

A Study on Computational Process in Gene Expression Data

E.Monica Sushil Cynthia^{*1} VairaprakashGurusamy^{#2} S.kannan^{#3}

¹Department of MCA, American College, Madurai, Tamilnadu, India.

¹Monicasushil777@yahoo.com

^{2,3}Department of Computer Applications, Madurai Kamaraj University, Madurai, Tamilnadu, India.

²vairaprakashmca@gmail.com

³skannanmku@gmail.com

Abstract---This context is commenced to examine the various methods and its challenges in Disease Identification of Gene Expression Data. The elemental responsibility of these techniques is classification and categorization of gene expression, analysis of the expression, Pattern Recognition, and Identification. This provides an inclusive survey of Micro Array Data analysis techniques and intends a processing component for disease identification. For the healthcare provider, it is essential to maintain the quality of data because this data is useful to provide cost effective healthcare treatments to the patients. Health Care Administration retains the Microarray data which is refined by expertise and is analyzed by the expertise to identify the disease. This process of analyzing this Microarray data as manual is complicated in identification and classification; due to this Microarray data some difficulties such as missing information, empty values, and incorrect entries. Exclusive of quality information there is no valuable consequences. For successful data mining, an impediment in health data is individual the major difficulty for examining medical data. So, it is essential to maintain the quality and accuracy data for data mining to making an effective decision. The major goal of this survey is focused on various techniques of data mining for developing a prediction model for disease susceptibility using Gene Expression Data. The microarray data is pre-processed to analyze the gene expression to classify the over-expression and under-expression data. Then the classified gene data is then clustered and the best feature selection is applied to discover a pattern. Finally, the association mining handled under the organized set of the gene expression data to the identification of the disease. This context provides efficient techniques to overcome the manual identification of diseases.

Keywords: Gene Expression Data, Disease Identification, Classification, Association Mining, Pattern Recognition

1. Introduction

DNA microarrays propose the capability to appear at the expression of thousands of genes in a particular research one of the significant relevance of microarray knowledge is disease identification and classification. Through microarray technology, researchers will be proficient in organizing special diseases according to dissimilar expression intensity in common and development cells, to determine the affiliation among genes, to recognize the critical genes in the development of disease [1]. The main task of microarray classification is to construct a classifier from chronological microarray gene expression data, and then it utilizes the classifier to categorize prospecting data. Appropriate to the rapid improvement of DNA microarray knowledge, gene ranking techniques and organization techniques are being figured for enhanced use of classification algorithm in microarray gene expression data. The study of oversized gene expression data sets is fetching a dispute in disease classification [2]. Thus gene selection is one of the significant characteristics. Proficient gene selection can considerably simplify computational burden of the consequent classification assignment and can yield a much smaller and more condensed gene set, not including the defeat of classification. In classifying microarray data, the key objective of gene selection is to explore for the genes, which remain the greatest amount of information about the set and decrease the categorization error. [3] Data mining techniques classically descend into either supervised or unsupervised classes. Microarray technologies afford a dominant tool by which the expression prototypes of thousands of genes can be examined concurrently whose relevance collection from disease diagnosis to treatment response. Gene expression is the renovation of the DNA progression into mRNA progression by dictation then transformed into amino acid sequences called *proteins*.

The key challenge in classifying gene expression data is the annoyance of dimensionality difficulty. There is a huge amount of genes (features) evaluated to small sample sizes [3]. To conquer this, feature selection is worn to recognize differentially articulated genes and to eliminate inappropriate genes. Gene selection remains a significant task to extend the exactness and speed of classification structures. In general, feature selection can be prepared into three kinds: Filter, Wrapper, and Embedded methods. They are classified based on how a feature selection method merges with the production of a classification form. An extensive quantity of literature has been available on gene selection techniques for construction of a valuable classification model. In this paper, we

present a review of feature selection techniques for disease identification and classification. In this study, it has been focused on Microarray gene association analysis from an association mining approach.

