International Journal of Computer Science Engineering & Technology
APPROVED BY
NATIONAL SCIENCE LIBRARY (NSL)
NATIONAL INSTITUTE OF SCIENCE COMMUNICATION AND INFORMATION RESOURCES (NISCAIR)
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR)- NEW DELHI INDIA .
ISSN :2455-9091

ANALYSIS OF DIFFERENT DATA MINING ALGORITHMS FOR BREAST CANCER

[1] P. Laurajuliet, [2] Dr. P. R. Tamilselvi,
[1] Assistant Professor of Computer Applications, Vellalar College for Women (Autonomous), Erode, India.
[2] Assistant Professor of Computer Science, Government Arts & Science College, Komarapalayam, India.

ABSTRACT- Breast cancer is one of the subsequent driving reasons for cancer demise in ladies. Regardless of the way that cancer is preventable and treatable in essential stages, the huge numbers of patients are determined to have cancer late. Set up strategies for recognizing and diagnosing cancer basically rely upon talented doctors, with the assistance of restorative imaging, to identify certain side effects that generally show up in the later phases of cancer. The target of this paper is to locate the littlest subset of features that can ensure exceedingly precise classification of breast cancer as either benign or threatening. At that point a relative report on various cancer classification approaches viz. Naïve Bayes (NB), Logistic Regression (LR), and DecisionTree (DT) classifiers are directed where the time multifaceted nature of every one of the classifier is likewise estimated. Here, Logistic Regression classifier is closed as the best classifier with the most noteworthy accuracy when contrasted with the other two classifiers.

Keywords- [Breast Cancer, Classification Accuracy, Feature Extraction, Supervised machine learning, benign.]

1. INTRODUCTION

As indicated by World Health Organization (WHO), breast cancer is the top cancer in ladies both in the created and the creating world. Expanded future, urbanization and appropriation of western ways of life trigger the event of breast cancer in the creating world. Most cancer occasions are analyzed in the late periods of the sickness thus, early location so as to improve breast cancer result and survival is exceptionally essential. In this examination, it is proposed to add to the early conclusion of breast cancer. An analysis on breast cancer analyze for the patients is given. For the reason, above all else, information about the patients whose cancers' have just been analyzed is accumulated and they are organized, and afterward whether different patients are in a difficult situation with breast cancer is attempted to be anticipated under front of those information. Expectations of different patients are acknowledged through various calculations and the correctness's of those have been given. Information mining is progressed toward becoming assuming an extraordinary job in figuring applications in the space zone of medication. Accessibility of information mining applications and its methods are appeared in the territories of healthcare organizations, persistent

consideration, the executives, and escalated care framework. Breast cancer is the second most normal and driving reason for death among all ladies accessible in this world. As per distributed statics breast cancer is influencing both created and creating nations. Still today there is not any more powerful approaches to forestall the breast cancer since its motivation is obscure. Regardless of whether there is no immediate reason for these maladies beginning time treatment of this sickness can give full recuperation from this infection. Right now, information digging is utilized for recovering productive report from bigger archives. There are a few information mining applications and systems which are utilized to examine tremendous information, those information mining strategies are Clustering, Classification, Association Rules, Prediction and Neural Networks Decisions Trees. Among these, some classification calculations and much referred to calculation, for example, naïve Bayes, bolster vector machine, fake neural system, decision tree (c5.0) and k closest neighbor calculation are giving most exact outcome in a great deal of research paper.

The term information mining has an alternate importance when various people groups are portraying it. Yet, the real and essential definition is breaking down a huge information so as to anticipate the future occasions. At present information mining is assuming incredible job in health industry to make healthy industry more productive than previously. Breast cancer is one of the most widely recognized cancers among ladies. Breast cancer is one of the real reasons for death in ladies when contrasted with every single other cancer. Cancer is a sort of ailments that makes the cells of the body change its attributes and cause strange development of cells. Most kinds of cancer cells in the long run become a mass called tumor. The event of breast cancer is expanding all around. It is a noteworthy health issue and speaks to a critical stress for some ladies. Early location of breast cancer is basic in lessening life misfortunes. Anyway prior treatment requires the capacity to distinguish breast cancer in beginning times. Early determination requires an exact and dependable finding technique that enables doctors to recognize benign breast tumors from threatening ones. The programmed conclusion of breast cancer is a significant, certifiable restorative issue. Consequently, finding a precise and successful determination technique is significant. Lately machine learning strategies have been broadly utilized in forecast, particularly in medicinal conclusion. Restorative finding is one of serious issue in therapeutic application. The classification of Breast Cancer information can be valuable to anticipate the result of certain infections or find the hereditary conduct of tumors. A noteworthy class of issues in restorative science includes the analysis of infection, in view of different tests performed upon the patient. Thus the utilization of classifier frameworks in restorative conclusion is bit by bit expanding.

