Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 9, Issue. 1, January 2020, pg.58 – 64

Knowledge Discovery in Heart Dataset using Classification Techniques

Suresh Kumar Pandey*; Dr. Bharat Mishra*; Dr. S.S.Gautam*
suresh.kumarpandey@rediffmail.com
*Department of Physical Sciences, Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, Satna (M.P.)-India

ABSTRACT: Data mining plays most important role in area of research with the goal of finding meaningful information from huge amount of data. Presently, data mining techniques and tools are used by researchers in the field of healthcare, particularly for prediction of disease. The focus of this paper is use of data mining techniques on healthcare issues, benefits, application and use on healthcare domain. We have created our dataset of heart patients with 2139 instances and analysed the dataset with WEKA engine and discussed the result and conclusion. We have proposed a new classifier model in order to minimize misclassified instances with adopting Hybrid Approach. After the evaluation of proposed hybrid classifier model we found that proposed algorithm is providing accuracy about 97.66%.

1. INTRODUCTION

Data mining is one of the most powerful and emerging technologies which is used to mine certain useful trends and patterns which are unknown, to enhance the performance of the organizations. Almost all the organizations are growing rapidly with the help of data mining functionalities. Data mining helps in finding out something in the massive data which is unknown and most profitable to the organization [1].

According to [2], Data Mining is defined as extracting information from vast sets of data. In other words, data mining is the procedure of mining knowledge from data. Information Industry have enormous data available. Which is of no use until it is converted into useful information. To get useful information we require to analyze this huge amount of data. Apart from extraction of information we also need to perform other data mining processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation”[2].

In Data mining, knowledge extraction or discovery is done in seven sequential steps viz. Data cleaning, Data integration, Data Selection, Data transformation, Data Mining, Pattern evaluation, and

Knowledge representation. The goal of knowledge discovery and data mining process is to discover the patterns that are unknown among the huge set of data and interpret useful knowledge and information[3]**.**

Data mining process is extraction of information from large data sets and transforms it into some understandable form for further uses. So it helps to achieve the specific objectives. The goal of a data mining effort is normally either to create a descriptive model or a predictive model [4]. A descriptive model presents the data in concise form which is essentially a summary of the data points, finds patterns in the data and understands the relationships between attributes represented by the data.

The descriptive model consists of tasks such as Clustering, Association Rules, Summarizations, and Sequence Discovery. The predictive model basically works by making a prediction about values of data, which uses known results found from different datasets [5]. Thus, predictive data mining model uses classification, prediction, regression and analysis of time series.

2. DATA MINING IN HEALTH CARE

Today Data mining is considered as one of the most challenging and top most research area in “Health Care” all over the world due to high importance of healthcare issues. Healthcare is an active research area due to its large scale potential and its impact on human. There are very large and gigantic data available in clinics, hospitals and medical institutions. Therefore, there is a need of dominant automated Data Mining tools for analysis and explanation of useful information from this data. This extracted useful information is very precious for healthcare specialist to understand the reason of diseases and for providing better and cost effective treatment to patients.

Extraction of healthcare dataset using Data Mining tools provides precious and sometimes hidden novel information regarding healthcare which in turn helpful for making administrative as well as medical decision which includes estimation of medical staff, decision regarding health insurance policy, choice of treatments, disease prediction *etc.*.

Data mining techniques are also used for both analysis and prediction of various diseases. Some research work proposed an enhancement in available Data Mining methodology in order to improve the result and some studies develop new methodology and framework for healthcare system. Many Data mining techniques such as classification, clustering and association are used by healthcare institutes to augment their capability for decision making on the subject of patient health. There are plenty of research resources available regarding Data Mining tasks which are presented with their advantages and disadvantages[6].

3. Related Work

As discussed above, Healthcare data is massive. It is related to patient centric data, resource management data and transformed data. Analysis of this massive data using data mining techniques can be used by Healthcare organizations in answering several important and critical questions related to health care.

Manish Shukla and Sonali Agarwal examined *a*pplication of classification technique and Data clustering machine learning approach. They presents an approach for centroid selection in k-mean algorithm for health datasets which gives better clustering results in comparison to traditional k-mean algorithm[7].

Rita Samikannu et al., examined Classification analysis to support medical diagnosis, improving quality of patient care. They propose a feature selection approach for finding an optimum feature subset that enhances the classification accuracy of Naive Bayes classifier [8].

