Performance Measure of Medical Disease Classification using Support Vector Machine

D.Haveela Bala
Research Scholar
Department of Computer Science
Rayalaseema University
Kurnool - 518007

Abstract - The classification of medical disease prediction has become an increasingly challenging problem, due to recent advances in data collection and medical mining technology. The Clinical organizations have collected large quantities of information about patients and diseases. In this paper we examine fundamental aspects of the SVM classification in medical diagnosis. We also discussed different Kernel methods, Kernel techniques have long been used in SVM to handle linearly inseparable problems by transforming data to a high dimensional space. And also presents a performance evaluation metrics with different parameters on the three different medical data sets namely, Wisconsin Breast cancer, Mammographic-Mass and Pima Diabetes in python Language from UCI Datasets. The empirical results demonstrate SVM classification is an efficient classification for medical disease prediction.

Keywords: SVM, Classification, Data Mining and Kernel

I. INTRODUCTION

Now a day's various Healthcare organizations are generating huge amounts of data which are difficult to handle for further processing and it also needs to provide diagnosis aid accurately. The Healthcare organizations have collected large quantities of information about patients, diseases and their clinical lab test results. Data mining is the search for relationships and patterns within this data that could provide useful knowledge for effective decision-making. Medical data mining is one of key issues to get useful clinical knowledge from medical databases. Many different data mining techniques exist for medical diagnosis such as classification, association rules and clustering are used by clinical organization to increase their capability for making decision regarding patient health.

.Data mining is the process of extracting valid, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions. The extracted information can be used to form a prediction or classification model, or to identify relations between database records [1]. In this paper, SVM is used to classify the Medical data.

The remainder of the paper is structured as follows. Section 2 discusses the overview of classification. Section 3 provides the theoretical background of Support Vector Machine for classification. In Section 4 we present experimental results for three real-world data sets and accuracy parameters are discussed. We conclude in Section 5 with some discussion of the potential significance of our results.

II. CLASSIFICATION OVERVIEW

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts. Constructing fast and accurate classifiers for large data sets is an important task in data mining and machine learning [1, 5]. Classification is the most commonly applied data mining technique, and employs a set of pre-classified examples to develop a model that can classify the population of records at large.

Classification is a two-step process. In the first step, which is called the learning step, a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database instances. Each instance is assumed to belong to a predefined class. In the second step, the model is tested using a different data set that is used to estimate the classification accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data instances for which the class label is not known. There are several algorithms that can be used for classification such as decision tree, Support Vector Machines, Bayesian methods, rule based algorithms, and Neural Networks [5].

The objective is to use the training data set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data set attributes. With classification, the generated model will be able to predict a class for given data depending on previously learned information from historical data.

III. SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) is a machine learning technique for classification and regression analysis based on statistical learning theory [3]. SVMs have been applied to number of real world problems such as medical diagnosis, hand-written recognition, pattern recognition and image processing.

The SVM classifies the input data by constructing an N-dimensional hyperplane that optimally separates the data into two categories. The subsets of data instances that actually define the hyperplane are called the "support vectors", and the margin is defined as the distance between the hyperplane and the nearest support vector [4] as shown in figure-1. By maximizing this separation, it is believed that the SVM better generalizes to unseen data instances, while also mitigating the effects of noisy data or over-training. Error is minimized by maximizing the margin, and the hyperplane is defined as the center line of the separating space, creating equivalent margins for each class. The main goal of SVM method is to separate input data into two classes using examples of each from the training data to define the separating hyperplane.

The SVM method draws a maximum margin hyper plane which separates a set of positive examples from a set of negative examples. In the case of examples not linearly separable, SVM uses a Kernel functions to map the examples from input space into high dimensional feature space. Using a Kernel function can solve the non-linear problem.

The purpose of SVM classification is to device a computationally efficient way of learning good separating hyper planes in a high dimensional feature space.

The simple procedure of the SVM algorithm is provided as follows. Suppose a training set (x_i, y_i), i =1, 2, m where $x_i \varepsilon R^n$, $y \varepsilon \{+1, -1\}$, the following conditions:

$x_i + b \geq +1$ for $y_i = +1$

$x_i + b \leq -1$ for $yi = -1$ (1)

Which is equivalent to:

$y_i (x_i . w + b) - 1 \geq 0 \ \forall i = 1, n$ (2)

