ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

A DETAILED SURVEY ON MEDICAL DATASETS APPLYING DATA MINING TECHNIQUES

Avinav Baruah
E-mail Id: avinavbaruah236@gmail.com
Department of Computer science and engineering, JKLU, Jaipur-302026, India

Abstract-Data mining is the extraction of knowledge from massive amounts of data. The data mining is applied in the medical and health sectors. The most often used data mining technique is classification. Classification is one of the data mining tasks, which is involved in assigning the object to predefined categories. There are various data mining classification techniques, which includes: Decision tree, K nearest neighbor (KNN), Support vector machine etc. The aim of the paper is to do survey of various medical datasets. The different data mining classification techniques are studied in the survey work.The various data mining tools which includes: R, RapidMiner, Matlab, Weka etc. are involved in the survey. From the survey work, it can be concluded that the data mining techniques play an important role in data mining, and they are applied in the medical and health fields.
Keywords: Classification, Rapid Miner, Weka

1. INTRODUCTION

Data mining is based on 'mining' of useful information from huge quantities of data. It is used in various fields such as: medical and health, insurance, education, manufacturing, banking etc.
Classification is the most commonly used data mining technique. Classification is a type of data mining task, which is based on assigning the object to predefined categories, by constructing a model. The model is linked with one or more categorical variables.
There are different types of data mining classification techniques, which includes: Decision tree, K nearest neighbor (KNN), Support vector machine etc.

2. SURVEY ON MEDICAL DATASETS USING VARIOUS DATA MINING TECHNIQUES

P. Suganya [1] proposed a novel metaheuristic data mining algorithm so as to detect and classify the Parkinson disease. The research makes us of various performance measure techniques: confusion matrix, precision, recall and error rate. Also, the confusion matrix is applied by making use of different attributes such as sensitivity, specificity, accuracy, positive and negative predictive values. The research work also involves comparison of five classification algorithms which includes: SCFW with KELM, ABO, FCM, PSO, ACO algorithms. The author concluded from the results that the ABO algorithm has the highest accuracy, sensitivity, specificity among all algorithms used in the study.
H. Ganesh [2] evaluated the accuracy level of Parkinson's disease of various classification algorithms. Here, four classification techniques are applied which includes: Decision table, J48, Naive bayes, and Random tree. The data mining tool used here is Weka. Comparison of the classification techniques are done, so as to find out which technique shows the highest accuracy. The accuracy of the classification technique is found out by applying the Confusion matrix. From the results, it can be observed that the random tree classification technique gives the best accuracy in comparison to all the classification techniques used.
H. Ganesh [3] did a survey of data mining techniques by applying the Parkinson's disease. The data mining techniques find application in the areas of medical, science, railway, business etc. They are also applied in cases of medical diagnosis and prediction of diseases. Various data mining techniques are used in Parkinson's disease.
T.V.S. Sriram [4] studied the voice characteristics and identification techniques which are applied to recognize PD people based on their voices. The work also explains the biometric methods along with their pros and cons. ParkDiag is a tool used to diagnose and predict the PD. It is a simple tool. Data mining techniques were used in the work. KStar and ADTree gives 100%, Naïve bayes shows 83% and Bayes Net gives 70%.
S.U. Khan [5] found an accurate model for detection of disease. The research work is based on doing data preprocessing, which includes data cleaning, recovering missing values. Next, data transformation is done. After that, clustering techniques are applied, which includes: KNN, Random Forest, Ada-Boost. Matlab and Weka are used here. From the results, it is seen that K-NN is the most appropriate technique for classification. Its accuracy is 90.26%.

R.G. Ramani [6] conducted a survey on data mining techniques for knowledge discovery in databases, involved in the classification of Parkinson Disease. Various classification techniques applied to the dataset are compared. Accuracy Analysis and feature relevance analysis are used here. The aim is to find the best classification technique. From the results, it is seen that Random Tree classifier gives 100% accuracy.

M. Metkari [7] proposed an approach for improving the accuracy and efficiency for heart disease diagnosis, involved in data mining. In this work, genetic algorithm and artificial neural network are used. For the improvement of accuracy of independent classifiers, discretization technique is applied. The ANN with K-means discretization shows the highest accuracy. ANN with discretization provides the minimum error rate.

