Data Analytics in Healthcare

A Project Report
Submitted in partial fulfillment of
the requirements for the degree of
Master of Technology
by

Arun R
IE615 Project Report
(163190013)

Industrial Engineering and Operations Research
Indian Institute of Technology Bombay
Mumbai 400076 (India)

24 November 2017

Table of Contents

List of Figures

List of Tables

0.1 Introduction

The amount of data produced by healthcare industries continue to grow exponentially. Bulk of the data comes from electronic health care, pharmacy, insurance claim, human tracking system and diagnostic instruments. This data can be leveraged using data analytics to provide better treatment to patients and reduce the operations cost. Healthcare data analytics can be used to,

- Diagnose disease
- Plan for disaster
- Understand patient flow
- Effectively manage resources and cost
- Reduce fraud

In this report, we will be discussing the various machine learning techniques that are proposed in literature to diagnose breast cancer and diabetes. We will also review a machine learning methodology which is used to predict the chances of readmission in hospital.

0.2 Support vector machines (SVM) combined with feature selection for breast cancer diagnosis

0.2.1 Aim

To diagnose breast cancer using a procedure which combines feature selection and SVM [1]. Experiments were conducted on the Wisconsin breast cancer dataset (WBCD). The performance of the procedure is evaluated using classification accuracy, significance, specificity, positive predictive values, negative predictive values and receiving operating characteristic (ROC) curves.

0.2.2 Procedure

WBCD consists of nine features: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nucleoi, bland chromatin, normal nuclei and mitoses. The breast cancer diagnosis procedure proposed by the authors consist of two steps. In the first step, features are selected, and in the second step, SVM-based prediction model is built using the selected features. The two steps are discussed in the following two sections.

Feature selection

F-score (ref equn. 1) of a feature conveys the importance of the feature in the model. Higher the F-score, higher the importance of the feature. Hence, the procedure proposed in this paper uses the features with higher F-score to build a prediction model.

$$F_i = \frac{(\bar{x}_i^+ - \bar{x}_i)^2 + (\bar{x}_i^- - \bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^+ - \bar{x}_i^+)^2 + \frac{1}{n_--1}\sum_{k=1}^{n_-}(x_{k,i}^- - \bar{x}_i^-)^2} \tag{1}$$

where F_i is the F-score of i^{th} feature; $\bar{x}_i$, $\bar{x}_i^+$, $\bar{x}_i^-$ are the averages of the i^{th} feature of the whole, positive and negative datasets, respectively; $x_{k,i}^+$ and $x_{k,i}^-$ are the i^{th} feature of k^{th} positive and negative instance, respectively.

SVM model

SVM with RBF kernel is used to build the predictive model for breast cancer diagnosis. 10-fold cross validation is used to find the best value for RBF parameters. For each $k \in \{1, 2, \cdots, m\}$, where m is the number of features, a SVM model is built using the top k features with highest F-score. Out of the m models, the one with highest accuracy is used for prediction.

0.3 Predictive modeling of hospital readmissions using metaheuristics and data mining

0.3.1 Aim

To predict the risk of hospital readmission using a method which combines particle swarm optimization (PSO) and SVM [2]. Performance of PSO with SVM is compared with other data mining techniques such as random forest and neural networks using classification accuracy, specificity and sensitivity as metrics. Experiments were done on a heart failure (HF) medical record dataset.

0.3.2 Procedure

The dataset consist of 1641 instances and 9 features: patient age, length of stay, admission acute, comorbidity index score, use of emergency rooms, gender, MS-DRG, patient readmission risk and insurance payer.

Data imbalance problem

Out of the 1641 HF patients, 316 of them is readmitted in hospital within 30 days. Clearly, the two classes are imbalanced. To address this problem, random oversampling technique, a sampling technique which adds randomly chosen instances from the underrepresented class, is used.

Particle swarm optimization based SVM

SVM with RBF kernel is used to build the predictive model. To chose the best value for RBF parameters, particle swarm optimization, a metaheuristic search algorithm which is based on the social behavior of flocks, is used.