2. Challenges in Analyzing Microarray Data

Numerous challenges in microarray require to be atreatywithpreviousto new information about gene expression can be exposed. Some of the problems are:

1. Microarray data is high dimensional data illustrated by thousands of genes in only someexample sizes, which reasonimportant problems such as inappropriate and blast genes, difficulty in creating classifiers [4], and severalmisplaced gene expression values due to indecentexamining. In addition, most of thestudies that functional microarray data are endured from information overfitting, which involvesextrasupport.
2. Mislabeled records or difficult tissues product by specialist also another form of negative aspect that could reduce the accuracy of experimental results and direct to anindefinite conclusion about gene expression patterns [4, 5].
3. Biological relevanceproduct is another essentialcondition that should be inuse into adescription in consideringmicroarray data rather than only focusing on theexactness of disease classification [6]. Even though there is no suspiciongaining high accuracy classification consequences are significant in microarray data study, but revealing thegenetic information through the development of categorization is also important.
4. Cross-platform evaluation of gene expression studies iscomplex to perform when microarrays were build using specialprinciples. Thus, the consequences cannot be imitated.

3. Processing Components

Disease Identification and classification is based on microarray data analysis classical methods like pre-processing, clustering, feature selection, pattern discovery, classification and association mining. In following sections, surveys of existing methods are analyzed. In the remaining sections, the importance of processing modules is discussed.

4. Overview of the existing techniques in disease identification

Feature selection methodsin excess of DNA microarray focus on filter techniques. The majority of the proposed techniques are univariate i.e. Each feature is measured individually. The featuresignificanceachieve is considered, and low attaining features are distant [8]. The top ordered genes are used to construct the classifier. Following are the feature selection methods.

(a) Information Gain (IG)

Entropy assessment is used in Information Gain, Gain Ratio, and Symmetrical improbability aspect grading methods [8]. Entropy of Y is

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

$p(y)$ is the trivial probability soliditypurpose for unsystematic variable Y. the provisional entropy of Y after examining 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 provisional probability of yspecified x. The information gained regarding Y after surveying X is

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

Information gain is symmetrical measure,

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

(b) Gain Ratio (GR)

In order to expect Y standardize information gain (IG) byisolating it by theentropy of X. The Gain_ratio is

$$Gain_ratio = \frac{IG}{H(X)}$$

The gain ratio is an asymmetrical evaluate. Gain_ratio falls in the range [0, 1] since of normalization. Gain_ratio=1 specify that X entirely predicts Y and Gain_ratio=0 designate that X and Y are independent.

(c) Fisher's Criteria

The equation used to rank the genes in fisher criteria is

$$fisher(g) = \frac{(m_1(g) - m_2(g))^2}{s_1^2(g) + s_2^2(g)}$$

m_1 and m_2 stand for the mean expression value of the j^{th} gene in excess of all samples in cancer and standard casing respectively. s_1 and s_2 indicate the standard deviation of the j^{th} gene over all samples in exaggerated and common case respectively [9].

(d) Clustering and Network analysis based Technique

Shang Gao, Omar Addam and colleagues in [10] projected these two feature selection systems. The proposed clustering based process uses Genetic Algorithm (GA) approach.

(e) Support Vector based Correlation Coefficient (SVcc)

This method chooses vector data spots using support vector machine (SVM). These preferred data points are extra used for ranking the genes using association coefficient. The top ordered genes are used for categorization.

Classification and clustering are in common measured as alike; the only variation is classification is a supervised learning technique whereas clustering is an unsupervised learning progression. In classification, the category label of each training tuple is recognized in progress hence called as supervised learning. The classifier constructs using the known instances (training set) to effectively forecast the group of the new instance (test set). The exactness of the classifier is resolute as the proportion of test tuples that are properly classified by the classifier from the test set. In other words, the classification is all about forecasting the class of the new instance by learning from identified instances and their class labels. KNN, DT, SVM, NN, NB are classification technique [4], [6], [8].