O. Chandrakar [8] focused on the comparison of the classification techniques used for the prediction of blood glucose level. To develop the model for classification, two cases are included: Records of patients are taken into account; records of patients are not taken into account. According to the results, when patient's records are taken into account: more accurate prediction and false Negative cases are less.

K.A. Shakil [9] discussed about many data mining algorithms that have been used for the prediction of dengue disease. In the work, Weka with 10 cross validation is applied, for the evaluation of data and comparison of results. Weka consists of large collection of various machine learning and data mining algorithms. Here, the classification of the dengue data set is done. Then the comparison of the various data mining techniques is done in Weka using Explorer, knowledge flow and Experimenter interfaces. The primary aim of the research work is the classification of data. It also helps the users to mine relevant information from data and to find an appropriate algorithm for accurate predictive model. From the results, it is seen that Naïve Bayes and J48 are the top most performance techniques for classified accuracy since they show highest accuracy i.e. 100%. The highest ROC = 1. It gives lowest mean absolute error.

G. Kaur [10] applied effective data mining method for the prediction of diabetes using medical records of patients. The modified J48 classifier is employed for increasing the accuracy rate of the data mining method. Weka is applied as an API of Matlab so as to produce the J-48 classifiers. From the results, it is seen that the modified J48 classifier is more improved than the existing J-48 algorithm. Also, it is observed that the algorithm used in the work shows accuracy up to 99.87 %.

K.R. Ananthapadmanaban [11] focused on applying various classification techniques used in data mining. Then, the comparison of the data mining classification techniques are done. The tool used in the work is RapidMiner. Naive bayes and Support Vector Machine are employed for the prediction of the early detection of eye disease diabetic retinopathy. From the results, it is observed that Naive bayes is more efficient than SVM. Sensitivity and specificity are applied to measure the performance.

R. Jothikumar [12] compared the accuracies of the classification techniques, which includes Random Tree, Naïve Bayes, Decision Tree and Random forest. RapidMiner is the tool used in the work. Here, the input to the classification techniques are the datasets collected. The results are evaluated. It is seen that Naïve Bayes has the highest accuracy of 79.25%.

T. Sharma [13] focused on the application of the different classification techniques of data mining. The tools applied are Weka and RapidMiner. The dataset used is the public healthcare dataset. The measurement parameters used in the work are: accuracy rate and error rate. Using the parameters, it can be said that the classifier possessing a higher accuracy rate and lower error rate classify the dataset more precisely and vice-versa. From the results, it can be said that the Decision Tree algorithm in RapidMiner is the top most classification method for the work.

S. Sharma [14] emphasized on the application of different machine learning algorithms to a medical diagnosis problem. The problem that is taken for the work is that of the chronic kidney disease diagnosis. Twelve classification algorithms are applied to the chronic kidney disease dataset, and then analyzed. There was a comparison made between the results of the candidate methods used for prediction and the actual medical results of the subject. This was done so as to find out the efficiencies. Different metrics are applied, so as to do performance evaluation, which includes: predictive accuracy, precision, sensitivity and specificity. From the results, it is observed that decision-tree is the best technique used. It gives the accuracy of 98.6%, sensitivity of 0.9720, precision of 1 and specificity of 1.

E. Venkatesan [15] focused on applying the various classification techniques namely J48, Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree) and Best First Tree (BF Tree) to the breast cancer dataset, and then analyzing them. The accuracy measures used in the work are: FP rate, TP rate, Recall, Precision, ROC Area and F-measure. From the results, it is observed that the J48 classifier has the highest accuracy of 99%. So, it is the best technique among all other techniques used in the research work.

K. Saravananathan [16] emphasized on the analysis of the classification techniques J48, CART, SVMs, and KNN, on the Diabetes medical dataset. Various performance indicators i.e. accuracy, specificity, sensitivity, precision, error rate are found out. The use of accusation along with a suitable data preprocessing technique can yield better

accuracy for the classifier. Data normalization improves the J48 performance. From the results, it is observed that J48 classification technique is better than all other techniques used in the work.