K.Prasanna Lakshmi & C.R.K.Reddy enlightened, Associative Classification approach which combines associative rule mining and classification. According to them associative classifiers are useful for application where maximum predictive accuracy is preferred. They proposed an efficient technique for heart disease prediction. They build a classifier with prediction rules of high interestingness values and results show that this work helps doctors in their diagnosis decisions [9].

Ranganatha S. et al., stored medical information of patients who come for hospitalization for heart disease and data mining algorithms are run on that information and result has been provided in the form of user understandable words and graph. They claim that ID3 outputs the result in the form of decision tree which is easy for understanding. Naïve Bayesian predicts the chances of heart disease based on given conditions[10].

Monika Gandhi & Shailendra Narayan Singh used different data mining techniques for prediction of heart disease. In their study, they have used and analysed data mining methods namely, Naive Bayes, Neural network and Decision tree algorithm on medical data sets[11].

Eiko Kai1 et al., used association rule technique to find common set of rules in order to build a clinical decision support system. They also showed examples of the meaningful information from the analysed data to build a better clinical decision support system[12].

In the paper [13] Aqueel Ahmed et al., proposed to find out the heart diseases through data mining techniques : Support Vector Machine, Genetic Algorithm, rough set theory, association rules and Neural Networks. They conclude that decision tree and SVM are most effective methods for the heart disease.

Aastha Joshi & Rajneet Kaur compares six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, OPTICS , and STING[14].

Malarvizhi and Dr. S. Ravichandran [15] considers the diabetic and breast cancer patients details for evaluating clustering algorithms that can be used in early diagnosis. Sometimes because of poor initialization once troids, k-means may give poor results. According to them Kmeans++ overcomes this problem by proposing a balanced initial value and is said to successfully overcome problems associated with initial defining of cluster-centers for k-means. k-means++ is an enhanced K-means. They have also analyzed predictions in healthcare domain using clustering to discover new range of information retrieval, specifically proving the usefulness of k- means++ clustering[15].

K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhan discussed and examined the potential use of classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihood of patients getting a heart disease [16].

According to Arvind Sharma and P.C. Gupta, Data mining can contribute with important benefits to the blood bank sector. They have used J48 algorithm and WEKA tool for their research. They claim that, classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% [17].