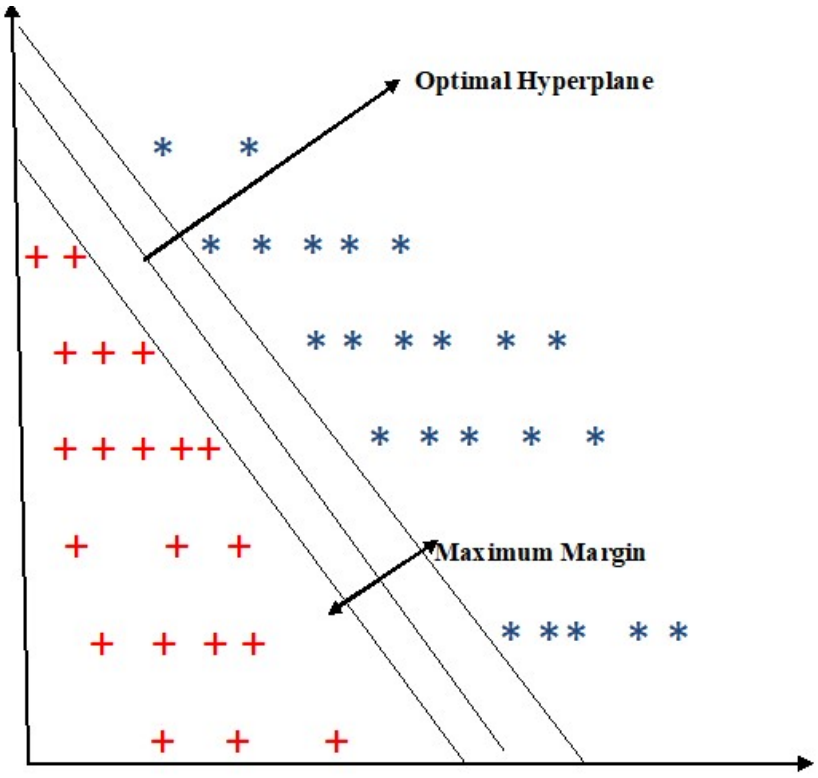

Figure 1: SVM Classification Optimal Hyper plane

This method draws a hyper-plane $(w. x_i) + b = 0$ to separate the data from classes label +1 and -1 with a maximal margin in the input feature space. The maximization of the margin is equivalent to minimize the norm of *w*. Thus, SVM can be trained to solve the following optimization problem:

$f(x) = \text{sign}(w. x) + b$ (3)

$$\text{Minimize } \frac{1}{2} w^T w + \sum_{i=1}^{N} \xi_i \quad (4)$$

Subject to $y_i (w. x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for i =1, n.

For linearly separable data, the support vectors are the subset of actual training tuples. The above equation tells us on which side of the hyper plane the test tuple x^T falls. If the sign is in positive, then x^T falls on or above the maximal margin hyper plane and SVM predicts that x^T belongs to class +1. If the sign is in negative, then x^T falls on or below the maximal margin hyperplane and the class prediction is -1. There are two types of SVMs, (1) Linear SVM, which separates the data points using a linear decision boundary and (2) Non-linear SVM, which separates the data points using a non-linear decision boundary [3, 4].

The Linear SVM performs well on datasets that can be easily separated by a hyperplane into two parts. But sometimes datasets are complex and are difficult to classify using a linear Kernel. Non-linear SVM classifiers can be used for such complex datasets. The concept behind non-linear SVM classifier is to transform the dataset into a high dimensional space where the data can be separated using a linear decision boundary.

IV. EXPERIMENTAL RESULTS

The SVM methods have been experimented with data taken from the UCI Machine Learning Repository [2] and used the Python Language to experiment the SVM algorithm. The Python Scikit-learn is a package for data classification, regression, clustering and visualization. The dataset we used in our experiment is briefly described in Table 4.1. The data is divided in two sets. The training set is 70% and the remaining 30% are used for testing. The experiments were conducted with complete feature set.

Table 4.1: provides the attribute information of three UCI datasets

SNO	Datasets	Features	Instances	Class
1	Wisconsin Breast cancer	11	699	2
2	Pima Diabetes	9	768	2
3	Mammographic-Mass	6	961	2

i. Measures for performance evaluation

There exist different measures that can be used to evaluate the performance of a classifier, such as confusion matrix, Accuracy, sensitivity, specificity, precision and recall, etc. Each of these evaluation measures have their own limitations and, as a result, an appropriate evaluation measure which best suits the problem should be selected. Due to the factors mentioned, and in order to have a reliable performance evaluation, the assessment should be considered based on the Cross validation.

Accuracy: Accuracy is a measure which determines the probability that how much results are accurately classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Sensitivity: Sensitivity is a measure which determines the probability of the results that are true positive such that person has the disease.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (6)$$

Specificity: Specificity is a measure which determines the probability of the results that are true negative such that person does not have the disease.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (7)$$

Precision: Precision represents how precise the classifier predictions are since it shows the amount of true positives that were predicted out of all positive labels assigned to the instances by the classifier. Precision is the proportion of positive predictions that are correct

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

Recall: Recall is the proportion of positive samples that are correctly predicted positive. It shows the amount of truly predicted positive classes out of the amount of total actual positive classes.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

Where,

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

Confusion matrix is a visualization tool which is commonly used to present the accuracy of the classifiers in classification that assist with performance evaluation purposes which consist of the concepts defined above measurements. This is illustrated in table 4.2. It is used to show the relationships between outcomes and predicted classes.

