International Journal of Future Generation Communication and Networking
Vol. 13, No. 2, 2020 pp.1238-1245

Performance Study Of Classification Algorithms Using The Microarray Breast Cancer Dataset

Ms.M.Pyingkodi [1], Dr.S.Shanthi[2] , Dr.T.M.Saravanan[1], K Thenmozhi[3],

K. Nanthini[1], D.Hemalatha[4], M. Muthukumaran[5], M. Dhivya [6]

1. *Assistant Professor, Department of Computer Applications, Kongu Engineering College, India.*
2. *Associate Professor, Department of Computer Science & Engineerig , Kongu Engineering College, India.*
3. *Assistant Professor, Department of Computer Applications, Selvam College of Technology, India.*
4. *Assistant professor, Department of Computer Technology, Kongu Engineering College, India.*
5. *Assistant Professor , Department of Computer Science & Engineerig, Sunway University, Malaysia*
6. *PG Student, Department of Computer Applicaions, Kongu Engineering College, India.*

Abstract

Breast tumour indicates one of the diseases that build a high digit of passing away every year. It is the mainly widespread sort of all malignancy and the main source of women's deaths universal. Categorization and data taking out process are an successful way to order data in particular in medical pasture, where those technique are broadly used in judgment and study to construct conclusion. A recital assessment between different machine learning algorithms: Support Vector Machine, k Nearest Neighbour, Random Forest, Linear Regression and Logistic Regression on the Wisconsin Breast disease datasets is demeanour. The most important purpose is to weigh up the accuracy in pigeonhole data with high opinion to good organization and usefulness of every one algorithm in terms of accurateness, exactitude, understanding and specificity. Breast sarcoma represents one of the diseases with the intention of cause a high digit of deaths each year. It is the most widespread type of cancer in addition to the chief cause of women's passing away wide-reaching. It has been agreed reasonably in the uncovering of breast malignant cells exactness charge, recall, exactness, understanding, and specificity among classifiers. For the length of the recognition period taxonomy is carry out and the grades were weighing up with the presentation judgment sandwiched between machine learning algorithms and present the unsurpassed effect depending on the data, correspondingly.

Keywords: *Support Vector Machine, k Nearest Neighbour, Random Forest, breast cancer, classification.*

1. INTRODUCTION

Organism the as a rule commonly going on cancer in women, breast cancer have an effect on around 10% of women at a quantity of face in their life. It is the instant important supplier to women's death subsequent to lung cancer. 25% of all cancers in women as well as 12% of all new-fangled personal belongings are grounds by breast malignancy Big Data has distinguished amount in charge due to it person being second-hand in source of industry intellect, industry analytics and statistics removal to get hold of intelligence and product prediction. Subject like health check knowledge increase quickly when convinced come near like figures mining is functional due to superior opportunity of calculation of diseases, dropping tablets costs, on the road to recovery health of patient

ISSN: 2233-7857 IJFGCN

by restore the superiority of healthcare all along through value by reduction people's lives from beginning to end real time judgment.

According to World health organization, Breast cancer is the for the most part common cancer in the midst of women and it is the second hazardous cancer after lung cancer. In 2018, from the study it is rough and ready that total 627,000 women mislaid their life unpaid to breast cancer so as to is 15% of all tumour deaths surrounded by women. In case of any indication, individuals appointment to oncologist. Doctors can without difficulty recognize breast cancer by means of Breast ultrasound, Diagnostic mammogram, Magnetic resonance imaging, Biopsy. Based on these experiment grades, doctor can advise more test or psychoanalysis. Near the beginning finding is extremely critical in breast cancer. If probability of cancer is predict at in the early hours period then survivability probability of unwearied may augment. An interchange way to make out breast tumour is using machine learning algorithms for guess of nonstandard tumours. Thus, the explore is established out for the appropriate verdict and cataloguing of patients hooked on hateful and benevolent assemblage.

2. RELATED WORK

Cancer verdict is one of the most premeditated efforts in the medicinal field. Quite a few researchers have determined in direct to pick up recital and get to obtain suitable results [2]. The judgment of this cancer is a big dilemma in cancer opinion researches. In reproduction bright, machine education is a control which agrees to to the machine to progress from beginning to end a development. Machine learning is broadly used in bioinformatics and above all in breast cancer diagnosis. One of the most standard methods is K-nearest neighbors which is a supervised learning method. By means of the K-NN in cancer diagnosis is widely used among the researcher. The value of the results depends on the value of parameter 'k' and distance. The 'k' represents the number of nearest neighbours. The performance of different distances that can be employ in the K-NN method. Our experiment will be executed on the WBCD database which is Wisconsin Hospital.

