International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Augmentation of Classifier Accuracy through Implication of Feature Selection for Breast Cancer Prediction

Deepa B G, Senthil S , Gupta Rahil M, Shah Vishakha R

Abstract: ***Breast Cancer Examination and Prediction are great provocations to the researchers in the medical applications. Breast Cancer Examination distinguishes benign from malignant breast lumps, Breast Cancer Prediction has great deal in foretelling when Breast Cancer is expected to reoccur in patients that have had their cancers excised. Feature Selection is considered to be the preliminary step used in process to find best subsets of attributes. In this paper authors confer about the performance of five classifiers Sequential minimal optimization (SMO), Multilayer Perceptrons, Kstar, Decision Table and Random Forest with and without feature selection. The results manifest that after implying two feature selection techniques such as Correlation based and information based with ranker algorithm there is an augmentation in the accuracy rate of the classifier. It has been observed that after through implication feature selection techniques accuracy of the classifiers such as SMO, Multilayer Perceptrons, Kstar, Decision Trees, and Random Forest are enhanced.***

Index Terms: Breast Cancer, Feature selection, Data Mining. SMO, Multilayer Perceptron's, Kstar, Decision Table, Random Forest, Information gain based feature selection, correlation based feature selection.

I. INTRODUCTION

Breast Cancer is one of the most seen cancer in India. Now a days incidence of Breast cancer in the age group between 30 -40 is exceeding. Researchers can extricate women's from breast cancer and models developed can be life savior by helping doctors in early detection of Breast cancer without side effects and with nominal charges.

Data Mining is the process of extracting meaningful information from large volume of databases for scientific decision making. In particular medical data mining plays a vital role in data analysis that helps the medical practitioners in predicting the concealed relations and patterns for decision making. It is an Interdisciplinary field and widely used by most of the researchers in Health sector for model building.

Classification is a supervised learning algorithm used by many researchers in diagnosing the diseases. The Classification algorithms used to setup the models in the topical paper are SMO, Multilayer Perceptrons, Kstar, Decision Trees, and Random Forest.

Feature selection plays a major role in increasing the accuracy rate of classifier within the minimum amount of time by reducing the irrelevant and noisy data. Selection of most relevant features by reducing the redundant features will help to increase the accuracy of classifier and it is one of the major area of research in data mining and knowledge discovery.

The flow of remaining research paper is as follows: section 2 speaks about literature survey of different feature selection algorithms and their performance. Section 3 describes about feature selection techniques, Section 4 briefs about the data sets used. Section 5 presents the proposed method of comparing classifier accuracy with and with two feature selection techniques and experimental results are shown. Conclusion is in section 6.

II. REVIEW OF LITERATURE

Significant amount of work has been carried out in prediction of Breast Cancer using feature selection and using Classification techniques.

Ahmed Iqbal Pritom , Shahed Anzarus Sabab [1] discussed about importance of feature selection algorithm on the classifiers accuracy. Feature selection algorithms used are Ranker algorithm, InfoGain AttributeEval and the classifiers are C4.5, Naive Bayes, Support Vector Machine (SVM) and Decision Tree. SVM results in better accuracy before and after feature selection, where in Naïve Bayes and Decision Tree results in better accuracy after feature selection method. Runjie Shen, Yuanyuan Yang and Feng Feng Shao [2] proposed a diagnostic model by using the INTERACT feature selection technique where in 9 features among 32 are selected from this method and selected features are passed to SVM classifier by using 10 fold cross validation to check the improvement in the accuracy of classifier, authors through the experiments proved that performance of SVM improved after feature selection. Hiba Asria,Hajar et.al [3] describes the effectiveness and efficiency of different classifiers SVM, C4.5, NB and k-NN in finding accuracy, precision, sensitivity and specificity on Wisconsin Breast Cancer WBC (original) from UCI repository datasets and concluded that SVM classifier performs better in Breast Cancer Prediction by achieving the results of 97.13% .

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

Mrs. Deepa*, B G, Assistant Professor, holds MCA in Computer Applications from VTU and B.Sc. in Computer Science from Kuvempu University.

Dr. S. Senthil, Professor and Director, School of Computer Science and Applications, REVA University, India.

Gupta Rahil M, Department of Computer Science and Applications, REVA University, India.

Shah Vishakha R Department of Computer Science and Applications, REVA University, India.

Subrata Kumar Mandal [4] main focus of this paper is to identify the cancer in early stage, for that authors used feature selection algorithm PCC and selected number of features passed to the classifiers Naïve Bayes, Logistic Regression and Decision Trees and classified the Breast Cancer tissues as Benign and Malignant, finally based on the time complexity and classification accuracy authors concluded that Logistic Regression algorithm results in best accuracy.