0.4 Predicting diabetes mellitus using SMOTE and ensemble machine learning approach

0.4.1 Aim

To compare the performance of various machine learning (ML) techniques in predicting diabetes mellitus [3]. The dataset used in experiments is obtained from Henry Ford Health Systems.

0.4.2 Procedure

The dataset consist of 32,555 patients and 62 feature. The 62 features are broadly classified into four categories: demographic characteristics, disease history, medication use history and exercise test data.

Feature selection and data imbalance problem

From the 62 features, 26 features are manually selected based on their importance. Further, from these 26 features, 13 features with the highest information gain value is chosen for building the ML models. The information gain of each feature is computed using WEKA software. Like the dataset discussed in previous section, this dataset also has data imbalance problem (out of 32,555 patients only 5,099 were diagnosed with diabetes). To address this problem, synthetic minority oversampling technique (SMOTE) is used.

ML Classification Models

The authors study the diabetis prediction of various machine learning techniques such as Decision Trees, Logistic Regression, Naive Bayes and Random Forest. Apart from these methods, an ensemble vote method which consist of three decision trees (Naive Bayes, Random Forest and Logistic Model Tree) is also used to build the prediction model. Performance of each model is validated using 10-fold cross validation.

0.5 Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms

0.5.1 Aim

To diagnose breast cancer using a procedure which combines K-means and SVM [4]. The SVM model classifies a given tumor data as benign or malignant. Since the computation time required to build the SVM model depends on the number of features, K-means is used to extract features from the original dataset which has a large number of features.

0.5.2 Procedure

Experiments were done on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The WDBC dataset consist of 30 features in 10 categories for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. In each category three values are measured: mean value, standard error and maximum value. Te dataset consist of 569 instances out of which 212 were diagnosed with malignant tumor. Since the number of features is large, constructing a SVM predictor model will take a long time. Hence, before building the model, K-means is used to extract features from the dataset. The extraction procedure is explained in next section. Table 1 lists the notations used in this work.

Feature extraction

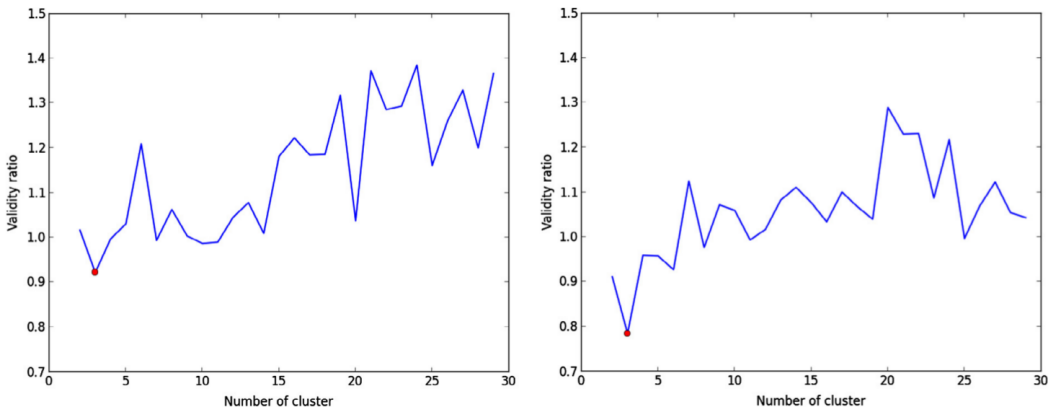
The feature extraction procedure starts by normalizing the data. After normalization, K-means is applied separately on the benign and malignant datasets to cluster the data. Validity ratio (see equn. 2) is used to find the optimal number of clusters. Lower the validity ratio, better is the quality of clusters.