2. DATA MINING ALGORITHMS

2.1 DISCRIMINANT ANALYSIS

Discriminant analysis in RapidMiner is connected with ostensible names and numerical traits. It is utilized to figure out which factors segregate between at least two normally happening gatherings, it might have a distinct or a prescient target. .Discriminant analysis is performed in three different ways as linear, quadratic, and regularized in RapidMiner. In linear case, a linear mix of features which best isolates at least two classes of models is attempted to be found. At that point, the resultant mix is utilized as a linear classifier. Linear Discriminant analysis is to some degree like the change analysis and regression analysis with some distinction. In these two strategies, the needy variable is a numerical worth while it is a clear cut an incentive in LDA (Linear Discriminant Analysis). LDA is additionally identified with principle component analysis (PCA) and factor analysis (both search for linear mixes of

factors which best clarify the information), yet PCA and different strategies does not think about the distinction in classes while LDA endeavors to display the contrast between the classes of information. Quadratic Discriminant Analysis (QDA) is firmly identified with linear discriminant analysis (LDA), where it is expected that the estimations are typically dispersed. Not at all like LDA nonetheless, in QDA there is no supposition that the covariance of every one of the classes is indistinguishable. The regularized discriminant analysis (RDA) is a speculation of the LDA and QDA. The two calculations are extraordinary instances of this calculation. On the off chance that the alpha parameter is set to 1, RDA administrator performs LDA. So also if the alpha parameter is set to 0, RDA administrator performs QDA.

2.2 Artificial Neural Networks (Multi-Layer Perceptron)

Multi-Layer Perceptron is a classifier that utilizations back spread to characterize occurrences. This system can be worked by hand, made by a calculation or both. The system can likewise be observed and altered during preparing time. The hubs in this system are on the whole sigmoid (with the exception of when the class is numeric in which case the yield hubs become unthresholded linear units).

2.3 Decision Trees

Rapid Miner produces a Decision Tree for classification of both ostensible and numerical information. A decision tree is a tree-like diagram or model. It is increasingly similar to a rearranged tree since it has its root at the top and it develops downwards. This portrayal of the information has the bit of leeway contrasted and different methodologies of being important and simple to decipher. The objective is to make a classification model that predicts the estimation of an objective quality (frequently called class or mark) in view of a few information traits of the Example Set.

In Rapid Miner a quality with name job is anticipated by the Decision Tree administrator. Every inside hub of tree relates to one of the information traits. The quantity of edges of an ostensible inside hub is equivalent to the quantity of potential estimations of the comparing info characteristic. Active edges of numerical traits are marked with disjoint reaches. Each leaf hub speaks to an estimation of the mark trait given the estimations of the info characteristics spoken to by the way from the root to the leaf.

Decision tree

INPUT: S, where S =set of classified instances
OUTPUT: Decision Tree
Require:S ≠ ∅, num_attributes > 0
1: procedure BUILDTREE
2: repeat
3: maxGain← 0
4: splitA← null
5: e ← Entropy (Attributes)
6: forallAttribuesa in S do
7: gain ←InformationGain(a,e)
8: if gain >maxGainthen
9: maxGain← gain
10: splitA← a
11: end if
12: end for
13: Partitions(S, splitA)
14: until all partions proceed
15: end procedure

Decision Trees are produced by recursive apportioning. Recursive dividing implies more than once part on the estimations of qualities. In each recursion the calculation pursues the accompanying advances:

An credit An is chosen to part on. Settling on a decent selection of ascribes to part on each stage is essential to age of a valuable tree. The characteristic is chosen relying on a determination measure which can be chosen by the basis parameter.

