RESEARCH ARTICLE OPEN ACCESS

An Efficient Survey on various Data Mining Classification Algorithms in Bioinformatics

Ch Narasimha Chary[1], A Krishna[2], N Abhishek[3], Dr R P Singh[4]

[1, 2] (Rescarch Scholer at Sri Satya Sai University of Technology & Medical Sciences, Bhopal, Mp)

[3](Asst Professor, Sree Dattha Institute of Engineering and Science)

[4](Research Supervisor at Sri Satya Sai University of Technology & Medical Sciences, Bhopal, Mp)

Abstract—

Data mining is utilized to separate the information from a lot of data. Data mining consist of two models, they are prescient and unmistakable. Classification is one of the data mining methods which go under prescient model. Classification is utilized as a part of many applications, for example, counterfeit intelligent, machine learning, insights and database framework. Data mining can be connected to these issues, to enhance the productivity of frameworks and the outlines of machines. This paper surveyed the some calculation gives the best outcome. The analysts utilized diverse classification calculation in which are to be specific K-Nearest Neighbor classifiers, Decision tree, Bayesian system, Support Vector Machine, Artificial Neural Networks. This paper additionally introduced the comparison of each of the five algorithms utilized as a part of Bioinformatics Research.

Key words: Classification, Decision tree, Bayesian network, k- nearest neighbor classifier, Support vector machine, artificial neural network

I. INTRODUCTION

Classification one of data mining errand which is utilized to foresee the qualities. In classification ought to have two classes and that classes are predefined. The input of the classification display is the quality of test data and the yield is which data test belongs to the class. Classification is the partition or ordering of articles into classes. In this strategy the classes are predefined and that will train the classification framework to allot items to the classes. The training depends on training test and that example will contains a set example data. In this training, test data classes are as of now known. In this classification systems testing and validatation puts an important part. Classifying the test data and comparing the outcome with the obscure outcome can determine the exactness.

A microarray database is a store of contains microarray quality expression data. The key employments of a microarray database are to store the degree data, manage an accessible index, and make the data accessible to different applications for analysis and interpretation. The models that determine to take care of an issue are delegated Predictive and Descriptive. Microarray innovation has turned out to be one of the significant devices that many scientists use to monitoring genome in wide expression levels of qualities in a given organism. A microarray is regularly a glass slide on to which DNA atoms are settled in a methodical manner at particular locations called spots. Classification is the way toward finding a model that portrays and distinguish data classes or concepts. The motivation behind this model used to anticipate the class of items and whose class name is obscure.

Bioinformatics is a combination of atomic science and software engineering. In this innovation the PCs are accustomed to storing, extracting, organizing, analyzing, interpreting and integrate natural and hereditary information. Bioinformatics is important for the utilization of identifying human maladies and genomic information. It manages algorithms, databases and information frameworks, web innovations, manmade brainpower and delicate computing, information and computation hypothesis, delicate engineering, data mining, picture processing, modeling and simulation, single processing, discrete science, control and framework hypothesis, circuit hypothesis, and insights. Bioinformatics produces new information and the computational apparatuses are likewise used to make that learning.

The paper organized as takes after: section 1 portrays the introduction on data mining and microarray, section 2 depicts the writing survey, section 3 depicts the various classification algorithms, section 4 portrays comparison of classification algorithms and finally the paper is concluded in section 5.

II. LITERATURE REVIEW

S.Archana and Dr. K.Elangovan et al. [1] discuss the classification algorithms can be implemented on different types of data sets like data of patients, financial data according to performances. Hence these classification techniques show how a data can be determined and grouped when a new set of data is available. On the basis of the performance of these algorithms can also be used to detect the natural disasters like cloud bursting, earth quake, etc.

David B.fogel et al. [2] had presented to develope, breast cancer by using neural network technique and the related works are also used in breast cancer diagnosis based on back propagation method with multilayer perception. In contrast to back propagation found that evolution computational method and algorithms were used often, perform more classic optimization techniques.

