IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Dimensionality Reduction and Comparison of Classification Models for Breast Cancer Prognosis

R. Garg[1*], V. Mongia[2]

[1*] Computer Science, Guru Nanak College, Moga, Panjab University Chandigarh, India
[2*] Computer Science, Guru Nanak College, Moga, Panjab University Chandigarh, India

**Corresponding Author: 87rajnigarg@gmail.com*

Available online at: www.ijcseonline.org

Received: 06/Jan/2018, Revised: 10/Jan/2018, Accepted: 25/Jan/2018, Published: 31/Jan/2018

Abstract- Cancer is a most prevailing problem in the society now days. Generally cancer specifically Breast cancer is a major problem in women. On among three cases of cancer is a Breast cancer. There are many factors that affect the cancer. All these factors and the symptoms in the patient can be recorded using hardware and software. Now days, due to advancement in technology data of patient is recorded and processed by using analytical method. Data mining provides various methods to process this data effectively and efficiently. This processed data can be proven very useful in earlier detection of diseases. The earlier detection of these symptoms can be proven helpful to save life of a patient. In our research, original data on Breast cancer from Winconsin has been taken. This data set has 10 attribute and 699 instances. In this study, a comparative model has been developed that compare performance of various data mining technique on the dataset. The study reveals that BayesNet is the best classifier that correctly predicts cancer survivability in the patient. Further, KStar is the fastest algorithm that takes lowest computation time for the classification. In the next step dimensionality reduction using gain ratio is performed to find out most dominant factors causing Breast cancer.

Keywords- Data Mining, Breast Cancer, Bayesian, SVM, Decision Tree, Regression Model

I. INTRODUCTION

Introduction: Breast Cancer is a most prevailing problem in the society. Every year million of cases of Breast cancer are reported worldwide. In the past decades many innovative methods and techniques has been developed for earlier detection of breast cancer but further advancement are required for detection, prevention or cure of this disease. Past data of Cancer patients can be very useful for prediction of Breast cancer in a patient. In this study data of 699 patients have been collected and ten data mining algorithms are applied on this dataset to identify the best classifier for prediction of Breast cancer in the patient. Furthermore, dimensionality reduction has been performed to find out most dominant factors causing breast cancer.

II. DATA MINING

To discover new information from the present data, different data mining techniques proposed by [1] are used. N.T.Nghe. et.al has made comparative study of these data mining techniques on WEKA tool [2]. Due to the computational efficiency and speed of WEKA [3] the same is used in this research. The most commonly used data mining methods are: Association, classification and clustering.

Association: association rule is used to find the relationship between one instance to another instance [4]. In the context of our research association rule can be used to find the linking between patient's attribute and breast cancer disease. If the patient's normal nucleoli, cell size and uniformity of cell size are not in range then he is most likely to chance of breast cancer.

Classification: in classification technique the whole dataset is divided into set of predefined classes. That's why classification is also called supervised learning [4]. For a instance cancer predictor's classifier classify class of patient into benign or malignant. This technique divides the whole process into two phases. In the first phase a model is built with the help of training data and in the next phase this model is tested with test tuples and its accuracy is determined. Backpropagation. K-nearest neighbor and

decision tree are good example of classification technique. In this research decision tree are used for the prediction.

Clustering: Clustering divides the dataset into different regions called clusters [4]. Cluster comes under un-supervised learning because classes are not predefined. Object under one cluster have similar values and this value differs from other clusters for instance in breast cancer dataset, cluster can be generated on the basis of patient's normal nucleoli, cell size and uniformity of cell size etc.

