ISSN:0975-9646

An SVM Based Approach to Breast Cancer Classification using RBF and Polynomial Kernel Functions with Varying Arguments

S.V.G.Reddy#1, K.Thammi Reddy*2, V. Valli Kumari#3 , Kamadi VSRP Varma#4

1,4Associate Professor, Department of CSE, GIT, GITAM University, Visakhapatnam
2Professor, Department of CSE, GIT, GITAM University, Visakhapatnam
3Professor, Department of CS & SE, College of Engineering, Andhra University, Visakhapatnam

Abstract— **Breast Cancer is the most dreadful disease in women which is leading to death. The medical data classification is acquiring lot of importance before the diagnosis of the disease. Few authors have worked in the field of Breast Cancer classification using standard SVM techniques. In this proposed work, the Breast cancer classification is done using RBF and polynomial Kernel functions of Support Vector Machines with different values of RBF_Sigma, Box Constraint and polyorder arguments which lead to high classification accuracy compared to the previous Results.**

Keywords— **Support vector machine, kernel function, Radial basis function, Polynomial, RBF_Sigma, BoxConstraint, Polyorder**

I. Introduction

SVM Classifier with RBF Kernel Function

Support Vector Machines are used for classification in machine learning which are supervised learning models, that associated with learning algorithms which is used to analyze data and recognize patterns. SVM training algorithm, given a set of training examples, each marked as belonging to one of the two categories, it builds a model that assigns new examples into one category or the other. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called as the kernel trick, implicitly mapping their inputs into high dimensional feature spaces. The (Gaussian) Radial basis function kernel or RBF kernel, is a popular kernel function used in support vector machine classification.

If the data of various classes can be separated [1] as in Fig.1, then the linear SVM is used. Otherwise if the data of the classes cannot be separated [1] , for example noise as in Fig.2, then the non linear SVM classifier is used. We use few mathematical expressions to generate the Hyper plane for the linear or non linear SVMs. For linear SVM as in Fig.1, If the mathematical expression is greater than zero, then the data is said to be located above the Hyper plane and the class is referred as “YES” and on the other hand if the mathematical expression is less than zero, then the data is said to be located below the Hyper plane and the class is referred as “NO”. For Non linear SVM as in Fig.2, If the mathematical expression is greater than zero, then the data is said to be located outside the Hyper plane and the class is referred as “YES” and on the other hand if the mathematical expression is less than zero, then the data is said to be located inside the Hyper plane and the class is referred as “NO”. For Non linear SVM, there are several inbuilt kernel functions such as RBF and Polynomial etc. with various arguments in Matlab to train and classify the data.