Notation	Definition
K	Number of clusters
F	Number of features in original dataset
S_c/S_k	Set of points in c^{th}/k^{th} cluster
X^i	i^{th} input in dataset
X^i_j	j^{th} feature in i^{th} input
X^{μ_k}	Center of k^{th} cluster
$X^{\mu_k}_j$	j^{th} feature of center of k^{th} cluster

Table 1: Notations and Definitions

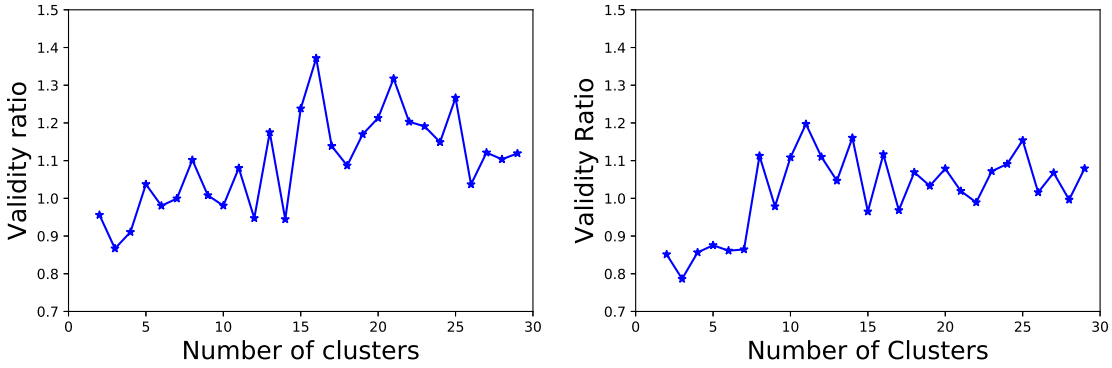
$$\text{Validity Ratio} = \frac{d_{avg}}{d_{min}} \tag{2}$$

where $d_{avg} = \frac{\sum_{k=1}^{K} \sum_{i \in S_k} \sqrt{\sum_{j=1}^{F} (X^i_j - X^{\mu_k}_j)}}{N}$ is the average distance between the data points and their cluster center. $d_{min} = \min \left[\sum_{j=1}^{F} \sqrt{(X^{\mu_{k_2}}_j - X^{\mu_{k_2}}_j)^2} \right] \forall k_1 \neq k_2$ is the minimum distance between two cluster centers. Figures 1a and 1b are the validity ratio variation figures reported by the authors for benign and malignant tumors, respectively. Clearly, the validity is low for both the classes when $K = 3$. We now present the results that we got in our experiments. Figures 2a and 2b show the validity ratio variation for benign and malignant and tumors, respectively. It is evident that the results that we got from our experiments does not match the results reported. However, the range and trend of the plots are almost the same. Even from our experiments, we observe that the validity ratio is low when $K = 3$. The difference in results could be due to random initialization of cluster centers in K-means.



(a) Benign tumors (b) Malignant tumors

Figure 1: Variation of validity ratio with number of clusters (original result)



(a) Benign tumors (b) Malignant tumors

Figure 2: Variation of validity ratio with number of clusters (new result)

Using K-means with $K = 3$ on the two classes of tumors gives six cluster centers. These centers given a symbolic representation for their respective clusters. A given tumor instance is compared with these cluster centers (as shown below) to extract six features.

$$f_c(X_j^i) = \begin{cases} 1 - \frac{|X_j^{\mu_c} - X_j^i|}{\max |X_j^{\mu_c} - X_j^n|}, & \text{if } \min(X_j^n) \leq X_j^i \leq \max(X_j^n),\ \forall n \in S_c \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$p_c = \frac{1}{F} \sum_{j=1}^{F} f_c(X_j^i), \quad 1 \leq c \leq K^m + K^b \tag{4}$$

where p_c is the extracted feature value corresponding to c^{th} cluster. K^m and K^b are the number of clusters used in malignant and benign tumor classes, respectively.

SVM model

SVM model is built using the new extracted features. The authors used sigmoid kernel in the SVM model. 10-fold cross validation is used to validate the prediction model.

0.5.3 Results

In this section, we compare the performance of the SVM models with and without feature extraction. Accuracy, specificity, sensitivity, confusion matrix and computation time are used as performance metrics. Linear kernel and sigmoid kernel are used to build the SVM model.

