International Journal of Latest Trends in Engineering and Technology
Vol.(11)Issue(3), pp.016-020
DOI: http://dx.doi.org/10.21172/1.113.03
e-ISSN:2278-621X

PATTERN CLASSIFICATION OF BREAST CANCER PATIENTS FOR PERSONALIZED MEDICAL DIAGNOSIS

Sunil Gupta[1], Dr.Dinesh Kumar[2]

Abstract- Computer assisted Pattern Classification has been used extensively in the field of Bioinformatics especially in prediction of diseases and their various stages. Breast Cancer has been a deadliest most disease among women. Among all types of cancer every fourth patient is actually a breast cancer patient. The situation is more severe in Indian women. Timely prediction, detection, progression and reoccurrence prediction can improve the life expectancy. Machine learning techniques have come as a boon for this purpose. This Paper not only presents a review of various facts about breast cancer but pattern classification techniques also with their analysis, bottlenecks and area of improvements.

Keywords – Machine Learning, Breast Cancer, Pattern Classification.

1. INTRODUCTION

The work proposed here deals with evaluation, comparison and improvement of Pattern Classification techniques used in Prognosis, Diagnosis and recurrence of Breast Cancer. The work deals with evaluation of major machine learning techniques used so far in order to detect various stages of Breast Cancer. The proposed study further deals with development of an improved model by reducing complexity and increasing accuracy of the Pattern Classification so that personalized medical treatment can be given to breast cancer patients, which is one of the most prevailing diseases among women. As per world health organization every fifth woman die of breast cancer. Pattern classification is being used for a long time to treat cancer for Prognosis (Prediction), Diagnosis (Detection), Staging, Recurrence Prediction and life expectancy of these patients specially breast cancer patients.

2. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Computer based Pattern Classification is a branch of Machine Learning which is in turn a branch of Artificial Intelligence. Artificial Intelligence has been originated since 1950 and has many definitions given and refined by many researchers. A most widely accepted and a working definition could be the one which was given by Schalkoff in 1990 as “A field of computer science which seeks to explain and emulate intelligent behavior in terms of Computational Process”[1]. An insight in depth of Artificial Intelligence is given by Elaine Rich and Kevin Knight in their book “Artificial Intelligence” which is world’s most widely referred book. There has been stated various task domain of Artificial Intelligence. The first task domain is mundane tasks which involves Perceptions. Second comes Formal tasks which cover those tasks which are formalized or rather rule based systems for example Mathematics. Third domain is Expert Task Domain which includes expert tasks for example Design, Diagnosis, learning, planning, reasoning etc. Leaning has been found to be a key ingredient to implement Artificial Intelligence. Most of the time now a day’s Learning is rather referred as Machine Learning. Most of the research is going on in the field of machine learning only, as it can give us breakthrough in many problems which have been area of experts only so far, for example diagnosis, legal advice, designs, image processing etc. When it comes to the definition of Machine Learning one of the most widely accepted definition is given by Tom M. Mitchell in his book “Machine Learning”. They have defined it as “A computer is said to learn from experience E with respect to some class of tasks T and performance P, if it’s performance at tasks in T, as measured by P, improves with experience E”[2]. Machine Learning basically deals with two types of problems. First Machine Learning Problems are referred as Classification Problems which is more of a supervised type of learning where past instances are available to learn or train the machine before we put the machine to actually perform some prediction related task. Second Machine Learning Problems are referred as Clustering Problems which are more of a Unsupervised type of learning where there is no past data available rather input itself has to be divided into cluster by machine based on their common features and thus arriving at some decision.

