

Heart Disease Prediction using Kernel based Support Vector Machines

Dr. D. Rajeswari¹, Mrs. K. Sathya², Ms. S. Indra³, Dr. G. Jothi⁴

^{1,2,3} *Department of Computer Science, Sona College of Arts and Science, Affiliated to Periyar University, Salem. Tamil Nadu.*

⁴ *Department of Computer Applications, Sona College of Arts and Science, Affiliated to Periyar University, Salem. Tamil Nadu.*

Abstract— [Font: Heart Disease is the first largest fatality disease all over the global. At the same period, it is curable if it is effectively predicted early. The prediction of heart disease using Kernel based Support Vector Machine (SVM) classifier is proposed in this paper. SVM was trained to classify and predicting the heart disease according to the features presented. Experiments have been conducted of various training-test partitions involving Statlog heart disease datasets from UCI repository. In the first stage, dimension of heart disease dataset has 13 explanatory variables is reduced to 6 variables using feature selection procedure namely Logistic Regression. In second stage, a support vector machine with different kernel functions namely linear, polynomial, Radial basis and Sigmoid was utilized as classifier. The performance of SVM classifier with each kernel function was evaluated by using performance indices such as accuracy, sensitivity, specificity, and f-measure. The kernel function namely Linear is used to find the best cross validation of training data with parameter (c). Linear kernel function obtained highest accuracy (86.41%) which verifies the efficiency and efficacy of LR-Linear strategy.

Index Terms— Logistic Regression, Support Vector Machine, Kernel Functions, Classification, Heart Disease

I. INTRODUCTION

Heart disease prediction is an important research problem in health care domains and medical communities. This disease affects the regular working mechanism of heart organ it may lead to sudden death. There are several risk factors for heart disease including lack of physical exercise, obesity, alcohol consumption, smoking, older age, diabetes. There are different types of heart disease. Therefore, it would be predicting early which would increase the life period.

It is mandatory one when building the SVM classifier to find the appropriate kernel learning function, such as linear, polynomial, radial basis function (RBF) and sigmoid. However, the superior performance of SVM classifiers constructed using different kernel functions.

In this paper, we propose the kernel-based Support vector machine (SVM) tool for learning the validate kernel function to be embedded into SVM classifier. This approach is used extensively in real life applications, especially for decision support system in the medical diagnostic system for earlier detection of some serious disease.

SVM is a Kernel based method of learning developed by Vapnik, Boser in the year 1992 [1] a method of solving both binary classification and regression tasks [2]. It is the effectively employed classifiers in the field of Bioinformatics for classifying and predicting the disease. In many decision-making domains SVM outperform with high degree efficiency with other machine learning techniques [20].

The performance of SVM is further improved by appropriate Kernel Function [3], which transform the data points to high dimensional space from original input space. With the help of kernel finding an optimal hyperplane in the higher dimensional space which separates two data points without any computational complexity [9]. In particular, to obtain high efficiency result, it is important to select the most appropriate kernel function among several methods. In respect to find the best solution for the classification problem, is to identify the optimal kernel function in SVM, by embedding various kernel methods to corresponding data [21]. In this paper, Logistic Regression model utilizes for feature selection procedure to determine

high ranking features which have strong impact on heart disease, the SVM Kernel based model selection is conducted on the selected features and the relative performance is validate by cross-validation (CV).

The content of this paper is as follows. Section 2 explains the review of literature section 3 presents a feature selection procedure and presents classification method making use of Support vector machine with kernel-based function. Then section 4 presents the experiments carried out and the results. Section 5 and 6 describes the conclusion and reference.

II. RELATED WORKS

According to the literature discusses several machine learning techniques that have been applied to develop heart disease prediction models, such as random forest, artificial neural networks, naïve Bayes, decision trees and support vector machines.

From the above-mentioned methods, have shown that more specifically SVM outperforms better than many other machine learning techniques [5-7].

