages 14–15, 2000.

[5] Kohavi , R., John, G.: Wrappers for feature subset selection. Artif. Intell. 97(12),273–324 (1997)

[6] Saeys, Y., Inza, I., Larra˜ naga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)

[7] R. Kohavi and G. H. John. Wrappers for feature subset selection.Artificial Intelligence, 97(1-2):273–324, 1997.

[8] P . Langley. Selection of relevant features in machine learning. In AAAIFall Symposium on Relevance, pages 140–144, 1994.

[9] Yi Zhang, Chris H. Q. Ding, Tao Li: A Two-Stage Gene Selection Algorithm by Combining ReliefF and mRMR. BIBE 2007: 164-171.

[10] Pengyi Yang, Zili Zhang: An Embedded Two-Layer Feature Selection Approach for Microarray Data Analysis. IEEE Intelligent Informatics Bulletin 10(1): 24-32 (2009)

[14] Hall, M.A., and Smith, L.A., "Practical feature subset selection for machine learning", Proceedings of the 21st Australian Computer Science Conference, 1998, 181–191.

[15] Mitchell T:Machine Learning McGraw Hill; 1997.

[16] Witten IH and Frank MD :Data Mining: Practical Machine Learning Tools and Techniques Elsevier; Second2005.

[17] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, and Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286(5439):531-537.

[18] http://rapid-i.com/content/view/396 .

[19] WEKA, http://www.cs.waikato.ac.nz/ml/index.html.

[20] T. Fawcett, "ROC Graphs: Notes and Practical Consideration for Researchers", HP Laboratories, March 16, 2004, Kluwer Academic Publishers, pages 1 -38.

AUTHORS PROFILE

Dr. Osama Abdo Mohamed is assistant professor in King abd El-Aziz University. He has got Msc. Degree in computer science in. 1998 & PH.D. degree in computer science in 2007. He has more than 17 years teaching experience and more than 17 years research experience in the field of signal processing and logic programming. He has published more than 10 national and international research papers in various refereed journals

Mr. Amr Hassan is Information systems specialist at a health insurance organization. He has a BSc in Mathematics and Computer Science in 2006, pre-MSc titles in computer science from Zagazig University, Zagazig, Egypt in 2010 .