ISSN:0975-9646

A study on Comparative Performance of SVM Classifier Models with Kernel Functions in Prediction of Hypertension

Rahul Samant[#1], Srikantha Rao[#2]

[#1] *Department of IT , SVKM'S NMIMS, Shirpur, Dhule, India*

[#2] *Department of Management Research, University of Mumbai, , Kandivali, Mumbai, India*

***Abstract*— This paper investigates the ability of several models of Support Vector Machines (SVMs) with alternate kernel functions to predict the probability of occurrence of Essential Hypertension (HT) in a mixed patient population. To do this a SVM was trained with 13 inputs (symptoms) from the medical dataset. Different kernel functions, such as Linear, Quadratic, Polyorder (order three), Multi Layer Perceptron (MLP) and Radial Basis Function kernel (RBF) were coded and tested. A detailed database, comprising healthy and diabetic patients from a university hospital was used for training the SVM for prediction. All five kernel function SVM structures tested showed reasonably good accuracy in prediction of disease (s), with linear kernel structure showing best prediction in 3 out of 4 datasets and Polyorder in one database. Thus the best choice appears to be situation specific.**

Keywords*— Support Vector Machine, Classification, Kernel functions, Hypertension

I. INTRODUCTION

Hypertension is the most commonly diagnosed condition in medical practice. Hypertension is deemed as a factor for Syndrome X that has been investigated for years in epidemiologic studies. Linked to a plethora of medical disorder, hypertension appears as a top risk factor for life threatening conditions, such as stroke and heart attack. [1, 3]. Hence the early detection is indeed needed to improve general health care. Although earlier identification of this disease is gaining importance in clinical research, the investigation of factors for prevention and intervention are also crucial issues in preventive medicine. Modifiable factors such as life-style variables and body measurements, for reducing risk of the disease are especially interesting for public health professionals. [5]. It has also been shown that employing computer aided diagnostic systems (CAD) as a "second opinion" has lead to improved diagnostic decisions and support vector machines (SVMs) have shown remarkable success in this area [6]. In contrast to logistic regression, which depends on a pre-determined model to predict the occurrence or not of a binary event by fitting data to a logistic curve, SVM discriminates between two classes by generating a hyper plane that optimally separates classes after the input data have been transformed mathematically into a high-dimensional space. Because the SVM approach is data-driven and model-free, it may have important discriminative power for classification, especially in cases where sample sizes are small and a large number of variables are involved (high-dimensionality space). This technique has recently been used to develop automated classification of diseases and to improve methods for detecting disease in the clinical setting. [4]. Clinical diagnostics has always depended on the clinician's ability to diagnose pathologies based on the observation of symptoms exhibited by the patient and then classifying his/her condition. Correct diagnosis can make the difference between life and death in the correct and timely intervention. Similar situations arise where the precise links between cause and effect is not yet established and one is predestined to process a certain amount of data to draw inferences to guide decisions.

Classification is challenging not only in respect to acquiring the relevant data through tests about factors known to be associated with the pathology, but also the data analytics adopted to lead to reliable and correct prediction. This present paper looks into one such data analysis technique, now about 20 years in use and known as *support vector machine* or SVM, that helps one to develop classification models based on statistical principles of learning. Like artificial neural networks, an SVM is data driven—it is trained using a dataset of examples with known class (label), and then utilized to *predict* the class of new examples.

Ture et al [5] compared performances of three decision trees, four statistical algorithms, and two neural networks in order to predict the risk of essential hypertension disease. MLP and RBF—two neural networks procedures—performed better than other techniques in predicting hypertension. Hsu et al [6] constructed a classification approach based on the hybrid use of case-based reasoning (CBR) and genetic algorithms (GAs). Hypertension detection was attempted using anthropometric body surface scanning data. The obtained result revealed the relationship between a subject's 3D scanning data and hypertension disease. GA was adopted to determine the optimum feature weights for CBR. The proposed approaches were compared with a regular CBR and other widely used approaches including neural nets and decision trees.

Zhang et al [7] developed an ANN based automated computer aided diagnosis system to help radiologist in detecting micro-calcifications in digital format mammograms. Dana [8] developed a ANN based AI system to detect breast cancer. The system was trained using eight input nodes represent features of calcification, areas in breast tissues where tiny calcium deposits built up and might indicate the presence of cancer.