Aalaei Sh, Shahraki H, Rowhanimanesh AR, Eslami S [5] presents the advantage of wrapper approach using GA-based in feature selection on 3 different data sets from Wisconsin WBC, WPBC and WDBC on different classifiers PS-Classifier and Artificial Neural networks (ANN), and concluded that classifiers results in improvement in accuracy after feature selection. In WBC data set PS- Classifier given better accuracy on the other end ANN results in better accuracy for WPBC and WDBC data sets.

Hajar Alharbi et.al [6] employs a new framework by reducing features by extracting and selecting best features among 1100 mammography images. The proposed approach uses five feature ranking methods like mRMR-Minimum Redundancy Maximum Relevance, Fisher Score, Sequential forward feature selection, genetic algorithms and relief-f. The authors take the data set from IRMA database, extracted 109 features among that selected 49 as important in classifying the tissues in to benign or malignant. The classifier used is Neural network and achieved accuracy of 94.27%, Specificity of 99.27% and Sensitivity of 98.36% by using novel feature selection approach.

M.F. Akay [7] proposed a method F-Score + SVM that results in accuracy of 99.51% , that was the highest accuracy rate achieved as compare to existing work, the features are selected based on the F-Score of each feature, and experiments are done on different percentage of training and test data like 50-50, 70-30 and 80-20 and results in better accuracy.

D.Lavanya et al [8] authors experimented with different features election method on different data sets by using Decision tree classifier CART and proved that the results obtained after applying the feature selection improved. Results has been shown in terms of Accuracy, Time and in tree size for different data sets. Authors in their paper clearly specified how to select the best feature selection algorithm among the existing one to improve the accuracy of the classifier in predicting the Breast Cancer.

Borges, Lucas Rodrigues [9] two machine learning algorithms J48 and Bayesian Networks algorithms are compared in the paper in terms of accuracy and the author discussed about classification breast cancer tissues in to benign or malignant. As a conclusion Bayesian algorithm results with 97.80% and J48 96.05 accuracy.

III. FEATURE SELECTION METHODS

Feature selection is the process of eliminating redundant and irrelevant attributes as much as possible to reduce the dimensionality of the data. Feature selection are broadly classified into three categories: Filter, Wrapper and Embedded approach. In this paper two feature selection techniques such as Correlation-based and information based with ranker algorithm are used to increase the classification accuracy.

In correlation based feature selection (CFS) algorithm attributes will be highly compatible with target variables but not consonant with each other [11]. Correlation between the attributes will be in the range of 0 and 1, where 1 indicates attributes are highly correlated and 0 indicates no correlation between attributes. CFS can be evaluated on natural and artificial datasets that quickly eliminate noisy, irrelevant and redundant data. CFS works faster for larger datasets and improves the performance of machine learning algorithms. CFS algorithms is fully automatic and won't expect the user to specify the number of attributes to be selected. CFS is one of the filter approach.

The following equation gives how the features are correlated [12][13].

$$Merits\ S_k = \frac{K\,\overline{r_{cf}}}{\sqrt{K+K(K-1)\overline{r_{ff}}}} \quad (1)$$

In (1) $\overline{r_{cf}}$ is the average value of feature- classification correlation, and $\overline{r_{ff}}$ represents feature –feature correlation.

Information-based ranker algorithm works on information gained from the attributes helps in attribute evaluation process [14], In this method researchers obtained the subset of features by using minimum redundancy and maximum relevance among the features. Information gain is mainly used to reduce the overall entropy. Formula to calculate the entropy D is given, m represents classes, Pi probability, log2 represents information can be encoded in bits [17].

$$Info(D) = -\sum_{i=1}^{m}(P_i)log_2(P_i) \quad (2)$$

If D is classified into some feature attributes K{ k1, k2, …, kv} , D will divide into v sub partitions that is {D1, D2, …., Dv}.

Information required in exact classification can be done using the formula (3),

$$Info_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|} \times Info(D_j) \quad (3)$$

Where modulus of Dj divided by modulus of D is weight of jth partition and info(Dj) represents entropy of the partition Dj.

Information partition on K attributes can be calculated using (4)

$$Information\ Gain(A) = Info(D) - Info_A(D) \quad (4)$$

The highest information gained features will be given highest rank and selection of features will be based on highest ranked attributes.

IV. DATA SETS

Data Sets are taken from Wisconsin –Breast –Cancer data set consisting of 699 instances of 10 attributes, characteristics of Data sets is Multivariate, Attributes are of integer type [18]. There are 16 Missing values and Percentage of malignant and benign are 241(34.5%) and 458(65.5%).