Examples in the Example Set are arranged into subsets, one for each estimation of the trait An if there should arise an occurrence of an ostensible property. If there should arise an occurrence of numerical characteristics,

subsets are shaped for disjoint scopes of trait esteems.

A tree is come back with one edge or branch for every subset. Each branch has a relative subtree or a mark worth delivered by applying a similar calculation recursively.

All in all, the recursion stops when every one of the models or occurrences have a similar name esteem, for example the subset is unadulterated. Or on the other hand recursion may stop if a large portion of the models are of a similar name esteem. This is a speculation of the principal approach; with some mistake edge. Anyway there are other stopping conditions, for example,

There are not exactly a specific number of cases or models in the current subtree. This can be balanced by utilizing the negligible size for split parameter.

No trait arrives at a specific limit. This can be balanced by utilizing the base addition parameter.

The maximal profundity is come to. This can be balanced by utilizing the maximal profundity parameter.

Pruning is a method wherein leaf hubs that don't add to the discriminative intensity of the decision tree are evacuated. This is done to change over an over-explicit or over-fitted tree to a progressively broad structure so as to improve its prescient power on concealed datasets. Pre-pruning is a kind of pruning performed parallel to the tree creation process. Post-pruning, then again, is done after the tree creation procedure is finished.

2.4 Logistic Regression

Logistic regression is a prevalent technique to anticipate a double reaction. It is a unique instance of Generalized Linear models that predicts the likelihood of the result.

Logistic Regression administrator is a Logistic Regression Learner. It depends on the inside Java execution of the myKLR by Stefan Rueping. myKLR is an apparatus for huge scale piece logistic regression dependent on the calculation of Keerthietal (2003) and the code of mySVM. For similarity reasons, the model of myKLR varies marginally from that of Keerthi et al (2003). As myKLR depends on the code of mySVM; the organization of model documents, parameter records and portion definition are indistinguishable. Definite data is given in the following section(Support Vector Machines). This learning strategy can be utilized for both regression and classification and gives a quick calculation and great outcomes for some, learning errands. mySVM works with linear or quadratic and even hilter kilter misfortune capacities.

2.5 Support Vector Machines

A fundamental depiction of the SVM is that the standard SVM takes a lot of information and predicts, for each given info, which of the two potential classes includes the information, making the SVM a non-probabilistic paired linear classifier. Given a lot of preparing models, each set apart as having a place with one of two classifications, a SVM preparing calculation assembles a model that doles out new models into one classification or the other. A SVM model is a portrayal of the models as focuses in space, mapped with the goal that the instances of the different classes are isolated by a reasonable hole that is as wide as could be allowed. New models are then mapped into that equivalent space and anticipated to have a place with a class dependent on which side of the hole they fall on.

All the more officially, a help vector machine builds a hyperplane or set of hyperplanes in a highor boundless dimensional space, which can be utilized for classification, regression, or different undertakings. Instinctively, a great partition is accomplished by the hyperplane that has the biggest separation to the closest preparing information purposes of any class (alleged practical edge), since when all is said in done the bigger the edge the lower the speculation mistake of the classifier. While the first issue might be expressed in a limited dimensional space, it regularly happens that

the sets to separate are not linearly distinct in that space. Therefore, it was suggested that the first limited dimensional space be mapped into an a lot higher-dimensional space, apparently making the detachment simpler in that space. To keep the computational burden sensible, the mapping utilized by the SVM plans are intended to guarantee that dab items might be figured effectively as far as the factors in the first space, by characterizing them as far as a bit capacity K(x,y) chose to suit the issue. The hyperplanes in the higher dimensional space are characterized as the arrangement of focuses whose inward item with a vector in that space is steady. A stage in SVM classification includes distinguishing proof as which are personally associated with the realized classes is called feature determination. Feature determination and SVM classification together have been utilized even, when expectation of obscure examples isn't important. They can be utilized to recognize key sets which are engaged with whatever procedures recognize the classes.

