International Journal of Engineering & Technology, 7 (4.20) (2018) 22-26

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers

Jabeen Sultana[1]*, Abdul Khader Jilani[2]

[1]*Department of Computer Science , Majmaah University , Kingdom of Saudi Arabia*
**Corresponding author E-mail:*j.sultana@mu.edu.sa

Abstract

The primary identification and prediction of type of the cancer ought to develop a compulsion in cancer study, in order to assist and supervise the patients. The significance of classifying cancer patients into high or low risk clusters needs commanded many investigation teams, from the biomedical and the bioinformatics area, to learn and analyze the application of machine learning (ML) approaches. Logistic Regression method and Multi-classifiers has been proposed to predict the breast cancer. To produce deep predictions in a new environment on the breast cancer data. This paper explores the different data mining approaches using Classification which can be applied on Breast Cancer data to build deep predictions. Besides this, this study predicts the best Model yielding high performance by evaluating dataset on various classifiers. In this paper Breast cancer dataset is collected from the UCI machine learning repository has 569 instances with 31 attributes. Data set is pre-processed first and fed to various classifiers like Simple Logistic-regression method, IBK, K-star, Multi-Layer Perceptron (MLP), Random Forest, Decision table, Decision Trees (DT), PART, Multi-Class Classifiers and REP Tree. 10-fold cross validation is applied, training is performed so that new Models are developed and tested. The results obtained are evaluated on various parameters like Accuracy, RMSE Error, Sensitivity, Specificity, F-Measure, ROC Curve Area and Kappa statistic and time taken to build the model. Result analysis reveals that among all the classifiers Simple Logistic Regression yields the deep predictions and obtains the best model yielding high and accurate results followed by other methods IBK: Nearest Neighbor Classifier, K-Star: instance-based Classifier, MLP- Neural network. Other Methods obtained less accuracy in comparison with Logistic regression method.

Keywords: *Breast Cancer Data, Classification, Decision Trees (DT), Logistic Regression, Multi-Layer Perceptron (MLP), and Prediction.*

1. Introduction

Breast cancer remains to be the outmost identified cancer in the whole universe and is the prime source of cancer demise amid women. Earlier detection of breast cancer can save many lives in a best effective manner. Without reaching for surgical biopsy, prompt diagnosis entails accurate and steadfast diagnosis process that permits medical practitioner to differentiate benign breast tumors from malignant tumors. Every single minute, anywhere all over the globe context of breast cancer is identified amongst females and every one minute, everywhere all over the globe and somebody expires from breast cancer. Breast tumors can be identified and further classified into three different categories known as benign breast cancers, in situ cancers, and invasive cancers. Benign breast tumors fall into the Chief category of tumors often detected by undergoing mammography. They cannot extent to external organs as by nature they are non-cancerous. By means of mammography in rare cases, this one is hard to discriminate specific bulk of benign from malignant lesions. In situ or noninvasive cancer, is fully confined in the ducts. Also, the basal membrane should be free from malignant cells. Coming to invasive, if the cancer spreads above the basal membrane and overreaches into the neighboring tissue. Consequently, early detection of breast cancer is crucial. To classify patients into either non-cancerous group called into ”benign” or cancerous group into ”malignant” is the prime purpose of these forecasts.

Machine learning permits processors to learn from prior instances, to intimate complex patterns from huge, noisy or compound data sets. It is a separation of artificial intelligence which employs a range of optimized, statistical and probabilistic techniques. In general, this expertise is fine suitable to health applications, definitely those which relies on difficult genomic and proteomic measurements. Currently machine learning methodologies are being applied to detect and classify tumors in broad range of medical solicitations. To diagnose and assist cancer, machine learning has been used initially and foremost. It permits interpretations or conclusions to be made that could not rather be made using standard statistical procedures. Therefore, it is intensely more influential because of this reason [1]. Remarkably, supervised learning is applied in more or less all algorithms of machine learning used in diagnosing prognosis and tumor prediction. Besides this, conditional probabilities are considered as the base for classification tasks or to a definite group of classifiers and used by most of the supervised learning algorithms.