II. THE SVM CLASSIFICATION METHODOLOGY

Since 1980 as the power of computing began to grow, automated learning aimed at modeling and understanding relationships among a set of variables derived from objects drew much interest. The goal became that of using supervised learning to model the relationship between some selected inputs and outputs. Artificial Neural Nets (ANN) and Support Vector Machines (SVM) are two such devices created in that period and these continue even today as state-of-the-art classification methods. Of late, in the last twenty or so years, SVM has been extensively used to target problems of classification where an input-output training dataset is presented to the algorithm, which in turn, when its learning is complete, becomes capable of classifying *yet new* input data. [2]. Most work on SVM and its applications have focused on the two-class pattern classification problem. [9, 10].

Briefly, the two-class SVM classifier may be described as follows, though comprehensive references on it are already extensive. [14, 15]

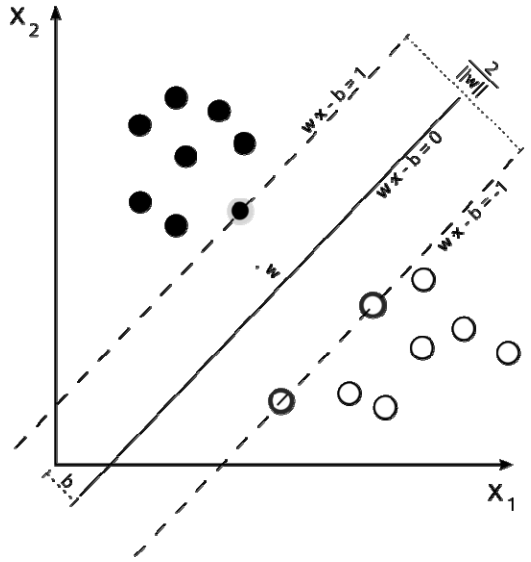

Figure 1: The Hard Margin SVM in the (X_1, X_2) feature space[10]

Let vector **x** of inputs be a pattern that we need to classify and let y (a scalar) denote its assigned class label, ±1. Let {($\mathbf{x}_i$, y_i), i = 1, 2, … *l*} be the training examples based on *l* patterns classified earlier by examining each example and tagging or labeling it as "+1" or "-1" earlier. The SVM's learning task then becomes constructing the classifier or a decision function f(**x**) that would be able to correctly classify a new input pattern **x** not included in the training set. Such classifiers may be linear, or nonlinear.

The different kernel functions are listed below. More explanation on kernel functions can be found in the book by Vapnik. [11]. The below mentioned ones are extracted from there and just for mentioning purposes.[13]

1] *Polynomial:* A polynomial mapping is a popular method for non-linear modeling. Intuitively, the polynomial kernel considers given features and combination of these features to determine their similarity. The second kernel is usually preferable as it avoids problems with the hessian becoming Zero.

$$K(x,x') = (s(x,x') + c)^d \quad (6)$$

$$K(x,x') = ((x,x') + 1)^d \quad (7)$$

2] *Gaussian Radial Basis Function*: Radial basis functions most commonly with a Gaussian form

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad (8)$$

3] *Quadratic*: This kernel function is used with non-linearly separable data.

$$K(x,x') = (s(x,x') + c)^2 \quad (9)$$

4] *Multi-Layer Perceptron*: The long established MLP, with a single hidden layer, also has a valid kernel representation.

$$K(x,x') = \tanh((x,x') + C) \quad (10)$$

5] *Linear*: This kernel function is used to classify linearly separable data.

$$K(x,x') = (x,x') \quad (11)$$

where s, c and σ are kernel-specific parameters.

If the training dataset *is* linearly separable, there will exist a linear function or hyper plane of the form

$$f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x} - \gamma \quad (1)$$

such that for each training example $\mathbf{x_i}$ the function yields $f(\mathbf{x}) \geq 0$ whenever $y_i = +1$, and $f(x) < 0$ when $y_i = -1$. Thus the training data are separated by a function $f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x} - \gamma = 0$, the equation representing the hyperplane in the **x** space. While there may be many such hyper planes existing that can achieve such separation of **x**, SVM aims at locating the hyper plane that maximizes the separation between the two classes of **x** it creates. Mathematically, this is achieved by finding unit vector **w** that minimizes a *cost function* $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2}\|\boldsymbol{w}\|^2$ subject to the separability constraints

