ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 11, November 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Classlification & Prediction based Data Mining Techniques

Neha Sharma[1]
Department of Computer Science & Engineering
India

Damyanti Sharma[2]
Department of Computer Science & Engineering
India

Abstract: ***Data mining is a process of extracting the meaningful information and pattern matching from the existing dataset. The mined data is used for the purpose of decision making in various fields. The data mining is applicable to fields like industries to increase the profitability by making some future policies, in commercial field to analyze the behavior of customer so that sales can be increased by attracting more and more customers, in medical field to analyze the type of disease from which the patient is suffering. This paper provides a review to the use of data mining in medical field to diagnose the patients who suffers from kidney related disease. The doctors analyze the health related data of the patient and hence as result conclude that which type of disease the patient is suffering from. The data mining can be applied to various data sets by using the classifiers and pattern matching techniques. The various types of data mining classifier is also introduced in this work.***

Keywords: Data mining, Knowledge Discovery, chronic Disease, Genetic Algorithms.

I. INTRODUCTION

Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of such systems would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Kidney related disease is most common among aging patients. The medical field has large datasets which are used for diagnosing the patient's health issues. The process of analyzing such a huge dataset is not n easy task. The medical data is mined by using the concept of data mining and pattern matching. The problem of data management in bioinformatics is sorted by using the data mining [1].

Kidney is an inseparable part of human body. Kidney is a pair of organs which is located at the lower part of the abdomen. It purifies the blood by extracting the toxin material from the body. From last few decades there is a increase in the number of patients with kidney related disease. The disease like kidney failure can even lead to a human to death. The kidney related disease also known as chronic disease or acute disease. The chronic kidney disease is referred to that situation when the kidney of one's is incapable of performing filtrations process in the body. This comprises circumstances that harm the kidneys. If kidney is unable to work then it can induce the amount of waste material in blood. And hence led to various health issues like high blood pressure, coronary artery disease, hypertension etc. the solution to kidney related disease is either dialysis or kidney transplant[10].

Data mining is an automatic process which is used to remove the meaningful information from the data storage and further use this removed information for various purposes. The extraction of meaningful data can be performed by matching patterns and it is achieved by cluster analysis, anomalies analysis, and dependencies analysis. Spatial indices are used to perform all above functions or processes. The matched pattern is a form of brief summary of data stored in the data warehouse and these patterns are used for future prediction and various decision making systems to take right decision. For example in case of machine learning systems this extracted information can be used for prediction analysis. Another example, data mining is a

process which find or investigates various groups of correlated data in the database which further can be used for predictive analysis in near future. Data analysis, data collection, compilation of data is not a connected to the data mining but still included in the process of KDD i.e. Knowledge Discovery. But all of these are performed as optional or additional steps in KDD. Some terms like data dredging, data snooping uses the data mining to uncover the pattern matching in large amount of data which is stored in warehouse[2].

The figure below describes the process of data mining step by step. First step is to clean the data which refers to eliminate the impurities like repetition, null values, and irrelevant data from dataset. Second step is to perform data integration which refers to integrate the related datasets in a set. Third step is to select the meaningful data and then the selected data is converted from source code format to destination code format which is known s data transformation process [7]. After this the actual process of data mining is performed in which the meaningful data is extracted from the datasets. Here meaningful data refers to that data which is helpful in decision making process in future. Then after the evaluated patterns are matched to find the existence of some constituent patterns. Then at last the received or extracted data is referred as knowledge representation.