Data mining applications are developed to Predict the consequences of ailments and is one of the most exciting and exigent tasks. Huge volumes of health data are being composed

and are easily accessible by the medical investigation groups since the use of computers powered with automated tools. In order to mine the knowledge from huge databases, knowledge discovery takes additional time by the standard Machine Learning Techniques [2]. Medical practioners mostly use Data Mining tools in order to classify and exploit patterns and associations amongst great quantity of variables. They help to predict the consequent results of disease using the past circumstances warehoused within datasets. As Extracting suitable information starting from the whole existing data is prominent and overwhelming task within the limited timespan. Various techniques in order to mine and extract valuable information or knowledge from data are being offered by Data mining techniques or methods. These methods are appropriate and best suitable for the whole data that are composed and pertaining to various arenas of discipline. Numerous investigations are carried out and published and also made available about data mining applications in various fields of sciences such as medicine, education, defense, telecommunications, banking, insurance, etc.

In this paper, a comparative study is done amongst different methods of data mining to predict breast cancer diagnosis, treatment & prognosis and a best Model is selected which is the prime objective of this paper. The remaining parts of this paper are divided into subsequent sections. Second section gives the details of Literature part and work done in the area of Breast Cancer. Third section presents data collection and preprocessing. Fourth section explains our methodology in predicting students' performance. The experimental results are presented in Section five, and lastly Section Sixth explains conclusion and future work.

2. Literature survey

This phase provides review of the present day studies being accomplished on breast cancer and the usage of various data mining techniques to predict and diagnoise the breast cancer. Presently, most of the physicians opt to make surgical biopsy in order to figure out different kinds of cancers for benign breast tumors from malignant). As biopsy could be very crucial challenge maximum of them believed that it should be stopped as much as possible. Thus, to recognize the kind of cancer and keep away from needless surgical biopsy, a smart system was presented which can be beneficial for both patients and physicians [3]. Subsequently, Vikas Chaurasia et al. carried out easy RBF, Logistic and REP Tree for prognosis of breast cancer [4]. To predict breast cancer comparative analysis was done by using neural network, decision tree, genetic algorithm and logistic regression by Wei-pin Chang et al. [5]. Their experimental outcomes found out that, amongst those applied techniques for predicting breast cancers lowest prediction accuracy was obtained by decision tree model and better accuracy rate was obtained by logistic regression model. In addition to this, genetic algorithm achieved maximum accuracy by generating standard classification rules inside the class of breast cancers. By making use of specific classification techniques for diagnosis of breast cancers. Shweta Kharya observed from a comprehensive survey and claimed that, decision tree yielded excessive accuracy rate and is the first-class predictor among the involved techniques and the Bayesian network, a well-known technique that's used in medical world [6].
Logistic regression, Decision Tree, Naïve Bayes, neural networks, multi-layer perceptron, and support Vector machine are some algorithms that were applied to diagnoise breast cancers by Senturk et al. [7]. Their experimental outcomes stated that, support Vector machine obtained high classification accuracy compared to other classifiers. Using C4.5 algorithm patients were Categorized into either "Carcinoma in situ" or "Malignant potential" group and confirmed that to diagnoise breast cancers, C4.5 obtained high accuracy by Rajesh et al. [8]. Observations were made from a survey by applying the numerous techniques which can also be utilized by several investigators to diagnoise breast cancers by Gupta and et al. [9]. Sooner or later they noted that, the optimal technique which can yield high rate of accuracy can be determined after developing numerous varieties of models using various algorithms on each instance, by trying distinctive techniques or algorithms.

