Leukemia Cancer Comparative Classifires Suite

Ahmed Abd El-Nasser,ahmed_a_nasser@hotmail.com, Modern Academy in Maadi
Mohamed Shaheen, cshaheen@hotmail.com, Arab Academy for Science, technology, and Maritime Transport
Hesham El-Deeb,hmeldeeb14@yahoo.com, Faculty of Computer Science, M.T.I University

Abstract

A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis of certain important information. For the ultimate diagnosis, normally, many tests generally involve the clustering or classification

Microarray data classification is used primarily to predict unseen data using a model built on categorized existing Microarray data.

The applications of microarray technology are able to utilize information and knowledge from human genome project to benefit human health. In the last few years, the remarkable progress achieved in microarray technology domain has helped researchers to develop the optimized treatment of cancer.

Human acute leukemia is used as test case to a generic approach to cancer classification, this classification approach is based on gene expression monitoring by DNA microarrays that distinct between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

The objective of this research is to investigate and compare the accuracy, time to build model, and errors of classification process using Locally Weighted Learning (LWL) algorithm with nine different classifiers (Bayes Network learning, Conjunctive Rule, NBTree, Voting Frequency Intervals (VFI), Random SubSpace, Naïve Bayes Updateable, DIMM, Kstar, and PART); to previous tested datasets after performing some preprocessing to the datasets to enhance the classification process.

The proposed approach and experiments showed that after conducting the preprocessing and the classification using Voting Frequency Intervals, Random Sub Space and Naïve Bayes Updateable algorithms through LWL approach it can be reached in 0.1 s time and accuracy of 94% which is outperform the other previous techniques for the same data when comparing with previous published studies.

Keywords Bioinformatics, Classification, Data Mining, DNA, Leukemia, LWL.

I. INTRODUCTION

Leukemia is a type of cancer of the blood or bone marrow characterized by an abnormal increase of immature white blood cells called "blasts." Leukemia is a broad term covering a spectrum of diseases. According to American Cancer Society (ACS) it is estimated that 48,610 persons (27,880 men and 20,730 women) will be diagnosed with and 23,720 men and women will die of leukemia in 2013 only.

In turn, it is part of the even broader group of diseases affecting the blood, bone marrow, and lymphoid system, which are all known as hematological neoplasms. Acute Lymphoblastic Leukemia (ALL) is the most common type of leukemia in young children and Acute Myelogenous Leukemia (AML) occurs more commonly in adults than in children, and more commonly in men than women.

Gene expression, also known as protein expression, is the process by which gene's coded information is converted into a final gene product. Gene expression data can help in better understanding of cancer.

To get the expression level data efficiently Microarray technology was invented to simultaneously monitor a large number of genes from biological samples. The data generated by microarray can be viewed as a two-dimensional array. Each row of the array represents a gene; each column represents a biological sample tested on the microarray.

Classification problem has been extensively studied by researchers in the area of data mining and machine learning. Many classification algorithms have been proposed in the past, such as the decision tree methods, the linear discrimination analysis, the Bayesian network, etc. For the last few years, researchers have started paying attention to the cancer classification using gene expression [1, 2].

The goal of this research is to build a classifier from categorized historical Microarray gene expression data, and then to use the classifier to categorize future in-coming data or predict the future trend of data. It has been reported that the results of microarray experiments can be nearly 100% accurate [3, 4].

There is a substantial amount of research with machine learning algorithm such as Bayes Network, Radial Basis Function, Decision tree and pruning, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm.

Xiaosheng Wanget al. in [5] reached highest accuracy in Leukemia data set with 97.22% using NB tree and C4.5 Algorithms and 92.013% with SVM and KNN. Peter J. Tan et al. [6] reached highest accuracy in Leukemia dataset with 94.3% with C4.5 and 95.7% with Ad C5.0 algorithm. Hong Hu. et al. [7] reached highest accuracy in Leukemia dataset with 79.2% with C4.5 and 86.1% with Ad C4.5 algorithm. Aik Choon Tan et al. [8] reached highest accuracy in Leukemia dataset with 91.18% using C4.5 and Ad. C4.5.

