Comparative Analysis of Data Mining Algorithms on EHR of Rheumatoid Arthritis of Multiple Systems of Medicine

Dr. Vaishali S. Parsania[1], Prof. Krunal Kamani[2], Prof. Gautam J Kamani[3]

[1]Asst. Prof., Department of MCA, Atmiya Institute of Technology & Science, Rajkot- Gujarat-India.
vvkaneria@aits.edu.in
[2]Assistant Professor (Computer Science), Sheth M. C. College of Dairy Science, Anand Agricultural University, Gujarat, India
[3]Assistant Professor, College of Agricultural Information Technology, Anand Agricultural University, Anand, Gujarat-India

Abstract: Data mining techniques are applied usually to uncover concealed knowledge from massive data stacked up in databases. One of the potential fields of Data mining application is healthcare systems in which the increasingly large amount of data are populated in the databases. Such populated databases needs the application of suitable data mining techniques to extract the knowledge patterns which are vital decision making as well as care taking systems. In the field of healthcare enormous amount of data is generated and populated in databases. These databases are vital for knowledge extraction and its uses for futuristic betterment of health of populace. The Electronic Health Record (EHR) database for a disease of Rheumatoid Arthritis is considered in the research work. It includes the data from multiple systems of medicine which include Ayurvedic system of medicine and Allopathic system of medicine. The classification algorithms - BayesNet, Naïve Bayes, ZeroR, JRip, OneR and PART are implemented on EHR of Rheumatoid Arthritis. Results are obtained for 100, 500 and 1000 instances of EHR to encompass a comparative approach for analytics.

Keywords: Data mining, Electronic Health Record (EHR), Rheumatoid Arthritis (RA), multiple systems of medicine, Classification algorithms, BayesNet, Naïve Bayes, ZeroR, JRip, OneR and PART

I. INTRODUCTION

To uncover the hidden information from these large databases data mining techniques come at the aid. The uncovered information usually includes relationship and patterns within these datasets subjected to clustering and classification. The organized uncover information takes the shape of a knowledgebase which is vital information.

There was also the introduction of new methods for knowledge representation in addition to traditional statistical analysis of data. It was recognized that information is at the heart of any field operations and decision-makers could make use of the data stored to gain valuable insight into it. Data Mining or Knowledge Discovery in Databases is the process which helps to fetch the knowledge from the bundle of information [1]. 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 [2].

Data mining has many different techniques like Association, Classification, Clustering, Prediction etc. out of that here classification techniques are applied on dataset as it is more relevant for retrieving result according to literature survey. 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. [3] Here the proposed system model will also classify the records based on the available dataset of EHR. The EHR used for the research study is including the multiple systems of medicine and is unique itself [4]. The Ayurvedic system of medicine and allopathic system of medicine is taken in EHR [5].

From the classification techniques algorithms - BayesNet, Naïve Bayes, ZeroR, JRip, OneR and PART are taken as benchmark for the study of the proposed model.

II. CLASSIFICATION TECHNIQUES

In this paper five classification techniques BayesNet, Naïve Bayes, ZeroR, JRip, OneR and PART are selected to apply on the database and deriving results from it using Weka tool. Each of this classification technique has its specialty as a classifier.

a. Navie 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. [6]

b. Bayes Net

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 $Val(X_i)$. [7]

c. 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. [8]

d. 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. [9]

e. 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. [10]

f. 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. [11]

III. IMPLEMENTATION OF CLASSIFICATION TECHNIQUES ON EHR FOR RA:

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.

IV. EHR FOR RA DATABASE STRUCTURE

Special attributes are selected for designing the EHR of RA with the help of medical experts. It includes general patients’ attributes and disease specific attributes. The attributes are selected by considering the vitality of both the system of medicine. The EHR dataset is taken in 3 different sizes of 100, 500 and 1000 for better evaluation purpose.

V. RESULTS & DISCUSSION

The implementation of above selected algorithm is done in Weka environment with 3 different EHR datasets of 100, 500 and 1000. The above Screenshots are for EHR with the dataset 1000 with all five algorithms.

The EHR dataset is subjected to BayesNet, Naïve Bayes, ZeroR, JRip, OneR and PART algorithms. The obtain results are tabulated and analyzed. In Table I Correctly classified instances are given for all this DM algorithms for the mentioned different dataset size. Table II shows the Average of Correctly Classified Instances with implemented DM Algorithms in Weka Environment.

