International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

A Review on Naive Baye's (NB), J48 and K-Means Based Mining Algorithms for Medical Data Mining

Rajbir Kaur[1], Rakesh Gangwar[2],

[1]M.Tech Scholar, Department of Computer Science & Engineering
Beant College of Engineering and Technology, Gurdaspur, Punjab, India
[2] Associate Professor, Department of Computer Science & Engineering
Beant College of Engineering and Technology, Gurdaspur, Punjab, India

---

Abstract - *Data mining can be defined as discovery of meaningful patterns of large quantity of data and it analyze and explore to data . This paper studies various data mining techniques for improve accuracy rate for prediction of various diseases. This paper reviews the techniques and various evaluation techniques that describe and distinguish various kind of techniques for detection of diseases and their treatment in medical data mining.*

Key Words*: *Data Mining techniques, Naive bayes, ANN, KNN

1. INTRODUCTION

Data Mining is one of the very motivating and critical part of study with desire to of removing data from significant amount of accumulated information sets..An transformative route has been experienced in the repository market in the progress of the next functionalities information col lection and repository formation, information management (including information storage and collection, and repository purchase processing), and information analysis and understanding (involving data warehousing and data mining). Merely said, information mining refers to removing or \mining" know corner from big amounts of data.We've been collecting a myriad of information, from easy exact proportions and text documents, to more complicated data such as for instance spatial information, multimedia programs, and hypertext documents. *Information Mining*, also generally known as *Knowledge Discovery in Sources* (KDD), refers to the nontrivial extraction of implicit, previously as yet not known and possibly of good use data from information in databases. While data mining and understanding discovery in sources (or KDD) are often treated as synonyms, data mining is clearly part of the understanding discovery method

2. DATA MINING TECHNIQUES

It is the process of turning raw data into useful information so that various pattern can be extracted. Various researchers have studied and work on data mining techniques to evaluate and classify the diseases for medical data

2.1 ANN (Artificial Neural Network)

ANN is a classification model which is grouped by interconnected nodes. It can be viewed as a circular node which is represented as an artificial neuron that reveals the output of one neuron to the input of another. The ANN model is helpful in revealing the hidden relationships in the historical data, thus facilitating the prediction and forecasting of diseases of patients.ANN model is accurate enough to make important and relevant decisions regarding data usage.

2.2 NAIVE BAYES

Naïve Bayes is a classification technique which is based on probability theories which fully embody the characteristics of data of medical science. Bayes model is easy to use for very large datasets. In simple terms, a Naive Bayes assumed that the value of a particular feature does not related to the presence or absence of any other feature, given in the class variable. It undergoes through following steps:

a) Extract, clean and classify the symptoms of diseases.
b) Remove large punctuations and split them.
c) Counting Tokens and calculating the probability. This probability is called as posterior probability which is calculated by the formula described in.
d) Adding the probabilities and then wrapping up.

2.3 DECISION TREE

Decision tree is one of the predictive modeling technique used in data mining. It aids to divide the larger dataset into smaller dataset indicating a parent-child relationship. Each internal node is labeled with an input feature. Different kind of attribute test are express by internal nodes, test result are represent by bifurcations and nodes of leaf express classification of that type. Decision tree can handle both numerical and categorical data. It is well suited with large datasets. Higher accuracy in decision tree classification technique depicts that the technique can simulate. Decision tree is able to deal and handle large quantity of input data such as text with numeric data only textual or nominal. It is a

successful supervised learning approach which has the capability of extracting the information from vast amount of data based on decision rules.

2.4BACK PROPOGATION NEURAL NETWORK (BPNN)

ANN consists of interconnected processing units. Neuron is its single unit. Single neuron receives an input from another neuron. Each neuron has assigned weight.Weights finds strength and type of nature of interlinked neurons. Each input has signals that are processed by weighted sum of inputs. The signals from each input are then processed through a weighted sum on the inputs .Back propagation neural network algorithm find the errors by using steepest descent method. The connected weights are adjusted by moving in the direction of the negative gradient of the energy function at each iteration for evaluating the network performance. Various performance metrics are used for calculating the network error based on specific formulas. BPNN algorithm follows four steps:

1. Computation of feed forward..
2. Apply BPNN at the output layer.
3. Apply BPNN to the hidden layer.
4. Updating of weights.

