International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Novel Method for Cancer Gene Prediction Using Back Propagation Algorithm

Annakkodi P. S[1], Manjula Devi B[2]

[1]Research Supervisor, Head, Department of Information Technology,
Sri Ramalinga Sowdambigai College of Science & Commerce, Coimbatore-109, Tamil Nadu, India

[2]Research Scholar, Department of Computer Science, Sri Ramalinga Sowdambigai College of Science & Commerce,
Coimbatore-109, Tamil Nadu, India

Abstract: ***Cancer is a dynamic disease, it has been estimated that a new genetic alteration occurs in one out of 10000 tumor cells at each cell division. This leads to an increasing number of genetic aberrations in cancer cells, some of which result in altered gene expression. These changes in gene ex-pression are at least partially responsible for the characteristic of a tumor or tumor-type. Genome-wide gene expression pro-files provide detailed "tumor signatures" and can potentially be used for molecular diagnosis and classification of tumors. Cancer is a complex family of diseases, from the view of molecular biology; cancer is a genetic disease resulting from abnormal gene expression. Cancer leads to all mortalities, making it the second leading cause of death in the United States. Early and accurate detection of cancer is critical to the well being of patients. Analysis of gene expression data leads to cancer identification and classification, which will facilitate proper treatment selection and drug development. Gene expression data sets for ovarian, prostate, and lung cancer were analyzed in this research. An integrated gene-search algorithm for genetic expression data analysis was proposed. This integrated algorithm involves a genetic algorithm and correlation-based heuristics for data preprocessing (on partitioned data sets) and data mining (decision tree and support vector machines algorithms) for making predictions. Knowledge derived by the proposed algorithm has high classification accuracy with the ability to identify the most significant genes.***

Keywords: Consulting, Testing, Requirements, Process Improvement.

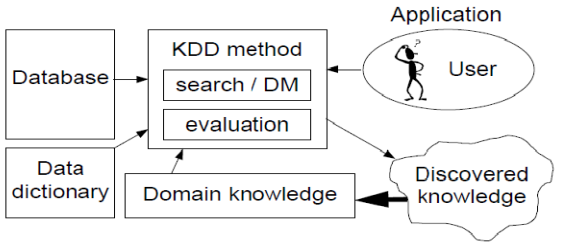
1. Introduction

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven

1.1 KDD Process

The idea of automatic knowledge discovery in large databases is first presented informally, by describing some practical needs of users of modern database systems. Several important concepts are then formally defined and the typical context and resources for KDD are discussed. Then the scope of KDD and DM is briefly presented in terms of classification of KDD/DM problems and common points between KDD and several other scientific and technical disciplines that have well-developed methodologies and techniques used in the field of KDD.

KDD methods often make possible to use domain knowledge to guide and control the process and to help evaluate the patterns. In such cases domain knowledge must be represented using an appropriate knowledge representation technique (such as rules, frames, decision trees, and the like). Discovered knowledge may be used directly for database query from the application, or it may be included into another knowledge-based program (e.g., an expert system in that domain), or the user may just save it in a desired form. Discovered patterns mostly represent some previously unknown facts from the domain knowledge. Hence they can be combined with previously existing and represented domain knowledge in order to better support subsequent runs of the KDD process



A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable.

Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Health care data is massive. It includes patient centric data, resource management data and transformed data.

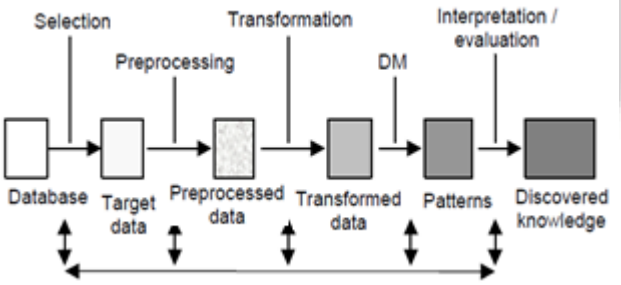
1.2 Databases and Machine Learning

In database management, a database is a logically integrated collection of data maintained in one or more files and organized to facilitate the efficient storage, modification, and retrieval of related information. In a relational database, for example, data are organized into files or tables of fixed-length records. Each record is an ordered list of values, one

value for each field. Information about each field's name and potential values is maintained in a separate file called a data dictionary. A database management system is a collection of procedures for retrieving, storing, and manipulating data within databases. In machine learning, the term database typically refers to a collection of instances or examples maintained in a single file.2 Instances are usually fixed-length feature vectors. Information is sometimes also provided about the feature names and value ranges, as in a data dictionary

A learning algorithm takes the data set and its accompanying information as input and returns a statement (for example, a concept) representing the results of the learning as output.

