Breast Cancer Classification using Nature-inspired Algorithm

J Daphy Louis Lovenia, S.Ezhilin Freeda, D.Darling Jemima, Nathaniel Christopher

Abstract-Breast Cancer is one among the dangerous ailments that roots up of deaths among women worldwide. Lots of risk factors have been identified through research though the exact reasons of breast cancer are not yet fully understood The Artificial Immune Recognition System classifier helps to classify the type of breast cancer on Wisconsin dataset which provides accurate prediction of the classes of breast cancer. i.e, Benign and Malignant. This system focuses on supervised classification with the help of clonal selection algorithm, Hierarchical Learning Vector Quantization, Multipass Self Organizing Map. The goal of this system can implement the algorithm to classify the cancer accurately and to compare the error rate, f-measure with previous classification algorithms.

Keywords: Classification, Clonal rate, Affinity Recognition Balls, Hierarchical LVQ, Multipass SOM.

I. INTRODUCTION

The International Agency for Research on Cancer by the World Health Organization released an article that the Cancer trouble got higher to 18.1 million new cases and 9.6 million cancer deaths as in September 2018. Female breast cancer positioned as the fifth leading cause of death (627 000 deaths, 6.6%) [1]. Cancer is starting with an unrestrained increase of cells in any of the tissues or parts of the human body. Cancer may arise in any parts of the body and may spread through the cells to other parts of the body. Later effects of cancer i.e malignant stage would cause much pain than the initial discovery of cancer at the benign stage. There are different types of cancer one such which occurs in women is breast cancer. There could be quite a lot of aspects that could influence people's affinity for cancer. Culture plays a significant part in socioeconomic grade along with its connection with the profession and standard of living factors. A good amount of research in urbanized countries have reacted that breast cancer occurrence varies between natives with the diverse culture [1]. There are several classification methodologies that include Concept descriptions, Classification, Prediction and frequent item set to find the helpful patterns [2].

Breast cancer is an unrestrained increase of breast cells. Malignant tumor are the most vulnerable attacks for death. Signs of carcinoma might embrace an swelling within the breast, amendment in breast muscles form, impling of the skin, fluid coming back from the mamilla, a recently inverted mamilla, or a red or scaly patch of skin. In those with distant unfold of the sickness, there is also bone pain, swollen body fluid nodes, shortness of breath, or yellow skin.

Breast Cancer constitutes a chief community health matter globally with over 1 million new-fangled cases spotted annually, which eventually resulting in over 400,000 annual deaths and about 4.4 million women living with the disease [2]. A tumor can be benign (not treacherous to health) or malignant (has the latent to be risky). This paper includes three different algorithms to accurately classifies whether breast cancer is benign or malignant.

II. LITERATURE SURVRY

Akinsola et al [3] used three selected classification algorithms for evaluation using the WEKA tool. These are the three algorithms C4.5, multi-layer perceptron and Naive Bayes which are tested. Experimental results show that C4.5 proves to be the best algorithm with the highest accuracy. The finest algorithm stand on the breast cancer dataset is C4.5 with an accurateness of 93.98% and the entire amount of time consumed to construct the model is at 0.28 seconds, followed by the Multilayer perceptron with the correctness of 83.8673% and an expected time consumed to build the model is at 12.68 seconds and Bayes network classifier with the exactness of 76.5037% and a full amount of time taken to build the model is at 0.03 seconds.

Meriem Amrane[4], have evaluated diverse classifier algorithms on the Wisconsin Breast Cancer diagnosis dataset. The Breast Cancer Dataset (BCD) that they used is donated to the University of California, Irvine (UCI). The observation of each category is classified based on their similarities. Naïve Bayesian Classifier and K nearest neighbor are the two machine learning classifiers that are used.

C Nalini T.Poovozhi[5] foretold carcinoma mistreatment classification formulas like the k nearest algorithm and also the call tree. Here, for examination the result, the parameters properly classified instances, incorrectly classified instances, time taken, statistic, relative absolute error, and root relative square error ar used. From the experimental results, it's determined that the performance of the choice tree is healthier than the K Nearest formula. Classification techniques ar accustomed predict carcinoma. They compared the choice tree and K nearest neighbor supported the assorted performance factors, classification techniques. From the results, it are often terminated that the choice tree achieves multiplied classifier performance and minimum price than the k nearest neighbor classifier formula. Thereby we have a tendency to conclude that the present algorithms offer ninety

Revised Version Manuscript Received on 30 May, 2018.

S. Selvi, Professor , PSG College of Arts & Science, Coimbatore, Tamil Nadu, India.

R. Sridevi , Professor , PSG College of Arts & Science, Coimbatore, Tamil Nadu, India.

seven.51% of accuracy for the classification of carcinoma prediction.

