Volume 6, Issue 3, March 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Survey on Biomarkers Classification in Data Mining

**S. B. Ishwarya*, R. Porkodi, (Asst. Professor)**
CS Department, Bharathiar University,
Tamilnadu, India

***Abstract—Data mining is the set of techniques and procedure of analyzing data from different ways and summarizing it into useful information. Classification is a data mining technique is used to classify each item in a set of data into one of predefined set of classes. This paper presents four data mining classification algorithms SVM, KNN, Random Forest, and Artificial Neural Network used in biomarkers classification. These four algorithms are the most significant data mining algorithms in the research of data mining. One of the main goals of microarray data analysis is monitoring of gene expression for thousands of genes in equivalent and producing large amounts of valuable data and also detection of biological knowledge, for example metabolic pathways. The papers also focus on biomarkers in bioinformatics and classification of biomarkers. Biomarkers is an assessable indicator of a particular biological state, typically one related to the risk, presence, severity, prognosis therapeutic reaction of disease. Even though physiological metrics like height or blood pressure are biomarkers the term is known as molecular biomarkers. The paper also focuses on different types of biomarkers such as prognostic biomarkers, predictive biomarkers, pharmacodynamic biomarkers, and surrogate endpoints.***

***Keywords— Bioinformatics, Microarray Dataset, Biomarkers, classification of Biomarkers, Support Vector Machine, K-Nearest Neighbors, Random Forest, Artificial Neural Network.***

## I. INTRODUCTION

Data mining has involved a great deal of interest in the information industry and in society as a complete in recent years, due to the large accessibility of vast quantity of data and the imminent want for turning such data into useful information and knowledge. Discovery of knowledge from this vast quantity of data is a challenge indeed. Data mining is an attempt to make sense of the information sudden increase in this huge volume of data [1].Data mining refers to extracting or "mining" knowledge from large quantities of data. Thus, data mining should have been more properly names knowledge mining from data. Knowledge mining, a shorter expression, may not replicate the importance on mining from large amount of data. Mining is a bright term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. It should be valid to any kind of data warehouse, as well as to passing data, such as data stream. It can be viewed as an effect of the natural development of information technology. Depending on the type of data, the data mining system may also included techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, web technology, bioinformatics etc. It constitutes following functions namely classification, regression, clustering, summarization, association rules, etc. In this study, I have further discussed about the following i.e., classification in data mining, data mining in bioinformatics, microarray dataset and biomarkers respectively.

The paper organized as follows: section I describe the introduction of data mining, section II describe the literature review, section III describe classification of biomarkers, section IV describe classification algorithm of biomarkers, section V describe comparison of biomarker classification algorithm and finally the paper is concluded in section VI

### *A. Classification in Data Mining*

Data mining tasks can be classified into two types: descriptive and predictive. Descriptive mining tasks discriminate the general properties of the data in the database. Predictive mining tasks perform assumption on the current data in order to build predictions. Classification is one type of data analysis that can be used to extract model describing important data classes or to predict future data trends [2]. It based on association rule mining is explored. Classification is the procedure of finding a model that explain and distinguished data classes, for the principle of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the examination of a set of training data. Classification patterns permit us to classify values of target variables from attribute variables. The TABLE I show the classification records based on attribute values in balloon dataset.

TABLE I CLASSIFICATION OF BALLOON DATASET

| Attribute Variables | | | | | Target Variable |
|---|---|---|---|---|---|
| **S.N** | **Color** | **Size** | **Act** | **Age** | **Inflated** |
| 1 | Red | Small | Extend | Adult | T |

 

| | | | | | |
|---|---|---|---|---|---|
| 2 | Red | Small | Extend | Child | T |
| 3 | Red | Small | Dip | Adult | T |
| 4 | Red | Small | Dip | Child | T |
| 5 | Red | Large | Extend | Adult | T |
| 6 | Red | Large | Extend | Child | F |
| 7 | Red | Large | Dip | Adult | F |
| 8 | Red | Large | Dip | Child | F |
| 9 | Green | Small | Extend | Adult | T |
| 10 | Green | Small | Extend | Child | F |
| 11 | Green | Small | Dip | Adult | F |
| 12 | Green | Small | Dip | Child | F |
| 13 | Green | Large | Extend | Adult | T |
| 14 | Green | Large | Extend | Child | F |
| 15 | Green | Large | Dip | Adult | F |
| 16 | Green | Large | Dip | Child | F |

**IF (Color=Red AND Size=Small) OR (Age=Adult AND Act=Extend), THEN Inflated=T; OTHERWISE, Inflated=F.**

The following relation of the attribute variable, Color, Size, Age and Act with the target variable, inflated taking the value T for true or F for false. This relation permits us to classify a given balloon into definite value of the target variable using a specific value of its Color, Size, Age, ad Act attributes. Therefore, the relation gives us data patterns that allow us to make the classification of a balloon. Even though extract this relation pattern by groping the 16 data records in the dataset, learning such a pattern manually from a much large set of data with noise can be hard task. A data mining algorithm allows us to learn from large data set mechanically.

