Neural Comput & Applic (2020) 32:5735–5744
DOI 10.1007/s00521-017-3242-y

ORIGINAL ARTICLE

Balanced training of a hybrid ensemble method for imbalanced datasets: a case of emergency department readmission prediction

Arkaitz Artetxe[1] · Manuel Graña[2] · Andoni Beristain[1] · Sebastián Ríos[3]

Received: 27 March 2017 / Accepted: 4 October 2017 / Published online: 14 October 2017

Abstract Dealing with imbalanced datasets is a recurrent issue in health-care data processing. Most literature deals with small academic datasets, so that results often do not extrapolate to the large real-life datasets, or have little real-life validity. When minority class sample generation by interpolation is meaningless, the recourse to undersampling the majority class is mandatory in order to reach some acceptable results. Ensembles of classifiers provide the advantage of the diversity of their members, which may allow adaptation to the imbalanced distribution. In this paper, we present a pipeline method combining random undersampling with bootstrap aggregation (bagging) for a hybrid ensemble of extreme learning machines and decision trees, whose diversity improves adaptation to the imbalanced class dataset. The approach is demonstrated on a realistic greatly imbalanced dataset of emergency department patients from a Chilean hospital targeted to predict patient readmission. Computational experiments show that our approach outperforms other well-known classification algorithms.

Keywords Class imbalance · Hospital readmission · Ensemble learning · Extreme learning machine

✉ Manuel Graña
manuel.grana@ehu.es

Arkaitz Artetxe
aartetxe@vicomtech.org

1 Vicomtech-IK4 Research Centre, Mikeletegi Pasealekua 57, 20009 San Sebastián, Spain

2 Computation Intelligence Group, Basque University (UPV/EHU), P. Manuel Lardizabal 1, 20018 San Sebastián, Spain

3 CEINE, Universidad de Chile, Av. República 701, Santiago, Región Metropolitana, Chile

1 Introduction

In supervised classification, we say that a dataset is imbalanced when the a priori probabilities of the classes are significantly different, i.e., there exists a minority (positive) class that is underrepresented in the dataset in contrast to the majority (negative) class [8, 23, 27]. The minority class can have the meaning of a rare event, such as an alert condition, an intrusion in a security system, or a disease in a population. Such situations appear in health care as well as in many other fields, e.g., fraud detection, cybersecurity, communications, fault diagnosis. Often the minority class is the target class to be predicted because it is related to the highest cost/reward events [15]. Most classification algorithms assume equal a priori probability for all the classes, or equivalently equal cost to errors in classification, so that when this premise is violated, we find that the resulting classifier is biased toward the majority class. It has a higher predictive accuracy over the majority class, but poorer predictive accuracy over the minority class. A measure of class imbalance is given by the imbalance ratio (IR), defined as the ratio of the number of instances in the majority class and the number of those in the minority class. Some studies have shown that conventional classifier performance deteriorates even with moderate imbalance ratios [17].

Readmissions to a hospital service are recurrent admissions of a patient with in between times are smaller than a set threshold. For instance, 30 days have been set as the standard threshold for readmission in the USA. Readmissions are becoming a strong concern of hospitals as a measure of the quality of given care [25], springing new interventions and measures, such as networking social care and health institutions, and developing predictive tools allowing to preventive measures [2]. Despite the number of studies carried out on readmission prediction, some authors claim that most of the current predictive models based on administrative and clinical data discriminate poorly on readmissions [12, 19], may be due to fact that the data are poorly discriminant by itself.

The current approach to readmission risk prediction is to build specific one-class or two-class classifiers from available data. However, readmission events are rare events; therefore, the training has to deal with a class-imbalanced dataset [4]. Unfortunately, still most of the literature on readmission prediction does not take into account this fact.

In this paper, to mitigate the effect of imbalanced data we employ an ensemble method that embeds a resampling strategy (random undersampling) within each of the bootstrapped replica of the original dataset. Then, each resampled dataset is used to train an ensemble classifier couple composed of a decision tree and a extreme learning machine. Extreme learning machines (ELM) [11] have a great appeal, despite their lack of stability, because of their very quick learning. In addition, decision trees, which have proven to perform well in a wide range of domain, produce human-interpretable models, which is a desirable characteristic in a clinical domain. The final classifier is an ensemble of these coupled classifiers. The effect is that we train the individual classifiers of the ensemble using almost balanced datasets, achieving competitive results on the imbalanced test dataset. Testing this approach on a highly imbalanced clinical dataset, we achieve competitive results in a experimental comparative study.

The paper is organized as follows. Section 2 presents some relevant work on the issue of readmission risk prediction. In Sect. 3 we present our dataset. Next, we briefly describe the evaluation metrics and classifiers that we have used to build our models. In Sect. 5 experimental results are presented. Finally, in Sect. 6 we discuss the conclusions and future work.