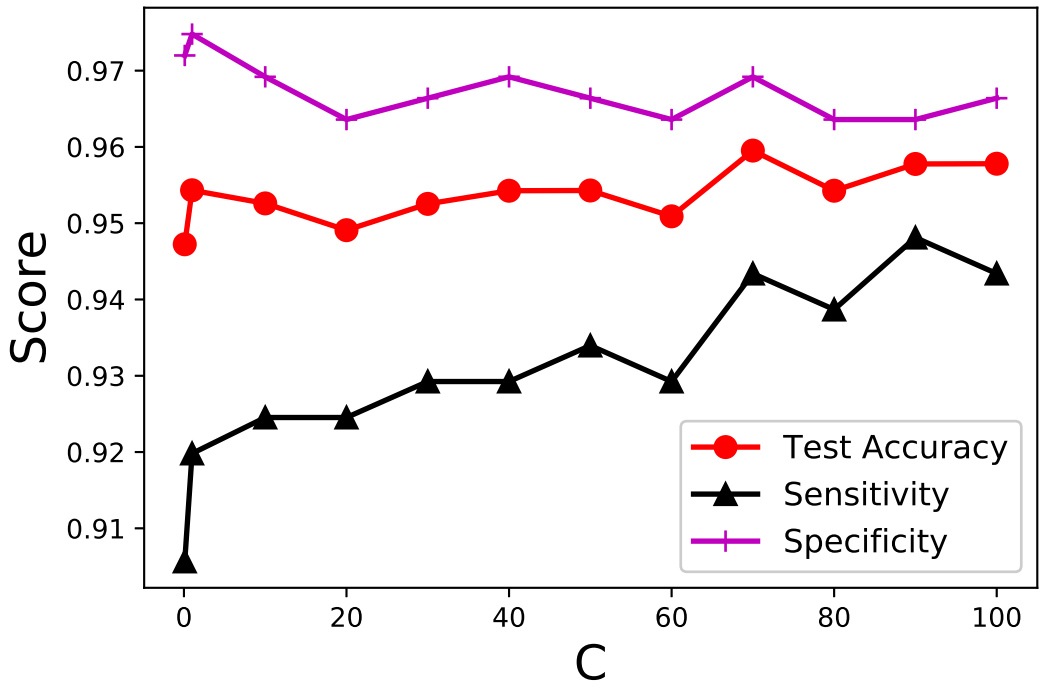


Figure 3: Variation of accuracy, sensitivity and specificity with change in SVM penalty parameter (C) for linear kernel SVM model built with all the 30 features

SVM with linear kernel

Figure 3 shows the variation of test accuracy with the SVM penalty parameter (C) for linear kernel SVM model which is build with all the 30 features. The figure also shows the variation of sensitivity and specificity. Clearly, the accuracy is high when C is 70. The confusion matrix corresponding to this C value is shown in figure 4. The accuracy value for C=70 is 0.96.

Next we show the performance of SVM model built with the 6 features extracted through K-means. Figure 5 shows the variation of test accuracy, significance and sensitivity with the SVM penalty parameter (C). The highest accuracy (0.81) is achieved when C=80. The confusion matrix for this C value is shown in figure 6. As expected, accuracy of the SVM model built with extracted features is less when compared to the accuracy when all the 30 features are used. However, the time required (see table 2) to build the SVM model with extracted features is less when compared with the time taken to build the model with all the 30 features.

