An Analytical Study of Supervised and Unsupervised Classification Methods for Breast Cancer Diagnosis

Mahua Nandy(Pal)
CSE Department
MCKV Institute of Engineering
243, G.T. Road(N), Liluah,
Howrah, India

ABSTRACT
In this work, ANN and SVM, two most popular supervised machine learning techniques, are considered as the representatives and k-means clustering is used as representative of unsupervised learning. By analyzing the diagnosis result using Wisconsin Breast Cancer Dataset (WBCD) which is commonly used among researchers who use machine learning methods for breast cancer diagnosis, it can be concluded that SVM outperforms in case of breast cancer diagnosis. The result is verified using two other breast cancer datasets. One is Breast Cancer Dataset from UCI Machine Learning Repository and another one is "Breast cancer dataset with Electrical Impedance Measurements in samples of freshly excised tissue". The purpose of the comparison is to choose the best solution in terms of performance. Another notable significance of the work is that accuracy of the recognition drops down severely if proper feature set is not used. One significant disadvantage of neural network is its time taken to build the model which is also evident from the work.

General Terms
Pattern Recognition.

Keywords
ANN, SVM, K-Means clustering

1. INTRODUCTION
We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

In computer science, the imposition of identity on input data, such as speech, an image, or a stream of text, is done, by the recognition and delineation of patterns it contains and their relationships. So, pattern recognition, in general, faces high demand for precision and speed, which is addressed in this paper.

In [1], a comparison of different classification techniques has been done for the task of classifying a speaker's emotional state into one of two classes: aroused and normal. The comparison was conducted using the WEKA (The Waikato Environment for Knowledge Analysis) open source software which consists of a collection of machine learning algorithms for data mining. In paper[2], the objective is to investigate and compare classification performances in case of image retrieval for different methods (Radial Basis Function networks, Support Vector Machines neural networks, Naïve Bayes and Decision Trees). In [3], The classifiers based on linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), kernel fisher discriminant (KFD), support vector machine (SVM), multilayer perceptron (MLP), learning vector quantization (LVQ) neural network, k-nearest neighbor (k-NN), and decision tree (DT), are compared in terms of classification accuracy. In paper [4], breast cancer diagnosis based on a SVM-based method combined with feature selection has been proposed. The paper [5] investigates the use of K- means clustering, an unsupervised classifier which does not depend on the label of the data, for classification. In [6], Given that neural networks have been widely reported in the research community of medical imaging, a focused literature survey has been done on recent neural network developments in computer-aided diagnosis, medical image segmentation and edge detection towards visual content analysis, and medical image registration for its pre-processing and post-processing, with the aims of increasing awareness of how neural networks can be applied to these areas and to provide a foundation for further research and practical development.

In this study, a comparison among three types of classification performances on breast cancer datasets has been conducted using WEKA, a Java based data mining tool so that appropriate classification method for breast cancer diagnosis can be chosen.

The rest of the paper is organized as follows. Section II summarizes the basic concept of supervised and unsupervised learning models. Section III reviews the proposed methodology. Section IV contains the descriptions of three datasets used in this paper. Section V is the implementation details. Section VI describes the experimental results of using the three classification methods on Wisconsin dataset for breast cancer which is a popular dataset among researchers for analyzing experimental results. Section VII shows comparative study of classification methods using another two datasets by analyzing the similarity in nature of classification. Finally, Section VIII concludes the paper.

2. PRELIMINARIES
2.1 General Issues

Supervised learning assumes that a set of training data has been provided; a learning procedure then generates a model. Unsupervised learning, on the other hand, attempts

to find inherent patterns in the data that can then be used to determine the correct output value for new data instances.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.

Artificial Neural Networks (ANN) are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modeled.

A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

2.1 Manhattan Distance

The Manhattan distance function computes the distance that would be travelled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components.

The formula for this distance between two vectors p and q is:

$$d(X,Y) = \sum_{i=1}^{n} |p_i - q_i| \ldots\ldots\ldots (1)$$

3. PROPOSED METHODOLOGY

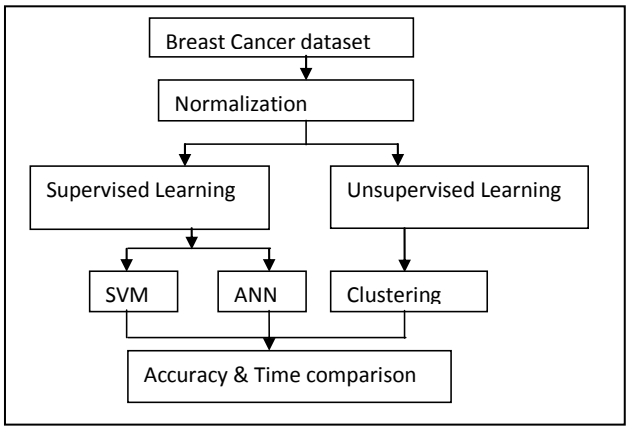

Fig. 1. Proposed methodology

The proposed methodology is explained in the figure 1 and the subsequent steps are mentioned as follows:
At first, data are normalized, because data may have large range values of different orders and magnitude, which may create problems during classification. So, the feature values are normalized with zero mean and unit standard deviation by applying normal transformation to each feature. The normal transformation is defined as

$$f_i = \frac{v_{i-} \mu_i}{\sigma_i} \ldots\ldots\ldots (2)$$

Where vi is the ith feature of each pixel, μi is the average value of the feature and σ i is the standard deviation of each feature.