5. Microarray Data Analysis

Microarray data is the center of a revolution in biotechnology. It is used to monitor the expression of tens of thousands of genes at the same time. Hence it is used to accomplish many genetic tests in parallel [10]. The result of a microarray experiment is a list of genes that are found to be differentially expressed in special kinds of tissues. The microarray dataset can be a variety of an $M \times N$ matrix D of expression values, where the row stand for genes g_1, g_2, g_3, \dots , in and the column correspond to different experimental conditions $s_1, s_2, s_3, \dots, s_n$. Each aspect $D[i, j]$ represents the expression stage of the gene g_i in the sample s_j . The matrix typically holds a large amount of data, so data mining techniques are used to extract useful information.

6. Pre-processing

Data pre-processing in microarray expertise is an essential preliminary step previous to data investigation is achieved. Various pre-processing techniques have been projected but nothing has confirmed perfect to date. Regularly, datasets are inadequate by laboratory restraints so that they require a strategy on feature and toughness, to notify advanced testing although data are yet classified. Another goal of preprocessing is to "clean" the raw data. The measured intensities are not only influenced by the actual RNA abundance but also by other sources of variation. Commonly [10, 11], preprocessing microarray data is a three-step procedure: background correction, normalization, and summarization. Therefore, many researchers developed alternative algorithms for each of the three preprocessing steps.

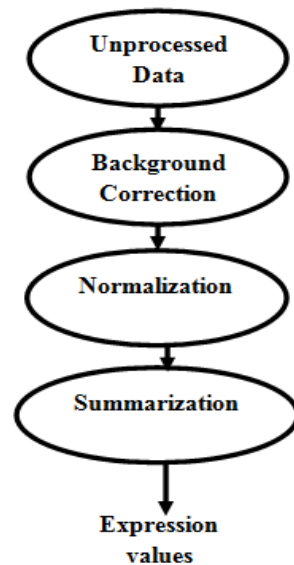


Figure 2: Microdata Pre-processing

The algorithms are implemented such that every step is self-contained, i. e. every method for one step can combine with every method for another step opening a large playground for all the combinatory. The number of possible combinations is even higher since the first three steps are optional; only summarization is mandatory to complete the preprocessing.

a) *Background Correction*

Background correction methods estimate the background portion of the probe signals and subtract it accordingly. In the case of DNA arrays, the background can be estimated from the area surrounding and separating the spots.

b) *Normalization*

In general, the initial transformation useful to expression data, referred to as normalization, regulates the entity hybridization intensities to stability them suitably so that significant genetic assessment can be made [11]. There is a quantity of basis why information must be standardized, including irregular measured of preliminary RNA, distinctions in classification or recognition efficiencies involving the luminous colorant used, and methodical biases in the considered expression stages. Abstractly, normalization is related to regulating expression levels considered by northern investigation or quantitative repeat transcript relative to the expression of one or more reference genes whose levels are assumed to be constant between samples.

c) *Summarization*

The process of reducing multiple measurements on the same gene down to a single measurement by combining some manner.

7. Expression Analysis

The standard genetic approach for investigating biological typically begins by recognizing transformations that source a phenotype of importance [14]. Over-expression or Under-expression of a wild-type gene product, though, can also basis distorted phenotypes, afforded that geneticists with another yet influential device to categorize components that may reside after disregarded using established loss-of-function study. The most admired two types of Differential gene expression conferred as follows

1. Statistical tests

- a) Statistical t-test: a two illustration position test of the null suggestion that the θ means of two usually circulated populations are equivalent.
- b) Welch's t-test: asymmetrical variation
- c) Mann-Whitney U test (also called Wilcoxon rank-sum test): θ nonparametric

2. U-test

- a) Robustness: U-test is further tough to outliers
- b) Effectiveness: When familiarity holds, the effectiveness of the U-test is as regards 0.95 when evaluated to the t-test. For distributions adequately distant from standard and for suitably huge sample sizes, the U-test can be greatly more proficient than the t-test.