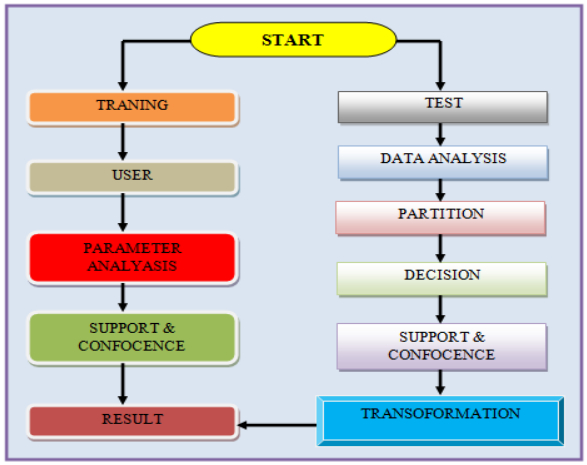

Figure 1: SVM Data Flow

2.6 Naïve Bayes

A Naive Bayes classifier is a straightforward probabilistic classifier dependent on applying Bayes' hypothesis (from Bayesian insights) with solid (innocent) freedom presumptions. An increasingly engaging term for the basic likelihood model would be 'free feature model'. In straightforward terms, a Naive Bayes classifier expect that the nearness (or nonattendance) of a specific feature of a class (for example trait) is inconsequential to the nearness (or nonattendance) of some other feature. For instance, an organic product might be viewed as an apple in the event that it is red, round, and around 4 creeps in breadth. Regardless of whether these features rely upon one another or upon the presence of different features, a Naive Bayes classifier considers these properties to autonomously add to the likelihood that this organic product is an apple.

Naive bayes

Input:

Training dataset T,

F = $(f_1,f_2,f_3\ldots,f_n)$ // value of the predictor variable in testing dataset.

Output:

A class of testing dataset

Step:

Read the training dataset T;

Calculate the mean and standard deviation of the predictor variables in each class;

Repeat

Calculate the probability of f_i using the gauss density equation in each class;

Until the probability of all predictor variables $(f_1,f_2,f_3\ldots,f_n)$ has been calculated.

Calculate the likelihood for each class;

Get the greatest likelihood;

Naïve Bayes Classifier is one of the every now and again utilized techniques for supervised learning. It gives a productive method for taking care of any number of characteristics or classes which is simply founded on probabilistic hypothesis. Bayesian classification gives pragmatic learning calculations and earlier information on watched information.

Let X be an information test : class name is obscure

Let H be a speculation that X has a place with class C

Classification is to decide P(H|X), (i.e., posteriori likelihood): the likelihood that the

speculation holds given the watched information test X

P(H) (earlier likelihood): the underlying likelihood

P(X): likelihood that example information is watched

P(X|H) (probability): the likelihood of watching the example X, given that the speculation holds

training information X, posteriori likelihood of a speculation H, P(H|X), pursues the Baye's hypothesis

$$P(X|H) = \frac{P(X|H)\,P(H)}{P(X)} = P(X|H) \times P(H)/\,P(X)$$

The benefit of the Naive Bayes classifier is that it just requires a modest quantity of preparing information to assess the methods and differences of the factors essential for classification. Since autonomous factors are accepted, just the differences of the factors for each mark should be resolved and not the whole covariance lattice.

2.7 K-Nearest Neighborhood

The k-Nearest Neighbor calculation depends on learning by relationship, that is, by contrasting a given test model and preparing models that are like it. The preparation models are portrayed by n properties. Every model speaks to a point in a n-dimensional space. Along these lines, the majority of the preparation models are put away in a n-dimensional example space. At the point when given an obscure model, a k-closest neighbor calculation scans the example space for the k preparing models that are nearest to the obscure model. These k preparing models are the k "closest neighbors" of the obscure model. "Closeness" is characterized as far as a separation metric, for example, the Euclidean separation.

Algorithm: k-nearest neighbors
Procedure: Find class labels
Input: k, the number of nearest neighbours; D, the set of test sample; T, the set of training sample
Output: L, the label set of test sample

1: read DataFile(Training Data)
2: read DataFile (Testing Data)
3: L ={}
4: **For** each d in D and each t in T **do**
5: Neighbors (d) = {}
6: **If**| Neighbors (d)|<k then
7: Neighbors (d) = Closest (d, t) U Neigbhors (d)
8: **End for**
9: **If** |Neighbors (d)| = k **then**
10: **Break**
11: L = test Class(Neighbors (d)) U L
12: **End for**

Neighbors (d) return the k nearest neighbours of d

Closest (d,t) return the closest elements of t in d

Test Claass (S) return the class label of S.