In the year 2013, Marcin Ciecholewski et.al presents a method based on SVM, Neural networks and the Osuna-Platt algorithm used to detect the ischemic heart disease, these techniques test against the SPECT image database shown that SVM with Gaussian kernel function found to be highly accurate [10]. In the year 2010 D. Conforti and R. Gudino describes the development of kernel-based support vector machine with three kernel functions namely Polynomial, Gaussian, Laplacian for early diagnosis of acute myocardial infarction (AMI) from the above kernel functions polynomial performs better than other [11]. In the year 2010 Ismail Babaoglu et.al used binary particle swarm optimization and genetic algorithm (GA) for feature selection models on determination of coronary artery disease using SVM classifier outperforms well [10]. In the year 2014, Hanaa Ismail Elshazly et.al proposed genetic based support vector machine (GA-SVM) for lymph disease diagnosis. Here GA used for reducing dimension of feature vectors and SVM with different kernel learning functions such as linear, quadratic and gaussian for classification. Experimental results verify that GA-Linear kernel strategy can obtained efficient outcome [12]. In the year 2009, Mehmet Fith Akay proposed SVM with feature selection was used to diagnose the breast cancer. F-score technique utilizes for choosing

input features. The experimental results shows that the proposed methods yield the high degree of classification [13]. Juanying Xie and Chunxia Wang developed a diagnosis model in the year 2011, for erythemato-squamous disease based on SVM with novel hybrid feature selection method named improve F-score and sequential forward search (IFSFS) these methods achieve high efficacy when compared to other techniques [14]. In the year 2003, H. X. Liu et.al the author used the SVM classifier for diagnosing breast cancer and also used to find missing values in the dataset, the SVM exhibits the better performance when compared with other prediction algorithms like back propagation neural network, probability neural network and Unsupervised learning method K-Means clustering [15]. Akshaya Kumar Mandal et al., evaluated the twelve machine learning models including, support vector machines (SVM) achieved the highest performance [22]. Wei Zhou et al., classified the heart disease dataset using SVM [23].

III. METHODS

A. Logistic Regression

Logistic Regression techniques extensively employed for addressing the association between the response variable and explanatory variables [3]. For example, the presence and absence of disease heart disease might be predicted from knowledge of the patients age, cholesterol level, blood sugar level, ECG reports of heart disease and any physical testing procedures. The dependent variables (outcome variables) can be both continuous and categorical. If $X_1, X_2, \dots, X_n$ denote n predictor variables (e.g.: patient age, gender, diabetes presence and so on), Y denotes the presence ($Y = 1$) or absence ($Y = 2$) of disease, and p denotes the probability of disease presence (i.e, the probability that $Y = 1$), the following equation describes the association between the predictor variables and p:

$$\text{Log} \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n \quad (1)$$

where β_0 is a constant and $\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients of the predictor variables $X_1, X_2, \dots, X_n$. The coefficients of regression are determined from the observation data. The probability of disease presence p can be determined with this equation.

Every coefficient of regression defines the scope of the involvement and its contribution to the corresponding

explanatory variable (predictors) to the outcome (class label). The effect of the explanatory variables on the target class is normally estimated by using the odds ratio of the explanatory variable, which represents the factor by which the odds of an outcome change for a one-unit change in the predictor variable. The odds ratio is estimated by taking the exponential of the coefficient (e.g., $\exp(\beta_1)$). For example, if β_1 is the coefficient of variable X ("age"), and p represents the probability of heart disease, $\exp(\beta_1)$ is the odds ratio corresponding to the age factor of heart disease. Logistic regression techniques generally include only the variables that are considered "strongly impact" in cause of heart disease. With use of P values, the most significant risk factors (variables) is defined in terms of the statistical significance of the coefficients for the variables. The significance criterion $P \leq .05$ is generally used when testing for the statistical importance of variables.