Table 1: Description of Dataset.

Sl.No	Attributes	Domain
1	Sample code number	Id number
2	Clump Thickness	1-10
3	Uniformity of cell size	1-10
4	Uniformity of cell shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial cell size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10

V. EXPERIMENTAL EVALUATION

Experiments are conducted by using WBCD and the model is evaluated as follows.

Table-2 Classifier accuracy without using feature selection techniques

Sl.No	Name of the Classifier	Prediction Accuracy	Time taken to build the model
1	Sequential minimal optimization(SMO)	96.995	0.01
2	Multilayer Perceptrons	95.279	0.61
3	KStar	95.422	0.01
4	Decision Table	95.279	0.14
5	Random Forest	96.566	0.26

Figure-1 Graphical representation of Classifier accuracy without using feature selection techniques

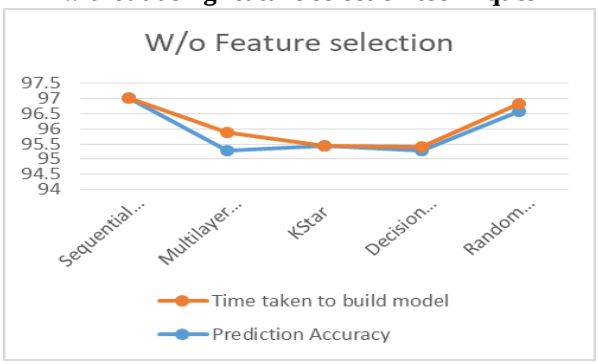


Table-3 Classifier accuracy using correlation based feature selection technique

Sl.No	Name of the Classifier	Prediction Accuracy	Time taken to build the model
1	Sequential minimal optimization(SMO)	96.995	0.01
2	Multilayer Perceptrons	95.994	0.42
3	KStar	95.994	0.01
4	Decision Table	95.422	0.03
5	Random Forest	96.281	0.08

Figure-2 Graphical representation Classifier accuracy using correlation based feature selection technique

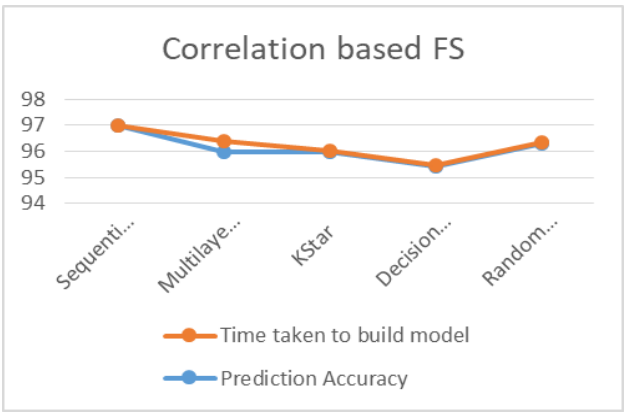
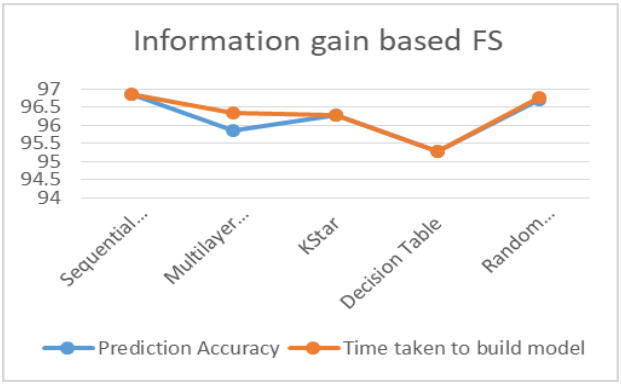


Table-4 Classifier accuracy using Information gain based feature selection technique

	Name of the Classifier	Prediction Accuracy	Time taken to build model
1	Sequential minimal optimization(SMO)	96.8526	0.01
2	Multilayer Perceptrons	95.8512	0.48
3	KStar	96.2804	0.01
4	Decision Table	95.279	0.01
5	Random Forest	96.709	0.06

Figure-3 Graphical representation Classifier accuracy using Information gain based feature selection technique



VI. CONCLUSION

Accuracy of the classifiers such as SMO, Multilayer Perceptrons, Kstar, Decision Table, and Random Forest before feature selection is 96.995, 95.279, 95.442, 95.279 and 96.566 respectively. After incorporating correlation based feature selection approach the classifier accuracy is 96.995, 95.994, 95.994 95.422, and 96.281 respectively. The result of Information gain based feature selection approach is 96.852, 95.815, 96.280, 95.279, and 96.709 respectively. Empirical outcomes demonstrate that 1% enhanced outcome in Multilayer Perceptrons after applying correlation based feature selection by calculating the value [(95.994-95.279)/95.994]*100, 1% improvement in the classifier accuracy in Kstar after applying Information gain based feature selection approach by calculating [(96.280-95.442)/96.280]*100 and 1% enhanced result in the classifier accuracy in Random Forest after applying the both feature selection by calculating [(96.281-95.566)/95.566]*100 and [(96.709-95.566)/95.566]*100.Classifiers accuracy are reinforced if it is deployed after the feature selection.