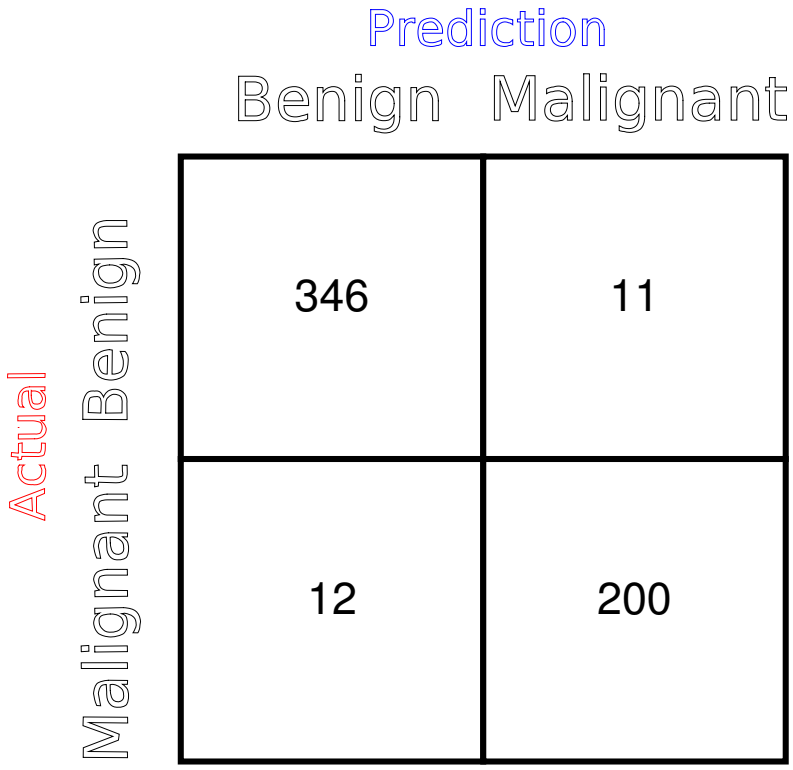


Figure 4: Confusion matrix for linear kernel SVM model built with all the 30 features and C = 70

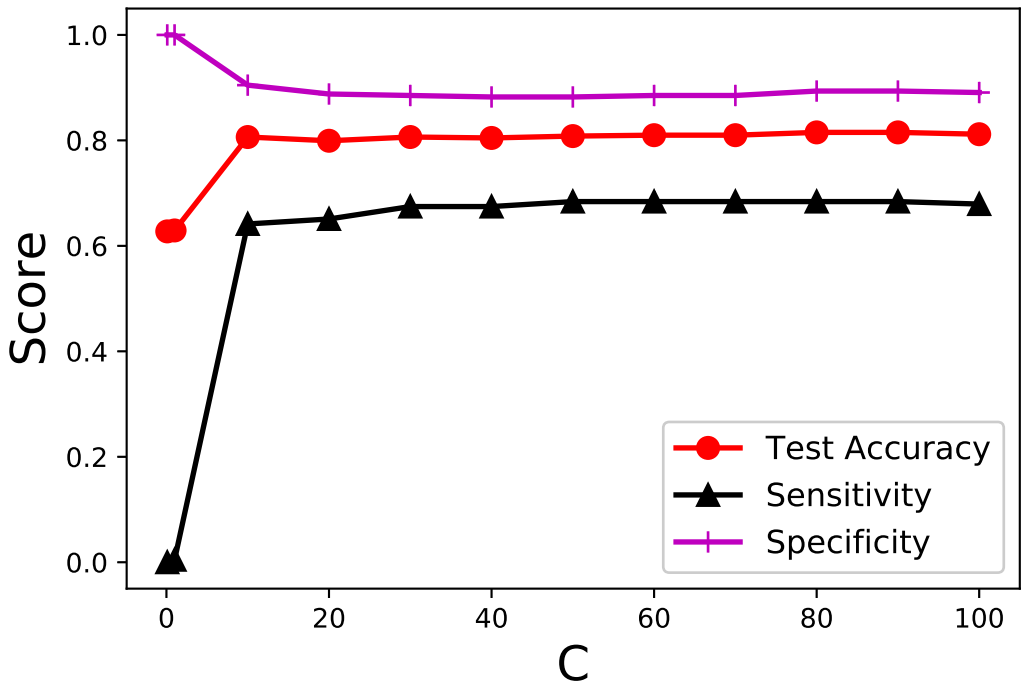


Figure 5: Variation of accuracy, sensitivity and specificity with change in SVM penalty parameter (C) for linear kernel SVM model built with the 6 extracted features