The impact variables can be selected with several kinds of methods. In forward selection, variables are sequentially added a classifier without explanatory variables if they are found to be statistically important variables. In vice versa, backward selection method, in presence of all independent variables in the model, and the variables are eliminated one after another as they are determining to be unimportant variables. The stepwise logistic regression method is a mixture of these two feature selection procedures and is used to estimate which variables to add to or eliminate from the model in a sequential fashion on the basis of statistical criteria [4].

B. Kernel-based SVM

Support Vector Machine supervised machine learning algorithm is a Kernel based learning techniques developed by Boser, Vapnik and Guyon in the year 1992 [1], which has been widely used for solves classification problems in the field of Bioinformatics and medical diagnosis and also deals with large dimensional data. Structural risk minimization (SRM) is the basic inductive principle for developing SVM [19]. SRM is the principle rooted in the statistical learning theory, it gives better performance on the test data (unknown data) [18].

$K(x_i, x_j)$ is the dot product of the input vectors x_i and x_j , a separation of two classes with the maximum margin, in all data points lying on one side of the

decision plane are labelled as +1 and another side are labelled as -1. The training cases (data points) lies near to the hyperplane are called Support Vectors (SVs). The SVM classifier consist in defining the function $f(x) = \text{sign}(\langle \omega, X \rangle + b)$ (2)

Which finds the hyperplane, where ω is the weights vector, b is the perpendicular distance from hyperplane to the origin, $f(x)$ is the label of the sample.

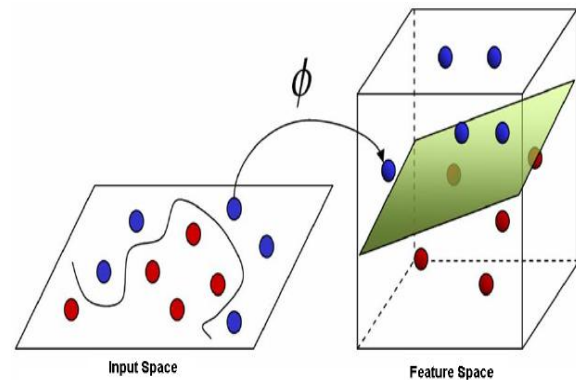


Figure 1: SVM model generation

The selection of Kernel function is the main job in SVM approach. Disparate amount of SVMs will obtain depend on selecting various kinds of kernel learning methods such as polynomial SVMs, radial basis function, linear SVMs and so on for the given data and it may produce different performance in the results [17]. A kernel method is a learning technique which uses a dot product to manage pattern analysis. With the help of different kernel function SVM classify by discovering non-linear relationships among the data.

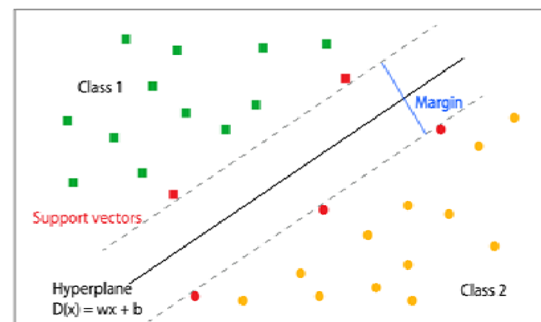


Figure 2: SVM classifier linear case

In our experiments, we choose the following kernel functions

- The linear kernel: $k(x, y) = x^T y + c$ (3)

- The polynomial kernel:

$$k(x_i, x_j) = (x_i \cdot x_j + c)^d \quad (4)$$

- The radial basis function:

$$k(x, y) = \exp(-\gamma \sum_{j=1}^p (x_{ij} - y_{ji})^2) \quad (5)$$

- The sigmoid kernel:

$$k(x, y) = \tanh(\alpha x^T y + c) \quad (6)$$

IV. METHODOLOGY AND EXPERIMENTS

A. Heart Disease dataset

In this paper, statlog heart disease dataset is used here has taken from publicly available dataset (availableat: [http://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](http://archive.ics.uci.edu/ml/datasets/statlog+(heart))). This dataset is commonly used among researchers who use machine learning techniques for heart disease classification. This dataset composed of 270 observations and 13 different features and 1 class label.