Locally Weighted Learning (LWL) technique uses an instance-based algorithm to assign instance weights which are then used by a specified weighted instances handler and then do a classification (e.g. using naive Bayes) or regression (e.g. using linear regression). For more info, see [9].

II. RELATED WORK

A lot of work was done In last 20 years in the domain of DNA microarray data classification and cancer disease classification, and several decision tree algorithms have been applied on microarray data, some of these algorithm presented acceptance performance results, but the other fail, as well as, some method present significant accuracy on several microarray dataset but there resulted insignificant accuracy on other datasets. Some related works in this area are mentioned in this section. But first some background about used algorithms will briefly described.

Bayesian networks are a powerful probabilistic representation, and their use for classification has received considers- able attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C [10].

Single conjunctive rule learner is one of the machine learning algorithms and is normally known as inductive learning. The goal of rule induction is generally to induce a set of rules from data that captures all generalizable knowledge within that data, and at the same time being as small as possible [13].

Nearest neighbors algorithm is considered as statistical learning algorithms and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbor does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest- neighbor classifier finds the closest training-point to the unknown point and predicts the category of that training- point accordingly to some distance metric [10]. The distance metric used in nearest neighbor methods for numerical attributes can be simple Euclidean distance.

The NBTree algorithm is a hybrid between decision-tree classifiers and Naive Bayes classifiers. It represents the learned knowledge in the form of a tree which is constructed recursively. However, the leaf nodes are Naive Bayes categorizers rather than no dis-predicting a single class [10]. For continuous attributes, a threshold is chosen so as to limit the entropy measure. The utility of a node is evaluated by discretizing the data and computing the fivefold cross-validation accuracy estimation using Naive Bayes at the node. The utility of the split is the weighted sum of utility of the nodes and this depends on the number of instances that go through that node. The NBTree algorithm tries to approximate whether the generalization accuracy of Naive Bayes at each leaf is higher than a single Naive Bayes classifier at the current node. A split is said to be significant if the relative reduction in error is greater than 5% and there are at least30 instances in the node [10]

The VFI algorithm is a classification algorithm based on the voting frequency intervals. In VFI, each training instance is represented as a vector of features along with a label that represents the class of the instance. Feature intervals are then constructed for each feature. An interval represents a set of values for a given feature where the same subset of class values is observed. Thus, two adjacent intervals represent different classes. A detailed explanation of both the above algorithms can be found in [11]

In the random subspace method, classifiers are constructed in random subspaces of the data feature space. These classifiers The Random Subspace Method (RSM) is the combining technique proposed by Ho [6]. In the RSM, it modifies the training data. However, this modification is performed in the feature space are usually combined by simple majority voting in the final decision rule. [12]

In Naive Bayes Updatable classifier given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, our aim is to construct a rule which will allow us to assign future objects to a class, given only the vectors of variables describing the future objects. Problems of this kind, called problems of supervised classification, are ubiquitous, and many methods for constructing such rules have been developed. One very important one is the naive Bayes method—also called idiot's Bayes, simple Bayes, and independence Bayes. This method is important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. [13]

PART is a Class for generating a PART decision list which uses separate-and-conquer. It builds a partial C4.5 decision tree for each step, and then makes the "best" leaf into a rule. This algorithm presents a rule-induction procedure that avoids global optimization but nevertheless produces for more information, see: Elbe Frank. [14]

DIMM which Re-implement the Diverse Density algorithm, changes the testing procedure.

Kstar(K*) is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

There is a lot of work done in this area here some examples which related to this work as a base for later discussion of results.

E. A. Manilich et al. [15] proposed a new framework, for optimized implementation of a random Forests classifier. The authors focus on the memory consuming and computational complexity, and they show acceptable computing performance while preserving predictive accuracy.