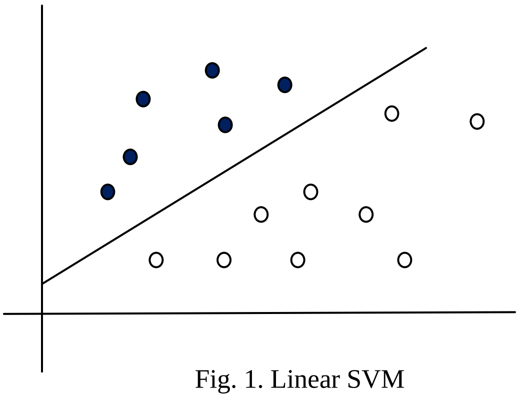

Fig. 1. Linear SVM

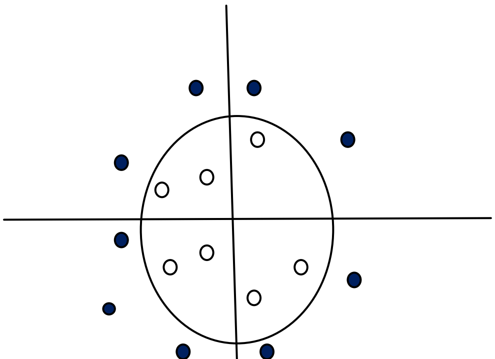

Fig. 2. Non Linear SVM

II. SVM-RBF WITH VARYING ARGUMENTS (PROPOSED MODEL)

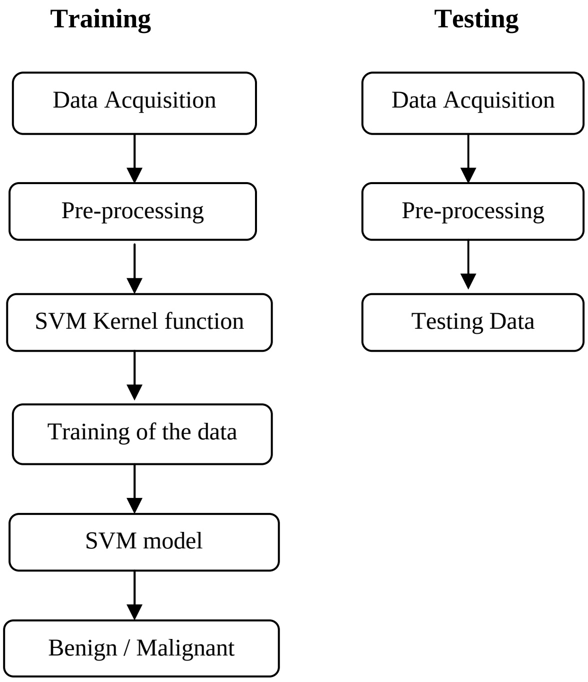

Fig. 3. SVM-RBF with Varying Arguments

The Wisconsin Breast cancer dataset [2] is taken from UCI machine learning repository and used for the training & testing. The dataset is comprised of 10 attributes, 699 records containing two classes such as Benign and Malignant. The 3-fold Cross validation is performed on the data set where 2 folds are used for the training and 1 fold for the testing. The Pre-processing is performed on the dataset. Then SVM kernel functions [3] such as RBF and polynomial (non linear) are applied to generate the hyper plane. And the two folds of data is used for the training purpose and the SVM model is built. Then, one fold of data is acquired, pre-processed and used for the testing purpose which is given as input to the SVM model. Then the SVM model tests the data and classifies as Benign or Malignant.

III. EXPERIMENTAL ANALYSIS

The following functions and the arguments [4] of Matlab are used to train and test the data.

svmtrain( ) - used to train the data set where SMO (Sequential minimal Optimization) method and the kernel function RBF is used.

SMO - Sequential minimal optimization is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines.

svmclassify( ) is the function used to classify the test data sample.

One of the kernel function used is **RBF** and the arguments such as RBF_Sigma, BoxConstraint are considered and their values are

RBFSigmaValue - Positive number that specifies the scaling factor, sigma, in the radial basis function kernel.

BoxConstraintValue - Box constraints for the soft margin. Choices are:

- Strictly positive numeric scalar.
- Array of strictly positive values with the number of elements equal to the number of rows in the Training matrix.

If BoxConstraintValue is a scalar, it is automatically rescaled by N/(2*N1) for the data points of group one and by N/(2*N2) for the data points of group two. N1 is the number of elements in group one, N2 is the number of elements in group two, and N = N1 + N2. This rescaling is done to take into account unbalanced groups that is cases where N1 and N2 have very different values.

If BoxConstraintValue is an array, then each array element is taken as a box constraint for the data point with the same index.

And the other kernel function used is **Polynomial** and the arguments such as Polyorder, BoxConstraint are considered and their values are

PolyorderValue - Positive number that specifies the order of a polynomial kernel and BoxConstraint is mentioned above.

The program is executed in Matlab to compute the sensitivity, specificity, classification accuracy and the respective confusion matrices are generated.

Sensitivity (Se) = TP / (TP + FN)
Specificity (Sp) = TN / (TN + FP)
Accuracy = TP + TN

Where TP – True Positive, FN – False Negative
TN – True Negative, FP – False Positive

Table 1 – Classification Accuracy for Three Folds of Data

Dataset	Sensitivity	Specificity	Accuracy
Fold1	98.7	100	99.1
Fold2	95.4	96.3	95.7
Fold3	95.4	98.8	96.6
Average Accuracy			**97.13**

Here, the arguments of RBF kernel function such as RBF_Sigma and BoxConstraint were randomly assigned with different values and tested, which lead to 97.13% classification accuracy. The training and testing is done on all the three folds of data, where accuracy achieved is 99.1 % on the first fold , 95.7% accuracy on the second fold , 96.6% accuracy on the third fold and the average accuracy is 97.13%.

The accuracy values are presented in Table 1 and Fig.4 which is promising when compared with the earlier classification Results[5] shown in Table 2.