8. Clustering

However, the huge number of genes and the difficulty of genetic seriously enhance the challenges of understanding and inferring the resultant group of data, which frequently consists of millions of dimensions [14, 15]. An initial step to addressing this disputing is the use of clustering methods, which is important in the information mining procedure to expose ordinary structures and recognize a motivating pattern in the essential data. Cluster analysis inquires about to separation a given data set into collections based on particular features so that the data ends in a group are more related to every other than the ends in dissimilar groups [15]. An extremely rich literature on cluster analysis has developed more than the past three decades.

Clustering techniques have established to be supportive to appreciate gene function, gene regulation, cellular processes, and subtypes of cells. Genes with related expression prototypes (co-expressed genes) can be clustered collectively with related cellular functions. Moreover, co-expressed genes in the equivalent cluster are probably to be concerned in the same cellular progressions and a strong association of expression patterns among those genes specifies co-regulation. Penetrating for general DNA sequences at the supporter expanses of genes in the same cluster permits regulatory motifs detailed to every gene cluster to be recognized and regulatory basics to be proposed.

Table 2: Clustering Techniques Comparison Table

| Techniques | Advantages | Disadvantages |
|-------------------------------------|---|---|
| K-Means Clustering | -Less Time Complexity -Less Space Complexity -Order-independent | -More iteration for optimal no. of clusters -Sensitive to noise when huge noise occurs |
| Hierarchical Clustering | -Embedded Flexibility -Ease of handling of any forms -More versatile | -Difficult to modify -Changes may cause huge complexity -High Computational Complexity |
| Self-organizing map(SOM) | -Combination of several map units allows the construction of non-convex clusters -Measuring and joining criteria I useful for big clusters -Sub-optimal partition provided if the initial weights are not properly chosen | -Not much effective if the patterns are merged -Not effective in highly noisy data |
| Expectation-maximization Clustering | -Robust to noisy data -Handle high dimensional data -Converges fast -Linear in data size -Accept desired no. of clusters | -Provide less accuracy in some noisy data -Very sensitive for noise - Low quality results when using random dataset |

(a) *K-Means Clustering*

The k-means algorithm [15] is one of the most extensively used methods for clustering. It begins by initializing the k cluster midpoint, where k is resolute prior to clustering. Then, each object (input vector) of the data set is assigned to the cluster whose center is the nearest.

(b) *Hierarchical Clustering*

Partitioning algorithms are based on identifying an original number of collections, and iteratively reallocating objects between sets to convergence. In distinguishing, hierarchical algorithms merge or split existing sets, generating a hierarchical construction that reproduces the classification in which groups are combined or divided. This method iterates in anticipation of all objects are in a particular group. The differential alternative of hierarchical clustering algorithms may use special rate purposes.

(c) *SOM (Self-Organization Map)*

Stimulated by neural networks in the brain, SOM uses an opposition and collaboration mechanism to attain unsupervised learning. In the traditional SOM, a set of nodes is approved in an arithmetic pattern, classically 2-dimensional pattern. Every node is related to a weight vector with the identical aspect as the input space. The intention of SOM is to locate a superior mapping from the high dimensional input space to the 2-D depiction of the nodes. One approach to using SOM for clustering is to observe the items in the input space

signify by the same node as assembled into a cluster. All through the training, every object in the input is accessible to the map and the best identical node is identified.

(d) *Expectation-maximization Clustering*

The EM algorithm is an allocation-based clustering algorithm. Distribution-based clustering algorithms suppose that objects are formed according to a prospectsupply. Special clusters can be measured created according to different possibility distributions. For each entity, themostprobability of an object belonging to anexplicit cluster is figured. A number of clusters, k, isresoluteproceeding to clustering. EM is measured one of the most admired distribution-based clustering algorithms.

