Ranking Predictors of Complications following a Drug Eluting Stent Procedure Using Support Vector Machines

RK Gouripeddi[1], VN Balasubramanian[1], S Panchanathan[1], J Harris[2], A Bhaskaran[2], RM Siegel[2]

[1]Center for Cognitive Ubiquitous Computing, Arizona State University, Tempe, AZ, USA
[2]Advanced Cardiac Specialists, Gilbert, AZ, USA

Abstract

Predictive and risk stratification models using machine learning algorithms such as Support Vector Machines (SVMs), have been used in cardiology and medicine to improve patient care and prognosis. In this work, we have used SVM based Recursive Feature Elimination (SVM-RFE) methods to select patient attributes/features relevant to the etio-pathogenesis of complications following a drug eluting stent (DES) procedure. With a high dimensional feature space (145 features, in our case), and comparatively few patients, *there is a high risk of 'over-fitting'. Also, for the model to* be clinically relevant, the number of patient features need to be reduced to a manageable number, to be used in patient care. SVM-RFE selects subsets of patient features that have maximal influence on the risk of a complication. In our results, when compared with our initial model with all the 145 features, we obtained better performance of the classifiers with 75 top ranked patient features, a 50% reduction in the original dimensionality of the data space. There was a universal improvement in performance of all SVMs with different kernels and parameters. This method of feature ranking helps to determine the most informative patient features. Use of these relevant features improves the prediction of complications following a DES procedure.

1. Introduction

Drug Eluting Stents (DES) have emerged as the de facto option of Percutaneous Coronary Intervention (PCI) for the treatment of Coronary Artery Disease (CAD), with distinct advantages over bare metal stents [1]. However, as with every new intervention methodology, unanticipated complications following a coronary DES procedure are being observed in recent years. In addition to Major Adverse Cardiac Events (MACE) and procedural complications associated with all PCI procedures, DES procedures have resulted in newer complications such as late Stent Thrombosis, increased incidence of early Stent Thrombosis [2] and late restenosis [3]. These complications have generated significant interest in cardiologists and stent manufactures, and have led to the development of post-PCI risk prediction models. With increasing rates of CAD (considered the leading cause of death in the United States for both men and women), these risk prediction models have the potential to lead to widespread impact on large populations of patients.

In [4], we mentioned some of the existing PCI models. However, most of these models study PCI procedures in general and treat mortality as the only outcome. Additionally, all these studies are based on statistical analysis, and the validity of statistical models to specific patient cases is questionable. In all of these models, risk scores are assigned to a very few patient attributes, and are weighted to obtain a cumulative risk score, which is thresholded to identify the risk for a particular patient. We developed a SVM based predictive model for complications following a DES procedure which gave a high accuracy (94% with a 0.97 AUC (area under ROC curve)) [4].

A known problem in machine learning is the risk of "over-fitting" which arises when the dimensionality of feature space is high and the number of training examples is comparatively small (145 vs. 2312, in our case). In such a situation, a decision function that separates the training data (even a linear decision function) can be found but will perform poorly on test data. Training techniques that use regularization [5] avoid over-fitting of the data to some extent without the requirement of space dimensionality reduction. Support Vector Machines (SVMs) is one such example method. Yet, it has been shown from experimental results that even SVMs benefit from space dimensionality reduction [6].

Many methods have been used to reduce the dimensionality of the feature space and for feature selection [7]. Projecting onto the first few principal directions of the data using principal component analysis is a commonly used method to reduce the dimensionality of the feature space. Here, new features that are linear

combinations of the original features are obtained [8]. The disadvantage of principal component methods is that none of the original input features can be discarded and the linear combinations of original features have no practical importance as their clinical significance or semantic meaning of individual risk features cannot be determined.

Feature ranking techniques are greedy algorithms that avoid the computational burden of exhaustive enumeration and are used for designing classifiers by selecting either the top ranked features or those features that exceed a set threshold based upon the ranking criterion. Multiple univariate classifiers that use correlation methods select features that individually classify the training data. These methods eliminate features that do not contribute to the discrimination (noise) and rank features depending on their correlation or anti-correlation (coefficients). A limitation of correlation based methods is it does not yield compact feature sets as it does not take into account any redundancies between features. Also, complimentary features that do not individually separate the classes but accomplish the separation as a group will be missed. Evaluating how well an individual feature contributes to the classification can produce a simple feature ranking system. Multivariate classifiers such as SVMs that train linear discriminant functions overcome these limitations of correlation based methods as they are optimized to use multiple variables (features) simultaneously. For more details about feature selection refer [6,7,9].