III. DATASET

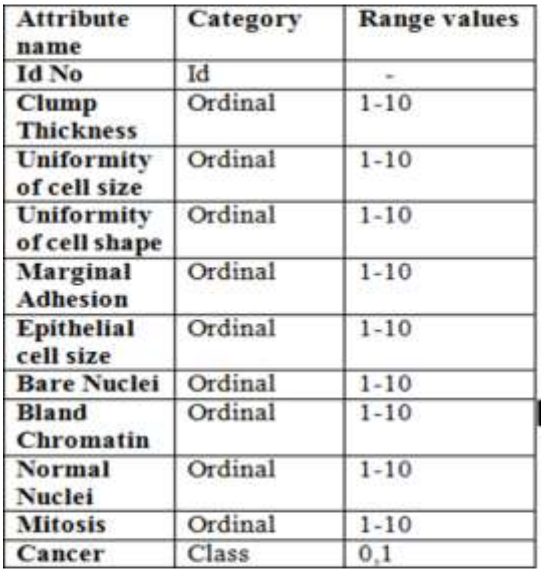
A. Dataset

The Breast Cancer Dataset (BCD) that we have utilized for the experimental use was donated to the University of California, Irvine (UCI). The Winsconsin Breast cancer Dataset has 11 attributes and the first attribute is the ID that can be de-identified because it's a clinical dataset that cannot be disclosed. Nine attributes areas discussed in the breast cancer classification section, these nine attributes can be used to identify whether the cancer is malignant or benign, the last label contains a binary value (2 for benign tumor and 4 for malign tumor). The winsconsin Breast Cancer Dataset consists of 699 clinical records of patients with either malignant or benign tumors. In the preprocessing steps, the missing data in Wisconsin Breast Cancer Dataset had been removed. The entire complete dataset is divided into training and testing subsets used for training and testing respectively.

B. Breast cancer classification (BCC)

The attributes that are considered for classification of breast Cancer includes the following

1. Clump Thickness-It determines the layered structures.
2. Uniformity of Cell Size- It evaluates the sample size and along its consistency.
3. Uniformity of Cell Shape- The shape of the cancer cell varies so this attribute provides the details of the cell shape.
4. Marginal Adhesion- It ensures to provide information about how much adhesion happens with other organs.
5. Single Epithelial Cell Size- It projects the cell size of the epithelial cells.
6. Bare Nuclei- In benign stage, nuclei are not surrounded by the cytoplasm
7. Bland Chromatin- It gives the nucleus structure.



Attribute name	Category	Range values
Id No	Id	-
Clump Thickness	Ordinal	1-10
Uniformity of cell size	Ordinal	1-10
Uniformity of cell shape	Ordinal	1-10
Marginal Adhesion	Ordinal	1-10
Epithelial cell size	Ordinal	1-10
Bare Nuclei	Ordinal	1-10
Bland Chromatin	Ordinal	1-10
Normal Nuclei	Ordinal	1-10
Mitosis	Ordinal	1-10
Cancer	Class	0,1

Table 1 Range values with Category

1. Normal Nucleoli- The nucleolus is usually invisible and very small. In cancer cells, there are more than one nucleolus and it becomes much more prominent.
2. Mitoses-When the number of mitosis increases, there is a high chance of getting the malignant tumor. [6]. Pathologists assigned to each of these characteristics a number from 1 to 10. The likelihood of malignancy needs the nine criteria, even if one of them is very large to classify Breast Cancer

IV. PROPOSED SYSTEM

In this system, nature inspired algorithm or clever algorithm is used to classify breast cancer. i.e., Benign and Malignant.

1. AIRS2 - Artificial Immune Recognition System

One of the supervised classification algorithms is the Artificial immune system (AIS) uses clonal selection, affinity maturation and affinity recognition balls (ARBs). Random number seed is used in for random number generatorMemory cell initialized as 1st is that the pool size that specifies the number of haphazardly designated coaching information instances that is that the seed the memory cell pool. This parameter should be within the vary [0, num coaching instances]. The calculation of affinity threshold to work out whether or not a candidate memory cell is employed by the Affinity threshold scalar (ATS) that might more replace the previous best matching memory cell. That happens as long as the similarity measuring between the applier and therefore the best match cell is <
(AT*ATS).

