A mixed-ensemble model for hospital readmission

Article *in* Artificial intelligence in medicine · August 2016

DOI: 10.1016/j.artmed.2016.08.005

2 authors, including:

Lior Turgeman
IBM
17 PUBLICATIONS **102** CITATIONS

SEE PROFILE

Artificial Intelligence in Medicine 72 (2016) 72–82

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A mixed-ensemble model for hospital readmission

Lior Turgeman *, Jerrold H. May

Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, United States

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form 20 July 2016
Accepted 30 August 2016

Keywords:
Decision trees
Support vector machine (SVM)
Ensemble learning
Imbalanced data set
Decision function
Error reduction
Hospital readmission

ABSTRACT

Objective: A hospital readmission is defined as an admission to a hospital within a certain time frame, typically thirty days, following a previous discharge, either to the same or to a different hospital. Because most patients are not readmitted, the readmission classification problem is highly imbalanced.
Materials and methods: We developed a hospital readmission predictive model, which enables controlling the tradeoff between reasoning transparency and predictive accuracy, by taking into account the unique characteristics of the learned database. A boosted C5.0 tree, as the base classifier, was ensembled with a support vector machine (SVM), as a secondary classifier. The models were induced and validated using anonymized administrative records of 20,321 inpatient admissions, of 4840 Congestive Heart Failure (CHF) patients, at the Veterans Health Administration (VHA) hospitals in Pittsburgh, from fiscal years (FY) 2006 through 2014.
Results: The SVM predictions are characterized by greater sensitivity values (true positive rates) than are the C5.0 predictions, for a wider range of cut off values of the ROC curve, depending on a predefined confidence threshold for the base C5.0 classifier. The total accuracy for the ensemble ranges from 81% to 85%. Different predictors, including comorbidities, lab values, and vitals, play different roles in the two models.
Conclusions: The mixed-ensemble model enables easy and fast exploratory knowledge discovery of the database, and a control of the classification error for positive readmission instances. Implementation of this ensembling method for predicting all-cause hospital readmissions of CHF patients allows overcoming some of the limitations of the classifiers considered individually, and of other traditional ensembling methods. It also increases the classification accuracy for positive readmission instances, particularly when strong predictors are not available.

1. Introduction

A hospital readmission is defined as an admission to a hospital within a certain time frame, following an original hospital discharge, either to the same or to a different hospital. The Congestive Heart Failure (CHF) diagnosis includes some of the highest percentages of patients who are readmitted to a hospital within thirty days of discharge [1–4], and is the leading cause of hospital admissions among patients over the age of 65 years [5]. CHF is also associated with high rates of mortality and morbidity [6]. Several previous papers used logistic regression to estimate the probability of hospital readmissions [7–10]. Another type of baseline model uses survival analysis (or hazard models) to estimate the time duration between consecutive patient admissions [11,12]. Although both approaches are useful in identifying readmission risk factors, they are not as useful for dealing with the non-stationary nature of patient readmissions, where the readmission propensity might change over time, depending on different conditions and treatments during prior admissions [13]. Also, most approaches are characterized by limited classification power when a large range of variables is considered.

A variety of reasons could lead to readmissions, such as early discharge of patients, improper discharge planning, and poor care transition [14–16]. Vinson et al. [4] found that the factors predictive of readmission of CHF patients included a prior history of heart failure, four or more admissions within the preceding eight years, and heart failure precipitated by an acute myocardial infarction or uncontrolled hypertension. Using subjective criteria, they indicated that factors contributing to preventable readmissions included non-compliance with medications or diet, inadequate discharge planning or follow-up, a failed social support system, and failure to seek medical attention promptly when symptoms recurred. Schwartz et al. [17] studied the severity of cardiac illness, cognitive functioning, and functional health of 156 patients within

* Corresponding author.
E-mail address: Tur.lior@gmail.com (L. Turgeman).

http://dx.doi.org/10.1016/j.artmed.2016.08.005
0933-3657/

seven to ten days of a patient's discharge from the hospital. They found that 44% of the patients were re-hospitalized during a three-month period. A patient's severity of cardiac illness, functional health status, and caregiver psychosocial and informal support factors influenced hospital readmissions during the three months post hospital discharge. Key predictors of readmissions in their regression analysis were the interaction of the severity of patient cardiac illness with functional status, the interaction of depression with stress, and informal social support. He et al. [18] developed an administrative claim-based algorithm to predict 30-day readmission using standardized billing codes and basic admission characteristics available before discharge. The algorithm works by exploiting high-dimensional information in administrative claims data, and performing logistic regression on the selected attributes. Shipeng et al. [19] experimented with a general framework for hospital-specific and condition-specific models for readmission risk prediction, by implementing a Support Vector Machine (SVM) as a classification approach, and Cox Regression as a prognostic approach. The institution-specific readmission risk prediction framework has been shown to have flexibility and to be effective, as compared to one-size-fit-all models. For a systematic review of statistical models and predictors for CHF patients' hospital readmissions, see [20,21]. For a survey of previous research on CHF predictive factors, we refer the reader to [22] and [23].

However, most readmission models do not provide clinically useful, interpretable rules that could explain the reasoning process behind their predictions. They typically only produce a score that describes the chance of readmission, based on the values of the predictors. Providing at least some level of explanation for the reasons behind a prediction may assist healthcare providers in their decisions to administer patient-specific, targeted interventions, in order to reduce patients' chances of readmission, as well as to facilitate proper resource utilization by the hospital. While actionable insights could be extracted from decision trees, trees may perform less well than other more sophisticated tools [24,25]. Furthermore, for datasets that are noisy, inconsistent, skewed, and have a significant amount of missing values, tree models may not be flexible enough to produce consistent predictions [24]. Other models that are characterized by greater classification accuracy, such as logistic regression based models [7–10], may not enable physicians, and clinical staff, to interpret the results in the context of their existing knowledge.

As discussed by Shmueli and Koppius [10], the design of a predictive analytics model involves a trade-off between its predictive power and its explanatory transparency. In this paper, we present a new way to combine two different types of data mining models, to address a challenging clinical predictive task, while supporting the goal of effective communication of the reasoning underlying a prediction. The ensemble model described in this paper integrates a boosted C5.0 tree model, the principal classifier, with a complementary Support Vector Machine (SVM) model as a secondary component. The idea of ensemble predictive modeling is to apply a collection of independent models for predicting a class for a case, rather than basing the prediction on a single model. Ensembles typically produce a consensus prediction, using a majority vote of the members of the ensemble. In such an approach, each model in the ensemble addresses the same original task, but in a different way. The motivation is that a composite model will produce more accurate and reliable decisions than would be obtained from a single model [26]. For example, random forests induce multiple trees, using different subsets of the given input variables, and combine the results by majority vote. Previous empirical work for ensembling has focused primarily on the total predictive error reduction from using multiple models, or on the exploration of novel methods for generating models and combining their predicted classes [27]. Along with simple combiners, there exist more sophisticated methods, such as stacking [28] and arbitration [29]. Ali and Pazzani et al. [27] use an empirical analysis to understand the reduction in generalization error that results from using multiple learning models. By comparing several combination methods, such as uniform voting, Bayesian combination, distribution summation, and likelihood combination, they show that the amount of observed error reduction is negatively related to the degree of correlation of errors in the individual models [30,31]. In particular, they found that when the limiting factor is not the noise or difficulty of the data, the multiple models approach provides an excellent way to achieve a large error reduction.

One situation in which the limiting factor is not noise or difficulty is where the data set includes many irrelevant attributes. As the number of irrelevant attributes is increased, the ensemble models approach does increasingly better than does a single model. However, beyond some point, adding irrelevant attributes reduces the accuracy of the model. Thus, when building a predictive model for a data set that has a large number of variables, the tradeoff between the error correlations between pairs of the included models, and the number of variables in those models, is the one that would most likely influence the amount of observed error reduction that would be achieved through ensembling. On the other hand, the degree to which the error patterns of models are correlated is related to the ability of those models to provide accurate predictions without having prior domain knowledge, and to how they deal with noisy labels. Reducing the amount of error is also a function of the unique characteristics of each included model. An additional complication is the degree of imbalance in class size in the data set. When the number of instances in the minority class is significantly less than the number of instances in the majority class, i.e. the data set is imbalanced [32], the amount of error for the default classifier, due to a skewed distribution of data, or lack of information, is increased. Common approaches for dealing with imbalanced data involve modifications either to the data distribution or to the classifier itself [24,25]. Considering any of those techniques, for the purpose of building an ensemble classifier, should also take into account the influence it has on the predictive power of each included model, in terms of the total error reduction.