In this work, we ranked the various features that were selected to develop our initial model to predict complications following a DES procedure [4], with SVM – Recursive Feature Elimination (RFE) [6], a type of iterative backward feature elimination [7]. SVMs are known to discover informative patterns and have also been shown to be effective in discovering informative features or attributes [10] with qualitative and quantitative advantages in classification performance [6].

SVM-RFE selects subsets of patient features that have maximal influence on the risk of a complication. Since SVMs employ kernel functions that transform the data into a different feature space, any correlations between attributes are compensated in this process to provide a more accurate ranking of the relevance of the individual patient attributes. The patient features are ranked by iteratively training the SVM, computing a ranking criterion for all the features based on the weights of the support vectors, and eliminating those features with the smallest criterion.

Feature selection apart from reducing the risk of 'over-fitting', it also makes the model more clinically relevant by reducing the patient features to a manageable number and making it fit for use in patient care. Section 2 details the data setup and the methods used. Section 3 discusses the results obtained with SVM-RFE, and describes the effectiveness of RFE in selecting features for the model. Section 4 lists the inferences obtained from this work, and its future directions.

2. Methods

2.1. Data setup

Data was obtained from the central PCI registry maintained at a cardiac care facility in Arizona. 2312 patient cases who had a DES procedure performed during the period 2003–2007, and who had followed up with the cardiac care facility during the 12 months following the procedure, were selected from the PCI registry as the dataset for the development of the model. The complications considered for this model included: Stent Thrombosis and Restenosis, which manifest as chest pain, myocardial infection and sometimes even death. Patients with these complications need to be treated with a repeat PCI procedure or Coronary Artery Bypass Graft (CABG). The dataset was extracted as a Comma Separated Value (CSV) format file from the PCI registry which was maintained in SPSS. All patient particulars

CATEGORY	ATTRIBUTES
Clinical & Presentation	Age, Gender, Ejection Fraction, Diabetes, Hypertension, Hyperlipidemia, Smoking, Race, Acute Coronary Syndrome (Acute MI, Unstable Angina), Chronic Stable Angina, Cardiogenic Shock, Congestive Heart Failure, Pulmonary Edema
History	Previous Myocardial Infarction (Acute MI, Silent MI), Unstable Angina, Chronic Stable Angina, Previous PCI, Previous CABG, Previous Stroke, Cardiogenic Shock, Congestive Heart Failure
Angiographic	Vessel, No. of Lesions treated, Bifurcation lesion, Narrowed Coronary Arteries, Multi-vessel Disease, Target Coronary Artery (Left Anterior Descending, Diagonal Left Circumflex, Obtuse Marginal, Right, Right Posterolateral, Right Posterior Descending, Saphenous vein Graft), Coronary Lesion Characteristics (Calcific, Eccentric, Diffuse Disease, Ostial Disease, Total Occlusion, Thrombus), Vessel Tortuosity, Reference Vessel Diameter, Lesion Length, Restenotic lesion, Lesion Type (A, B1, B2, C), Thrombus, Pre-procedure TIMI = 0
Procedural	Urgent / Emergent, Balloon Predilatation (Diameter, Length, Balloon to artery ratio, Maximal Predilatation Inflation Pressure), Stent Implantation (Stent Length, Diameter, 2.25 mm stent, Stent length / Lesion length ratio, Maximal Stent balloon inflation pressure) Postprocedure TIMI flow < 3, Left main Stenting, Multiple stents, Dissection, Acute reocclusion

Table 1. Attributes used in the development of the predictive model

including demographics, clinical parameters, patient history, angiographic, procedural and follow-up details were obtained as available in the registry. All patient data was handled in compliance with the U.S. Food and Drug Administration's (FDA) Protection of Human Subject's Regulations 45 CFR (part 46) and 21 CFR (parts 50 and 56) and the U.S. Department of Health and Human Services Health Insurance Portability and Accountability Act (HIPAA) of 1996, with appropriate approval from the Institutional Review Board (IRB).