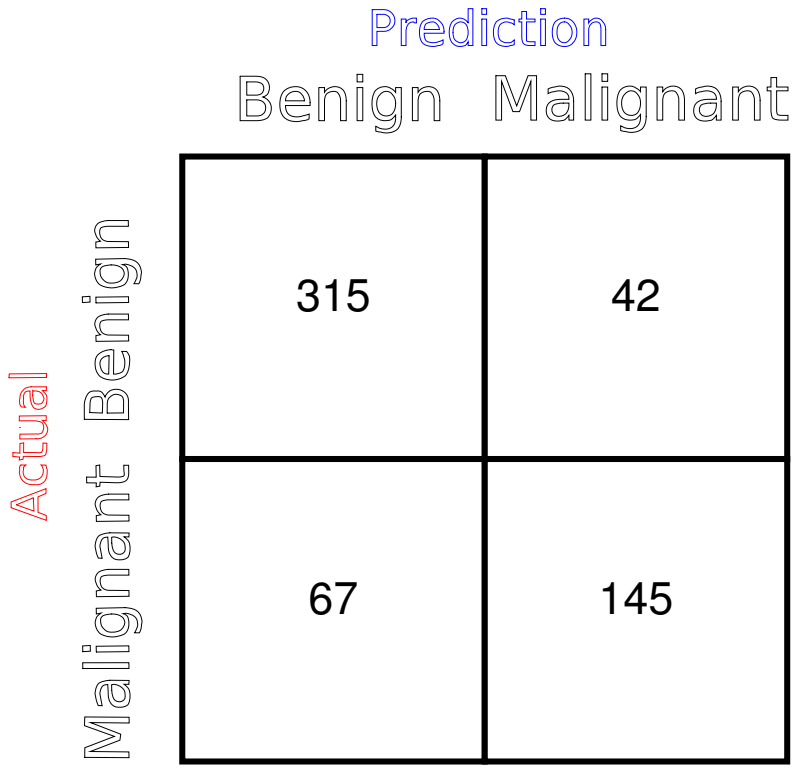


Figure 6: Confusion matrix for linear kernel SVM model built with the 6 extracted features and C = 80

	0.1	1	10	20	30	40	50	60	70	80	90	100
With all 30 features	2.8	11.3	47.9	49.1	64.3	79.2	94.5	67.8	71.8	71.34	74.9	75.9
With the 6 extracted features	0.03	0.04	0.04	0.05	0.04	0.05	0.05	0.06	0.05	0.06	0.06	0.07

Table 2: Computation time (in sec) of linear kernel SVM

SVM with sigmoid kernel

Figure 7 shows the variation of test accuracy with the SVM penalty parameter (C) for sigmoid kernel SVM model which is build with all the 30 features. The figure also shows the variation of sensitivity and specificity. Irrespective of the C value, the accuracy of the model is same. The confusion matrix of the model is shown in figure 8. The model predicts all the tumors to be malignant. The accuracy of this model is 0.63. Setting a higher C value could increase the accuracy of the model. Note: The accuracy value did not change when the gamma value of sigmoid function was changed.