In this paper, we suggest combining both approaches. The base classifier, a boosted C5.0 decision tree, is able to handle high-dimensional data, and its representation of the knowledge induced from the data is highly intuitive. By using a boosting algorithm, we obtain an ensemble classifier that exhibits less classification error that would a single C5.0 classifier. C5.0 associates, with each prediction, a confidence level that quantifies its trustworthiness. Based on the dataset characteristics, we define a confidence threshold such that records for which the C5.0 tree reports an insufficient confidence are further analyzed by an SVM model. The SVM is characterized with strong power in evaluating information in the case of non-regularity in the data, and is able to provide good out-of-sample generalization. Based on the considerations above, we suggest an optimization approach for the mixed-ensemble model which takes into account the unique characteristics of the learned dataset.

2. Materials and methods

2.1. Data

We were granted access to the anonymized administrative records of 20,321 inpatient admissions, of 4840 patients, to Veterans Health Administration (VHA) hospitals in Pittsburgh, from fiscal years (FY) 2006 through 2014. All patients in that data set had been diagnosed with CHF during this time period, although the admissions were for all causes. Each admission record is considered as a

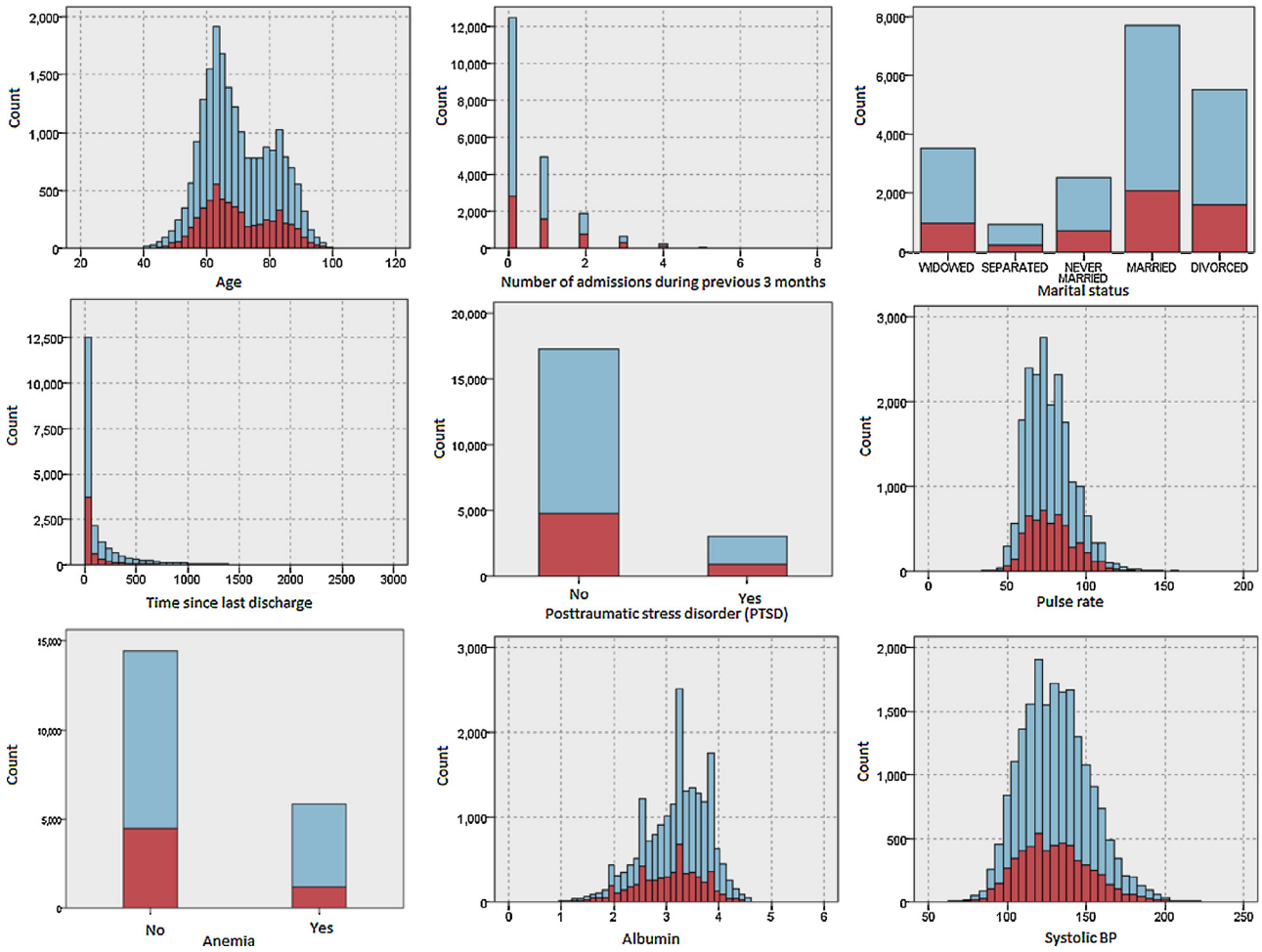

Fig. 1. Readmission class distribution among different attributes. Positive readmission instance (red), negative readmission instance (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

unit of analysis. The number of admissions for each patient ranged between 1 and 32. The time elapse between the previous inpatient admission and the current index admission, for each patient, was calculated by the date difference between them. This dataset is highly imbalanced. Only 27.99% of the admissions are of readmitted patients. Fig. 1 shows the readmission class distribution among different attributes. "Readmission" is the target variable, where the red color stands for positive readmission instances and the blue color stands for negative readmission instances.

2.2. Modeling approach

Our modeling approach for hospital readmissions suggests controlling the trade-off between reasoning transparency and predictive accuracy by combining a decision tree, as a base classifier, with an additional machine learning component. We ran a large number of experiments. Each experiment assigned a different learning algorithm and a different feature selection technique to our dataset. Table 1 compares the results obtained by different ensemble and stand-alone models, applied to our data. Among the decision trees, C5.0 is characterized with the highest accuracy. The SVM was chosen as the secondary classifier because it attained the greatest accuracy among the "black-box" classifiers, and because of the flexibility of the resulting classifier by that could be obtained by adjusting the kernel parameters. As discussed below, controlling the amount of observed error reduction through minimizing the correlation between the included models could be achieved by controlling the SVM structure, in combination with choosing the appropriate C5.0 confidence threshold. After selecting the classifiers, the predictive model was built by using the steps described in Fig. 2. Based on the considerations that are explained in the introduction, we suggest an optimization approach for the resulted mixed-ensemble model, which takes into account the unique characteristics of the learned dataset.

2.3. Data pre-processing

A few pre-processing steps were taken before analyzing the data:

1. Cohorts of CHF patients were identified based upon the ICD-9 codes of the previous principal (or secondary) inpatient discharge diagnoses. Patients may have had hospital admissions prior to their first CHF diagnosis.
2. All records of the same patient were merged if the patient was recorded as having had multiple admissions on the same day to the same medical unit. That applied to both medical and surgical inpatients.
3. Admissions that were recorded as having occurred within 24 h of the immediately prior index discharge were excluded from the admission set.
4. Inconsistent and/or error components, such as, age discrepancies, or a discharge date that preceded the admission date, were excluded.

Table 1
Performance comparison of different machine learning methods for both training and test sets.