All patient attributes available in the PCI registry were used in developing the model. These were matched with the attributes considered in earlier works done on predictors for complications following PCI/DES procedures [4,11-13]. All the 145 variables that were considered, have been listed in Table 1.

The data was cleaned and missing values were handled in the most clinically relevant manner, where appropriate. The data was subsequently normalized. Of the selected patient cases, only 182 had a complication at 12 months following DES. To handle class imbalance (approximately, 92% to 8%) in the patient data, our experiments illustrated the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) [14] to obtain good performance with imbalanced data. All steps of data extraction, pre-processing, and model development were carried out in MATLAB R2007b and Weka. The SVM-KM toolbox [9] was used for the algorithm implementation.

2.2. SVM recursive feature elimination

Linear SVMs are maximum margin classifiers; that is, the decision boundary is positioned to leave the largest margin on either side. Of interest in SVMs are the weights $\mathbf{w_i}$ of the decision function $\mathbf{D(x)}$ are a function of a small subset of examples called "support vectors" [5]. These are examples that are closest to the decision boundary. The concept of support vectors forms the basis of the computational properties of SVMs and their standout classification performance. The resulting decision function of an input vector $\mathbf{x}$ is [5]:

$$\mathbf{D(x) = w \cdot x + b}$$
$$\text{with } w = \Sigma_k \alpha_k y_k x_k \text{ and } b = \langle y_k - \mathbf{w \cdot x_k} \rangle$$

where, the weight vector $\mathbf{w}$ is a linear combination of training patterns. Most weights α_k are zero, those with non-zero weights are support vectors. SVM-RFE is an application of RFE uses the magnitude of these weights as ranking criterion (see Figure. 1).

3. Results

SVM-RFE was used to rank all the features (Figure 2). To determine the significance of the ranking, multiple

Algorithm SVM-RFE:
Inputs:
Training examples
$X_0 = [x_1, x_2, \ldots x_k, \ldots x_l]^T$
Class labels
$y = [y_1, y_2, \ldots y_k, \ldots y_l]^T$
Initialize:
Subset of surviving features
$s = [1, 2, \ldots n]$
Feature ranked list
$r = [\,]$
Repeat until s = [ ]
Restrict training examples to good feature indices
$X = X_0(:, s)$
Train the classifier
α = SVM-train(X, y)
Compute the weight vector of dimension length(s)
$w = \Sigma_k \alpha_k y_k x_k$
Compute the ranking criteria
$c_i = (w_i)^2$, for all i
Find the feature with smallest ranking criterion
f = argmin(c)
Update feature ranked list
r = [s(f), r]
Eliminate the feature with smallest ranking criterion
s = s(1: f − 1, f + 1: length(s))
Output:
Feature ranked list r.

Figure 1. The linear SVM-RFE algorithm [7].

SVM based predictive models were used to predict the risk of complications by recursively eliminating one feature at a time based on the obtained SVM-RFE ranking. The dataset was randomly divided into training and testing data, with 75% and 25% of the case instances respectively. Polynomial and RBF kernels were used to study the performance of the SVM in predicting the risk of complication in patients. A grid search was performed for model parameter selection. The spread function for the RBF kernel function ranged from 2^{-4} to 2^4. The exponent of the polynomial kernel was varied from 1 to 6. The performance of each of models was evaluated using accuracy and AUC. When compared with our initial model with all the 145 features, we obtained better performance of the classifiers with 75 top ranked patient features, a 50% reduction in the original dimensionality of the data space. There was a universal improvement in performance of all SVMs with different kernels and parameters. The models with Gaussian kernel of spread function 2^1 gave the best performance (Figure. 2) when one feature was recursively eliminated based on the SVM-RFE ranking.

4. Discussion and conclusions

SVM-RFE selected ranked clinically more significant attributes higher up. The performance of SVM-RFE demonstrates the utility in using it for clinical feature selection, wherein features cannot be implicitly assumed to be orthogonal to each other. This method of feature ranking helps to determine the most informative patient features. Use of these relevant features improves the prediction of complications following a DES procedure. Further, the high ranked patient features that contribute to the complications can be aggressively controlled and clinically managed in the patient population to prevent the occurrence of a complication. One of the main limitations of this work is that the patient data used for this model was limited to this particular cardiac care facility, and the RFE would need to be re-trained with data from other centers to generalize the ranking. In future work, we wish to evaluate and assess the model with expert interventional cardiologists. While the present work has used linear SVM for RFE, SVMs of different kernels could also be used [9].