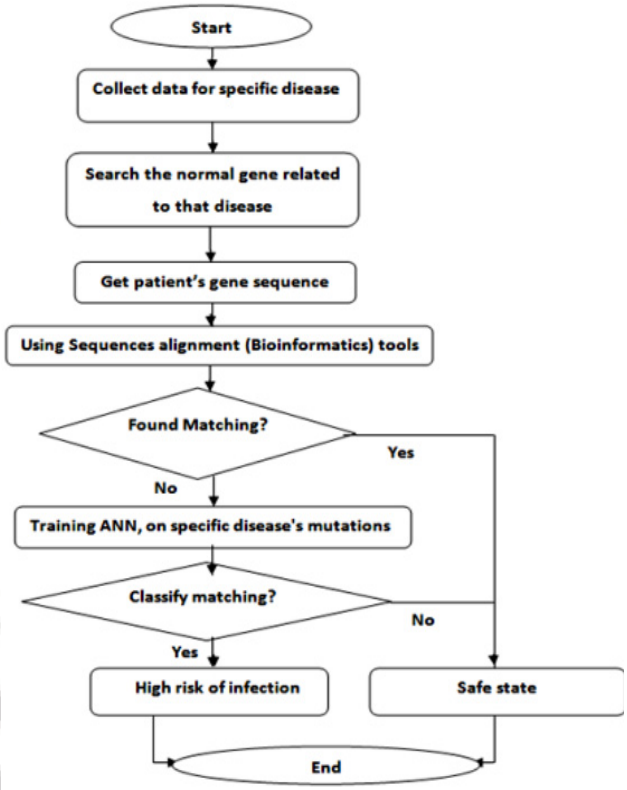
Knowledge discovery is a *process*, and not a one-time response of the KDD system to a user's action. As any other process, it has its environment, its phases, and runs under certain assumptions and constraints Fig.1.2 shows typical data sets, activities, and phases in the KDD process [12]. Its principal resource is a database containing a large amount of data to be searched for possible patterns. KDD is never done over the entire database, but over a representative target data set, generated from the large database. In most practical cases, data in the database and in the target data set contain noise, i.e. erroneous, inexact, imprecise, conflicting, exceptional and missing values, as well as ambiguities. By eliminating such noise from the target data set, one gets the set of preprocessed data. The set of transformed data, generated from the preprocessed data set, is used directly for DM. The output of DM is, in general, a set of patterns, some of which possibly represent discovered knowledge



Phases in the KDD process

1.3 Clustering

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur. This paper presents an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners.



Flowchart of Novel method

2. Review of Literature

Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Data mining techniques form a group of heterogeneous tools and techniques and are used for different purposes. These techniques and methods are based on statistical techniques, visualization, machine learning, etc. Data mining algorithms try to fit a model closest to the characteristics of data under consideration. These Models can be descriptive or predictive [6]. Descriptive models are used to identify patterns in data, clustering, association rules, and visualization are some are the tasks of descriptive modeling

2.1 Micro Array Technology

Micro array technology provides a tool for estimating expressions of thousands of genes simultaneously. Many supervised learning methods have been proposed with the steps as follows: Firstly, three-fourth of the samples of data is used to train the Classifier and secondly, the trained classifier is used to predict or test the one-fourth of the samples. The challenges in such problem were discussed in as 1.Scarce of training data, 2.High dimensionality. The answer to the challenge lies in predicting cancer by using a small subset of important genes from wide collection of gene expression data. With thousands of genes and small amount of samples ranking the genes according to their importance in contributing to classifier's prediction strength is a crucial problem.