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i - \gamma) \geq 1; i = 1, 2, 3, \ldots, l. \quad (2)$$

Sometimes the training data is not completely separable by a hyper plane. In such situations a slack variable ξ_i is added to relax the strict separability constraints in (2) as follows:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i - \gamma) \geq 1 - \xi_i; \xi_i \geq 0; i = 1, 2, 3, \ldots, l. \quad (3)$$

The new cost function that now must be minimized becomes

$$J(\boldsymbol{w}, \xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}\xi_i \quad (4)$$

Vapnik called *C* a user-specified, positive "regularization" parameter. In the general sense, not all situations comprising training examples {($\mathbf{x}_i$, y_i), i = 1, 2, … *l*} can be effectively modeled by the linear relationship (1), for the relationship may be nonlinear. To handle these SVM utilizes *kernels*—functions that can easily compute dot products of two vectors, a key requirement to achieve computational efficiency (Ng 2013).

In (1) $\boldsymbol{w}$ is a weight vector and b is the bias. The hyper plane {**x**: $f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x} - \gamma$} devides the input space of **x** into two and the sign of f(**x**), the discriminant function of the classifier, denotes the side of the hyper plane a point **x** is on. The decision boundary is the demarcation between the two

regions classified as positive and negative. When the decision boundary is a linear function of the input examples, it is called a liner classifier. In general, this boundary can be nonlinear. If we assume that the input data space spanned by **x** *is* linearly separable, a linear decision boundary (a hyper plane) exists in it. Indeed, many such hyper planes may exist. The goal of SVM learning is to use the input data to design an optimum hyper plane (f(**x**)) that will maximize the geometric distance (the "margin") between the examples in the two classes. This is achieved as stated earlier by finding unit vector **w** that minimizes the *cost function* $\frac{1}{2}w^T w = \frac{1}{2}\|w\|^2$ subject to the separability constraints

$$y_i(w^T x_i - \gamma) \geq 1; i = 1, 2, 3, \ldots, i.$$

These constraints here ensure that the classifier f(**x**) classifies each example $\mathbf{x_i}$ correctly. Under the just stated assumption of linear separability being possible, the ***hard margin SVM*** (Figure 1, source Stackoverflow.com 2013) can be constructed to help classify unseen examples. Note that γ is computed once (4) has been minimized. [9, 11]. Mathematically this problem is one of optimization:(5)

$$minimize\ \frac{1}{2}\|w\|^2\ using\ w, \gamma\ subject\ to\ y_i(w^T x_i - \gamma) \geq 1\ i = 1, 2, 3, \ldots, n$$

III. Methodology

The database used for analysis in this study has been compiled as a part of an earlier study entitled Early Detection Project (EDP) conducted at the Hemorheology Laboratory of the erstwhile Inter-Disciplinary Programme in Biomedical Engineering at the School (now Department) of Biosciences and Bioengineering, Indian Institute of Technology Bombay (IITB), Mumbai, India. Spanning over a period from January 1995 to April 2005, it compiled 981 records, each with 30 parameters, which encapsulated the biochemical, hemorheological and clinical status of the individuals. We note that the Hemorheology Laboratory has pioneered the research in the field of Clinical Hemorheology by conducting the baseline hemorheological studies in the Indian population and correlating various hemorheological parameters with several disease conditions.

We apply KNN-imputations to impute the missing values in the dataset used in this study [19]. We also used well established method of Principal Component Analysis (PCA) of feature reduction technique to select 13 important features which were also clinically accepted in the literature to predict hypertension [20].

In all, 13 parameters were noted for each respondent. Table 1 describes the symptom (input) variables used for the present study. They include age, health indicators (e.g. systolic blood pressure (BP1), diastolic blood pressure (BP2)) and biochemical parameter like Serum Proteins (SP), Serum Albumin (SALB), Hematocrit (HCT), Serum Cholesterol (SC), Serum Triglycerides (STG), along with various hemorheological (HR) parameters (e.g.; Whole Blood Viscosity(CBV), Plasma Viscosity(CPV), using a Contraves 30 viscometer, and Red Cell Aggregation (RCA).We used this database to develop and validate SVM models for four classification schemes: Classification Scheme I (healthy vs. diabetic) , Classification Scheme II (diabetic vs. hypertensive), classification scheme III (diabeteic vs diabetic+hypertensive) and classification scheme IV (healthy vs. diabetic) with knn-imputed data for missing values . The SVM models were used to select thirteen input variables that would yield the best classification of individuals into these diabetes categories.