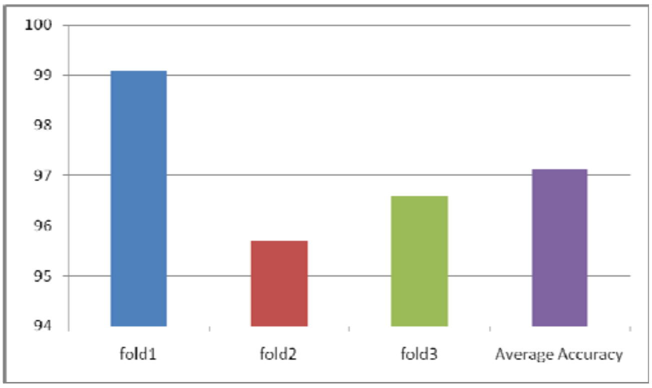

Fig . 4. Classification Accuracy Graph

Table.2 - Classification Accuracy Results

Reference	Classifier	Accuracy
[5]	SVM – RBF	96.84
Proposed Model	SVM-RBF with varying arguments	97.13

The accuracy of the classification is depending upon kernel function and also on its arguments. The maximum classification accuracy is achieved by using both the RBF and Polynomial kernel functions. When the RBF kernel function is used, the arguments considered are RBF_Sigma and BoxConstraint. When RBF_Sigma is 1 and BoxConstraint is 1.2 the maximum classification Accuracy is achieved.

For the kernel function RBF, the effect of the arguments RBF_Sigma and BoxConstraint for the classification is demonstrated below [6]. When the BoxConstraint is 1.2, the testing is done for varying values of RBF_Sigma i.e. 0.4, 0.6, 0.8, 1, 1.2, 1.4 where classification Accuracy 97, 97.9, 98.3, 99.1, 98.3, 98.3 (in percentage) is achieved respectively. So, when the value of RBF_Sigma is 0.4, the classification value of 97% is observed and gently increased for the increase of RBF_Sigma value and the maximum classification accuracy is achieved at RBF_Sigma value of 1 and decreased when the RBF_Sigma is 1.2 and so on which is showed in Table.3 and Fig.5.

Table.3 - RBF_Sigma Vs Classification Accuracy

RBF_Sigma	Classification Accuracy
0.4	97
0.6	97.9
0.8	98.3
1	99.1
1.2	98.3
1.4	98.3

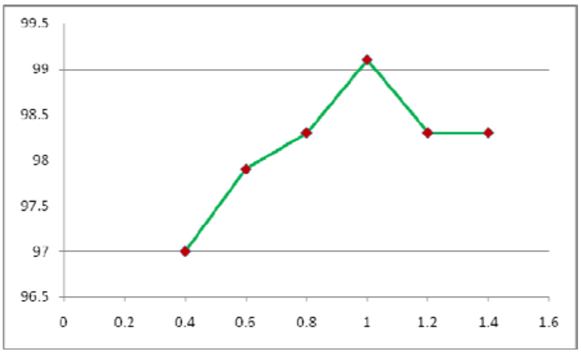

Fig. 5 – RBF_Sigma Vs Classification Accuracy Curve

When the RBF_Sigma is 1, the testing is done for varying values of BoxConstraint i.e. 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.8, 2 where classification accuracy 98.3, 98.3, 97.9, 98.3, 98.3, 99.1, 99.1, 99.1, 99.1 in percentage is achieved respectively. So, when the value of BoxConstraint is 0.2, the classification accuracy of 98.3% is observed and gently increased for the increase of BoxConstraint value and when BoxConstraint is 0.6, the accuracy got down to 97.9% and the maximum classification accuracy is achieved at BoxConstraint value of 1.2 and maintained the same accuracy when the BoxConstraint is increased and so on which is showed in Table.4 and Fig 6.

Table.4 -BoxConstraint Vs Classification Accuracy

BoxConstraint	Classification Accuracy
0.2	98.3
0.4	98.3
0.6	97.9
0.8	98.3
1	98.3
1.2	99.1
1.4	99.1
1.8	99.1
2	99.1

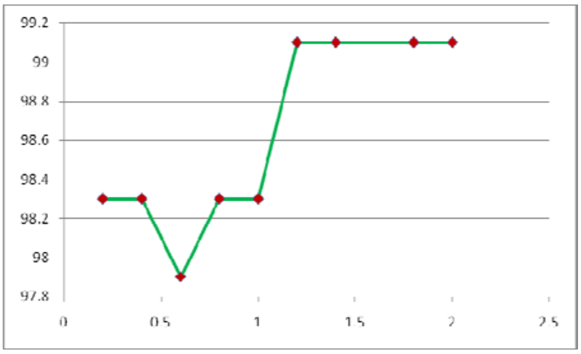

Fig (6) – BoxConstraint Vs Classification Accuracy Curve

Next, by applying the Polynomial kernel function the maximum classification accuracy is achieved when the arguments Polyorder value is 2, BoxConstraint value is 0.002. For the kernel function Polynomial, the effect of the arguments polyorder and BoxConstraint for the classification is demonstrated below [6].