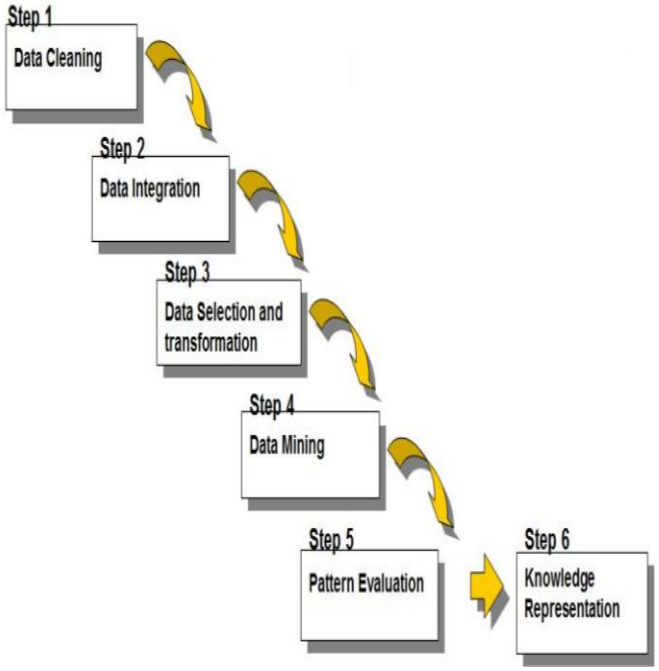

Fig1. Process of Data Mining

Data mining includes various issues and hence it is a most preferable area of research. There are many issue related to the data mining such as theoretical issues. There are some problems related to the practical implementation of the mining such as investigation of interesting and unknown knowledge set from real world databases. Following is the list of main problems or issues related to the data mining along with their solution or algorithms [18]:

- Massive datasets and high dimensionality.
- Over decent and assessing the statistical significance.
- Understandability of patterns.
- Non-standard partial data and data integration.
- Mixed changing and redundant data.

Data mining is a process of extracting the interesting information from the large amount of data sets available. Data mining particularly perform the following tasks:

- Classification

- Estimation
- Prediction
- Affinity grouping or association rules
- Clustering
- Explanation and visualization

Classification, Estimation and Prediction is performed for directed data mining. Directed Data Mining is a term which describes the process when the given data in database is used to create patterns or model that defines the single or multiple meaningful attributes as compare to rest of the attributes [21].

II. CLASSIFIERS OF DATA MINING

There are various techniques used for data mining. Some of them are explained as follows:

- Statistics
- Artificial Intelligence technique for data mining.
- Decision Tree approach.
- GA i.e. Genetic Algorithm.
- Visualization.

Statistics: It is an important process to perform for data mining. It is used for selection of data and extracting knowledge. Statistics is performed on the output of the data mining in order to separate the meaningful data from irrelevant data. Statistics is used in the data cleaning in order to clean the data by locating the values which are irrelevant and outliers or does not belongs to the data sets.

Artificial intelligence (AI) techniques: Data mining also uses the concept of AI i.e. Artificial Intelligence. The techniques like machine learning, neural Networks come under the topic of Artificial Intelligence. There is a list of some other techniques also such as pattern recognition, Knowledge presentation, Knowledge Acquisition also a type of artificial intelligence techniques and perform by data mining. Classification leads to the major lacking point in data mining. The classification is a process in which the data in database is classified on some basis such as mutual exclusion. Mutual exclusion refers to the how close the members of each group to each other and how far they are from members of another groups [16].

Decision tree approach: Decision tree method works on the basis of tree structure. The decisions are presented in the form of tree structure. Decision tree structure is useful for decision support systems. There are some methods which are specifically used for decision tree such as follows:

- CART
- CHAID

CART stands for Classification and Regression Tree. CHAID stands for Chi Square Automatic Interaction Detection. Both of these techniques are used for data classification.

Genetic algorithm [1]: The idea behind the development of Genetic Algorithm is Darwin's Theory of evolution. A population of rules is used to find out the solution of specific problems. The population rule is created initially and randomly. A genetic algorithm (GA) is a hunt procedure utilized as a part of figuring to discover correct or estimated answers for improvement and pursuit issues. [13] Hereditary calculations are sorted as worldwide hunt heuristics. Hereditary calculations

are a specific class of developmental calculations (otherwise called transformative reckoning) that utilization methods motivated by transformative science, for example, legacy, change, determination, and hybrid (additionally called recombination).Genetic calculations are actualized as a PC reproduction in which a populace of dynamic representations (called chromosomes or the genotype or the genome) of competitor arrangements (called people, animals, or phenotypes) to an enhancement issue advances toward better arrangements.

Visualization: refers to the pictorial representation of any problem, system or technique. It provides the batter understanding to the problem or system. Visualization plays an important role in data mining also. Data mining by using visualization can be achieved with the help of human brain or help.

III. Related Work

[1] **Neha Sharma(2016) [1],** This paper describes the importance of data mining in various fields and author implements the data mining process for kidney disease prediction. The proposed work is based on the combination of two techniques i.e. Neuro and fuzzy approach of data mining hence the proposed work is named as Neuro-fuzzy system. The proposed method work on the basis of mathematical computations and theory of probability. The results of proposed work prove the efficiency of the work.