S.K. Devi [17] emphasized on analyzing the different data mining techniques used to predict heart disease and diagnosis. A heart disease prediction model helps to detect the heart disease status, with the help of the clinical data of the patients. The various data mining classification techniques used in the work are: Decision trees, Naive Bayes, Neural Networks and Support Vector Machines. Combination of any of these algorithms are used to make decisions faster and more accurate.

S.P. Shukla [18] emphasized on using the two techniques, namely, K-means and fuzzy C-means and applying them to the cancer dataset. The cancer data is classified using two different methods and comparison of the classification techniques are done. The two methods applied for data classification are: Supervised and unsupervised. Comparison of EBPA and FCM is done, taking performance into account. FCM gives less performance since it uses unsupervised manner of classification.

N. Amin [19] compared the various classification techniques, applying Weka. It is also seen which technique is the best for hematological data. From the results, it is seen that the best classification technique is J48. It gives accuracy of 97.16%. The time required to construct the model is 0.03 seconds. Naïve Bayes gives the lowest average error of 29.71%.

V. Pellakuri [20] emphasized on analyzing the performance of various classification techniques using Orthopaedic dataset. The work is based on the data mining techniques, which are employed in medical research, especially in orthopaedic diagnosis. Various classification techniques are compared by applying two data mining tools: Weka and Tanagra. Prediction of the orthopaedic problems is done by applying twenty algorithms. The accuracy of the algorithms are estimated. The attribute ranking is developed so that a decision can be made on the orthopaedic problems. It is seen that the results have more accuracy in Tanagra in comparison to Weka.

T.A. Shaikh [21] measured the performance of the algorithms, which include: Naïve bayes, Decision tree and Artificial neural network, on the medical datasets. From the results, it is observed that in case of Parkinson's disease, Artificial neural network gives the highest accuracy. In case of Primary tumor, Naïve bayes shows the greatest accuracy.

Saloni [22] emphasized on classifying the healthy people from the people affected by Parkinson's disease, by applying data mining of voice features. The classifier used in the work is support vector machine. Different subsets are prepared using the voice features. The important features of the classification are: DFA (Detrended fluctuation analysis) and PPE (pitch period entropy).

S. Vijayarani [23] laid emphasis on the analysis of the performance of three classification rule generation algorithms, which include C4.5, PART, RIPPER. The algorithms are involved in producing both sensitive and non-sensitive classification rules. The datasets used are: breast cancer dataset and heart disease dataset. From the results, it is observed that PART algorithm performs better than all the algorithms used in the research work.

S. Joshi [24] compared the two classification techniques i.e. J48 and Naïve bayes. J48 is linked with decision tree. Naïve bayes is linked with probability. The dataset used is diabetes dataset. The parameters applied are accuracy and cost analysis. From the results, it is seen that Naïve bayes technique is better than J48. Naïve bayes has higher accuracy. It takes less time to construct the model, when using Naïve bayes technique.

L. Jena [25] laid emphasis on the prediction of chronic kidney disease. Six classification techniques have been used and compared in the work, which include: Decision table, J48, Conjuctive rule, Multilayer perceptron, SVM and Naïve bayes. The data mining tool applied in the paper is Weka. From the results, it is seen that Multilayer perceptron shows highest accuracy and best performance for prediction. So, it is the best classification technique, and is used for the prediction of chronic kidney disease.

CONCLUSION

In this paper, a detailed survey on medical datasets is done, by applying data mining techniques. Initially, a brief explanation about data mining and classification is given. Next, the paper focuses on surveying medical datasets using different data mining techniques. The survey work is based on the study of 25 research papers. The research papers apply different data mining techniques on various medical datasets. From the survey, it is seen that the data mining techniques play a vital role in data mining. The techniques have applications in the medical and health sectors.

REFERENCES

[1] P. Suganya, C.P. Sumathi , "A Novel Metaheuristic Data Mining Algorithm for the Detection and Classification of Parkinson Disease," Indian Journal of Science and Technology, Vol 8(14), DOI: 10.17485/ijst/2015/v8i14/72685, July 2015.

