PREDICTION OF ACUTE MYELOID LEUKEMIA CANCER USING DATAMINING-A SURVEY

Durairaj M
Assistant Professor
School of Comp. Sci., Engg. & Applications,
Bharathidasan University, Trichy, TN, India
durairaj.bdu@gmail.com

Deepika R
Research Scholar
School of Comp. Sci., Engg. & Applications,
Bharathidasan university, Trichy, TN, India
deepikar9191@gmail.com

ABSTRACT
Cancer is a life threatening disease for human being. It has claimed many breaths away over the years. This paper reviews several articles related in diagnosing myeloid dysplastic syndrome (MDS) and acute myeloid leukemia (AML) pathogenesis. Though there is much better medical advancement in investigating the presence of cancer, it still remains uncharacterized, and their diagnosis relies strongly on the idiosyncratic factors. This article address the issue of applying medical data mining, an emerging research trend which helps in finding accurate solutions in many fields. This paper highlights the various data mining techniques used for the diagnosis of cancer.

Keywords
Cancer Classification, Feature Selection, Microarray Data, Machine Learning, Acute Myeloid Leukemia, Data Mining

1. INTRODUCTION
Accurate cancer prediction is significant for the fruitful application of specific medications. Though the prediction of Cancer has developed in the recent years, there is still a necessity for a completely computerized and less subjective method for cancer diagnosis. Investigations demonstrated that DNA microarrays may offer helpful information for cancer classification at the gene expression level due to their ability to measure the plenty of messenger ribonucleic acid (mRNA) transcripts for thousands of genes at the same time. Quite a lot of machine learning algorithms have previously been implemented to classify tumors using microarray data. Voting machines and self-organizing maps (SOM) were used to analyze acute leukemia. Support vector machines (SVMs) were applied to multi-class cancers Biological data processing. This possess numerous challenges like sequence discovery, sequence ending, revealing unknown relationship with regard to structure, drug discovery and performance of genes to know biological systems. This field expects demands for instantaneous prediction and classification due to the accessibility of DNA cancer details, structure organization information of proteins and microarray technology to provide dynamic information about thousand of genes in data.

The aims of Bioinformatics are:
1. To arrange data in such way that allows the researchers and practitioners to obtain available information to submit new entries they produce
2. To create software tools and resources that helps in analysis and management of data
3. To custom the data to examine and understand the results in biological method
4. To aid pharmaceutical industry physicians to understand and predicting the microarray cancer data structures

In this work, the recent articles published on the application of data mining tools on medical data analysis are reviewed. The different data mining application and techniques are classified and compared. This review aims to propose an effective methodology for prediction and analysis of Acute Myeloid Leukemia cancer.

This paper divided as, the section 1 describes the medical data mining and its uses. In the Section 2, the different data mining tools and applications are discussed. Section 3 presents the review of recent articles on the application of data mining tools on medical data. Section 4 discusses the outcome of this review and compares the tools. Section 5 concludes with defining future work on this direction.
2. MICROARRAY IN DATA MINING

Microarray is one of the recent technologies that ensure the investigators to research and identify the problems which were assumed to be untraceable by assisting the concurrent measurement of thousands of gene expression levels. Microarray fixes the DNA molecules on the crystal slide in an ordered manner at specific area called probes or spots. The spots are then traced on the crystal slide using different methodologies like photolithography to robot spotting. There are only two possible kinds of DNA spots can either be short stretch of oligo-nucleotides corresponding to the gene or a complete copy of genomic DNA.

This paper focuses on biologist’s perspective to induce information concerning the many tools and programs out there for microarray data processing tasks. With this motivation at the tip of every knowledge mining task, this work is to provide the list of tools with its underlying algorithms, internet resources and relevant reference. Analysis of Microarray data Microarray datasets are generally very huge, analytically accurate and is influenced by a number of variables. It is tremendously beneficial to decrease the dataset to those genes that are best distinguished between the two cases or classes (e.g. normal vs. diseased). Such investigation yields a list of genes whose expression is restrained to modify and recognized as differentially expressed genes. Identification of disparity in gene expression is the primary task of an extensive microarray analysis.

In this work, the articles published recently in the reputed journals are reviewed. This work is performed to list out suitable Data mining techniques to apply for the efficient diagnosis and prediction of cancer. This work also aims to device a suitable mechanism to early defection of the cancer disease.