For inputs to the SVM model, the first 13 columns of data represent the patient's health parameters. The 14th column represented the diagnosis made by the doctor for the patient. Dataset DS1 is a mixed data set, having samples of diabetic and healthy patients. Dataset DS2 is a dataset which stores data about hypertensive and diabetic patients. DS3 is a dataset, having diagnosis information about patients who are diabetic and hypertensive as well as diabetic. Dataset DS4 is KNN imputed dataset for missing values having information about patients who are diabetic and healthy.

Table 1: Diagnosis variables of datasets used in the study

Num.	Symptom variable name	Data Type
1.	AGE	Numeric, Range(19-73)
2.	BSF	Numeric, Range(48,311)
3.	BSP	Numeric, Range(61,383)
4.	SC	Numeric, Range(90,389)
5.	STG	Numeric, Range(41,456)
6.	SALB	Numeric, Range(3.1,6.45)
7.	SP	Numeric, Range(0.83,10.68)
8.	CPV	Numeric, Range(1.069,1.785)
9.	CBV	Numeric, Range(2.448,8.695)
10.	HCT	Numeric, Range(22,60)
11.	RG	Numeric, Range(1.374,6.174)
12.	BP1	Numeric, Range(98,240)
13.	BP2	Numeric, Range(60,116)

IV. Results and Discussion

Five different kernel functions namely *linear, quadratic, Polyorder, MLP and RBF*, were evaluated in terms of their discriminative classification accuracy. The liner kernel function performed best in Classification Scheme -- I, III and IV, and the quadratic linear kernel function performed best in Classification Scheme II. Performance parameters such as the accuracy, sensitivity and specificity were presented in Table. The SVM was train using four datasets. For the first dataset,DS1—SVM with *linear* kernel recorded the best classification accuracy of 84.83% with sensitivity 87.32 % and specificity 83.22%. The best classification accuracy of 85.33% for DS2 was shown by *quadratic* kernel function. The sensitivity and specificity were 79.35 % and 82.12% respectively. For datasets, DS3 and DS4 all the kernel functions displays satisfactory level of accuracy, but the *linear* kernel function was a better choice due to slightly better accuracy level. The classification accuracy for first three datasets was higher than that of fourth dataset due to the fact that the fourth dataset was KNN-imputed for missing values and first three datasets were cleaned datasets.

Table I Experimental results of SVM classifier accuracy (sensitivity, specificity).

	SVM classification accuracy with kernel functions				
	Linear	**Quadratic**	**Poly (3)**	**RBF**	**MLP**
DS1	85.2 (87.3, 83.2)	82.7 (80.8, 82.6	76.4 (80.1, 83.7)	75.1 (84.3, 85.6)	67.3 (71.7,62.6)
DS2	83.8 (81.1, 85.1)	84.8 (79.3, 82.1)	74.5 (81.3, 88.6)	72.2 (80.4, 88.6)	68.5 (72.3,68.1)
DS3	88.4 (88.5, 84.4)	86.2 (84.2, 80.4)	75.1 (79.3, 88.5)	63.4 (71.3, 78.6)	64.1 (63.3,68.6)
DS4	81.2 (75.6, 78.2)	80.5 (71.1, 68.9)	68.37 (71.5, 72.8)	66.13 (61.3, 71.6)	63.5 (65.3,68.6)

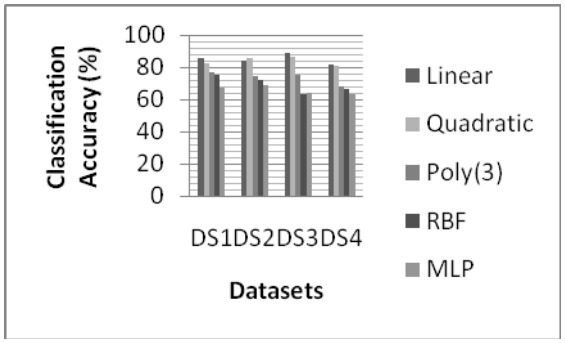

Fig. 2 Classification accuracy for different SVM kernel function models

V. CONCLUSIONS

Support vector machine modeling is a promising classification approach for detecting persons with common diseases such as diabetes and pre-diabetes in the population. In this study we implemented SVM with five different kernel functions and investigated the appropriate choice of kernel function for the prediction of diabetes.