Ng Ee Ling et al. [16] compared between several classifiers including decision tree (C4.5), on three cancer datasets, there

result shows that there is no one classifier that works best for all datasets. They obtained classification accuracy between 79.13 and 86.80 for default (C4.5).

The weighted voting method is proposed by Golub and Slonim et al. [1, 17] for classifying binary class data. The GS method is a correlation based classifier. The assignment of classes is based on the weighted voting of the expression values of a group of "informative genes" in the test tuple.

In [18] they used the Naive Bayes algorithm for gene classification. In applying Naive Bayes method to gene classification, the method models each class as a set of Gaussian distributions: one for each gene from the training samples.

Neural networks is used in [19] for cancer type prediction. The method consists of three major steps: principle component analysis, relevant gene selection and artificial neural network prediction.

Decision tree, also known as classification trees, is a well know classification method [20]. It has been widely used in classification applications and many extensions/variations.

III. THE COMPARATIVE STUDY

Global View

The proposed approach in this research is to test some classification algorithms through getting datasets and choose a model to build, train this model, and then test and evaluate the output from the classification process from each algorithm.

The phases of the proposed approach can be categorized as shown in figure 1; the proposed approach is to get the microarray dataset and get the gene expression for the dataset. As an enhancement step a simple pre-processing step is performed. From each gene expression value, its mean is subtracted. Then a classifier is chosen to be used and the evaluation process to get the classification results as shown later in the experimental results section.

Figure 1: Experiment Flow Milestones

Leukemia Datasets

Three data sets are presented; each dataset consists of a matrix of gene expression vectors obtained from DNA micro-arrays [22] for a number of patients. The datasets were obtained from cancer patients with two different types of leukemia(ALL,AML).The three datasets contains 7130 Genes, The first dataset is 72 Sample (47 ALL, 25 AML), the second dataset is 38 Sample (27 ALL ,11 AML). The third and last dataset is 34 Sample (20 ALL, 14 AML) was obtained from Affymetrix oligonucleotide microarrays. Table 1 summarizes the three datasets information.

TABLE 1: LEUKEMIA DATASETS

Dataset	Sample NO.	Genes No.	Categories	
			ALL	AML
Data Set 1	72	7130	47	25
Data Set 2	38	7130	27	11
Data Set 3	34	7130	20	14

This research will use Locally Weighted Learning (LWL) approach with nine different classification algorithms (Bayes Network learning, Conjunctive Rule, NBTree, VFI, Random Subspace, Naïve Bayes Updateable, Kstar, DIMM, and PART)

Performance Measures

For each algorithm as a performance measure procedure; computation of the accuracy, time elapsed to build the model, and error statistics for each dataset classification process is performed. [16]

Classification accuracy =number of correct classified instances /total number of instances.

True positives = TP; False positives = FP

True Negatives = TN; False negatives = FN

Recall = TP / (TP + FN) // true positives / actually positive

Precision = TP / (TP + FP) // true positives / predicted positive

F-measure = 2TP / (2TP + FP + FN)

ROC curve is used to evaluate the discriminative performance of binary classifiers. This is obtained by plotting the curve of the true positive rate (Sensitivity) versus the false positive rate for a binary classifier by varying the discrimination threshold.

Root Mean-Squared Error= Square root of (Sum of Squares of Errors / number of predictions)

Mean Absolute Error= (Sum of Absolute Values of Errors / number of predictions)

Root Relative Squared Error=

Square root of (Sum of Squares of Errors / Sum of Squares of differences from mean)

Relative Absolute Error=Sum of Absolute Values of Errors / Sum of Absolute Values of differences from mean)

IV. IMPLEMENTATION

To measure and investigate the performance on the selected classification algorithms mentioned before, the same experiment procedure as proposed in the proposed approach section is used.

Two-third of the data set is used for training and the remaining is for testing purposes.

Sample of the results report generated can be seen in figure 2.