3. RELATED WORK

Kaishi Li, et.al. [1] Presented a methodology for Feature extraction of microarray genes which has a greater impact on its classification and clustering as it is taken as input to any network. The use of gene expression data in discriminating two types of very similar cancers acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) presented in Classification results are reported in applying methods other than neural networks. This paper explored the role of the feature vector in classification. This work concludes that in order to achieve best results in learning algorithm, feature subset selection method should be applied on to the dataset.

A. Zibakhsh, et.al. [2] Presented a new memetic algorithm that has the ability of filtering interpretable and precise fuzzy if–then rules over cancer data. The concept memetic algorithms with the Multi-View fitness function were introduced as a first approach. Multi-View fitness function that was presented considered two kinds of evaluating procedures. Procedure one, was placed within the main progressive structure of the algorithmic rule. This evaluates every single fuzzy if–then rule per the required rule quality. The second procedure determines the standard of every fuzzy rule per the full fuzzy rule set performance. Compared to classic memetic algorithms, these forms of memetic algorithms enhance the rule discovery method considerably.

Gopala Krishna Murthy Nookala, et. al. [3] created a comprehensive comparative analysis of fourteen totally different classification algorithms and their performance has been evaluated by victimization three different cancer information sets. The results indicated that none of the classifiers outperformed all others in terms of the accuracy once applied on all the three information sets. Most of the algorithms performed higher because the size of the info set is augmented. They counseled the users to not follow a selected classification methodology and to judge totally different classification formulas and choose the classic algorithm.

Shweta Kharya, et.al [4] mentioned various data processing approaches that are used for carcinoma identification and prognosis. Carcinoma identification is distinctive of benign from malignant breast lumps and carcinoma prognosis predicts once carcinoma is to recur in patients that have had their cancers excised. This study paper summarized varied review and technical articles on carcinoma identification and prognosis conjointly they targeted on current analysis being done out victimization the info mining techniques to reinforce the carcinoma identification and prognosis.

Cheng-Mei Chen, et.al. [5] established a survival prediction model for cancer of the liver victimization data processing technology. The information were collected from the cancer registration data source of a medical center in Northern Taiwan between 2004 and 2008 comprised of 227 patients were recently diagnosed with cancer of the liver throughout this point. 9 variables relating cancer of the liver survival were analyzed victimization t-test and chi-square takes a look at. Six variables showed vital. Artificial Neural Network (ANN) and Classification and Regression Tree (CART) were adopted as prediction models. The models were tested in 3 conditions; one variable of clinical stage alone and other variables such as
significant and non-significant. The result guaranteed five year endurance with the output prediction.

Xiangchun Xiong, et. al. [6] discussed on three methods to diagnose breast cancer. Mammography, FNA (Fine Needle Aspirate) and surgical biopsy. They used FNA with a Data Mining & Statistics method to get an easy way to achieve a best result. They combined some statistical methods with data mining methods to find the unsuspected relationships. They explored that statistics and data mining techniques can offer great promise in helping us uncover patterns in the data.

Chun-Hui Wu, et. al. [7] employed some of the data mining techniques to explore hidden knowledge among meridian energy of prostate cancer from 213 patients’ health examination data including patient demographics and evaluations for the prostate-specific antigen (PSA) blood test as well as the meridian energy. The findings were considered as helpful reference in diagnosis and treatment of prostate cancer for TCM physicians. This study provided new scientific and quantitative information for TCM physicians in clinical practice of prostate cancer. TCM physicians would benefit from a better understanding of the relationship between meridian energy system and prostate cancer.

Soltani Sarvestani, et. al. [8] collected datasets for breast cancer knowledge discovery and invoked various data mining techniques to find out the percentage of disease development. Thus, the result helped in selecting a reasonable treatment of the patient. This work also indicated that statistical neural networks can be effectively used for breast cancer diagnosis to help oncologists.

Yao Liu and Yuk Ying Chung [9] implemented a classifier for the prediction of lung and breast cancer, which are the most common cancers for both men and women. The author also examined the usefulness of the new rule pruning procedure, and showed the proposed procedure which has a positive influence on the accuracy for both DPSO and PSO. This work also compared their approach with popular data mining algorithms on classifying breast cancer and lung cancer, and demonstrated that DPSO combined with the proposed rule pruning is effective in predicting common types of cancer. Experiment showed that the new pruning method increased the classification accuracy, and this approach is found to be effective in cancer prediction.

Nashat, et.al. [10] presented a method to find a clustering pattern of the genes involved in breast cancer. They designed a growing Hierarchical Self-Organizing Map (GHSOM) to mine gene microarray data. They applied their technique on 24,481 genes of DNA microarray of breast tumor samples. Result revealed 17 genes that are likely to be correlated with four breast cancer marker genes.

