ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

A Review on Different Data Mining Techniques and Their Comparative Analysis

Gulshan Bleem, Ishpreet Singh Virk

Department of computer Science and Engineering, BBSBEC, Fatehgarh Sahib, Punjab, India

Department of computer Science and Engineering, BBSBEC, Fatehgarh Sahib, Punjab, India

Abstract: Data mining is the current stream where companies are giving more emphasis on. Various types of applications are there in today world which are producing large amount of data. These data items need to be processed for the extraction of the useful data. This data later on can be used for the purpose of decision making purpose. In current scenario we have taken the data for the diabetic prediction based on various health parameters like Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2-Hour serum insulin, Body mass index, Diabetes pedigree function, Age (years). Based on these parameter the prediction for the diabetics is performed.

KEYWORDS: Prediction, data mining, Jrip.

I. INTRODUCTION

With the ever-increasing data in every field, there is a need to utilize the obtained information in a productive. Focusing on usage and advantages of various technologies and their applications, they can result in refined data analysis and better overall predictions. These technologies and platforms can provide provisions for improved solutions, better strategies and improve the process of decision making.

Data mining is an emerging technology that is designed to efficiently handle big data and provide useful implications in various fields. The term "Data mining" was first coined in 1990's and is often considered as a synonym of Knowledge discovery in databases. But data mining is actually a crucial part of KDD where the details analyzed to obtain connections and find patterns within the data and using them predictions can be made or the recent trends can be uncovered and analyzed whereas KDD is the overall process of knowledge extraction. These tasks are broadly classified into supervised, unsupervised and reinforcement learning. Supervised learning refers to process of training the model using the data which is already labeled, that is, data is already classified correctly. In unsupervised learning, the model is not trained because the instances are neither labeled nor classified initially. The model itself classifies the data based on the patterns observed and their similarities. In reinforcement learning, the machine learns from the feedback of the previous input and outputs.

Various search engines and social networking sites collect huge amounts of data and by using different data mining techniques, they try to find the hidden patterns within the data. Data mining is widely used in diverse areas such as banking, e-commerce, health and medicine, genetics, education, stock exchange and various other fields.The data is available in structured, semi-structured and unstructured format which is initially processed and then analyzed using different techniques. These techniques are broadly classified into two main categories, namely, predictive and descriptive. The predictive data mining techniques focus on understanding the future, making predictions using the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

available datasets whereas the descriptive techniques summarize and analyze the past data and properties to make it useful in predicting new ones.

Various fields like marketing, education, banking sector, bio-informatics, healthcare make use of these techniques. But nowadays, a lot of focus is given to the healthcare sector to improve the process of decision making and provide better facilities to the patients. The data for analyzing the health data can be collected from various sources like paper records, x-rays, etc and can be stored in electronic health records. The accumulated data can then be processed and analyzed. The predictive analysis considers the various symptoms, their effects on health and can help in early detection and prevention of these diseases. Different data mining techniques have been implemented to obtain the results regarding breast cancer, heart disease and diabetes to analyze the current status and make future predictions regarding the occurrence of disease, early detection and preventable patient deaths.

1.1 Background

Turkoglu (2002) researched about lung-valued disease detection by using wavelet entropy and short time Fourier transform in case of discovering lung signal features. Then, the researchers obtained the accuracy rate of 95.9% for pathological sounds and 94% for normal sounds according to Artificial Neural Network. After processing segmentation, feature vectors were developed using Daubechies-2 wavelet detail to use in classification. Neural Network classification algorithm produced the success rate up to 90%. For the purpose of extracting the imperative measurable qualities of lung sounds comparing to four main lung diseases, Babaei and Geranmayeh (2009) proposed a multiresolution wavelet based algorithm. It was found out that Daubechies wavelet channel with multilevel disintegration was the most uncertain in sickness discovery. Then, the most elevated precision accomplished was 94.42% with the use of multilayer perceptron with hidden back propagation algorithm.

Lung-valued diseases was planned (Maglogiannis, Loukis, Zafiropoulos, & Stasis, 2009) for recognizable proof by computerized determination framework. The strategy contains an underlying pre-preparing step took after by diagnosis phase considering SVM classifiers done in 3 stages: arrangement of lung sound as expected or obsessive, discovery of the lung mumble sort, detection if the sound compares to a mitral or aortic sickness, with the achievement accuracy rate of over 90% in each step. In addition, for both of detecting a systolic disease and diastolic, the accuracy is over percentage of 76. Therefore, the performance of SVM is better than other classification algorithms such as Neural Networks, K-nearest-neighbour and Naïve Bayes. Pavlopoulos (2004) proposed a decision tree classification algorithm for lung sounds. In the calculation of lung sounds signals, the result is accuracy of 90% and more partial AS and MR accuracy of 91.6% and 88.5% after extracting 100 features using decision tree algorithm.