Breast cancer is well thought-out to be the second important source of cancer deaths in women nowadays[6,10]. One of the main struggles is to envisage repeated and non-recurrent actions, in all probability more imperative than the original breast cancer judgment. The goal of this paper is to inspect the probable gift of the Naive Bayesian classification line of attack as are liable hold in computer-aided diagnosis of such actions, with the well-known Wisconsin Prognostic Breast Cancer dataset. The grades illustrate that the Naive Bayes classifier supply act equivalent to other contraption learning practice with low computational effort and far above the soil speed Naive Bayes classifier is a probabilistic classifier foundation on the Bayes' theorem, taking into consideration a strong (Naive) sovereignty postulation. Thus, a Naive Bayes classifier believes that all characteristic in competition contribute to the likelihood of a certain conclusion. Attractive into story the environment of the original prospect model, the Naive Bayes classifier can be qualified incredibly resourcefully in supervised learning surroundings, effective much superior in many composite real-world circumstances, in particular in the computer-aided judgment than one power look forward to. Computer mock-up in medical opinion are being urbanized to help general practitioner discriminate stuck between healthy patients and patients with ailment[7]. This representation can aid in flourishing conclusion manufacture by consent to computation of disease possibility on the basis of acknowledged patient description and clinical experiment grades. Two of the most commonly used processor representation in medical risk inference are logistic deterioration and an artificial neural network. A swot up was demeanour to reassess and measures up to these two models, explicate the compensation and drawback of each, and make available criterion designed for representation range. The two models were used for inference of breast cancer peril on the basis of mammographic descriptors and demographic threat feature. Even though they confirmed similar recital the two representations have only one of its kind description potency as well as boundaries that have got to be well thought-out and may confirm harmonizing in causative to enhanced clinical result production.

The use of statistics withdrawal draw near in medical sphere of influence is growing rapidly [8]. This is for the most part because the usefulness of these come within reach of arrangement and guess

ISSN: 2233-7857 IJFGCN

arrangement has enhanced, above all in relative to portion medical practitioners in their decision creation. This type of make enquiries has develop into crucial for ruling ways to recover patient product diminish the charge of medication, and extra move ahead clinical studies. Consequently, records pre-processing RELIEF feature variety, and Modest AdaBoost algorithms, are used to haul out familiarity from the breast cancer endurance record in Thailand. The recital of these algorithms is scrutinize by means of arrangement exactness, understanding and specificity, uncertainty matrix and stratified 10-fold cross-validation process. Computational grades explain that Modest AdaBoost outperforms Real and kind AdaBoosts. The author used data mining approach to find the abnormalities in mammogram images and they classified benign or malignant. The authors first identify the region of interest (ROI) using Intutionistic fuzzy clustering method and extracted statistical features and multi scale surrounding region dependence method features[12,13]. They classified the RoI into benign or malignant using Self Adaptive Resource Allocation Network and conducted the experiment on real and benchmark datasets.

3. MACHINE LEARNING APROACHES

Machine learning is a grassland of reproduction cleverness that uses arithmetic modus operandi to give computer systems the capability to "be trained" from data, devoid of being overtly programmed. Machine learning is division of Data knowledge which incorporates a large set of numerical techniques. Researchers make use of machine learning for tumor forecast and prediction[3,14,15].

I. KNN ALGORITHM

K-Nearest Neighbor is a manage machine education algorithm as the data prearranged to it is labelled. It is a non-parametric process as the sorting of experiment data indicate relies in the lead the adjoining education data face to a certain level than taking into consideration the scope of the dataset. It is in employment in explain both cataloguing and deterioration responsibilities. In taxonomy procedure, it catalogue the substance based on the k adjoining teaching illustration in the quality breathing space

The functioning attitude in the rear KNN is it presumes that alike data indicate lie in equal setting. It reduces the weigh downstairs of construction a model, get a feel for a digit of restriction, or structure in addition postulation. It clutch the idea of closeness based on geometric procedure call as Euclidean detachment, estimate of aloofness between two points in a flat surface.