Model			Specificity	Sensitivity	Precision	AUC
Stand-alone models						
Naïve-Bayes net		train	0.841	0.442	0.580	0.699
		test	0.823	0.489	0.250	0.676
Logistic regression		train	0.930	0.189	0.574	0.642
		test	0.917	0.281	0.293	0.699
Neural net		train	0.973	0.056	0.516	0.589
		test	0.960	0.089	0.053	0.639
SVM		train	0.935	0.282	0.684	0.768
		test	0.929	0.233	0.286	0.643
Ensemble models						
C5	boosted	train	0.943	0.352	0.709	0.714
		test	0.787	0.435	0.197	0.693
CART decision tree	boosted	train	0.921	0.196	0.557	0.529
		test	0.922	0.226	0.261	0.556
	bagged	train	0.959	0.116	0.593	0.557
		test	0.973	0.099	0.309	0.579
CHAID decision tree	boosted	train	0.897	0.295	0.590	0.671
		test	0.875	0.303	0.227	0.691
	bagged	train	0.960	0.131	0.624	0.678
		test	0.969	0.105	0.292	0.707
Quest decision tree	boosted	train	0.929	0.175	0.554	0.494
		test	0.920	0.203	0.236	0.487
	bagged	train	0.961	0.096	0.556	0.555
		test	0.977	0.072	0.283	0.579
Naïve net+ Logistic regression	voting	train	0.888	0.321	0.587	0.635
		test	0.869	0.382	0.261	0.653
Naïve net+ Neural-Net	voting	train	0.905	0.246	0.564	0.632
		test	0.896	0.263	0.234	0.635
Naïve net+ SVM	voting	train	0.890	0.368	0.626	0.65
		test	0.878	0.358	0.263	0.649
Logistic regression+ Neural-Net	voting	train	0.951	0.125	0.559	0.559
		test	0.942	0.168	0.261	0.59
Logistic regression+ SVM	voting	train	0.933	0.239	0.641	0.602
		test	0.924	0.262	0.297	0.607
Neural-Net + SVM	voting	train	0.957	0.167	0.662	0.602
		test	0.950	0.165	0.287	0.577

5. Missing values were imputed by the hot-deck method, using the same variable in prior or later observations of the specific patient in the data set. If the patient had no prior recorded inpatient admissions, his record was excluded.
6. Because our dataset included 1800 different diagnostic ICD-9 codes, too many to use in a predictive model, we categorized each principal (or secondary) inpatient discharge diagnostic code into one of the following categories, based on their association with possible readmission outcomes: Abnormality (lab abnormality), Acute, Cancer, Chronic, Complication, Exam, Fracture (FX or Fracture), Immunization (PNA, Flu, etc.), Infection, Pain, Procedure, Social, or none of the above.
7. The data set was separated into a training set of 15,481 admissions (75%), and test (holdout/validation) set of 4840 admissions (25%). Both C5.0 and SVM were trained using only the training set.

Table 2 shows a statistical summary of the attributes that were selected for use in the final dataset.

2.4. Feature selection and identification of outliers

We used a frequency-based approach to reduce the number of features, by applying Pearson's chi-square test of independence, which is suitable for unpaired data from large samples. The importance value of each variable was calculated as 1-p, where p is the p-value for the Pearson's chi-square test of association between each candidate predictor and the binary target variable, i.e., whether the patient was readmitted. The first 38 variables with Pearson's correlation coefficients greater than 0.6 were chosen. In addition, a sensitivity analysis for both C5.0 and SVM structures was further applied, in order to avoid undesirable bias. Table 3 shows the predictor importance, within each of the algorithms.

2.5. C5.0 algorithm

C5.0 works by recursively splitting the sample, using the feature that provides the maximum information gain [33]. For each record in our dataset, C5.0 generates predictions and confidence levels. In C5.0, each record is scored with a class and the confidence of the terminal node that classifies that record. If a rule set is directly generated from the C5.0 tree, and the misclassification costs are equal, then the confidence for the rule is calculated by:

$$C_{c5} = \frac{\text{Number of records correctly classified in leaf} + 1}{\text{Total number of records in leaf} + 2}$$

The C5.0 tree was optimized by running different combinations of the maximum depth and the minimum split size parameters. We applied 10-fold cross-validation to obtain a reliable, unbiased estimate of predictive accuracy.

2.6. SVM algorithm

SVM is based on the principle of structural risk minimization, in which the training set is mapped into a high-dimensional feature space, using a nonlinear transformation, referred to as the kernel. We used the Radial Basis Function (RBF) kernel [34]. Given a set of training vectors $x_i \epsilon R^l$, $i = 1,\ldots, l$, in two classes, and a vector $y \in R^l$, such that $y_i \epsilon \{-1, 1\}$, the LIBSVM algorithm [35,36] solves the following optimization problem:

$$\min f(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$

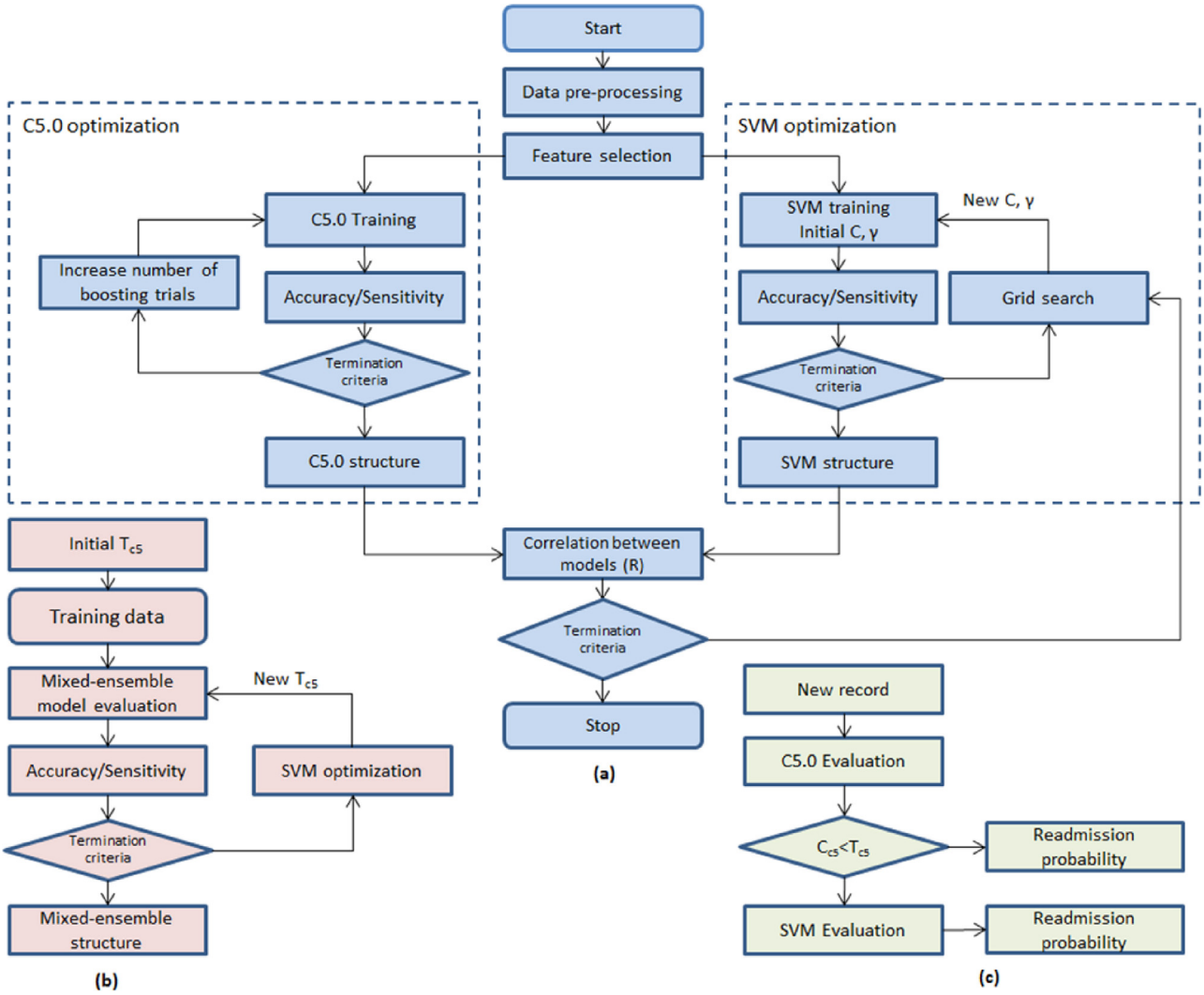

Fig. 2. (a) Optimization of the C5.0 (left) and SVM (right). (b) Optimization of the mixed-ensemble model. (c) mixed-ensemble model structure.

subject to $0 \le \alpha_i \le C$, and $y^T\alpha = 0$. The vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)^T$ contains the coefficients for the training samples (the coefficient is zero for non-support vectors).$Q\left(x_i \cdot x_j\right) = y_i y_j K\left(x_i, x_j\right)$. $K\left(x_i, x_j\right)$ is the transformation kernel. The decision function is defined as:

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i K\left(x_i, x_j\right) + b$$

where b is a constant term. The posterior probabilities are approximated using the sigmoid function

$$P_{A,B}(x) = \frac{1}{1 + \exp\left(Af(x) + B\right)},$$

where the optimal parameters A and B are estimated by solving a regularized maximum likelihood problem [37].