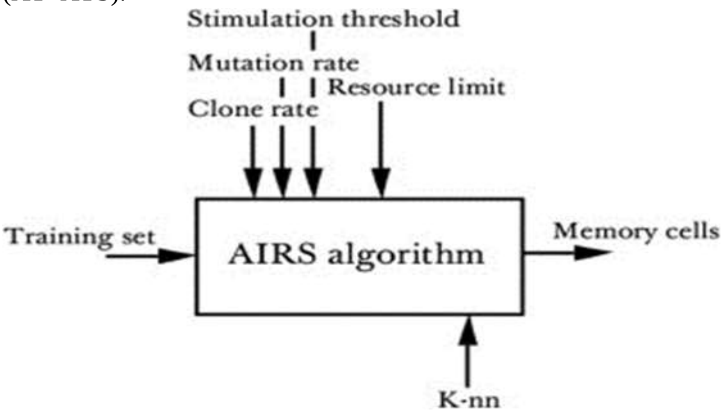


Figure.4.1.1 AIRS Work flow

The determination of the mutated clones to create associate man of affairs at the time of the man of affairs refinement stage that can be calculated as (stimulation * being rate) within the being rate calculation. the amount of best match memory cells used throughout the classification within the k-Nearest Neighbor stage to majority votes the classification of unseen knowledge instances. variety the amount the quantity} of decimal places to be used for the output of numbers within the model also can be fixed Total AL locatable resources specifies the greatest number of resources (B-cells) that may be owed to man of affairs within the ARB pool. Those ARBs with the weakest stimulation area unit aloof from the pool till the entire allotted resources area unit but the most allowable resources. the perfect range of instances to method if a batch prediction is being dead. a lot of or fewer instances is also given, however this provides implementations an opportunity to specify a most well-liked batch size.

The stimulation threshold will be enforced to make a decision once to prevent purification for Associate in Nursing substance within the pool of ARBs. this happens once the mean normalized businessman stimulation worth is larger than or capable the stimulation threshold. It should be set within the vary of [0,1].

Hypermutation rate is employed with the being rate to work out the amount of clones a best matching memory cell will produce to then seed the arbitrager pool with. this can be calculated as (stimulation * being rate * hypermutation rate). the full coaching instances to calculate the affinity threshold (AT) additionally specifies the amount of coaching knowledge instances accustomed calculate the affinity threshold (AT). it's the mean similarity between knowledge instances.[6] a price of -1 indicates to use the complete coaching dataset

1. Hierarchical Learning Vector Quantization

HLVQ is a neural network classifier to classify breast cancer as malignant or benign. It consists of a layered network with two-staged layers. The first layer is useful to prune the categories of the candidate easier. The possible output classes can be classified as malignant or benign in the second layer. Consider A to be the First layer of HLVQ, and let B be the second layer of HLVQ Wi initializes the code block vectors and the learning rate ? for A haphazardly choose an input vector X1. We must calculate the winner unit nearby to the input vector (i.e. the codebook vector Wa with the smallest Euclidean distance concerning the input vector X): i.e. alter the weights of the winner unit: If Wa and X belong to the same class the classification is right.

Wa(t+1) =Wa(t)+?(t)[X(t)?Wa(t)] (1)

If Wa and X belong to different classes (the classification is not considered to be correct)

Wa(t+1)=Wa(t)??(t)[X(t)?Wa(t)]. (2)

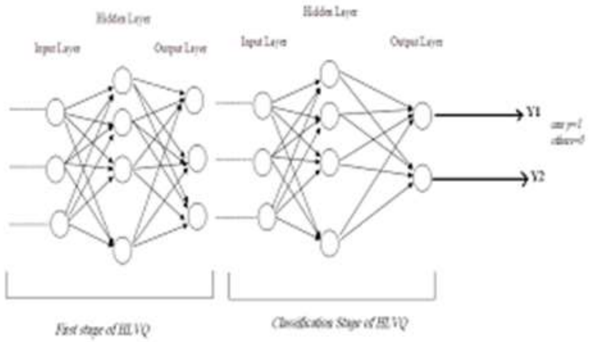


Figure 4.2.1 Flow diagram of HLVQ

Attribute info square measure known to be (M = malignant, B = benign). virtually 10 real-valued options square measure computed for every organelle, radius, texture, perimeter, area, smoothness, compactness, concavity, umbilicate points, symmetry, form dimension.

3) Mutlipass Self Organizing Map (SOP)

Multi-SOMs algorithm is obtainable with several small maps [10]. As the size of the map is small, it is quick to train a SOM model. Thoroughly, we have a huge number of neurons with haphazardly allocated weights. Because the size of neurons is huge enough that we can think that the allotment of the weights on Multi-SOMs is close to the true distribution and its bias is statistically small enough to be determined.

Steps to perform algorithm is as follows as initially generate several Self Organizing Map. Then for each Self Organizing Map and for each input neuron try to check the similarity in the sample X with the weight vector of w_a

On the SOM. Then immediately save the most similar weight vector and immediately update the weight vectors of the neighborhood. When the number of iterations is less than maximum iteration times then classify the input neutron and also calculate the probability for each class k=1, 2…, n. For every neuron, record the neuron which belongs to the class of benign or malignant.