The most prominent methods used in mining of data are Classification and Prediction. supervised learning is performed for classification tasks. Some portion of data called is taken as training set comprising of instances. Further each instance comprises collection of features and further these features describes one single entity known as as class. The foremost objective of classification technique is to generate a model with proficiency of forecasting the class label as accurate as feasible in earlier hidden records. Furthermore, to predict the correctness or accuracy of the created model, a test set is used [10]. Classifying tumor cells, studying the effectiveness of remedies are some applications of classification in medical diagnosis. Numerous algorithms for classification are proposed such as [11-15]: Decision Tree Induction, Rule-Based Methods, k-Nearest-Neighbor which are Memory-Based Methods, Genetic Programming [14,15]. Grounded on the other attributes values, Regression similar to classification tries to predict the value of an attribute [16]. Separating the data set into two continuous variables and associating them to discover the quality of attributes is the main aim of regression. To simply state regression predicts the quality of one specific attribute grounded on the value of other attribute thereby resulting in feature selection with best quality. Regression can be applied in various medical diagnosis and cancer diagnosis or prognosis [17].

3. Data preprocessing

Processing of data is essential to sort the dataset suitable for several classifiers in order to perform classification. Data in the real world is dirty; it can be cleaned by removing the incorrect and nosy data which is then integrated from multiple sources, known as meta data that removes duplicates and redundant data. Data preprocessing can be achieved by various subsequent means such as reduction of data, data cleaning, data integration, data transformation and data discretization. The data set comprises of the following attributes such as Evenness of Cell Size, Evenness of Cell Shape, Clump Thickness, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. Data set, Wisconsin breast cancer is chosen from the UCI Machine Learning Repository [18]. It is essential to have a suitable Data Mining tool in order to carry out classification, prediction and extracting rules on the data available. 10-fold cross validation is applied on all the dataset using different classifiers. The obtained results are calculated by evaluating the performance of the models using various classifiers for breast cancer using measurements like Accuracy, RMSE, Specificity, Sensitivity, Time taken to build the model, F-measure, ROC curve area and Kappa Statistics.

4. Methods and proposed approaches

4.1 Brief overview about the methodology to predict breast cancer

Mainly, three important phases are proposed in this approach.

- In the first phase, a statistical method called as Cross validation is applied which splits the data in two parts: one used to train a model and the second used to test the model. Also, evaluates and compares by obtained learning models using various algorithms [19].
- Breast Cancer dataset is trained by the classifiers namely, Simple Logistic-regression method, IBK, K-star, MLP, Random Forest, Decision table, Decision Trees, PART, Multi-Class Classifiers and REP Tree, resulting in appropriate Models.
- In the second phase, test data set is applied on the Model obtained from trained dataset and the results are obtained.
- In the third phase, the obtained results are evaluated in terms of measures like Accuracy, RMSE, Specificity, Sensitivity, Time taken to build the model, F-measure, ROC curve area and Kappa Statistics among various classifiers used here and a comparison is drawn leading to selection of the best method.

4.2 Overall description and various classification methods used

In this Paper we have selected the open source software WEKA, proficiently works with limited data [20]. Few tasks offered by data mining like pre-processing of data, Classification, Clustering, Association and Visualization can be performed. Data set is fed in the form of Comma Separated Values-CSV format. In general WEKA is used in medical analysis for preliminary collection of data [21]. A decompositional method is applied to discover information from the Breast Cancer dataset. The extracted knowledge is used for prediction purposes [22]. The framework illustrated in Fig. 1 summarizes the suggested framework for Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers. Experimentations are carried out on the Breast Cancer dataset taken from the UCI machine Learning Repository, fed to various classifiers like Simple Logistic-regression method, IBK, K-star, MLP, Random Forest, Decision table, Decision Trees, PART, Multi-Class Classifiers and REP Tree. The obtained results are calculated and evaluated in terms of measures like Accuracy, RMSE, Specificity, Sensitivity, Time taken to build the model, F-measure, ROC curve area and Kappa Statistics.