2.8. Our hybrid classifier

We combined the two models by predefining a threshold for the C5.0 tree confidence (C_{c5}). Records that are predicted with C_{c5} below the predefined confidence threshold, T_{c5} (typically those that do not fire the C5.0 rules), are further classified by the SVM, as described in Fig. 2.

2.9. Model optimization

2.9.1. Boosting experiments

Because our data set is highly imbalanced, it was inherently difficult to achieve high sensitivities. There is an incentive for a tree model to predict that most records are members of the majority class (in our case, that the patient will not be readmitted), as over 70% of the training data consists of such instances. Building a sequence of models, by using a boosting algorithm, has the potential for reducing the misclassification error for imbalanced datasets [38]. By using the original boosting algorithm, suggested by Freund and Schapire [39,40], the C5.0 algorithm classifies each record by applying to it the entire set of created models, using a weighted voting procedure to combine the separate predictions into one. Although boosting methods are known to be powerful in terms of avoiding over-fitting [41], using a large number of trials has led to over-fitting [41,42]. The maximum sensitivity for the test set, and lowest total accuracy difference between the training and test sets, were obtained after five boosting trials (data not shown).

2.9.2. SVM parameter tuning

The SVM algorithm is typically characterized by good in-sample and good out-of-sample performance. However, it is very sensitive to the parameters of the kernel function, and could easily over-fit the training set. Avoiding over-fitting could be achieved by a careful

Table 2
Summary statistics of selected attributes.

Field	Mean	Std. Dev	Skewness
Demographics and historical			
Age	69.294	11.019	0.222
Length of stay (LOS)	5.767	10.959	38.65
Number of admissions during the previous 3 months	0.608	0.957	2.065
Number of emergency visits during the previous 12 months	1.874	2.76	2.594
Time since last discharge	148.91	285.272	3.352
Completed outpatient appointment rate (%)	20.8	31.4	116.9
Number of bed days in all hospitals during the previous year	14.739	24.039	18.859
Vitals and lab values			
Systolic blood pressure (BP)	130.896	21.988	0.431
Diastolic blood pressure (BP)	70.946	12.363	0.458
Heart rate (Pulse)	77.033	15.144	0.964
Respiration rate	18.593	3.274	11.205
Albumin (g/DL)	3.18	0.615	−0.533
BUN (mmol/L)	25.762	43.43	16.696
Potassium (mEq/L)	4.246	2.605	18.074
White blood cell count (WBC) 10^9 x cells/L	7.906	5.64	25.349
Hematocrit (HCT) (%)	36.59	3.422	11.03
Glucose	8.903	39.651	5.726
Comorbidities			
Anemia (%)	71.1		
Asthma (%)	13.8		
Chronic Obstructive Pulmonary Disease (COPD) (%)	66.6		
Chronic Renal Failure (CRF) (%)	54.9		
Cerebrovascular Accident stroke (CVA) (%)	27.5		
Diabetes (%)	69.8		
Dementia (%)	21.7		
Depression (DPRSN) (%)	57.8		
Hypertension (HTN) (%)	96.5		
Ischemic Heart Disease (IHD) (%)	83.9		
Post-Traumatic Stress Disorder (PTSD) (%)	14.8		
Peripheral Vascular Disease (PVD) (%)	52.3		

Table 3
Predictor importance from SVM and C5.0.

Predictor	Importance(C5)	Importance (SVM)
Number of admissions during previous 3 months	0.1422	0.0422
Albumen category	0.1294	0.0301
Anemia	0.0771	0.0043
Category for number of bed days in a year	0.055	0.0169
Source of admission	0.0394	0.0002
Completed follow-up appointment	0.0376	0.0497
Time since last discharge	0.0367	0.0016
Hospital bed days category	0.0314	0.0155
Number of Emergency visits during previous year	0.0307	0.008
Chronic Obstructive Pulmonary Disease	0.027	0.0034
Category for number of hospital/bed day in a year	0.0256	0.018
Category for systolic reading	0.0245	0.0143
Category for respiration rate reading	0.0209	0.0049
Length Of Stay	0.0191	0.002
Ischemic Heart Disease	0.0187	0.006
Category for number of Other Non-Face visits in a year	0.0183	0.0018
Number of drugs filled for the year	0.0169	0.0307

tuning of the regularization parameter, C, and, in the case of nonlinear SVMs, a careful choice of the kernel and appropriate tuning of the kernel parameters.

Recent developments demonstrate the potential for using different tuning approaches for SVM parameters [43–45]. However, some of them are too sensitive to the initial parameters, may result in the optimization routine terminating in a local minimum value rather than in a global minimum, or tend to have result in complex computation problems and exhibit low efficiency [45,46]. Combining SVM with a particle swarm optimization (PSO) technique, implemented for parameter tuning, was recently applied for predicting hospital readmission [47].

Selection of C and gamma (γ) could be achieved by grid search [48,49]. The reasons why we used grid-search for the initial parameter settings are as follows. Although the method could be time consuming with low efficiency [50,51], it is considered to have satisfactory results when performed using continuous experiments [46,49]. The grid-search can be easily parallelized, because each pair (C, γ) is independent [48]. Furthermore, in our case, we use grid search only for setting the initial range of feasible values for C and for gamma (γ) (Fig. 2a). Further turning of C and γ, which is the heart of our modeling approach, is applied in combination with adjusting T_{c5}, such that the correlation between the SVM and C5 is minimized, and the performance of the mixed-ensemble structure is maximized, as is explained further below.

2.9.3. Decision boundaries of the SVM

The Gaussian RBF kernel creates a separating surface, based on the combination of bell-shaped surfaces centered at each support vector. The width of each bell-shaped surface is inversely

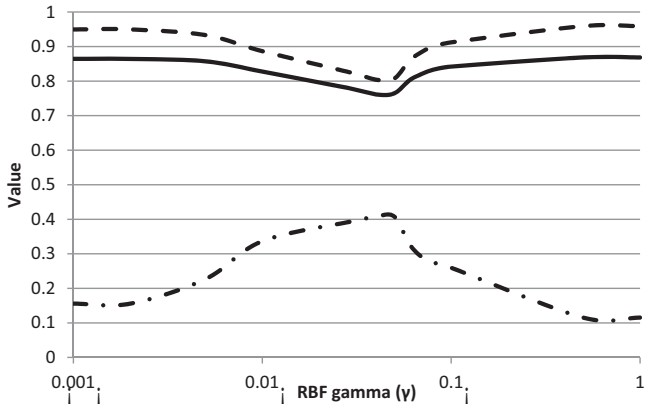

Fig. 3. Specificity (dashed line), sensitivity (dashed-dot line), and total accuracy (solid line), as a function of γ.