Understand the two aim in a seaplane are A(x0,y0) and B(x1,y1) then the Euclidean remoteness sandwiched between them is calculated as follow

$$\sqrt{(x0 - x1)^2 + (y0 - y1)^2}$$

An article to be confidential is agreed to the individual class which correspond to the greater digit of its bordering neighbors. If *k* takes the cost as 1, then the information point is confidential into the sorts that surround only one next-door neighbour. Given a new say data point, the space between that points to all the data points in the schooling dataset are work out. Foundation on the distances, the preparation set data indicate through shorter reserve from the test data point are painstaking as the adjacent neighbours of our test data. Finally, the analysis data point is top secret to one of the module of its near neighbor. Thus the logging of the test data point turning point on the classification of its nearest neighbors. Desire the importance of K is the decisive step in the functioning of KNN algorithm. The value of K is not unchanging and it varies for every dataset, depending on the category of the dataset. If the assessment of K is less the firmness of the calculation is a reduced amount of. In the matching behaviour if boost its value the ambiguity is reduced, direct to smoother limitations and

ISSN: 2233-7857 IJFGCN

boost solidity. In KNN, conveying a new figures point to a grouping lock, stock and barrel depends upon K's value. K correspond to the integer of nearby teaching data points during the closeness of a specified test numbers point and then the test statistics point is agreed to the course group contain peak number of next-door neighbours.

Pick a value for the limitation k:
Input : Give a trial of N graphic and their program.
The classed of an taster x is c(x):
Give a original illustration y:
Conclude the k-nearest neighbors of y by manipulative the distances.
Merge classes of these y case in point in one class c
Production: The class of y is c(Y)= c

Class	**Precision**	**Recall**	**F1 Score**	**Support**
0(Low)	1.67	0.50	0.57	4
1(High)	0.56	0.83	0.67	12
Macro avg	0.52	0.49	0.47	24
Weight-edavg	0.50	0.54	0.49	24

Table:1 KNN Classification Report

II. SVM ALGORITHM

Support Vector appliance is a supervised machine culture algorithm which is undertaking well within example acknowledgment struggle and it is worn as a preparation algorithm for learning sorting and deterioration policy from data. SVM is on the whole specifically used what time the digit of character and integer of occurrence are sky-scraping. A double classifier is put together by the SVM algorithm. This twofold classifier is constructing with a excited seaplane everywhere it is a stripe in supplementary than 3-dimensions.The hyper flat surface does the vocation of untying the constituent into one of the two curriculum.

Excited level surface of SVM is manufacture on arithmetic equations. The equation of hyper even is WTX=0 which is comparable to the line equation y= ax + b. Here W and X correspond to vectors everywhere the vector W is until the end of moment in time regular to the hyper aircraft.WTX correspond to the dot item for consumption of vectors. As SVM agreement through the dataset when the integer of description are more so, necessitate to exercise the equation WTX=0 in this holder as a substitute of by means of the procession equation y= ax + b.

If a set of guidance data is specified to the apparatus, each data piece drive be dispensing to one or the former uncompromising variables, a SVM teaching algorithm builds a representation with the principle of plots new data item in the direction of one or the previous sort. In an SVM demonstration, each data piece is be a symbol of as points in an n-dimensional freedom where n is the quantity of description where each attribute is correspond to as the value of a scrupulous synchronize in the n-dimensional space. Sorting is passed out by result a hyper-plane that partitions the two-classes competently. Later, original data item is atlas into the same freedom and its class is predict based on the face of the hyper-plane they revolve up.

ISSN: 2233-7857 IJFGCN

Class	Precision	Recall	F1 Score	Support
0(Low)	1.50	0.75	0.60	4
1(High)	0.56	0.75	0.60	12
Macro avg	0.52	0.54	0.48	24
Weight-edavg	0.53	0.54	0.49	24

Table: 2 SVM Classification Report

III. RANDOM FOREST ALGORITHM

Random forests also known as random decision forests creates a large quantity of trees that accomplish their amount fashioned through collection erudition process for sorting, deterioration. Carrier and attribute unpredictability are the description it bring into play to create those trees. The haphazard reforest has an advantage over the resolution tree which, is so as to it does not greater than in shape the data.
It is in addition use for arrangement and falling off. It activate by put up many conclusion tress.
Concluding assessment is pedestal on the greater part of the trees.

Class	Precision	Recall	F1 Score	Support
0(Low)	1.00	0.75	0.86	4
1(High)	0.59	0.83	0.69	12
Macro avg	0.70	0.61	0.63	24
Weight-edavg	0.63	0.62	0.63	24

Table:3 RandomForest Classification Performance

VI. LINEAR REGRESSION ALGORITHM

Linear Regression is an appliance erudition algorithm stand on administer learning. It achieve a decay duty. Linear degeneration execute the mission to envisage a charge unpredictable value (y)

ISSN: 2233-7857 IJFGCN

foundation on a specified self-regulating patchy (x). By means of Least square process to check its kindness and fit. Mathematically, the affiliation can be symbolize with these equation.

$$y=mx + c$$

The profession is to diminish the detachment between the definite values and envisage value. Demonstration with least blunder will be line at Linear decay or Best fit Line. Exactness and honesty of fit precise by R-Squared value. This cost used to determine how close the data be to the integral waning line.

