Retrieving Information from Special EHR of Multiple Systems of Medicine by Applying Naïve Bayes, BayesNet, PART, JRip and OneR Classification Algorithms

Dr. Vaishali S. Parsania
Asst. Prof., Department of MCA, Atmiya Institute of Technology & Science, Rajkot, Gujarat, India

Abstract: Data mining is a procedure to find out hidden valuable and beneficial knowledge by analyzing huge amounts of data, which is pile up in databases. Nowadays there are gigantic amount of data in the field of healthcare. Without extracting the knowledge from it those data are merely a collection of facts if not used for any futuristic betterment of health of populace. Here the Electronic Health Record (EHR) database is of multiple systems of medicine which include Ayurvedic system of medicine and Allopathic system of medicine. In this paper various classification algorithms of data mining are applied on this special EHR. Results are taken for different size of databases and also with different classification algorithms (BayesNet, Naïve Bayes, ZeroR, JRip, OneR, PART) to encompass a comparative approach. The comparative graphical representation of various classification algorithms applied on different size of EHR gives an insight to critic the result and derive the conclusion.

Index Terms - **Data mining, multiple system of medicine, EHR, Classification algorithms, BayesNet, Naïve Bayes, ZeroR, JRip, OneR, PART**

I. INTRODUCTION

Data mining is a technology to enable data exploration, data analysis and data visualization of very large databases at a high level of abstraction. "DM is a process of non-trivial extraction of novel, implicit, and actionable knowledge from large datasets." [1]

There are many data mining techniques like association, classification, clustering, prediction etc. There are various data mining techniques available out of that here some techniques are selected which are suitable for the proposed application model. Here classification techniques are selected for the application on EHR databases of various sizes. Electronic health record is probably one of the most significant contributions of Information Communication Technology (ICT) in present healthcare [2]. The EHR taken over here is containing multiple systems of medicine that is Ayurvedic system of medicine and allopathic system of medicine. The EHR is the amalgamation of both the system of medicine with different attributes. Weka is selected for applying classification algorithms on EHR database. Clinical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within these data could provide new medical knowledge. [3]

Classification is a task of predicting the value of a categorical variable (target or class) by building a model based on one or more numerical and/or categorical variables (predictors or attributes).

Classification is a data mining function that assigns items in a group to target classes. The purpose of classification is to accurately envisage the target class for each case in the data. Here the proposed system model will also classify the records based on the available dataset of EHR.

II. CLASSIFICATION TECHNIQUES IN DM

Classification is a data mining technique that assigns items in a group to target class [4]. The purpose of classification is to precisely predict the target class from the substance dataset.
Here from the classification techniques some algorithms like BayesNet, Naïve Bayes, ZeroR, JRip, OneR, PART are taken as benchmark for the study of the proposed model.

2.1 BayesNet

"Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. Given a finite set $X = \{X_1,\ldots,X_n\}$ of discrete random variables where each variable X_i may take values from a finite set, denoted by $Val(X_i)$." [5]

2.2 Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. "A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods." [6]

2.3 ZeroR

"ZeroR is a learner used to test the results of the other learners. ZeroR chooses the most common category all the time. ZeroR learners are used to compare the results of the other learners to determine if they are useful or not, especially in the presence of one large dominating category. In the ZeroR method, the result is the class that is in majority when the attributes are categorical and, when they are numerical. Thus the ZeroR is always considered as the base case for data mining." [7]

2.4 JRip

"This implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is proposed by William W. JRip is an inference and rules-based learner (RIPPER) that tries to come up with propositional rules which can be used to classify elements." [8]

2.5 OneR

"OneR, short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, we construct a frequency table for each predictor against the target. It has been shown that OneR produces rules only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret." [9]

2.6 PART

This is a class for generating a PART decision list. "It uses separate-and-conquer approach and builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule." [10]

III. DISTRIBUTION OF EHR DATASET

The distribution of database of EHR is shown to have an insight about the attribute dispersion in the EHR used. Here total numbers of records are 1000.

3.1 Distribution According to Disease:

The data shown in the figure show how the data are assemblage in each attribute per disease. Different color reflects the types of data accommodated by each of the attribute.