Table 4.2: confusion matrix

		Predicted	
		Positive	**Negative**
Actual Class	**Positive**	TP	FN
	Negative	FP	TN

ii. Results

The confusion matrix of SVM classification method is presented in the table 4.3 of three data sets. The values to measure the performance of the methods (i.e. accuracy, sensitivity and specificity) are derived from the confusion matrix and shown in table 4.4 and same shown in graphical representation in figure 2

Table 4.3: Confusion Matrix of three UCI Test Data sets

Wisconsin Breast cancer Test Data (205)					**Pima Diabetes Test Data (231)**			
		Predicted					**Predicted**	
		Benign	**Malignant**				**Negative**	**Positive**
Actual Class	**Benign**	125	5		**Actual Class**	**Negative**	141	16
	Malignant	5	70			**Positive**	34	40

Mammographic-Mass Test Data (289)			
		Predicted	
		Benign	Malignant
Actual Class	Benign	140	11
	Malignant	51	87

Based on the above confusion matrices, we calculated Accuracies, Sensitivity, Specificity, Precision and Recall as shown in table 4.4 same shown in graphical representation in figure- 2. It can be seen that the SVM algorithm of all features of accuracy on Breast cancer (95%), Pima Diabetes (78%) and Mammographic-Mass (79%) of accuracies.

Table 4.4: Performance of SVM Classification

S.No	Dataset	Accuracy	Sensitivity	Specificity	Precision	Recall
1	Wisconsin Breast cancer	95	96	93	95	96
2	Pima Diabetes	78	95	88	72	54
3	Mammographic-Mass	79	92	64	88	63

Table 4.4 shows comparative results of classification accuracy and the same shown in bar graph in figure 2. It can be seen that the SVM algorithm of all features of accuracy on Breast cancer (95%) and Pima Diabetes (75%) data sets.

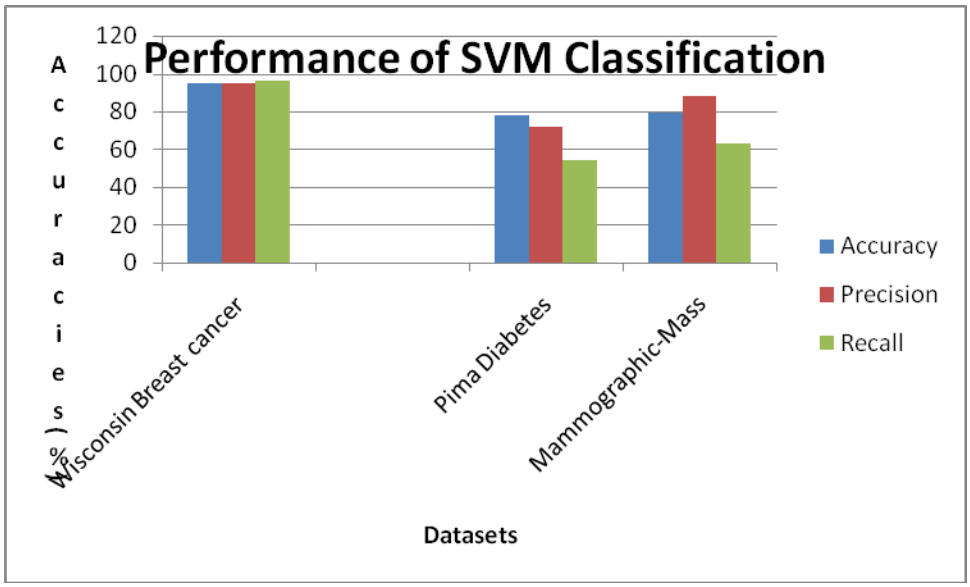

Figure-2: Performance of SVM Classification

V. CONCLUSION

In this research, we applied three medical databases to test the efficiency of the SVM algorithm. In our research various performance metrics are chosen to be compared and evaluated. This paper aims at studying the SVM

classification for medical disease prediction in various aspects. The problem of learning efficient model from data with high dimensionality can cause trouble to most algorithms. From experimental results, it has been revealed that the SVM based classification method can increase the accuracy of data classification in all three datasets. Thus, SVM is an effective method for Medical data classification.

REFERENCES

[1] J.Han and M.Kamber,"Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006

[2] UCI machine learning repository. http://archive.ics.uci.edu/ml/.

[3] Vapnik V.N, "Statistical learning Theory", John Wiley and Sons, New York, USA, 1998.

[4] Vapnik, V.N. The Natural of Statistical Learning theory. Springer–Verleg, NewYork, USA1995.

[5] Written H.I, Frank E, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann Publishers, 2005.