IJDI-ERET

INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH AND EMERGING TECHNOLOGIES
Vol. 3, No. 1, 2014
www.ijdieret.in

(Research Article)

Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis

Dr. Vaishali S. Parsania*[1], Dr. N. N. Jani[2], Navneet H Bhalodiya[3]

[1]Asst. Prof., Department of MCA, Atmiya Institute of Technology & Science, Rajkot- Gujarat-India.*
[2]Dean CS, KSV University,Director, SKPIMCS, Gandhinagar,Gujarat-India.
[3]Student, Department of MCA, Atmiya Institute of Technology & Science, Rajkot- Gujarat-India.

Abstract

This research paper intends to provide comparative analysis of Data Mining classification algorithms. Some benchmarking classification algorithms like Naïve Bayes, Bayesian Network, JRip, OneR and PART are selected based on literature survey.

These classification algorithms are applied on Hypothyroid health database for the purpose of finding better techniques for classification.

The multiple parameters taken into considerations for analytical purpose are accuracy, sensitivity, Precision, False positive Rate and f-measure. Results of all these parameters are taken for all the described classification techniques. At the last the results are provided in tabular form to facilitate comparative analysis for the hypothyroid database.

Keywords: Data Mining, Knowledge Discovery in Databases (KDD), Naïve Bayes, Bayesian Network, JRip, OneR and PART

1. Introduction

Data mining also known as 'Knowledge Discovery in Databases' process or KDD is the computational process of identifying patterns from the large data sets[1].

World grows in complexity, to handle the data. Data mining becomes essential for analysing the patterns that underlie it. Healthcare is one of the sectors, where Data mining can be very efficiently apply for knowledge generation.

Health sector deals with thousands of records each and every day. To manage these records is very tedious task and specially to derive some useful knowledge from those records is more difficult. Using Data Mining techniques these health data can be utilized to generate knowledge. Data mining involves the use of refined data analysis tools to find out previously unknown, valid patterns and relationships from large datasets.

These tools can include the interdisciplinary research area like statistical models, arithmetical algorithm, machine learning methods, intelligent information systems, database systems, expert systems etc [2].

Classification techniques are selected for the processing of hypothyroid database. From the classification techniques bench marking algorithms Naïve Bayes, Bayesian Network, JRip, OneR and PART are selected based on literature survey.

2. Classification Techniques

Classification is a data mining technique that assigns items in a group to target class. The purpose of classification is to accurately envisage the target class for each case in the data [4].

Five classification techniques are taken as benchmarking algorithms to be studied for the taken health database of Hypothyroid. Briefings of these algorithms are as under.

2.1 NAÏVE BAYES: The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions which assumes all of the features are equally independent. It uses a Bayesian algorithm for the total probability procedure, the principle is according to the probability that the text belongs to a category of prior probability, the text would be assigned to the category of posterior probability. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [5].

[*] *Corresponding Author: e-mail: vvkaneria@aits.edu.in,*

ISSN 2320-7590

2.2 BAYESIAN NETWORK: A Bayesian network is a structure that shows the conditional dependencies between domain variables and may also be used to illustrate graphically the probabilistic underlying relationships among domain variables. A Bayesian network consists of a directed acyclic graph and probability tables. The nodes of the network represent the domain variables and an arc between two nodes indicates the existence of a underlying relationship or dependency among these two nodes[3].

2.3 JRIP: JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in growing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular decision in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered [7].

2.4 OneR: OneR, short for "One Rule", is a simple classification algorithm that generates a one-level decision tree. OneR is able to deduce typically simple, yet precise, classification rules from a set of instances. OneR is also able to handle missing values and numeric attributes showing flexibility in spite of simplicity. The OneR algorithm creates one rule for each attribute in the training data, then selects the rule with the minimum error rate as its 'one rule'[8].

2.5 PART: PART is a separate-and-conquer rule learner. The algorithm producing sets of rules called 'decision lists' which are planned set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the class of the first matching rule. PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule [9].

3. Database Structure: HYPOTHYROID [6]

The above classification techniques are applied on the database hypothyroid having following structure.