### *B. Data Mining in Bioinformatics*

Data mining and bioinformatics are fast growing and closely related research frontiers. The various bioinformatics research areas are computation biology, genetics, genomic, system biology, statistical genetics, microbial genetics, etc. A gene is a basic foundation of any living organism. Sequence of genes in a human body represents the mark of the person. The genes are piece of the deoxyribonucleic acid (DNA). Phosphate and deoxyribose sugar molecules joined together by covalent bonds [3]. Nitrogenous base is attached to each sugar molecular. There are four bases: adenine [A], cytosine[C], guanine [G] and thymine [T]. The DNA can be measured as series of symbols. Each symbol is one of the four bases A, C, G, or T. Entire stretch of the DNA is called genome of an organism. Classically, a DNA series may be base pair long. Such lengthy stretch of DNA is broken up into small fragments. These fragments are sequenced experimentally, and then reassembled simultaneously to reconstruct the novel DNA sequence. Understanding what parts of the genome encode which genes is a main area of study in computational molecular biology or bioinformatics [4, 5].

### *C. Microarray Dataset*

Microarray dataset means collecting data from microarray chip. In this electronic chip consists of thousands of holes. Each and every hole filled with same or different kind of genes. Then applied hundreds of different experiment on that gene simultaneously. Finally the result of gene expressions for each experiment were collected and stored in dataset format. In modern years there has been a blast in the rate of achievement of biomedical data. Advances in molecular genetics technologies, such as DNA microarrays allow us for the first time to obtain a "global" view of the cell [6, 7]. Investigation of microarrays presents an amount of exclusive challenges for data mining. The first generation of microarray analysis methodologies developed over the last 5 years has demonstrated that expression data can be used in a variety of class discovery or class prediction biomedical problems including those relevant to tumor classification [8, 9].Deoxyribonucleic acid (DNA) microarray Technology provides tools for monitoring the expression levels of huge number of different genes simultaneously [10]. It is possible for the biologists to concurrently evaluate the expressions of thousands of genes in a single experiment by the aid of microarray technologies [11] [12] [13]. A microarray method also plays an important role in personalized medicine because it can be used to identify the individual's unique genetic vulnerability to treat the diseases [14]. Microarrays dataset is one of the latest breakthroughs in experimental molecular biology, which allow monitoring of gene expression for tens of thousands of genes in parallel and are producing huge amounts of valuable data [15].

### *D. Biomarkers*

Biomarkers are molecular substances in the body that can be used to specify healthy or diseased. These biomarkers can be found in tissue, blood, urine and other body fluids. Biomarkers have played an increasingly important role in drug discovery, understanding the mechanism of action of a drug, investigating efficacy and toxicity signals at an early stage of pharmaceutical development, and in identifying patients likely to respond to treatment. Although physical character or physiological metrics like height or blood pressure are biomarkers, the term is now normally shorthand for molecular biomarkers.

 

## II. LITERATURE REVIEW

Durges K. Srivastava, Lekha Bhambhu [16], The author proposed, a novel learning method, Support Vector Machine (SVM) in this paper, is applied on different data which have two or multi class. The comparative results using different kernel functions linear, polynomial, and sigmoid and RBF for all data samples. The experiment results are encouraging .It can be seen that the choice of kernel function and best value of parameters for particular kernel is critical for a given amount of data. Then the best kernel is RBF for infinite data and multi class.

S.V.N. Vishwanathan, M. Narasimha Murty [17], The author said Simple SVM algorithm works by maintaining a candidate Support Vector set. It uses a greedy approach to pick points for inclusion in the candidate set. To speed up convergence initialized algorithm with the nearest pair of points from opposite classes. Then use an optimization based approach to increment or prunes the candidate Support Vector set. The algorithm makes repeated passes over the data to satisfy the KKT constraints. Simple SVM algorithm currently does not use any kind of kernel cache to reuse kernel computations. Currently investigating methods to speed up the algorithm using some efficient caching scheme.

Vaishali P Khobragade, Dr.A.Vinayababu [18], The author discusses an efficient classification technique was introduced to classify the microarray genes into their specified cancer class type. The performance of the proposed classification technique was analyzed by performing statistical measures in terms of true positive and true negative values and compared with the existing classifier and old statistical features. For the performance analysis process the GA parameters was tuned. The parameter tuning process best and worst case results were analyzed to acquire the better result.

Masahiko Gosho, Kengo Nagashima, YasunoriSato [19], The suggestion of author in this paper was, highlight several important aspects related to study design and statistical analysis for clinical research incorporating biomarkerss. The importance of biomarkerss in medical diagnosis, prevention, and therapy of diseases is increasing. This article provides an overview on the study designs for biomarkers research.

R.Porkodi [20], In this paper the author implemented the five classification algorithms names Naïve Bayes, KNN, CN2, SVM and Random Forest and found the KNN outperforms almost all five algorithms. For target class 2, Random forest algorithm outperforms well than the remaining algorithms during the validation carried by AUC. In the outset, the three algorithms KNN, CN2, Naïve Bayes and Random forest gives better performance and the SVM classification algorithm obtained poor result for this data set.

Vrushali Y Kulkarni,Pradeep K Sinha [21], In this paper the author said, improvement in learning time of Random Forest by proposing a new approach called Disjoint Partitioning. Conclude that this approach works well with datasets that are imbalanced in nature & have binary classification. It reduces learning time notably while achieving comparable accuracy as that of original Random Forest.

Saranya Vani.M, S.Uma, Sherin.A, Saranya.K [22], The author suggested in this paper provides a review of traditional classification techniques used for data mining. The main attention is on classification techniques like decision tree induction, Bayesian networks, rule based classification, k-nearest neighbor classification techniques which are used to mine databases. A comprehensive review is done on the issues, recent advancements and research works on these techniques.

