Feature Selection and Diagnose of Healthcare Issues using Classification Algorithms

Vikas Mongia[1], Dr. Deepak Mehta[2]

[1]Research Scholar,[2]Assistant Professor

Guru Kashi University, Talwandi Sabo, Bathinda (India)

ABSTRACT

Health care is very basic need of everyone in today's society. Health care data includes patient data, their treatment and resource management data. Health care data grow exponentially day by day. So it becomes difficult to take the right decision at the right time from the large healthcare dataset. Data mining play a vital role to discover the hidden pattern of healthcare. In this paper systematic study on "data mining in healthcare" has been done. Various application of healthcare is thoroughly studied. To process the data, CRISP-DM model has been considered that consists of six phases. In this study nine data mining algorithms are executed on heart disease dataset. Quality of these algorithms is measured on the basis of eight parameters like correctly classified instance, incorrectly classified instance, execution time etc... The experiment is done using 10 fold cross validation method. The study has proven that accuracy of SMO classification algorithm is being considered highest i.e. 84.074% and minimum execution time is taken by random tree i.e. 0.005 seconds. Feature reduction technique is used for identifying and removing those attributes that do not contribute towards classification of the dataset. In this work, gain ratio technique is used to evaluate the worth of an attribute with respect to the class. Then ranker algorithm is used to arrange these attributes in descending order according to their gain ratio and last three attributes having zero gain ratio are removed without affecting the correctly classified instance.

Keyword: Healthcare, Data Mining, Classification, Feature Reduction, WEKA

I. INTRODUCTION TO HEALTHCARE

Healthcare is used to improve the patient care and reduce the cost. Sound growth of every nation depends upon the health of their citizen. Healthcare covers complex processes of the diagnosis, treatment, and prevention of disease, injury, and other physical and mental impairments in humans [1]. The healthcare industries in most countries are evolving at a rapid pace. The healthcare industry can be regarded as place with rich data as they generate massive amounts of data including electronic medical records, administrative reports and other benchmarking finding [2].These healthcare data are however being under-utilized. Data mining in healthcare are being used mainly for predicting various diseases as well as in assisting for diagnosis for the doctors in making their clinical decision.

II. DATA MINING IN HEALTH CARE

Data mining in healthcare has become increasingly popular because it offers benefits to care providers, patients, healthcare organizations, researchers, and insurers.

Care providers can use data analysis to identify effective treatments and best practices. By comparing causes, symptoms, treatments, and their adverse effects, data mining can analyze which courses of action are most effective for specific patient groups. It can also identify clinical best practices to help develop guidelines and standards of care.

Patients can receive better, more affordable healthcare services. This is especially true when healthcare managers use data mining applications to identify and track chronic diseases and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims.

Healthcare organizations can use data mining to make better patient-related decisions. For instance, it provides information to guide patient interactions by determining patient preferences, usage patterns, and current and future needs—all of which helps to improve patient satisfaction. With healthcare organizations under increasing financial pressure, data mining can also influence revenues, costs, and operating efficiency while maintaining high-quality care.

III. REVIEW OF LITERATURE

Getew sahle [5] applied J48 Decision tree and rule induction JRIP algorithm on a data set of 6568 records from Ethopia Demographics & health Survey(EDHS) 2011. In this study they tried to find the determinant factors that affect postnatal care visit. To balance the imbalance data they applied SMOTE (Synthetic Minority Oversampling Technique). SMOTE is an oversampling approach in which the minority class is oversampled by taking each minority class sample and introducing synthetic examples. Comparison has been performed before and after SMOTE using accuracy, precise, f-measure and recall. A total of 22 interesting rules generated (15 in J48 Decision tree algorithm and 7 in JRIP rule induction algorithm) with an accuracy 93.97% and 93.93% respectively. This research revealed that delivery place, prenatal health professional and age as well as residence are important variable to predict postnatal health care services.

In [6] Dussun Delen et.al. Revealed that why many United State do not have health coverage. In this study 193,373 records and 23 variables are considered. Two main classification techniques were used i.e. Artificial Neural Network (ANN) and Decision Tree. In their research experimental results showed Artificial Neural Network produce most accurate result as compare the decision tree. Employment status, education and marital status are the most important predictive factor. Original data set contained 303,822 records. Only those records were considered in which patient age lies between 18-64 and rest of the records were removed. Some record which had some missing values or inaccurate data were also removed. The final data set remained in hands were 178,274 records. Out of all records 84% respond yes and 16% were respond no. To balance the imbalanced data, size of the data set has been reduced. Final data set which was ready for compilation was 54,524 records. After the compilation of result, ANN produced 78.45% and decision tree produced 74.11% respectably the overall accuracy.

