TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 11390 - 11394

Augmented Breast Cancer Classification and Prediction Using Machine Learning

[1] Apsha Banu S, [2]L Sudha Rani
[1]Assistant Professor of CSE Department, G. Pulla Reddy Engineering College (Autonomous): Kurnool,Andhra Pradesh, India
[2]Post-Graduation student of CSE Department currently pursuing IIIrdsemester,G. Pulla Reddy Engineering College (Autonomous): Kurnool, Andhra Pradesh, India
[1]sudha1021@gmail.com, [2]apshu786@gmail.com

Article Info
Volume 82
Page Number: 11390 - 11394
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 21 February 2020*

Abstract
The mostly occurring cancer in Indian women is breast cancer.Fifty percent of people who are suffering from this disease facing death due to lack of proper treatment. Cancer is a disease which causes due to the change in cells of the body and increase beyond normal growth and control. Breast cancer is one among the mostly occurring of cancer.Test has to be done to find out the reoccurrence of disease in already diagnosed people(Prognosis).It is highly required to raise the survival rate of patient suffering from breast cancer. With the helpof technologyand machine learning the cancer diagnosis and detection accuracy has improved.Rule based classification algorithm plays a vital role in modern breast cancer diagnosis. A classifier is said to be good if it can acquire high accurate classification rules from historical diagnosis. Each diagnosis consists of a large amount of data, it is the aim to form minimal high accurate classification rules from the available past data. Generally, feature reduction techniques help to reduce classification rules. But the challenge is classification performance.

However, if we could able to obtain a technique of feature reduction giving high classification accuracy, it would help obtain minimal high accurate classification rules.

Keywords: *Machine learning, Prognosis, Big Data, Bioinformatics, breast cancer, classification Algorithms.*

I. INTRODUCTION

In the field of medical, daily a huge amount of data has been generated,processing it and acquiring new knowledge from it will improve the healthcare and medical services.

However, it would overcome the price issues of opposing and alleviate diseases.

Machine learning has widely used in the field of computer science due to its effectiveness and accuracy, machine learning is a field where machine can learn by itself from its past experiences known as past data and extract data to carry out the task on upcoming data .machine learning can be done in three ways :supervised, unsupervised and Reinforcement learning.Every unique type of learning has different machine learning techniques .The identity of data determine the type of machine learning technique has to be used to get required information.

As per WHO the most powerful problems in the science research is breast cancer diagnosis. Number of reporting of diseases are growing very high as per today's survey report. Womenare losing their lives due to breast cancer. So, it is necessary to forecast the breast cancer on early stages.

January - February 2020
ISSN: 0193 - 4120 Page No. 11390 - 11394

The aim of this paper is to detect the breast cancer using machine learning classifiers in term of Accuracy, precision and recall.

II. LITERATURE SURVEY

The survey had done on the Breast cancer patients to calculate the frequency of disease spread over women.Breast cancer is verycommon disease in Indian women where enhanced stages at detectionand mortality rates, makes it necessary to understand cancer in women. We conducted a literature survey to evoluate the awareness on breast cancer among women.

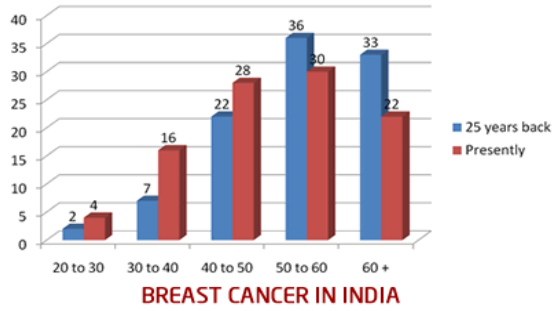

Figure 1.Increasing incidence of BC in the age groups (30's & 40's)

According to fig1, India has the highest BC diagnosed rate in the age of 30's & 40's. in the age of 50's and 60's , the BC affected incidence are low but deaths due to BC's are high, so performance enhancement should be required on early breast cancer detection.Therefore, development in current techniques are needed to forecast breast cancer at early stage.