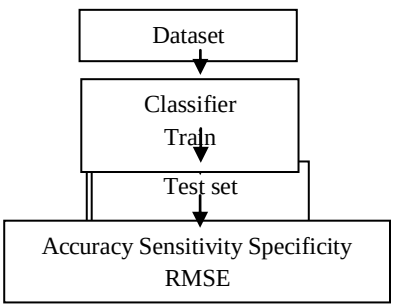

Fig. 1 Frame work representing proposed approach

4.2 Methods used

Logistic Regression: Multinomial logistic regression is used to build a trained model for predicting with a ridge estimator. [23].

IBK: It is a K-nearest neighbor's classifier and calculates distance weighting to select appropriate value of K based on cross-validation. [24].

K-Star: K*, an instance-based classifier. As determined by some similarity function classifies test instances based upon the model obtained after training instances similar to it. [25].

MLP: A neural network, each entity accomplishes a biased weighted sum of inputs to them and passes this activation level through a transfer function to generate output. Logistic and hyperbolic tangent sigmoid functions are the most common activation functions in MLP [26].

Random Forest: Random trees are formed in randomly and resembles like a forest and class is predicted from that [27].

Decision Table: Class is predicted from a simple decision table constructed by majority classifiers [28].

Decision Trees: J48 algorithm is employed to construct the decision tree, starting from the root of the tree and proceeding down to its leaf nodes. Class label for a test item is obtained from a decision tree by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance [29].

PART: A partial C4.5 decision tree is built in each repetition and makes the "best" leaf into a rule Class for generating a PART decision list by using separate-and-conquer strategy [30].

Multi-class Classifier: Multi-class datasets with 2-class classifiers are handled by a meta classifier.

REP Tree: A decision tree is built by means of information gain and prunes it using reduced-error pruning [31].

Datasets are huge in dimensions to solve actual domain classification problems. Knowledge discovery consumes major part of time in order to mine the knowledge from those databases by the standard Machine Learning Techniques [2]. In this paper the hidden knowledge to predict patient's suffering from Breast Cancer is extracted by analyzing the results obtained and are evaluated by considering various parameters and are explained in detail here.

1. The percentage of test instances that are correctly classified

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances by rules}}{\text{Total number of instances in test data}} * 100$$

on a given test set is determined as the accuracy of a classifier.

2. Sensitivity and Specificity is calculated from Confusion Matrix obtained in the model.

3. Sensitivity is the proposition of the positive instances that are correctly identified TP = (TP/ TP +FN) * 100.

4. Specificity is the proposition of the negative instances that are correctly identified
TN = (TN/TN + FP) * 100.

5. RMSE Root Mean Square Error is a measure of the difference between values predicted by a model and the

values actually observed.

$$\text{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

6. F-measure, ROC curve area and Kappa Statistics are also calculated using Confusion Matrix with the help of WEKA tool.

4.3 Confusion Matrix

A comparison is drawn between the actual class labels and the predicted class labels based on the class labels by the classifiers. The following describes the case when we deal with two-class classification problem [33]. The generated confusion matrix is 2 * 2 matrixes.

TP	FN
FP	TN

5. Experimental result and discussions

Classifiers are applied to available data set with the help of WEKA explorer tool. There are 40 classifiers divided into 4 groups in WEKA approximately. In this paper we have chosen few different classifiers and compared the performance of classifiers. The dataset pertaining to Breast Cancer is chosen from UCI-Machine Learning repository, and fed to various classifiers namely Simple Logistic-regression method, IBK, K-star, MLP, Random Forest, Decision table, Decision Trees, PART, Multi-Class Classifiers and REP Tree. The Experiments are carried out on different classifiers, the data set is trained well on each classifier and a Model is obtained then validated with test data, results are obtained. The obtained results are calculated and evaluated in terms of measures like Accuracy, RMSE, Specificity, Sensitivity, Time taken to build the model, F-measure, ROC curve area.