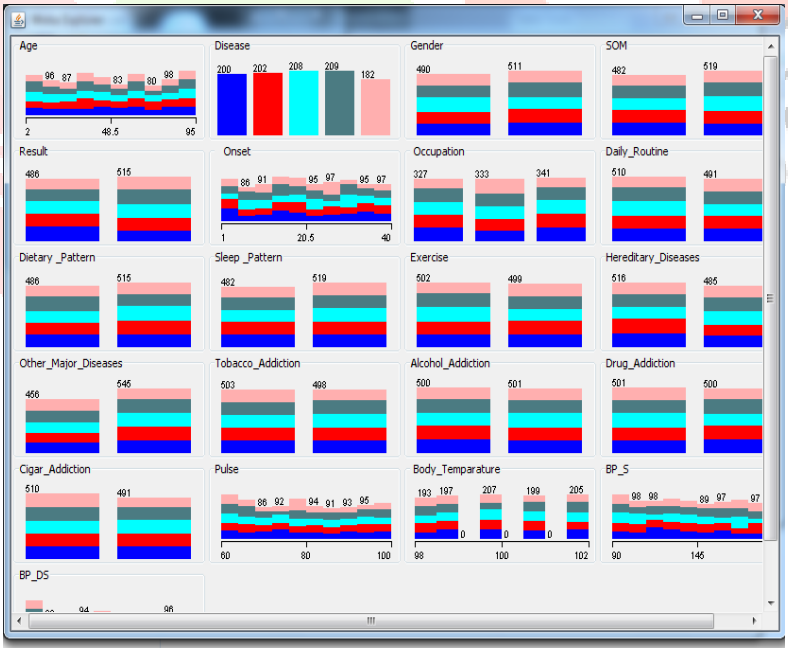

Fig. 1: Distribution of data according to Disease

3.2 Distribution According to SOM

The data shown in the figure show how the data are grouped in each attribute per SOM (System of Medicine). Different color reflects the types of data accommodated by each of the attribute.

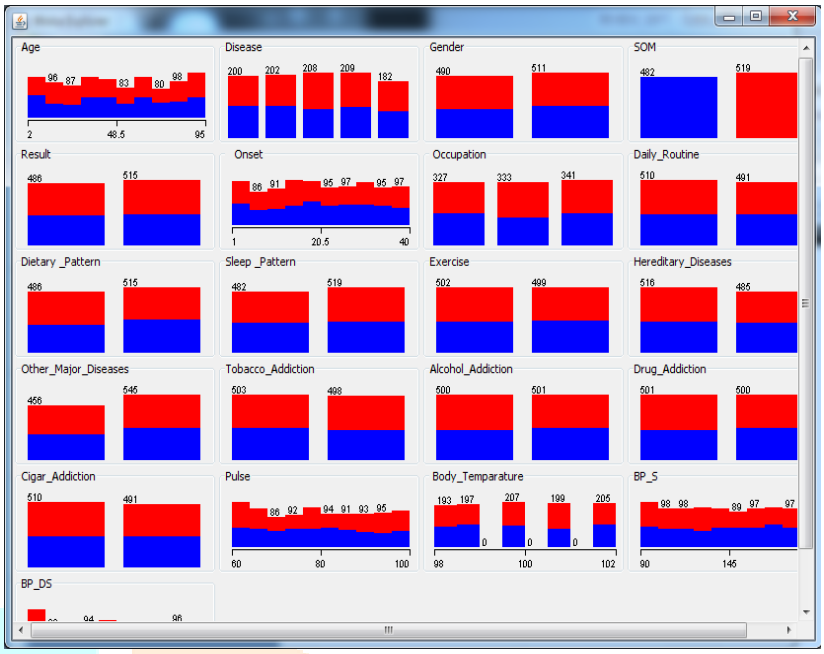

Fig. 2: Distribution of data according to SOM

IV. IMPLEMENTATION OF CLASSIFICATION ALGORITHMS

The above described classifiers are implemented with the Weka tool. This work is targeted to obtain the results in terms of correctly classified instances from the supplemented dataset. Here the EHR dataset is taken of various sizes as to check the consistency in result with respect to number of records in dataset. In the following figure the EHR of 5000 instances are shown for the result analysis.

4.1 Implementation of BayesNet

BayesNet classification algorithm is applied on EHR and following result is being derived in Weka environment.