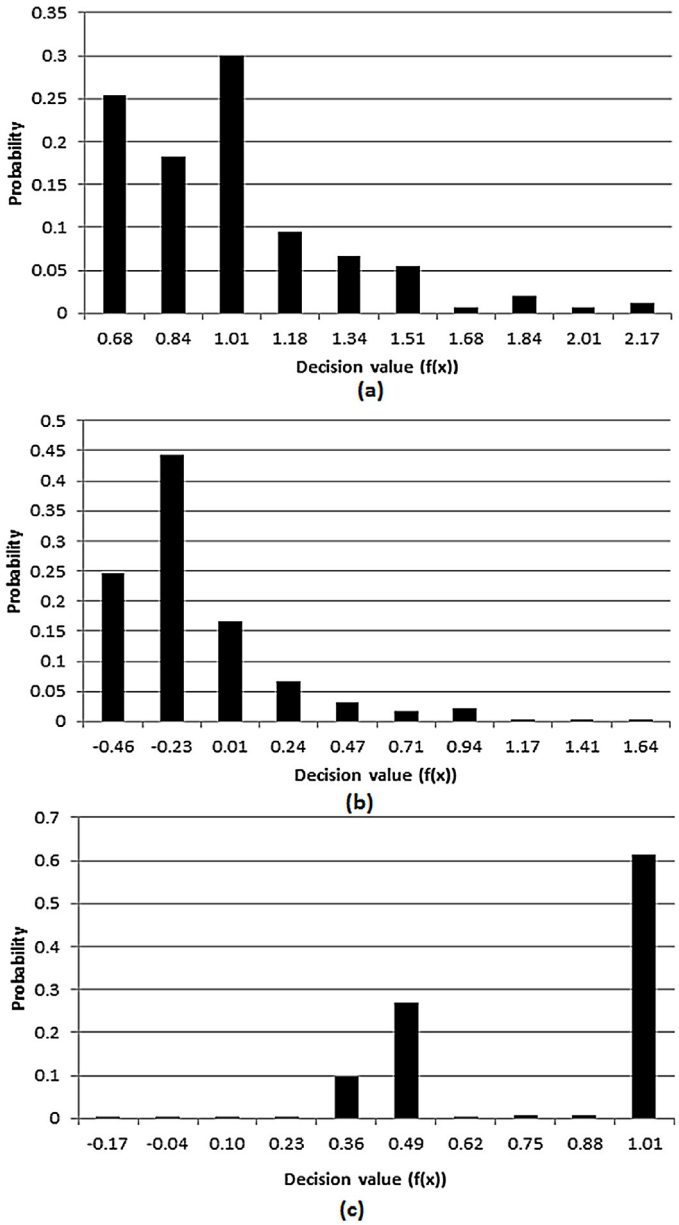

Fig. 4. (a) Decision values histogram (SVM), for $T_{c5} = 0.55$ (γ = 0.012). (b) Decision values histogram (SVM), for $T_{c5} = 0.85$ (γ = 0.012). (c) Decision values histogram (SVM), for $T_{c5} = 0.65$ (γ = 0.045).

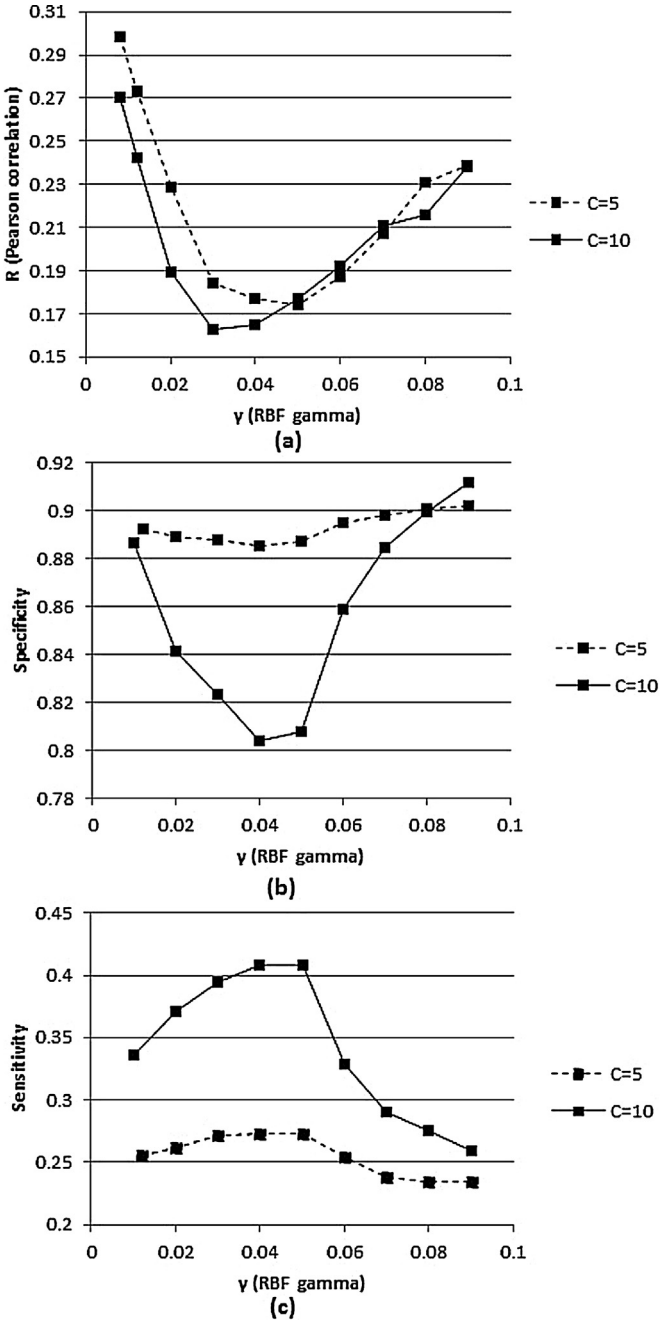

Fig. 5. (a) Pearson correlation between the models (R), as a function of γ, for C = 5 (dashed line), and for C = 10 (solid line). (b) Specificity as a function of γ, for C = 5 (dashed line), and for C = 10 (solid line). (c) Sensitivity as function of γ, for C = 5 (dashed line), and for C = 10 (solid line).

proportional to the value of the RBF parameter gamma (γ). Using too small a value for γ, as compared to the minimum pairwise distance between data values, might result in over-fitting. Values for γ that are too large may result in all points falling into a single class. Fig. 3 shows the sensitivity, specificity and total accuracy for different γ values, for the test set. Within the range of $0 < \gamma < 0.04$, as γ increases, the sensitivity increases and the specificity decreases. For $\gamma > 0.04$, as γ increases, the sensitivity decreases and the specificity increases. For lower sensitivity values, increasing the regularization parameter (C) would reduce the margin width for the training set, and increase the sensitivity for the test set. Fig. 4 shows the histogram for the SVM's decision values (Eq. (3)) for γ = 0.012 and $T_{c5} = 0.55$ (Fig. 4a), and for γ = 0.012 and $T_{c5} = 0.85$ (Fig. 4b). Because

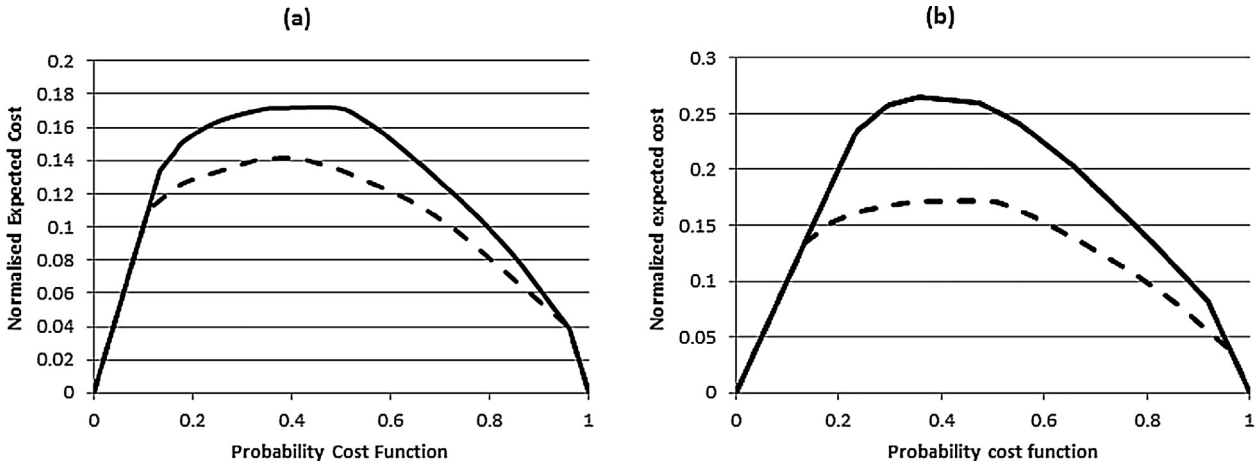

Fig. 6. (a) Cost curves for C = 10. (b) Cost curves for C = 5. Where, T_{c5} = 0.65 (solid line), and T_{c5} = 0.85 (dashed line).