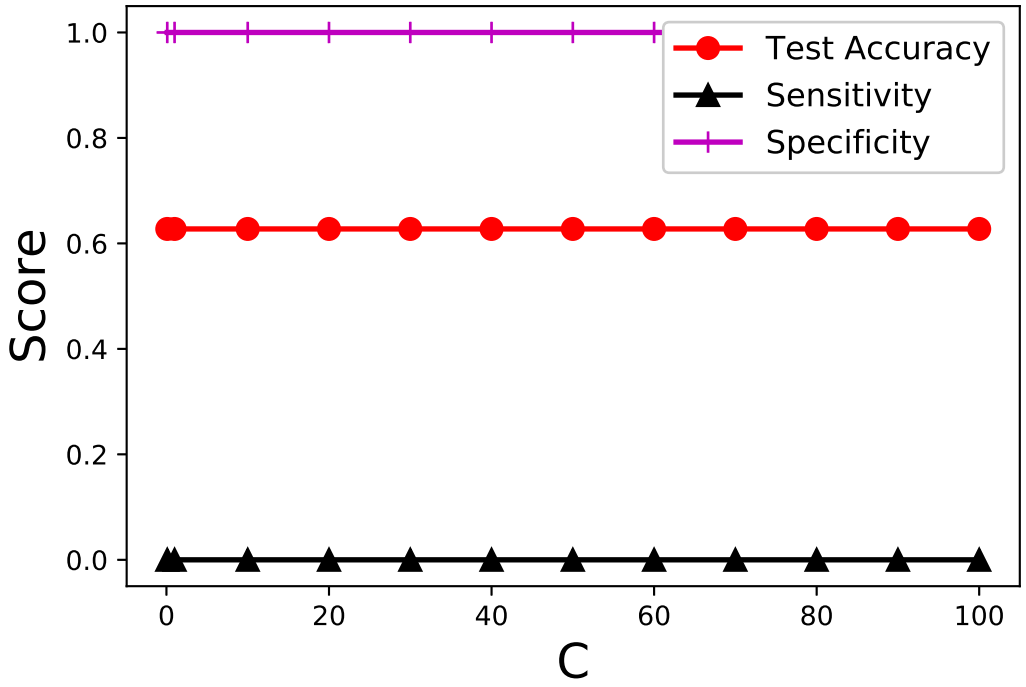


Figure 7: Variation of accuracy, sensitivity and specificity with change in SVM penalty parameter (C) for sigmoid kernel SVM model built with all the 30 features

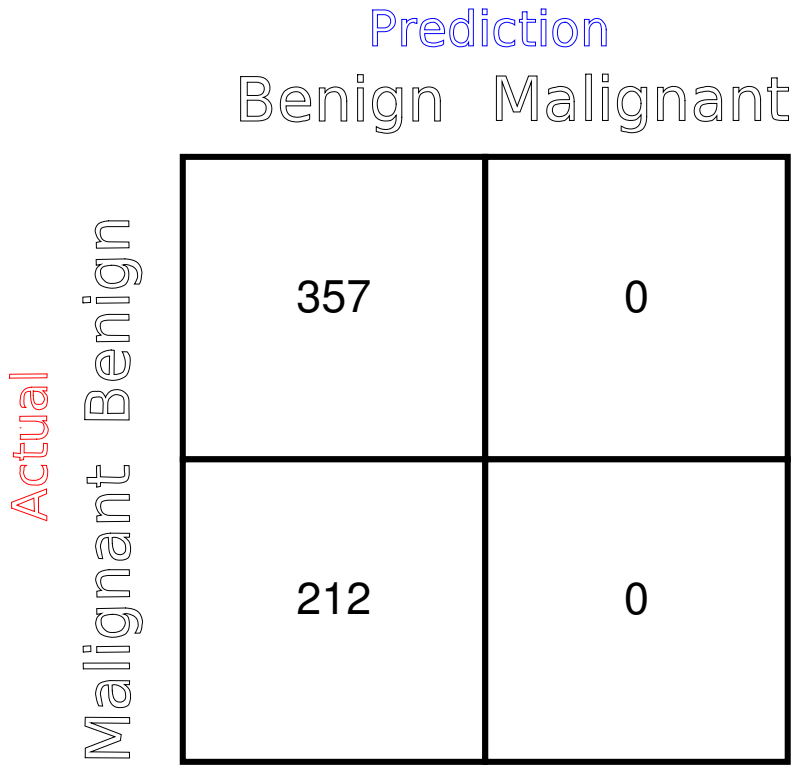


Figure 8: Confusion matrix for sigmoid kernel SVM model built with all the 30 features