Table 1. Hypothyroid Database

Database	**Hypothyroid**
No. of Instances	**3163**
No. of Attributes	**26**

Sr. No.	**Attributes**	**Data Type**	**Values**
1	Age	int	Real
2	Sex	boolean	M,F
3	On_Thyroxine	boolean	F,T
4	Query_On_Thyroxine	boolean	F,T
5	On_Antithyroid_Medication	boolean	F,T
6	Thyroid_Surgery	boolean	F,T
7	Query_Hypothyroid	boolean	F,T
8	Query_Hyperthyroid	boolean	F,T
9	Preganant	boolean	F,T
10	Sick	boolean	F,T
11	Tumor	boolean	F,T
12	Lithium	boolean	F,T
13	Goitre	boolean	F,T
14	TSH_Measured	char	Y,N
15	TSH	int	Real
16	T3_Measured	char	Y,N
17	T3	int	Real
18	TT4_Measured	char	Y,N
19	TT4	int	Real
20	T4U_Measured	char	Y,N
21	T4U	int	Real
22	FTI_Measured	char	Y,N
23	FTI	int	Real
24	TBG_Measured	char	Y,N
25	TBG	int	Real
26	Class	string	Hypothyroid, Negative

4. Parameters For Measuring Performance Of Classification Techniques:

For measuring performance of Naïve bayes, BayesNet, PART, JRip and OneR classification techniques the following parameters are taken. In Classification techniques parameters to be examined are accuracy, sensitivity, precision, specificity and f-measure.

Following graph shows the measure of the individual parameter for the each algorithm.

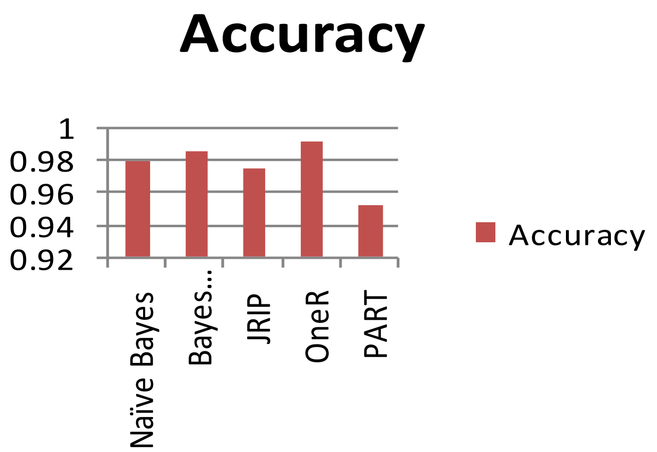


Figure 1. Accuracy measured for Classifiers

This graph shows that the accuracy is comparatively better in OneR techniques and that is 0.99.

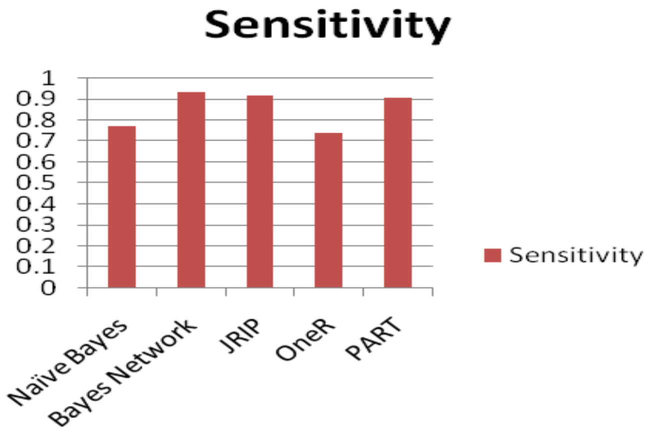


Figure 2. Sensitivity measured for Classifiers

This graph shows that the sensitivity is comparatively better in Bayes Network techniques and that is 0.934.

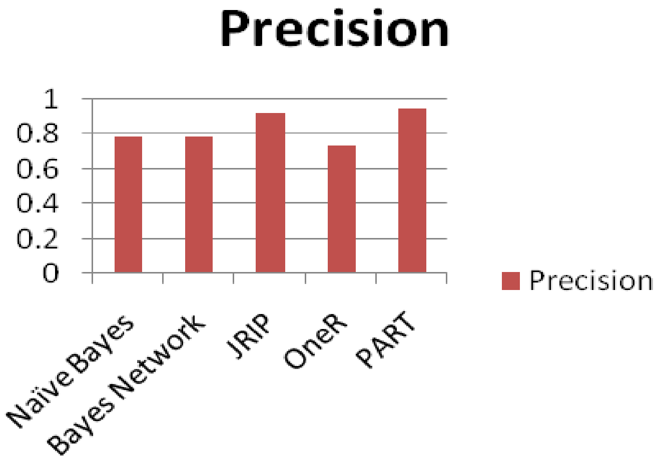


Figure 3. Precision measured for Classifiers

This graph shows that the precision is comparatively better in PART techniques and that is 0.945.