9. Feature Selection

There are someconditions that can delay the development of feature selection, such as the existence of unrelated and unnecessary features, noise in the data or interfaceamong attributes. In the presence of hundreds or thousands of features, such as DNA microarray investigation, researchers perceivethat is general that a huge number of features is not revealingsince they are either unrelated or unnecessary with esteem to the class perception. Furthermore, when the number of features is high excluding the number of examples is undersized, machine learning gets mainlycomplexbecause the search space will be lightlyinhabited and the reproduction will not be capableof distinguishingproperly thesignificant data and the noise [9, 10].

In addition,the classification, feature selection techniques can also be categorized into three methods: filters, wrappers, and embedded techniques. By means of such anenormous body of feature selection methods, the demandoccurs to find out various criteria that permit users to sufficientlymake a decision which algorithm to use in assured situations [7]. This work evaluates several feature selection techniques in the prose and ensures their concert in a simulatedprohibited experimental situation, distinct the capacity of the algorithms to choose the appropriate features and to get rid of the irrelevant ones not including authorizing noise or redundancy to hinder this development.

Table 3: Feature Selection Methods Summary


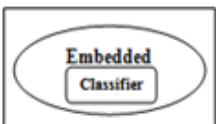
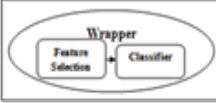
| Methods | Description | Advantages | Disadvantages | Techniques |
|---|--|---|--|--|
| Filter methods  | -Rely on the general characteristics of training data -Carry out the feature selection process as a pre-processing step with independence of the induction algorithm. | -Independence of the classifier -Lower computational cost than wrappers -Fast -Good generalization ability | No interaction with the classifier | -The Consistency-based Filter -Information Gain -ReliefF -CFS |
| Embedded  | -Perform feature selection in the process of training -Usually specific to given learning machines. | -Interaction with the Classifier -Lower computational Cost -Captures feature Dependencies | Classifier-dependent Selection | -FS-Perceptron -SVM-RFE |
| Wrapper  | involve optimizing a predictor as a part of the selection | -Interactive -Feature Dependencies | -Expensive -Risk of over fitting -Classifier dependent selection | -Wrapper SVM -Wrapper-C4.5 |

Table 3 provides a summary of the characteristics of the three feature selection techniques, representing the mainly prominent advantages and disadvantages, as fine as various examples of every method that will be promote xplicated. Within filters, one can discriminate among *univariate* and *multivariate* processes [13]. **Univariate** techniques are prompt and scalable, although disregard feature dependencies. On the other hand, **multivariate** filters customs feature dependencies, but at the rate of being slower and take away scalable than univariate methods.

Table 4: Feature Selection Methods Evaluation

| Techniques | Description | Advantages | Disadvantages |
|---|---|--|---|
| Information Gain(IG) | -This filter provides an ordered ranking of all the features and a threshold value is required -Univariate filter method | -Scalable -Fast -Classifier independent | Ignores feature dependencies |
| Correlation-based feature selection (CFS) | -That ranks feature subsets according to a correlation-based heuristic evaluation function. -Multivariate filter method | -Redundant features should be screened out - Irrelevant features should be ignored -Faster computation | Classifier-independent Selection |
| mRMR (minimum Redundancy Maximum Relevance) | Selects features that have the highest relevance with the target class and are also minimally redundant | - Eliminates redundancy -Tests the relevance of genes in combination with other genes | Time complexity more as compared to filter methods |
| Embedded methods (SVM-RFE Recursive Feature Elimination for Support Vector Machines, FS-P Feature Selection—Perceptron) | -SVM embedded method performs feature selection by iteratively training a SVM classifier with the current set of features - A perceptron is a type of artificial neural network that can be seen as the simplest kind of feed-forward neural network | -Tests the predictive power of genes -Less computational complexity compared to wrapper method -Less prone to over-fitting | Heavily dependent on the model, so they can fail to fit the data well |
| Wrapper methods (SVM, Wrapper-C4.5) | Evaluates attribute sets by using a learning scheme. Cross-validation is used to estimate the accuracy of the learning scheme for a set of attributes. | Carries out exhaustive search, generating optimal solutions | -Exponential time complexity -Complex -Doesn't take enough measures to eliminate redundancy -Prone to Over-fitting |