Table 1 Result analysis of breast cancer on various classifiers

Methods	Accuracy	RMSE	TP	FP	F. Measure	ROC	Kappa Statistics	Time to build the Model
Simple Logistics	97.18	0.14	0.97	0.03	0.97	0.99	0.93	0.65
IBK	95.78	0.20	0.95	0.05	0.95	0.95	0.90	0
K-Star	95.43	0.20	0.95	0.05	0.95	0.98	0.90	0
MLP	95.25	0.19	0.95	0.05	0.95	0.98	0.89	5.01
Random Forest	95.25	0.18	0.95	0.05	0.95	0.98	0.89	0.06
Decision Table	94.02	0.21	0.94	0.07	0.94	0.97	0.87	0.16
PART	93.49	0.25	0.93	0.07	0.93	0.93	0.86	0.05
Decision Tree	93.14	0.25	0.93	0.07	0.93	0.92	0.85	0.03
Multi-Class Classifier	93.14	0.25	0.93	0.06	0.93	0.97	0.85	0.11
REP Tree	92.44	0.25	0.92	0.91	0.92	0.96	0.83	0.04

The table 1 describes the Result analysis of breast cancer dataset on various classifiers. Results are deeply analyzed in terms of accuracy and Simple logistic regression yields maximum classification accuracy 97.18, IBK-95.78%, K-star-95.43%, MLP-95.25 %, Random forest-95.25%, Decision table-94.02%, Decision tree-93.14%, PART-93.49, Multi-class Classifier-93.14%, and REP tree-92.44% yields minimum accuracy is given by REP tree. The other observation are made in terms of root mean square error, Specificity, Sensitivity, Time taken to build the model, F-measure, ROC curve area and Kappa Statistics as shown in below Table-1. It is observed that Simple logistic regression achieves high accuracy to predict class or type of cancer to which Cancer patients belong to.

Result analysis reveals that among all the classifiers Simple Logistic Regression yields the deep predictions and obtains the best model yielding high and accurate results with accuracy of 97.18% and low 0.14 rmse followed by other methods IBK: Nearest Neighbor Classifier, K-Star: instance-based Classifier, MLP- Neural network. Other Methods obtained less accuracy in comparison with Logistic regression method.

6. Conclusion and future work

Data Mining could be used to predict the kind of tumor the patient is suffering from in the Medical field as there is increase in the number of Women suffering from Breast cancer. To predict the class of cancer to which a patient may classified, we need to extract the hidden knowledge pertaining to various attributes that could be used to boost the efficiency in general by utilizing the best resources available. In our paper, Comparing the efficiency of different classifiers, namely; Simple Logistic regression, MLP, Multi-Class Classifiers, DT, REP tree, K-star, IBK, Decision table, PART and Random Forest. Results conclude that Simple Logistic regression method obtains the Best Model to predict breast cancer by means of different data mining techniques. Results indicate that Simple Logistic regression obtained best performance in general compared to the other classifiers in terms of classification accuracy, |RMSE, specificity and sensitivity, F-measure, ROC curve area, time taken to build the model and Kappa Statistics. The obtained results conclude that the DT, REP Tree has a loss of accuracy, sensitivity and specificity in knowledge discovery. Compared to Random forest and REP TREE though DT is generally a superior classifier, there is loss of accuracy while making the knowledge acquired by DT. Therefore, it is better to use Simple Logistic regression method directly for knowledge discovery in comparison with DT and REP TREE as efficient methods

In future work rule extraction part can be improved, also try to improve accuracy in DT, REP Tree. To obtain best possible rules with the help of RSES tool, Rough Sets algorithms like LEM2 can be used in DT, NBTREE. Future research work also suggests that this work can be extended to three-class Classification.

References

[1] Cruz JA, Wishart DS, Applications of Machine Learning in Cancer Prediction and Prognosis, Departments of Biological Science and Computing Science, University of Alberta Edmonton,AB, Canada.Vol.2, 2-21 (2006).

[2] Han J., Kamber M., Data Mining Concepts and Techniques. Morgan Kaufman Publishers, 2001.