V. LOGISTIC REGRESSION

Logistic regression is a supervised learning system. It makes use of an equation comparable to Linear Regression but the product of logistic regression is a uncompromising unpredictable while it is a cost for former regression model. Binary conclusion can be expect from the self-determining variables. The conclusion of ward variable is disconnected. Logistic Regression uses a uncomplicated equation which give you an idea about the linear relative between the autonomous variables. These self-determining variables along with their coefficients are amalgamated linearly to form a linear equation that is worn to expect the amount produced.

The equation use by fundamental logistic copy is

—— Ln () = a0+a1*x+a2*x

Class	Precision	Recall	F1 Score	Support
0(Low)	0.50	0.75	0.60	4
1(High)	0.60	0.75	0.67	12
Macro avg	0.59	0.58	0.54	24
Weight-edavg	0.61	0.58	0.55	24

Table:4 Logistic Regression Classification Performance

4. EXPERIMENTAL RESULT

Subsequent to generate analytical model, good organization can be chequered. For this, the representation can be measure up to based on precision and time addicted. It was in actuality hard to desire the algorithm which have elevated concert, larger precision and effectiveness, since all of them finished very close in exactness.

Algorithms	**Accuracy**
KNN	88.73%
SVM(Linear classifier)	100%
SVM(RBF classifier)	98.59%
RANDOM FOREST	91.66%
LOGISTIC REGRESSION	100%

ISSN: 2233-7857 IJFGCN

Table: 5 Performance of Accuracy Vs Algorithms

5. CONCLUSION

Using data mining classification techniques, a research method for predicting the lung cancer disease is developed. These program draws a hidden knowledge from a database of historic breast cancer disease. Here, we can conclude that Support vector machine and Logistic regression gives better accuracy in breast cancer prediction. The method used to predict breast cancer may be further enhanced and extended. We would like to build web-based software in Future Work to evaluate the output of various classifiers where users can simply upload their data set and evaluate the results rapidly.

REFERENCES

1. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR, " Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence", Journal of Health & Medical Informatics, 2013.

2. Seyyid Ahmed Medjahed, TamazouztAitSaadi,"Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules", International Journal of Computer Applications , 62(1):1-5, January 2013.

3. M. Pyingkodi and R.Thangarajan, "Informative Gene Selection for Cancer Classification with Microarray Data Using a Met heuristic Framework", Asian Pacific Journal of Cancer Prevention Vol 19, 2018.

4. AbdelghaniBellaachia, ErhanGuven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", SIAM Conference on Data Mining, 2006

5. Cuong Nguyen, Yong Wang, Ha Nam Nguyen, "Random forestclassifier combined with feature selection for breast cancer diagnosisand prognostic", J. Biomedical Science and Engineering, 2013, 6,551-560

6. Diana Dumitru, "Prediction of recurrent events in breast cancer using the Naive Bayesian classification", Mathematics Subject Classification, 2000.

7. Turgay Ayer, MS JagpreetChhatwal, "Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation", Radio Graphics, 2010.

8. 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

9. 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.

10. Diana Dumitru, "Prediction of recurrent events in breast cancer using the Naive Bayesian classification", Annals of the University of Craiova - Mathematics and Computer Science Series, Volume 36(2),2009,pages 92-96.

11.Wenbin Yue, Zidong WangMachine "Learning with Applications in Breast Cancer Diagnosis and Prognosis", Designs, Vol.2,13, 2018.

12. S.Shanthi and Murali Bhaskaran, "A Novel Approach for Detecting and Classifying Breast Cancer in Mammogram Images", International Journal of Intelligent Information Technologies, Vol.9, Issue 1, pp.21-39,January-March 2013.

13. S.Shanthi and Murali Bhaskaran, "A Novel Approach for classification of abnormalities in digitized mammograms", Sadhana, Vol.39, Issue 1, pp.1141-1150, 2014.

ISSN: 2233-7857 IJFGCN

14.M.Pyingkodi, S. Shanthi, M.Muthukumaran, K.Nanthini, K.Thenmozh, "Hybrid bee colony and weighted ranking firefly optimization for cancer detection from gene regulatory sequences", International Journal of Scientific and Technology Research, 2020.

15. Thenmozhi, K., Karthikeyani Visalakshi, N, "Distributed ICSA clustering approach for large scale protein sequences and cancer diagnosis", Asian Pacific Journal of Cancer Prevention, 2018.

ISSN: 2233-7857 IJFGCN