Shadab Adam Pattekari et al. [3] developed a prototype Heart Disease Prediction System (HDPS) using Decision Trees, Naive Bayes and Neural Networks. In this system user answers the predefined questions. Then it retrieves hidden data from stored database and it compares the user's values with trained dataset.

Endo et al. [4] had implemented common machine learning algorithms to predict survival rate of breast cancer patient. Logistic regression had the highest accuracy; artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

Sonali Agarwal, G. N. Pandey, and M. D. Tiwari et al. [5] had proposed Support Vector Machines (SVM) is established as a best classifier with maximum accuracy and minimum root mean square error (RMSE). This is aimed to develop a faith on Data Mining techniques so that present education and business system may adopt this as a strategic management tool.

K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhan et al.[6]discussesd examine the potential use of classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. It can predict the likelihood of patients getting a heart disease.

Shweta Kharya et al.[7]discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis Decision tree is found to be the best predictor with 93.62% Accuracy on benchmark dataset and also on SEER data set.

Tina R. Patil, Mrs. S. S. Shereka et al.[20] had proposed that to make comparative evaluation of classifiers Naive Bayes and J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool.

N. Poomani, R.Porkodi et al.[21] had compared on various supervised learning algorithms to predict the best classifier. The experimental result shows that the highest accuracy is found in J48graft classifier and the lowest error rate 0.9587 among various classification algorithms. Based on the experimental result, it proves that the probabilistic model is not much suitable for classify breast cancer dataset.

R. Porkodi and G. Suganya [36] had implemented classification algorithm in colon cancer dataset. The experimental result shows that the highest accuracy is found in both KNN and Neural Network classifier gives the result Finally KNN and Neural Network classifier produces the good accuracy than the Support Vector Machine, Random Forest and Naïve Bayes classifier.

III. CLASSIFICATION ALGORITHMS

Classification is one of the most widely used methods
of data mining in healthcare. The classification algorithms can be useful to forecasting the outcome of some diseases or its discover the genetic performance of growth. This model is used to build the relating a predefined set of classes or ideas. The classification model is used to construct by analysing database tulles are described by attributes and also used to predict categorical class labels and classify the data based on the training set.

Classification techniques in data mining are capable of processing a large amount of data and it can be used to classifying newly available data. The classification algorithm is a method procedure which takes some value or set of value as input and generates some value or set as output. The result of a given problem is the output that we got after solving the problem. If the given algorithm is considered to be correct for every input occurrence, then it will generate the correct output and it gets completed or otherwise it does not considered as a correct algorithm. This paper gives the detailed description of five algorithms namely Decision tree, Bayesian network, K-nearest neighbor, Support vector machine and artificial neural network.

A. Decision Tree

A decision tree is a predictive modeling technique from the field of data mining that builds a simple tree-like structure. Decision Tree (DT) is one of the classification techniques in data mining. Decision tree builds classification in the form of tree structure. It divides whole training set into smaller subsets and at the same time decision tree incrementally developed. The result is a tree with decision nodes and leaf nodes. Decision tree classifier is helps to implement complex decision into easy process and the complex decision is subdivided into simpler decision. In decision
Process, a decision tree can be used to visually and explicitly represent decisions and decision making.

A decision tree describes data but not a decision relatively the resulting classification tree can be an input for decision making. Decision tree is one of the popular algorithms which are able to handle both categorical and numerical data and perform less computation. Decision trees are often simpler to interpret. Decision tree is a directed tree with a node and cannot having

incoming edges called root. All the nodes have one exact incoming edge. Each non-leaf node called internal node or splitting node and it contains a decision and most correct target value assign to one class is represented by leaf node. Decision tree can be used to analyze and represent classifiers models. On the other hand, decision trees also referred to a hierarchical model of decisions and their cost. When a tree is used for classification, then it is said to be as a classification tree. There are some specific decision tree algorithms, namely ID3 (Iterative Dichotomiser3), C4.5 Algorithm, CART (Classification and Regression Tree).