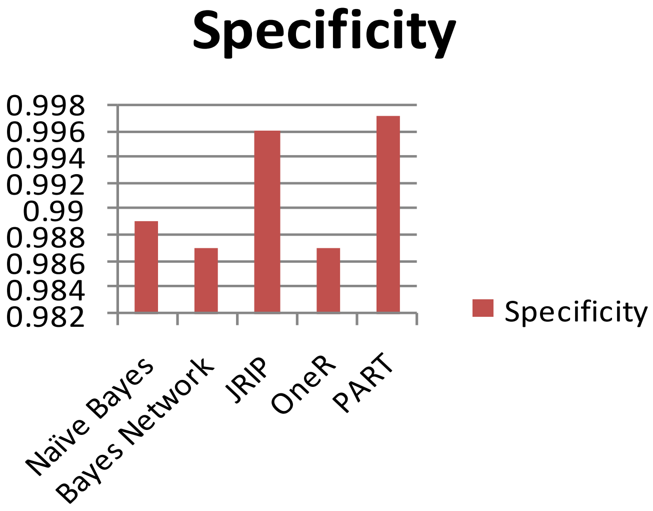


Figure 4: Specificity measured for Classifiers

This graph shows that the specificity is comparatively better in PART techniques and that is 0.997.

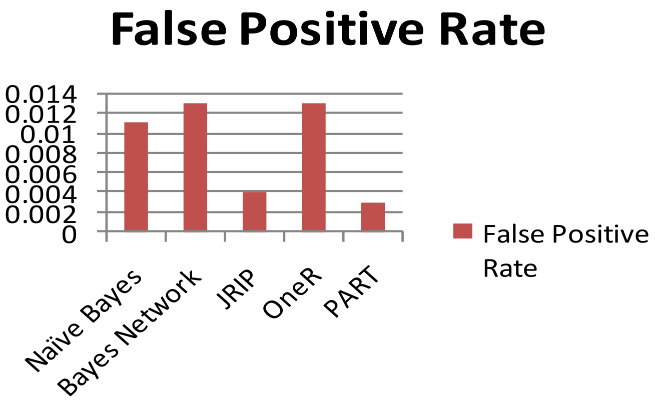


Figure 5. FPR measured for Classifiers

This graph shows that the false positive rate is minimum in PART techniques and that is 0.003.

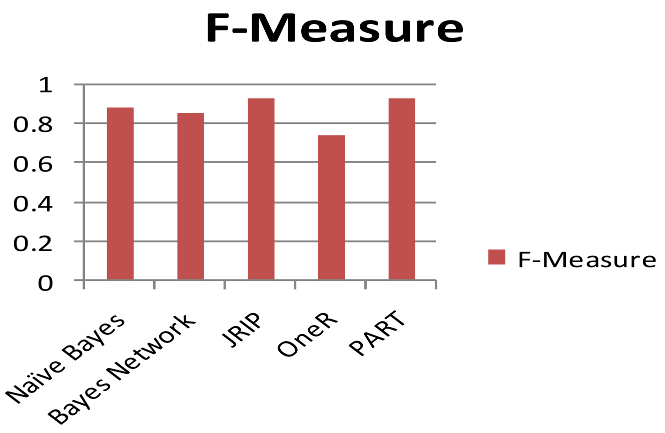


Figure 6: F-measure measured for Classifiers

This graph shows that the f-measure is comparatively better in PART techniques and that is 0.927.

