ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Comparison of Data Mining Techniques for Feature Reduction using Evolutionary Search

[#]R. Prasanka, [*]M. Balamurugan

[#]Research Scholar, School of Computer Science, Engineering and Applications,
Bharathidasan University, Tiruchy
prasanka22@gmail.com
*Associate Professor, School of Computer Science, Engineering and Applications,
Bharathidasan University, Tiruchy
mbala@bdu.ac.in

***Abstract*—Feature Reduction is a data mining techniques which is used to remove unwanted attributes and also known as variable selection, subset selection. It is the process of choosing the subset of relevant features for obtaining best accuracy result. In this paper, Evolutionary Search, the existing algorithms Naïve Bayes, J48 and Classification via Regression is compared. The result obtained before and after the Evolutionary Search is discussed. Comparing the three, Naïve Bayes Algorithm performs well with increased accuracy.**

***Keywords*—Data Mining; Feature Reduction; Weka Tool; Evolutionary Search; Naïve Bayes; J48; CvR.**

I. INTRODUCTION

Data mining which means discovery of extracting large hidden data, previously unknown patterns and relationships that are difficult to detect with traditional statistics. In data mining, the hefty data sets are a computational progression to ascertain patterns. Data mining techniques are the result of a long process of research and product development [1]. It is the examination to excerpt concealed and previously mysterious patterns, to perceive the Traditional statistics relationships and knowledge that are difficult for further use of the overall goal of data mining process extract information and transform it into a comprehensible format from the data set. The role of Data Mining health care data is massive. The major challenge in the medical data mining is imprecision and uncertainty. Nowadays, offering cost worthy service is major problem faced by the health care organizations. It is essential to check more number of diagnose test, to predict a disease. By using intelligent diagnostic tool in the health care, there is possibility to predict the disease. The researchers in the medical field have succeeded in identifying and predicting the disease with the aid of Data mining techniques. This paper compares three existing algorithm. Evolutionary Search (ES) is used to optimize the features. The features obtained by using the ES are then tested with three algorithms. The results obtained are discussed to find the better algorithm for Feature Reduction.

Section II describes the Feature Selection process. Section III gives on overview of the Existing Techniques. Section IV discusses the results and Section V concludes the paper.

II. FEATURE REDUCTION

Feature reduction is removing unwanted attributes which is also known as variable selection attribute selection or variable subset selection and it is the process of selecting subset of relevant features for use in model construction. Feature selection reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data. The feature selection technique is used in various research fields. Feature selection techniques are to be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection in supervised learning has a main goal of finding feature subset that produces higher classification accuracy.

III. RELATED WORKS

Heon Gyu Lee et al [2] proposed a technique to enhance the characteristics of heart rate variability using multi-parametric feature mechanism. M. A. Jabbar et.al [3] developed a heart disease prediction mechanism with the support of associative classification technique in data mining. Sellappan Palaniappan et al. [4] proposed an Intelligent Heart Disease Prediction System (IHDPS) with the help of Neural Network, Decision Trees and Naïve Byes. Liu H et al. [5] discussed the feature reduction technique which is used to find out the accuracy of heart diseases. They also discussed Feature reduction algorithm for classification and clustering.

Niti b Guru et al.proposed a mechanism with the support of neural network for predicting the heart disease, blood pressure and sugar[6]. Charly, K used classification and prediction algorithms for decision trees, neural networks; clustering, association rules and regression [7].From the medical point of view association rules are developed and

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

preliminary results are defendable [8]. Carlos et al, formulate the interrelation technique and decision tree algorithm [9].

Ordonez et al. [10] proposed an association rules based constrained improved mapping algorithm for predicting heart diseases. This algorithm concentrates on finding useful constrains and mapping heart diseases data to applicable format for interrelation technique. The association rule algorithm employs various significant constraints which are used to reduce the rules count and improve the mining process speed.

IV. TECHNIQUES USED

A. *Naive Bayes*

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [11][17].