A comparative three mostly used machine learning techniques where execute on Wisconsin Breast Cancer Dataset to forecast the breast cancer outbreak are.

- Decision Tree
- Support Vector Machine(SVM)
- K-Nearest neighbor(K-NN)

III. METHOD

3.1 Decision Tree

Decision Tree is a Supervised Machine Learning algorithm i.e, you have to teach the system what the input is and what the related output is in the training data where the data is constantly break as claimed by conditions applied. The tree is represented by two features, namely decision nodes and leaves. The leaves are the decisions or the conditions and the decision nodes are where the information breakers. There are two main types of Decision Trees:

- Classification trees (Binary types)
- Regression trees (Continuous valuedata types)

ENTROPY:

Entropy(messy data or unordered data)also denoted by H(S) for a finite set S, it is the measure of the quantity of uncertainty or randomness in data.

Inherently, it explainthe predictability of a certain event.

SI units of entropy is joules per Kelvin (J.k^-1)

In SI base units:kg.m^2.s^-2.k^-1

In particular, lower values gives very less uncertainity and vice versa

INFORMATION GAIN:

Information gain is nothing but the reduction in entropy after dataset got split .it is mostly used to build decision trees. selecting a variable and reduces the dataset for effective classification indicates high information gain with automatic reduction in entropy.

It is calculated by difference in entropy before and after transformation of dataset.

The formula for Infirmation gain is:

January - February 2020
ISSN: 0193 - 4120 Page No. 11390 - 11394

$IG(S,a) = H(S) – H(S \mid a)$

Here IG is information Gain,H(s) is Entropy before dataset got transformed and H(S | a) is the conditional Entropy given the variable a for the reduction of dataset.

3.2 Support Vector Machine

Support Vector Machine isa supervised machine learning which includes both classification and regression algorithms. But mostly preferred for classification .

Support vector Machine separates two different classes by means of hyperplane .

There are many chances for finding different hyper planes but SVM's challenge is to find optimistic one. A hyper plane is said to be optimistic ifand .only if the distance between hyper plane and its support vectors is as far as possible.

Based on the distance between new data to the hyper plane, SVM classifies the data to its related class.

3.3 K-Nearest Neighbor

K-nearest neighbor is a simple supervised learning algorithm which implements both classification and regressionalgorithms .

The letter K in k-nearest neighbor indicates the number of nearest neighbors to the new data .

This algorithm classifies the new data based on majority of neighbor it has among two available classes.This two classes contains cancerous tumors and non-cancerous tumors.The distance between input and nearest neighbors can be found using Euclidian distance and manhattan distance.

KNN is simple algorithm which stoes all available cases and classifies the new data or class based on similarity measures.

Main objective of this experiment is to increase the working production of machine learning techniques in terms of accuracy, precision, recall.

- Precision: precision is the ratio between true positives and all corresponding positives of predicted data.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall: recall is the ratio between true positives and total actual positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- Accuracy: it is the ratio between true positives , true negatives and All Samples

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

IV.CONCLUSION

Breast cancer became very frequent among women. So ,the improvements in existing technologies must be needed to easy prediction of disease.

Machine learning makes system to learn from its past experiences it makes an contribute to types early detection and prediction of cancer.

Machine Learningmodels are still in the testing and experimental phase for cancer prognoses. As datasets are getting larger and of higher quality .it is very necessary to build increasingly accurate models.

REFERENCES

[1]International Agency for Research on Cancer (Iarc) And World Health Organization (Who). Globocan 2018: Age Standardized (World) Incidence and Mortality Rates, Breast. Accessed: Sep. 1, 2018.[Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf

January - February 2020
ISSN: 0193 - 4120 Page No. 11390 - 11394

[2] Y. Lu and J. Han, ‘‘Cancer classification using gene expression data,’’ Inf. Syst., vol. 28, no. 4, pp. 243–268, 2003.