## III. CLASSIFICATION OF BIOMARKERS

An specialist functioning group at the National Institutes of Health (NIH) has defined a biological marker or biomarkers as ‘a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention’ [23]. According to this classification, biomarkers cover a rather large array of data types, for example, biochemistry laboratory tests on blood, function testing, electrocardiographic testing, and image information such as computed tomography (CT), magnetic resonance imaging (MRI) and positron-emission tomography (PET). Typical examples of such biomarkers are listed in Table II [24–31]. Biomarkers can be broadly classified into prognostic biomarkers, predictive biomarkers, pharmacodynamic biomarkers, and surrogate endpoints [26- 32].

### *A. Prognostic Biomarkers*

Prognostic biomarkers classically identify patients with differing risks of a specific outcome, such progression or death [33, 34]. Recently, the prognostic biomarkers were defined as a single trait or signature of traits that separates a population with respect to the outcome of interest, regardless of the types of therapies or treatments [35]. For example, under this description, if particular biomarkers were prognostic, the result of patients with biomarkers-positive status would be improved than that of patients with biomarkers-negative status in both the test and standard treatments. The prognostic biomarkers can decide populations into groups whose result will be poor or good following the test and standard treatments, but it cannot direct the choice of a particular treatment. The preliminary knowledge necessary to propose a validation study of prognostic biomarkers [34].

### *B. Predictive Biomarkers*

Biomarkers predict the differential outcome of a particular therapy or treatment [35]. In addition, Chakravarty *et al.* [36] state that predictive biomarkers are a baseline characteristic which categorizes patients by their degree of response to a particular treatment. In this case, for example, biomarkers-positive patients make moderately better than do biomarkers-negative patients when standard treatment is administered, whereas test treatment may be more successful in the biomarkers-positive group. Currently used predictive biomarkers are, irinotecan-treated patients who were

 

homozygous for the uridine diphosphoglucuronosyl transferase allele had a greater risk of hematologic toxic effects than did patients who had one or two copies of the wild-type allele [37–40]. As this example demonstrates, validated biomarkers can prospectively classify patients who are likely to have a positive clinical outcome from a specific treatment; therefore, predictive biomarkers could direct the alternative of treatment in one of several ways.

### *C. Pharmacodynamic Biomarkers*

According to Jenkins *et al.* [39], when the change in biomarkers is the parameter that is to be understood, explained, or controlled, then the biomarkers is considered an endpoint. The biomarkers could be used in this sense as a biomarkers of the drug activity to display proof of attitude and be used to optimize the dosing plan of the drug during the previous phases of the drug expansion list, while clinical biomarkers are used in clinical trials. Pharmacodynamic biomarkers are indicators of drug result on the object in an organism. The biomarker can be used to check the link between drug treatment, target result and biological tumor response. For example, inflammatory markers such as C-reactive protein (CRP) or erythrocyte sedimentation rate (ESR) may be used to pick a dose in rheumatoid arthritis treatment or can form part of a clinical composite such as disease activity score (DAS) and be used for the same purpose [39].

### *D. Surrogate Endpoints*

A surrogate endpoint is intended to be a substitute for a clinical endpoint. It is expected to predict clinical benefit based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence [41, 42]. In clinical trials, a surrogate endpoint is a evaluate of the result of an assured treatment that may compare with a true endpoint but does not essentially have a guaranteed relationship with it. According to the Biomarkers Working Group [43], a surrogate endpoint is defined as a biomarkers intended to substitute for a clinical endpoint. A clinical researcher uses epidemiological, therapeutic, pathophysiological, or other scientific confirmation to select a surrogate endpoint that is probable to predict clinical advantage destruction, or require of gain or destruction.

TABLE II BIOMARKERSS REPORTED IN LITERATURES

| Biomarkers | Current Use | Classification | Purpose |
|---|---|---|---|
| Human epidermal growth factor receptor 2 (HER2),epidermal growth factor receptor (EGFR),V-Ki-ras2 Kirsten rat sarcoma viral oncogenehomolog (KRAS) mutations [24,26,27,29] | Directing treatment | Predictive biomarkers | Used to classify subpopulations of patients who are most probable to answer to a given therapy |
| BCR-ABL (*i.e.*, Philadelphia chromosome in CML) [27] | Directing treatment of imatinib | Predictive biomarkers | Used to classify subpopulations of patients who are most probable to answer to a given therapy |
| Estrogen receptor (ER) and progesterone receptor (PR) [25,27] | Selection for hormonal therapy | Predictive biomarkers | Used to classify subpopulations of patients who are most probable to answer to a given therapy |
| Promyelocytic leukemia-retinoic acid receptor α(PML/RARα) translocation [30] | Prescribing arsenic trioxide foracute promyelocyticleukaemia | Predictive biomarkers | Used to classify subpopulations of patients who are most probable to answer to a given therapy |
| Amyloid β peptide (AB) 1-42 [24,26] | Diagnosis of prodromal Alzheimer's disease | Prognostic biomarkers | Obliging in selecting patients for adjuvant complete treatment. |
| Gene signature chips (e.g., Oncotype, MammaPrint) [26] | Prognosis prediction in oncology | Prognostic biomarkers | Obliging in selecting patients for adjuvant complete treatment. |
| B-type natriuretic peptide (BNP) [31] | Screening and diagnosis in heart Failure | Prognostic biomarkers | Obliging in selecting patients for adjuvant complete treatment. |
| C-reactive protein (CRP), Interleukin-6 (IL-6),Tumor necrosis factor (TNF-α) in blood samples [24,26] | Proof of principle in inflammatory diseases | Pharmacodynamic biomarkers | Inspect the relation between drug regimen, target effect & biological tumor response. |
| FDG-PET (SUVmax) functional imaging [24,26] | Proof of concept (e.g., in tumourMetabolism) | Pharmacodynamic biomarkers | Inspect the relation between drug regimen, target effect & biological tumor response. |