our dataset is highly imbalanced, both distributions are biased. For lower T_{c5} values, the variance of the distribution is greater and the mean decision value is greater. That is, records that are classified with a lower C5.0 confidence are characterized by greater distances from the decision boundaries of the SVM. The optimization problem for those records will then become more complicated. Increasing T_{c5} can be considered, as long as the degree of correlation between the predictions of the two models, R, would not increase. Increasing γ can increase the total accuracy, by hardening the SVM's margins, but might also lead to over-fitting (Fig. 3). For $\gamma = 0.045$, the histogram of the SVM's decision values becomes bimodal, where the two peaks represent two ranges of decision values, corresponding to the two possible predictions (Fig. 4c). That is, the trade-off between maximizing the SVM's margins and minimizing the training error term should be analyzed in the context of each predefined threshold for C5.0 (T_{c5}) and its influence on the correlation between the models.

2.9.4. Correlation between the two models (R)

As the number of records that are classified by the SVM algorithm increases, the degree of correlation between the predictions of the two models increases in significance. Because records that are predicted with a lower C5.0 confidence have greater distances from the decision boundaries of the SVM, minimizing the margins' width would decrease the degree of correlation, but the total sensitivity will be affected [27]. Therefore, for each dataset, one would need to find the optimum balance between wide margins and a lesser degree of correlation between the predictions of the two models. Fig. 5a shows the Pearson correlation value, R, between the SVM and the C5.0 predictions, as a function of γ, for different C values. Clearly, R is lesser for the greater values of C (R = 0.174 for C = 5, R = 0.163 for C = 10). For greater values of C, the correlation reaches its minimum at lesser γ values ($\gamma = 0.05$ for C = 5, $\gamma = 0.03$ for C = 10) (Fig. 5a). For that minimum value, sensitivity reaches its maximum and specificity its minimum (Fig. 5b, Fig. 5c). However, both sensitivity and specificity are more sensitive to changes in γ as the value of C increases. Thus, the amount of error reduction due to the degree of correlation should be further analyzed, because of the influence of the value of γ. Increasing the model's sensitivity for records with uncertain C5.0 predictions by minimizing the SVM's margins width should also contribute to the model's stability. In certain cases, one might prefer to lower the model's sensitivity to γ differences, at the cost of total accuracy reduction. Due to the reasons explained above, the tradeoff between the model's stability and the level of correlation can be controlled by T_{c5}. As shown in Fig. 6, although the normalized expected cost of the SVM is lesser for C = 10, it is more sensitive to T_{c5} differences for C = 5.

3. Results

3.1. Group characteristics and predictors' importance

We found that, in both models, historical variables, such as the number of admissions before the current index admission, play an important role. Significant predictors of readmissions risk include two laboratory values, the readings for albumin and for white blood cell (WBC) count; comorbidities, such as anemia and COPD; and the source of admission. A history of recent prior admissions also predisposes a patient to readmission. Of the records having a C5.0 confidence that is lower than T_{c5} (denoted as Group-A, T_{c5} = 0.65), 37.1% had no inpatient admissions during the previous three months; whereas 66.8% of the records classified by the C5.0 model with a confidence greater than T_{c5} (denoted as Group-B, T_{c5} = 0.65) had no inpatient admissions during the previous three months. 63.3% of Group-A records are predicted to result in a readmission, as compared to 19.6% of Group-B records. Those differences imply that the value T_{c5} must be chosen carefully, due to its influence on the minority class analysis.

3.2. Performance metrics

Table 4 includes the values of performance metrics for our hybrid model, including its sensitivity, its specificity, its F1 score, its number of positive predictive values (PPV), and its number of negative predictive values (NPV), for different thresholds (T_{c5}), both for the test set and for the training set. For the test set, for $T_{c5} > 0.5$ as T_{c5} increases, specificity increases and sensitivity decreases. For $T_{c5} > 0.9$, the model is characterized by greater total accuracy for both sets, although with a lesser sensitivity for the test set. Fig. 7 shows the proportions of all cases that are TP (true positives), TN (true negatives), FP (false positives) and FN (false negatives) as a function of the cut off value, for T_{c5} = 0.65 (upper panel) and T_{c5} = 0.85 (lower panel), for Group-A. At lower cut-off values, the model predicts more cases to be readmissions, so the relative proportions of TP and FP predictions are greater. At higher cut-off values, the model predicts fewer cases to be readmissions, so the relative proportions of TN and FN are greater. For greater cutoff values, the relative proportion of TP + TN for SVM is greater than that for C5.0, for both threshold values. The relative proportion of TP is greater for T_{c5} = 0.65 than it is for T_{c5} = 0.85. Fig. 8 shows the

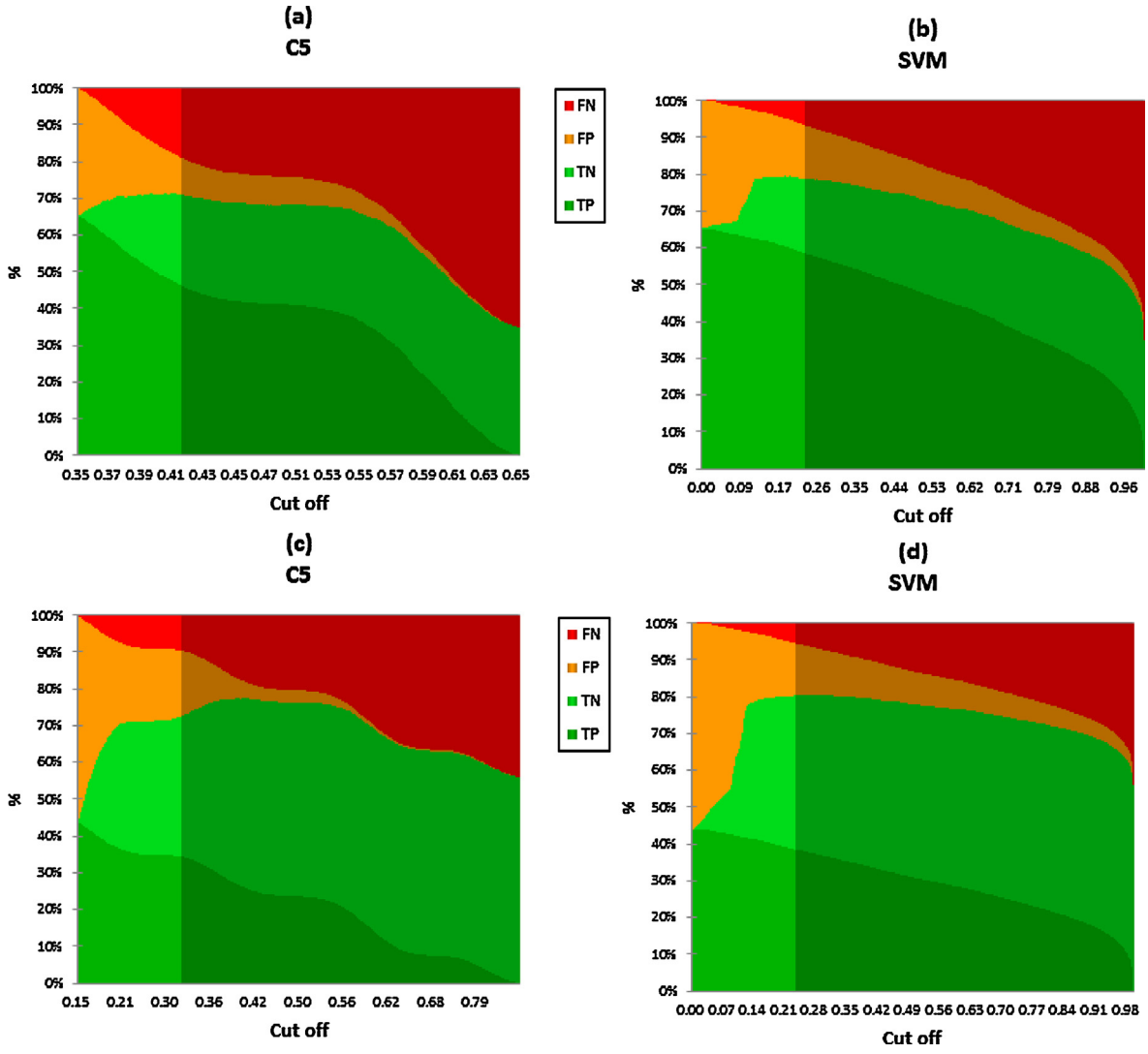

Fig. 7. Proportional breakdown of the number of TP (true positives, green), TN (true negatives, light green), FP (false positives, orange) and FN (false negatives, red), as a function of the chosen cut-off value, for $T_{c5} = 0.65$ ((a) and (b)), and $T_{c5} = 0.85$ ((c) and (d)), for Group-A (train + test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