In summary, Neural Networks produced 94% of normal and 95.9% of pathological sounds in the extraction of diseased and normal sounds for the purpose of detecting lung disease and it gave the result of 98.50% in grow and learn and 97.01% in multilayer perceptron back propagation for detecting lung murmur. The decision tree algorithm produced 90% for mitral regurgitation.

1.2 Classification Approaches

Classification is designed for the purpose ofestimating a definite result according to a given input. Based on the outcome from prediction, the algorithm forms a training set including a pair of attributes and the individual result, mostly known as prediction or objective attribute. At that point, it tries to discover the associations among the attributes if there should be an occurrence of anticipating the result.After that, the algorithm receives a new data set, called prediction set containing the same arrangement of characteristics, aside from the prediction attribute. The algorithm

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

analyses the input and gives a predicti1on as the output. The prediction accuracy characterizes how great the algorithm is.

1.2.1 Artificial Neural Networks

Neural networks approach the problems in a different way. Neural networks, with their remarkable ability to obtain meaning from tangled or imprecise data, can be used to derive the patterns and determine trends that are too complicated to be observed by other computer techniques or either humans. A trained neural network can be thought of as a machine in the category of information it has been given to analyses.

Some of the benefits are adaptive learning, self-organization, real time operation and fault tolerance or redundant information coding.

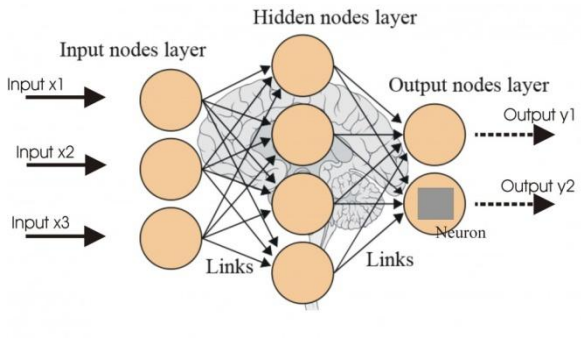

Figure 1: Basic Structure processing Artificial Neural Network[2]

1.2.2 Decision Trees

Decision trees support a group of rules relating to a new dataset in the place of predicting the result. The structure is formed where each branch node states a choice among the alternatives and each leaf indicates a classification. By thinking the branch node as the query, the branch as the rule and the leaf as the statement. The number of levels or branches on the tree is flexible by explain the users which can produce the results. By processing the depth of the tree incorrectly, it can go forward to under or overfitting. Then it shows the model hardly in order to fit any other data sample in overfitting case. In addition, Decision trees can depict as the procured information with a visual frame and they are adaptable for translating unless they are not greatly vast that can prompt the issue of over fitting.Decision trees are also good at reducing the noise and issuing the incomplete data and also, they can control both continuous and discrete data.It is the best classification and clustering technique. it includes the classfication of the different attributes values In tree like format. As the attribute will be the parent element and the value of the attribute will be the child elements of the parent elements. The driving of the parent child relation will helps in determining the total relationship of different attributes and their interrelationships.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

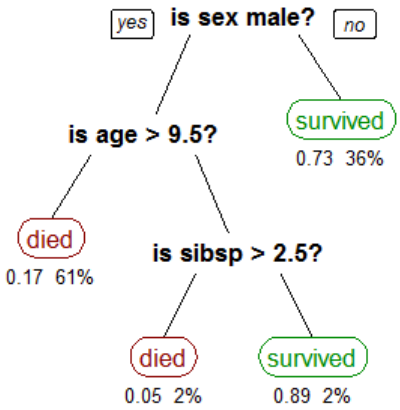

Figure 2(a): Basic Structure of Decision Tree[3]

Figure 2(b): Basic Structure of Decision Tree[8]

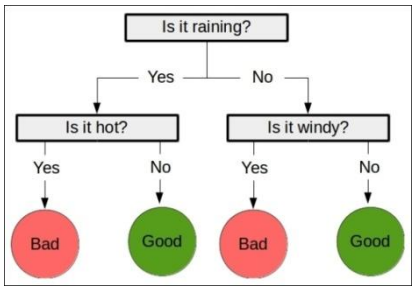

Figure 2(c): Basic Structure of Decision Tree[8]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

1.2.3 One-R

In most cases, One-R is defined as the ordinary. It places the attributes based on the error rate on the training set. It handles all numerically-esteemed attributes consistently and utilizes a clear strategy to partition the scope of quality into a few disjoint intervals. It treats missing data by handling missing as a true value.