This algorithm will continue its processing until the value of error function becomes too small.

2.5 KNN

K-nearest neighbor is an algorithm which is used for regression and classifying the quality problems. It considers various parameters which results in the ease of calculation time and predictive power. It uses a vast amount of classes to calculate the likelihood score. When several KNNs share a class, then the weights of other neighbours to it also added together. Result of such added weights is considered to be the likelihood score. These scores are then sorted in order to find the ranked list. Therefore, KNN is a very simple and effective algorithm.

2.6 J48

J48 is a C4.5 decision tree. J48 decision tree is developed by Ross Quinlan in 1993. This is classifier that is in flowchart structure, which is used to present different models of classification and just because of its nature it reveals the human reasoning. Different decision tree algorithms have many advantages over various learning algorithms like noise robustness, low computational cost for model generation and have ability to different redundant attributes and its modules.

3. RELATED WORK

Yuanyuan Gao,et al.(2017)[1] addressed the features of data mining with medical data. Various researchers in advance fields not be aware of the problems like heterogeneity privacy etc in medical science. Another aspects like legal and ethical in medical data mining are discussed ,another facts like benefits predication , fear lawsuits and data control. Estimation and hypothesis of medical data mathematically totally different from routines of data collection. Data in medical science is especially helpful for urgent condition like death etc..

Tapas Ranjan Baitharu, et al.(2016)[2] discussed that particular prediction of breast cancer survivability has been a new challenging research problem with regard to many researchers. Since the particular early dates of the particular related research, much development has been recorded within several related fields. The key objective of this manuscript is to record over a research project exactly where they took benefit of all those available technological advancements in order to develop prediction models regarding breast cancer survivability.

Ahmed Mohamed Ahmed,et al.(2016)[3]offers a survey of the obtainable literature on data mining using soft processing. This category possesses various equipments of soft computing and applied hybridizations of theses equipment, particular functions ,the criteria's that are required by each unit. For uncertainly in data rough pieces are used to handle them. In soft processing many challenges in data mining are indicated.

IsraAl Turaiki,et al(2016)[4].-The widespread option of new computational strategies and tools for info examination and predictive modeling necessitates medical informatics experts and practitioners to systematically choose the most appropriate technique to cope with scientific prediction problems. Specifically, the assortment of methods referred to as 'data mining' provides methodological and technical answers to cope with the examination of medical info and engineering of prediction models. A big variety of these procedures requires general and basic guidelines that might help practitioners in the correct collection of data mining tools, development and validation of predictive versions, combined with the dissemination of predictive designs within clinical environment.

Ila Dutta,et al(2017)[5] Most data of medical data is very high dimensional .Some of data is more relevant than other data ,it can be obtain after the utilization. Processing of medical data is selecting the proper subset of medical data features is one of important task because it increases the use of model and decreases the cost of particular model. From this paper we can easy find how analysis of medical data

with fuzzy strategies of fuzzy modeling can be determined by some methods or its indices.

Selina S.Y. Ng,et al(2014)[6] information of data mining and large databases expertise recognizes by many researches in systems of data source and learning of equipments by different commercial companies from where they can get revenues .A lot of applications in services of different information providing such as web applications on internet data warehousing and ways of data mining better understand the patterns of data .all of this process increases the demand of information and from where revenue can be increase it also helpful for business opportunities. In response to such a demand, a survey is supplied by this article, from a database researcher's perspective, on the info mining techniques lately developed. A classification of the available data mining techniques is provided and a comparative study of such techniques is presented

Alejandro Peña-Ayala, et al(2014)[7]-This paper surveys the application of data mining to traditional educational systems, particular web-based courses, well-known learning content management systems, and adaptive and intelligent web-based educational systems. Each of these systems has different data source and objectives for knowledge discovering. After preprocessing the available data in each case, data mining techniques can be applied: statistics and visualization; clustering, classification and outlier detection; association rule mining and pattern mining; and text mining. The success of the plentiful work needs much more specialized work in order for educational data mining to become a mature area.