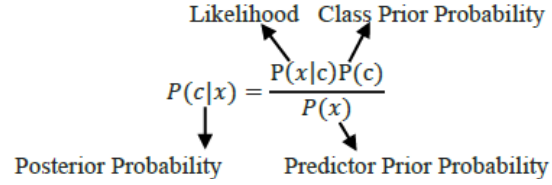
ID3 is one of the most important decision tree algorithms. In this method, information gain in advance and generally to determine suitable property for each node of a generated decision tree. We can select the attribute with the highest information as the test attribute based on current node. Therefore, the use of an information approach will effectively reduce the required dividing number of object classification. ID3 is a supervised learning algorithm, based on information entropy. It is developed from several classes and set of datasets. The algorithm planned a set of rules that allows predicting the class of an item and also used to identify the attribute of the class and then differentiate from others class. ID3 know the all dataset values and that dataset used to determine which attribute are important and which decision tree need to be included at which position is situated.

C4.5 algorithm is the successor of ID3 algorithm. It is used to reduce the error rate by replacing the internal node with a leaf node.C4.5 algorithm accepts both continuous and categorical attributes to build the decision tree.C4.5 has an enhance method of tree pruning and reduce the misclassification error due to noise. C4.5 algorithm handle the attribute with different costs and also handling training data with missing attribute values.

CART stands for Classification and Regression Trees. It is characterized by the each internal node which has exactly two outgoing edges. The splits are selected and the obtained tree is pruned by cost–complexity. When provided, CART can consider misclassification costs in the tree instructed and also users to provide the prior probability distribution. The major characteristic of CART is capacity to generate regression trees. Regression trees are predicted a real number and not a class.

B. Bayesian Network

The Naive Bays algorithm is a simple probabilistic classifier that is used to calculate a set of probabilities by using combinations of values in a data set. In all class variable attribute should be indecent in bays theorem. Bayesian network (BN) is a graphical model for probability relationships among a set of variables. This BN consist of two components. First component is mainly a directed Acyclic which contains nodes are called the random variables and the edges between the nodes or random variables. Second component which contain a set of parameters that describe the conditional probability of each variable given its parents. A naive Bayes classifier assumes the presence or absence of a particular feature and unrelated to the presence or absence of any other feature of class variable. Naive Bayes classifiers can be trained very well in a supervised learning and this method is important for several reasons.



LikelihoodClass Prior Probability

P (C|X) = P (X_1|C) ×P (X_2|C) ×……..×P (Xn|C) ×P(C)

P (C|X) is the posterior probability of class (target) given predictor (attribute).

P(C) is the prior probability of class.

P (X|C) is the likelihood which is the probability of predictor given class.

P(X) is the prior probability of predictor.

POSTERIOR = PRIOR×LIKELIHOOD/EVIDENCE Where Posterior is the predicting the event will

Occur, Prior is past experience, Likelihood is possible of chance and Evidence is total number of event will occur.

C. K- Nearest Neighbors

KNN Algorithm is based on similarity measure and used to Store all accessible cases and used to identify the unknown Data point based on the nearest neighbor. It is easy to Understand but has an unbelievable work in fields and Practice especially in classification. KNN is a supervised Classification technique which is used extensively. It is an Easy to implement classification technique and Training is Very fast. KNN is particularly well suited for multimodal Classes. In this method the training tulles are represented in N-dimensional space and given an unknown tuple, k-nearest Neighbor classifier searches the k training tulles that are Closest to the unknown sample and places the sample nearest Class.

The K nearest neighbor method is simple and Implement to the small sets of data, but when applied to Large of data and high dimensional data the results in Slower performance. Accuracy in data classification is a major issue in data mining and in order to improve the Accuracy of classification. The improvements have been Made to the K nearest neighbor method. Weighted nearest Neighbor classifier (wk-NNC) is one such method which Adds a weight to each of the neighbors in a classification.

