Integrated Intelligent Research (IIR)
International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Pages: 7-10
ISSN: 2278-2419

Classification and Prediction of Disease Classes using Gene Microarray Data

Sujata Joshi, A Deeptha, K Prathibha, N Hema, J Priyanka
Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore
Email : sujata_msrp@yahoo.com, adeeptha1993@gmail.com, hema612@gmail.com, prathibharcr@gmail.com, priyapriyanka13891@gmail.com

Abstract - In the year 1999, when T. R Golub first presented an idea for classifying cancer at the molecular level, this boosted research in cancer diagnosis to a whole new level. The researchers began to analyze the disease at the genetic level with the help of microarray databases. Then there were many new algorithms designed by researchers to classify different types of cancer. The objective of this paper is to present a tool designed exclusively to predict and classify leukemia into its types. The leukemia dataset published by Golub is used for this purpose. The first step is to identify the most significant genes causing cancer from the training set. These selected genes then are used to build the classifier based on decision rules, and eventually to predict the type of leukamia. This classifier which is modeled based on decision rules is found to work with an accuracy of 94%. The algorithm is quite simple in terms of complexity. It is possible to use a minimum number of genes for classification purposes rather than using a large set of genes. The genes that are responsible for prognosis of cancer are mainly selected for designing the classifier.

Keywords- Decision Tree, Leukemia, Microarray, Disease

I. INTRODUCTION

Cancer is considered to be one of the most deadly human diseases. Conventionally morphological diagnosis of tumor is not always effective as revealed by frequent occurrences of misdiagnoses[3]. Recent biological studies carried out at the molecular level have revealed that cancer is a disease that involves dynamic changes in the genome. Abundant explorations and experimentations have been conducted to carry out cancer diagnosis and prediction based on gene microarray data since the pioneering work of Golub et al.[1][3].Leukemia is a type of cancer caused by high numbers of abnormal white blood cells. There are many types of leukemia based on whether they are acute or chronic in nature and the kind of cell they affect (myelogenous or lymphocytyic).

A **DNA microarray database** consists of expression levels of many numbers of genes. Microarray takes the advantage of getting the data of over several thousands of genes in a single experiment. Hence, it acts as a repository for doctors to simultaneously analyze and interpret the values and use the same to predict diseases. Such a database would be helpful in understanding the behavior of human cells and diagnose fatal diseases, such as cancer, at an early stage. In spite of all these benefits, there are two major challenges in the analysis of microarray. [9]. They are:

- Very few samples are available (often less than hundred)
- The dataset is of high dimensionality (nearly thousands or tens of thousands of genes)

In order to diagnose a disease the gene expression data is used. But there is high redundancy in the microarray data as not all the genes contain information relevant to a particular disease. Selection of relevant gene is most important in microarray gene expression analysis[5]. The gene selection method reduces a large dimensional microarray dataset into a small set of genes which can classify the samples.The process of disease classification generally includes two key procedures: gene selection and classifier design. [11]. The former procedure is particularly crucial as the number of genes irrelevant to classification may be huge. Therefore, prediction will turn out to be accurate only if gene selection is performed aptly, that is, the most informative genes that play a significant role in pathogenesis of cancer (in this case, leukemia), are correctly identified from a large number of candidates. Once such genes are chosen, the next procedure is to develop a classifier based on those genes.

Many classifiers that use a large number of genes for prediction have already been designed by researchers. But such multi-gene models are disadvantageous due to the fact that it is not easy to assess which gene is more important in the models. As a result, it becomes hard to detect the significant biomarkers of related cancers. In addition, multi-gene models are complex to understand.Recently, a soft computing method based on rough sets was proposed to conduct cancer classification using single or double genes[3]. This method was improved to achieve sufficient accurate classification and to find important biomarkers with ease by using single-gene models. [3]. Microarrays allow monitoring of gene expression of tens of thousands of genes in parallel. Micro array analysis includes extracting samples from the cells, getting the gene expression matrix from the raw data, and data normalization. A method called Gene filtering is applied to identify the clusters[13].

In order to develop a useful model from a data set, one has to properly understand the values represented by the data. This also helps to perform all kinds of data preprocessing.The Golub dataset used in this research work was taken from the official website of Broad Institute[12]. This dataset consists of 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Each of the 72 patients had bone marrow samples obtained at the time of diagnosis. These observations have been examined with Affymetrix Hgu6800 chips, which have resulted in 7129 gene expressions (Affymetrix probes)[4]. This entire dataset has been divided into train and test datasets. The train data consists of 38 samples, out of which 27 are ALL cases and 11 are AML cases, whereas the test data contains 34 samples, of which 20 are ALL and 14 are AML. We aim to design and train a classifier using the training set and validate the same by predicting the classes of the test set.