The k-closest neighbor calculation is among the least difficult of all machine learning calculations: a model is grouped by a lion's share vote of its neighbors, with the model being doled out to the class most normal among its k closest neighbors (k is a positive number, regularly little). On the off chance that k = 1, at that point the model is essentially doled out to the class of its closest neighbor. A similar technique can be utilized for regression, by basically doling out the mark an incentive for the guide to be the normal of the estimations of its k closest neighbors. It very well may be valuable to weight the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more far off ones. The neighbors are taken from a lot of models for which the right classification (or, on account of regression, the estimation of the name) is known. This can be thought of as the preparation set for the calculation, however no express preparing venture is required. The fundamental k-Nearest Neighbor calculation is made out of two stages:

Find the k preparing models that are nearest to the concealed model.

Take the most regularly happening classification for these k models (or, on

account of regression, take the normal of these k mark esteems).

2.8 C4.5

C4.5 was created by Quinlan Ross which is an expansion to ID3. It is chiefly utilized for producing decision tree. The part zone characterized here is gain proportion. C4.5 classification utilizes entropy and data gain for tree part. It is reasonable for dealing with both all out just as consistent information. A limit worth is fixed with the end goal that every one of the qualities over the edge are not thought about. The underlying advance is to ascertain data gain for each trait. The property with the most extreme increase will be favored as the root hub for the decision tree. Given a set S of cases, C4.5 first grows an underlying tree utilizing the gap andconquer calculation as pursues:

If every one of the cases in S have a place with a similar class or S is little, the tree is a leaf marked with the most regular class in S. Otherwise, pick a test dependent on a solitary trait with at least two results. Make this test the base of the tree with one branch for every result of the test, parcel S into comparing subsets S1, S2,... as per the result for each case, and apply a similar strategy recursively to every subset.

KNN algorithm:

Algorithm: Simple k nearest neighbors pseudo code
Procedure: Find class labels
Input: k, the number of nearest neighbours; D, the set of test sample; T, the set of training sample
Output: L, the label set of test sample
1: read DataFile(Training Data)
2: read DataFile (Testing Data)
3: L ={}
4: **For** each d in D and each t in T **do**
5: Neighbors (d) = {}
6: **If**| Neighbors (d)|<k then
7: Neighbors (d) = Closest (d, t) U Neigbhors (d)
8: **End for**
9: **If** |Neighbors (d)| = k **then**
10: **Break**
11: L = test Class(Neighbors (d)) U L
12: **End for**

Notes:
Neighbors (d) return the k nearest neighbours of d
Closest (d,t) return the closest elements of t in d
Test Claass (S) return the class label of S.

Decision tree

INPUT: S, **where** S =set of classified instances
OUTPUT: Decision Tree
Require:S ≠ ∅, num_attributes > 0
1: **procedure** BUILDTREE
2: **repeat**
3: maxGain← 0
4: splitA← null
5: e ← Entropy (Attributes)
6: **forall**Attribuesa in S **do**
7: gain ←InformationGain(a,e)
8: **if** gain >maxGain**then**
9: maxGain← gain
10: splitA← a
11: **end if**
12: **end for**
13: Partitions(S, splitA)
14: **until** all partions proceed
15: **end procedure**

Naive bayes

Input:
Training dataset T,
F = (f_1,f_2,f_3…,f_n) // value of the predictor variable in testing dataset.
Output:
A class of testing dataset
Step:
1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat
Calculate the probability of f_i using the gauss density equation in each class;

Until the probability of all predictor variables (f_1,f_2,f_3…,f_n) has been calculated.