KNN using distance function.

Distance function

Encliden	$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{k}\lvert x_i - y_i\rvert$
Minkowski	$\left[\sum_{i=1}^{k}(\lvert x_i - y_i\rvert)^q\right]^{1/q}$

Hamamoto's bootstrapped training set can also be used as a substitute of the training patterns. The training outlines are replaced by a weighted mean of a few of its neighbors from its own class of training patterns. This method proves to improve the accuracy of classification.

However the time to create the bootstrapped set is O (n2) where n is the number of training patterns. K-Nearest Neighbor Mean Classifier (k-NNMC). Finds k nearest neighbors for each class of training patterns separately. The classification is done based to the nearest mean pattern. This improvisation proves to show better accuracy of classification in when compared to other techniques using Hamamoto's bootstrapped training set. Hamming Distance

$D_H = \sum_{i=1}^{k}\lvert x_i - y_i\rvert$

X=Y→ D=0

X≠Y→ D=1

D. Support Vector Machine

Support vector machines (SVM) are also a type of machine learning tool. A support vector machine constructs a hyper plane in infinite-dimensional space, and which can be used to classification, regression, or other tasks. SVMs were first applied to protein sequence classification and have been applied to remote homology detection also. SVMs are supervisied binary classifiers used to find a linear separation between different classes of points in 3-D space. In 2D space, this separator is a line and in 3-D, it is a plane. This find an optimal separating hyper plan between members and non- members of a given class in an abstract space. SVM'S as applied to gene expression data begin with a collecting of known classifications of genes. One could build a classifier capable of disc riming between members and non-members of a given class. This would be useful in recognizing new members of a class, among genes of unknown function. The classifier could be applied to original set of training data of identify outliers that may have been previously unrecognized. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. The equation shown below is the hyper plane:

Hyper plane, ax + by = C

The main idea in SVM is an optimal hyper plane and which can be used in classification, for separation of linear patterns. The optimal hyper plane is select from the set of hyper planes. The set of hyper planes are classifying patterns from the margin of the hyper plane. The distance from the hyper plane to the nearest point of each pattern. The major purpose of SVM is to maximize the margin so the classifying given patterns is correctly and large margin size classifies patterns also correctly. The given pattern can be mapped by kernel function, Φ(x).i.e. x Φ(x). The different kernel function is an important aspect in the SVM-based classification. The kernel functions commonly used for LINEAR, POLY, RBF, and SIGMOID. For e.g.: the equation for Poly Kernel function is given as:

K(x, y) =<x, y>^p

E. Artificial Neural Networks

Neural Networks are used in prototype recognition and classification. A neural network is combination of nodes that are connected in a topology with each node has input and output connections to other nodes. Neural Networks are also called connectionist models because they are represented by weighted functions. The neural networks which are working with simple individual processing elements can perform

Complex method. The observation in a single layer neural network whose weights and biases is trained to produce a correct output when presented with the corresponding input vector. Artificial neural networks are connected by artificial neurons. Artificial neural networks is used to understand the biological neural networks and for solving artificial intelligence problems. These problems can be solved without using a biological system because the real, biological nervous are highly complex. Artificial neural network algorithms attempt to summary this complexity and focus on theoretically but most of the information are from processing point of view.