The objective of this study is to develop a classification based predictive model for the chosen dataset and develop a tool for the same. The proposed model is developed based on the values in the training dataset and can predict the classes of the test dataset.In this paper, after a brief elaboration on the objectives of the research, the feature selection and classification techniques that have been used for the proposed system are described. This is followed by the results that were obtained from applying these techniques and then the concluding remarks.

II. DATA MINING TECHNIQUE USED

A. Classification

Classification is a data mining technique that takes a set of pre-classified instances to develop a model that can predict the classes of another set of instances, whose class labels are unknown. The classification process consists of learning and classification. Training data and test data are used for this purpose. Training set consists of records which are fed with known class labels. Using train dataset we build a classification model, which is analyzed using the test set records, whose class labels are unknown. [7].

B. Decision Tree

A decision tree is a hierarchical structure consisting of nodes and directed edges. The decision tree has three types of nodes:

- A root node – It has no incoming edges and has zero or more outgoing edges.
- Internal nodes or non-terminal nodes - Each of the internal node has one incoming edge and two or more outgoing edges.
- Leaf or terminal nodes- Every leaf node has exactly one incoming edge and no outgoing edges.

Leaf nodes in a decision tree are assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions, i.e. threshold values to separate records that have different characteristics. [10].

III.PROPOSED FRAMEWORK

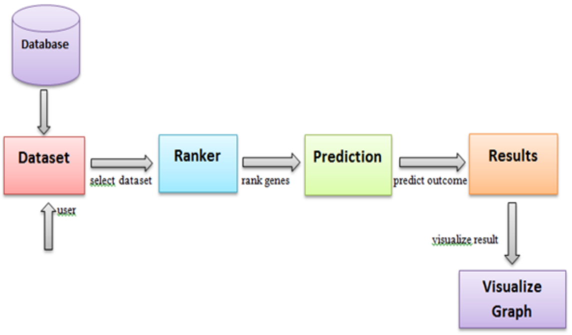


Figure 1. Architecture of Leukos Hiema Carcinoma Classifier

Leukos Heima Carcinoma Classifier is the name of the tool that has been developed as a part of this research work. This software is designed to classify Acute Leukemia cancer into two classes viz. Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). The workflow of LHC classifier is as shown in Figure 1. Firstly, user fetches the required dataset from the file system. Then ranking of genes based on information gain is done using Ranker. From the top 25 genes obtained after ranking, the two most important genes are chosen for prediction. These two genes are Zyxin (X95735_at) and Cystatin A (D88422_at) . The classification of each sample can be viewed in results. Also, total number of cases in each class is displayed. The accuracy of the classifier is shown along with the confusion matrix. The results can be visualized using bar graph.

IV.METHODOLOGY

The proposed tool, Leukos Heima Carcinoma Classifier (LHCC), mainly deals with the prediction of leukemia classes, which are ALL and AML. In order to choose the best genes, a feature selection technique called Ranker has been described in Step 1. For classification purpose, decision tree algorithm has been discussed in Step 2. The description of both the steps is as follows.

Step 1: Feature Selection

In this step, each gene is ranked based on its information gain, which is a symmetrical measure. Information gain ratio is the ratio of information gain to the intrinsic information. Information gain is the change in information entropy H from a prior state to a state that takes some information as given: [5]

IG(T, a) = H(T) – H(T|a)

Entropy is used in calculation of information gain. It is the sum of the probability of each label times the log probability of the same label. Entropy of Y is

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y))$$

p(y) is the marginal probability density function for random variable Y and H(Y) is the Entropy of Y. The conditional entropy of Y after observing X is

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x))$$

p(y|x) is the conditional probability of y given x. The information gained about Y after observing X is

$$IG = H(Y) - H(Y|X)$$

Where IG is the information gain, H(Y) is entropy of attribute Y an H(Y/X) is the conditional entropy of Y. Using the above method, the gene with highest information gain ratio is ranked first and accordingly the remaining genes are ranked. The top twenty-five genes ranked in this manner are chosen for the next step.

Step 2: Finding the most relevant genes

Out of the top twenty-five genes, only two of them were found to be biologically most relevant leukemia-causing genes. They are ZYXIN (X95735_at) and CYSTATIN A (D88422_at).

Zyxin is a protein encoded by ZYX gene in humans, which is the most important gene causing leukemia. Cystatin A is another protein encoded by the CSTA gene. It is correlated with the prognosis and diagnosis of cancer and is up-regulated in AML patients.