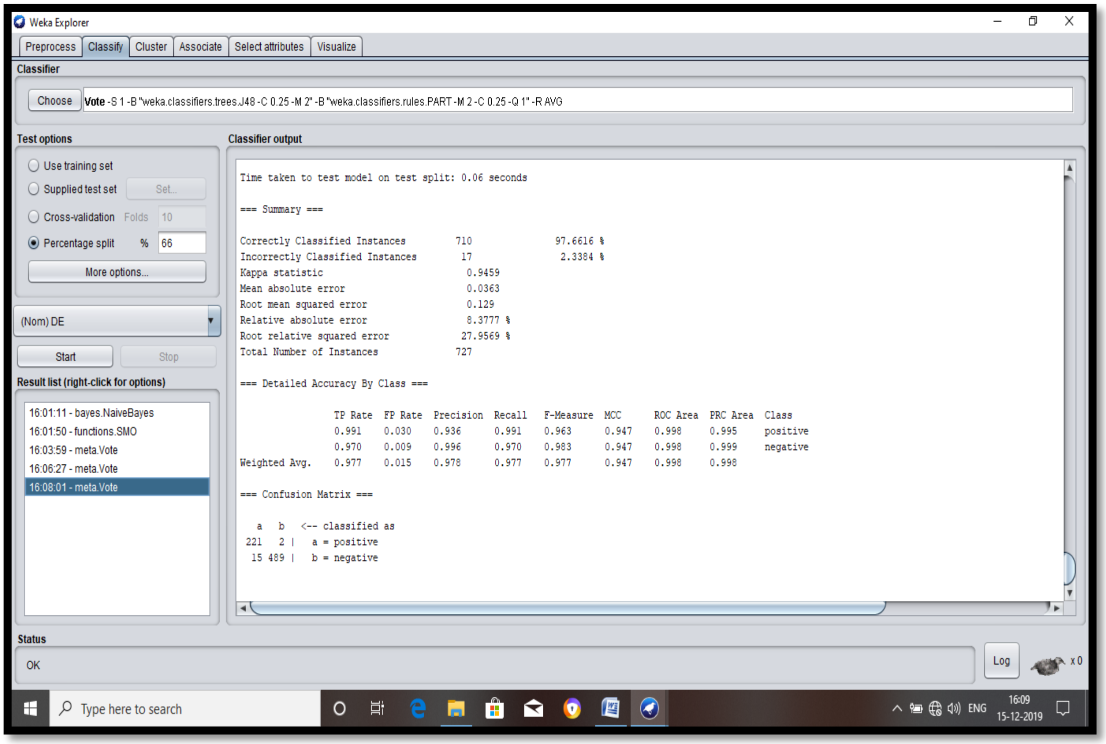
4. Experimental Result

For our experiment we have collected data from the clinic of Dr. A.P. Pathak, cardiologist of Rewa, MP-India. We have created dataset of 2139 patients of cardiac disorder. The heart dataset contains data of 1028 female and 1111 male patients. The age range of patients were 15 to 66 years. The 1456 patients were in disease state while others were negative. The data ware collected for nine attributes viz- Sex, Age, Chest Pain, Blood Presser, Blood Sugar, ECG, Heart rate, TMT and Disease.

After preprocessing of data, we used WEKA[18]. In WEKA engine for classification we have used following classifiers for prediction :

- Bayesian Network (Naïve Bayes)
- Support Vector Machine (SMO)
- C4.5 Decision Tree (J48)
- K-nearest Neighbour (IBK)
- Partial Decision Tree(PART)
- Proposed Vote(J48+PART)

Figure 1 shows the Weka output of classifier Vote(J48+PART).



Vote -S 1 -B "weka.classifiers.trees.J48 -C 0.25 -M 2" -B "weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1" -R AVG

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) DE

16:01:11 - bayes.NaiveBayes
16:01:50 - functions.SMO
16:03:59 - meta.Vote
16:06:27 - meta.Vote
16:08:01 - meta.Vote

```
Time taken to test model on test split: 0.06 seconds

=== Summary ===

Correctly Classified Instances         710               97.6616 %
Incorrectly Classified Instances        17                2.3384 %
Kappa statistic                          0.9459
Mean absolute error                      0.0363
Root mean squared error                  0.129
Relative absolute error                  8.3777 %
Root relative squared error             27.9569 %
Total Number of Instances              727

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.991    0.030    0.936      0.991    0.963      0.947    0.998     0.995     positive
                 0.970    0.009    0.996      0.970    0.983      0.947    0.998     0.999     negative
Weighted Avg.    0.977    0.015    0.978      0.977    0.977      0.947    0.998     0.998

=== Confusion Matrix ===

   a   b   <-- classified as
 221   2 |   a = positive
  15 489 |   b = negative
```

OK

Figure 1 : Heart dataset Weka output – classifier Vote(J48+PART)

Table 1 shows the confusion Matrix of Vote(J48+PART) Classifier for Heart dataset.

	Predicted			
	a	**b**	**Total**	**Actual**
	221	2	223	a= Positive
	15	489	504	b= Negative
Total	236	491	727	

Table 1 : Confusion Matrix of Vote(J48+PART) Classifier for Heart dataset

Time taken by different algorithm for generating the classifications (time taken to test model on test split) for heart dataset are given in Table 2.

Classifier Name	Time in Secon
Naïve Bayes	0.07
SMO	0.07
IBK	0.1
PART	0.1
J48	0.09
PROPOSED(PART+J48)	0.06

Table 2 : Time taken to test model on test split

In Table 3 we have shown the combined classifier instances (correctly and incorrectly classified) for different classifiers.

Classifier Name	Correctly Classified		Incorrectly Classified	
	Instances	%	Instances	%
Naïve Bayes	536	73.7276	191	26.2724
SMO	591	81.293	136	18.707
IBK	693	95.3232	34	4.6768
PART	704	96.8363	23	3.1637
J48	706	97.1114	21	2.8886
PROPOSED(PART+J48)	710	97.6616	17	2.3384

Table 3 : Combined classifier instances for Heart dataset

In our experiment attempt is made to discover the heart diseases through five classification model as Naïve Bayes, SMO (Support Vector Machine), IBK (K-nearest Neighbor), J48(C4.5 Decision Tree), PART (Projective Adaptive Rejunance Theory). For the evaluation of the above classification, we have used Matrix Accuracy, Precision and Recall. After making comparison it has been observed that classifier J48 provides good result for Heart diseases. The accuracy of J48 is 97.11% while the percentage of incorrectly classified instances is 3% (21 instances).