Naive Bayesian classifier is based on Bayes' theorem and the theorem of total probability. The probability that a document d with vector x = < x1,...,xn> belongs to hypothesis h is[12][13].

$$P(h1|x) = \frac{P(x|h1)\,P(h1)}{P(x|h1)\,P(h1) + P(x|h2)\,P(h2)}$$

Here, P(h1|xi) is posterior probability, while P(h1) is the prior probability associated with hypothesis h1. For m different hypotheses, we have

$$P(x) = \sum_{j=1}^{n} P(x|h)\,P(h)$$

Thus, we have

$$P(h1|x) = \frac{P(x|h1)\,P(h1)}{P(x)}$$

B. *J48 DECISION TREE:*

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple[14][15].

Algorithm [1] J48:

INPUT:

D //Training data

OUTPUT

T //Decision tree

DTBUILD (*D)

{

T=φ;

T= Create root node and label with splitting attribute;

T= Add arc to root node for each split predicate and label;

For each arc do D= Database created by applying splitting predicate to D;

If stopping point reached for this path, then T'= create leaf node and label with appropriate class;

Else T'= DTBUILD (D);

T= add T' to arc;

}

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them [16][18][19][20].

V. RESULTS AND DISCUSSIONS

The three existing algorithms Naïve Bayes (NB), J48 Algorithm and Classification Via Regression Algorithm is compared. The Heart Disease data set with 13 attributes is taken for testing the algorithms.

TABLE I. COMPARISON OF THE EXISTING ALGORITHMS BEFORE APPLYING EVOLUTIONARY SEARCH

	Naïve Bayes	Classification Via Regression	J48
Accuracy	83.33	68.71	78.57
Kappa Statistic	0.6	0.3	0.5
Mean Absolute Error	0.17	0.33	0.3
Root Mean Squared Error	0.37	0.49	0.4
Relative Absolute Error	38.61	71.99	66.97
Root Relative Squared Error	77.89	102.28	85.42
True Positive Rate	0.83	0.68	0.78
False Positive Rate	0.21	0.38	0.27
Precision	0.832	0.68	0.78
Recall	0.832	0.68	0.78
ROC Area	0.9	0.72	0.75

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Evolutionary Search (ES) is used to reduce the unwanted Features from the Data set. The ES will work as follows:

1) Generatetheinitial population of individuals randomly – (first generation)
2) Evaluate the fitness of each individual in that population.
3) Repeat on this generation until termination (time limit, sufficient fitness achieved, etc.):
    1. Select the best-fit individuals for reproduction – (parents)
    2. Breed new individuals through crossover and mutation operations to give birth to offspring.
    3. Evaluate the individual fitness of new individuals.
    4. Replace least-fit population with new individuals.

After Applying the Evolutionary Search the attributes number is reduced to 6. The existing algorithm is then classified for Accuracy. The results obtained are tabulated in Table 2.

TABLE II. COMPARISON OF THE EXISTING ALGORITHMS AFTER APPLYING EVOLUTIONARY SEARCH

	Naïve Bayes	Classification Via Regression	J48
Accuracy	84.69	83.33	77.21
Kappa Statistic	0.7	0.62	0.4
Mean Absolute Error	0.18	0.21	0.3
Root Mean Squared Error	0.34	0.35	0.4
Relative Absolute Error	39.04	53.87	67.42
Root Relative Squared Error	71.49	72.93	87.65
True Positive Rate	0.84	0.83	0.77
False Positive Rate	0.19	0.23	0.29
Precision	0.84	0.83	0.76
Recall	0.84	0.83	0.76
ROC Area	0.84	0.83	0.76

The Figure 1 (a), (b), (c) and 2 (a), (b), (c) shows the performance of the three algorithms before and after applying the Evolutionary Search.