Microarray technology has provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment. The vast amount of raw gene

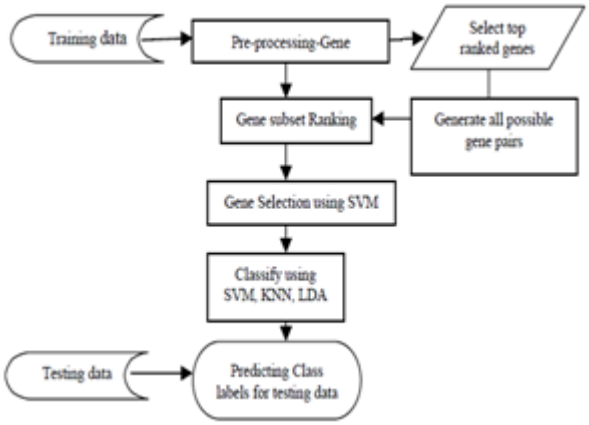
expression data leads to statistical and analytical challenges. One challenge area in the studies of gene expression data is the classification of the expression dataset into correct classes. The goals of classification are to identify the differentially expressed genes that may be used to predict class membership for new samples. First, supervised classification identifies a set of genes that can differentiate different classes of samples by using the training dataset with known classes. Then, the selected set of discriminative genes, or predictive genes, is used to identify the class of unknown samples.

2.2 Ovarian cancer

Ovarian cancer is particularly lethal with a long-term survival rate of only 29% [13]. The current biomarker that is used for detecting the cancer's presence is correlated with tumor volume. Thus, the cancer remains undetected at its early stage, where the cure rate is high, for a large number of patients. Petricoin et al. [13] applied genetic algorithm and clustering techniques to analyze 100 equally distributed training samples (i.e., 50 cancer and 50 normal) with 15,154 genes each (Table 1). The coding scheme for genetic algorithm was the logical chromosomes while the fitness function was the ability of a logical chromosome to specify a lead cluster map (i.e., generates homogeneous clusters). Their analysis resulted in 97.4% prediction accuracy when applied to 116 separate test samples. Five significant genes (M/Z values 534.82277, 989.15067, 2111.7119, 2251.1751, 2465.0242) were identified as ovarian cancer indicators.

2.3 Noval Method

The noval method for mutational disease prediction using bioinformatics tools and datasets for diagnosis the malignant mutations with powerful Artificial Neural Network (Backpropagation Network) for classifying these malignant mutations are related to gene(s) (e.g. BRCA1 and BRCA2) cause a disease (breast cancer). This noval method didn't take in consideration just like adopted for dealing, analyzing and treat the gene sequences for extracting useful information from the sequence, also exceeded the environment factors which play important roles in deciding and calculating some of genes features in order to view its functional parts and relations to diseases. This paper is proposed an enhancement of a novel method as a first way for diagnosis and prediction the disease by mutations considering and introducing multi other features show the alternations, changes in the environment as well as genes, comparing sequences to gain information about the structure/function of a query sequence, also proposing optimal and more accurate system for classification and dealing with specific disorder using backpropagation with mean square rate 0.000000001



3. Methodology

The collection datasets relate to gene caused a disease is very important stage in the proposed of enhancement a noval method, because the noval method didn't care about it, however as show it can play an important role for mutational disease predication, as referring to section 1. This proposal of enhancement a noval method focus on the important feature in Bioinformatics tools which is BLAST to reach to best Homology sequence related to the environment which is different from region to another. The algorithm of main tasks for enhancement a novel method of maturational disease prediction, which based on overcome the drawbacks at novel method, this enhancement will refer to it in red as follow:

- Algorithm of main tasks for Enhancement a Novel method
- Input: DNA sequence (normal and person's gene sequence)
- Output: Classification of tumor gene mutations (for the patient)

