European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 08, Issue 1, 2021

Classification of WBC dataset using supervised learning techniques

I.Jeena Jacob[1], A. Usha Ruby[2], Prasanna Venkatesan[3], VamsidharYendapalli[4], D.Sathya[5]
[1,2,3,4]*Department of CSE, GITAM University, Bengaluru Campus*
[5]*Department of CSE, Kumaraguru College of Technology, Coimbatore*

Abstract
Many women of mid-age suffer with breast cancer. Study says, this disease is the major reason behind the mortality. Since this became a main issue, lot of researches were done in identifying the cause of this. If a woman can predict the possibility of this disease based on her other characteristics, she can be vigilant and can treat early. The identification of malignancy of this disease is the need of the hour. This can be achieved by using classification and prediction algorithms. This paper implements different machine learning techniques for this aim and analyses efficiency of those algorithms. The analysis was made based on Wisconsin breast cancer dataset (WBCD).

Keywords- Wisconsin breast cancer dataset, Support Vector Machine, K-Nearest Neighbor

1. INTRODUCTION

Classification and prediction algorithms help the researchers in taking the decisions in many issues. The classification algorithms are used in many applications like biometrics, security, disease diagnosis, agriculture, communication etc. The classification algorithms are used for different disease diagnosing systems using dependable datasets in the form of record, image, video, audio, signal etc. If we utilize such prediction or classification systems properly, even harmful diseases even can be cured easily by predicting or identifying before it becomes severe. These algorithms work well because they can access datasets easily. This work applies the classification algorithms for predicting breast cancer [1-2]. The modeled system of this work utilizes various conditions of the particular person and predicts the possibility of the breast cancer. If the preliminary decision is there, it will help the physicians to take a prompt decision and also it will help to avoid the errors.

2. RELATED WORK

Classification algorithms work by learning the training dataset. The learning is nothing but extracting the features of training dataset. In the same way, the test dataset is compared by comparing its extracted features with that of training dataset. The property of dependent attribute is dependent on that of independent variables. Classification is done based on these relevant attributes. In this work, our aim is to classify cells as benign or malignant. The algorithm which is being used for classification and prediction is the one which decides the output. Classifications can be done by supervised, unsupervised and semi-supervised learning. When knowledge of data is used along with its class information, then that learning is called as supervised one. When only partial information is available, then that learning is called as semi-supervised and when no information is available, then that learning is unsupervised one.

Support Vector Machine (SVM) [3-4] is implemented using a hyperplane. This plane will have a bias term and normal vector. In Neural Network [5] analysis is based on statistical point. When overlapping of boundaries of classes need to be defined, fuzzy-based classification [6] is used. Multinomial logical regression [7] is another classification technique which can attain and classify many classes. Probabilistic models are used for classifying in Bayesian classifier [8]. Decision tree [9] is another classification technique where partitioning of the dataset is done based on some criteria. This classification is done by applying the conditions and rules. Similarly so many other classification algorithms are also researched by the researchers.

Another classification system is rule-based expert system. Fuzzy sets [6] are the classification systems which is an alternative framework for classification. If-then rules are use for taking decisions in these systems.

Bayesian classifier [24] is based on using label for predicting features from a specified class. Grouping of common values into a particular class is used in this work. Researchers worked on many other classification techniques [10-22] which also used for predicting the variable.

3. PROPOSED WORK

The proposed work used Logistic Regression, K-Nearest Neighbor (KNN) Classifier, Support Vector Machines (SVM), Kernel SVM and Random Forest Classification algorithms for classifying the WBCD data [23]. Figure 1 shows the proposed architecture. Training data is used to create a model where the feature is extracted and based on which the classification is learned. These extracted features are indexed in a proper way to retrieve it for matching. In the similar way, test data are used for testing the results. These data are fed into the same algorithm and the test data are learned. In this dataset, classes will not be available. The aim of this step is to identify the class from the given data. Logistical regression models binary variables using logistic function using regression analysis. KNN is a classifier classifies the object based on the voting of neighbours. Mostly the classification is done, if the number of neighbours is more than k. SVM is another supervised learning technique which can be used for regression and outlier analysis also. This is very effective for spaces with more dimensions. Kernal SVM is used with different kernel functions. Random forest algorithms are constructed by using many decision trees.