When the Polyorder value is 2, the testing is done for varying values of BoxConstraint i.e. 0.002, 0.02, 0.2, 2, 20, 200 where classification accuracy achieved is 98.7, 98.7, 97, 96.6, 95.7, 95.7 (in percentage) respectively. So, when the value of BoxConstraint is 0.002 the maximum classification value of 98.75% is observed and gently decreased for the increase of BoxConstraint value and when BoxConstraint is 20, the Accuracy got down to 95.7% and the classification accuracy is maintained at the same level which is showed in Table.5 and Fig.7.

Table.5 - BoxConstraint Vs Classification Accuracy

BoxConstraint	Classification Accuracy
0.002	98.7
0.02	98.7
0.2	97
2	96.6
20	95.7
200	95.7

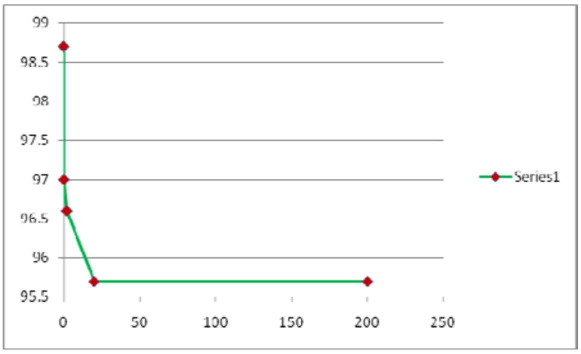

Fig.7 – BoxConstraint Vs Classification Accuracy Curve

IV. Conclusion

The classification techniques are emerging as the key factors for the diagnosis of Breast cancer disease. We have proposed SVM model using RBF and Polynomial kernel functions with varying arguments such as RBF_Sigma, BoxConstraint & Polyorder and achieved promising results of 97.13 percent classification accuracy. We may go for few Optimization techniques [7] which will select the arguments randomly of various kernel functions to enhance the classification accuracy. When the disease is classified as malignant, we would like to focus on the clinical factors such as the proteins (drug target) which are causing the disease [8] and to design the necessary drug (ligand) [9] for the respective drug target by using Insilco drug discovery techniques which may arrest the Breast cancer.

Acknowledgment

My Heartfelt Gratitude to my parents for their moral support and Prof. K.Thammi Reddy and Prof. Valli Kumari for their unique guidance. I am very thankful to KVSRP Varma who has assisted in producing this journal

References

1) A Computational Intelligence Technique for Better Diagnosis of Diabetes Disease using Support Vector Machines with RBF Kernel Function. Kamadi VSRP Varma, Dr. Allam Apparao, Dr. T.Sitamahalaxmi, Dr. P.V.Nageswar Rao, Kalagotla Satish Kumar. Proceedings of the National conference on “Advances in Computing & Networking” [ISBN: 978 93 83038 11 4]
2) Gouda I. Salama, M.B.Abdelhalim, Magdy Abd-elghany Zeid “Breast Cancer Diagnosis on Three Different Datasets”. International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012
3) http://www.cs.columbia.edu/~kathy/cs4701/documents /jason_svm_tutorial.pdf
4) http: //www.mathworks.in/help/stats / svmtrain.html
5) S. Aruna et al. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer.
6) Seeja.K.R, Shweta. Microarray Data Classification Using Support Vector Machine. International Journal of Biometrics and Bioinformatics (IJBB), Volume (5) : Issue (1) : 2011 10-15.
7) Minaei-Bidgoli, B., Punch, W. 2003. Using genetic Algorithms for data mining optimization in an educational web-based system, Genetic and Evolutionary Computation, pp: 2252-2263.
8) David E. Misek and Evelyn H. Kim. Protein Biomarkers for the Early detection of Breast cancer.International Journal of Proteomics Volume 2011 (2011), Article ID 343582, 9 pages
9) Bevan Kai-Sheng Chung1, Thomas Dick, Dong-Yup Lee. In silico Analysis for the discovery of Tuberculosis drug targets. Journal of Antimicrobial chemotherapy 10.1093/jac/dkt273.