In [7] Sergey Shishlen et.al. Built a model that predicts the accessibility of healthcare on the large variable data set. In which four decision tree algorithms are used. In this study data set has been fetched from Behaviour Risk Factor Surveillance System. The number of participants is 504,408 with 400 variables. Some variables which showed their most contribution is insurance coverage, physical and mental well being, employment status, gender etc.

In [8] reviewed various application of data mining for healthcare. The major applications that covered in this paper are infection control surveillance, treatment of various diseases customer Relationship Management, healthcare administration, and hospital management fraud and anomaly detection. In this study some data mining function data mining algorithm were also considered in the control of healthcare management and divided the healthcare network into three sections i.e. primary, secondary and tertiary.

In [9] reviewed various researcher articles on Brest Cancer diagnosis and prognosis problem and applied data mining techniques to uncover hidden patterns that can helps clinician in decision making. This study also revealed that ANN produced the highest accuracy in comparison to other classification techniques.

In [10] introduced a model that helps mothers to take care of their children. The model explored the applicability of ICT to make schedule of child’s vaccination to know vaccination dates and details about overall growth of child. The model also suggested next vaccination alert, next due vaccine with name vaccine and at which age. Model has been analysed by chi-square analysis.

In [11] suggested a hybrid approach that minimize the effect on an imbalanced healthcare data set. The feature selection technique took data from combined score. In this study 63 attribute of brain tumor with oligodendroglima are obtained. The experimental result showed that optimal new approach produce better results and imbalanced characteristics of medical data. The ensemble classification algorithm ANNIGMA with bagging and decision tree outperform all other existing algorithm.

In [12] developed a prototype model named Intelligent Heart Disease Prediction System (IHDPS) which predict the likelihood of patients getting a heart diseases. The IHDPS used three data mining techniques i.e., Decision tree, Naïve Bayes and Neural Network.

IV. DATA MINING PROCESS

a. Problem Definition: In the absence of adequate work done on the contribution of computer to cope up with the problem of health care issues in general and specifically upon heart disease. Following study is focusing to find solution in this regard.

b. Data Source: in this step relevant data set which is suitable for the problem are fetched from database. Data set is collected from online UCI Repositories. In heart disease dataset total 270 instances with 13 medical attributes(independent) and one target variable were obtained.[20]. Following table shows all these attributes.

Sr. No	Independent variables	Sr. No	Independent variables
1	Age	8	maximum heart rate achieved
2	Sex	9	exercise induced angina
3	Chest Pain type	10	oldpeak = ST depression induced by exercise relative to rest
4	resting blood pressure	11	the slope of the peak exercise ST segment
5	serum cholestoral in mg/dl	12	number of major vessels (0-3) colored by fluoroscopy
6	fasting blood sugar > 120 mg/dl	13	thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
7	resting electrocardiographic results (values 0,1,2)	14	Predictable attribute Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing(has heart disease))

c. Understanding the data: Due to the growth of information technology healthcare data has been increased exponentially. So there are many sources primary as well as secondary of healthcare database. So it becomes very difficult to collect relevant data only. To carry out data analysis relevant and correct data shout be considered.

d. Preparation: Once data has been understood next step is to prepare the data for analysis. It deals with missing values, imbalanced data and outlier detection. Preparation phase deals with solution of these types of issues.

e. Data mining techniques: different data mining techniques are available for analysis of data. Like association rule, classification and clustering etc... According to the requirement one or more techniques can be used. Association refers to linking of one incident to another. Association method can make the relationship between variable in heart disease problem. Classification method classifies whole heart disease data according to these known classes. Cluster analyses make the group of similar patient.

f. Model Building: in model building a data mining algorithm is applied on the heart disease database. In this study only decision tree is applied on the given dataset.

V. EXPERIMENT AND RESULT

This section applies the nine different classification algorithms on the heart disease dataset. All these algorithms use under the WEKA machine learning tool. Weka is a set of machine learning algorithms that can be applied to a data set directly, or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

The accuracy rate, execution time, mean absolute error, root mean squared error etc… are different of each classification algorithm.