3. PATTERN CLASSIFICATION

As stated above Pattern Classification comes under supervised Machine Learning, the efforts are put to give a formal definition to pattern classification also. One of the best definition is given by Richard O.Duda, Peter E.Hart and David G. Stork in their book “Pattern Classification”. They have defined Pattern Classification as “The assignment of a physical object

[1] School of Computer and System Sciences, Jaipur National University, Jaipur, Rajasthan-302004, India
[2] Assistant Professor, Computer Science, Jaipur National University,Jaipur, Rajasthan-302004, India

or event to one of several pre specified categories"[3]. Where a pattern is an object, process or event that can be given a name. A pattern class (or category) is a set of patterns sharing common attributes and usually originating from the same source. During recognition (or classification) given objects are assigned to prescribed classes. A classifier is a machine which performs classification. Over the time Pattern Classification has developed some etymologies. Some of the useful etymologies will be good to know:

3.1 Models and Approaches

Statistical Pattern Classification: it is based on underlying statistical model of patterns and pattern classes for example Bayesian Classification or Bayesian Network(BN).

Structural (or syntactic) Pattern Classification: pattern classes represented by means of formal structures as Decision Tree(DT), grammars, automata, strings, etc.

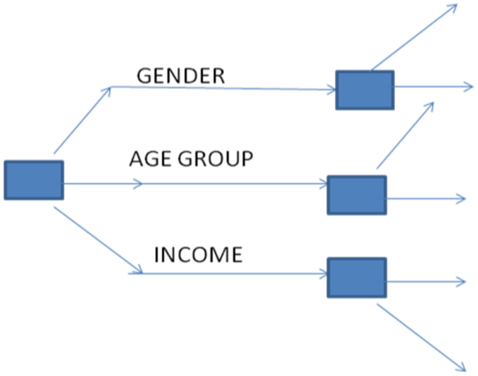

Figure 1 Decision Tree(DT)

Neural networks(ANN): classifier is represented as a network of cells modeling neurons of the human brain (connectionist approach).

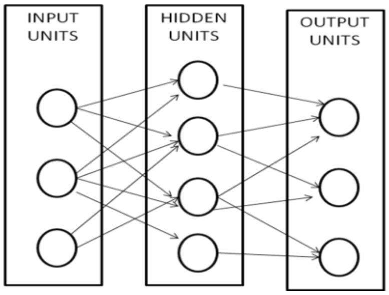

Figure 2 Artificial Neural Network(ANN)

Linear Separators: Classifier is represented as a set of linear equation which has to be optimized to classify. For example Support Vector Machin(SVM).

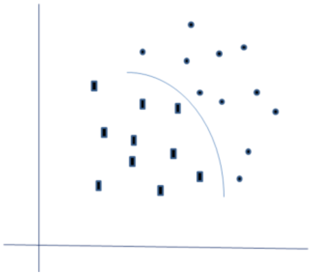

Figure 3 Support Vector Machine (SVM)

Hybrid Models: A combination of above. For example CANFIS(Co-Active Neuro Fuzzy Inference System)

(B)Pattern Classification: Phases and Components

Training Phase: In training phase past data with known classification is given to the machine so that a model can evolve into a decision making model.

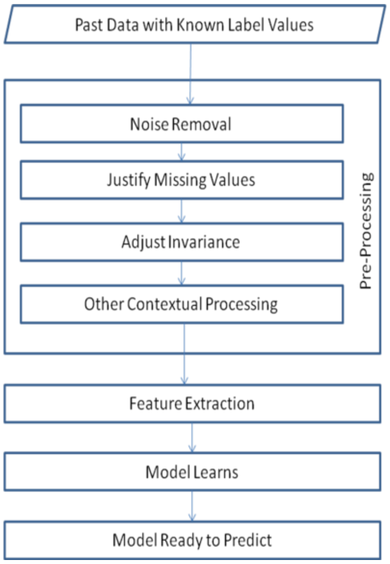

Figure 4 Training Phase

Testing Phase: Once the machine is trained it can be passed with test data with known classes. The classification done by machine is compared with the actual classes of the test data to arrive at accuracy of the machine and its learning power.