LITERATURE REVIEW

In [5] analyzed the prediction of survivability rate of breast cancer using different data mining techniques. In his study 151,886 record and16 attribute are considered. Three main decision tree algorithm were used for their study namely Naïve Bayes, back-propagated neural network, and the C4.5. Finally, they conclude that C4.5 algorithm has a much better performance than the other two techniques

[6] Reviewed that decision tree is best technique of data mining to detect high risk of breast cancer. In his study he used the dataset produced by department of Genetics of faculty of Medical Science of University Nova Lisboa with 164 control and 94 cases. The study reveals that there is significant relationship with Breast cancer by driving a decision tree and selecting the best leaf. All study has been performed under the WEKA machine learning tool. They found that high risk breast cancer group composed of 13 cases and only 1 control with fisher test value of $9.7 * 10^{6}$ and p-value of 0.017

In [7] author reviewed various researcher articles on Brest Cancer diagnosis and prognosis problem and applied data mining techniques to uncover hidden patterns that can helps clinician in decision making. It also discovered various most contributed attribute that leads to Brest Cancer, so this research helps the medical professional in decision making for early diagnosis and avoid biopsy. This study also revealed that ANN produced the highest accuracy in comparison to other classification techniques.

In [8] researcher compare three classification techniques of data mining named decision tree, artificial neural network and support vector machine. The authors applied classification algrothims on Breast Cancer dataset which was collected from Iranian Center for Breast Cancer (ICBC) program from 1997-2008. There were 1189 records of patients with 22 attributes. Result of their research revealed the accuracy performance of SVM, ANN and DT are 0.957, 0.947 and 0.936 respectively. The accuracy performance of SVM classification model is higher than DT and ANN.

[9] Has improved the accuracy and execution time taken by the Artificial Neural network using the island based model. This island based model reduced the training time taken by the ANN. In this model researcher run a standard sequential evolution algorithm on different island. The migration process is responsible for communication between the sub populations. This model is further divided into many phase i.e migration topology, migration policy, and migration interval and migration size. In this paper author has proposed two different migration topologies and compared results of these two.

[10] Applied different classification and clustering algorithm of data mining on the breast cancer dataset. In their study author showed that classification techniques performed better than clustering technique. C5.0 and SVM two classification techniques are produced 81% accuracy on contrary clustering algorithm fuzzy c-mean gave 37% accuracy which is lowest among other algorithms.

III. DATA MINING TECHNIQUES:

In this research five classification techniques are used: Decision Tree, Bayesian, Support Vector Machine, Regression Model and lazy learner. These models are used in data mining to improve prediction in our research these algorithms are compared to find out best algorithm that predicts the possibility of breast cancer in a patient. In the next section brief description of these classification models is provided.

Decision Tree: Decision Trees are most powerful classification models used in prediction [11]. These models construct tree like structure with class labels as its leaf node. The decision tree classification models used in our research are: J48, ADTree, RandomTree, RandomForest. These algorithms use some mathematical model like: information gain. Gain ration and ginni index. These mathematical models are used to find out splitting attribute from the input parameter.

Bayesian Classification: It is a predicted model which is based on bayes' theorem. NaiveBayesian classifier works

on assumptions "class conditional independence" which means that the effect of an attribute value on a given class is independence of the values of other attributes. In Bayesian classification, a tuple X only belongs to a class C_i only in the class has highest posterior probability condition on X i.e

$$P(Ci/X) > P(Cj/X) \; \forall \; 1 <= j <= m, \; j \neq i$$

$$\text{Where } P(Ci/X) = \frac{P(X/Ci)\,P(Ci)}{P(X)}$$

In this research, BayesNet and NaiveBayes Bayesian classification algorithms are used.

Support Vector Machine: SVM is a classification algorithm which is based on supervised machine learning. In SVM every data element is plotted in n-dimensional space. Number of dimension is equal to number of attributes. After plotting all data elements a line is drawn in such a way that separate two classes completely. Sequential Minimum Optimization (SMO) is used in this research to classify breast cancer training data.