Feature selection method	No of selected features	Selected features
Logistic Regression	6	Sex, Chest pain type, Resting blood pressure, Maximum heart rate achieved, number major vessel colored by fluoroscopy, Thal

Table 1: Selected features of Statlog heart disease dataset

Table 1 shows the impact features selected by the logistic regression techniques which is the reason for arising heart disease.

The performance of SVM classifier with each kernel learning method was evaluated by performance metrics such as sensitivity, specificity, accuracy, positive predicted value, negative predicted value and F-Measure.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} * 100\%$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} * 100\%$$

$$\text{Specificity} = \frac{TN}{FP + TN} * 100\%$$

$$\text{Positive predicted value} = \frac{TP}{TP + FP} * 100$$

$$\text{Negative predicted value} = \frac{TN}{FN + TN} * 100$$

$$F - \text{Measure} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Table 2 shows the results of different SVM classifiers using the statlog heart datasets. As we can see that the LR SVM classifiers perform better on the datasets.

Performance metrics	Linear	Polynomial	Radial basis function	Sigmoid
Accuracy	0.864	0.851	0.839	0.839
Specificity	0.828	0.80	0.80	0.828
Sensitivity	0.891	0.891	0.869	0.847
Positive prediction value	0.872	0.854	0.851	0.866
Negative prediction value	0.852	0.848	0.823	0.805
F-measure	0.881	0.872	0.86	0.857

Table 2: Performance measures of SVM classifier corresponding to kernel functions

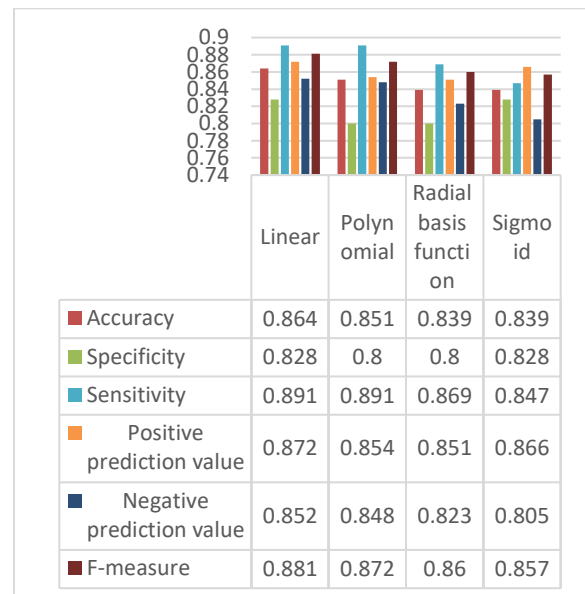


Figure 3: Performance measures of SVM classifier corresponding to kernel learning methods

V. CONCLUSION

In this paper, the performance of logistic regression and SVM with different combination of kernel functions in terms of heart disease prediction. One of the aspects in this paper to find the most significant features. In this study, a hybrid method for predicting

heart disease based on LR and kernel based SVM is presented. Logistic Regression is used for dimension reduction of heart disease dataset and SVM with kernel functions is used for classification. The proposed LR-Linear system is compared with other kernel functions such as polynomial, radial basis and sigmoid, among them, the LR+ linear SVM performing better than other three classifiers. The results showed that LR-Linear classifier achieved 89.1%, 82.8%, 87.2%, 85.2%, 88.1% for sensitivity, specificity, accuracy, positive prediction value, negative prediction value and f-measure accurately for heart disease data. Hence the result verifies efficiency of LR-Linear strategy compared with results generated from the other kernel functions. The size of feature space is reduced from 13 to 6 by using LR.