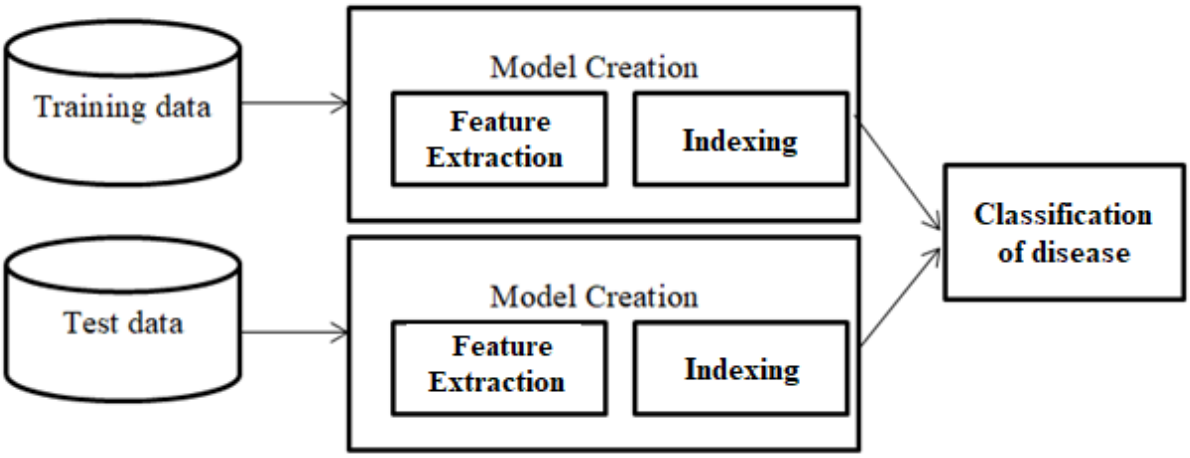

Fig.1: Proposed Architecture

Algorithm:

1. The data is trained by using algorithms. Here we used Logistic Regression, K-Nearest Neighbor (KNN) Classifier, Support Vector Machines (SVM), Kernel SVM and Random Forest Classification algorithms
    a. The learning is done by extracting the relevant features and indexing
    b. Features are indexed in order for easier access
2. The data to be tested is fed into corresponding model created based on the trained algorithm.
3. Both the features of query data is matched with training data features.
4. Test image is categorized into that of training to which the difference is minimum

4. EXPERIMENTAL ANALYSIS

The analysis of the proposed work is done by implementing it inBreast Cancer Wisconsin (Diagnostic) Data Set. This dataset has eleven columns. It is used to classify the cells as either benign or malignant. This classification is aided by ten features of cell nuclei. The characteristics may be compactness, concave points, concavity, symmetry, texture, radius, smoothness, area, perimeter and fractal dimension. The dataset will have total 30 features because it takes mean, standard error and average of largest three numbers. Women with age 40 to 45 have more chance to get this disease.

Different models are used for classification. Ratio between correct prediction with total number is calculated as the classification accuracy. Figure 2 gives the analysis of classification accuracy with percentage of test to train dataset. Figure 3 shows the analysis of classification accuracy with number of iterations. The figures show better result for random forest algorithm. When test to train is 30%, result gives better result for all the algorithms. Also the result gives better result when the iteration is 200.

