ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

DATA MINING TECHNIQUES FOR MEDICAL DIAGNOSIS

Linda Rose Dr.R. Maruthi*
Dhanraj Baid Jain college, SSN College of Engineering
g.lindarose@gmail.com, ,rmaruth2014@gmail.com

ABSTRACT-Data mining is a powerful technique to help the people in their business process and decision making strategies. Those techniques extract the hidden information from the large databases which helps to find the relationships and patterns from the data. Data mining is used for various applications such as business organizations, e-commerce, health care industry, scientific and engineering. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. This research paper provides an insight to the current techniques of knowledge discovery in databases using data mining techniques in the medical field.

***Keywords:** Data mining, medical diagnoses, Naive Bayes, K-means clustering, Support Vector Machines*

1.INTRODUCTION:

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous.

Data mining techniques in medical data includes association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. Recently it is finding its importance in healthcare domain for data analysis. Various studies in the literature reveal the development of intelligent decision support systems in the healthcare domain for accurate diagnosis of diseases. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on. The data mining techniques that have been applied to medical data include Apriori and FPGrowth [1,2,3,4], unsupervised neural networks [5,6], linear genetic programming [5], Association rule mining [7,[8], Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering Yang [9], decision tree algorithms like ID3, C4.5, C5, and CART [10,11,12,13], outlier prediction technique [14], Fuzzy cluster analysis [15], classification algorithm [16,17], Bayesian Network algorithm [10,18], Naive Bayesian [19], combination of K-means, Self Organizing Map (SOM) and Naïve Bayes [20], Time series technique [21,22], combination of SVM, ANN and ID3 [11], clustering and classification [23],SVM [11,24], FCM [22],k-NN [17] and Bayesian Network [10].

Medical diagnosis is a complicated task that needs to be executed professionally and precisely. Automating such diagnostic system is extremely important because most of the health care units do not have enough specialist and infrastructure. Therefore, automated medical diagnosis and predictive system is more useful in certain situations. It is not an easy task to develop such systems. An exhaustive study of various techniques in this field has to be studied and analysed. Once designed clinical test can be made at low cost. This paper aims to analyze the different predictive/ descriptive data mining techniques proposed in recent years for the diagnosis of heart disease, breast cancer, lung disease.

2. DATA MINING IN HEART DISEASES

Data mining techniques such as Naive Bayes, Decision Tree, Linear Regression and Association Rule are used to predict the heart diseases. The data classification is based on MAFIA algorithms in which the data is evaluated using entropy based cross validations. It provides extra information of knowledge for predicting and diagnosing the heart disease with the help of the already existing database.

Naive Bayes or Bayes' Rule is the basis for many machine learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities which provide novel ways for exploring and understanding data. This rule based classifier is more efficient and it outperforms more sophisticated classifiers. The feature independence

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

assumptions of this technique are advantageous to improve the prediction..

A conditional probability is the likelihood of some conclusion say C, given some evidence/observation, E, where a dependence relationship exists between C and E. This probability is denoted as P(C |E) in eq.1 as follows

$$p(C|E) = \frac{P(E|C)P(C)}{P(E)} \quad (1)$$

Heart Disease Prediction System is developed for decision making using Naive Bayesian Classification [25]. The system mine hidden knowledge from a historical heart disease database. Given a set of transaction T, the goal of association rule mining is to find all rules having support ≥ minsup threshold, confidence ≥ minconf threshold. Computational Complexity Given d unique items, Total number of item sets=2d Total number of possible association rule [26]

$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right] \quad (2)$$

K-means clustering is used for identifying the attributes which is one of the most popular, accepted and well known clustering techniques because of its efficiency. In generating the initial K centroids using the inlier method, following equations are used:

$$Ci = Min(X) - i \text{ where } 0 \le i \le k \quad (3)$$

$$Cj = Min(Y) - j \text{ where } 0 \le j \le k \quad (4)$$

Where C (ci, cj) is the initial centroid and min (X) and min (Y) is the minimum value of attribute X and Y respectively. K represents the number of clusters [27]

C4.5 is an algorithm for building decision trees .It is an extension of ID3 algorithm. C4.5 is one of widely-used learning algorithms. C4.5 algorithm builds decision trees from a set of training data using the concept of information entropy. C4.5 is also known as a statistical classifier. This algorithm checks for base cases and for each element x, discover the normalized information gain from dividing on x. Let x_best be the element with the highest normalized information gain. Create a decision node that breaks on a best. Repeats on the sublists obtained by dividing on x_best, and add those nodes as children of node. [28].