While filter techniques indulge the difficulty of finding a good feature separation separately of the form collection step, wrapper methods implant the form suggestion exploration within the feature subset search. In this association, a search process in the space of probably feature subsets is distinct, and assorted subsets of features produce and estimated. The assessment of a precise subset of features is acquired by training and testing a detailed classification model, interpretation this approach customized to a specific classification algorithm [13, 18]. Though, as the break of feature subsets develops exponentially with a number of features, heuristic search techniques are used to conduct the search for the best subset. These explore methods can be separated into two classes: deterministic and randomized search algorithms.

10. Classification

Classification segregates data sections into objective classes. The classification method forecasts the intention set for every data points. For instance, the patient can be classified as a high threat or low threat patient on the source of their disease prototype using data classification approach. It is a supervised learning approach having known class categories. Binary and multi-level are the two techniques of classification. [16]. Dataset is divided as training and testing dataset. By means of training dataset trained the classifier. The exactness of the classifier possibly will be tested using test dataset. Hu et al. used special classification scheme such as decision tree, SVM and ensemble approach for investigating microarray data [17]. Further, utilize of the classifier in the health field is discussed by Hatice et al., to analysis the skin diseases with weighted KNN classifier [22].

The survey work exposed that there is no only best algorithm which defers improved effect for each dataset. Classification methods are also used for expecting the behavior rate of healthcare services which is enhanced with speedy development every year and is fetching a key anxiety for everyone. Following are the various categorization algorithms used in healthcare:

Table 5: Classification Techniques Evaluation Table

| Techniques | Advantages | Disadvantages |
|-------------------------|---|---|
| K-NN | It is easy to implement. Training is done in faster manner. | 1. It requires large storage Space. 2. Sensitive to noise. 3. Testing is slow. |
| Decision Tree | It minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. It can easily process the data with high dimension. It is easy to interpret. | 1. It is restricted to one output attribute. 2. It generates categorical Output. 3. It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset. 4. If the type of dataset is numeric than it generates a complex decision tree |
| Support Vector Machine | 1. Better Accuracy as compare to other classifier. 2. Easily handle complex nonlinear data points. 3. Over fitting problem is not as much as other methods. | 1. Computationally expensive. 2. The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. 3. As compare to other methods training process take more time. |
| Neural Network | 1. Easily identify complex Relationships between dependent and independent variables. 2. Able to handle noisy data. | 1. Local minima. 2. Over-fitting. 3. The processing of ANN network is difficult to interpret and require High processing time if there are large neural networks. |
| Bayesian Belief Network | 1. It makes computations process easier. 2. Have better speed and accuracy for huge datasets. | 1. It does not give accurate Results in some cases where there exists dependency among variables. |

(a) *K-Nearest Neighbour (K-NN)*

K-Nearest Neighbour (K-NN) is one of the simplest classifier that notices the unrevealed data end by means of the previously recognized data points and confidential data points according to the selection system. K-NN orders the data points using additional than one nearest neighbor. K-NN has a number of functions in dissimilar regions such as health datasets, image domain, cluster study, pattern identification, online advertising etc. Shuman et al. used K-NN classifier for investigating the patients suffering from heart disease. The statistics were composed of UCI and experimentation was executed using exclusive of selection or with selection K-NN classifier and it is established that K-NN attains improved exactness without selection in the analysis of heart diseases as evaluate to with selection K-NN.