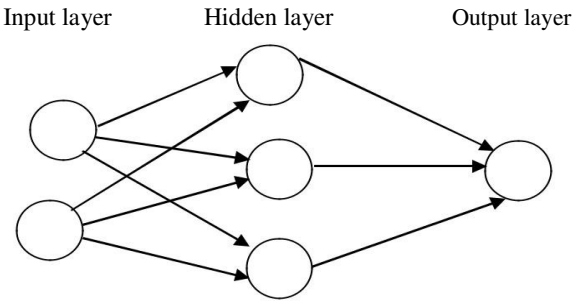


Fig. 1: Example of Artificial neural network

A neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of neural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. Basic topology of neural network consists of feed forward neural network and recurrent network. In feed forward neural network information flow starts from the input node .The information flow is one direction only from input node to hidden node and finally leads to the output node. In each node one or more processing elements (PE) may be active.PE is used to simulate the neurons in the brain.PE receive input from the outside world or from the previous layer. No cycles or loops in this network. But in recurrent neural network data flows bi-directionally and feedback connections exists here. Neural network consist of three parts architecture, learning algorithm and the activation function. Neural networks are programmed to store, recognize and retrieve patterns or database entries for solving ill-defined problems, to filter noise from measured data.

CONCLUSION

This paper manages classification strategies in data mining. Data mining consists of various fields, and one of that is bioinformatics. Classification is accustomed to predicting the qualities. In Data Mining, Classification methods has various algorithms in particular Decsion Tree,Naïve Bayes, K-Nearest Neighbors, Support Vector Machine and Artifical Neural Networks. Contrast with K-Nearest Neighbors, Decision Tree and Bayesian Network (BN), Support Vector Machine and Artificial Neural Networks for the most part have distinctive operational profiles. The classification procedure is to create more exact and precise framework comes about.

The goal of this paper is to enhance the Accuracy and performance of the Classifier. In the Classification algorithms, Decision Tree Classifier bolster a portion of the dataset recorded in Table 1. The comparison table shows how classification algorithms performed in various datasets and distinguished which one is gives the best exactness among the distinctive classifier. In future, it works with quality identification, quality prediction and quality analysis.

REFERENCES

[1] S.Archana, Dr. K.Elangovan, " Survey of Classification Techniques in Data Mining", Vol.2 Issue. 2, February- 2014, pg. 65-71.

[2] David B.Fogel, Eugene C, Wasson, Edward M.Boughton "Evolving neural networks for detecting breast cancer". 1995 Elsevier Science Ireland Ltd.

[3] Shadab Adam Pattekari and Asma Parveen ,"prediction system for heart disease using naive bayes", International Journal of Advanced Computer and Mathematical Sciences,ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[4] Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-[16].

[5] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari ,"Data Mining in Education: Data Classification and Decision Tree Approach".

[6] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "Applications of Data Mining Techniques in Health care and Prediction of Heart Attacks" International Journal on Computer Science and Engineering (2010).

[7] Shweta Kharya," Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[8] Amin, S.U., Agarwal. K., Beg.R., "Genetic neural network based data mining in prediction of heart diseaseusingriskfactors",Information& Communication Technologies (ICT), 2013 IEEE Conference on , vol., no., pp.1227,1231, 11-12 April.

[9] S.Ghorai, A.Mukherjee and P.K.Dutta, "Cancer Classification from Gene Expression Data by NPPC Ensemble ",IEEE/ ACM Transactions On Computational Biology and Bioinformatics,vol.8,No.3,May/June 2011.

[10] Topon Kumar Paul and Hitoshi Iba, "Prediction of cancer class with majority voting genetic programming classifier using gene expression data", IEEE/ACM
Transactions on Computational Biology and Bioinformatics, vol. 6, No. 2, April-June 2009.

[11] Kharrat, A., Gasmi, K. Messaoud, M. B., Benamrane, N. & Abid, M. (2010). A hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine",Leonardo Journal of science.,ISSN-1582-0233, pp. 71-82.

[12] Nidhi Bhatla Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), 2012.

[13] Romeo. M., F. Burden, M. Quinn, B. Wood and D. McNaughton. "Infrared Micro spectroscopy And Artificial Neural Networks In The Diagnosis Of Cervical Cance".U.S. National Library of Medicine C National Institutes of Health ,Vol.44(1),pp179-87,1998