3. DATA MINING IN BREAST CANCERS

. Breast cancer is one of the most common disease among women and it is the leading cause of cancer death. The practitioners and the uses Predictive model technique which is useful in designing the structure of decision models for procedures such as prognosis, diagnosis and treatment planning. Christo Ananth et al. [18] proposed a system, in which a predicate is defined for measuring the evidence for a boundary between two regions using Geodesic Graph-based representation of the image. The algorithm is applied to image segmentation using two different kinds of local neighborhoods in constructing the graph. Liver and hepatic tumor segmentation can be automatically processed by the Geodesic graph-cut based method. This system has concentrated on finding a fast and interactive segmentation method for liver and tumor segmentation. In the preprocessing stage, the CT image process is carried over with mean shift filter and statistical thresholding method for reducing processing area with improving detections rate. Second stage is liver segmentation; the liver region has been segmented using the algorithm of the proposed method. The next stage tumor segmentation also followed the same steps. Finally the liver and tumor regions are separately segmented from the computer tomography image.

Decision tree [7] is one of the classification methods for predicting the breast cancer , which classifies the labeled trained data into a tree or rules. The data is taken from the training set, then the rules are derived in learning phase to tet the accuracy of the classifier. Then the unlabeled data is classified using the tree or rules obtained in the learning phase.

4.DATA MINING IN LUNG DISEASES

A systematic study of symptoms and risk factors is needed in finding the possibility of lung diseases and predicting the way of treatment viz.surgery, chemotherapy or radiotherapy. Prediction technique is based on systematic study of the statistical factors, symptoms and risk factors associated with Lung disease attributes. The commonly used classification algorithms are IFTHEN Rule, Decision tree, Bayesian classifiers and Neural networks

Naïve Bayes to unfocused probabilities that aims at delivering robust classifications when dealing with small or incomplete data sets. Naive Bayes classifier is known to be a robust method, which shows on average good performance in terms of classification accuracy, also when the independence assumption does not hold .[31]

Decision tree algorithm is used to mine frequent patterns from the data set. The frequent item sets that occur throughout the data base are extracted as significant patterns. The data is fed into

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

the decision tree algorithm to obtain the significant patterns related to cancer and non cancer data sets.[32]. SFP=Sw (n) ≥φ for all values of n (2),where SFP denotes significant frequent pattern and φ denotes significant weightage.

5.CONCLUSION

This paper gives some of the methods and algorithms that are used in diagnosing and treating the various diseases on lung, heart and Breast. It also gives an insight of data mining techniques for finding the hidden patterns that helps in decision making. From this study it is observed that the Data Mining classification techniques are applied for accurate diagnosis of diseases. Implementations of the techniques are highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. In Future, work can be planned to perform experiments with dataset and algorithms to improve the classification accuracy for various diseases.

6.REFERENCES

[1] Umair Abdullah (2008). Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining1.IEEE.p1-5.

[2] Sunil Joshi and Dr. R. C. Jain.(2010). A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database.IEEE.0 (0), p498-501.

[3] Thanh-Trung Nguyen. (2010). An Improved Algorithm for Frequent Patterns Mining Problem. IEEE.0 (0), p503-507.

[4] Xiaoyong Lin and Qunxiong Zhu. (2010). Share-Inherit: A novel approach for mining frequent patterns. IEEE.0 (0), p2712-2717.

[5] Markus Brameier and Wolfgang Banzhaf.(2001). A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining.IEEE.p1-10

[6] DoronShalvi and Nicholas DeClaris., (n.d).An Unsupervised Neural Network Approach to Medical Data Mining Techniques.IEEE.0 (0), p1-6.

[7] AdepeleOlukunle and SylvanusEhikioya, (n.d).A Fast Algorithm for Mining Association Rules in Medical Image Data.IEEE.p1-7.