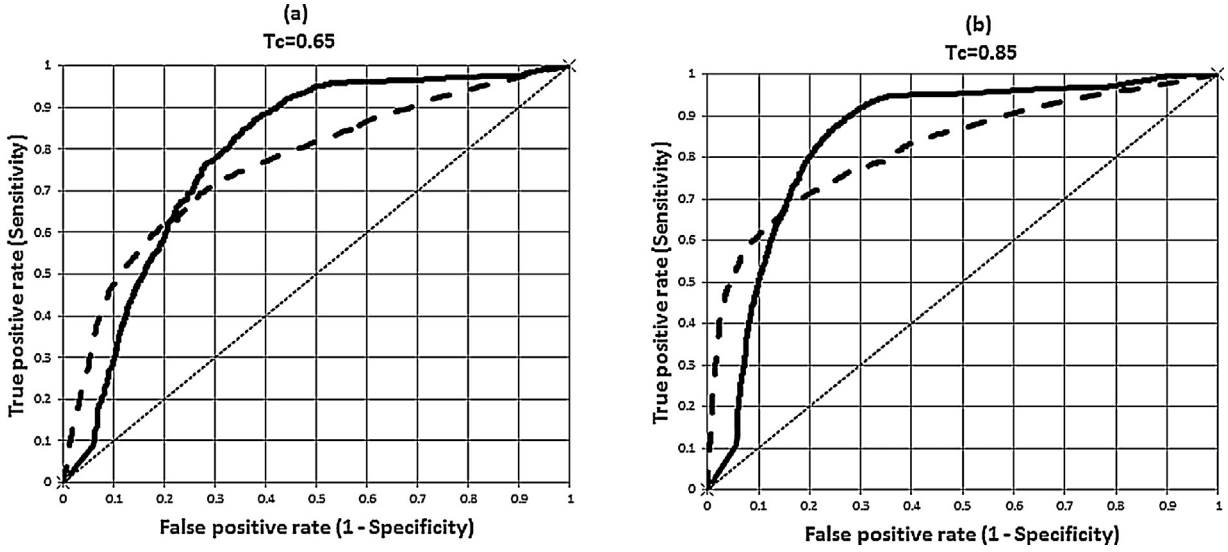

Fig. 8. ROC curves of SVM (thick – thin line) and C5.0 (long – dashed line), of Group A, for $T_{c5} = 0.65$ (a), and $T_{c5} = 0.85$ (b) (train + test).

Table 4
Performance metrics.

Threshold (C_{c5})	True positive	False positive	True negative	False negative	Sensitivity	Specificity	PPV	NPV	Accuracy	F1 score
Train set										
0.9	4675	21	10290	492	0.905	0.998	0.996	0.954	0.967	0.948
0.8	4169	43	10268	998	0.807	0.996	0.990	0.911	0.933	0.889
0.7	3932	68	10243	1235	0.761	0.993	0.983	0.892	0.916	0.858
0.65	3786	102	10209	1381	0.733	0.990	0.974	0.881	0.904	0.836
0.6	3314	123	10188	1853	0.641	0.988	0.964	0.846	0.872	0.770
0.5	2936	323	9988	2231	0.568	0.969	0.901	0.817	0.835	0.697
0.4	2936	323	9988	2231	0.568	0.969	0.901	0.817	0.835	0.697
Test set										
0.9	132	405	3916	388	0.254	0.906	0.246	0.910	0.836	0.250
0.8	129	378	3943	391	0.248	0.913	0.254	0.910	0.841	0.251
0.7	137	394	3927	383	0.263	0.909	0.258	0.911	0.839	0.261
0.65	134	381	3939	386	0.258	0.912	0.260	0.911	0.842	0.259
0.6	136	394	3927	384	0.262	0.909	0.257	0.911	0.839	0.259
0.5	151	532	3789	369	0.290	0.877	0.221	0.911	0.814	0.251
0.4	151	532	3789	369	0.290	0.877	0.221	0.911	0.814	0.251
0.3	151	532	3789	369	0.290	0.877	0.221	0.911	0.814	0.251

ROC curves for the two models, for Group-A. For both thresholds, the SVM predictions are characterized by greater sensitivity values (true positive rate) than are the C5.0 predictions, for a wider range of cut off values (0–0.65 for SVM, as compared to 0.51–0.65 for C5.0), for both threshold values. For $T_{c5} = 0.85$, that range is narrower (0-0.61 for SVM, as compared to 0.38–0.85 for C5.0).

4. Discussion

This study presents a mixed-ensemble model for estimating the probability that a hospitalized patient will be readmitted within 30 days following discharge, either to the same or to a different hospital. The mixed-ensemble model combines (1) a boosted C5.0 model (five trees) as the base ensemble classifier, which enables exploratory knowledge discovery about the learned readmission database; and (2) a support vector machine (SVM) as a secondary classifier, which allows control of the classification error for the minority class, i.e., positive readmission instances. We illustrate our approach for the prediction of all-cause hospital readmissions of patients who have been diagnosed with Congestive Heart Failure (CHF) over a nine year period. The timing of the CHF diagnosis during that time period is not analyzed in this research. That is, we assume that the risk of readmission for patients who were diagnosed with CHF toward the end of this period, as compared to patients who were diagnosed with CHF at the beginning of the period, is the same. Because most patients, even those with CHF, are not readmitted, the readmission classification problem is highly imbalanced. Thus, most models for predicting readmission are characterized by low sensitivity (true-positive rate). In order to deal with low sensitivity, as well as with optimizing the model's transparency, we suggest a new optimization approach, which takes into account the degree of correlation between the models, the distance of the minority instances to the decision boundaries of the SVM, the penalty for misclassification errors for patients who were actually readmitted (positive readmission instances), and generalization power. We found that historical variables, such as the number of admissions before the current index admission, are used by a significant portion of the extracted rules. That result is not surprising, because the more frequently a patient is admitted to a hospital, the greater are his chances of being readmitted. Interestingly, patient records that are correctly classified by C5.0 were less likely to have any type of comorbidity, as compared to records that are classified well by the SVM. However, the latter group is characterized with closer-to-normal lab values than is the first group. Such differences may indicate that the relationship of unusual lab values to readmission may be expressible using IF-THEN rules, but that the relationship of comorbidities to readmission can only be modeled with the more complex, nonlinear structure of a support vector machine. The use of a mixed-ensemble approach enables us to discover such insights.

5. Conclusions

We developed a new, dynamic mixed-ensemble model for predicting hospital readmission. To the best of our knowledge, our model is the first readmission model that deals with the potential conflict between predictive accuracy and reasoning transparency. Our results indicate that a cautious optimization of the model structure would support an effective communication of the reasoning underlying its prediction, as well as controlled-sensitivity classification of the minority class. Desirable extensions of this work include exploring whether a single model is appropriate for all types of patients, or if there need to be different models for different groups of patients, expanding the model to predict how the chances of readmission change over time for each patient, and incorporation of knowledge-based clinical information in the model structure.