5. Conclusion

This study presents a systematic review of the application of Data Mining methods in healthcare domain, with a focus on the application and the techniques used to optimize the results. In this study we present an overview of the current research being carried out using the data mining techniques for various diseases. Analysis shows that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the algorithms together results more effective.

We have tried to develop a new classifier model in order to minimize misclassified instances with adopting Hybrid Approach. We have made a new classifier model by the combination of PART and J48 algorithms. After the evaluation of proposed hybrid classifier model we found that proposed algorithm is providing accuracy about 97.66% in which misclassified instances are only 2.3%(17 instances). Thus, from this hybrid approach, accuracy have been increased by 0.55% (97.6616-97.114=0.55) in comparison to J48 classification. In this way it is evident that our proposed hybrid classifier for heart diseases is more efficient in comparison to pre-defined approaches.

REFERENCES

1.	Priyanka Gupta Rajan Gupta, PhD “Data Mining Framework for IoT Applications”International Journal of Computer Applications (0975 – 8887) Volume 174 – No.2, September 2017.
2	Han, J.,Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006
3	Deepashri, “Survey on techniques of data mining and its applications ”, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-2) , 2017
4	Pradnya P. Sondwale, “Overview of predictive and descriptive data mining techniques” IJARCSSE, Volume 5, Issue 4, April 2015 .
5	Nikita Jain, Vishal Srivastava “Data mining techniques: a survey paper” IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013.
6	Prof. Samiksha H. Zaveri Ph. D and Dr. Narayan Joshi, “A comparative study of data analysis techniques in the domain of medicative care for disease prediction”, Volume 8, No. 3, March – April 2017 International Journal of Advanced Research in Computer Science .
7	Manish Shukla and Sonali Agarwal,”Hybrid Approach for Tuberculosis Data Classification using Optimal Centroid Selection Based Clustering”, IEEE, 2014
8	Rita Samikannu, Nickolas Savarimuthu, Ramaraj Narayanaswamy, Sarojini Balakrishnan, “SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases , IEEE , 2008
9	K.Prasanna Lakshmi ,Dr. C.R.K.Reddy, Fast Rule-Based Heart Disease Prediction using Associative Classification Mining”,IEEE International Conference on Computer, Communication and Control (IC4-2015).
10	Ranganatha S., Pooja Raj H.R., Anusha C., Vinay S.K.,”Medical data mining and analysis for heart disease dataset using classification techniques”, IET Conference Proceedinges, 2013.
11	Monika Gandhi , Dr. Shailendra Narayan Singh,” Predictions in Heart Disease Using Techniques of Data Mining “1st International Conference on Futuristictrend in Computational Analysis and Knowledge Management (ABLAZE 2015)IEEE,2015
12	Eiko Kai1, Andrew Rebeiro-Hargrave, Sozo Inoue, Yasunobu Nohara, Rafiqul Islam Maruf , Naoki Nakashima and Ashir Ahmed, “Empowering the healthcare worker using the Portable Health Clinic”, IEEE 28th International Conference on Advanced Information Networking and Applications,2014
13	Aqueel Ahmed, Shaikh Abdul Hannan, “Data Mining Techniques to Find Out Heart Diseases: An Overview “, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012
14	Aastha Joshi & Rajneet Kaur,”A Review: Comparative Study of Various Clustering Techniques in Data

	Mining, "International Journal of Advanced Research in Computer Science and Software Engineering, 2013
15	Ms. A. Malarvizhi and Dr. S. Ravichandran," Data mining's role in mining medical datasets for disease assessments – a case study", International Journal of Pure and Applied Mathematics ,Volume 119, No. 12 , 2018
16	K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks", International Journal on Computer Science and Engineering (2010).
17	Elias Lemuye,"HIV status predictive modeling using data mining technology", LAP LAMBERT Academic Publishing, 2012.
18	Waikato Environment for Knowledge Analysis (WEKA), developed at the University of Waikato, New Zealand, www.cs.waikato.ac.nz/~ml/weka/