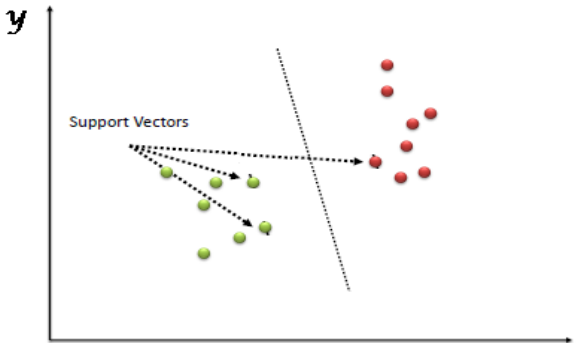

Fig. Lineally Separated using SVM

Lazy Learner: It is a classification model which is also called 'learning from your neighbors'. Lazy Learner classification model is so called because unlike either learner these model do not create classification model when the training data is provided. It simply stores the training data. When a test tuple is inserted then this algorithm compares it with the similar tuple of training data. KStar classifier and cased based reasoning classifier are lazy learners. In our research KStar algorithm is used. This algorithm compares the test tuple to the training dataset that are very similar to the test tuple. The closeness can be measured by the following formula.

$$\text{Dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x1i - x2i)2}$$

Where X_1, X_2 are tuples and x_{1i}, x_{2i} are attributes of these tuples.

IV. DATA MINING PROCESS

Problem Definition: Breast Cancer is a tumor in which cells divide and grow abnormally. It is a major problem prevailing in women worldwide. Till now, exact factor responsible for this problem are unknown. But some attribute like age, life style and family history can play significant role in this disease. Other factors like presence of clumps, its thickness, uniformity of cell size and its shape also be dominating attributes for diagnose of Breast Cancer. In our research the aim is to find the best classifier that predicts the possibility of Breast Cancer on the basis of some attribute of patient in advance so that necessary action can be taken to save life of a patient.

Data Source: To perform research, data is collected from UCI repository. It is an online repository having 412 different data set. The aim of this repository is to provide data to machine learning community. The UCI program is known for its completeness in data and its accuracy.

Understanding the Data: The Dataset on Breast Cancer is collected from UCI repository. It is donated by University of Winconsin Hospital. It has 10 attributes and 699 instances of patients. First nine attributes represents instances and last attribute is a class attribute with two possible outcomes: Benign and Malignant. This dataset has 65% instance of benign cancer and 35% instances of malignant cancer. The attributes of the dataset are shown in the table.

Sr. No	Attribute	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
11	Class:	(2 for benign, 4 for malignant)

Table: Representing Attributes of Data Set

V. EXPERIMENT AND RESULT

In this section, ten different classification algorithms applied on Breast Cancer dataset.

In this section, ten different classification algorithms are applied on Breast Cancer dataset. WEKA machine learning tool is used to analyze the result of these algorithms. Some parameters like correctly classified, time taken, Kappa statistic etc. are observed of each classification algorithms which is shown in the following table.

A) ACCURACY AND COMPUTATION:

In this research BayesNet classifier gave 97.14% accuracy, which is the highest among the other classification algorithms. Other classifier algorithms like SMO and SimpleLogistic also gave accuracy near to the BayesNet. On the contrary KStar took minimum time i.e .005 second to compute 699 instances with 10 attributes. But the accuracy rate of KStar is not as good as BayesNet.

The following table represents result of ten classification algorithms on Breast Cancer Dataset on the basis of classification accuracy, time and error. Error associated with the classification is determined as mean absolute error, root mean squared error and relative absolute error and root relative squared error.

B) DIMENSIONALITY REDUCTION

Dimensionality reduction is features of classification technique in which those attribute are removed from the studies which don't contribute to the result or which affect very less. So if those attribute are removed from the study then overall result can be improved. In this study gain ration technique is used to find the most contributing attribute and then ranker algorithm is applied to rank the attribute according to their importance in descending order according and