In datasets with consistent esteemed attributes, there is an issue of overfitting. In differentiating the scope of values into a limited number of interims, it is appealing for making each interval clear having examples that are the greater part of the same class. Nevertheless, normal overfitting may produce the result from deepening a decision tree until all the processes are clear, therefore, extreme overfitting may come out because of subdividing an interval until all the sub trees are clear. To avoid this case, One-R needs all intervals apart from the rightmost, to contain more than a predefined number of cases in the same class.

1.2.4 SVM (Support Vector Machines)

Support vector machines are a group of supervised learning techniques used for classification, regression and outlier's detection. It is productive in high dimensional spaces. As the requirement, it is needed to draw the margin line among the clusters. The working procedure of SVM algorithm is dependent in case of discovering the hyperplane, which can give the greatest minimum distance comparing to the training exemplifiers. The optimal hyperplane is existing in the half way of separating clusters. Therefore, it describes the maximum margin length between the positive and negative support vectors. In case of dividing the points, it can be allowed to separate linearly but it is necessary to consider in which the larger in margining the points, the better result is possible.

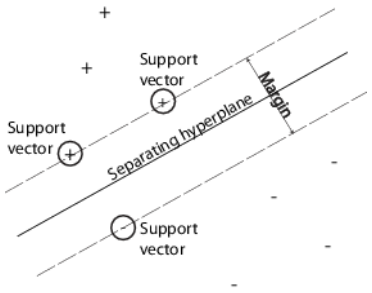

Figure 3: Expression of Support Vector Machine

1.3.5 Naïve Bayes

Naïve Bayes is a form of supervised learning approach and it is a statistical way for classification. It computes the probabilities of each attribute belonging to each class for making a prediction. It gives a strong assumption but results in a fast and effective method. Instead of the fact that it can perform better for more advanced classification strategies. In this algorithm, attributes are standing themselves without depending on others. As the great point, it helps decrease in computation cost only counting the class distribution. The positive points of using this Naïve Bayes algorithm are that it is easy for implementation and it can capture good results in many of the cases. Furthermore, it does not need to separate in testing, training and validating data. On the other hand, itcan judge that there is not influence among the inputs each other and the result is rarely true.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

1.3.6 JRIP

It is the proportional rule learner. It is repeated incremnetal pruning to produce error reduction. It sub divide the total dataset into two parts one is the training set and other ios the pruning set. At each stage of the iteration the prunning set operator will remove the error generating rule. At the initial the different rules are prepared in growing manner and further there is a phase of shrinking set. Which will reduce the rule set by applying the prunning set operator. It will sinmply reduces the or delete the single rule set prunning operator is the set which reduces the error rate to greatest level.

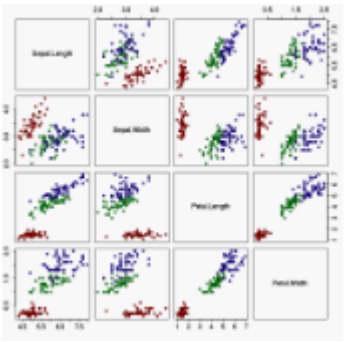

Figure 4 J-Rip classification

II. LITERATURE SURVEY

H. Fatemidokh et al. presented a comparative analysis of three different models for prediction of diabetes and prediabetes. Based on 12 different features and one outcome variable, the models, based on logistic regression, ANN, decision tree, were implemented. Results showed that the highest accuracy of 77.87% was obtained using decision tree. The logistic regression model attained an accuracy of 76.13% whereas ANN model had the least accuracy of 73.23%.

P. Meli et al. compared three different machine learning algorithms for prediction of diabetes using the PIMA Indian diabetes dataset. The preprocessed data is classified using naïve bayes, decision tree, SVM. Experimental results showed that naïve bayes achieved the highest accuracy of 76.30%. This work can be extended for prediction of various other diseases.

Archana et al. presented a comparative study of the performance of k-means clustering using distance metrics, namely, Euclidean, Manhattan and Minkowski distance. Results showed that performance of k-means clustering is affected by the selection of distance metrics. It is also concluded that euclidean distance metric provides the best result as compared to others whereas Manhattan distance metric is the worst.