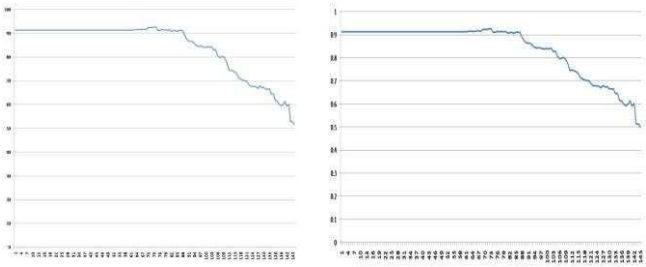

Figure 2. Accuracy (left) and AUC (right) of SVM Models (Gaussian kernel parameter 1) developed with recursive elimination of a feature based on the RFE ranking.

Rank	Feature	Rank	Feature
1	TIMI Grade Flow	14	Lesion Length
2	Pre Procedural Stenosis	15	Prior PTCA
3	Age	16	Balloon Length
4	Cath Time	17	Sudden Death
5	Dye Load	18	Prior ICD
6	Stent Type	19	Balloon Diameter
7	Previous Vessel Treated	20	Smoker
8	Discharge Days	21	Current Diagnosis
9	Race	22	Congestive Heart Failure
10	Fluoro Time	23	Peripheral Vascular Disease
11	Maximum Balloon Pressure	24	Success/ Failure of Procedure
12	Vessel Type	25	Diabetes
13	Number of Lesions		

Figure 3. Top 25 ranked features by SVM-RFE.

References

[1] Stone GW, Ellis SG, Cox DA, Hermiller J, O'Shaughnessy C, Mann JT, A polymer-based, paclitaxel-eluting stent in patients with coronary artery disease. N Engl J Med. 2004 Jan 15;350(3):221-31.

[2] Weisz G, Moses J, Schofer J. A "Liberal" Definition of Stent Thrombosis: Sirolimus-Eluting Stent vs. Bare Metal Stents. Circulation; 2005. p. (Suppl II):II-683.

[3] Stettler C, Wandel S, Allemann S, Kastrati A, Morice MC, Schömig A. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. Lancet. 2007 Sep 15;370(9591):937-48.

[4] Gouripeddi R, Balasubramanian V, Panchanathan S, Harris J, Bhaskaran A, Siegel RM. Predicting risk of complications following a drug eluting stent procedure: A SVM approach for imbalanced data. In: Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on. 2009. p. 1-7.

[5] Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed. Springer; 2000.

[6] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002 Jan 1;46(1):389-422.

[7] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997 Dec ;97(1-2):273-324.

[8] Duda RO, Hart PE. Pattern Classification and Scene Analysis. John Wiley & Sons Inc; 1973.

[9] Rakotomamonjy A. SVM and Kernel Methods Matlab toolbox. Available from: http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html

[10] Guyon I, Matic N, Vapnik V. Discovering informative patterns and data cleaning. In: Advances in knowledge discovery and data mining. American Association for Artificial Intelligence; 1996. p. 181-203.

[11] Godino C, Furuichi S, Latib A, Morici N, Chieffo A, Romagnoli E, et al. Clinical and angiographic follow-up of small vessel lesions treated with paclitaxel-eluting stents (from the TRUE Registry) Am J Cardiol. 2008 Oct 15;102(8):1002-8.

[12] Romagnoli E, Burzotta F, Trani C, Siviglia M, Biondi-Zoccai GGL, Niccoli G, et al. Euroscore as predictor of in-hospital mortality after percutaneous coronary intervention. Heart. 2008 Jan 20;hrt.2007.134114.

[13] Biondi-Zoccai GG, Agostoni P, Sangiorgi GM, Airoldi F, Cosgrave J, Chieffo A, et al. Incidence, predictors, and outcomes of coronary dissections left untreated after drug-eluting stent implantation. Eur Heart J. 2006 Mar 1;27(5):540-546.

[14] Chawla NV, Hall LO, Bowyer KW, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 ;16321-357.

Address for correspondence:

Ramkiran Gouripeddi
699 S Mill Avenue Suite 370AB
Tempe AZ 85287
rgouripe@asu.edu