[2] H. Ganesh, G. Annamary, "Comparative study of Data Mining Approaches for Parkinson's Diseases," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 9, September 2014.
[3] H. Ganesh, G. Annamary, "A Survey of Parkinson's Disease Using Data Mining Algorithms," Hariganesh S et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 4943-4944.
[4] V.S. Sriram, M.V. Rao, G.V. SatyaNarayana, D.S.V.G.K. Kaladhar, "ParkDiag: A Tool to Predict Parkinson Disease using Data Mining Techniques from Voice Data," International Journal of Engineering Trends and Technology (IJETT) – Volume 31 Number 3- January 2016.
[5] S.U. Khan, "Classification of Parkinson's Disease Using Data Mining Techniques," Parkinsons Dis Alzheimer Dis July 2015 Vol.:2, Issue:1.
[6] R.G. Ramani, G. Sivagami, "Parkinson Disease Classification using Data Mining Algorithms," International Journal of Computer Applications (0975 – 8887) Volume 32– No.9, October 2011.
[7] M. Metkari, M. Pradhan, "Improve the Classification Accuracy of the Heart Disease Data Using Discretization," International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 10, Volume 2 (October 2015).
[8] O. Chandrakar, J.R. Saini, "Comparative Analysis of Prediction Accuracy of General and Personalized Datasets Based Classification Model for Medical Domain," International Journal of Advanced Networking Applications (IJANA).
[9] K.A. Shakil, S. Anis, M. Alam, "Dengue disease prediction using Weka Data mining tool."
[10] G. Kaur, A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," International Journal of Computer Applications (0975 – 8887) Volume 98 –No.22, July 2014.
[11] K. R. Ananthapadmanaban, G. Parthiban, "Prediction of Chances - Diabetic Retinopathy using Data Mining Classification Techniques," Indian Journal of Science and Technology, Vol 7(10), 1498–1503, October 2014.
[12] R. Jothikumar, R.V. Sivabalan, "Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies," Middle-East Journal of Scientific Research 24 (Recent Innovations in Engineering, Technology, Management& Applications): 200-206, 2016.
[13] T. Sharma, A. Sharma, V. Mansotra, "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2016.
[14] S. Sharma, V. Sharma, A. Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis".
[15] E. Venkatesan, T. Velmurugan, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification," Indian Journal of Science and Technology, Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/84646, November 2015.
[16] K. Saravananathan, T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining," Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874, November 2016.
[17] S.K. Devi, S. Krishnapriya, D. Kalita, "Prediction of Heart Disease using Data Mining Techniques," Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016.
[18] S.P. Shukla, R. Dwivedi, "Clustering and Classification of Cancer Data Using Soft Computing Technique," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278- 0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. I (Jan. 2014), PP 32-36.
[19] N. Amin, A. Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data," American Journal of Engineering Research(AJER) e-ISSN : 2320-0847 p-ISSN : 2320-0936 Volume-4, Issue-3, 2015, pp-55-61.
[20] V. Pellakuri, D. Gurram, D.R. Rao, M.R.N. Rao, "Performance Analysis and Optimization of Supervised Learning Techniques for Medical Diagnosis Using Open Source Tools," VidyullathaPellakuri et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 380-383.
[21] T.A. Shaikh, "A Prototype of Parkinson's and Primary Tumor Diseases Prediction Using Data Mining Techniques," International Journal of Engineering Science Invention ISSN(Online): 2319 – 6734, ISSN (Print): 2319 – 6726, Volume 3 Issue 4, April 2014, PP.23 – 28.
[22] Saloni, R.K. Sharma, A.K. Gupta, "Detection of Parkinson Disease Using Clinical VoiceData Mining," International Journal Of Circuits, Systems And Signal Processing,Volume 9, 2015.
[23] S. Vijayarani, M. Divya, "An Efficient Algorithm for Generating Classification Rules,"IJCST Vol. 2, Issue 4, Oct . - Dec. 2011.

[24] S. Joshi, B. Pandey, N. Joshi, " Comparative analysis of Naive Bayes and J48Classification Algorithms," International Journal of Advanced Research in ComputerScience and Software Engineering, Volume 5, Issue 12, December 2015.

[25] L. Jena, N.K. Kamila, "Distributed Data Mining Classification Algorithms forPrediction of Chronic- Kidney-Disease," International Journal of Emerging Researchin Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11), November2015.