Finally, these two genes were used to design the decision tree

Step 3: Decision Tree

In order to build the decision tree using these two genes, the threshold value of each of the gene has to be calculated. [8]. This can be achieved using the following formulae: [6]

$$\mathrm{Info}(S) = -\sum_{i=1}^{k} \frac{\mathrm{freq}(C_i, S)}{|S|} \log_2 \left(\frac{\mathrm{freq}(C_i, S)}{|S|} \right).$$

where,

S is any set of samples

freq(Ci, S) stand for the number of samples in S that belong to class Ci

|S| denote the number of samples in the set S

$$\mathrm{Info}_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \mathrm{Info}(T_i).$$

where,

T is the set of training samples

Info(T) is the information content of T

x is the chosen attribute based on which T has been partitioned.

$$\mathrm{Gain}(x) = \mathrm{Info}(T) - \mathrm{Info}_x(T)$$

The above formula measures the information that is gained when T is partitioned in accordance with x. The aim is to select x which maximizes Gain(x), i.e. the highest information gain.Based on these formulae, the threshold values for Zyxin and Cystatin A are obtained as 938 and 558 respectively. The decision tree for genes is constructed considering all the parameters as shown in Figure 2.

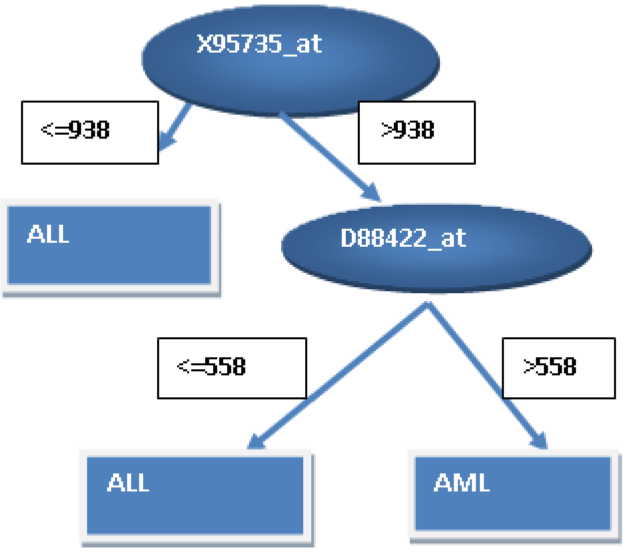
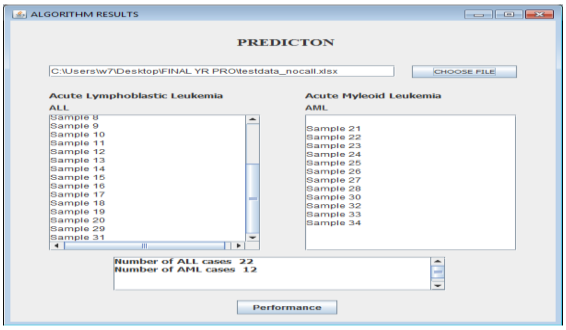


Figure 2. Decision Tree

From the tree it can be seen that all those samples whose values for the gene X95735_at are lesser or equal to 938 are classified as ALL. For those samples with values greater than 938, the next gene D44822_at is to be considered. In this case, those samples whose gene expression values are lesser or equal to 558 are classified as ALL and the remaining samples are classified as AML

V.Results

The test data consist of 34 samples and is stored in excel file (.xlsx format). Once this file is chosen from the prediction page, the classifier is made to run automatically and the predicted output is displayed .Samples classified as ALL and AML are shown in Figure 3 along with the cases belonging to each class. Originally it was found that test data contains 20 ALL and 14 AML cases. Our LHC classifier predicted the test data as 22 ALL and 12 AML i.e. two of the AML samples were wrongly classified as ALL.



C:\Users\w7\Desktop\FINAL YR PRO\testdata_nocall.xlsx

Acute Lymphoblastic Leukemia
ALL

Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
Sample 16
Sample 17
Sample 18
Sample 19
Sample 20
Sample 29
Sample 31

Acute Myleoid Leukemia
AML

Sample 21
Sample 22
Sample 23
Sample 24
Sample 25
Sample 26
Sample 27
Sample 28
Sample 30
Sample 32
Sample 33
Sample 34

```
Number of ALL cases  22
Number of AML cases  12
```

Figure 3. Decision Tree

The confusion matrix for the prediction output is shown in Table 1.

