Support Vector Machine Based Conformal Predictors for Risk of Complications following a Coronary Drug Eluting Stent Procedure

VN Balasubramanian[1], R Gouripeddi[1], S Panchanathan[1], J Vermillion[2], A Bhaskaran[2], RM Siegel[2]

[1]Center for Cognitive Ubiquitous Computing (CUbiC), School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287, USA
[2]Advanced Cardiac Specialists, Gilbert, AZ 85206, USA

Abstract

Drug Eluting Stents (DES) have distinct advantages over other Percutaneous Coronary Intervention procedures, but have been associated with the development of serious complications after the procedure. There is a growing need for understanding the risk of these complications, which has led to the development of statistical risk evaluation models. Conformal Predictors are a recently developed set of machine learning algorithms that allow not just risk classification on new patients, but add valid measures of confidence in predictions for individual patients. In this work, we have applied a novel Support Vector Machine (SVM) based conformal prediction framework to predict the risk of complications following a coronary DES procedure. This predictive model helps to risk stratify a patient for post-DES complications, and the valid measures of confidence can be used by the physician to make an informed, evidence-based decision to manage the patient appropriately.

1. Introduction

Machine learning algorithms such as Support Vector Machines [1], genetic algorithms [2], and neural networks [3] [4] have been used in cardiology to improve the quality of care, stratify risk, and provide prognostications. Traditional learning algorithms learn from data of past patients, and provide predictions on new patients, without convincing information of the reliability or confidence in the predictions. In medical diagnosis/prognosis, it is extremely essential to evaluate the performance of such algorithms on the risk of possible error in supporting the decision-making process. A new set of machine learning algorithms called Conformal Predictors have been recently developed that, unlike many conventional classification systems, allow us not just to risk classify new patients, but add *valid* measures of confidence in our predictions for every *individual* patient. In this work, we have applied a novel conformal predictor framework based on Support Vector Machines (SVM) to the problem of predicting the risk of complications following a coronary Drug Eluting Stent procedure (DES), using patient data provided by Advanced Cardiac Specialists, a cardiology practice based in Arizona, USA.

Drug Eluting Stents (DES) have emerged as the de facto option for Percutaneous Coronary Intervention (PCI), with distinct advantages over bare metal stents [5]. Since restenosis rates are less than 10% with DES, there has been an explosive growth in their use over a very short period. However, unanticipated complications have been increasingly observed following a coronary DES procedure. In addition to standard Major Adverse Cardiac Events (MACE) and procedural complications associated with all PCI procedures, DES have resulted in additional complications, including late Stent Thrombosis, increased incidence of early Stent Thrombosis, and late restenosis, which could result in myocardial infarction or death.

Existing models in this scope (such as the Boston Scientific DES Thrombosis score [6] and the Mayo Risk score [7]) are rule-based and derived from correlation analysis. The validity of such statistical models to *specific patient cases* is questionable. For example, age ≥ 65 is used as a common patient attribute in such models, and this may be set to zero for a patient with age 64. This increases the possibility of incidence of false positives and false negatives in the predictions, thereby limiting the scope of their applicability. Predictive models based on machine learning techniques have the ability to consider each patient as a unique entity, and predict outcomes for a particular patient case in question, unlike statistical models.

The predictive model proposed in this paper helps to stratify the risk for a specific patient for post-DES complications, and thereby stratify patient populations according

CATEGORY	ATTRIBUTES
Demographic and 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 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

to healthcare requirements, reducing the need for unnecessary invasive procedures with their attendant risks and significant costs. The *valid* measures of confidence can be used by the physician to make an informed, evidence-based decision to manage a patient, choosing the most appropriate option from repeat PCI, Coronary Artery Bypass Graft surgery (CABG), and/or maximized medical therapy to minimize the possibility of occurrence/recurrence.

2. Methods

2.1. Data setup

Data was obtained from the central Percutaneous Coronary Intervention registry maintained at Advanced Cardiac Specialists (ACS), consisting of patient cases across the state of Arizona (including cases of different genders, races and ethnic groups). 2312 patient cases who had a DES procedure performed during the period 2003 to 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 infarction and sometimes even death. All patient particulars including demographics, clinical parameters, patient history, angiographic, procedural and follow-up details (a total of 165 patient attributes) were obtained as available in the registry. These attributes are listed in Table 1. The dataset was extracted as a Comma Separated Value (CSV) format file from the PCI registry which was maintained in SPSS. All patient data was handled in compliance with the U.S. Food and Drug Administration's (FDA) Protection of Human Subjects 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.

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 (only 7.87% of the total data) 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) [8] to obtain good performance with imbalanced data. All steps of data extraction, pre-processing, and model development were carried out in MATLAB R2007b. The SVM-KM toolbox [9] was used for the algorithm implementation.