REFERENCES

1. Pritom, Ahmed Iqbal, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab. "Predicting breast cancer recurrence using effective classification and feature selection technique." In Computer and Information Technology (ICCIT), 2016 19th International Conference on, pp. 310-314. IEEE, 2016.
2. Shen, Runjie, Yuanyuan Yang, and Fengfeng Shao. "Intelligent breast cancer prediction model using data mining techniques." In Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on, vol. 1, pp. 384-387. IEEE, 2014.
3. Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.
4. Mandal, Subrata Kumar. "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree." International Journal Of Engineering And Computer Science 6, no. 2 (2017).
5. Aalaei, Shokoufeh, Hadi Shahraki, Alireza Rowhanimanesh, and Saeid Eslami. "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets." Iranian journal of basic medical sciences 19, no. 5 (2016): 476.
6. Alharbi, Hajar, Gregory Falzon, and Paul Kwan. "A novel feature reduction framework for digital mammogram image classification." In Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on, pp. 221-225. IEEE, 2015.
7. Akay, Mehmet Fatih. "Support vector machines combined with feature selection for breast cancer diagnosis." Expert systems with applications 36, no. 2 (2009): 3240-3247.
8. Lavanya, D., and Dr K. Usha Rani. "Analysis of feature selection with classification: Breast cancer datasets." Indian Journal of Computer Science and Engineering (IJCSE) 2, no. 5 (2011): 756-763.
9. Borges, Lucas Rodrigues , "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection", Proceedings of XI Workshop de Visão Computacional - October 05th-07th, 2015.
10. Karabulut, Esra Mahsereci, Selma Ayşe Özel, and Turgay Ibrikci. "A comparative study on the effect of feature selection on classification accuracy." Procedia Technology 1 (2012): 323-327.
11. https://www.cs.waikato.ac.nz/~mhall/thesis.pdf
12. M. Hall 1999, Correlation-based Feature Selection for Machine Learning
13. Senliol, Baris, Gokhan Gulgezen, Lei Yu, and Zehra Cataltepe. "Fast Correlation Based Filter (FCBF) with a different search strategy." In Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on, pp. 1-4. IEEE, 2008.
14. Dinakaran, S., and P. Ranjit Jeba Thangaiah. "Role of attribute selection in classification algorithms." International Journal of Scientific & Engineering Research 4, no. 6 (2013): 67-71.
15. Jiawei Han and Micheline Kamber, "Data mining concepts and techniques" Morgan Kaufman Publishers, 2006 Elsevier pp. 297- 298
16. https://www.eecs.yorku.ca/tdb/_doc.php/userg/sw/weka/doc/weka/attributeSelection/InfoGainAttributeEval.html.
17. Sharma, Anuj, and Shubhamoy Dey. "Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis." IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications 3 (2012): 15-20.
18. https://www.kaggle.com/roustekbio/breast-cancer-csv.

AUTHORS PROFILE

Mrs. Deepa, B G, Assistant Professor, holds MCA in Computer Applications from VTU and B.Sc. in Computer Science from Kuvempu University. Having 7.7 years of teaching experience. Presented papers in National level conferences, published technical papers in International journals. Pursuing Ph.D in REVA University in Data Mining (Medical diagnosis).

Dr. S. Senthil is Professor and Director, School of Computer Science and Applications, REVA University, India. Previously, he worked as Associate Professor and Head, Department of Computer Science, Vidyasagar College of Arts and Science, India. He obtained his B.Sc. (Applied Science – Computer Technology) from PSG College of Technology, India, in 1995, Master of Computer Applications from Bharathidasan University, India, in 1999, M.Phil. in Computer Science from Manonmaniam Sundaranar University, India in 2002, and Ph.D. in Computer Science from Bharathiar University , India, in 2014. Another achievement is clearing SET in 2012. He has published 40 research papers in various reputed National and International Journals. He has presented a paper entitled "Lossless Preprocessing Algorithms for better Compression" in an IEEE International Conference at Zhangjiajie, China. His interest is in Database Systems, Data Mining, Data Compression, and Big Data Analytics.