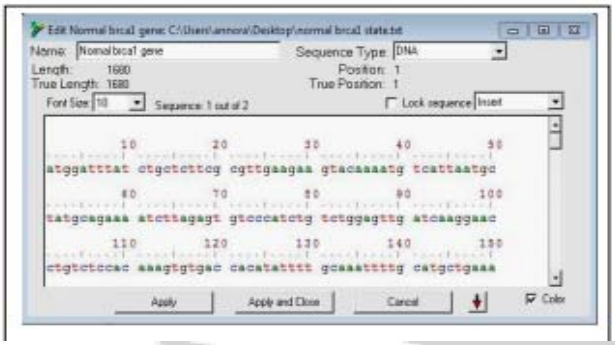
3.1 Bioinformatics Techniques

Implement the proposed enhancement of noval method will apply to diagnosis the mutations at patient gene and its protein is malignant or not. In order to know how to access and get helpful bioinformatics tools this modified approach to reach an enhancement of a novel method consider the collection dataset related to gene caused the disease is very important stage in the pro-posed of enhancement a noval method, because the noval method which not care about it for mutational disease predications, so in this stage will consider additional parts as follow:

1. The data set adapted to depend on the environment, as it plays roles in alternations and differences in genes shapes also genes mutations.
2. The researchers or biologists adopt a normal gene for analysis and comparison should be compatible with the environment of the study. The improvement of this point can be shown in Fig. 2. The normal gene can be obtained from NCBI or EBI, Ensemble, COS-MIC. and each of these databases provide different form of normal gene adopted in the local database but when search for homology sequences at NCBI, the 100% match found between the two normal gene sequences. When normal gene of COSMIC is entered to NCBI at blast it shows the

Refuses (normal) as the first one in homology list with max identities=100.

3. Calculating GC%content and AT% content for gene stability and ability of doing amplification can be done by using BioEdit package to determine ratios and some amino acid and protein features. Needed to calculate GC% and AT% of normal BRCA1 gene to determine is the genome around 38% GC to depend it otherwise search about another normal BRCA1 gene form another database, shown in Fig. 3.1, Fig. 3.2 and Fig. 3.3 respectively.