2.2. SVM-based conformal predictions

Support Vector Machines (SVMs) [10] are algorithmic implementations of statistical learning theory which build consistent models from data to classify newer data into identified categories. We use SVMs in this work to classify patients who have undergone a DES procedure into high-risk and low-risk categories, based on the patient/procedural attributes listed in Table 1. By transforming the data from its original coordinate space to a new space using a kernel function, a non-linear decision boundary in the original space can be converted to a linear boundary in the new transformed feature space. The cumbersome computations of this optimization problem are simplified by means of the kernel trick by which the dot product between data vectors in the transformed space (which measures the similarity) can be expressed in terms of the sim-

ilarity in the original space. For example, the Gaussian kernel function is given by:

$$K(u,v) = e^{\frac{-(\|u-v\|^2)}{2\sigma^2}}$$

where σ, the spread of the Gaussian kernel, is the primary parameter and u and v are the input feature vectors. The polynomial kernel function of degree d is given by:

$$K(u,v) = (u.v + 1)^d$$

The recently introduced theory of conformal predictions (CP) [11] [12] is based on theoretical concepts from algorithmic randomness, transductive inference and hypothesis testing. When classifying a new test instance (patient record), CP assigns a p-value to each possible class label (high-risk and low-risk, in this work) which are used the confidence regions of prediction. More importantly, the performance (confidence level) can be set prior to classification, and the predictions are well-calibrated i.e. the accuracy rate is exactly equal to the preset confidence level. In this paper, we describe how the CP framework can be applied using SVMs.

The CP framework relies on the definition of a non-conformity measure, which captures the extent to which a given patient data conforms to each class (high-risk and low-risk) of patients. This non-conformity measure is defined uniquely for every classification algorithm, and plays an important role in the performance of the framework. Given a kernel function for a SVM, $\phi(x)$, the decision boundary of a kernel-based binary SVM, $w.\phi(x) + b$, and assuming that a patient, x_i, belongs to a particular class label $y_p \in \{-1, +1\}$, we define the non-conformity measure for the binary SVM as follows. The distance of the patient x_i from the separating hyperplane in the SVM is given by:

$$d_i^h = \frac{|y|}{\|w\|}$$

where $|y|$ is the output of the SVM for the patient x_i, and $\|w\|$ is the weighted sum of the support vectors (using the dual formulation of the SVM). Then, the distance to the margin boundary of the class under consideration is given by:

$$d_i^m = \frac{|y| - 1}{\|w\|}$$

since the class labels are assumed to be $\{-1, +1\}$. Now, we define the non-conformity measure in this work as (where a is a parameter that is chosen empirically):

$$\alpha_i^{y_p} = e^{-a * d_i^m}$$

For patient data that is used to train the SVM, the non-conformity measure is computed with respect to its own known class label. For an unseen test patient case, the non-conformity measure is computed with respect to both class labels in the predictive model. A p-value function is then defined as:

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{m}$$

where $\alpha_{n+1}^{y_p}$ is the non-conformity measure of x_{n+1}, assuming it is assigned the class label y_p, and m is the number of data instances that belong to the class label y_p. Based on a preset confidence level $1 - \epsilon$ (which can be chosen by the user depending on the confidence level he/she would like to see in the results under a particular situation), the conformal predictions framework outputs the prediction set containing all labels with p-values greater than ϵ. The *validity* of the confidence values indicates that the obtained accuracy of prediction always corresponds to the given level of confidence up to a statistical fluctuation. For example, if a user set the confidence level of the system at 95%, the algorithm would ensure there are *at least* 95 correct predictions on a set of 100 unseen patients. This property of the framework is practically very valuable when a physician is supported by a risk prediction model in decision-making.

3. Results

The dataset was randomly divided into training and testing data, with 75% and 25% of the case instances respectively. The model was trained only with the training data, and was not exposed to the test data at any stage, until the evaluation of the model. Polynomial and Gaussian kernels were used to study the performance of the SVM in predicting the risk of complication in patients. We achieved an accuracy of 94% on unseen patients with the Gaussian kernel with a spread of 1 (with an Area Under ROC Curve (AUC) = 0.97). The sensitivity and specificity for this model were also high - 0.9271 and 0.9518, respectively. For more details of our results with the SVM classification algorithm, please refer to [8].