 

| | | | |
|---|---|---|---|
| Hemoglobin a1c (HbA1c) [26] | Represents glycaemic control in Diabetics | Surrogate endpoint | Measure of effect of a specific treatment that may correlated with a real clinical endpoint |
| Prostate-specific antigen (PSA) [27] | Screening and monitoring in prostate cancer | Surrogate endpoint | Measure of effect of a specific treatment that may correlated with a real clinical endpoint |

## IV. CLASSIFICATION ALGORITHMS OF BIOMARKERS

Classification is the product of applying a machine learning algorithm to recognize the relationship between patters of variables and classes represented in the training data set. Parameters of the classifier are learned from the training data, but the objective is to intend a generalizable classifier-one accomplished of accurate prediction of class memebership for new data point. Support Vector Machine, Random Forests, K-Nearest Neighbors, and Artificial Neural Network methods are commonly used as learning algorithm in biomarkers classification.

### *A. Support Vector Machine*

Support Vector Machines (SVM) has recently gained prominence in the field of machine learning and pattern classification [44]. Its uses a nonlinear mapping to change training data into higher dimension. Inside this new dimension, it searches for the linear optimal separating hyperplane. Let the data set D be given as(**X1,** y1), (**X2**, y2),….., (**Xn,** yn), where **xi** is the set of training tuples with connected class labels yi. Each yi can take one of two values, either +1 or -1, corresponding to the classes, respectively. A straight line can be drawn to separate all of the tuples of class +1 forms all of the tuples of class-1. To find the best one, that is one that will have minimum classification error on previously unseen tuples. A separating hyperplane can be written as

$$\boldsymbol{W\_X+b = 0;}$$

Where $\boldsymbol{W}$ is a weight vector, namely, $\boldsymbol{W} = \{w1, w2, \ldots\ldots, wn\}$; $n$ is the number of attributes and $b$ is a scalar, often referred to as a bias. Training tuples are 2-D, e.g., $\boldsymbol{X} = (x1, x2)$, where $x1$ and $x2$ are the values, respectively, for $\boldsymbol{X}$.
In solving the quadratic optimization trouble of the linear SVM, the training tuples emerge only in the form of dot products, $\Phi(\boldsymbol{Xi}).\Phi(\boldsymbol{Xj})$.kernel function, K(**Xi**, **Xj**), to the original input data. That is,

$$\mathrm{K}(\mathbf{Xi}, \mathbf{Xj}) = \Phi(\mathbf{Xi}).\ \Phi(\mathbf{Xj}).$$

An SVM with a small number of support vectors can have high-quality simplification, even when the dimensionality of the data is high.

### *B. K-Nearest Neighbors*

K-nearest neighbor classifier first locates k data points that are most parallel to the data point as the k-nearest neighbor of the data point and then uses the aim classes of these k-nearest neighbors to verify the aim class of the data point. To find out the k-nearest neighbors of the data point, we must to use a measure of match or contrast between data points. Many measure of match or contrast exist, including the Euclidean distance, the Minkowski distance, the Hamming distance, Pearson's correlation coefficient, and cosine similarity.
The Euclidean distance is defined as

$$\boldsymbol{d(xi, xj)} = \sqrt{\sum_{\boldsymbol{k=0}}^{\boldsymbol{n}} (\mathbf{xi,l} - \mathbf{xj,l})} \quad , \boldsymbol{i \neq j}$$

The Euclidean distance is a measure of dissimilarity between two data point xi and xj.
The Minkowski distance is defined as

$$\boldsymbol{d(xi, xj)} = [\sum_{\boldsymbol{l=1}}^{\boldsymbol{p}} |\boldsymbol{xil - xj, l}|^{\mathbf{r}}]^{\mathbf{1/r}}, \boldsymbol{i \neq j}$$

This distance is also measure of dissimilarity.
Pearson's correlation coefficient is also a distance measure defines as follow

$$\boldsymbol{\rho_{xi\,xj} = s_{xixj}/s_{xi}\ s_{xj}}$$

Where$s_{xixj}$ ,$s_{xi}$ , and $s_{xj}$ are the expected covariance of xi and xj. KNN is an easy to understand and easy to implement classification technique.

### *C. Random Forest*

Random Forest is a classifier consisting of a collection of tree-structured classifiers {*h(x, Θk) k=1, 2, ….*},
Where the {Θ*k* } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x* [46]. Random Forest generates an ensemble of decision trees. To achieve diversity among base decision trees, Breiman selected the randomization approach which works well with bagging or random subspace methods [45], [46], [47]. To produce each single tree in Random Forest Breiman following steps: If the number of records in the training set is N, then N records are sampled at random but with substitution, from the original data, this is bootstrap sample. This sample will be the training set for growing the tree. If there are M input variables, a number m << M is selected such that at each node, m variables are selected at random out of M and the best split on these m attributes is used to split the node. The value of m is held constant during forest growing. Each tree is grown to the largest extent possible. There is no pruning.