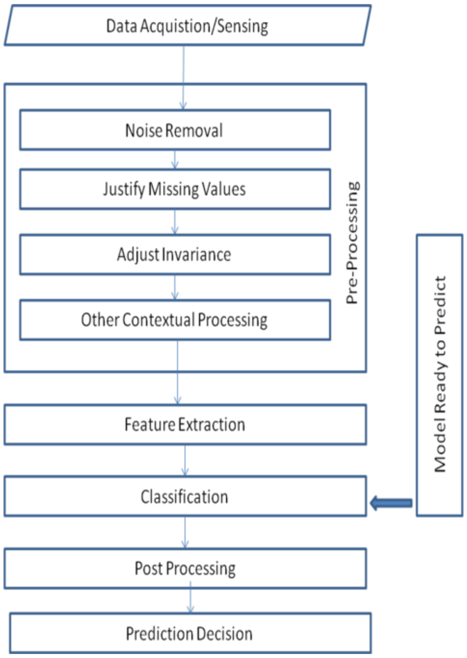

Figure 5 Testing Phase

(C)Following are the sub problems Pattern Classification [3]:
Feature Extraction: The problem deals with deciding the features useful for the purpose of classification. The features selected on one hand must have strong correlation with labels or classes on other hand features themselves must have no correlation with each other. Correlation among features will unnecessarily increase dimensions.
Noise: The unnecessary data records or features will mislead the classification so noise data which has very week correlation must be removed.
Over fitting: Data may be trained on a specific set of instances.
Model Selection: There are various choices of models are available. One has to select a model which gives minimum error rate with less complexity. So one has to try several models
Prior Knowledge: sometimes it is useful to use known facts to reduce the complexity and dimensionality. Training data itself can be called prior knowledge.

Missing Features: It is very possible that features which has strong correlation are not known or given at all. So listing out maximum possible features should be the first effort.
Mereology: Some times a part of the feature may show a strong relation rather than feature as a whole. It can even be vice versa that is two or more feature combined may show a strong coorelation rather then individual. Thus deciding part or whole is a problem to be solved by experiments or some other way.
Segmentation: deals with dividing the problem into subparts.
Context: deals with the objectivity of the classification.
Invariance's: Deals with various forms of same data like orientation, size etc.
Evidence Pooling: deals with the diversity of data on which model has to be trained and tested.
Costs and Risks: deals with cost of classification and risk associated with wrong classification due to error rate.
Computational Complexity: deals with space and time complexity.

4. PATTERN CLASSIFICATION OF BREAST CANCER

4.1 Breast Cancer:

Breast Cancer is one of the most deadly and fastest growing epidemics. Its growth and spread is worldwide as 12% of the worlds female population is suffering from Breast Cancer. Breast Cancer accounts to 25.8% of the total cancer cases. In case of India the situation is more worse as up to 31% of the female cancer patients in India is actually suffering with Breast Cancer as per World Health Organization [4]. Following are the problems related to pattern classification of breast cancer patients:
To find out whether the pattern can be classified as cancer pattern or not in the early stages as in early stages it is tough to recognize a pattern as cancerous pattern.
Accurately finding out the infected areas and amount of infection.
Accurately finding out the nature of infection.
To decide the type of therapy and the dosage in that particular therapy as per the cancer pattern.

4.2 Pattern Classification of Breast Cancer:

Pattern Classification is used extensively to solve various problems in Breast Cancer as given below:
Prediction or Prognosis of Breast Cancer.
Detection or Diagnosis of Breast Cancer.
Severity or staging for the Course of Treatment of Breast Cancer.
Recurrence susceptibility of Breast Cancer.
Life Expectancy of Breast Cancer Patient.
Pattern Classification techniques have improved cancer prediction significantly. A Review by Konstantina Kourou et al presented an in depth study of theses pattern classification techniques recently used in modeling cancer progression. They have reviewed supervised learning techniques most on different features and different data sets. The techniques they have surveyd claim to improve accuracy by 15% to 20%. Though there are complexities involved in all these techniques with some problems. These techniques include ANN, BN, SVM and DTs and applied on mix of set of data like genomic data, biomarkers data, Imaging Data to name a few [5].