REFERENCES

- [1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [2] H. Yu and S. Kim, "SVM tutorial—Classification, regression and ranking," in *Handbook of Natural Computing*, 2012.
- [3] G. M. Fitzmaurice, "Regression diagnostic histopathology," *Diagnostic Histopathology*, vol. 22, no. 7, 2016.
- [4] T. Ayer *et al.*, "Comparison of logistic regression and artificial neural network models in breast cancer risk estimation," *Radiology*, vol. 30, no. 1, pp. 13–22, 2010.
- [5] K. Ashfaq Ahmed *et al.*, "Comparative prediction performance with support vector machine and random forest classification techniques," *International Journal of Computer Applications*, vol. 69, no. 11, pp. 0975–8887, 2013.
- [6] H. Byun and S.-W. Lee, "A survey on pattern recognition applications of support vector machines," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, no. 3, pp. 459–486, 2003.
- [7] R. A. Collazo *et al.*, "A comparative study between artificial neural network and support vector machine for acute coronary syndrome prognosis," *Pesquisa Operacional*, vol. 36, no. 2, 2016.
- [8] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, pp. 783–789, 1999.
- [9] E. A. Zanaty and A. Afifi, "Support vector machines (SVMs) with universal kernels," *Applied Artificial Intelligence*, vol. 25, pp. 575–589, 2011.
- [10] M. Ciecholewski, "Ischemic heart disease detection using selected machine learning methods," *International Journal of Computer Mathematics*, vol. 90, no. 8, pp. 1734–1759, 2013.
- [11] Babaoglu, O. Findik, and E. Ulker, "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine," *Expert Systems with Applications*, vol. 37, pp. 3177–3183, 2010.
- [12] D. Conforti and R. Guido, "Kernel based support vector machine via semidefinite programming: Application to medical diagnosis," *Computers & Operations Research*, vol. 37, pp. 1389–1394, 2010.
- [13] H. I. Elshazly, A. M. Elkorany, and A. E. Hassanien, "Lymph diseases diagnosis approach based on support vector machines with different kernel functions," in *Proc. IEEE International Conference on Computer Engineering & Systems (ICCES)*, 2014.
- [14] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, pp. 3240–3247, 2009.
- [15] J. Xie and C. Wang, "Using support vector machines with a novel hybrid feature-selection method for diagnosis of erythematous-squamous diseases," *Expert Systems with Applications*, vol. 38, pp. 5809–5815, 2011.
- [16] H. X. Liu, R. S. Zhang, F. Luan, X. J. Yao, M. C. Liu, Z. D. Hu, and B. T. Fan, "Diagnosing breast cancer based on support vector machines," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 900–907, 2003.
- [17] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "SVM and SVM ensembles

- in breast cancer prediction,” *PLoS ONE*, vol. 12, no. 1, 2017.
- [18] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: John Wiley & Sons, 1998.
 - [19] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, “Evolving support vector machines using fruit fly optimization for medical data classification,” *Knowledge-Based Systems*, vol. 96, pp. 61–75, 2016.
 - [20] Y. Luo, Z. Li, H. Guo, H. Cao, C. Song, X. Guo, and Y. Zhang, “Predicting congenital heart defects: A comparison of three data mining methods,” *PLoS ONE*, vol. 12, no. 5, 2017.
 - [21] S. Amari and S. Wu, “Improving support vector machine classifiers by modifying kernel functions,” *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.
 - [22] K. Mandal, S. Dehuri, and P. K. D. Sarma, “Analysis of machine learning approaches for predictive modeling in heart disease detection systems,” *Biomedical Signal Processing and Control*, vol. 106, p. 107723, 2025.
 - [23] W. Zhou, H. Liu, R. Zhou, J. Li, and S. Ahmadi, “An optimal method for diagnosing heart disease using combination of grasshopper evolutionary algorithm and support vector machines,” *Heliyon*, vol. 10, p. e30363, 2024.