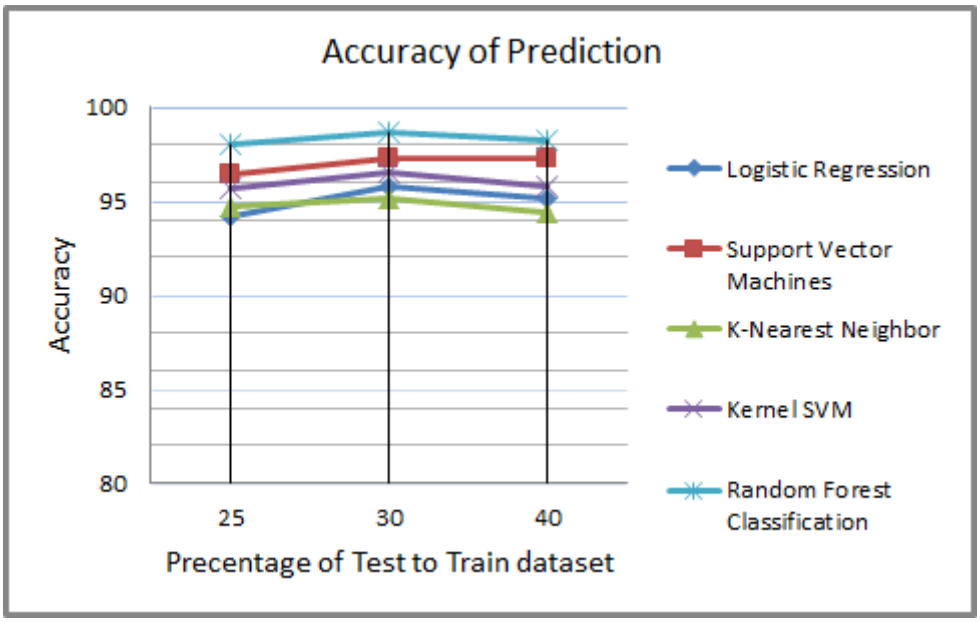

Fig.2: Classification accuracy with percentage of test to train dataset

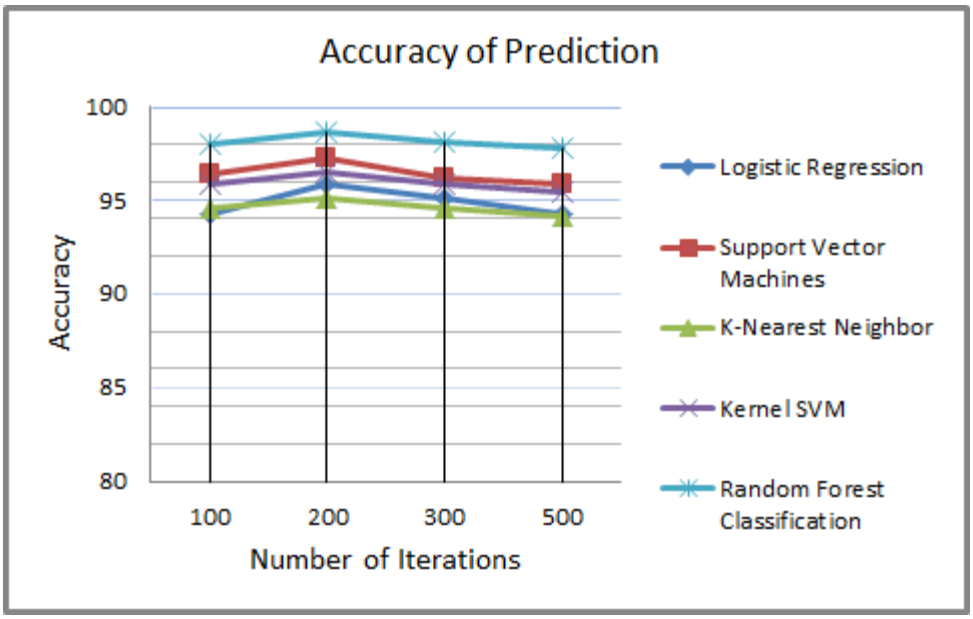

Fig.3: Classification accuracy with number of iterations

The classification is done based on different algorithms. The random forest gives better result than Logistic Regression as +2.8, Support Vector Machines as 1.4,K- Nearest Neighbor as 3.5 and Kernel SVM as 2.1.

