Performance Analysis of Breast Cancer Risk Prediction and Diagnosis

M. Karthika, D. Deepika, R. Kanmani, Dr. K. Sangeetha

Department of Computer Science and
Engineering,Kongu Engineering
College,Perundurai,Erode,India

Abstract— Breast cancer is one of the second leading causes of cancer death in women. Despite the fact that cancer is preventable and curable in primary stages, the huge number of patients are diagnosed with cancer very late. Research efforts says that the Support Vector Machine (SVM) have higher accurate diagnosis ability. In this work, experiment was carried out using Wisconsin (Original) Breast Cancer Dataset (WBCD) to classify the breast cancer as either benign or malignant. The performance of this method is evaluated using classification accuracy, sensitivity, specificity, positive and negative predictive values and confusion matrix. Experimental result show that SVM gives the highest accuracy with lowest error rate. Experiment is executed within a simulation environment and conducted in R data mining tool.

Keywords— *Breast cancer, SVM, Classification, Benign, Malignant.*

*****

I.INTRODUCTION

Breast cancer (BC) is that the most typical cancer in women, poignant concerning 10% of all women at some stages of their life. It's one among the foremost unsafe varieties of cancer among women within the world. The world health organization's International Agency for analysis on cancer (IARC) estimates that quite 400,000 women expire annually with carcinoma.[1]

According to the World Health Organization, more than 1.2 million women will be diagnosed with breast cancer each year worldwide. Fortunately, the mortality rate from breast cancer has decreased in recent years with an increased emphasis on diagnostic techniques and more effective treatments. [2]

Breast analysis strategies have been stepped forward over the past decade. Wide variety of automated classification systems has been evolved over last years. Extraordinary strategies have various consequences. But, there nonetheless are problems to be solved: developing new and better techniques. The contrast among specific structures enables us to realize higher system with high performance; this can help radiologists to take accurate outcomes regarding the disease. [1]

Radiologists still produces some variation in reading images. So, there's a requirement for automatic interpretation of pictures or automatic system, and for this purpose classifier is needed. Nowadays several techniques are used for classification however Support Vector Machine shows higher ends up in several instances. This paper offers performance analysis of SVM. [1]

II.SUPPORT VECTOR MACHINE

SVMs are set of supervised learning strategies used for classification and regression [1].They belong to a family of generalized linear classification. A special property of SVM is, SVM at the same time minimize the empirical classification error and maximize the geometric margin. Thus SVM referred to as most Margin Classifiers. SVM relies on the Structural risk Minimization (SRM). SVM map input vector to the next dimensional area wherever a highest separating hyperplane is built. Two parallel hyperplanes are created on all sides of the hyperplane that separate the info. The separating hyperplane is that the hyperplane that maximize the space between the two parallel hyperplanes. Associate degree assumption is formed that the larger the margin or distance between these parallel hyperplanes the higher the generalization error of the classifier are going to be [1]

Two key parts within the implementation of SVM area unit the techniques of mathematical programming and kernel functions. The parameters area unit found by finding a quadratic programming drawback with linear equality and difference constraints; instead of by finding a non-convex, at liberty optimization drawback. The flexibleness of kernel functions permits the SVM to look a good kind of hypothesis areas. [3]

This machine is given with a group of coaching examples, (xi and Yi) wherever the xi area unit the important world information instances and also the yi area unit the labels indicating that category the instance belongs to. For the two category pattern recognition drawback, Yi= +1 or Yi = -1. A coaching example (xi, Yi) is named positive if yi =+1 and negative otherwise. SVM construct a hyper plane that separates two categories and tries to attain most separation between the categories. Separating the categories with an outsized margin minimizes a sure on the expected generalization error. the only model of SVM referred to as top Margin classifier, constructs a linear extractor (an optimum hyper plane) given by (WTXi - γ) = 0 between two categories of the examples. The free parameters are a vector of weights W. that is orthogonal to the hyper plane and a threshold value. These parameters are obtained by determining the subsequent improvement drawback exploitation Lagrangian duality [1]

$$\text{Minimize } 1/2\|W\|^{2} \text{ -------------- (1)}$$

$$\text{Subject to Dii (WTXi } - \gamma) \geq 1,\ i = 1,\ldots\ldots,\ I \text{ ---- (2)}$$