	Correctly Classified	Incorrectly classified	Time	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
NaïveBayes	83.7	16.3	0.02	0.6683	0.1835	0.3598	37.163	72.4
BayesNet	81.11	18.19	0.01	0.6172	0.1982	0.3666	40.1338	73.7855
Bagging	80	20	0.05	0.595	0.2915	0.376	59.0278	75.6655
J48	76.67	23.33	0.01	0.5271	0.274	0.4661	55.4788	92.5962
SMO	84.074	15.926	0.04	0.6762	0.1593	0.3991	32.2467	80.3119
Logistic	83.7	16.3	0.02	0.6683	0.2247	0.3322	45.4919	70.8153
Multilayer Perceptorn	77.4	22.6	0.94	0.5444	0.2328	0.4438	47.1437	89.3044
LMT	83.33	16.67	0.25	0.6617	0.2302	0.3558	46.6077	71.5948
Random Tree	77.037	22.963	0.005	0.535	0.2296	0.4792	46.4953	96.4325

Table 1 shows execution result of 9 classification algorithm

The main comparison of between different algorithms is made by on the basis of computation time and accuracy.

a. Accuracy: Table 2 shows the accuracy results of various data mining algorithms that applied on heart disease dataset. The result shows that SMO classification algorithm gave 84.074% accuracy, which is the highest among the entire participating algorithm.

SMO	NaïveBayes	Logistic	LMT	BayesNet	Bagging	Multilayer Perceptorn	Random Tree	J48
84.074	83.7	83.7	83.33	81.11	80	77.4	77.037	76.67

Table 2 shows the accuracy of various classification algorithms

b. Computation Time: Table 3 shows the time taken by all the algorithms on heart disease dataset. There are 270 instances with thirteen independent variables and one dependent variable. It is observed that random tree took the minimum time .005 second to execute the given data set.

Random Tree	BayesNet	J48	NaïveBayes	Logistic	SMO	Bagging	LMT	Multilayer Perceptorn
0.005	0.01	0.01	0.02	0.02	0.04	0.05	0.25	0.94

Table 3 shows the execution time taken by classification algorithms

c. Feature reduction

Feature reduction is a technique of identifying and removing those attributes that do not contribute towards classification of the dataset. In this work, gain ratio technique is used to evaluate the worth of an attribute with respect to the class. Then rankers algorithm is used to arrange these attributes in descending order according to their gain ratio and last three attributes having lowest gain ratio can be removed.

Rank	GainRatio	Attribute	Rank	GainRatio	Attribute
1	0.203	Thalassemia	8	0.0669	Sex
2	0.1904	Cpain	9	0.0567	Age
3	0.1659	Mvasal	10	0.0241	Rer
4	0.1299	Exercise	11	0	Srm
5	0.1203	Heartrate	12	0	RBP
6	0.1196	Oldpeak	13	0	Sugar
7	0.1099	Slop			

Table 4: Representing Gain Ratio of all attributes

The attributes srm, RBP, sugar has minimum gain ratio. These three attributes will be removed because they do not contribute toward data classification. This will result in less computation time and less memory requirement.