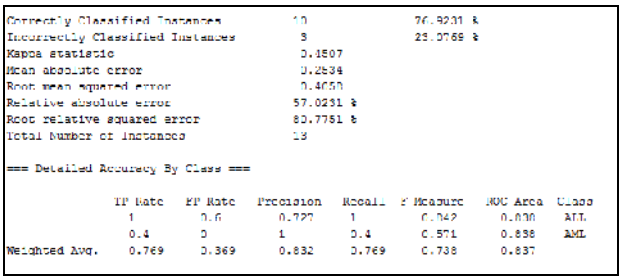

Figure 2: Sample Output Results

V. EXPERIMENTAL RESULTS

Based on both sections proposed approach and Performance measurements table 2 contains horizontally the used algorithms(,2,3,4,5,6,7 which are Bayes Network learning, Conjunctive Rule, NBTree, VFI, Random Sub Space, Naïve Bayes Updateable, and PART) and vertically the accuracy ,time to build the model, MAE, RMSE, RAE, RRSE, Precision ,Recall , and ROC.

TABLE2: SIMULATION RESULTS OF EACH ALGORITHM ON ALL DATASETS

Used Classifier	Data set	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Accuracy	#1	91.66	87.5	91.66	54.16	95.83	94	91.66
	#2	84.61	76.92	84.61	76.92	84.61	84.61	84.61
	#3	91.66	75	91.66	94	94	91.66	91.66
Time	#1	0.03	0.02	0.01	0	0	0.1	0.01
	#2	0.03	0.02	0.01	0	0	0.1	0.01
	#3	0.03	0.02	0.01	0	0	0.1	0.01
MAE	#1	0.083	0.136	0.083	0.459	0.124	0	0.095
	#2	0.153	0.253	0.153	0.440	0.305	0.1538	0.153
	#3	0.083	0.330	0.083	0.263	0.263	0.083	0.083
RMSE	#1	0.288	0.358	0.288	0.481	0.209	0	0.275
	#2	0.392	0.405	0.392	0.447	0.376	0.392	0.392
	#3	0.288	0.525	0.288	0.300	0.300	0.288	0.288
RAE	#1	18.48	30.423	18.486	102.1	27.68	0	21.18
	#2	34.61	57.023	34.60	99.12	68.76	34.61	34.61
	#3	16.66	66.173	16.66	52.78	52.78	16.66	16.66
RRSE	#1	62.88	78.194	62.88	105.0	45.74	0	60.11
	#2	78.07	80.775	78.04	88.99	74.96	78.07	78.07
	#3	56.01	101.97	56.01	58.31	58.31	56.01	56.01
Precision	#1	0.938	0.884	0.938	0.838	0.964	0.9	0.938
	#2	0.877	0.832	0.877	0.784	0.877	0.877	0.846
	#3	0.929	0.833	0.929	0.9	0.96	0.929	0.929
Recall	#1	0.917	0.875	0.917	0.542	0.958	0.92	0.917
	#2	0.846	0.769	0.846	0.769	0.846	0.84	0.846
	#3	0.917	0.75	0.917	0.96	0.96	0.917	0.917
ROC	#1	0.993	0.865	0.993	0.95	0.97	0.95	0.96
	#2	0.894	0.837	0.887	0.9	0.925	0.8	0.837
	#3	0.99	0.646	0.94	0.94	0.96	0.917	0.917

Table 2 also summarizes the result based on correctly classified instances, time to build the model, error rates which discussed with equation in the previous section. Mean absolute error, root mean squared error, relative absolute error, root relative squared error, and Total Number of Instances for each simulation for the three datasets.

The used classifies in

According to classification process while implementing the two algorithms Re-implement the Diverse Density algorithm (MIDD) and Kstar algorithms because these classifiers are not creating a weighted instance handler to pass to the LWL algorithm.

Table 2 shows that the best classifying for data set 1 with algorithm is Naïve Bayes Updateable which reached94% accuracy in 0.14s, the Bayes Network learning NBTree, and PART with the same accuracy 91.67 and error rates0.0832, 0.0832, 0.0953 respectively.