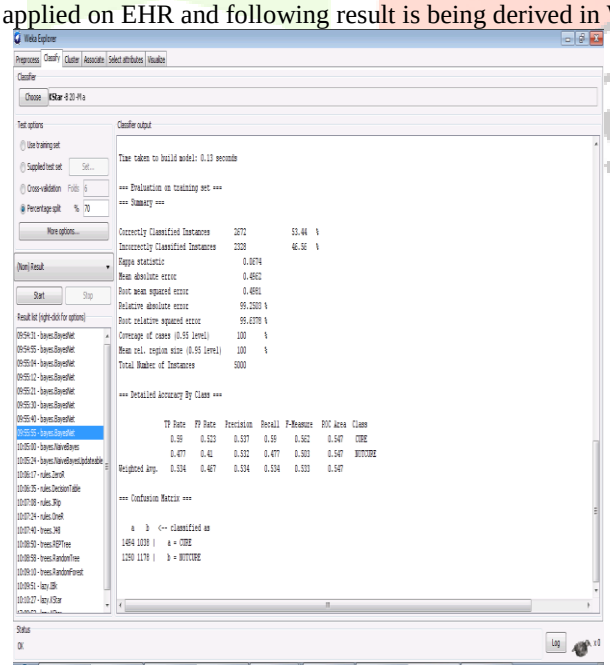

Fig. 3: Results of BayesNet implementation

4.2 Implementation of Naïve Bayes

Naïve Bayes classification algorithm is applied on EHR and following result is being derived in Weka environment.

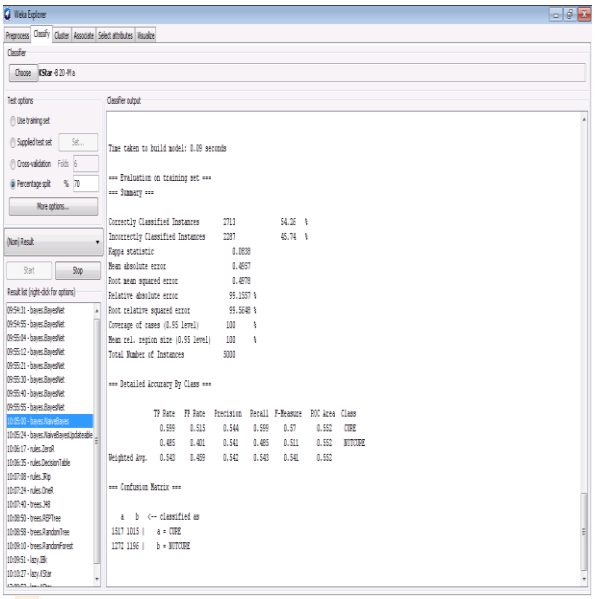

Fig. 4: Results of Naïve Bayes implementation

4.3 Implementation of ZeroR

ZeroR classification algorithm is applied on EHR and following result is being derived in Weka environment.

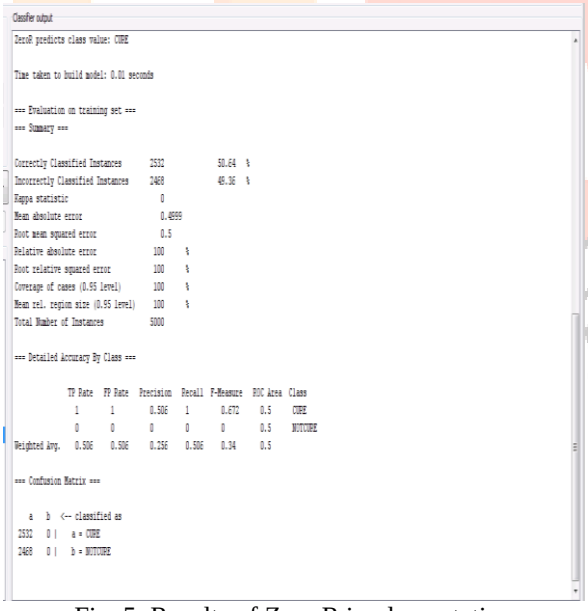

Fig. 5: Results of Zero R implementation

4.4 Implementation of JRip

JRip classification algorithm is applied on EHR and following result is being derived in Weka environment.