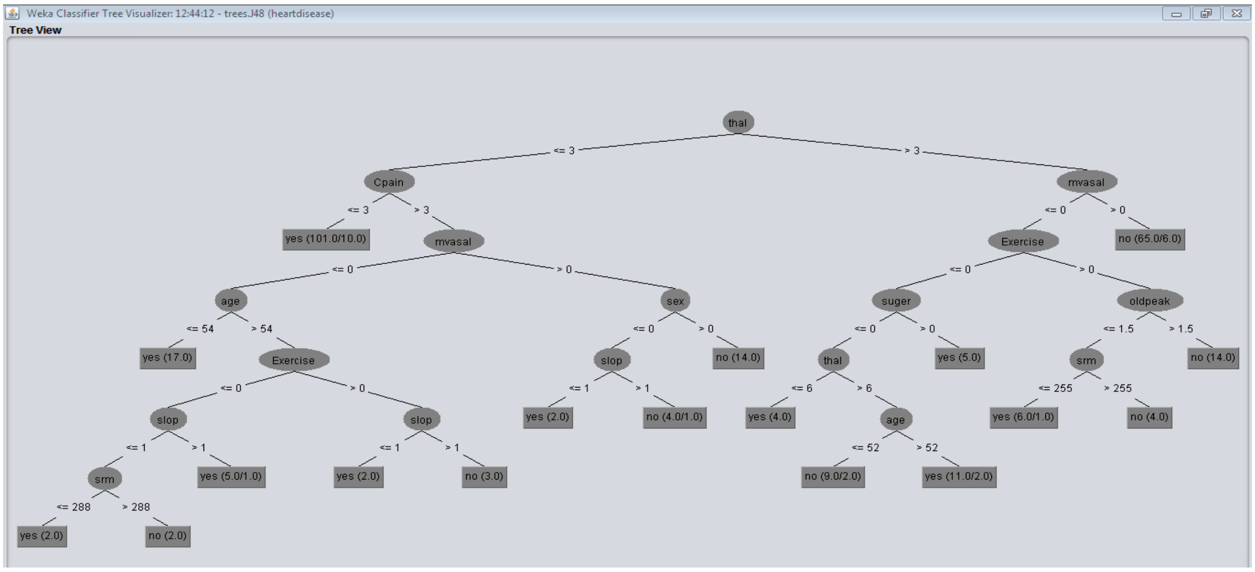

Fig 1:Tree Constructed on the basis of Gain Ratio

VI. CONCLUSION

The experimental results have shown that different classification algorithms behave differently on the same dataset. Some algorithms are good in correctly classification, some are good in execution time and some algorithms are good in mean squared error etc… Some attributes do not contribute to the target variable and if remove these attribute from the data set, overall performance of the algorithm can be improved. In our experiment there are three variable named srm, RBP, sugar do not contribute the class attribute and if we remove these attribute, correctly classified instance will remain same but execution speed of the algorithm sure increase.

REFERENCES

[1] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, “Emerging Information technologies for enhanced healthcare,” *Comput. Ind.*, vol. 69, pp. 3–11, 2015.

[2] N. Wickramasinghe, S. K. Sharma, and J. N. D. Gupta, “Knowledge Management in Healthcare,” vol. 63, pp. 5–18, 2005.

[3] Data Mining: Concepts and Techniques Second Edition Jiawei Han *University of Illinois at Urbana-Champaign* Micheline Kamber

[4] www.confidenceconnected.com/blog/2012/10/09/data_mining_in_a_healthcare_setting/

[5]Geletaw Sahle “Ethiopic maternal care data mining:discovering the factors that affect postnatal care visit in Ethiopia” Sahle Health Inf Sci Syst (2016) 4:4 DOI 10.1186/s13755-016-0017-2

[6]Dursun Delen*, Christie Fuller, Charles McCann, Deepa Ray “Analysis of healthcare coverage: A data mining approach” Available online at www.sciencedirect.com Expert Systems with Applications 36 (2009) 995–1003

[7]Sergey Shishlenin, Gongzhu Hu, “Predicting Access to Healthcare Using Data Mining Techniques”, Software Engineering Research, Management and Applications Volume 578 of the series Studies in Computational Intelligence pp 191-204 Date: 02 November 2014

[8] Anand Sharma, Vibhakar Mansotra, “Emerging Applications of Data Mining for Healthcare Management - A Critical Review ”, *2014 International Conference on Computing for Sustainable Global Development (INDIACom),* **DOI:** 10.1109/IndiaCom.2014.6828163

[9] Shelly Gupta, Dharminder Kumar,Anand Sharma, “DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS “ Vol. 2 No. 2 Apr-May 2011

[10] Siddhi Shah, Shefali Naik and Vinay Vachharajani “Child Growth Mentor—A Proposed Model for Effective Use of Mobile Application for Better Growth of Child “,Advances in Intelligent Systems and Computing Volume 409 pages 153-160.

[11] SHAMSUL HUDA , JOHN YEARWOOD , HERBERT F. JELINEK,, (Member, IEEE), MOHAMMAD MEHEDI HASSAN , (Member, IEEE), GIANCARLO FORTINO , AND MICHAEL UCKLAND “A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis”, VOLUME 4, 2016,pages 9145-9154

[12]Sellappan Palaniappan, Rafiah Awang “Intelligent Heart Disease Prediction System Using Data Mining Techniques ” AICCA ’08 Proceeding of the 2008 IEEE/ACS International conference on computer Systems and Application pages 108-115.

[13] *Gavin Brown.* “UCI Machine Learning Databases”, http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29, 2004