ISSN 2320 - 2602
Volume 3, No.3, March 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst10332014.pdf

SURVEY OF GENE-EXPRESSION-BASED CANCER SUBTYPES PREDICTION

G. Mohana Lakshmi[1], K.Mythili[2]
[1]Research Scholar, Hindustan college of Arts and Science, India, gmohanalakshmi90@gmail.com
[2] Associate Professor, Hindustan college of Arts and Science, India, mythiliarul@gmail.com

ABSTRACT

Classification of data has been successfully applied to a wide range of application areas, such as scientific experiments, medical diagnosis, credit approval, weather prediction, customer segmentation, target marketing and fraud detection. A major challenge in clinical cancer research is the prediction of prognosis at the time of tumor discovery. Accurate prediction of different tumor types can help in providing better treatment and toxicity minimization on the patients. To detecting the tumor data mining algorithms are applied; in other words data mining algorithms are applied to predict cancer data like a classification problem. Some of the data mining techniques are explained in this survey article. The objective of this study is to summarize various review and technical articles on diagnosis and prognosis of cancer. In this research we provided an overview of the current research being carried out on various cancer datasets using the data mining techniques to enhance the cancer diagnosis and prognosis**.** This demonstration of these techniques is used to obtain the efficient scheme for cancer subtype prediction. We can obtain the more efficient method or we may propose the new technique to overcome the problems in these existing approaches. This survey article is intended to provide easy accessibility to the main ideas for non-experts.

Keywords: Micoarray, Gene Expression, classification, cancer subtype prediction, Breast cancer, Feature selection (FS) and SEER (Surveillance Epidemiology and End Results).

1. INTRODUCTION

Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labeled data (i.e., data from a sample with clinical follow-up) can be exploited for learning. They are also useful for identifying potential gene markers for each cancer subtype that helps in successful diagnosis of particular cancer type existing system developed a classification system by identifying potential gene markers and subsequently applying the proposed technique on the selected genes for the classification of human cancer. Recent research in the area of cancer diagnosis suggests that unlabeled data, in addition to the small number of labeled data, can produce significant improvement in accuracy, a technique called semisupervised learning [5]. Indeed, semisupervised learning has proved to be effective in solving different biological problems including protein classification [6], prediction of transcription factor–gene interaction [7], and gene- expression based cancer subtype discovery [8], [9].

The advent of microarray technology has made it possible to study the expression profiles of a large number of genes across different experimental conditions. Microarray- based gene expression profiling has shown great potential in the prediction of different cancer subtypes [1], [2], [3]–[4]. Nevertheless, small sample size remains a bottleneck in obtaining robust and accurate prediction models. The number of samples in microarray based-cancer studies is usually small because microarray experiments are time consuming, expensive, and limits based on sample availability.

Cancer classification using microarray data poses another major challenge because of the huge number of features (genes) compared to the number of examples (tissue samples). This is an important problem in machine learning which is known as feature selection [10]. Successful gene identification involves 1) dimension reduction to reduce computational cost; 2) reduction of noise to increase classification performance; and 3) identification of more interpretable features. Only a small number of genes in the microarray data consisting of thousands of genes that shows strong correlation with the target phenotypes. Only a few small selected genes have their biological relationship with the target diseases. A survey on the classical and computational intelligence methods for gene identification can be found in [14].

Gene expression refers to the level of production of protein molecules defined by a gene. Gene expression monitoring is one of the most fundamental approach in measuring gene expression is to measure the mRNA instead of proteins, since mRNA sequences hybridize with their complementary RNA or DNA sequences while this property lacks in proteins[11],[12],[15]. The DNA arrays, pioneered are novel technologies that are designed to measure gene expression of tens of thousands of genes in a single experiment. The measuring ability of gene expression for a very large number of genes, entire genome covering for some small organisms, increases the issue of characterizing cells in terms of gene expression, in other words using gene expression to determine the fate and functions of the cells[20]. The major fundamental of the characterization problem is that of identifying a set of genes and its

expression patterns that either characterize a certain cell state or predict a certain cell state in the future.

Gene selection aims to find a set of genes that best discriminate biological samples of different types. The selected genes are "biomarkers," and they form a "marker panel" for analysis. Most gene selection schemes are based on binary discrimination using rank-based schemes such as information gain that reduces the entropy of the class variables given the selected features. One significant issue in these rank-based methods is data sparseness[18]. In the view of example, the estimation of the traditional information gain is an empirical estimation directly on the data. Suppose we select the 11th gene for a data set. The 10 selected genes split the training data groups (assuming that each gene does a binary split). Because we have very few samples in most groups, the estimations of mutual information between the 11th gene and the target in each group are not accurate[19]. Thus, the information gain, which is the sum of the mutual information over all groups, is not accurate.