(b) *Decision Tree (DT)*

DT is related to the flowchart in which all non-leaf nodes stand for an analysis on a particular feature and every division represents a conclusion of that test and each leaf node has a category label. The node at the top most labels in the tree is known to be the root node. Constructing a decision for every difficulty doesn't require any kind of field knowledge. Decision Trees is a classifier that uses trees similar to the graph. Khan et al., used decision tree for forecasting the survivability of breast cancer patient and Chien et al., proposed a widespread hybrid decision tree classifier for categorizing the movement of a patient having persistent disease.

(c) *Support Vector Machine (SVM)*

The support vector machine (SVM) classifier generates an overexcited level surface or several hyperplanes in the high dimensional gap that is valuable for classification, degeneration and other proficient responsibilities. SVM have lots of attractive features appropriate to this it is gaining esteem and has a capable experimental recital. SVM builds an overexcited level surface in unique input space to split the data ends. An amount of time it is complicated to achieve partition of data points in original input space, so to construct partition easier the innovative restricted dimensional space plotted into original privileged dimensional space.

Soliman et al. used SVM categorization approach for classification of a range of diseases and SVM mutually with k-means clustering were functional on microarray data for recognizing the diseases. SVM is one of the most accepted approaches that are used by the researcher in healthcare domain for classification.

(d) *Neural Network (NN)*

This is an algorithm for categorization that utilizes a descent method and based on a biological nervous system having numerous consistent handling out essentials identified as neurons, carrying out in concord to a task explicitly. Regulations are extracted from the educated Neural Network (NN) assist in developing interoperability of the educated network. To resolve a particular difficulty NN used neurons which are structured handling out elements. Neural Network is used for classification and model recognition. An NN is adaptive in nature since it alters its formation and regulates its power in order to reduce the fault. A collection neural network method is projected by Das et al., for analysis of heart disease in organize to extend the efficient decision support system.

(e) *Bayesian Methods*

The classification based on Bayes premise is acknowledged as Bayesian classification. It is a straightforward classifier which is accomplished by using classification algorithm. Bayes theorem supplies source for Naive Bayesian Classification and Bayesian Belief Networks (BBN). The major crisis with Naive Bayes Classifier is that it is supposed that all characteristics are self-regulating with each other whereas in medical field attributes such as patients symptoms and their fitness status are interrelated with each other. Bayesian Belief Network is extensively used by numerous researchers in the healthcare area.

11. Association Mining

Gene expression data necessitates a few steps of data processing prior to it can be investigated for association rules. In market basket analysis an entry is either obtained or not obtained but microarray data includes of uninterrupted statistical data. The initial step is to discretize the data, renovate it to a Boolean or tertiary record. The majority of the request of mining association rule on microarray gene expression still relies on discretization responsibilities prior to pertaining any data mining method. The standardized microarray dataset is frequently signifying as sequences of constant numbers [19]. Discretization is the progression of conversion from continuous data into distinct data.

The threshold process used to discretize the information. This technique is appropriate for microarray study. Genes with chronic expression values superior to a fastidious value are considered as overexpressed, if not as below expressed. By means of threshold method, every gene expression is transformed into one of the two separate values 1, 0 for over-expressed and under-expressed. Association rules are a significant group of techniques of pronouncement samples in data. Association rule mining method remove appealing associations along with a collection of items (genes) in a huge quantity of information. One of the most illustrious purposes of these techniques is market basket analysis [19, 20] where the key goal is to discover interactions between the acquired items under different operation. An association rule is functional on microarray dataset in regulate to locate the associations between genes under different illustration.

12. Conclusion

In this paper, we analyzed various image analysis techniques used in bioinformatics sector to identify and classify the diseases using microarray data. The system based on association mining and classification for categorize the diseases from the defected genes. In connection with this, important phases such as microarray data pre-processing, clustering, feature selection, classification and association mining. Here microarray data is attained by bioinformatics then pre-processed using normalization and summarization. Pre-processed data are further used to analyze the expression data using expression analysis tests such as statistical tests, U-test. The analyzed gene expression is clustered using clustering techniques like K-means, SOM or Hybrid method. The feature selection is the main objective to predict the various diseases using this gene expression. Finally, the threshold value is applied on association rule mining to predict and classify the diseases using microarray data. One can extend this work to generate most significant feature vectors for efficient and accurate classification.