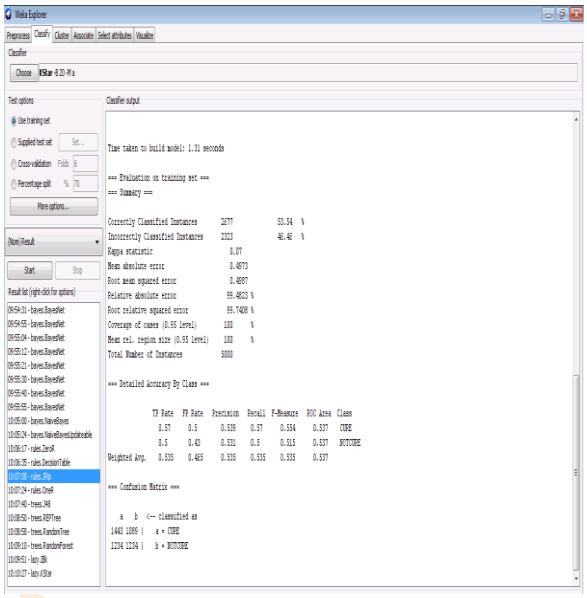

Fig. 6: Results of JRip implementation

4.5 Implementation of OneR

OneR classification algorithm is applied on EHR and following result is being derived in Weka environment.

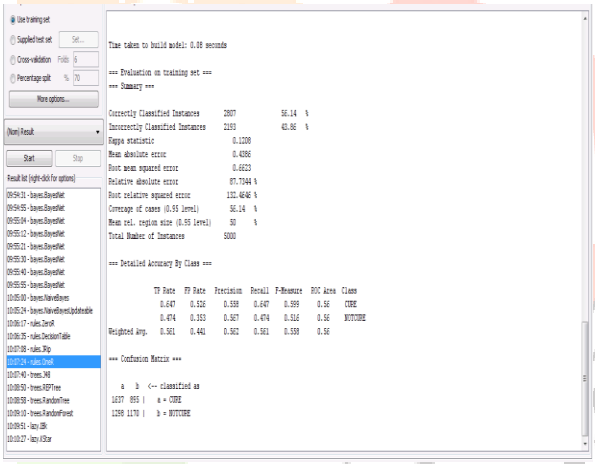

Fig. 7: Results of One R implementation

4.6 Implementation of PART

PART classification algorithm is applied on EHR and following result is being derived in Weka environment.

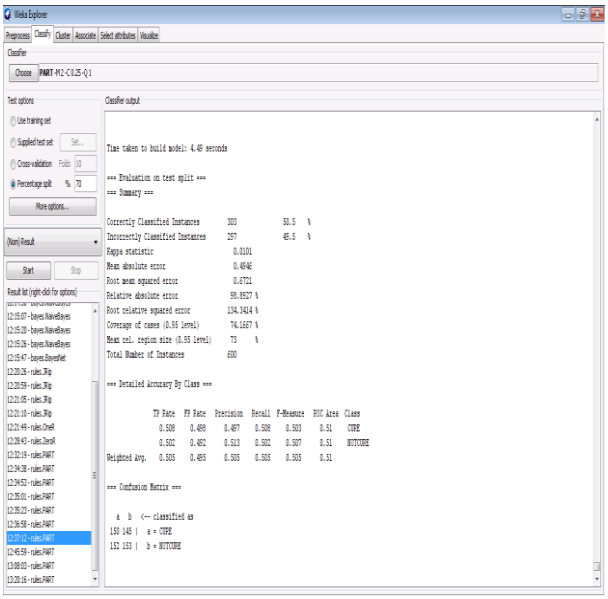

Fig. 8: Results of PART implementation

V. RESULTS

The implementation of above selected algorithm is done in Weka environment. Four EHR datasets of 500, 1000, 2000 and 5000 size are taken for implementation. The above algorithm results are for EHR with the dataset of 5000.

The EHR dataset is subjected to BayesNet, Naïve Bayes, ZeroR, JRip, OneR and PART algorithms. The obtain results in terms of correctly classified instances are tabulated and analyzed.