2. VARIOUS TECHNIQUES FOR CANCER PREDICTION

2.1 Data mining techniques for diagnosis and prognosis of cancer disease

Breast cancer has become the primary reason of death in women in developed countries. The most valuable way to reduce breast cancer deaths is to detect it earlier. In the early, diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The main goal of these predictions is to assign patients to one of the two group either a "benign" that is noncancerous and cancerous such as "malignant"[13][14]. The prognosis problem is the long-term care for the disease for patients whose cancer has been surgically removed. Foretelling the outcome of a disease is one of the most interesting and challenging tasks where to develop applications of data mining. The use of computers with automated tools, huge volumes of medical data are being collected and made available to the medical research groups.

Data mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and prepared them able to predict the outcome of a disease using the historical datasets. The goal of this study is to summarize various review and technical articles on diagnosis and prognosis of breast cancer. In this research, provides an overview of the current research being carried out on various breast cancer datasets using the data mining techniques to enhance the breast cancer diagnosis and prognosis**.**

Decision Tree learning is one of the most widely used and practical methods for classification. In this approach, learned trees can be represented as a set of if-then rules that improve human readability. In addition, Decision trees are very simple to understand and interpret by domain experts. A decision tree contains the nodes that have exactly one incoming edge, without the root node that has no incoming edges[16][17]. An internal node is a node with outgoing edges, whereas the other nodes are called leaves or terminal nodes or decision nodes. The evaluation process is conducted using Weka J48, C4.5 decision tree is generated. A number of parameters were tested such as the confidence factor 6 O. utilized for pruning, whether to use binary splits or not, then whether to prune the tree or not and the minimum number of instances per leaf.

2.2 Analysis of feature selection with classification

Classification is extensively used in various application domains: retail target marketing, fraud detection, design of telecommunication service plans, Medical diagnosis, etc. In the domain of medical diagnosis classification plays an important role. Because large volume of data maintained in the medical field, in this field classification is extensively used to make decisions for diagnosis and prognosis of patient's disease. For diagnosis of diseases such as breast cancer, ovarian cancer and heart sound diagnosis Decision tree classifiers are used extensively.

Classification is a data mining task which assigns an object to one of several pre-defined categories based on the attributes of the object. The input to the problem is a data-set termed as the training set, which contains the number of examples each having a number of attributes. The attributes are either continuous, while the attribute values are ordered, or categorical while the attribute values are unordered. The class label is the One of the categorical attributes which is called or the classifying attribute. The main goal is to use the training set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training dataset. In addition, Classification has been studied extensively in statistics, machine learning, neural networks and expert systems over decades. There are various classification methods:

- Decision tree algorithms
- Bayesian algorithms
- Rule based algorithms
- Neural networks
- Support vector machines
- Associative classification
- Distance based methods
- Genetic Algorithms

Feature selection (FS) plays a major role in classification. It is one of the Preprocessing techniques in data mining.

Feature selection is comprehensively used in the fields of statistics, pattern recognition and medical domain area. Feature Selection means reducing the number of attributes. By removing the irrelevant and redundant attributes the attributes are reduced, which do not have significance in classification task. The performance of the classification techniques is improves with the feature selection approach. The process of feature selection is

• Candidate subsets of attributes are generated from original feature set using searching techniques.

• Estimation of each candidate subset to determine the relevancy towards the classification task using measures such as distance, dependency, information, consistency, classifier error rate.

• To determine the relevant subset or optimal feature subset, termination condition is used.

• To check the selected feature subset, validation process is performed.

2.3 Breast Cancer Diagnosis in Different Datasets Using Multi-Classifiers

This work presents a comparison among the different classifiers decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method. Also, here introduce a fusion at classification level between these classifiers to get the most suitable multi-classifier approach for each data set.

It consists of two phases namely: training and testing phases. In the training phase the four steps are included. They are: acquisition, preprocess, feature extraction and feature selection, whereas the testing phase includes the same four steps in the training phase in addition to the classification step. The sensor data are subject to a feature extraction and selection process for determining the input vector for the subsequent classifier in the acquisition step. This builds a decision regarding the class associated with this pattern vector. Derived from either feature selection or feature extraction, Dimensionality reduction is accomplished. In the preprocessing step, the image is prepared and filtered to clear the noise and improve the quality of the images. Whereas, feature extraction considers the whole information content and maps the useful information content into a lower dimensional feature space. Feature selection approach is based on omitting those features from the available measurements which do not contribute to class separability. In other words, redundant and irrelevant features are ignored. In the Classification step different classifiers are applied to get the best result of diagnosing and prognosing the tumor.