Where Dii corresponds to class labels, which assumes value +1 and –1. The instances with non-null weights are called support vectors. In the presence of outliers and wrongly classified training examples it may be useful to allow some training errors in order to avoid over fitting. A vector of slack variables is that measure the amount of violation of the constraints is introduced and the optimization problem referred to as soft margin is given below [1]

$$\text{Minimize } C\sum I\ 1{=}1\ \mathcal{E}i + ½ \|W\|^2 \text{ ------------ (3)}$$

$$\text{Subject to } Dii\ (WTXi - \gamma) \geq 1,\ i = 1, \ldots\ldots ,I \text{ ---------- (4)}$$

The step-down of the target perform causes most separation between two categories with minimum variety of points crossing their individual bounding planes. The parameter C could be a regularization parameter that controls the trade-off between the 2 terms within the objective perform. the right alternative of C is crucial permanently generalization power of the classifier. the subsequent call rule is employed to properly predict the category of latest instance with a minimum error. The advantage of the twin formulation is that it permits AN economical learning of non–linear SVM separators, by introducing kernel functions. Technically, a kernel perform calculates a real between 2 vectors that are (nonlinearly) mapped into a high dimensional feature house. Since there's no have to be compelled to perform this mapping expressly, the coaching remains [1]

$$Ƒ(x) = \text{sgn } [WT\ X\text{-}Y] \text{ ----------- (5)}$$

Feasible although the dimension of the real feature space can be very high or even infinite. The parameters are obtained by solving the following nonlinear SVM dual formulation (in Matrix form), [1]

$$\text{Minimize } LD\ (U) = 1/2\ uTQu - et\ u \text{ --------- (6)}$$

$$Dtu = 0,\ 0 \leq Ce$$

Where Q=DKD and K is kernel matrix. The kernel function K(AAT) may be polynomial or RBF (Radial Basis Function) is used to construct hyper plane in the feature space, which separates two classes linearly, by performing computations in the input space. The decision function in this nonlinear case is given by [1]

$$Ƒ(x) = \text{sgn } [(K\ (x, xT) * u - y \text{--------- (7)}$$

Where u, the Lagrangian multipliers.

When the number of classes is more than two, then the problem is called multiclass SVM. There are two types of approaches for multiclass SVM the first method is called indirect method, several binary SVM's are constructed and the classifier's output are combined for finding the final class. In the second method called direct method, a single optimization formulation is considered. The formulation of one of the direct methods called Crammer and Singer Method is [1]

$$\text{Minimize } ½ \sum_{k=1}^{N} (W^{T}_{k}) W_{k} + C\sum i^{1} = 1\mathcal{E}i \text{ ------- (8)}$$

Subject to the constraints

$$W^{T}_{wi}\, \emptyset(Xi\ ) - W^{T}_{k} \emptyset(Xi\ ) \geq e^{t}_{k} - \varepsilon i \forall k \neq ki \text{ ----- (9)}$$

where ki is the class to which the training data xi belong,

$$e^{i}_{k} = 1 - C^{i}_{k} \qquad \text{---------} \qquad (10)$$

$$C^{i}_{k} = \{ \begin{smallmatrix} 1\ if\ ki=k \\ 0\ if\ ki\neq k \end{smallmatrix} \qquad \text{--------} \quad (11)$$

