IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 3, March 2019

Medical Expert System using Data mining and Machine Learning

Suhas A Bhyratae[1], Sumukha J Sharma[2], Tarun Kumar K[3], Yathish Kumar R[4]

Assistant Professor, Department of Information Science & Engineering,
Atria Institute of Technology, Banglore, India[1]

Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[2,3,4]

Abstract: A vast amount of data is generated in the fields of healthcare and diagnostics, doctors have to make a direct contact with patients to determine the wounds, injuries and diseases by which the patient is affected. This paper highlights the application of classifying and predicting a specific disease by implementing the operations on medical data generated in the field of medical and healthcare. The proposed system can solve difficult queries for detecting a particular disease and also can assist medical practitioners to make smart clinical decisions which traditional decision support systems were not able to. The decisions taken by medical practitioners with the help of technology can result in effective and low cost treatments. In this paper, data mining methods namely, Naive Bayes and J48 algorithms are compared for testing their accuracy and performance on the training medical datasets.

Keywords: Prediction, Classification, KNN, Naïve Bayes, J48, neural network.

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The difference between data analysis and data mining is that data analysis is to summarize the history such as analysing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data.

II. RELATED SURVEY

Data mining in highly visible fields like e-business, marketing has successful application and retail has to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still information rich but poor in knowledge. Data availability is very rich in health care systems. However there are very less analysis tools that helps in discovering trends in data and hidden relationships. This paper intends to give a survey of known techniques of knowledge discovery in data mining that are used in today's medical field particularly in Heart Disease Prediction. Various experiment have been conducted to check the performance of predictive datamining technique on the identical dataset and the result shows that Decision Tree can outperform and sometime Bayesian classification has similar accuracy as of decision tree and other predictive methods like KNN, Neural Networks. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. It is a huge advantage to automate this system. All doctors may not possess expertise in every subject specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would be a huge benefit by bringing all of them together. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced efficiency and a comparative study of various techniques available can give an accurate implementation of automated system.

The data mining classification techniques viz. K-NN, Decision Trees, and Naive Bayes are utilized. Three different supervised machine learning algorithms i.e. Naive Bayes, K-NN, Decision tree algorithm have been utilized for analysing the data set. Tanagra tool is utilized to classify the data and the data is evaluated using I O-fold cross validation and the results are compared. Tanagra is a data mining suite built around graphical user interface algorithms. The main objective of Tanagra project is to provide researchers and students an easy-to-utilize data mining software, and permitting to analyse either real or synthetic data. Tanagra is powerful system that contains clustering, supervised

learning, Meta supervised learning, feature selection, data visualization, supervised learning assessment, insights, and feature selection and construction algorithms. Decision Tree is a classifier which is straight forward and simple to implement. It asks for no domain knowledge or parameter setting and can handle high dimensional data. The outcomes obtained from Decision Trees are simple to read and interpret. The drill through feature to access detailed patients data is only available in Decision Trees. Naive Bayes is a statistical classifier which accepts no dependency between attributes. It tries to maximize the posterior probability in determining the class. The benefit of using naïve Bayes is that one can work with the naive Bayes model without using any Bayesian techniques.

Naive Bayes classifiers works well in many complex real-world situations. The k-nearest neighbour (k-NN) is a method for classifying objects based on closest training data in the feature space. K-NN is a type of instance-based learning. The k-nearest neighbour algorithm is one of the least complex machine learning algorithms. But the accuracy of the k-NN algorithm can be severely degraded by noisy or irrelevant features, or if the feature scales are not consistent with their importance. The experiment is done using training data set consisting of 3000 instances with 14 different attributes. The data set is divided into two parts that is 70% of the data are utilized for training and 30% are utilized for testing. Based on the experimental outcomes, it is clear that the classification accuracy of Naive Bayes algorithm is better compared to other algorithms.

This paper has dissected prediction systems for Heart disease utilizing more number of input attributes, the framework uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to foresee the probability of patient getting a Heart disease. Until now, 13 attributes are utilized for prediction. This research paper included two new attributes i.e. obesity and smoking. The data mining classification techniques such as Decision Trees, Naive Bayes, and Neural Networks are analysed on Heart disease database. The performance of these methods is compared, based on precision. As per the outcomes precision of Neural Networks, Decision Trees and Naive Bayes are 100%, 99.62% and 90.74%. Our analysis shows that out Of these three models Neural Networks predicts Heart disease with highest precision. Various studies have been done that have centre on diagnosis of heart disease. They have applied diverse data mining techniques for diagnosis & accomplished different probabilities for different methods. Using data mining techniques Naive Bayes, Neural Network and Decision Trees, an Intelligent Heart Disease Prediction System (IHDPS) was developed and was proposed by Sellappan Palaniappan et al. Each method has its own strength to get appropriate outcomes. To fabricate this system hidden patterns and relationship between them is utilized. It is web-based, easy to understand and expandable. To develop the multi-parametric feature with direct and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique was proposed by HeonGyu Lee et al. To achieve this, they have utilized several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and Support Vector Machine. Today, numerous hospitals manage healthcare information using healthcare information system. As the system contains huge amount of data, utilized to extract hidden information for making intelligent medical diagnosis. The main objective of this research is to build Heart Disease Prediction System that provides diagnosis of heart disease using historical heart database. To develop this system, 13 input attributes such as sex, blood pressure, cholesterol, etc. are utilized. To get more appropriate outcomes, two more attributes i.e. obesity and smoking are utilized, as these attributes are considered as vital attributes for heart disease.