V. PERFORMANCE METRICS & RESULTS

Based on the classifier of Hierarchical LVQ, AIRSV2, Multipass SOM, TP Rate FP Rate Precision Recall F-Measure, ROC Area Class can be measured.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.987	0.008	0.996	0.987	0.991	0.989	Beg
0.992	0.013	0.976	0.992	0.984	0.989	Mal

Table.5.1 F-Measure calculation using Hierarchical LVQ

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.965	0.012	0.993	0.965	0.941	0.997	Beg
0.988	0.035	0.937	0.988	0.941	0.988	Mal

Table.5.2 F-Measure calculation using AIRS V2

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.983	0.041	0.978	0.983	0.943	0.973	Beg
0.959	0.017	0.967	0.959	0.943	0.941	Mal

Table.5.3 F-Measure calculation using Multipass SOM

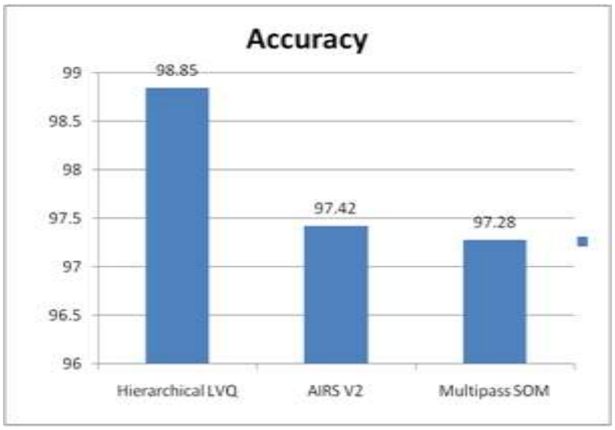


Figure:5.1 Accuracy for the Algorithms

The accuracies for the three algorithms such as hierarchical LVQ are measured high of 98.85% whereas the AIRS provide the accuracy of 97.42% and 97.28 % accuracy is achieved through the multipass SOM algorithm. The figure 5.2 shows the correctly classified instances and incorrectly classified instances of the three different algorithms. of three algorithms. Finally, the performance metrics are very clearly shown that Hierarchical LVQ provides 98.85% of highest accuracy. Thus, this algorithm outrages the other two algorithms.

VI. FUTURE ENHANCEMENT AND CONCLUSION

This system could play a vital role in the prior analysis procedure for different types of cancer and supply a useful defensive approach. In the future, the number of attributes enclosed by the classifier which can be further increased by increasing the sample size of the training set and hence the development model will be more exact. Thus, in this paper, three different algorithms are used to classify breast cancer as benign and malignant. After the evaluation, it is identified as Hierarchical LVQ provides the highest accuracy of 98.85%. Thus, nature-inspired algorithms provide good accuracy to classify breast cancer.

REFERENCES

1. World Health Organization, International Agency for Research on Cancer, Press release no: 263, September 2018.
2. P.Ramachandran, N.Girija, T.Bhuvaneswari Early Detection and Prevention of Cancer using Data Mining Techniques, International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014.
3. Akinsola Adeniyi F, Sokunbi M.A, Okikiola F.M, Onadokun I.O, Data Mining For Breast Cancer Classification, International Journal Of Engineering And Computer Science ISSN:2319-7242, Volume 6 Issue 8 August 2017, Page No. 22250-22258, Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i8.06.
4. Meriem Amrane, Ikram Gagaoua, Breast Cancer Classification Using Machine Learning, 78-1-5386-5135-3/18/$31.00 ©2018 IEEE.
5. C Nalini, T.Poovozhi, DATA-MINING Classification Technique Applied for Breast Cancer, International Journal of Pure and Applied Mathematics Volume 119 No. 12 2018, 10935-10945, ISSN: 1314-3395.
6. M. Sugiyama, "Introduction to Statistical Machine Learning "1ed, ed. T. Green: Morgan Kaufmann,2006.
7. L. Adi Tarca, V.J.C., X. Chen, R. Romero, S. Dr.ghici, "Machine Learning and Its Applications to Biology", PLoS Comput Biol, Vol. 3, pp. 116-122, 2007.
8. Uma Ojha, Dr. Savita Goel, A Study on Prediction of Breast Cancer Recurrence using Data Mining Technique, 978-1-5090-3519-9/17/$31.00_c 2017 IEEE.
9. Meriem Amrane, Tolga Ensarİ, “Breast cancer classification using machine learning”, 2018 Electric Electronics, Computer Science, Biomedical Engineering’s' Meeting (EBBT), April 2018.
10. Shen Lu, Richard S seagull, “Multi-SOM an algorithm for high dimensional, small size dataset, International Journal on Systemics, Cybernetics and Informatics, Vol:11, no:2,2013.