Contributors

We thank the U.S. Department of Veterans Affairs for providing financial support for this research, through master contract numbers VA244-13-C-0581 and VA240-14-D-0038 with the University of Pittsburgh. This work is an outcome of a continuing partnership between the Katz Graduate School of Business and the Pittsburgh Veterans Engineering Resource Center (VERC). Inpatient admissions data were pulled from the VA corporate data warehouse by Dr. Youxu C. Tjader. The categorization at Step 6 of the data pre-processing was created by Ms. Roberta Sciulli. We thank the reducing readmission project group of the Pittsburgh Veterans Engineering Resource Center (VERC), for useful discussions.

References

[1] Aggarwal S, Gupta V. Demographic parameters related to 30-day readmission of patients with congestive heart failure: analysis of 2,536,439 hospitalizations. Int J Cardiol 2014;176:1343–4.
[2] Islam T, O'Connell B, Lakhan P. Hospital readmission among older adults with congestive heart failure. Aust Health Rev 2013;37:362–8.
[3] Krumholz HM, Parent EM, Tu N, Vaccarino V, Wang Y, Radford MJ, et al. Readmission after hospitalization for congestive heart failure among Medicare beneficiaries. Arch Intern Med 1997;157:99–104.
[4] Vinson J, Rich M, Sperry J, Shah A, McNamara T. Early readmission of elderly patients with congestive heart failure. J Am Geriatr Soc 1990;38:1290–5.
[5] Parmley WW. Pathophysiology and current therapy of congestive heart failure. J Am Coll Cardiol 1989;13:771–85.

[6] Schocken DD, Arrieta MI, Leaverton PE, Ross EA. Prevalence and mortality rate of congestive heart failure in the United States. J Am Coll Cardiol 1992;20:301–6.
[7] Muus K, Knudson A, Klug M, Gokun J, Sarrazin M, Kaboli P. Effect of post-discharge follow-up care on re-admissions among US veterans with congestive heart failure: a rural-urban comparison. Rural Remote Health 2010;10:1447.
[8] Philbin EF, DiSalvo TG. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. J Am Coll Cardiol 1999;33:1560–6.
[9] Silverstein MD, Qin H, Mercer SQ, Fong J, Haydar Z. Risk factors for 30-day hospital readmission in patients≥ 65 years of age. Proc (Baylor University Medical Center) 2008;21:363–72.
[10] Shmueli G, Koppius O. Predictive analytics in information systems research Robert H. Smith School Research Paper No RHS; 2010. p. 6–138.
[11] Krumholz HM, Chen Y-T, Wang Y, Vaccarino V, Radford MJ, Horwitz RI. Predictors of readmission among elderly survivors of admission with heart failure. Am Heart J 2000;139:72–7.
[12] Alexander M, Grumbach K, Remy L, Rowell R, Massie BM. Congestive heart failure hospitalizations and survival in California: patterns according to race/ethnicity. Am Heart J 1999;137:919–27.
[13] Bardhan I, Kirksey K, Oh J-H, Zheng E. A predictive model for readmission of patients with congestive heart failure: a multi-hospital perspective. In: Proceedings of the thirty-second international conference on information systems. 2011. p. 1–39.
[14] Williams EI, Fitton F. Factors affecting early unplanned readmission of elderly patients to hospital. BMJ 1988;297:784–7.
[15] Kent S, Yellowlees P. Psychiatric and social reasons for frequent rehospitalization. Psychiatr Serv 1994;45:347–50.
[16] Annema C, Luttik M-L, Jaarsma T. Reasons for readmission in heart failure: perspectives of patients, caregivers, cardiologists, and heart failure nurses. Heart Lung: J Acute Crit Care 2009;38:427–34.
[17] Schwarz KA, Elman CS. Identification of factors predictive of hospital readmissions for patients with heart failure. Heart Lung: J Acute Crit Care 2003;32:88–99.
[18] He D, Mathews SC, Kalloo AN, Hutfless S. Mining high-dimensional administrative claims data to predict early hospital readmissions. J Am Med Inform Assoc 2014;21:272–9.
[19] Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnapuram B. Predicting readmission risk with institution-specific prediction models. Artif Intell Med 2015;65:89–96.
[20] Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, et al. Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch Intern Med 2008;168:1371–86.
[21] Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011;306:1688–98.
[22] Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, et al. Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch Intern Med 2008;168:1371–86.
[23] Kim SM, Han H-R. Evidence-based strategies to reduce readmission in patients with heart failure. J Nurse Pract 2013;9:224–32.
[24] Hilbert JP, Zasadil S, Keyser DJ, Peele PB. Using decision trees to manage hospital readmission risk for acute myocardial infarction, heart failure, and pneumonia. Appl Health Econ Health Policy 2014;12:573–85.
[25] Hosseinzadeh A, Izadi M, Verma A, Precup D, Buckeridge D. Assessing the predictability of hospital readmission using machine learning. In: Proceedings of the twenty-seventh (association for the advancement of artificial intelligence) AAAI conference. 2013. p. 1532–8.
[26] Rokach L. Ensemble methods for classifiers Data Mining and Knowledge Discovery Handbook. Springer; 2005. p. 957–80.
[27] Ali KM, Pazzani MJ. Error reduction through learning multiple descriptions. Mach Learn 1996;24:173–202.
[28] Wolpert DH. Stacked generalization. Neural Netw 1992;5:241–59.
[29] Chan PK, Stolfo SJ. A comparative evaluation of voting and meta-learning on partitioned data. ICML 1995:90–8.
[30] Gal-Or M, May JH, Spangler WE. Assessing the predictive accuracy of diversity measures with domain-dependent, asymmetric misclassification costs. Inf Fusion 2005;6:37–48.
[31] Kuncheva LI. Combining pattern classifiers: methods and algorithms. Hoboken, New Jersey: John Wiley & Sons; 2004. p. 275–93.
[32] He H, Garcia EA. Learning from imbalanced data. Inst Electr Electron Eng (IEEE) Trans Knowl Data Eng 2009;21:1263–84.
[33] Quinlan JR. C4 5: programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.
[34] Vapnik VN, Vapnik V. Statistical learning theory. New York: Wiley; 1998. p. 389.
[35] Chang C-C, Lin C-J. LIBSVM a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2:27.
[36] IBM SPSS Modeler 17 Algorithms Guide. IBM Corporation 1994, 2015; 2012. p. 367.
[37] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 1999;10:61–74.
[38] Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn 2007;40:3358–78.
[39] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Saitta L, editor. Machine learning, proceedings of the thirteenth international conference (ICML '96). 1996. p. 148–56.
[40] Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. Computational learning theory. Springer; 1995. p. 23–37.
[41] Rätsch G, Onoda T, Müller K-R. Soft margins for AdaBoost. Mach Learn 2001;42:287–320.
[42] Rätsch G, Onoda T, Müller KR. An improvement of adaboost to avoid overfitting. In: 5th international conference on neural information processing. 1998. p. 506–9.
[43] Keerthi SS. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. Inst Electr Electron Eng (IEEE) Trans Neural Netw 2002;13:1225–9.
[44] Lau K, Wu Q. Leave one support vector out cross validation for fast estimation of generalization errors. Pattern Recognit 2004;37:1835–40.
[45] Aydin I, Karakose M, Akin E. A multi-objective artificial immune algorithm for parameter optimization in support vector machine. Appl Soft Comput 2011;11:120–9.
[46] Wang S, Li R. A novel classification method based on improved SVM and its application international. J Database Theory Appl 2015;8:281–90.
[47] Zheng B, Zhang J, Yoon SW, Lam SS, Khasawneh M, Poranki S. Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Syst Appl 2015;42:7110–20.
[48] Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. Technical Report. Department of Computer, Science and Information Engineering, National Taiwan University; 2004.
[49] Min JH, Lee Y-C. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst Appl 2005;28:603–14.
[50] Zhu Y, Li C, Zhang Y. A practical parameters selection method for SVM. In: International symposium on neural networks. 2004. p. 518–23.
[51] Lin J, Zhang J. A fast parameters selection method of support vector machine based on coarse grid search and pattern search. 2013 fourth global congress on intelligent systems: IEEE 2013:77–81.