TABLE I

Correctly Classified Instances (in percentage) with implemented DM Algorithms

	DM Algorithms					
	BayesNet	**Navie Bayes**	**ZeroR**	**JRip**	**OneR**	**PART**
Size of Dataset						
100	**53**	**57**	**57**	**49**	**54**	**49**
500	**53.8**	**53.4**	**53.4**	**52.8**	**55.2**	**48.4**
1000	**51.6**	**52.1**	**50.4**	**51.8**	**50.4**	**52.8**

VI. Graphical representation of the Results with different classification algorithms

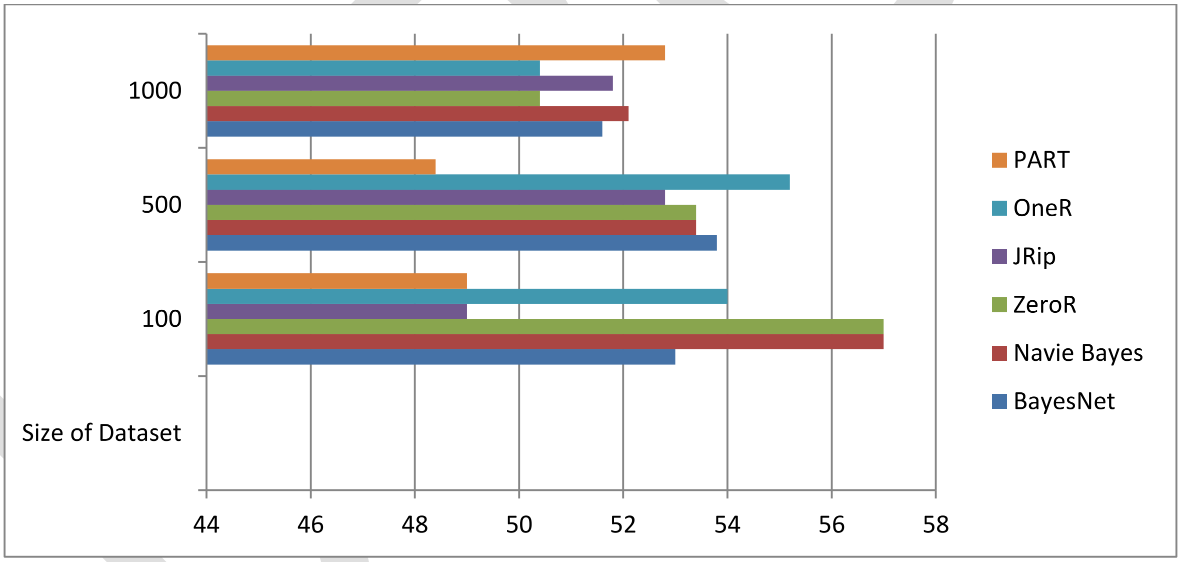

Fig. 6: Graphical representation of the Results with different classification algorithms

TABLE III

Average of Correctly Classified Instances with implemented DM Algorithms

	DM Algorithms					
	BayesNet	**Navie bayes**	**ZeroR**	**JRip**	**OneR**	**PART**
Average of correctly classified instances	**52.8**	**54.167**	**53.6**	**51.2**	**53.2**	**50.067**