4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

Algorithm 1.1 C4.5(D)
Input: an attribute-valued dataset D
1: Tree = {}
2: **if** D is "pure" OR other stopping criteria met **then**
3: terminate
4: **end if**
5: **for all** attributes a ∈ D **do**
6: Compute information-theoretic criteria if we split on a
7: **end for**
8: a_{best} = Best attribute according to above computed criteria
9: Tree = Create a decision node that test a_{best} in the root
10: D_v = Induced sub-datasets from D based on a_{best}
11: **for all** D_v **do**
12: $Tree_v$ = C4.5(D_v)
13: Attach $Tree_v$ to the corresponding branch of Tree
14: **end for**
15: **return** Tree

CONCLUSION

Contrasting with every other cancer, breast cancer is one of the real reasons for death in ladies. Along these lines, the early identification of breast cancer is required in lessening life misfortunes. In this paper we have connected methods in particular information cleaning, feature choice, feature extraction, information discretization and classification for anticipating breast cancer as precisely as could be allowed. Our examination uncovers that Logistic Regression Classifier gives the most extreme accuracy with decreased subset of features (four) and time multifaceted nature of this calculation is least contrasted with other two classifiers. This work can further be improved by distinguishing proof of specific phase of breast cancer, should be possible in not so distant future.

REFERENCES

[1]. Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, Omkar Tadakhe, "Data Mining Techniques for Diagnosis And Prognosis of Cancer", Int. Journal of Advanced Research in Computer and Communication Engg., Vol. 4, Issue 3, 2015, pp. 613-614.

[2] K.R.Lakshmi, M.Veera Krishna, S.Prem Kumar, "Performance Comparison of Data Mining Techniques for Prediction and Diagnosis of Breast Cancer Disease Survivability", Asian Journal of Computer Science and Information Technology, Vol. 3, 2013, pp. 81 - 87.

[3] Joshi, Miss Jahanvi, and Mr.RinalDoshi, Dr.Jigar Patel. "Diagnosis And Prognosis Breast Cancer Using Classification Rules", Int. Journal of Engineering Research and General Science, 2014, Vol. 2, Issue 6, pp. 315-323.

[4] VasanthaM., &Bharathy, V. S. "Evaluation of Attribute Selection Methods with Tree Based Supervised Classification-A Case study with Mammogram Images", Int. Journal of Computer Applications, Vol. 8, No. 12, 2010, pp. 35.

[5] VikasChaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", Int. Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, 2014, pp. 2464.

[6] Zarei, S., Aminghafari, M., HakimehZali, "Application and comparison of different linear classification methods for breast cancer diagnosis", International Journal of Analytical, Pharmaceutical and Biomedical Sciences, Vol. 4, Issue2, 2015, pp. 123-128.

[7] Y.Ireaneus Anna Rejani, Dr.S.ThamaraiSelvi, "Early Detection Of Breast Cancer Using Svm Classifier Technique", Int. Journal on Computer Science and Engineering, Vol. 1, Issue 3, 2009, pp. 127-130.

[8] Rajesh,k., Dr.SheilaAnand,"Analysis of SEER Dataset for Breast Cancer Diagnosis usingC4.5 Classification Algorithm", IJARCCE,Vol.1, Issue 2, 2012, pp. 2278-1021.
[9] Vanaja, S., K. Rameshkumar, "Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey", Journal of Computer Science, Vol. 11, 2015, pp. 32-33.
[10]Chandrahasan, R. Kalaichelvi, et al. "An Empirical Comparison of Boosting and Bagging Algorithms", International Journal of Computer Science and Information Security, Vol. 9, Issue 11, 2011, pp. 147-152.
[11]Sugimoto, Masahiro, Masumi Takada, and Masakazu Toi, "Comparison of robustness against missing values of alternative decision tree and multiple logistic regressions for predicting clinical data in primary breast cancer",IEEE, 2013, pp. 3054-3057.
[12]Nithya, R., and B. Santhi, "Decision tree classifiers for mass classification", Int. Journal of Signal and Imaging Systems Engineering, Vol. 8, No. 1-2, 2015,pp. 39-45.
[13]J.S.Saleema, N.Bhagawathi, S.Monica ,P.DeepaShenoy, K.R.Venugopal and L.M.Patnaik, Int. Journal on Soft Computing, Artificial Intelligence and Applications, Vol.3, No. 1, 2014, pp. 9-10.
[14]Lavanya, D., and Dr K. Usha Rani. "Analysis of feature selection with classification: Breast cancer datasets." Indian Journal of Computer Science and Engineering, Vol. 2, No. 5, 2011, pp. 756-763.
[15]H.S.Hota, "Diagnosis of Breast Cancer Using Intelligent Techniques", Int. Journal of Emerging Science and Engg.,ISSN: 2319–6378, Vol.1, Issue-3, 2013, pp.48-49.