The decision function for a new input data xi is given by

$$\widetilde{di} = argmax\{fk\ (xi\ )\} \text{ ---------------- (12)}$$

$$fk\ (xi\ ) = w^{T}_{k} \emptyset(xi\ ) - \gamma k \text{ ----------- (13)}$$

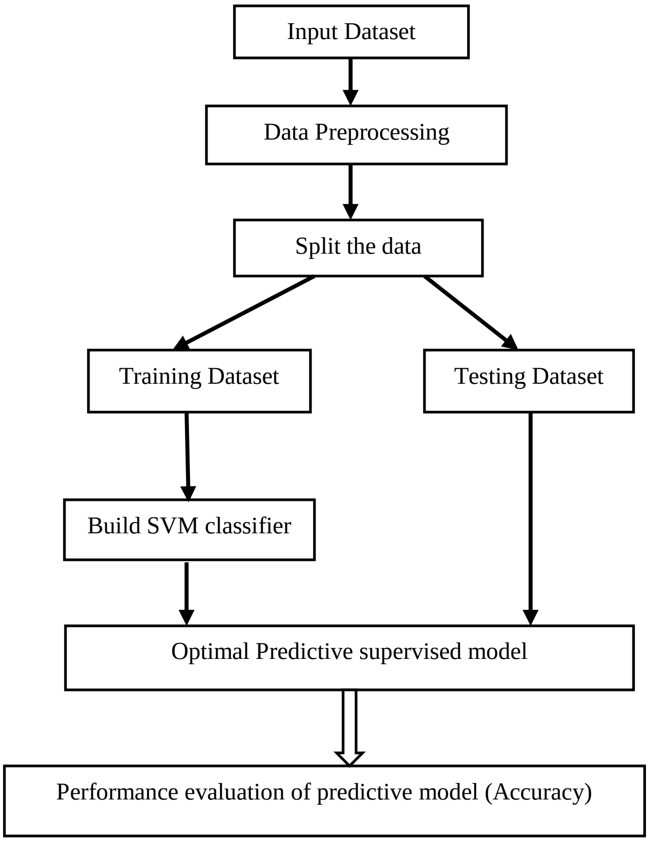

FIG.1:SVM MODEL

III.METHODOLOGY AND EXPERIMENTS

Breast cancer dataset

Wisconsin (Original) Breast Cancer Dataset (WBCD) is used for the experiments . This dataset is taken from the UCI machine learning repository. The dataset contains 699 samples taken from needle aspirates from human breast cancer tissue. It consists of eleven features, each of which is represented as an integer between 1 and 10. The features are; sample code number (F_1), clump thickness (F_2), uniformity of cell size (F_3), uniformity of cell shape (F_4), marginal adhesion (F_5), single epithelial cell size (F_6), bare nuclei (F_7), bland chromatin (F_8), normal nuclei (F_9), mitoses (F_{10}) and class (F_{11}). Four hundred and fifty eight (458) samples of the dataset belong to benign class, and the rest are of malignant class. [2]

IV.MEASURES AND PERFORMANCE EVALUATION

The measures are classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value and confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Table 1 shows the representation of confusion matrix for a two class classifier. Table 2 shows the values of two classes. Classification accuracy, sensitivity, specificity, positive predictive value and negative predictive value can be defined by using the elements of the confusion matrix as [2]

Classification accuracy (%) = TP+ TN / TP+FP+FN+TN

Sensitivity (%) = TP / TP+FN * 100

Specificity (%) = TN / FP+TN *100

Positive predictive value = TP / TP+FP *100

Negative predictive value = TN / FN+TN *100

Table 1. Confusion matrix representation

	Predicted	
Actual	**Positive**	**Negative**
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

Table 2. Confusion matrix

	Benign	**Malignant**	**Class**
SVM	448	7	**Benign**
	10	234	**Malignant**

V.LIMITATIONS OF SVM

- The biggest limitation of SVM lies within the alternative of the kernel (the best choice of kernel for a given problem is still a research problem).
- A second limitation is speed and size (mostly in training - for large training sets, it typically selects a small number of support vectors, thereby minimizing the computational requirements during testing).
- The optimum style for multiclass SVM classifiers is additionally a research space[4]

VI.CONCLUSION

The support vector machine has been introduced as a robust tool for several aspects of data mining including classification, regression and outlier detection. The SVM uses applied math learning theory to look for a regularized hypothesis that matches the available data well without over-fitting. The SVM has only a few free parameters, and these are often optimized victimisation generalization theory while not the necessity for a separate validation set throughout training. Therefore the SVM provides higher accuracy (97%). In the proposed system, RVM is employed to give higher accuracy compared to SVM.