 

The Generalization error (PE*) of Random Forest is given as,

$$PE^* = P_{x,y}(mg(X, Y)) < 0$$

Where mg(X, Y) is Margin function. The Margin function procedures the extent to which the average number of votes at (X, Y) for the correct class exceeds the average votes for any other class. Here X is the predictor vector and Y is the classification. The Margin function is given as,

$$mg(X, Y) = avk\, I(hk(X) = Y) - \max_{j \neq Y} avk\, I(hk(X) = j)$$

Here I (.) is Indicator function. Margin is directly proportional to assurance in the classification.
Strength of Random Forest is given in terms of the predictable value of Margin function as

$$S = E_{X,Y}(mg(X, Y))$$

The generalization error of ensemble classifier is bounded above by a function of mean correlation between base classifiers and their average strength (s) [48]. If ρ is mean value of correlation, an upper bound for generalization error is given by,

$$PE^* \leq \rho(1 - s2) / s2$$

Hence, to yield better accuracy in Random Forest, the base decision trees are to be diverse and accurate.

### *D. Artificial Neural Network*

Artificial neural network [49] is organization (network) composed of a number of interconnected units (artificial neurons).Each unit has an input/output (I/O) attribute and implements a restricted computation or function. The output of any unit is determined by its I/O characteristic, its interconnection to other units, and external inputs. ANN does not form one network, but a various relations of networks. In general function or functionality achieved is determined by the network topology. The overall computational form consists of a reconfigurable interconnection of simple elements, or units. The neurons [50] are implicit to be arranged in layers, and the neurons in the same layer perform in the same manner. All the neurons in a layer typically have the same activation function.

(***1) Input layer:*** The neurons in this layer accept the external input signals and execute no computation, but basically transfer the input signals to the neurons in another layer.

***(2) Output layer:*** The neurons in this layer collect signals from neurons either in the input layer or in hidden layer.

***(3) Hidden layer*:** The layer of neurons that are connected between the input layer and the output layer is known as hidden layer. Neural nets are often classified as signal layer networks or multilayer networks.

*1) Single Layer Network*

A single layer network [50] consists of one layer of connection weights. The net consists of a layer of units called output layer from which the response of the net can be obtained. This type of network can be used for pattern classification problems.

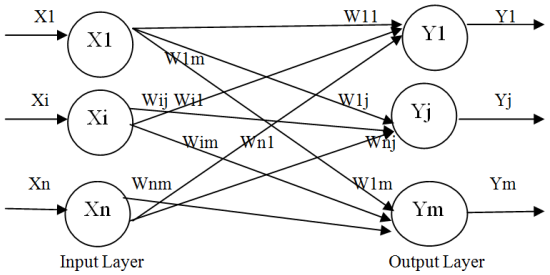


Figure 1: Single Layer Neural Network

*2) Multilayer Network*

A multilayer network [50] consists of one or more layers of unit (called hidden layers) between the input and output layers. Multilayer networks may be created by simply cascading a collection of layers; the output of one layer provides the input to the following layer. A multilayer net with nonlinear activation function can solve any type of problem.

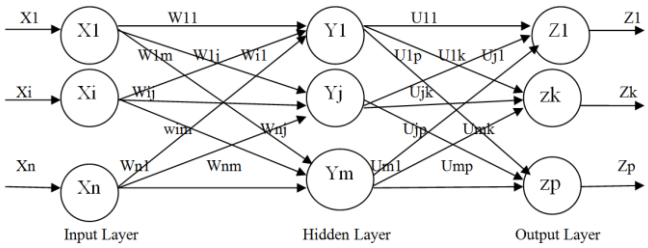


Figure 2: Multilayer Neural Network

*3) Activation Function*

The principle of nonlinear activation function is to guarantee that the neuron's response is enclosed. In order to get the advantages of multilayer network compared with the restricted capabilities of single layer networks, nonlinear function are required. The various activation functions are

Identity Function (Linear Function):

 

**f(x) = x for all x.**

Binary Step Function:

**f(x) =1 if x ≥ θ**

**f(x) = 0 if x < θ**

Sigmoidal Function:

$$\mathbf{f(x)} = \frac{1}{1+e^{-x}}$$

Sigmoidal function is extremely accepted because it is repetitive, bounded and has a simple derivative, f'(x) = f(x) [1-f(x)].A logistic or a binary sigmoid function with a range from 0 to 1.

Bipolar Sigmoid Function:

$$\mathbf{f(x)} = \frac{2}{1+e^{-x}} \mathbf{-1}$$

Bipolar sigmoid is used as an activation function when the desired range of output value is between -1 and +1.

## V. COMPARISON OF CLASSIFICATION ALGORITHMS USED IN BIOMARKERSS CLASSIFICATION

The extensive survey has been conducted in classification algorithms in data mining for biomarkers classification. The outcome of the survey produces the comparison of various classification algorithms based on experimental dataset used, accuracy of their research and demerits are listed in Table III.