Figure 1 presents the results of applying the SVM-based CP framework. Each of the sub-figures demonstrate the performance at a particular preset confidence level. Evidently, the number of errors in each of the trials are always bounded by the specified confidence level. In the CP framework, it is essential to minimize: (i) the number of multiple predictions, i.e., the number of test patient instances when the framework provides more than one class label in the output prediction set; and (ii) the number of empty predictions, where the system does not provide any class label in the output prediction set. As illustrated in Figure 1, the number of multiple and empty predictions increase as the required confidence level increases. However, in this work, even at a high confidence level of 90%,

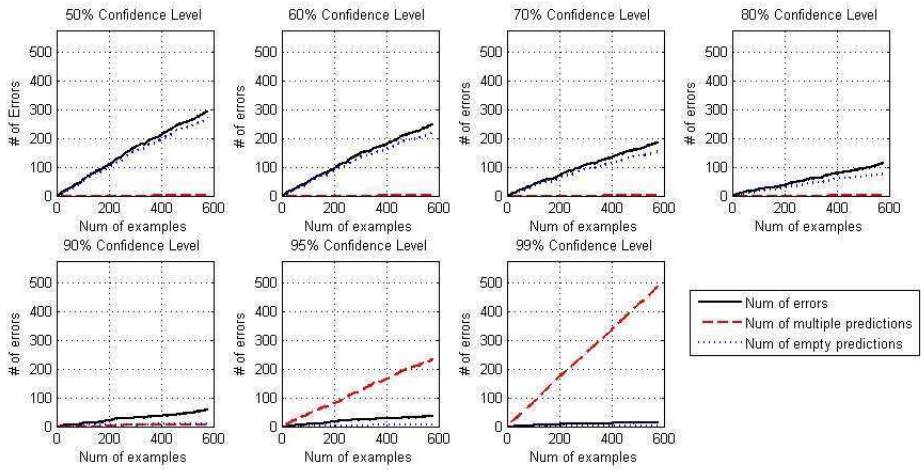

Figure 1. Results of applying the CP framework. Note that for each of the user-defined confidence levels $1-\epsilon$, the number of errors are always consistently lesser than $(100 \times \epsilon)\%$.

we observed a very low number of multiple and empty predictions. The number of empty predictions continued to be low even at the 99% level.

4. Discussion and conclusions

The performance of the presented SVM model on real-world patient data demonstrates the applicability of the proposed framework in providing patient-specific risk stratification. The novel SVM-based conformal predictors not only provided high accuracies, but were also calibrated for their error rates (controlled by the preset confidence level). Such a model can be extremely effective to risk stratify a patient for post-DES complications, and the valid measures of confidence can be used by the physician to make an informed, evidence-based decision to manage the patient appropriately. The proposed approach can also be very valuable in many other predictive models in cardiology and medicine.

References

[1] Matheny ME, Resnic FS, Arora N, Ohno-Machado L. Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality. J of Biomedical Informatics 2007;40(6):688–697.

[2] Vinterbo S, Ohno-Machado L. A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. AMIA Symposium 1999;984–988.

[3] Harrison RF, Kennedy RL. Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Annals of Emergency Medicine November 2005;46(5):431–439.

[4] Ohno-Machado L, Musen MA. Sequential versus standard neural networks for pattern recognition: an example using the domain of coronary heart disease. Computers in Biology and Medicine July 1997;27(4):267–281.

[5] Stettler C, Wandel S, Allemann S, et al. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. The Lancet September 2007; 370(9591):937–948. ISSN 0140-6736.

[6] Baran KW, Lasala JM, Cox DA, Song A, Deshpande MC, Jacoski MV, Mascioli SR. A clinical risk score for prediction of stent thrombosis. The American Journal of Cardiology September 2008;102(5):541–545.

[7] Singh M, Lennon RJ, Holmes DR, Bell MR, Rihal CS. Correlates of procedural complications and a simple integer risk score for percutaneous coronary intervention. Journal of the American College of Cardiology August 2002; 40(3):387–393.

[8] Gouripeddi R, Balasubramanian V, Harris J, Bhaskaran A, Siegel R, Panachanathan S. Predicting risk of complications following a drug eluting stent procedure: a svm approach for imbalanced data. 22nd IEEE International Symposium on Computer Based Medical Systems Aug 2009;984–988.

[9] Canu S, Grandvalet Y, Guigue V, Rakotomamonjy A. Svm and kernel methods matlab toolbox. Perception Systmes et Information, INSA de Rouen, Rouen, France, 2005.

[10] Vapnik V. The Nature of Statistical Learning Theory. 2nd edition. Springer, November 1999. ISBN 0387987800.

[11] Shafer G, Vovk V. A tutorial on conformal prediction. J Mach Learn Res 2008;9:371–421.

[12] Vovk V, Gammerman A, Shafer G. Algorithmic Learning in a Random World. 1st edition. Springer, March 2005.

Address for correspondence:

Vineeth Nallure Balasubramanian
699 S Mill Avenue Suite 380
Tempe AZ 85287
vineeth.nb@asu.edu