Sivagowry, S, et. al. [11] presented a method of Medical Data Mining is a domain of challenge which involves lot of imprecision and uncertainty. Provision of quality services at affordable cost is the major challenge faced in the health care organization. This work proposed a methodology, which applied Data mining technique for the prediction of heart disease. This proposed methodology successfully diagnosed heart disease in early response time.

4. DISCUSSION AND ANALYSIS
Some of works recently published on the Data Mining application for the diagnosis of cancer disease are briefly reviewed in this paper. It is observed from the review that some of the works are produced more encouraging results.

The objectives of these reviewed papers are to understand the essential features emphasizing the development and progress of life and also to discover the genetic causes of irregularities befalling in the functioning of the human organs. Hybridization stands as the core principle of microarrays amongst two DNA strands, the complementary property of nucleic acid sequences precisely paired with each other by creating hydrogen bonds between complementary nucleotide base pairs. Conversely, with the population of voluminous amounts of microarray data, it has now developed gradually more important to discourse the challenges of data quality and calibration related to this technology.

The recent advancement of the microarray technology has supported for a really high resolution mapping of chromosomal aberrations with the employment of their application array platform. Computational knowledge analysis tasks like data processing which incorporates classification and agglomeration accustomed extract helpful information from microarray data. Additionally, relating organic phenomenon knowledge with alternative biological information; it'll give quite biological discoveries like transcription factor binding web site analysis, pathway analysis, and macromolecule protein interaction network analysis.

There are two common approaches for depth microarray analysis of, they are, classification and clustering. Clustering is one of the unsupervised
approaches to cluster the data into groups of genes or test data with similar arrangements that are individual to the group. Classification is supervised learning and additionally called category prediction or discriminant analysis. Generally, classification may be a method of learning-from-examples. Given a group of pre-classified examples, the classifier learns to assign associate unseen action to one of the categories.

This literature review shows that the Data mining can be an effective tool in diagnosing diseases in early stages. Review shows that poor clinical decision may lead to disastrous consequences. Health care data is massive and clinical decisions are often made based on doctor’s experience rather than on the knowledge rich data hidden in the data base. This in some cases will result in errors, excessive medical cost which affects the quality of service to the patient’s. It is possible to gain the advantage of Data mining in health care by employing it as an intelligent diagnostic tool.

4.1 COMPARATIVE ANALYSIS
This section describes the various classification techniques used in the reviewed papers and the accuracy they were able to achieve by applying those algorithm on their selected dataset. Besides, a dataset was derived from UCI (University of California Irvin) machine learning repository for breast cancer to trace out the accuracy of the classification algorithms. The dataset contains ten attributes namely class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiate and 287 instances. The dataset was converted into .csv file format and inputted into the popular data mining tool WEKA 3.6 (Waikato Environment for Knowledge Analysis). Classification algorithms that were used in the surveyed papers were executed through WEKA to get the precision of the algorithms. The accuracy of reviewed works are compared and displayed in TABLE 1. There by providing the researcher the most standard classification algorithms that maintains its credibility almost in all the datasets.

![Table 1. Comparative Analysis of Reviewed Vs Derived Dataset Over Classification Algorithms](image)

<table>
<thead>
<tr>
<th>Citation</th>
<th>Algorithms</th>
<th>Accuracy in % (Reviewed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Bayes Net</td>
<td>90.95%</td>
</tr>
<tr>
<td>[2]</td>
<td>JRip</td>
<td>40.58%</td>
</tr>
<tr>
<td>[3]</td>
<td>Multilayer Perceptron</td>
<td>96.5%</td>
</tr>
</tbody>
</table>

The Table 1 clearly shows that the classification algorithms is not stable and fluctuates between the datasets. From the results it is proven that the P90 + 7IBk algorithm has the highest correct accuracy in reviewed paper with 97.9% did not proved to be best with the derived data set. It is observed that the Decision table algorithm obtained 80% accuracy.

There are 11 different Data Mining applications used in the reviewed work for the diagnosis of the disease, where as IBK and Decision Table techniques provided more accurate results than the other techniques.

5. CONCLUSION
This paper discusses three important things that are considered to be unique from other works; firstly it presents the importance of data mining approaches on medical data mining. Furthermore it, thoroughly surveys the related work that has been carried out so far on cancer prediction. Finally, this paper compares the correct accuracy level of the reviewed papers and discusses the key facts that are obtained from the results. Through the results, it was noted, that the correct accuracy of the classification algorithms are not stable and they differ from one another. As a future direction, this work will be focused to bring the stability in classification algorithms.

REFERENCES