References

[1] Ebrahim Edriss Ebrahim Ali[1], Wu Zhi Feng[2] : Breast Cancer Classification using Support Vector Machine and Neural Network,
[1,2]School of Information Technology and Engineering,Tianjin University of Technology and Education, Dagu Nanlu Road Tianjin, China, 2014

[2] Mehmet Faith Akay : Support vector machines combined with feature selection for breast cancer diagnosis,Department of Electrical and Electronics Engineering,Cukurova University,Adana 01330,Turkey ,2009

[3] Himani Bhavsar, Mahesh H. Panchal: A Review on Support Vector Machine for Data Classification, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 10, December 2012

[4] Robert Burbidge, Bernard Buxton: An Introduction to Support Vector Machines for Data Mining, Computer Science Dept., UCL, Gower Street, WC1E 6BT, UK.

[5] B.M.Gayathri.[1], C.P.Sumathi[2] and T.Santhanam[3] : BREAST CANCER DIAGNOSIS USING MACHINE LEARNING ALGORITHMS –A SURVEY, [1]Department of Computer Science, SDNB Vaishnav College for Women, Chennai, India gayathri_bm2003@yahoo.co.in, [2]Department of Computer Science, SDNB Vaishnav College for Women, Chennai, India drcpsumathi@gmail.com , [3]Department of Computer Application, D.G.Vaishnav College for Men, Arumbakkam, Chennai, India santhanam_dgvc@yahoo.com, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013

[6] Kathija[1], Shajun Nisha[2] M.Phil. (PG Scholar) : Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques , Dept of Computer Science, Sadakathullah Appa College, India[1] Prof& Head, P.G Dept of Computer Science, Sadakathullah Appa College, India[2] , International Journal of Innovative Research in Computer and Communication Engineering , Vol. 4, Issue 12, December 2016

[7] Haowen You[1] and George Rumbe[2] : Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data, [1]Department of Systems and Information Engineering, University of Virginia, Charlottesville, Virginia ,[2]Department of Systems Science and Industrial Engineering, Binghamton

University, Binghamton, New York, International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 1, Nº 3.

[8] Animesh Hazra, Computer Science & Engineering Department, Jalpaiguri Govt. Engg. College , Jalpaiguri, West Bengal ,India, Subrata Kumar Mandal, Information Technology Department , Jalpaiguri Govt. Engg. College , Jalpaiguri, West Bengal ,India, Amit Gupta, Information Technology Department, Jalpaiguri Govt. Engg. College, Jalpaiguri, West Bengal, India: Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms , International Journal of Computer Applications (0975 – 8887) Volume 145 – No.2, July 2016

[9] David Meyer FH Technikum Wien: Support Vector Machines, The Interface to libsvm in package e1071, Austria, February 1, 2017

[10] P.Dhivyapriya [1], Dr.S.Sivakumar [2] : Classification of Cancer Dataset in Data Mining Algorithms Using R Tool , Research cholar [1], Assistant professor [2] ,Department of Computer Science [1] ,Department of Computer Applications [2] ,Thanthai Hans Roever College, Perambalur,Tamil Nadu–India, International Journal of Computer Science Trends and Technology (IJCST)-Volume 5 Issue 1, Jan–Feb 2017

[11] Jung-Ying Wang: Data Mining Analysis (breast-cancer data), Register number: D9115007, May, 2003