With dataset 2 the algorithms NBTree, Random Sub Space, Naïve Bayes Updateable, and PART are the same accuracy 84.615% with times 0.01s, 0s, 0.14s, 0.01s respectively and the error rates 0.1538, 0.3056, 0.1538, and 0.1538 respectively.

And finally for dataset 3 Random Sub Space and VFI Algorithms are the height accuracy with 100% accuracy with no times.

As appeared in table 2 also that the time to build the model is the same for each used algorithm because the time needed to build the model not related to the number of genes or cases.

Figure2, figure 3, and figure 4 shows the simulation errors results percentage value and subsequently, mean absolute error and root mean squared error will be in numeric value only. The relative absolute error and root relative squared error in percentage for references and evaluation also shown.

The 94% accuracy achieved using Naïve Bayes Updateable for dataset1 and accuracy 84.6 % with NBTree for dataset 2.

Finally an accuracy of 94% reached for both algorithms VFI and Random Sub Space. Which are better results for the same datasets compared with Xiaosheng Wanget al. [5], Peter J. Tan et al. [6], Hong Hu. et al. [7], and Aik Choon Tan et al. [8].

As a type of comparison for previous results for the same dataset and algorithms; the obtained results are better in some

cases; Hong Hu. has reached 79.2% in 2006,Peter J. Tan reached 94.3%in 2007, and Xiaosheng Wanget in 2010 reached 92.0.13 while this work reached 94% using Naïve Bayes Updateable, VFI and Random Sub Space classifiers.

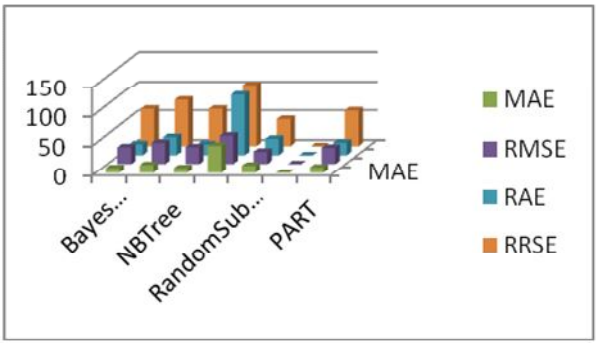

Figure 2 :simulation errors results for dataset 1 for used algorithms

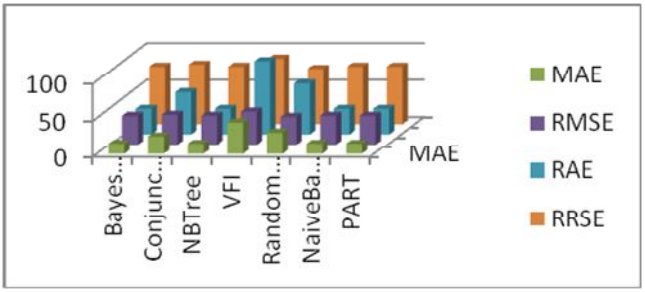

Figure 3 :simulation errors results for dataset 2 for used algorithms

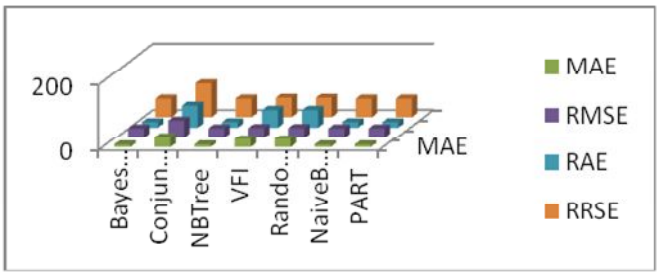

Figure 4 :simulation errors results for dataset 3 for used algorithms

VI. CONCLUSION

This research is conducted to highlight on the concept of selection the appropriate algorithm for a certain dedicated dataset to acquire a high precision in diagnose of the Leukemia disease.

In the same context, the investigation showed that each type of cancer has its own appropriate classifying algorithms that get the best results, so it's important to choose the best algorithm with each cancer type. Also it's important to choose the dataset carefully before testing process and it's not applicable to use the same datasets with more than one cancer type to get the best results.