5. CONCLUSION

Breast cancer is a very prominent health issue of the women nowadays. According to WHO, this is the main reason for mortality of women. Breast cancer predictionis done based on attributes of the independent variables in the dataset. The breast cancer prediction is working well for many supervised algorithms. In this work, we use Logistic Regression, Support Vector Machines, K- Nearest Neighbor, Kernel SVM and Random Forest classifications. Random Forest classification works better than other algorithms.

References

[1] RasoolFakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013
[2] Cancer Statistics, 2016. CA: A Cancer Journal for Clinicians
[3] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7.
[4] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694- 701.
[5] Bishop, C. Improving the Generalization Properties of Radial Basis Function Neural Networks. Neural Comput. 1991, 3, 579–588
[6] Angelov, P.; Zhou, X. Evolving Fuzzy-Rule-Based Classifiers from Data Streams. IEEE Trans. Fuzzy Syst. 2008, 16, 1462–1475.
[7] Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme Learning Machine for Regression and Multiclass Classification. IEEE Trans. Syst. Man Cybern. 2012, 42, 513–529.
[8] John, G.H.; Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In Proceedings.
[9] Rokach, L.; Maimon, O. Top-Down Induction of Decision Trees Classifiers—A Survey. IEEE Trans. Syst. Man Cybern. 2005, 35, 476–487.
[10] Setnes, M.; Roubos, H. GA-fuzzy modeling and classification: Complexity and performance. IEEE Trans. Fuzzy Syst. 2000, 8, 509–522.
[11] Setnes, M.; Babuška, R. Fuzzy relational classifier trained by fuzzy clustering. IEEE Transa. Syst. Man Cybern.1999, 29, 619–625.
[12] Anderson, T. The Theory and Practice of Online Learning; Athabasca University Press: Athabasca, AB, Canada, 2008.
[13] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
[14] Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine. 2015;2015
[15] Liu L, Deng M. An evolutionary artificial neural network approach for breast cancer diagnosis.InKnowledge Discovery and Data Mining, 2010.WKDD'10. Third International Conference on 2010 Jan 9 (pp. 593-596).
[16] Seyyid Ahmed Medjahed, TamazouztAitSaadi, AbdelkaderBenyettou. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications (0975 - 8887)
[17] Cuong Nguyen, Yong Wang, Ha Nam Nguyen Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomedical Science and Engineering, 2013, 6, 551-560
[18] Diana Dumitru. Prediction of recurrent events in breast cancer using the Naive Bayesian classification.2000 Mathematics Subject Classification.
[19] Turgay Ayer, MS; JagpreetChhatwal, PhD; OguzhanAlagoz, PhD; Charles E. Kahn, Jr, MD, MS; Ryan W. Woods, MD, MPH; Elizabeth S. Burnside, MD, MPH, MS. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. RadioGraphics 2010
[20] JarceThongkam, GuandongXu, Yanchun Zhang and Fuchun Huang. Breast Cancer Survivability via Adaboost Algorithm. HDKM '08 Proceedings of the second Australasian workshop on Health data and knowledge management
[21] Mitra, S.; Hayashi, Y. Neuro-Fuzzy rule generation: Survey in soft computing framework. IEEE Trans. Neural Netw. 2000, 11, 748–768.
[22] Nauck, D.; Kruse, R. Obtaining interpretable fuzzy classification rules from medical data. Artif.Intell. Med. 1999, 16, 149–169.
[23] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
[24]Ben-Gal, I. Bayesian Networks. In Encyclopedia of Statistics in Quality and Reliability; John Wiley & Sons, Ltd.: Malden, MA, USA, 2008.