[3] McCarthy et al. Applications of Machine Learning and High - Dimensional Visualization in Cancer Detection, Diagnosis, and Management. (2004).

[4] Chaurasia V, Pal S. Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. International Journal of Computer Science and Mobile Computing. Vol3, 10–22 (2014).
[5] Chang PW, Liou MD, editors. Comparison of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancerdata.AvailableOnline: http://edoc.ypu.edu.tw:8080/paper/ha/Other/%E5%BC%B5%E5%81%89%E6%96%8C_comparison%20of%20data%20mining%20in%20breast%20cancer.pdf.
[6] Kharya S. Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT);2:55–66 (2012).
[7] Senturk ZK, Kara R. Breast Cancer Diagnosis via Data mining: Performance Analysis Of Seven Different Algorithms. Computer Science & Engineering: An International Journal (CSEIJ); 4:35–46 (2014).
[8] Rajesh K, Anand S. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 1:72–77 (2012).
[9] Gupta S, Kumar D, Sharma A. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering. 2 (2011).
[10] Kumar R, Verma R. Classification Algorithms for Data Mining: A Survey. International Journal of Innovations in Engineering and Technology (IJIET) 1:7–14 (2012).
[11] Kesavaraj G, Sukumaran S. A Study on Classification Techniques in Data Mining. 1 4th ICCCNT (2012).
[12]Soundarya M, Balakrishnan R. Survey on Classification Techniques in Data mining. International Journal of Advanced Research in Computer and Communication Engineering Vol.3:7550–7552 (2014).
[13] Li J, Wong L. Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains; 15th European Conference on Machine Learning (ECML) (2004).
[14] Kumar D, Beniwal S. Genetic Algorithm and Programming Based Classification: A Survey. Journal of Theoretical and Applied Information Technology. 54:48–58 (2013).
[15] Mansuri AM, Verma M, Laxkar P. A Survey of Classifier Designing Using Genetic Programming and Genetic Operators. International Journal of Engineering Research and Reviews (IJERR) Vol. 2:16–22 (2014).
[16] Loh WY. Encyclopedia of Statistics in Quality and Reliability. Ruggeri, Kenett & Faltin, Wiley; Classification and Regression Tree Methods; pp. 315–323 (2008).
[17] Li Y, Zhu J. Analysis of array CGH data for cancer studies using fused quantile regression. Bioinformatics. Vol.23:2470–2476 (2007).
[18]-UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/ (5-15) several classification algo.
[19] Refaeilzadeh P., Tang L., Liu. H. Cross Validation. In Encyclopedia of Database Systems, 532538, Springer, U.S, (2009).
[20]"WEKA Data Mining Book" (n.d.) http://www.cs.waikato.ac.nz/~ml/weka/book.html.
[21] "WEKA 3: Data Mining Software in Java" (n.d.) Retrieved March 2010 from http://www.cs.waikato.ac.nz/ml/weka/.
[22] Kusiak A. Decomposition in Data Mining: An Industrial Case Study in IEEE Transactions On Electronics Packaging Manufacturing, Vol. 23, No. 4, 87-97, (2000).
[23] le Cessie, S., van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201.D. Aha, D. Kibler Instance-based learning algorithms. Machine Learning. 6:37-66 (1991).
[24]Aha D., Kibler D., Instance-based learning algorithms. Machine Learning. 6:37-66 (1991).
[25] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114, (1995).
[26] Walter H. Delashmit and Michael T. Manry, 2005. Recent Developments in Multilayer Perceptron Neural Networks. Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC. 699 (2005).
[27] Leo Breiman, Random Forests, Machine Learning:45 (1):5-32 (2001).
[28] Kohavi R. The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, (1995).
[29] Quinlan R., Induction of decision trees. Machine Learning, vol. 1, 81-106, (1986).
[30] Frank E., Ian H. Witten: Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151, (1998).
[31] Kohavi R., Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision Tree Hybrid. In Proceedings of KDD-96, Portland, USA, 202-207, (1996).