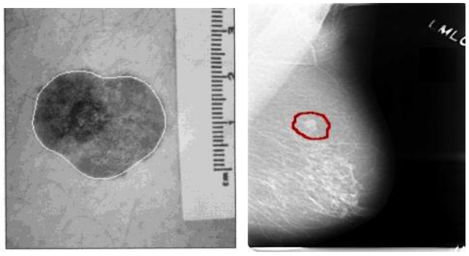

Figure 5 Detecting Breast Cancer (Mamograms)

Tao Lei et.al. have further used ANN for cancer prediction with justification where he extract pieces of input text as justification as prediction without justification has little applicability [6]. Chris McIntosh has gone one more step ahead by dose prediction accuracies of Whole Breast 78.68%, Breast Cavity 64.76%, and Prostate 86.83% [7]. Lamia et.al. have done same job with the help of Breast Cancer Ultra Sound Images sequence to be used in and CAD that is Computer Aided Diagnosis [8]. A. Pardo et. al. Have proposed Directional Kernel Density Estimation for Breast Cancer Classification which has remarkable sensitivity of 98% and specificity of 97% in Breast Conserving Therapy[9]. M.J. Gangeh et.al. also have done remarkable work on noninvasive Computer Aided Theragnosis using ultrasound spectroscopy and maximum mean discrepancy in locally advanced breast cancer [10]. Zhilli Chen et. al. has worked on an important primary sign of breast cancer that is microcalcification. They have worked on microcalcification clusters to reduce complexity and increase accuracy to decide whether the sign is benign or malign[11]. Ang Jun Chin with his co-researchers have worked on the problem of feature selection and dimensionality reduction though their review is based on gene selection but it is applicable to Breast Cancer therognosis also.[12]. Heewon Park and friends have worked on an adaptive penalized logistic regression for uncovering biomarker associated with anti-cancer drug sensitivity.[13].

Thus we can conclude there is a lot which still need to be done for example is

Complexity of DT, ANN, SVM, BN etc. is still a problem and it becomes more complicated due to the curse of dimensionality.

Accuracy in estimating size, area, type of infection still need the support of invasive methods.

Staging, Recurrence susceptibility and life expectancy are almost untouched from the point of view of accuracy.

Although a marvelous work is presented on Machine Learning related algorithms by Jang J.S.R., Sun C.T., Mizutani E. in their book “Neuro-Fuzzy AND Soft Computing”[14]. A deep study is also present in “Pattern Recognition and Image Analysis” by Earl Gose, Richard Johnsonbaugh, Steve Jost[15]. At the same time for preprocessing and many other components “Introduction to Algorithms” by Thomas H. Cormen, Charles E. Leiserson, Ronald L. Reivest, Clifford Stein “ is an unsparable compilation[16]. At last Machine Learning has come as boon to assist in decision making related to Breast Cancer.

5. CONCLUSIOS

Pattern Classification techniques are good candidate techniques to predict, detect, diagnosis and prognosis of Breast Cancer Patients. They have already helped in increasing the accuracy of prediction, prognosis and diagnosis more then 90% in some of the cases. By the time improvements in underlying modals of pattern classification will not only increase the accuracy but reduce the cost of decision making. This all will ultimately help in reducing breast cancer cases in advance stage but help in more accurate classification also of breast cancer patients for personalized medical treatment of these patients. To conclude we need to extend pattern classification techniques to reduce cost and increase accuracy.

6. FUTURE SCOPE

Next effort will be to improve the underlying modals such that the efficiency and accuracy of Pattern Classification techniques can be improved so that Breast Cancer Classification with the help of Machine Learning techniques can be more helpful for a medical expert to treat Breast Cancer Patients.