The normalized feature vector is passed to different classification methods.

Results are computed for three different datasets.

Outcomes are compared to draw the inference and are represented graphically.

4. DATASET DESCRIPTION

For experimental purpose, Wisconsin Breast Cancer Dataset (WBCD) [8], which is commonly used by researchers, is evaluated against machine learning algorithms. Then the results have been established by using another renowned breast cancer dataset from UCI Machine Learning (ML) Repository [9]. Another data set called "Breast cancer dataset with electrical impedance Measurements (EIM) in samples of freshly excised tissue" [7] has also been used to observe the trend, which is similar to the previous. ANN, SVM and K-Means Clustering classification methods have been applied on the data sets. Dataset descriptions are as follows,

4.1 Wisconsin Breast Cancer Dataset

Number of Instances: 699
Number of Attributes: 10 plus the class attribute
Number of classes: 2

Table 1. Class Description

Class	Description	# of Cases
2	Benign	458
4	Fibro-adenoma	241

Table 2. Feature description

Features	Description
AI	Code number
A2	Clump thickness
A3	Uniformity of cell size
A4	Uniformity of cell shape
A5	Marginal adhesion
A6	Single epithelial cell size
A7	Bare nuclei
A8	Bland chromatin
A9	Normal nucleoli
A10	Mitoses

4.2 Breast Cancer Dataset from UCI Machine Learning Repository

Number of Instances: 286
Number of Attributes: 9 plus the class attribute
Number of classes:2

Table 3. Class Description

Class	# of Cases
recurrence-events	85
no-recurrence-events	201

Table 4. Feature description

Features	Description
1	age
2	menopause
3	tumor- size
4	inv-nodes
5	node-caps
6	deg-malig
7	breast
8	breast-quad
9	irradiat

4.3 Breast cancer dataset with electrical impedance measurements in samples of freshly excised tissue

Number of Instances: 106
Number of Attributes: 9 plus the class attribute
Number of classes: 6

Table 5. Class Description

Class	Description	# of Cases
Car	Carcinoma	21
Fad	Fibro-adenoma	15
Mas	Mastopathy	18
Gla	Glandular	16
Con	Connective	14
Adi	Adipose	22

Table 6. Feature description

Features	Description
	Impedivity (ohm) at zero frequency
PA500	High frequency slope of phase angle
HFS	Impedance distance between spectral ends
DA	Impedance distance between spectral ends
AREA	Area under spectrum
A/DA	Area normalized by DA
MAXIP	Maximum of the spectrum
DR	Distance between I0 and real part of the maximum frequency point
P	Length of the spectral curve

5. IMPLEMENTATION

MATLAB Version 7.6 (2008a) has been used to make interface with WEKA which is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute, types, attribute values and the data).

To make third party and user defined Java classes available in MATLAB, they are placed on MATLAB path by adding their locations to the file classpath.txt. Since Weka is a Java library, it can directly be used with the API it exposes to read ARFF files. Here Weka 3.7.8 has been used to evaluate classification results.

6. EXPERIMENTS AND RESULTS

Experiments were first carried out using Wisconsin Breast Cancer Dataset. After that it was verified using two other renowned datasets. WBCD data are represented below vividly.

6.1 Result Analysis of ANN

Experiments were carried out with 10 fold cross validation. Learning rate and momentum has been set to 0.3 and 0.2 respectively. Number of hidden layer has been selected by using the formula:

No. of hidden layers = (attribute + classes)/2.

Different performance measure of classification are tabulated below,

Table 7. Result of ANN

Classification statistics	Value
Correct classification	95.8512%
Incorrect classification	4.1488%
Kappa statistic	0.9086
Mean absolute error	0.0472
Root mean squared error	0.1915
Relative absolute error	10.4335%
Root relative squared error	40.2805%
Coverage of cases(0.95 level)	97.9971%
Mean relative region size(0.95 level)	52.7182%

Table 8. Confusion Matix of ANN

a	b	classified as
441	17	a=2
12	229	b=4

6.2 Result Analysis of SVM

Polynomial kernel function has been used during implementation. Different performance measure of two class SVM classification are tabulated below,

Table 9. Result of SVM

Classification Statistics	Value
Correct classification	96.7096%
Incorrect classification	3.2904%
Kappa statistics	0.9274
Mean absolute error	0.0329
Root mean square error	0.1814
Relative absolute error	7.2802 %
Root relative squared error	38.1642%
Coverage of cases (0.95 level)	96.7096%
Mean rel. region size (0.95 level)	50 %

Table 10. Confusion Matrix of SVM

a	b	classified as
445	13	a=2
10	231	b=4

6.3 Result Analysis of K-Means Clustering

Manhattan distance is used for distance calculation while implementing K-Means clustering. Different performance measure of classification are tabulated below:

Table 11. Result of Clustering

Class	Clustered instances	Percentage
0	220	31%
1	479	69%

Table 12. Confusion Matrix of Clustering

0	1	assigned to cluster
10	448	2
210	31	4

7. COMPARATIVE STUDY

Comparison among three classification methods has been carried out considering different datasets to establish the contribution of this work to the challenge of obtaining the precision and speed for breast cancer detection. The results obtained are shown below.