[2] **Dr. T.Karthikeyan [2],** This paper describes a technique based on Ant colony Optimization known as Ant Miner. To discover rules in the database MAX-MIN ant system is optimize through Ant-Miner+. As a result, soil classification has been done based on different characteristics of different category of soil. In this paper, proposed technique is used to compare Ant miner and Ant miner+ algorithm where evaluation has been performed on training and soil dataset to association rule. Results shows that Ant-Miner + is a better approach than Ant miner.

[3] **S. Vaishnavi(April, 2014) [3],** Data mining has been used in various fields for several purposes due to its extraction of useful knowledge property. Thus in this paper, prediction of diseases in health care industry has been described where medical data is extracted through data analysis tools so that useful knowledge can be accessed and used. As medical data is flourishing source of information due to which it is mandatory to keep it up to date and extraction of data at the right time is also an important part of medical industry. Thus in this project the main focus is on the medical decision based upon the symptoms of the diseases that look similar and rare so that a good decision can take by the doctors. Diagnosis Decision Support System that can take patients data and can result into an appropriate prediction. This system is use to extract hidden knowledge from a heart disease database recorded earlier. This model is also helpful in answering the complex queries based on its own strength.

[4] **Jyoti Soni (June, 2011) [4],** health care systems are rich in information or data but poor in knowledge. Data that has been generated from health care systems are huge in volume but there is no availability of the tool which can discover the hidden relationship between data. Thus this paper focuses on GUI based interface which helps to enter the patient record with a single click and helps to evaluate that whether patient is infected with the heart disease or not. Method that follows this criterion is known as weighted association rule which is based on classifier. It is also called as Weighted Associative Classifier (WAC). Prediction is performed on the basis of the data stored in the database named as data repository. WAC helps to initialize weight to different attributes according to their predicting capability. Thus it has proven than proposed classifier or associative classifier performs better than traditional classifiers (decision tree and rule induction). Experiments has been performed which results that WAC accuracy is better than existing associative classifiers. Rule base generates through WAC acts as data mining technique. Java platform has been used for the implementation of the system and trained using benchmark data obtained from UCI machine learning repository. Thus system is flexible for new dataset.

[5] **Hongjun Lu(1996) [5],** Data mining is emerging as one of the active research field for many researchers . In this paper the author has presented the approach for the data mining. In this the neural networks have been introduced that is used for the

classification purpose as the classification of the data is one of the major problems in the data mining. The rules that are called as the symbolic rules are classified with the help of the neural network that is used data mining. Earlier these systems were not considered suitable for the data mining purpose. In this approach with the help of the neural network the symbolic rules are extracted these rules will help in classification. This proposed approach is divided into various steps. The network is first trained so that the accuracy required is achieved after that an pruning algorithm is used that will remove the redundant connections that are present in the network. After this the values of the hidden units that are present in the network is analyzed and by using this the classification rules are generated that are obtained by using the results of the analysis done earlier. from the results obtained from the experiments performed it is concluded that the proposed algorithm is quite efficient in solving the data mining problem .

[6] **Ghosh, Soumadip, et al. [6],** In this study the author represents a technique for classification named as novel Neuro fuzzy classification for data mining. In this the input to the system is fuzzyfied by using a function which is generalized function. In this the proposed technique uses a matrix which is the representation of degree of membership of various classifications. The pattern is classified on the basis of degree of membership. The proposed technique is applied to the ten benchmark data sets which is from UCI learning classification. The main aim of the proposed technique is to evaluate the technique by comparing this technique with RBFNN i.e. Radial Basis Function Neural Network and Adaptive Neuro-fuzzy Inference System. The performance of this technique is analyzed on the basis of various parameters like false positive rate, accuracy, root-mean square error, precision recall etc. In this paper the each and every parameter of proposed technique to prove the efficiency is compared with the ANFIS and RBFNN only.

IV. CONCLUSION

Data mining is used in various fields and medical field is one of those areas. The implementation of data mining in medical area is a trendy topic for research work now days. Hence this paper highlights implementation of data mining in predicting the kidney related disease. The data mining is also used for predicting various disease like heart disease, cancer etc. The disease prediction can be done by using various classifiers and pattern evaluation techniques. Hence this work can provide an introduction to the data mining in disease prediction and helpful to various researchers to select the disease prediction as a topic of their keen interest.

After having a review over related work it is concluded that in future various classifiers and association rules can be applied to the data mining in order to implement the accurate disease prediction.