7. REFERENCES

[1] “Artificial Intelligence” 2ND edition Elaine Rich and Kevin Knight, TMGH, ISBN: 0-07-460081-8

[2] “Machine Learning”, Tom M. Mitchell, Published by McGraw Hill Education, ISBN: 97-800-70428072

[3] “Pattern Classification” 2nd Edition, Richard O.Duda, Peter E.Hart and David G. Stork, Published by John Willey and Sons.

[4] N.V.S.Sree Rathna Lakshmi, 'Detection and Classification of Lesions in Mammograms Using Run length Features', International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.34 (2015).

[5] Konstantina Kourou , Themis P. Exarchos , Konstantinos P. Exarchos et.al., “Machine learning applications in cancer prognosis and prediction”, ELSEVIER Computational and Structural Biotechnology Journal 13 (2015) 8–17

[6] Tao Lei, Regina Barzilay and Tommi Jaakkola, “Rationalizing Neural Predictions”, MIT CSAIL, generaarXiv:1606.04155v1 [cs.CL] 13 Jun 2016.

[7] Chris McIntosh and Thomas G. Purdie, “Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy”, IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 35, NO. 4, APRIL 2016.

[8] Lamia Sellami , O. Ben Sassi, khalil Chtourou, and A. Ben Hamida, “Breast Cancer Ultrasound Images' Sequence Exploration Using BI-RADS Features' Extraction:Towards an Advanced Clinical Aided Tool for Precise Lesion Characterization”, IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 14, NO. 7, OCTOBER 2015.

[9] A. Pardo; E. Real; V. Krishnaswamy; J. M. LÃ³pez-Higuera; B. W. Pogue; O. M. Conde, 'Directional Kernel Density Estimation for Classification of Breast Tissue Spectra', IEEE Transactions on Medical Imaging Vol.36 Issue 1 Page 64 ISSN:0278-0062;02780062 DOI:10.1109/TMI.2016.2593948.

[10] M. J. Gangeh; H. Tadayyon; L. Sannachi; A. Sadeghi-Naini; W. T. Tran; G. J. Czarnota, 'Computer Aided Theragnosis Using Quantitative Ultrasound Spectroscopy and Maximum Mean Discrepancy in Locally Advanced Breast Cancer', IEEE Transactions on Medical Imaging Vol.35 Issue 3 Page 778 ISSN:0278-0062;02780062 DOI:10.1109/TMI.2015.2495246.

[11] Zhili Chen, Harry Strange, Arnau Oliver, Erika R. E. Denton, Caroline Boggis, and Reyer Zwiggelaar, “Topological Modeling and Classification of Mammographic Microcalcification Clusters”, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 62, NO. 4, APRIL 2015 Pg. 1203.

[12] Ang Jun Chin, Andri Mirzal, Habibollah Haron, Senior Member, IEEE, Haza Nuzly Abdull Hamed, “Supervised, Unsupervised and Semi-supervised Feature Selection: A Review on Gene Selection”, IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, DOI 10.1109/TCBB.2015.2478454.

[13] Heewon Park,Yuichi Shiraishi, Seiya Imoto,Satoru Miyano, A novel adaptive penalized logistic regression for uncovering biomarker associated with anti-cancer drug sensitivity, DOI 10.1109/TCBB.2016.2561937, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[14] “Neuro-Fuzzy AND Soft Computing” Jang J.S.R., Sun C.T., Mizutani E., Published By PHI ISBN-978-81-203-2243-1

[15] “Pattern Recognition and Image Analysis”, Earl Gose, Richard Johnsonbaugh, Steve Jost, Published by Prentice Hall of India, ISBN:81-203-1484-0

[16] “Introduction to Algorithms” Thomas H. Cormen, Charles E. Leiserson, Ronald L. Reivest, Clifford Stein, PHI, 3rd Edition ISBN-978-81-203-4007-7.