[8] Cindy L. Bethel and Lawrence O. Hall and Dmitry Goldgof (n.d). Mining for Implications in Medical Data.IEEE.p1- 4

[9] Jeong-Yon Shim, Lei Xu (n.d). Medical Data Mining Model For Oriental Medicine Via Byy Binary Independent Factor AnalysiS.IEEE.p1-4. *

[10] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). THE APPROACH OF DATA MINING METHODS FOR MEDICAL DATABASE.IEEE.p1-3. *

[11] Yanwei Xing, Jie Wang and Zhihong Zhao (2007).Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease.IEEE.p1-5. *

[12] Sam Chao, Fai Wong, "An Incremental Decision Tree Learning Methodologyregarding Attributes In Medical Data Mining". Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[13] My ChauTu AND Dongil Shin (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms.IEEE.p1-5. *

[14] ViliPodgorelec, MarjanHerikoMaribor, (n.d). Improving Mining of Medical Data by Outliers Prediction.IEEE.p1- 6

[15] Shuyan Wang Mingquan Zhou GuohuaGeng (n.d). Application of Fuzzy Cluster Analysis for Medical Image Data Mining.IEEE.p1-6.

[16] AshaGowdaKaregowdaM.A.Jayaram (2009).Cascading GA & CFS for Feature Subset selection in Medical Data Mining.IEEE.p1-4.

[17] Graduate Institute of Applied Information Sciences (2009). Medical Data Mining Using Bga And Rga For Weighting Of Features In Fuzzy K-Nn Classification.Ieee.P1-6.

[18] Christo Ananth, D.L.Roshni Bai, K.Renuka, A.Vidhya, C.Savithra, "Liver and Hepatic Tumor Segmentation in 3D CT Images", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3,Issue-2, February 2014,pp 496-503

[19] Ranjit Abraham, Jay B.Simha, Iyengar (n.d). A

comparative analysis of discretizationmethods for Medical Datamining with Naïve Bayesian classifier.IEEE.p1-2.

[20] Syed Zahid Hassan and BrijeshVerma,(n.d). A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases.IEEE.p1-6

[21] Michele Berlingerio (n.d). Mining Clinical Data with a Temporal Dimension: a Case Study. IEEE.p1-8.

[22] WojciechFroelich, AlicjaWakulicz-Deja (2009). Mining Temporal Medical Data Using Adaptive Fuzzy Cognitive Maps.IEEE.P1-8

[23] Michael Barnathan, Jingjing Zhang, Vasileios (n.d). A Web-Accessible Framework For The Automated Storage And Texture Analysis Of Biomedical Images.IEEE.p1-3.Khaleel et al., International Journal of Advanced Research in Computer Science and Software Engineering 3(8), August - 2013, pp. 149-153 © 2013, IJARCSSE All Rights Reserved Page | 153

[24] SarojiniBalakrishnan (n.d). SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases.IEEE.p1-6. International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014 Copyright to IJARCCE www.ijarcce.com 6787 Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing

[25] Shabana Asmi P And Dr. S. Justin Samuel(2015), An Analysis And Accuracy Prediction Of Heartdisease With Association Rule And Other Datamining Techniques Journal of Theoretical and Applied Information Technology20th September 2015. Vol.79. No.2

[26] Aswathy Wilson, Gloria Wilson, Likhiya Joy K Heart Disease Prediction Using the Data mining Techniques International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 1, Jan-Feb 2014

[27] M.A.NisharaBanu, B Gomathy Disease Predicting System Using Data Mining Techniques, International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 41-45

[28] S.Kharya , D. Dubey , and S. Soni Predictive Machine Learning Techniques for Breast Cancer Detection S.Kharya et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 2013, 1023-1028

[29] D.Lavanya and Dr.K.Usha Rani Ensemble Decision Tree Classifier For Breast Cancer Data International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012

[30] V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra Diagnosis of Lung Cancer Prediction SystemUsing Data Mining Classification Techniques V. Krishnaiah et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45

[31] P.RamachandranN.Girija, Ph.DT.Bhuvaneswari, Ph.DEarly Detection and Prevention of Cancer using Data Mining Techniques International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014