	Correctly classified	Incorrectly classified	Time	Kappa Statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
BayesNet	97.14	2.86	0.04	.9374	.0289	.1628	6.3926	34.26
NaïveBayes	95.99	4.01	0.02	.9127	.0403	.1983	8.926	41.742
SMO	96.71	3.29	0.14	.9274	.0329	.1814	7.2802	38.164
SimpleLogistic	96.14	3.86	0.37	.9143	.0521	.169	11.5331	35.553
VotedPerceptron	65.52	34.48	0.02	0.0	.3448	.5872	76.2835	123.54
KStar	81.26	18.74	0.005	.5484	.1865	.4213	41.2667	88.630
J48	94.56	5.4	0.04	.8799	.0691	.2228	15.2992	46.8739
ADTree	95.28	4.72	0.07	.8956	.0764	.1879	16.9071	39.4067
RandomTree	93.99	6.01	0.01	.8657	.0606	.2453	13.4084	51.6111
RandomForset	94.56	5.4	0.01	.8799	.0691	.2228	15.2992	46.8739

last three attributes having lowest gain ratio can be removed.

The attributes Marginal Adhesion , Clump Thickness, Sample code number has minimum gain ratio. These three attributes will be removed because they do not contribute toward data classification. This will result in less computation time and less memory requirement

Sr. No	Ranked	attributes:
1	0.399	Normal Nucleoli
2	0.395	Single Epithelial Cell Size
3	0.386	Uniformity of Cell Size
4	0.374	Bare Nuclei
5	0.314	Uniformity of Cell Shape
6	0.303	Bland Chromatin
7	0.299	Mitoses
8	0.271	Marginal Adhesion
9	0.21	Clump Thickness
10	0	Sample code number

Table Representing Gain Ratio of all attributes

VI. CONCLUSION:

The experimental results have shown that different classification algorithms behave differently on the same dataset. Some algorithms are good in correctly classification, some are good in execution time and some algorithms are good in mean squared error etc… Some attributes do not contribute to the target variable and if these attribute are removed from the data set then overall performance of the algorithm can be improved. In our experiment there are three variable named Marginal Adhesion , Clump Thickness, Sample code number do not contribute the class attribute and if we remove these attribute, correctly classified instance will remain same but execution speed of the algorithm will surely increase.

REFERENCES

[1] H. Trevor, T. Robert, and F. Jerome, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., vol. 2. Springer: New York, 2009, pp. 32-36.

[2] N.T.Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance", ASEE/IEEE Frontiers in Education Conference, pp. T2G7-T2G12, 2007.

[3] M. Lichman, UCI Machine Learning Repository, http://www.cs.waikato.ac.nz/ml/weka,2013

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. University of Illinois at Urbana-Champaign Elsevier San Francisco, 2009, pp. 285-306

[5] Bellaachia, Abdelghani, and Erhan Guven, "Predicting breast cancer survivability using data mining techniques", Age, Vol. 58, Issue 13, 2006, pp. 10-110

[6] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose, "A Data Mining approach for detection of high-risk Breast Cancer groups," Advances in Soft Computing, vol. 74, pp. 43-51, 2010.

[7] Shelly Gupta, Dharminder Kumar,Anand Sharma, "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS "
Vol. 2 No. 2 Apr-May 2011

[8] Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence" Health and Medical Informatics 2013, 4:2

[9] Htet Thazin Tike Thein1 and Khin Mo Mo Tun "An Approach For Breast Cancer Diagnosis Classification Using Neural Network" Advanced Computing: An International Journal (ACIJ), Vol.6, No.1, January 2015

[10] Uma Ojha, Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques" Cloud Computing, Data Science & Engineering - Confluence, 2017 , Noida Inida.

[11] K. Saravanapriya and J. Bagyamani, "Performance analysis of Classification Algorithms on Diabetes DataSet" . International Journal of Computer Science and Engineering (IJCSE), Vol. 5, Issue-9, pp 15-20, sept-2017

Authors Profile

Ms. R. Garg pursed Master of Computer Application from GNIMT in year 2010. She is currently working as Assistant Professor in Department of Computer Science, Guru Nanak College, Moga. She has 5 years of teaching experience.

Mr V. Mongia pursed Master of Computer Application from GHCMT Punjabi University Patitala in year 2005. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science, Guru Nanak College Moga. He has 10 years of teaching experience and 2 years of Research Experience.