References

- [1] Jun S. Liu Department of Statistics Harvard University "Bioinformatics: Microarrays Analyses and Beyond".
- [2] Alireza Osareh and Bitad Shadgar "Microarray Data Analysis for Cancer classification" IEEE Antalya, Turkey 2009, pp.125-132.
- [3] R. D. Canlas Jr., "Data Mining in Healthcare: Current Applications and Issues", (2009).
- [4] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
- [5] F. Hosseinkhah, H. Ashktorab, R. Veen, M. M. Orange O., "Challenges in Data Mining on Medical Databases", IGI Global, (2009), pp. 502-511.
- [6] K. Srinivas, B. Kavitha Rani, and Dr. A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal of Computer Science and Engineering, vol. 02 no. 02, (2010), pp. 250-255.
- [7] C. Shang and Q. Shen, "Aiding classification of gene expression data with feature selection: a comparative study." vol. 1, 2005, pp. 68-76.

- [8] Piyushkumar A. Mundra and Jagath C. Rajapakse “Support Vectors Based Correlation Coefficient for Gene and Sample Selection in Cancer Classification” IEEE 2010.
- [9] Yvan Saeys, Inaki In and Pedro Larranaga “A review of feature selection techniques in bioinformatics” 2005 pp 1–10.
- [10] K. Blekas, Nikolas P. Galatsanos, and Ioannis Georgiou, “an unsupervised artifact correction approach for the analysis of DNA microarray images”, IEEE, 2003.
- [11] Jianhua Xuan, Eric Hoffman, Robert Clarke, Yue Wang, “Normalization of Microarray Data by Iterative Nonlinear Regression”, IEEE conference on IBBE, 2005.
- [12] Bhanot, G., et al. (2006) A robust meta-classification strategy for cancer detection from MS data. *Proteomics*, 6, 592–604
- [13] Blanco, R., et al. (2004) Gene selection for cancer classification using wrapper approaches. *Int. J. Pattern Recognition. Artif. Intell.*, 18, 1373–1390.
- [14] Dudoit, S., et al. (2002), Comparison of discriminant methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97, 77–87.
- [15] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A., — Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci.*, 1999, 96 (12), 6745–6750.
- [16] Antonov, A.V., Tetko, I.V., Mader, M.T., Budczies, J., Mewes, H.W., — Optimization models for cancer classification: extracting gene interaction information from microarray expression data, *Bioinformatics* 20, 2004, 644–652.
- [17] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, — A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, Vol. 21, 2005, No. 5, pp 631–643.
- [18] A. E. Akagi, A. Amine, A. E. Ouardighi, D. Aboutajdine, — Feature selection for Genomic data by combining filter and wrapper approaches, *INFOCOMP Journal of computer science*, 2009, vol. 8, no. 4, pp. 28–36.
- [19] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;95:14863–8.
- [20] Azadeh Mohammadi, Mohammad H Saraei, Mansoor Salehi, “Identification of disease-causing genes using microarray data mining and Gene Ontology” Mohammadi et al. *BMC Medical Genomics* 2011.
- [21] Shang Gao, Omar Addam and colleagues, “Robust Integrated Framework for Effective Feature Selection and Sample Classification and Its Application to Gene Expression Data Analysis” IEEE 2012 pp. 112–119.
- [22] One Huey Fang, Norwati Mustapha, Md. Nasir Sulaiman “Integrating Biological Information for Feature Selection in Microarray Data Classification” 2010 Second International Conference on Computer Engineering and Applications IEEE, pp. 330–334.