TABLE III CLASSIFICATION ALGORITHMS COMPARATIVE TABLE

| Author Name & Year | Dataset | Algorithms/Methods | Accuracy | Demerit |
|---|---|---|---|---|
| Yu Shi , Daoqing Dai, Chaochun Liu, Hong Yan & 2009 [51] | Breast cancer dataset | Sparse diagonal discriminant analysis | High classification accuracy 54.59 | Exploring the application of the L1 penalized technique to other classifiers and other penalty terms |
| Rolf Søkilde, Martin Vincent, Anne K. Møller, Alastair Hansen, Poul E. Høiby,Thorarinn Blondal, Boye S. Nielsen, Gedske Daugaard, Søren Møller, and Thomas Litman & 2014 [52] | miRNA expression data set. | Least absolute shrinkage and selection operator (LASSO) algorithm | Classification accuracy 88% | Trimming of the number of discriminatory miRNAs, and prospective clinical trials |
| XavierRobin, Natacha Turck, Alexandre Hainard , Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez,Markus Müller & 2013 [53] | aSAH data set | Iterative combination of Biomarkerss and thresholds(ICBT) Method | Threshold sensitivity, specificity & Area under the ROC curve 95% ,90% & 95% | Classification power of the resulting panel is superior to that of single biomarkers. But, to be strictly validated these findings need to be replicated in larger, independent cohorts of patients. |
| Manli Zhou,Youxi Luo, Guoquan Sun ,Guoqin Mai,and Fengfeng Zhou & 2014 [54] | GSE5406,GSE1869 data set | SVM, Naïve Bayes, DTree, LASSO, KNN | Classification Accuracy, (SVM)1.000, (NBayes)0.998 | Feature Selection Based on Constraint Programming (FsCoP) improves all the other three feature selection algorithms. |
| Manju R Mamtani, Tushar P Thakre, Mrunal Y Kalkonde ,Manik A Amin, Yogeshwar V Kalkonde, Amit P Amin and Hemant Kulkarni & 2006 [55] | OvCa, LuMe, LLML, BrCa data sets | Proposed Statistical Algorithm, KNN, SVM | Diagnostic accuracy, (Proposed Statistical Algorithm)100 % | Diagnostic accuracy of other algorithms was not high. |
| Malik Yousef,Mohamed Ketany,Larry Manevitz, Louise C Showe and Michael K Showe & 2009 [56] | CTCL(I), CTCL(II), Lymphocyte dataset | SVM-RNE, SVM-RCE and SVM-RFE | CTCL(I):100%, 96, 89% CTCL(II):91%, 76, 84% Lymphocyte:80 %,96%, 81% | Expression level of gene on multipath ways lack in accuracy |
| Shaoning Pang, Ilkka Havukkala, and Nikola Kasabov & 2006 [57] | Cancer dataset | 2-SVMT algorithm | Classification accuracy 80.7% | Data was not reuse in correct classification way |

## VI. CONCLUSION

The data mining and bioinformatics are two interdisciplinary research areas. Biomarkers are used to analysis molecular reaction which helps to identify the diseases and the patient treatment. The paper presents the biomarkers such

 

as Prognostic Biomarkers, Prognostic Biomarkers, Pharmacodynamic Biomarkers and Surrogate Endpoints and also presents four types of classification algorithm such as Support Vector Machine, K Nearest Neighbors, Random Forest and Artificial Neural Network used in biomarkers classification. This paper also presents the comparative study on the above mentioned classification algorithms based on their accuracy and demerits. The outcome of the study shows that the Support Vector Machine classification algorithms give better accuracy than other algorithms in biomarkers classification. Autism spectrum disorder is a serious neurodevelopment disorder or group of complex disorders of brain development. These disorders are characterized by following symptoms such as difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors. Disorders are mostly occurred on gene expression in blood of children. Identifying these symptoms in gene expression is quite necessary. This extensive survey gives better idea to extend the research to identify the Children autism spectrum disorder using biomarker classification techniques.

**REFERENCES**


[1] P.Piatetsky-Shapiro & W.J.Frawley,eds.,KDD. *Menlo Park ,CA:AAAI/MIT Press*, 1998.

[2] Jiawei Han, Micheline Kamber, *Data Mining:Concepts and Techniques*, Elsevier Inc., 2006

[3] J.D.Watson and F.H.*Crick proposed a structure of DNA* in 1953.

[4] S.L.Salzberg, D.B.Searls, & S.Kasif, eds., *Computational Methods in Molecular Biology*. Amsterdam: Elsevier Science B.V., 1998.

[5] ”Special Issue on Bioinformatics, Part I: advances and Challenges,” *Proceedings of the IEEE*, Vol.90, Nov 2002.

[6] Chipping Forecast 1999, 2002, *The Chipping Forecast.Special Supplement*. Nature Genet. 21, Jan. 1999.

[7] Hegde P. et al. *A concise guide to cDNA microarray analysis. Biotechniques*.2000Sep;29(3):548-50, 552-4, 556.

[8] Tamayo P. and S. Ramaswamy.“Cancer Genomics and Molecular Pattern Recognition” in Expression profiling of human tumors:diagnostic and research applications. *Marc Ladanyi and William Gerald eds. Humana Press* (2003).

[9] Golub T.. *Genome-Wide Views of Cancer*, N Engl J Med 2001; 344:601-602.