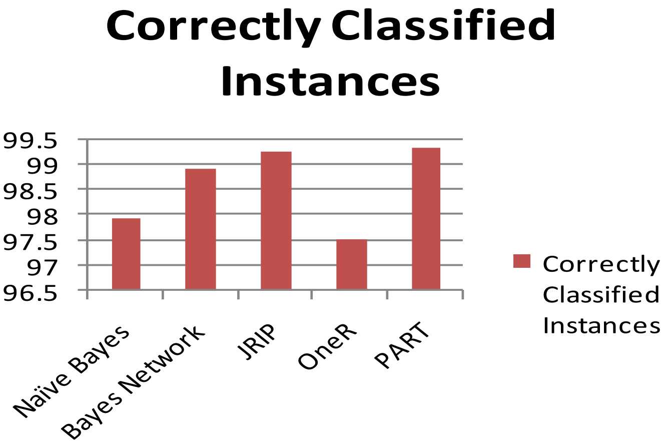


Figure 7. Correctly Classified Instances measured for each Classifier

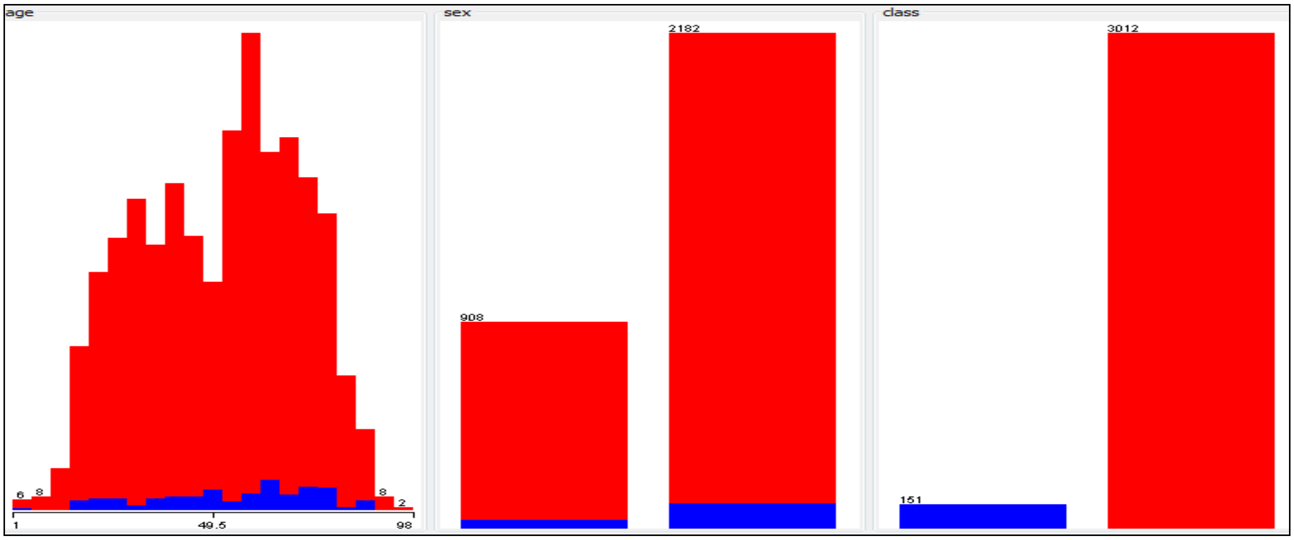
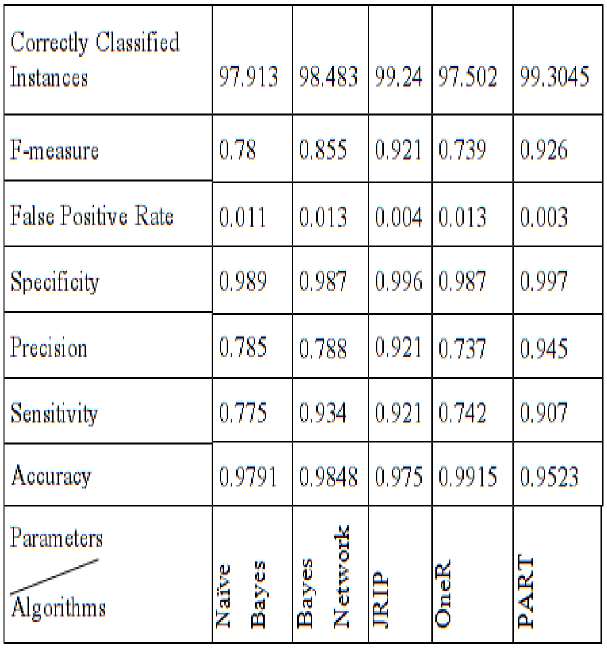


Figure 8: Visualization Graph of distribution of attributes age, sex, class

5. Testing Results:

Results are derived by applying classification algorithms on the database Hypothyroid. Data mining tool Weka 3.6 is used for the purpose of testing and taking results. Results of various data mining classifiers are given in terms of performance measures in table 1. Comparison amongst various classifiers results can be made to find out the better data mining classification algorithm.