Table 13. Comparative Study-Accuracy(%)

Methods	Wisconsin	UCI	EIM
SVM	96.71	76.9231	71.69
ANN	95.85	72.028	64.15
K-Means Clustering	94.13	67.77	53.77

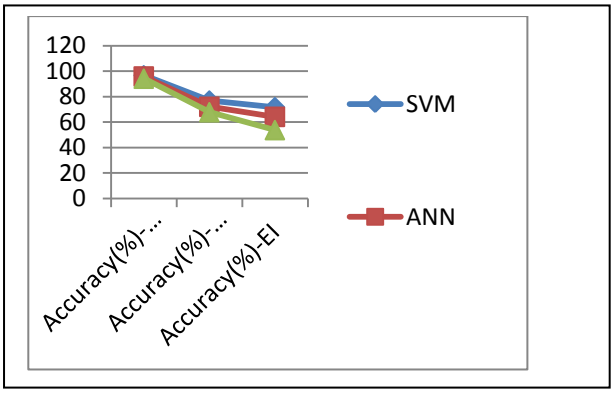

Table 14. Comparative Study-Time(sec)

Methods	Time-Wisconsin	Time-UCI	Time-EIM
SVM	0.21	0.33	0.09
ANN	1.39	2.197	0.42
k-Means Clustering	0.05	0.03	0.03

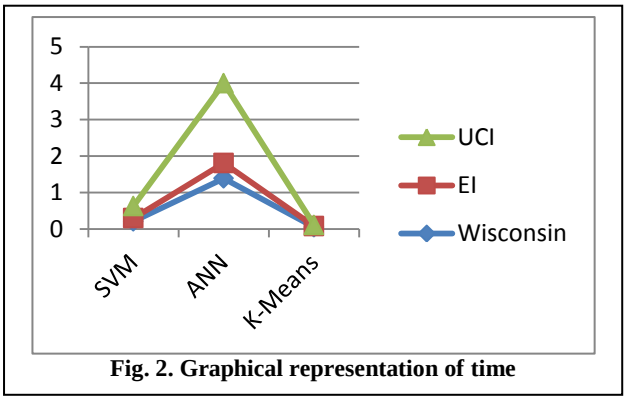

Fig. 2. Graphical representation of time

8. CONCLUSION

In the experiment it is observed that the classification result of SVM is better than that of ANN and clustering performs the least amongst three. The experimental result on two other datasets also supports the prediction, though accuracy falls due to poor feature set selection. Accuracy notably increases with perfect selection of feature set. More effective feature set would lead to a more accurate classification. So, effective feature set selection is also another issue of immense importance for a particular problem domain. This phenomenon is also evident from the work.

One significant disadvantage of neural network is its execution time which is also evident from the work. But the classification accuracy of ANN is comparable with SVM when a proper feature set is taken into consideration.

The main contribution of this paper is to provide a fair and extensive comparison of some commonly employed classification methods under the same conditions so that the assessment of different classifiers can be more convincing. As a result a guideline for choosing appropriate classification method for breast cancer classification task is provided.

9. ACKNOWLEDGEMENT

I am grateful to CSE department of my college for providing the necessary environment to carry out this project work.

10. REFERENCES

[1] T. Justin, R. Gajsek, V.Struc, Dobrisek, " Comparison of different classification methods for emotion recognition" S. MIPRO, 2010 Proceedings of the 33rd International Convention , pp. 700 -770.

[2] I. Mironica, R. Dogaru, "A comparison between various classification methods for image classification stage in CBIR", Interntional Symposium on Signals, Circuits and Systems (ISSCS), 2011 10th, pp. 1-4.

[3] Boyu Wang , Chi Man Wong, Feng Wan, Peng Un Mak, Pui In Mak, Mang I Vai, "Comparison of different classification methods for EEG-based brain computer interfaces: A case study", International Conference on Information and Automation, 2009. ICIA '09, pp. 1416 – 1421.

[4] Mehmet Fatih Akay, "Support vector machines combined with feature selection for breast cancer diagnosis", Expert Systems with Applications: An International Journal, volume 36 Issue 2, March, 2009 pp. 3240-3247.

[5] Lee, N. Gobert, Fujita, Hiroshi, "K-means Clustering for Classifying Unlabelled MRI Data", Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image, 3-5 Dec, 2007, pp. 92-98.

[6] J. Jiang, P. Trundle, J. Ren, "Medical image analysis with artificial neural networks", Journal of Computerized Medical Imaging and Graphics vol. 34, issue 8 (2010) pp. 617–631.

[7]Dataset:http://archive.ics.uci.edu/ml/datasets/breast+tissue

[8] Dataset:http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

[9] Dataset:http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/