Giorgio Maria Di Nunzio, et al(2014)[8]-Categorization of medical images means selecting the appropriate class for a given image out of a set of pre-defined categories. This is an important step for data mining and content-based image retrieval (CBIR). So far, published approaches are capable to distinguish up to 10 categories. In this paper, the author evaluates automatic categorization into more than 80 categories describing the imaging modality and direction as well as the body part and biological system examined. Based on 6231 reference images from hospital routine, 85.5% correctness is obtained combining global texture features with scaled images. With a frequency of 97.7%, the correct class is within the best ten matches, which is sufficient for medical CBIR applications.

S. Padmavathi, E. Ramanujam, et al(2015)[9].- Diabetes is really a significant health issue in the United States. The author has studied one particular diabetic data warehouse, featuring a technique of using data mining methods, and a number of the data issues, evaluation issues, and results. Techniques for translating a complex relational database eventually and sequencing data to a flat file||ideal for data mining are challenging. The author has discussed two parameter in detail|, a comorbidity index and the HgbA1c, a way of measuring glycemic control linked to outcomes. The author applied the classification tree method in Classification and Regression Trees (CART®) with a binary target variable of HgbA1c >9.5 and 10 predictors: age, sex|, emergency department visits, office visits, comorbidity index|, dyslipidemia, hypertension, cardiovascular disease, retinopathy, end-stage renal disease.

Vitali Loseu ,et al(2014)[10].-The healthcare industry gathers huge amounts of healthcare information which, however, {are not|"; mined"; to find out hidden data for efficient decision making. Advanced data mining method can provide solution to this situation. This research work has created a model Intelligent Heart Disease Prediction System (IHDPS) applying data mining methods, namely, Decision Trees, Naive Bayes and Neural Network. Results depict that each method has its distinctive strength in recognizing the objectives of the identified mining goals. IHDPS may answer complex "; what if"; queries which traditional decision support methods cannot. Using medical profiles such as for instance age|, sex, blood pressure and blood sugar levels it may predict the likelihood of patients getting a heart disease.

4. COMPARISION TABLE

Ref No.	Authors Name	Title	Year	Technique	Advantages	Limitations
1	Yuanyuan Gao	Incorporating association rule networks in feature category-weighted naive Bayes model	2017	Naive Bayes model	To represent and learn linear and non-linear relationships from the data being modeled.	The size of dataset is small. More data can be helpful to provide better results and better quality of predictions
2	Tapas Ranjan Baitharu	Healthcare Decision Support System Using Liver Disorder Database	2016	Liver Disorder Dataset	This find a non linear relationships to obtain data set quality and its robust attributes.	BPNN algorithm is not taken into use with other water quality parameters.
3	Ahmed Mohamed Ahmed	Predict Instructor Performance	2016	Naïve Bayes model	The method is effective as it measures the evaluation precision.	The method is not suitable for the evaluation of instructions quality.
4	IsraAl Turaiki	predictive models for MERS-CoV infections	2016	Predictive models	Forecasting results are good.	Some branches and rules in the decision tree is still can't explained.
5	Ila Dutta	Detecting financial restatements	2017	Expert Systems with Applications,	The CWQII is effective in evaluating the class of water quality.	It limited the use of comparing the class with other methods.
6	Selina S.Y. Ng	A naive Bayes model for robust remaining useful life prediction of lithium-ion battery.	2014	Naïve Bayes Model	It is simple and effective method and thus have low computational complexity.	Eutrophication problem is still a research area in the future of reservoir management.
7	Alejandro Peña-Ayala	A survey and a data mining-based analysis of recent works	2014	Naïve Bayes Model	FCE-EW is easy to operate as it evaluates the water quality.	Training samples are few.
8	Giorgio Maria Di Nunzio	A Visual Data Mining Approach to Parameters Optimization	2014	Optimization	Kstar algorithm has the best accuracy to classify water quality.	It limited the use of selection of models to find the most robust classification model for water quality.
9	S. Padmavathi, E. Ramanujam	Abnormalities Using Multivariate Maximal Time Series	2015	Multivariate Maximal Time Series Motif.	Effective performance metrics is evaluated while compared with the previous work.	It limited the use of user-centric approach for better accuracy prediction.
10	Vitali Loseu	Body Sensor Network Data Repository, In Wearable Sensors	2014	Body Sensor Network	It can be used with both seasonal and non-seasonal time series data.	It limits the use of monthly and day time series data.