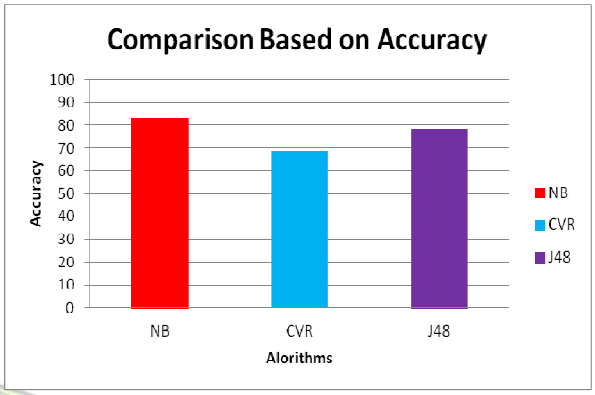

Figure 1 (a)

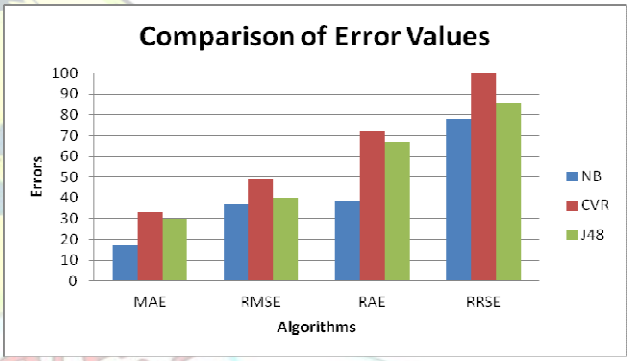

Figure 1 (b)

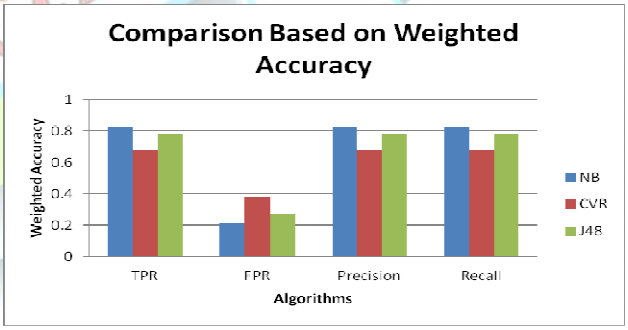

Figure 1 (c)

Figure 1 (a), (b), (c) : Comparison of the Existing Algorithms before Applying ES

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

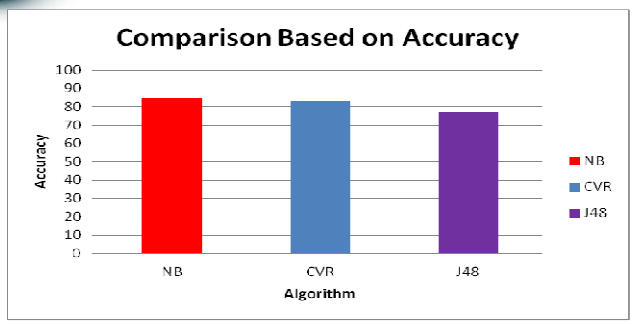

Figure 2(a)

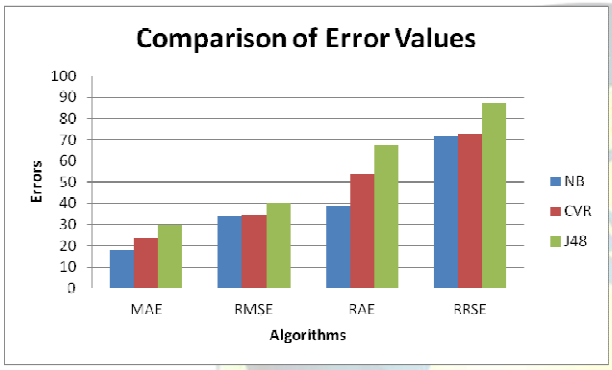

Figure 2 (b)

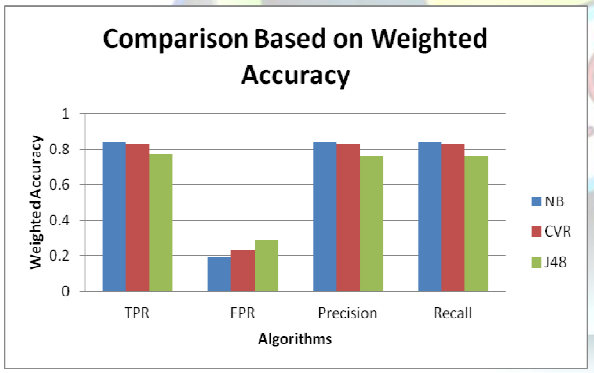

Figure 2 (c)