[10] James J. Chen and Chun-Houh Chen, "Micro array Gene Expression", *Encyclopedia of Biopharmaceutical Statistics*, 2nd Edition, Marcel Dekker, Inc., pp. 599-613, 2003

[11] Yuh-Jye Lee and Chia-Huang Chao, "A Data Mining Application to Leukemia Micro array Gene Expression Data Analysis", *International Conference on Informatics, Cybernetics and Systems* , Kaohsiung, Taiwan, 2003

[12] Seeja and Shweta, "Microarray Data Classification Using Support Vector Machine", *International Journal of Biometrics and Bioinformatics (IJBB*), Vol. 5, No. 1, pp. 10-15, 2011

[13] Yee Hwa Yang and Natalie P. Thorne, "Normalization for Two-color cDNA Microarray Data", *Science and Statistics: A Festschrift for Terry Speed*, Vol. 40, pp. 403- 418, 2003

[14] Mark A. Iwen, Willis Lang and Jignesh M. Patel, "Scalable Rule-Based Gene Expression Data Classification", *In Proceedings of the IEEE 24th International Conference on Data Engineering*, 2008

[15] V.Bhuvaneswari1 and .Vanitha2 REF.*Classification of Microarray Gene Expression Data by Gene Combinations using Fuzzy Logic (MGC-FL)*

[16] Durges K. Srivastava, Lekha Bhambhu., *Data Classification Using Support Vector Machine*., JATIT, 2005 -09.

[17] S.V.N. Vishwanathan, M. Narasimha Murty SSVM : *A Simple SVM Algorithm.*, Dept. of Comp. Sci. and Automation, Indian Institute of Science, Bangalore.

[18] Vaishali P Khobragade,Dr.A.Vinayababu *A Classification of Microarray Gene Expression Data Using Hybrid Soft Computing Approach*.International Journal of Computer ScienceVol 9,Issue 6,Nov2012,.ISSN:1694-0814.

[19] Masahiko Gosho , Kengo Nagashima, Yasunori Sato.*Study Designs and Statistical Analyses for Biomarkers Researc.* Sensors 2012, 12, 8966-8986; doi:10.3390/s120708966, ISSN 1424-8220

[20] Dr.R.Porkodi.A Study on Performance *Analysis of Data Mining Classification Algorithms over Lung Cancer Dataset.* International Journal of Research in Information Technology,Vol 2,Issue 3,March 2014, ISSN 2001-5569.

[21] Vrushali Y Kulkarni,Pradeep K Sinha,.*Efficient Learning of Random Forest Classifier using Disjoint Partitioning Approach*. Proceedings of the World Congress on Engineering 2013 Vol II,WCE ,July 3 - 5,2013, U.K.

[22] Saranya Vani.M, S.Uma, Sherin.A, Saranya.K, *Survey on Classification Techniques Used in Data Mining and their Recent Advancements*,International Journal of Science, Engineering and Technology Research, Vol 3, Issue 9, Sept 2014 , Pg:2380-2385,ISSN: 2278 – 7798

[23] Biomarkerss Definitions Working Group. Biomarkerss and surrogate endpoints: Preferred definitions and conceptual framework. *Clin. Pharmacol. Ther*. 2001, 69, 89–95.

[24] Frank,R.;Hargreaves,R.*Clinical biomarkerss in drug discovery and development.Nat.Rev.Drug Discov*.2003, 2,.

[25] Hayes, D.F.; Bast, R.C.; Desch, C.E.; Daniel, F.; Fritsche, H.; Kemeny, N.E.; Jessup, J.M.; Locker, G.Y.; Macdonald, J.S.; Mennel, R.G.; et al. *Tumor marker utility grading system: A framework to evaluate clinical utility of tumor markers*. J. Natl. Cancer Inst. 1996, 88, 1456–1466.

[26] Jenkins,M.;Flynn,A.;Smart,T.;Harbron,C.;Sabin,T.;Ratnayake,J.;Delmar,P.;Herath,A.;Jarvis,P.;Matcham,J.;On behalf of the PSI Biomarkers Special Interest Group.*A statistician’s perspective on biomarkerss in drug development.* Pharm. Stat. 2011, 6, 494–507.


 

[27] Ludwig, J.A.; Weinstein, J.N. *Biomarkerss in cancer staging, prognosis and treatment selection. Nat. Rev. Cancer* 2005, 5, 845–856.

[28] Mega,J.L.;Close,S.L.;Wiviott,S.D.;Shen,L.;Hockett,R.D.;Brandt,J.T.;Walker,J.R.;Antman,E.M.;Macias,W.;Braunwald,E.;Sabatine,M.S.*CytochromeP-450polymorphisms and response to clopidogrel*.N.Engl.J.Med.09,354–362.

[29] Simon, R. *The use of genomics in clinical trial design.* Clin. Cancer Res. 2008, 14, 5984–5993.

[30] Wang,Z.Y.;Chen, Z. *Acute promyelocytic leukemia:From highly fatal to highly curable.Blood*.2008, 1, 2505-15.

[31] Berger, R.; Huelsman, M.; Strecker, K.; Bojic, A.; Moser, P.; Stanek, B.; Pacher, R. *B-type natriuretic peptide predicts sudden death in patients with chronic heart failure*. Circulation 2002, 21, 2392–2397.

[32] Buyse, M.; Michiels, S.; Sargent, D.J.; Grothey, A.; Mateson, A.; de Gramont, A. *Integrating biomarkerss in clinical trials.* Expert Rev. Mol. Diagn. 2011, 11, 171–182.

[33] Hayes, D.F.; Trock, B.; Harris, A.L. *Assessing the clinical impact of prognostic factors: When is "statistically significant" clinically useful? Breast. Cancer.* Res. Treat. 1998, 52, 305–319.