Table1: Results of classifiers in Weka

	Classifier					
	Bayes Net	Navie bayes	ZeroR	JRip	OneR	PART
Size of Dataset						
500	49.67	49.03	49.09	49.09	49.69	49.69
1000	49.35	49.55	51.44	49.85	50.64	49.37
2000	53.15	55.25	50.45	49.35	53	50.5
5000	53.44	54.26	50.64	53.54	51.18	50.5

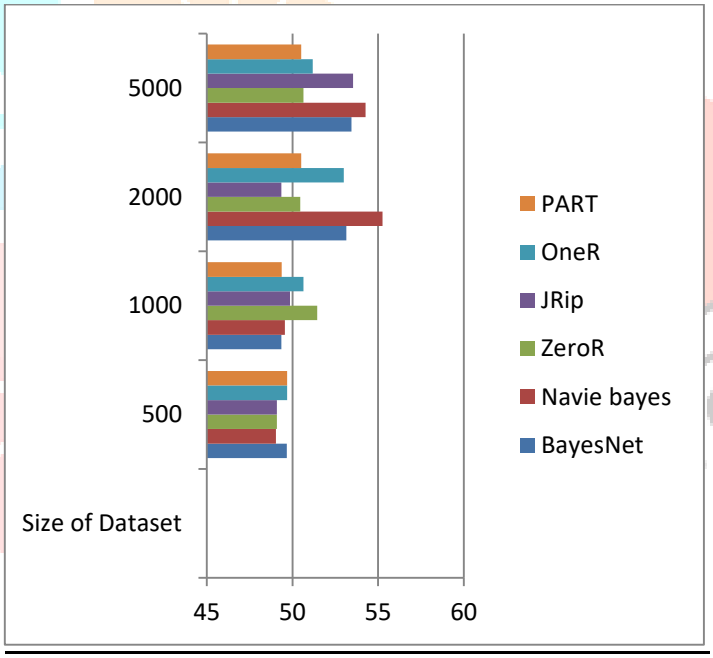

Fig. 9: Graphical representation of the Results obtain in Weka with different classification algorithms

From the above table and chart the result can be seen that the selected classifiers are not generating consistent results. As the data size grows the results are also good in terms of correctly classified instances. Comparatively Naïve Bayes is generating good results and if the data size grows result can be still better. Also BayesNet and JRip is generating good results if the data size is high.

VI. CONCLUSION

The result of classification algorithms applied on different size of EHR databases can be evaluated from the above table and chart. These results clearly show that classification algorithms are not generating consistent results. As the data size grows the correctly classified instances are improved in terms of percentage.

The size of the database matters when classification algorithms are used. As the model which is built for the data to be targeted to some classes can be more concrete if the number of data are more the result is also better in terms of testing data to be evaluated based on that built model.

From all these algorithms Naïve Bayes is producing comparative good results. In Naïve Bayes also as the data size grows the correctly classified instances gives more result in terms of percentage.

To achieve the better results and consistency, improved algorithms can be designed to achieve improved level of results and consistency.

REFERENCES:

[1] M. Khajehei and F. Etemady, "Data Mining and Medical Research Studies," 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, pp. 119–122, Sep. 2010.
[2] Vaishali V Kaneria et. al., "Designing of ICT framework for e-knowledge based HC services", 'International Journal of Computer Science and Management Research' (ISSN: 2278-733X), Volume 1 Issue V, December 2012
[3] J. W. Hales, D. Ph, M. L. Hage, and W. E. Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse," Proc AMIA Annu Fall Symp., no. PMCID: PMC2233405, pp. 101–105, 1997.
[4] Dr. Vaishali S Parsania et. al. , "Applying Naïve bayes, BayesNet, PART, JRip and OneR Classification Techniques on Hypothyroid Database for Comparative Analysis", (IJDI-ERET) (ISSN 2320-7590), Vol. 3, No. 1, June- 2014
[5] "Introduction Bayes Net" , URL: http://bayesnets.com/
[6] "Naive Bayesian", URL: http://chem-ng.utoronto.ca/~datamining/dmc/naive_bayesian.htm
[7] http://mydatamining.com /2008/04/14/rule-learner-or-rule-induction/
[8] "One R Algorithm", URL: http://chem-eng.utoronto.ca/~datamining/dmc/oner.htm
[9] K. Sartipi, M. Najafi, and R. S. Kazemzadeh, "Data and Mined-Knowledge Interoperability in eHealth Systems," no. December, 2008.
[10] Anita Shet, Chief Executive Officer, pinkWhaleHealthcare Link: http://ehealth.eletsonline.com/2010/07/11407/