Table 1: Confusion Matrix

2x2 Confusion Matrix		Predicted Class	
		Class1=ALL	Class2=AML
Actual Class	Class1 =ALL	22	2
	Class2=AML	0	12

Out of 34 samples,32 of them were correctly classified with an accuracy of 94.11%. 2 instances were incorrectly classified. The results are visualized using bar graph shown in Figure 4.

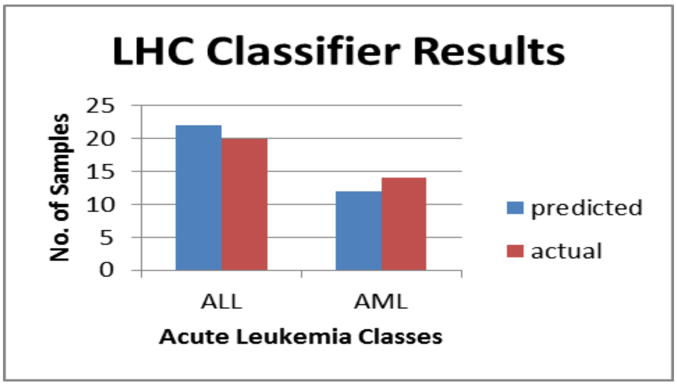


Figure 4. Graph showing the Predicted Results

It can be seen that all the 20 ALL instances are correctly classified whereas 2 of the AML instances are misclassified.

VI.Conclusion

The LHC classifier is simple nd works efficiently. It uses only two genes and thus the complexity in its execution is reduced.

Out of all the test samples, only two of them are misclassified. This gives a reasonable accuracy of 94.11%. Also, the working of the tool, LHC Classifier, is found to be quite consistent. The tool developed exclusively for Golub's dataset is found to work with satisfactory speed. Since this particular dataset has been used most often for research in leukemia, and precisely, cancer diagnosis, the main idea was to design a classifier and develop a tool for classifying the same dataset.The efficiency of the algorithm may be improved if the dataset is populated with additional instances. Different prediction techniques can be used which may boost the accuracy of classification. Increasing the number of classes to Chronic Lymphocytic Leukemia (CLL) and Chronic Myelogenous Leukemia (CML) also provide scope for future work in classifier design.

REFERENCES

[1] Golub, T.-R., Slonim, D.-K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov J.-P., Coller, H., Loh M.-L., owning, J.-R., Caligiuri, M.-A., et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 286(5439):531-537, 1999

[2] Wang, X., and Gotoh, O., "Microarray-Based Cancer Prediction Using Soft Computing Approach", Cancer Informatics, 7:123–139, 2009.

[3] Wang, X., and Gotoh, O., "Cancer Classification Using Single Gene", Cancer Informatics, 7:123–139, 2011.

[4] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed, "Summaries of Affymetrix GeneChip probe level data", Nucleic Acids Research, Feb 15;31(4)e15,2003.

[5] Kshipra Chitode, Meghana Nagori, "A Comparative Study of Microarray Data Analysis for Cancer Classification", International Journal of Computer Applications, November (0975 – 8887) Volume 81 – No 15, November 2013

[6] K. Ming Leung, "Decision Trees and Decision Rules", Polytechnic University Department of Computer Science / Finance and Risk Engineering, December 2007.

[7] A. Bharathi, A.M. Natarajan, "Cancer Classification using Support Vector Machines and Relevance Vector Machine based on Analysis of Variance Features", Journal of Computer Science, 7 (9): 1393-1399, ISSN 1549-3636, 2011.

[8] S. K. Shevade, S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression", Bioinformatics 19 (17): 2246-2253, 2003

[9] R. L. Somorjai, B. Dolenko and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions", Bioinformatics 19 (12): 1484-1491,2003

[10] Oscar Picchi Netto, Sergio Ricardo Nozawa, Rafael Andres Rosales Mitrowsky, Alessandra Alaniz Macedo, Jose Augusto Baranauskus, "Applying Decision Tree to Gene Expression Data from DNA Microarrays: A Leukemia Case Study" ISSN 2175-2761,2010

[11] Dr. S. Santhosh Baboo, Mrs. S. Sasikala, "Multicategory Classification Using Relevance Vector Machine for Microarray Gene Expression Cancer Diagnosis", International Journal of Advanced Research in Computer Science, Vol 1,No. 4, ISSN. 0976-5697, 462-473, 2010.

[12] http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

[13] P.Venkatesan and Jamal Fathima .J.I, "Identification of differentially expressed genes by unsupervised learning method", International Journal of Data Mining Techniques and Applications ISSN: 2278-2419 Vol 02, Issue 01,121-125,June 2013