[3] S. Gokhale., "Ultrasound characterization of breast masses", The Indian journal of radiology &imaging, Vol. 19, pp. 242-249, 2009.

[4] N. Bhatia, "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

[5] Biolab.si. (2018). Bioinformatics Laboratory. [Online].Available:http://www.biolab.si/supp/bi-cancer/projections/

[6] Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction.PloS one, 12 (1).

[7] Riedmiller, Martin, and AG MaschinellesLernen. "Multi-Layer Perceptron." Machine Learning Lab Special Lecture, University of Freiburg (2014)

[8] Tsirogiannis, G. L., et al. "Classification of medical data with a robust multi-level combination scheme." Neural Networks, 2004.Proceedings.2004 IEEE International Joint Conference on.Vol. 3.IEEE, (2004).

[9] B.M. Gayathri, Dr. C.P. Sumathi, “Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer” 2016.

[10] T Choudhury, V Kumar, D Nigam,Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm - International Journal of Advanced Research in Computer Science and Software Engineering, 2015

[11] S. Butler and G. Webb, “A case study in feature invention for breast cancer diagnosis using X-Ray scatter images” Lecture Notes in Artificial Intelligence. Vol. 2903, pp. 677-685, 2003.

[12] . S. Belciug, F. Gorunescu, A. B. Salem, and M. Gorunescu., “Clustering-based approach for detecting breast cancer recurrence” In Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA), (2010):533–538.

[13] Chen, Jianguo, et al. "A parallel random forest algorithm for big data in a Spark cloud computing environment." IEEE Transactions on Parallel and Distributed Systems 28.4 (2017): 919-933 .

[14] UCI Breast Cancer Wisconsin (Original) Dataset, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Origina l%29. Last Access: 30.01.2019 .

[15] Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications, 38 (7), pp. 9014-9022.

[16] WHO breast cancer statistics [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breastcancer/en/

[17].] T Choudhury, V Kumar, D Nigam, B Mandal ,Intelligent classification of lung & oral cancer through diverse data mining algorithms, International Conference on Micro-Electronics and Telecommunication Engineering 2016.

[18]. Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction.PloS one, 12 (1).

[19] A. Hong and S. Cho, “Lymphoma cancer classification using genetic programming with SNR features” Lecture Notes on Computer Science. vol. 3003, pp. 78-88, 2004.

[20] J. Abonyi and F. Szeifert “Supervised fuzzy clustering for the identification of fuzzy classifiers” Pattern Recognition Letters. vol. 24, pp. 2195-2207, 2003.

[21]. JF McCarthy, M.K., PE Hoffman, "Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management", Ann N Y AcadSci, Vol.62, pp. 10201259, 2004.

[22]M.H. Asri, H.A Moatassime, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis". ProcediaComputSci, Vol. 83, pp. 1064– 1073, 2016.

January - February 2020
ISSN: 0193 - 4120 Page No. 11390 - 11394

[23] R. B. Ray, M. Kumar, and S. K. Rath, ‘‘Fast in-memory cluster computing of sizeable microarray using spark,’’ in Proc. Int. Conf. Recent Trends Inf. Technol. (ICRTIT), Chennai, India, 2016, pp. 1–6.

[24] W. Wolberg and W. Street, “Computerized breast cancer diagnosis and prognosis from fine-needle aspirates” Arch. Surg. vol. 130, pp. 511-516, 1995.

[25] Z. Zhou, Y.J., Y. Yang, S.F. Chen, "Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles Artificial Intelligence", Medicine Elsevier, Vol. 24, pp. 25-36, 2002.

[26] A. Jemal, R.S., E. Ward, Y. Hao, J. Xu, T. Murray, M.J. Thun, "Cancer statistics", A Cancer Journal for Clinicians, Vol. 58, pp. 71-96, 2008.