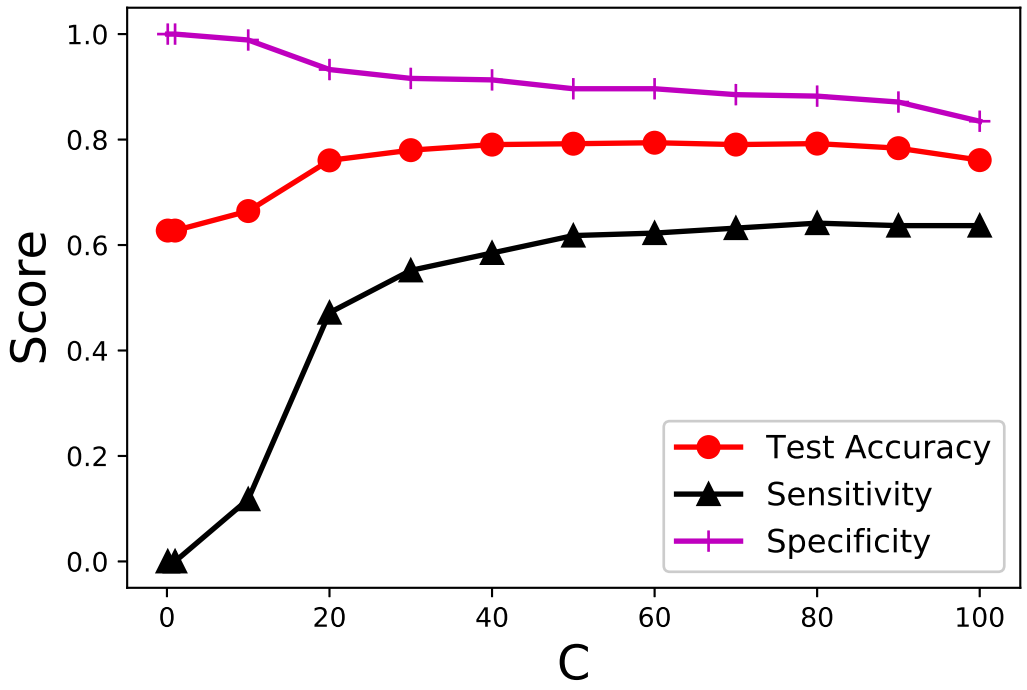


Figure 9: Variation of accuracy, sensitivity and specificity with change in SVM penalty parameter (C) for sigmoid kernel SVM model built with the 6 extracted features

Next we show the performance of SVM model built with the 6 features extracted through K-means. Figure 9 shows the variation of test accuracy, significance and sensitivity with the SVM penalty parameter (C). The highest accuracy (0.79) is achieved when C=80. The confusion matrix for this C value is shown in figure 6. Clearly, the accuracy of the SVM model built with extracted features is more when compared to the accuracy when all the 30 features are used.

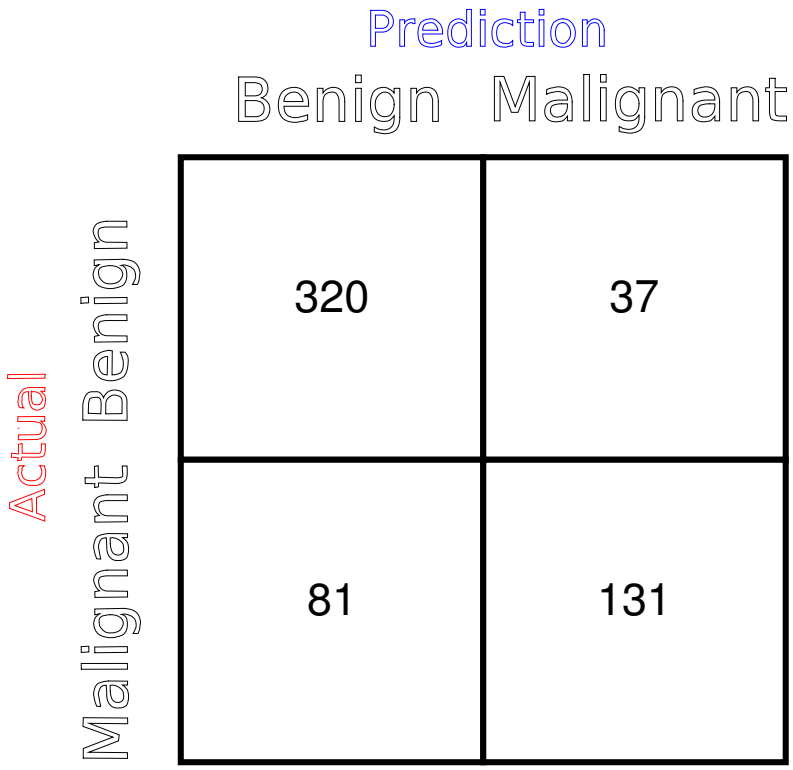


Figure 10: Confusion matrix for sigmoid kernel SVM model built with the 6 extracted features and C = 80

References

[1] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, 36(2):3240–3247, 2009.

[2] Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S Lam, Mohammad Khasawneh, and Srikanth Poranki. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 42(20):7110–7120, 2015.

[3] Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian, Clinton Brawner, Jonathan Ehrman, and Sherif Sakr. Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project. *PLoS One*, 12(7):e0179805, 2017.

[4] Bichen Zheng, Sang Won Yoon, and Sarah S Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482, 2014.