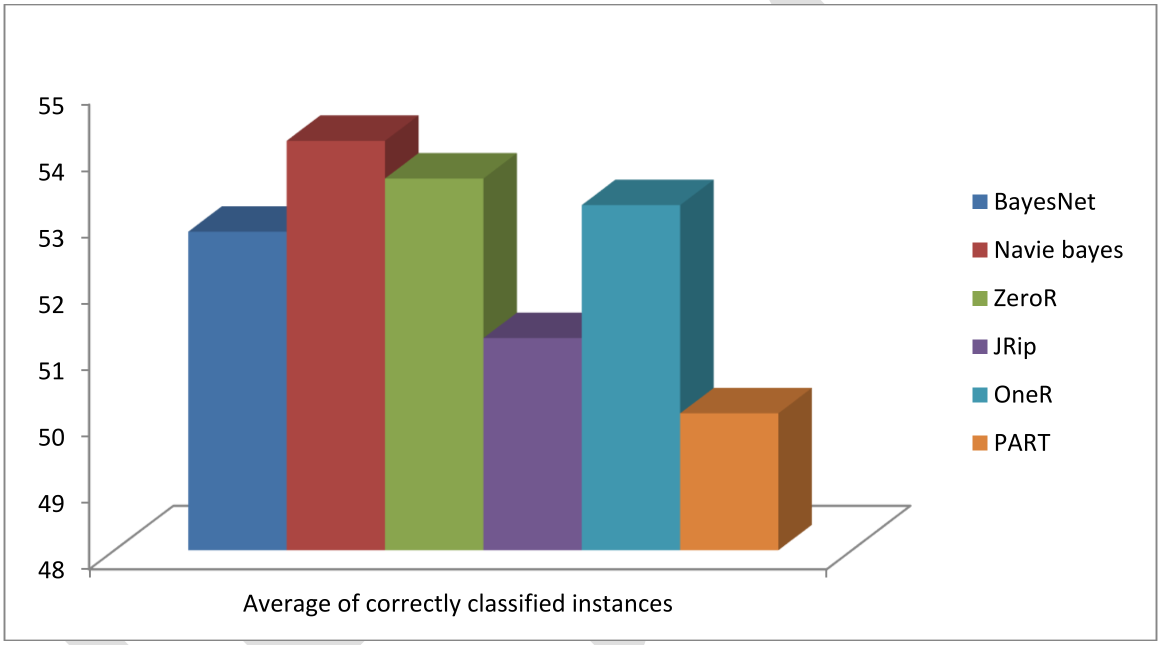

Fig. 7: Average of correctly classified instances with different classification algorithms

Result can be analyzed from the above tested data. The results of BayesNet are improved from 53 to 53.8 if the size of dataset is increased but again when data size becomes 1000 results are decreased to 51.6. In Naïve Bayes results are decreases from 57 to 53.4 when data size becomes 500 and again it decreases to 52.1 when data size grows to 1000. In ZeroR results are decreased from 57 to 53.4 when data size is increased to 500 and again it decreases to 50.4 when data size grows to 1000. In JRip results are increased from 49 to 52.8 when data size becomes 500 from 100. But again when data size increases to 1000 results are decreased to 51.8. In OneR results are increased from 54 to 55.2 when data size increases to 500. But again the result decreases to 50.4 when data size grows to 1000. In PART result decreases from 49 to 48.4 when data size increases up to 500, but when data size grows to 1000 the result is increased from 48.4 to 52.8.

The result analysis reveals that Naïve Bayes has comparative good results as compare to other algorithms in terms of consistency and average of correctly classified instances. But still improved results are desired to benefit more no of patients.

VII. CONCLUSION

The result of classification algorithms applied on different size of EHR of Rheumatoid Arthritis has been evaluated from the above tables and charts. These results show that Naïve Bayes has good results as compare to other algorithms in terms of consistency and average of correctly classified instances. By modifying the Naïve Bayes algorithm some improved algorithms can be designed to achieve better level of consistency and better results based on the EHR of single disease. The modified algorithm SSOM has been designed for the improved results and will disclose in the subsequent research paper. As SSOM has good results more number of patients can be benefited in the selection of system of medicine. By following Optimal Data Analysis (ODA) technique the accuracy can be refined further [12] [13].

ACKNOWLEDGMENT

I would like to thank Dr. Sanjay Parsania, Dr. Dhirendra Joshi, Dr. Paresh Parsania and Dr. Ukani for providing data of patients and their valuable inputs in this research.

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. 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.
3. P. Andreeva, M. Dimitrova, and P. Radeva, “DATA MINING LEARNING MODELS AND ALGORITHMS.”
4. Vaishali V Kaneria, Dr. N N Jani, “Designing of ICT framework for e-knowledge based HC services”, International Journal of Computer Science and Management Research, Vol 1 Issue 5 December 2012
5. Vaishali V Kaneria, Dr. N N Jani, “ICT Framework for e-Knowledge based Healthcare Services: Study & Analysis”, International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 1, February 2012
6. “Naive Bayesian”, Internet: http://chem-ng.utoronto.ca/~datamining/dmc/naive_bayesian.htm, [Jan, 2013]
7. “Introduction Bayes Net" , Internet: http://bayesnets.com/, [Jan, 2013]
8. “ZeroR”, Internet: http://www.saedsayad.com/zeror.htm, [Feb , 2013]
9. “One R Algorithm”, Internet: http://www.saedsayad.com/zeror.htm, [Feb, 2013]
10. K. Sartipi, M. Najafi, and R. S. Kazemzadeh, “Data and Mined-Knowledge Interoperability in eHealth Systems,” no. December, 2008.
11. “PART”, Internet: http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/PART.html, [Feb, 2013]

12. Paul R. Yarnold, Ph.D., and Robert C. Soltysik, M.S., “Optimal Data Analysis: A General Statistical Analysis Paradigm”, Volume 1, Release 1, September, 2013
13. Paul R. Yarnold, Ph.D., and Robert C. Soltysik, M.S., “Maximizing Accuracy of Classification Trees by Optimal Pruning”, Volume 1, Release 1, September, 2013