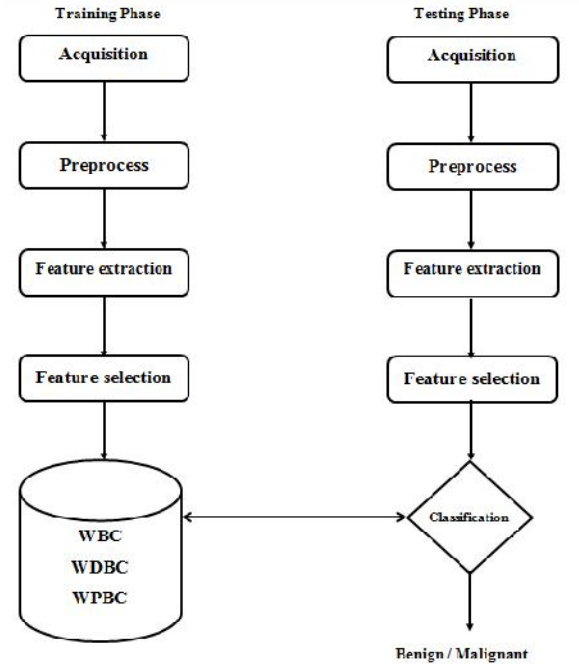

Figure 1. Breast Cancer Diagnosis Model

2.4 Predicting Breast Cancer Survivability

In this research, present data mining techniques to predict the survivability rate of breast cancer patients. In this study, used the SEER data and have introduced a pre-classification approach that take into account three variables: Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD). The data used is the SEER Public-Use Data. In the preprocessing step, the data set consists of 151,886 records and it have all the available 16 fields from the database of SEER. Here investigated three data mining techniques: the Naïve Bayes method, the back-propagated neural network approach and the C4.5 decision tree algorithms. Several experiments were conducted using these techniques.

The Naïve Bayes technique depends on the famous Bayesian approach following a simple, clear and fast classifier. It has been called 'Naïve' due to the fact that it assumes mutually independent attributes. In actual fact, this is almost never true but is achievable by preprocessing the data to remove the dependent categories. This method has been used in many areas to represent, utilize, and learn the probabilistic knowledge and significant results have been achieved in machine learning. The second technique uses artificial neural networks. In this study, a multi-layer network with back-propagation (also known as a multi layer

perceptron) is used. The third technique is the C4.5 decision-tree generating algorithm. C4.5 is based on the ID3 algorithm. It has been shown that the last two techniques have better performance.

In order to have a fair measure of the performance of the classifier; here used a cross validation with 10 folds. For the most elementary form, cross-validation consists of dividing the data into k subgroups. Each subgroup is predicted via the classification rule constructed from the remaining (k-1) subgroups, and the estimated error rate is the average error rate from these k subgroups. In this way, the error rate is estimated in an unbiased manner. The final classifier rule is calculated from the entire data set.

2.5 Robust gene list for predicting outcome in cancer

Considerable effort has been devoted recently to outcome prediction for several kinds of cancer on the basis of gene expression profiling (2–8), along with special emphasis on breast carcinoma (9–13). Numerous of these studies reported considerable predictive success. These successes were, though, somewhat thwarted by two problems: (*i*) while one group's predictor was tested on another group's data (for the same type of cancer patients), the success rate decreased extensively; and (*ii*) comparison of the predictive gene lists (PGLs) discovered by different groups revealed very minute overlap. These problems indicate that the currently used PGLs suffer from instability of their membership and of their predictive performance. These declarations are well illustrated by two prominent studies of survival prediction in breast cancer.

These intriguing problems have received great attention by the community of cancer research and have been addressed in various topical studies. The observable and most straightforward explanation of these apparent discrepancies is to attribute them to (i) different groups using cohorts of patients that differ in a potentially relevant factor (such as age), (ii) different microarray technologies used, and (iii) different methods of data analysis.

Lack of stability of these PGLs has been either ignored or demonstrated for a particular experiment by reanalysis of the data. Here, propose a mathematical framework to define a quantitative measure of a PGL's stability. Moreover, here present a method that uses existing data of a relatively small number of samples to project the expected stability one would obtain for a larger set of training samples, by this means helping to design an experiment that generates a list that has a desired stability.