5. CONCLUSION

Data Mining is one of the very motivating and critical part of study with desire to of removing data from significant amount of accumulated information sets. This specific paper shows about the comparison of various techniques based on medical data set in data mining. Various algorithms have been reviewed for predicting the diseases in medical data and their treatments and as a result of analyses. This review reveals various techniques of medical data mining to find various diseases and their detection so that their treatment can be easy. These techniques are very helpful for patients to detect their diseases.

6. REFERENCES

[1] Yuanyuan Gao, Anqi Xu, Paul Jen-Hwa Hu, Tsang-Hsiang Cheng, Incorporating association rule networks in feature category-weighted naive Bayes model to support weaning decision making, Decision Support Systems, Volume 96, April 2017, Pages 27-38.

[2] Tapas Ranjan Baitharu, Subhendu Kumar Pani, Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset, Procedia Computer Science, Volume 85, 2016, Pages 862-870.

[3] Ahmed Mohamed Ahmed, Ahmet Rizaner, Ali Hakan Ulusoy, Using data Mining to Predict Instructor Performance, Procedia Computer Science, Volume 102, 2016, Pages 137-142.

[4] Isra Al-Turaiki, Mona Alshahrani, Tahani Almutairi, Building predictive models for MERS-CoV infections using data mining techniques, Journal of Infection and Public Health, Volume 9, Issue 6, November–December 2016, Pages 744-748.

[5] Ila Dutta, Shantanu Dutta, Bijan Raahemi, Detecting financial restatements using data mining techniques, Expert Systems with Applications, Volume 90, 30 December 2017.

[6] Selina S.Y. Ng, Yinjiao Xing, Kwok L. Tsui, A naive Bayes model for robust remaining useful life prediction of lithium-ion battery, Applied Energy, Volume 118, 1 April 2014, Pages 114-123.

[7] Alejandro Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, Expert Systems with Applications, Volume 41, Issue 4, Part 1, March 2014, Pages 1432-1462.

[8] Giorgio Maria Di Nunzio and Alessandro Sordoni, Chapter 2 - Picturing Bayesian Classifiers: A Visual Data Mining Approach to Parameters Optimization, In Data Mining Applications with R, edited by Yanchang Zhao and Yonghua Cen, Academic Press, Boston, 2014, Pages 35-6.

[9] S. Padmavathi, E. Ramanujam, Naïve Bayes Classifier for ECG Abnormalities Using Multivariate Maximal Time Series Motif, Procedia Computer Science, Volume 47, 2015, Pages 222-228.

[10] Vitali Loseu, Jian Wu and Roozbeh Jafari, Chapter 5.2 - Mining Techniques for Body Sensor Network Data Repository, In Wearable Sensors, Academic Press, Oxford, 2014, Pages 383-407.

[11] Jonathan F. Easton, Christopher R. Stephens, Maia Angelova, Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: A data mining approach, Computers in Biology and Medicine, Volume 54, 1 November 2014, Pages 199-210.

[12] Helen R. Marucci-Wellman, Mark R. Lehto, Helen L. Corns, A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms, Accident Analysis & Prevention, Volume 84, November 2015, Pages 165-176.

[13] Guido Carvajal, David J. Roser, Scott A. Sisson, Alexandra Keegan, Stuart J. Khan, Modelling pathogen log10 reduction values achieved by activated sludge treatment using naïve and semi naïve Bayes network models, Water Research, Volume 85, 15 November 2015, Pages 304-315.

[14] Cheng-Huang Hung, Hua-Min Chiou, Wei-Ning Yang, Candidate groups search for K-harmonic means data clustering, Applied Mathematical Modelling, Volume 37, Issue 24, 15 December 2013, Pages 10123-10128.

[15] Kaiyang Liao, Guizhong Liu, Li Xiao, Chaoteng Liu, A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval, Knowledge-Based Systems, Volume 49, September 2013, Pages 123-133.

[16] Liang Bai, Jiye Liang, Chao Sui, Chuangyin Dang, Fast global k-means clustering based on local geometrical information, Information Sciences, Volume 245, 1 October 2013, Pages 168-180.