The proposed strategy in this research work is an extended form of the model that combines the genetic algorithms for feature selection and fuzzy expert system for effective classification. Healthcare knowledge based system for diagnosis of diseases can be developed using fuzzy logic and set theory. Experiments are conducted in Matlab using fuzzy tool. For this, Mamdani model of fuzzy system is used. The fuzzy rules are generated based on specialist's learning in this space. The dataset from UCI machine learning store is utilized, and only 6 attributes are observed to be effective and vital for heart disease prediction. In the proposed system, the input is the set of all the selected features and the output of the system is to achieve a value O or 1 that indicates the absence or appearance of heart disease in patients. In fuzzy logic process, initially fuzzification is performed by collecting the crisp set of input data and converting it to a fuzzy set using fuzzy linguistic variables, fuzzy linguistic terms and membership functions. After that, an inference is made based on a set of rules and lastly, defuzzification step is performed. The system produces the fuzzy rules based on the support sets obtained. The goal of our work is to give a study of different data mining techniques that can be used in heart disease prediction systems. Different techniques and data mining classifiers are characterized in this work which has emerged in recent years for productive and powerful heart disease diagnosis. The analysis demonstrates that Neural Network with 15 attributes has displayed the highest precision until now. On the other hand, Decision Tree likewise has also performed with 99.62% precision using 15 attributes. Also, in combination with Genetic Algorithm and 6 attributes, Decision Tree has demonstrated 99.2% efficiency. The Healthcare industry is by and large "data rich", yet unfortunately not all the data are mined which is required for finding hidden patterns & powerful decision making. Propelled data mining techniques are used to discover knowledge in database and for medical research, particularly in Heart disease prediction.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 3, March 2019

III. CONCLUSION

From the above we conclude that the prediction of the disease suffered by the patient can be done using the symptoms mentioned by them. In the previous years, researchers and practitioners in the field of medical and healthcare have achieved success in predicting a specific disease using data mining algorithms and techniques. The ideal strategy is to analyse and test various data-mining algorithms and to implement the algorithm which gives out highest degree of accuracy. The Datasets selected for implementation purpose contains more than 20 medical related attributes for each disease and maximum up to 500 algorithms training data examples. The algorithm used to implement in this project is Naive Bayes, which is selected by evaluating the predictive accuracy and latency analysis results.

REFERENCES

[1]. Ankita Dewan, Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2015 2ndlnternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE 2015.
[2]. Adam Sadilek, Henry Kautz, Lauren DiPrete, Brian Labus, Eric Portmanl, Jack Teitell, and Vincent Silenzio," Deploying nEmesis: Preventing Food bome Illnessby Data Mining Social Media" ewY ork.
[3]. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" , International Journal of Computer Applications (0975 8887) Volume 17 No. 8, March 2011.
[4]. Abhishek Taneja. "Prediction of heart diseases using data mining techniques". Oriental Journal of computer science and technology. December 2013. Vol. 6, No. (4): Pgs. 457-466.
[5]. T. Revathil, S. Jeevitha, "Comparative Study on Heart Disease Prediction System Using Data Mining Techniques", international Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013).
[6]. Ajinkya Kunjir, Harshal Sawant, Nuzhat Shaikh," A Survey on Prediction of Multiple Diseases and Performance Analysis using Data Mining and Visualization Techniques" , International Journal of Computer Applications (0975 —8887) Volume 155 — No 1, December 2011.
[7]. HICAP: Hierarchial Clustering with Pattern Preservation (2004). Hui Xiong, Michael Steinbach, Pang-Ning Tan, and Vipin Kumar, In Proc. of the Fourth SIAM International Conf. on Data Mining (SDM'04), Florida, USA, 2004.
[8]. Nuzhat F. Shaikh, Dharmpal D. Doye, "Improving the Accuracy of Iris Recognition System using Neural Network and Particle Swarm Optimization" International Journal of Computer Applications (0975 — 8887)Volume 79 — N03, October 2013
[9]. Monika Gandhi, Dr..Shailend1u Narayan Singh , "Predictions in Heart Disease Using Techniques of Data Mining" , 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015), IEEE 2015.
[10]. www'.cs.waikato.ac.nz/ml/weka/
[11]. Prof. Pier Luca Lanzi, "Visualization Techniques in Data Mining".
[12]. archive.ics.uci.edu/ml