[34] Simon,R.;Altman,D.G.Statistical aspects of prognostic factor studies in oncology.*Br.J.Cancer* 1994,69,979–985.

[35] Sargent, D.J.; Conley, B.A.; Allegra, C.; Collette, L *Clinical trial designs for predictive marker validation in cancer treatment trials*. J. Clin. Oncol. 2005, 23, 2020–2027.

[36] Chakravarty, A.G.; Rothmann, M.; Sridhara, R. *Regulatory issues in use of biomarkerss in oncology trials.* Stat. Biopharm. Res. 2011, 3, 569–576.

[37]……Ando,Y.;Saka,H.;Ando,M.;Sawa,T.;Muro,K.;Ueoka,H.;Yokoyama,A.;Saitoh,H.;Shimokata,K.;Hasegawa,Y.*Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity:A pharmacogenetic analysis.Cancer* Res. 2000, 60, 6921–6926.

[38] Innocenti,F. Undevia,S.D.;Iyer,L.;Chen,P.X.;Das,S.;Kocherginsky,M.;Karrison,T.;Janisch,L.;Ramírez,J.;Rubin, C.M.;et al.*Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of irinotecan.* J. Clin. Oncol. 2004, 22, 1382–1388.

[39] Marcuello, E.; Altés, A.; Menoyo, A.; Del Rio, E.; Gómez-Pardo, M.; Baiget, M. *UGT1A1 gene variations and irinotecan treatment in patients with metastatic colorectal cancer.* Br. J. Cancer2004, 91, 678–682.

[40] Rouits,E.;Boisdron-Celle,M.Dumont,A.;Guerin,O.;Morel,A.;Gamelin,E.*Relevance of different UGT1A1 polymorphisms in irinotecan-induced toxicity:A molecular and clinical study of 75 patients.Clin.Cancer* Res.2004, 10, 5151–5159.

[41] Ellenberg, S.S.; Hamilton, J.M. *Surrogate endpoints in clinical trials: Cancer. Stat. Med.* 1989, 8, 405–413.

[42] Prentice,R.L.*Surrogate endpoints in clinical trials: Definitions and operational criteria.Stat*. Med.1989,8,431-440.

[43] *Biomarkerss Definitions Working Group. Biomarkerss and surrogate endpoints: Preferred definitions and conceptual framework*. Clin. Pharmacol. Ther. 2001, 69, 89–95.

[44] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2nd edition, 2000.

[45] Breiman L, Bagging Predictors , *Technical report* No 421, (1994)

[46] Brieman L, *Random Forests, Machine Learning*, 45, 5-32, (2001)

[47] Opitz D,Maclin R,*Popular Ensemble Methods:An Empirical Study,Journal of Artificial Intelligence* 11.

[48] Prenger R, Lemmond T, Varshney K, Chen B, Hanley W, *Class-Specific Error Bounds for Ensemble Classifiers, KDD*'10, Washington DC, USA, (2010)

[49] Robert J.Schalkoff , *Artificial Neural Networks*., McGraw-Hill International Editions.,1997

[50] Dr.S.N.Sivanandam and Dr.M.Paulraj, *Introduction to Artificial Neural Network.*,Vikas Publishing House Pvt Ltd., 2003.

[51] Yu Shi , Daoqing Dai, Chaochun Liu , Hong Yan. *Sparse discriminant analysis for breast cancer biomarkers identification and classification.*Elsevier,. Progress in Natural Science 19 (2009) 1635–1641.

[52] Rolf Søkilde, Martin Vincent, Anne K. Møller,Alastair Hansen,Poul E. Høiby,Thorarinn Blondal,Boye S. Nielsen,Gedske Daugaard,Søren Møller, and Thomas Litman *Efficient Identification of miRNAs for Classification of Tumor Origin.*The Journal of Molecular Diagnostics, Vl. 16, No. 1, January 2014.

[53] XavierRobin,NatachaTurck,AlexandreHainard,Natalia Tiberti,Frédérique Lisacek,Jean-Charles Sanchez,Markus Müller.PanelomiX:*A threshold-based algorithm to create panels of biomarkerss.*Translational Proteomics I.2013.

[54] Manli Zhou,Youxi Luo,Guoquan Sun,Guoqin Mai,and Fengfeng Zhou. Constraint Programming Based Biomarkers Optimization. *Hindawi Publishing Corporation.BioMed Research International*.Volume 2015, Article ID 910515, 5 pages.

[55] Manju R Mamtani,Tushar P Thakre,Mrunal Y Kalkonde,Manik A Amin, Yogeshwar V Kalkonde, Amit P Amin and Hemant Kulkarni. *A simple method to combine multiple molecular biomarkerss for dichotomous diagnostic classification.* BMC Bioinformatics 2006, 7:442,doi:10.1186/1471-2105-7-442

[56] Malik Yousef,Mohamed Ketany,Larry Manevitz, Louise C Showe and Michael K Showe. *Classification and biomarkers identification using gene network modules and support vector machines*. BMC Bioinformatics 2009, 10:337 ,doi:10.1186/1471-2105-10-337.

[57] Shaoning Pang, Ilkka Havukkala, and Nikola Kasabov.*Two-Class SVM Trees (2-SVMT) for Biomarkers Data Analysis.* J.Wang et al.:ISNN 2006, LNCS 3973, pp. 629–634, 2006.c©Springer-Verlag Berlin Heidelberg,2006

 