References

1. Neha Sharma, ”Prediction of Kidney disease by using Data Mining Techniques”, IJARCSSE, vol 6(9), Pp 66-70, 2016.
2. Dr. T. Karthikeyan , “ A Study on Ant Colony Optimization with Association Rule” IJARCSSE, vol 2(5), Pp 486, 2012.
3. S. Vaishnavi, “Artificial Intelligence Approach for Disease Diagnosis and Treatment”,IJIRCCE, Vol. 2, Issue 4, pp 4000-4007, 2014.
4. Jyoti Soni Junw, “Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers”, IJCSE, Vol. 3 No. 6, pp 2385-2392, 2011.
5. Hongjun Lu, “Effective Data Mining Using Neural Networks” IEEE TRANSACTIONS ON knowledge and data engineering, vol. 8, NO. 6,Pp 957- 961, 1996.
6. Ghosh, Soumadip, "A Novel Neuro-Fuzzy classification technique for data mining." Egyptian Informatics Journal 15.3, Pp129-147, 2014.
7. K. Amarendra, “Research on Data Mining Using Neural Networks” Special Issue of International Journal of Computer Science & Informatics (IJCSI), , Vol.- II, Issue-1, 2, Pp2231–5292.
8. Dr. Yashpal Singh, “Neural Networks in Data Mining” Journal of Theoretical and Applied Information Technology, Pp 3742, 2005.
9. Rafael S. Parpinelli, “Data Mining with an Ant Colony Optimization Algorithm” IEEE, Vol 6(4), Pp 321-332, 2002.
10. Dr. S vijayarani, ”Data mining classification algorithm for kidney disease prediction”, IJCI, Vol 4(4), Pp 13-22, 2015.
11. Bijaya Kumar Nanda, “Class Based Rule Mining using Ant Colony Optimization” Special Issue of International Journal on Advanced Computer Theory and Engineering (IJACTE), Vol 2(3), Pp 25-30,2013.
12. Lambodar jena, ”distributed data mining classification algorithm for prediction of chronic kidney disease”, IJERMT, vol 4(11), Pp 1-9, 2015.

13. K.Rajeswari, "A Novel Risk Level Classification of Ischemic Heart Disease using Artificial Neural Network Technique - An Indian Case Study "ternational Journal of Machine Learning and Computing, Vol. 1, No. 3,Pp 231-235, 2011.
14. Miss. Manjusha B. Wadhonkar, " Classification of Heart Disease Dataset using Multilayer Feed forward backpropogation Algorithm " (IJAIEM) Volume 2, Issue 4,Pp 214-220, 2013.
15. Manish kumar, "Prediction of chronic disease using random forest machine learning algorithm", IJCSMC, Vol 5(2) Pp 24-33, 2016.
16. V.Kiruba, "Survey on data mining algorithms in disease prediction", IJCTT, vol 38(3), Pp 124-128, 2016.
17. D.Shanthi, "Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke" International Journals of Biometric and Bioinformatics (IJBB), Volume (3) : Issue (1),Pp 10-18, 2009.
18. Beant Kaur, "Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC, Volume: 2 Issue: 10, pp 3003-3008, 2014.
19. Abdel-Motaleb, "Artificial intelligence algorithm for heart disease diagnosis using Phonocardiogram signals", IEEE, ISSN :2154-0357, pp 1-6, 2012.
20. Ankita R. Mokashi, "Review on Heart Disease Prediction using Ann and Classifier", IJERT, Vol 4(9), 2015.
21. M. Akhil Jabbar, "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection", Global Journal of Computer Science and Technology Volume 13(3), pp 5-14, 2013.
22. R.Chitra, "Heart Attack Prediction System Using Fuzzy C Means Classifier", IOSR-JCE, Volume 14, Issue 2, pp 23-31, 2013.
23. Dimple , "Comparative Study of Different Techniques for Heart Disease Prediction System", SSIJMAR, Vol. 2, No. 2.
24. Neelamadhab Padhy, "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3,Pp 43-58, 2012.
25. Anand V. Saurkar, "A Review Paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4,Pp 98-101, 2014.
26. Abhishek, "Heart Disease Prediction System Using Data Mining Techniques "ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY Vol. 6, No. (4): Pgs. 457-466, 2013.
27. Boshra Bahrami, " Prediction and Diagnosis of Heart Disease by Data Mining Techniques "Journal of Multidisciplinary Engineering Science and Technology (JMEST) ,Vol. 2 Issue 2,Pp 164-168, 2015.
28. Deepali Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1678-1680 ,2014.