Figure 2 (a), (b), (c) : Comparing the Existing Algorithms after applying Evolutionary Search

VI. CONCLUSION

Feature Reduction is a crucial and important step in Data Mining. The features should be reduced without affecting the knowledge extracted from the data set. The three algorithms NB, J48 and CVR is tested before applying Feature Reduction by using the Evolutionary Search. It is observed that before and after applying the ES, the NB algorithm maintains the Accuracy level. And comparing all the other metrics, the performance of NB is good. To improve further accuracy, the algorithm can be hydride with other techniques which is the future scope of this work.

References

[1] Bhagyashree Ambulkar and Vaishali Borkar "Data Mining in Cloud Computing", MPGINMC, Recent Trends in Computing, ISSN 0975-8887, pp 23-26,2012.

[2] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2013

[3] M.A. Jabbar et.al, "Knowledge discovery using associative classification for heart disease prediction", AISC Vol.182, pp29-39, 2012.

[4] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

[5] Liu, H. and Yu, L. (2005), "Toward integrating feature selection algorithms for classification", IEEE Trans. Knowledge Data Engineering, vol. 17, pp, 491-502, 2005.

[6] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review , Vol. 8, No. 1,January - June 2007.

[7] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer", 7, 295-304, 1998.

[8] C.Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction", pp. 17-24, 2006.

[9] David Cooke, Carlos Ordonez, Ernest V. Garcia, Edward Omiecinski, Elyzabeth Krawczynska, Russell Folks, Cesar Santana, Levien de Braal, and Norberto Ezquerra." Data mining of large myocardial perfusion spect (mps) databases to improve diagnostic decision making", Journal of Nuclear Medicine, 40(5), 1999.

[10] C.Ordonez, "Mining Constrained Association Rules to Predict Heart Disease", In IEEE ICDM Conference, pp. 433–440, 2001.

[11] Margaret H. Danham, S. Sridhar, " Data Mining Introductory and Advanced Topics", Pearson Education, 1st edition, 2006.

[12] Aman Kumar Sharma, Suruchi Sahni, " A Comparative Study of Classification Algorithm for Spam email data analytics", IJCSE, Vol 3 (5), 2011, pp 1890-1895

[13] http://www.jstor.org/discover/10.2307/

[14] http://stackoverflow.com/questions/10317885/decision-tree-vs-naive-bayes-classifier.

[15] George Dimitoglou, James A. Adams and Carol M. Jim, "Comparison of the C4.5 and a Naïve Bayes Classifier for the prediction of Lung Cancer Survivability",volume-4-Issue-8-August-2012.

[16] Seongwook Youn, Dennis McLeod, " A Comparative study for Email Classification".

[17] S. Vijaykumar, M. Balamurugan, S.G. Saravanakumar, Unique Sense: Smart Computing Prototype, Procedia Computer Science, Volume 50, 2015, Pages 223-228, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.04.056.

[18] Vijaykumar, S., Saravanakumar, S., & Balamurugan, M. (2015). Unique Sense: Smart Computing Prototype for Industry 4.0 Revolution with IOT and Bigdata Implementation Model. *Indian Journal Of Science And Technology, 8*(35). doi:10.17485/ijst/2015/v8i35/86698

[19] Vijaykumar S, Dr. M. Balamurugan, Ranjani K, Big Data: Hadoop Cluster Deployment on ARM Architecture, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 4, Special Issue 1, June 2015, ISSN 2278-1021 & 2319-5940

[20] Durairaj. M, Sivagowry. S,"Feature Diminution by Using Particle Swarm Optimization for Envisaging the Heart Syndrome", IJITCS, vol.7, no.2, pp.35-43, 2015. DOI: 10.5815/ijitcs.2015.02.05