SVM is a model-free method that provides efficient solutions to classification problems without any assumption regarding the distribution and interdependency of the data. In epidemiologic studies and population health surveys, the SVM technique has the potential to perform better than traditional statistical methods like logistic regression, especially in situations that include multivariate risk factors with small effects (e.g., genome-wide association data and gene expression profiles), limited sample size, and a limited knowledge of underlying biological relationships among risk factors [18]. This is particularly true in the case of common complex diseases where many risk factors, including gene-gene interactions and gene-environment interactions, have to be considered to reach sufficient discriminative power in prediction models. Our work provides a promising proof of principle by demonstrating the predictive power of the SVM with just a small set of variables. This approach can be extended to include large data sets, including many other variables, such as genetic biomarkers, as data from different domains become available.

REFERENCES

[1] WHO/IDF 2006 (2007, Jan.). *Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia,* World Health Organization [Online]. Available: http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf

[2] Ban Hyo-Jeong, Jee Yeon Heo, Kyung-Soo Oh and Keun-Joon Park (2010). Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine, *BMC Genetics,* 11:26 http://www.biomedcentral.com/1471-2156/11/26

[3] M. Uusitupa, "Lifestyle matter in prevention of type 2 diabetes," *Diabetes Care*, vol. 25, no. 9, pp. 1650–1651, 2002.

[4] Rudiger W. Brause,(2000) Medical Analysis and Diagnosis by Neural Networks , URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.2901&rep=rep1&type=pdf.

[5] Boser, BE. I M Guyon and V N Vapnik (1992). *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ed. D Haussler, ACM Press.

[6] Lijun Cheng, Yongsheng Ding , SVM and statistical technique method applying in Primary Open Angle Glaucoma diagnosis, Intelligent Control and Automation (WCICA), 2010 8th World Congress, pp- 2973 - 2978

[7] Shashikant Ghumbre, Chetan Patil, and Ashok Ghatol,(2011) , Heart Disease Diagnosis using Support Vector

[8] Machine , International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011 ,pp 84-88.,

[9] Xiao-Peng Zhang*, Zhi-Long Wang, Lei Tang, Ying-Shi Sun, Kun Cao and Yun Gao, (2011) Support vector machine model for diagnosis of lymph node metastasis in gastric cancer with multidetector computed tomography: a preliminary study URL: http://www.biomedcentral.com/1471-2407/11/10

[10] A. Kampourakia, D. Vassisa, P. Belsisb, C. Skourlasa, (2013), e-Doctor: A Web based Support Vector Machine for Automatic Medical Diagnosis, *Procedia - Social and Behavioral Sciences* Volume 73, 27 February 2013, Pages 467–474

[11] Yu-Len Huang , Kao-Lun Wang , Dar-Ren Chen (2005) , Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines URL: http://web.thu.edu.tw/ylhuang/www/publish/NCA200604.pdf

[12] R.priya and P Aruna. Article: SVM and Neural Network based Diagnosis of Diabetic Retinopathy. *International Journal of Computer Applications* 41(1):6-12, March 2012

[13] Chunquan Huang The Research on Evaluation of Diabetes Metabolic Function Based on Support Vector Machine 2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010)

[14] Vapnik, VN (1998). *Statistical Learning Theory*, John Wiley.

[15] El-naqa, Isham (2012). Machine learning methods for predicting tumor response in lung cancer, Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, Vo. 2 (2), 173-181.

[16] Han, J and M Kamber (2006). *Data Mining Concepts and Techniques*, 2nd ed., Morgan Kaufman.

[17] Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[18] Tapan Bagchi, Rahul Samant, Milan Joshi (2013), "SVM Classifiers Built Using Imperfect Training Data", International Conference on Mathematical Techniques In Engineering Applications, ICMTEA 2013-BM-003

[19] Rahul Samant, Srikantha Rao . " *Effects of Missing Data Imputation on Classifier Accuracy* ", Vol.2 - Issue 11 (November - 2013), International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181 , pp 264-266, URL: www.ijert.org

[20] Rahul Samant, Srikantha Rao . " *A study on Feature Selection Methods in Medical Decision Support Systems* ", Vol.2 - Issue 11 (November - 2013), International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181 , pp 615-619, URL: www.ijert.org