So, the experimental study compares classification performance of different nine classifier algorithms via three cancerous microarray datasets.

The experimental results show that the Naïve Bayes Updateable algorithm has the higher accuracy when used with dataset 1while VFI and Random Sub Space classifiers has 84.6 % as accuracy measure when used with dataset 3.

The future work could be directed to apply the same approach to other types of cancer diseases and enhance results through modifying the preprocessing process of algorithms.

REFERENCES

[1] *Ying Lu and Jiawei Han, "Cancer Classification Using Gene Expression Data," Information Systems, Data Management in Bioinformatics, pp.243–268 2003, vol. 28.*

[2] *A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. "Tissue classification with gene expression profiles," In Proc. of the Fourth Annual Int. Conf. on Computational Molecular Biology, 2000.*

[3] *Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. "Use of proteomic patterns in serum to identify ovarian cancer,". Lancet 2002, pp. 359-572, vol.7.*

[4] *Wei Zhu, Xuena Wang, Yeming Ma, ManlongRao, James Glimm, and John S. Kovach."Detection of cancer-specific markers amid massive mass spectral data, "PNAS 100 pp. 14666-14671, 2003, vol. 25.*

[5] *Xiaosheng Wang and Osamu gotoh,"A Robust Gene selection Method for Microarray-based cancer Classification," Cancer Informatics, pp. 15–30, 2010.*

[6] *J. Tan Peter, L. Dowe David and I. Dix Trevor, "Building classification models from microarray data with tree-based classification algorithms", Clayton School of Information Technology, Monash University, Melbourne, Australia 2007.*

[7] *Hong Hu, Jiuyong Li, Hua Wang, Grant Daggard and Mingren Shi "A Maximally Diversified Multiple Decision Tree Algorithm for Microarray Data Classification", Conferences in Research and Practice in Information Technology (CRPIT), Vol. 73, 2006.*

[8] *Tan AikChoon and Gilbert David, "Ensemble machine learning on gene expression data for cancer Classification", Open Mind Journals Limited 2003.*

[9] *Eibe Frank, Mark Hall, Bernhard Pfahringer," Locally Weighted Naive Bayes," In: 19th Conference in Uncertainty in Artificial Intelligence, pp. 249-256, 2003.*

[10] *Ben-Gal I., Bayesian Networks, in Ruggeri F., Faltin F. &Kenett R., "Bayesian Networks", Encyclopaedia of Statistics in Quality & Reliability, Wiley & Sons 2007*

[11] *S. V. Sabnani. "Computer security: A machine learning approach". Master's thesis, Royal Holloway, University of London,2007*

[12] *Sotiris Kotsiantis," Combining bagging, boosting, rotation forest and random subspace methods," Artificial Intelligence Review Springer US , Volume 35, Issue 3 , pp 223-240,2011*

[13] *iZheng,Geoffrey I. Webb ,P.Suraweera,,L. Zhu"Subsumption resolution: an efficient and effective technique for semi-naive Bayesian learning," Machine Learning , Springer US, 2012, Vol. 87, Issue 1, pp 93-125.*

[14] *Mazid, M.M; Ali, A.B.M.S, Tickle, K.S., "A comparison between Rule based and Association Rule Mining Algorithms" Third International Conference on Network Security 2009, NSS'09.*

[15] *E. A. Manilich , Z. M. Ozsoyoglu, V. Trubachev, T. Radivoyevitch, "classification of Large Microarray Data Sets using Fast Random Forest Construction", J. Bioinformatics and Computational Biology, 2011,vol. 9, no. 2, pp. 82-91.*

[16] *Ng Ee Ling and Yahya Abu Hasan, "Classification on microarray Data", 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications UniversitSains Malaysia, Penang, 2006,pp. 13-21.*

[17] *D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander. Class prediction and discovery using gene expression data. In Proc. 4th Int. Conf. on Computational Molecular Biology (RECOMB) 2000, pp. 263–272.*