Table 1. Results of Data Mining classifiers



Correctly Classified Instances	97.913	98.483	99.24	97.502	99.3045
F-measure	0.78	0.855	0.921	0.739	0.926
False Positive Rate	0.011	0.013	0.004	0.013	0.003
Specificity	0.989	0.987	0.996	0.987	0.997
Precision	0.785	0.788	0.921	0.737	0.945
Sensitivity	0.775	0.934	0.921	0.742	0.907
Accuracy	0.9791	0.9848	0.975	0.9915	0.9523
Parameters / Algorithms	Naïve Bayes	Bayes Network	JRIP	OneR	PART

Visualization Graph of distribution of data in given database attributes: This figure is the visualization graph of attributes age, sex and class. It indicates the main three attributes age, sex and class for the diseases "Hypothyroid". It gives the insight about the distribution of data values for the given attributes in Hypothyroid database.

This graph shows that the correctly classified instances are comparatively better in PART techniques and that is 99.3024.

6. Conclusion

From the above Results taken for the database Hypothyroid the parameters values for classification techniques Naïve bayes, BayesNet, PART, JRip and OneR can be compared. The resultant table reveals that the best result considering majority of the parameters is found in algorithm PART. If Precision, Specificity, False Positive Rate, F-Measure and Correctly Classified Instances are the key parameters to be considered for the classification, then PART is preferable to apply. The result also reveals that, if Accuracy is the important parameter then OneR is more preferable over other algorithms. The Sensitivity is best found in Bayes Network, so preferred when Sensitivity is key parameter in the selection of classification technique.

References

1. "Data Mining", Internet: http://en.wikipedia.org/wiki/Data_mining, [July, 2013]
2. S.Deepajothi,Dr.S.Selvarajan, "A Comparative Study of Classification Techniques On Adult Data Set", IJERT, Vol.1 - Issue 8, October - 2012
3. Aysegul Cayci et. al," Bayesian Networks to Predict Data Mining Algorithm Behavior in Ubiquitous Environments." Universidad Politecnica, Madrid, Spain

4. P. Andreeva, M. Dimitrova, and P. Radeva, "Data Mining Learning Models And Algorithms For Medical Application", Proceedings Of The 18-Th Conference On Saer, Pp 11-18, 2004
5. Lingling Yuan, "An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing." ISCSCT '10, 14-15, AUGUST 2010
6. "Hypothyroid Database", Internet: http://www.hakank.org/weka/, [July, 2013]
7. Anil RAJPUT, Ramesh Prasad Aharwal, Meghna Dubey, S.P. Saxena,(2011) "J48 and JRIP Rules for E-Governance Data." IJCSS-448
8. Gaya Buddhinath and Damien Derry, "A Simple Enhancement to One Rule Classification." Department of Computer Science & Software Engineering University of Melbourne, Australia, 2006
9. Eibe Frank, Ian H. Witten, "Generating Accurate Rule Sets Without Global Optimization". In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

Biographical notes

Dr. Vaishali S. Parsania is currently working as an assistant professor in the department of MCA at Atmiya Institute of technology & science, Rajkot, Gujarat, India. She has experience of 10+ years in academic. She has completed her Ph.D. from KSV, Gandhinagar, Gujarat, India. She has presented 9 papers in seminars/conferences at national level, published 4 papers in international peer-reviewed journals and attended 30+ workshops/symposiums/conferences at various levels. She is member of Indian Society for Technical Education (ISTE), IDES, CSTA, IACSIT, IAENG and IEDS. Her research areas of interest are Data Mining, Knowledge Extraction & Management and open source technologies.

Dr. N. N. Jani is dean- faculty of computer science at KSV, Gandhinagar. He has rich experience of 38 years of teaching in computer science and Applications. He has achieved degrees M. Sc., Ph.D. DCATTP, MCATTP. He is a research guide in various universities and 25 research scholars have completed Ph.D. under his supervision. He has written 15+ books in the field of Computer Science and Applications, one of his books on MEMS (Micro Electro Mechanical System) has been appreciated by Ex-honorable president Dr. A.P.J. Abdul Kalam. He has organized 25+ workshops/ seminars/ conferences/ symposiums at state, national and International levels.

Navneet Bhalodiya has completed his MCA from Atmiya Institute of technology & science, Rajkot. He is working as software engineer at Capital Novus, Gandhinagar.