3. CONCLUSION

We reach the end of this survey on cancer subtype prediction. Our goal has been to present and explain the main ideas behind the existing techniques, to classify them according to the type of approach proposed, and to show how they perform in practice in a subset of possible practical scenarios. In this survey, various cancer detection techniques are presented. This survey is used to analyze a variety of tumor prediction methods. In other words, different approaches are presented and analyzed. We focus the cancer prediction techniques that are mostly related to the gene-expression. We can obtain the more efficient method or we may propose the new technique to overcome the problems in these existing approaches. This survey article is intended to provide easy accessibility to the main ideas for non-experts.

REFERENCES

[1] A. J. Gentles, S. K. Plevritis, R. Majeti, and A. A. Alizadeh, "Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia," *J. Amer. Med. Assoc.*, vol. 304, pp. 2706–2715, 2010.

[2] H.K. Kim, I. J. Choi, C. G. Kim, A. Oshima, and J. E. Green, "Gene expression signatures to predict the response of gastric cancer to cisplatin and fluorouracil," *J. Clin. Oncol.*, vol. 27, no. 15S, 2009.

[3] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Combining Pareto-optimal clusters using supervised learning for identifying coexpressed genes," *BMC Bioinformat.*, vol. 10, no. 27, 2009.

[4] U. Maulik and A. Mukhopadhyay, "Simulated annealing based automatic fuzzy clustering combined with ANN classification for analyzingmicroarray data," *Comput. Oper. Res.*, vol. 37, no. 8, pp. 1369–1380, 2010.

[5] O. Chapelle, V. Sindhwani, and S. S. Keerthi, "Optimization techniques for semi-supervised support vectors," *J.Mach. Learn. Res.*, vol. 9, pp. 203– 233, 2008.

[6] J.Weston, C. Leslie, E. Le, D. Zhou, A. Elisseeff, andW. S. Noble, "Semisupervised protein classification using cluster kernels," *Bioinformatics*, vol. 21, pp. 3241–3247, 2008.

[7] J. Ernst, Q. K. Beg, K. A. Kay, G. Bal´azsi, Z. N. Oltvai, and Z. Bar-Joseph, "A Semi-supervised method for predicting transcription factor–gene interactions in escherichia coli," *Plos Comput. Biol.*, vol. 4, p. e1000044, 2008.

[8] D. C. Koestler, C. J. Marsit, B. C. Christensen, M. R. Karagas, R. Bueno, D. J. Sugarbaker, K. T. Kelsey, and E. A. Houseman, "Semi-supervised recursively partitioned mixture models for identifying cancer subtypes," *Bioinformatics*, vol. 26, pp. 2578–2585, 2010.

[9] I. Steinfeld, R. Navon, D. Ardig`o, I. Zavaroni, and Z. Yakhini, "Clinically driven semi-supervised class discovery

in gene expression data," *Bioinformatics*, vol. 24, pp. 190–197, 2008.

[10] A. Blum and P. Langley, "Selection of relevant features and examples in machine learing," *Artif. Intell.*, vol. 97, no. 1/2, pp. 245–271, 1997.

[11] S. Bandyopadhyay, U. Maulik, and D. Roy, "Gene identification: Classical and computational intelligence approaches," *IEEE Trans. Syst., Man, Cybern. C*, vol. 38, no. 1, pp. 55–68, Jan. 2008.

[12] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[13] D.Lavanya, Dr.K.Usha Rani,..," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

[14] Gouda I. Salama, M.B.Abdelhalim, and Magdy Abd-elghany Zeid, "*Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers*", International Journal of Computer and Information Technology (2277 – 0764), Volume 01– Issue 01, September 2012.

[15] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques".

[16] Hitoshi Iyatomi, Hiroshi Oka, M.Emre Celebi, Masahiro Hashimoto, Masafumi Hagiwara, and Masaru Tanaka, Koichi Ogawa, "An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm", Computerized Medical Imaging and Graphics 32, 566–579, 2008.

[17] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Nat. Acad. Sci. USA*, vol. 103, pp. 5923–5928, 2006.

[18] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques".

[19] Hitoshi Iyatomi, Hiroshi Oka, M.Emre Celebi, Masahiro Hashimoto, Masafumi Hagiwara, and Masaru Tanaka, Koichi Ogawa, "An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm", Computerized Medical Imaging and Graphics 32, 566–579, 2008.

[20] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Nat. Acad. Sci. USA*, vol. 103, pp. 5923–5928, 2006.