Vaishali et al. Implemented multi objective Evolutionary fuzzy algorithm for classification of PIMA Indian diabetes dataset. Initially, before classification genetic feature selection is performed to remove the redundant features which helps in improving the accuracy and speed of the classifier. The performance of classification algorithm with and without using feature selection were compared and results showed that the combination of genetic feature selection and MOE fuzzy classifier is better than the others.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Babaei et al. proposed an optimized hierarchical clustering method which aimed at reducing the computation cost. Further, features were selected using Genetic algorithm and data was classified using SVM. The accuracy obtained using PIMA Indian diabetes dataset was improved by 1.351%.

Chen et al. Proposed an approach in which the processed data is fed to k-means clustering for data reduction. Then the instances obtained are classified using decision tree. The J48 decision tree was implemented using WEKA toolkit. Classification of 532 instances obtained after data reduction, using J48 decision tree, an accuracy of 90.04% was obtained.

III. CONCLUSION

Based on the discussion on different data mining techniques on different types of dataset for different types of purpose it is clear that in current times the data generated by various applications need different levels of processing to extract the useful data. This data can be used for different purposes in different fields. One such field is the prediction model. Where the data prediction based on the current data for the future actions. Medical is such field where the prediction is required for the different types of disease prediction based on the current parameters of the person health. Different researchers are putting their effort in this line. Each type of technique so far applied is for the prediction with some success rate. The work can be enhanced further by incorporating the hybrid approach based technique. which is having additional rule sets.

IV. FUTURE WORK

Based on the research of current research papers different techniques exists used for the prediction of the diabetes. The work can be enhanced further by simply using hybrid approach based classification for the classifying the results in different classes of true positive and true negative.

REFERENCES

1. H. Fatemidokht and M. K. Rafsanjani(2018) "*Development of a hybrid neuro-fuzzy system as a diagnostic tool for type 2 diabetes mellitus*," in Fuzzy and Intelligent Systems (CFIS), 2018 6th Iranian Joint Congress on, pp. 54-56, IEEE.
2. P. Melin, E. Ramirez, and G. Prado-Arechiga(2018) "A new variant of fuzzy k-nearest neighbor using interval type-2 fuzzy logic," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-7, IEEE.
3. R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri(2017) "Genetic algorithm based feature selection and moe fuzzy classification algorithm on pima indians diabetes dataset," in *Computing Networking and Informatics (ICCNI), International Conference on*, pp. 1-5, IEEE.
4. Babaei, S., &Geranmayeh, A. (2009). Lung sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals. Computers in biology and medicine, 39(1), 8-15.
5. Griffel, B., Zia, M. K., Fridman, V., Saponieri, C., &Semmlow, J. L. (2013). Path length entropy analysis of diastolic lung sounds. Computers in biology and medicine, 43(9), 1154-1166.
6. Gupta, C. N., Palaniappan, R., Swaminathan, S., & Krishnan, S. M. (2007). Neural network classification of homomorphic segmented lung sounds. Applied Soft Computing, 7(1), 286-297.
7. Maglogiannis, I., Loukis, E., Zafiropoulos, E., & Stasis, A. (2009). Support vectors machine-based identification of lung valve diseases using lung sounds. Computer methods and programs in biomedicine, 95(1), 47-61.
8. Pavlopoulos, S. A., Stasis, A. C., &Loukis, E. N. (2004). A decision tree–based method for the differential diagnosis of Aortic Stenosis from Mitral Regurgitation using lung sounds. Biomedical engineering online, 3(1), 1.
9. Turkoglu, I., Arslan, A., &Ilkay, E. (2002). An expert system for diagnosis of the lung valve diseases. Expert Systems with Applications, 23(3), 229-236.
10. Bhatia, K. and Syal, R., 2017, October. Predictive analysis using hybrid clustering in diabetes diagnosis. In 2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE) (pp. 447-452). IEEE.
11. Bora, D. (2019). Big Data Analytics in Healthcare: A critical analysis. In: Big Data Analytics for Intelligent Healthcare Management. [online] Mara Conner. Available at: https://www.sciencedirect.com/book/9780128181461/big-data-analytics-for-intelligent-he althcare-management [Accessed 16 May 2019].
12. Chen, W., Chen, S., Zhang, H. and Wu, T., 2017, November. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 386-390). IEEE.
13. Choudhury, A. and Gupta, D., 2019. A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. In Recent Developments in Machine Learning and Data Analytics (pp. 67-78). Springer, Singapore.