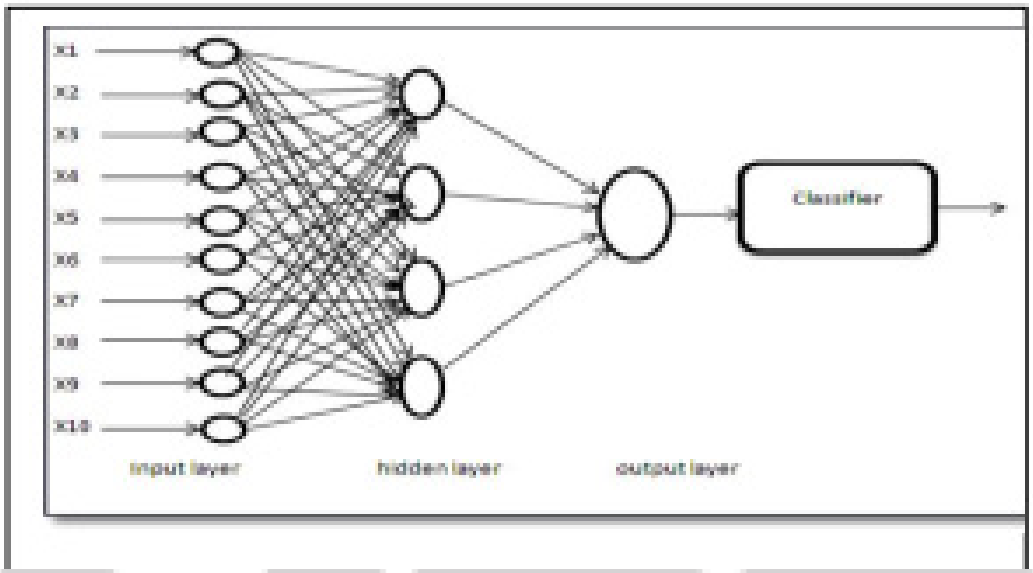


Gene base positions

3.2 Back propagation Algorithm

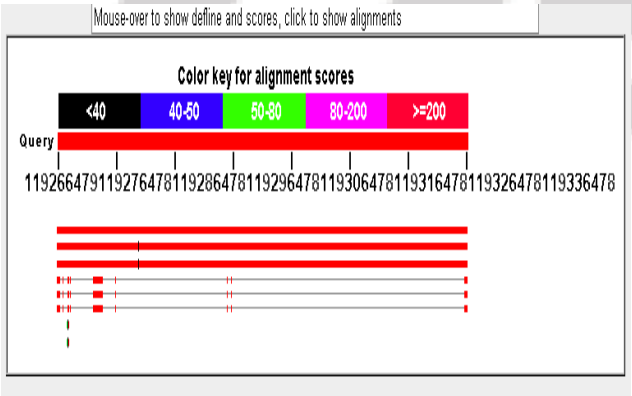
The Back Propagation Algorithm is a multi-layered Neural Networks for learning rules [4], credited to Rumelhart and McClelland. It produces a prescription for adjusting the initially randomized set of synaptic weights such that to maximize the difference between the neural network's output of each input fact and the output with which the given input is known (or desired) to be associated. Back propagation is a supervised learning algorithm and is mainly used by Multi-Layer- perceptron to change the weights connected to the net's hidden neuron layer(s).

The back propagation algorithm uses a computed output error to change the weight values in backward direction [12]. To get this net error, a forward propagation phase must have been done before. The neurons are being activated using the sigmoid activation function while propagating in forward direction



New Structure of BPN

4. Experimental Results



Distribution of hits on the query sequence

5. Conclusion

This work is our preliminary research on predicting survival times for cancer patients based on clinical data, demo graphic data, blood test results, and weight-loss assessment. We combined all data set and use feature selection to extract useful attributes. We also tried mixture of Gaussians to divide patients into several survival groups. We solve this prognosis problem with both regression and classification approaches, each approach contains one or more models. Our experiment shows that regression and classification are both prone to error, and there is no perfect solution for this data set thus far.

However, we found that feature selection is helpful in selecting relevant attributes, and a combination of feature

selection and mixture of Gaussian give us the best result. Hence, a statistics show that the expert suggests is not necessarily the best choice. Even if the current model cannot be put into practice today, any improvement is a great contribution to our future study later.

6. Future Work

Obvious feature work includes discovering a more accurate and appropriate model and determining whether our model can be better than what is practicing in the hospital. Also, it will be better if we can gather more available data for analysis, which will definitely, improve both the consistency of our data set and the usability of our results. For future works, we are looking to try different kind of feature selection, e.g., using Lasso algorithm. The drawback with lasso is it cannot reduce number of features to less than number of data points. However, this does not apply tour data set as we have more data points comparing to number of features. The other future work is applying neural network for the prediction similar to the approach discussed in [3]. However, setting the parameters in the neural network is always really hard. Finally, it worth to try to use semi-supervised learning algorithm instead of removing the unlabeled data from the data set.

Reference

[1] Bellaachia, A., and Guven, E. Predicting Breast Cancer Survivability Using Data Mining Techniques.
[2] Bittern, R., Dolgobrodov, D., Marshall, R., Moore, P., Steele, R., And Cuschieri, A. Artificial Neural Networks In Cancer Management. E-Science All Hands Meeting 19 (2007),
[3] Djebbari, A., Liu, Z., Phan, S., And Famili, F. International Journal Of Computational Biology And Drug Design (Ijcbdd). 21st Annual Conference on Neural Information Processing Systems (2008).
[4] Figueiredo, M., And Jain, A. Unsupervised Learning Of finite Mixture Models. Ieee Transactions On Pattern Analysis And Machine Intelligence 24 (2002),
[5] Zupan, B., Demsar, J., Kattan, M., Beck, R., and Bratko, I. Joint European Conference On Artificial Intelligence In Medicine And Medical Decision Making. 21st Annual Conference On Neural Information Processing Systems 1620 (1999),

Author Profile

Manjula Devi.B received her B.Sc(CS) and M.Sc(CS) from Sri Ramalinga Sowdambigai College of Science & Commerce, Coimbatore Affiliated to Bharathiar University, Coimbatore, Tamil Nadu, India in 2005 & 2007 respectively. Currently pursuing her M.Phil in computer science at Sri Ramalinga Sowdambigai College of Science & Commerce, Coimbatore affiliated to Bharathiar University, Coimbatore, Tamil Nadu, and India.

Annakkodi P.S Working as Assistant Professor in the Department of Information Technology, Sri Ramalinga Sowdambigai College of Science & Commerce, Coimbatore. Having 12 years of teaching experience in the specializations Computer science and information technology and guided several PG and M.Phil Projects.