2 Related work

There are two basic approaches to deal with imbalanced data [8]. One is to manipulate the data, either by undersampling the majority class or by oversampling (e.g., SMOTE [7]) the minority class, achieving a balanced training dataset. Both approaches have been tried in the literature with some variations. For instance, clustering the data prior to train conventional classifiers allows to efficiently reduce the data of the majority class without damaging the statistical distribution [28]. On the other hand, performing a Voronoi tessellation of the input space [29] allows to identify the regions where the generation of new samples by random resampling provides consistent data generalization. In any case, the trained classifiers must be validated against independent data sample following the same distribution (i.e., imbalance ratio) as the original dataset. The other approach is to tailor the classifier architecture to cope with the imbalance data. For instance, the approach proposed in [30] carries out the adaptation of support vector machine (SVM) the kernel by conformal transformations guided by the chi square test. Many approaches rely on ensemble classifiers, such as the Adaboost and its combinations/variations [26]. Ensembles have been used to improve the predictive power of binary classifiers in the decomposition of multiple class-imbalanced problem into one-to-one decision problems [31]. Most studies about fundamental research on imbalanced datasets have been carried out over small academic datasets; there are few studies involving as large as the one considered here [8].

Readmission risk modeling has both economic and health-care impacts. Identifying patients with higher risk of returning after discharge before a critical time span, allows to apply preventive interventions such as phone calls, home visits or online monitoring. Most studies consider 30 days as the threshold to qualify an admission as a readmission, because of political criteria. Logistic regression is a standard tool used for readmission prediction [1]. A systematic review of risk prediction models for hospital readmission [12], found that most studies are focused in specific conditions or diseases, such as acute myocardial infarction (AMI) or heart failure (HF) [1], concluding that readmission risk prediction remains poorly understood and has achieved limited success. A comparison of traditional statistical techniques and machine learning methods to predict 30- and 180-day readmissions (all-cause and HF-caused) [19] concludes that ML methods can improve both discrimination and range of prediction over traditional statistical techniques. The study identifies the existence of class imbalance problems, discussing different techniques, namely random oversampling, random undersampling and weight variation. The final approach they took was varying the weight of the minority class, i.e., readmit cases. Similarly, an ensemble composed of a boosted tree and a SVM classifier was applied to predict all-cause readmission of congestive heart failure patients [24], employing a boosting algorithm as a way to reduce the misclassification errors for imbalance datasets. Other authors [32] use random oversampling to improve the performance of several data mining approaches (neural network, random forest and SVM) to predict hospital readmissions, while others [21] apply undersampling and bagging to create the training data. Some works [3, 18, 32] apply simple data preprocessing approaches such as oversampling and undersampling. Finally, some works [13, 16] combine boosting with data resampling, such as the boosting and random undersampling (RUSBoost).

3 Dataset description

The readmission dataset that we will treat in this paper was collected at the EHR of the Hospital José Joaquín Aguirre of the Universidad de Chile between 2013 and 2016. The attributes recorded in the dataset are divided into three main groups: (1) sociodemographic and administrative data, (2) health status and (3) reasons for consultation or diagnoses made at admission. After removing inconsistent and missing samples, the dataset was composed of 99,858 instances. Missing values were imputed using the arithmetic mean for continuous variables and the mode for categorical variables. Table 1 shows the characteristics of the dataset and the distribution of 72-h readmissions among different variables.

For each admission of a patient to the ED, we calculated the number of days elapsed since his last visit. In order to build our model following a binary classification approach, the target variable was set to readmission/not readmission. Those patients returning to the ED within 72 h after being discharged were considered as a readmission; otherwise, they were considered as a normal admission.

Figure 1 shows the distribution of readmission class among some attributes of our dataset. Readmissions (shown as green columns) are much less frequent than normal admissions. Class distribution shown in Table 3 indicates a imbalance ratio (IR) of approximately 28, which is a strong case of class-imbalanced data.

4 Methods

In this section, we present the classification algorithms that we have used as well as the evaluation metrics employed for measuring the performance of our models.

4.1 Classifiers

Decision trees (DT) and random forests (RF) [6, 20] are built by recursive partitioning of the data space using some quantitative criterion (e.g., mutual information, gain-ratio, gini index). Tree leaves correspond to the probabilistic assignment of data samples to classes. Often, a pruning process is applied in order to reduce both tree complexity and training data over-fitting. Ensembles of DT classifiers were among the first been proposed, i.e., bagging [5] and random forests [6]. Random forests are ensembles of DT, where each individual DT is built on a bootstrapped training data subset over a random subset of the input variables. The majority voting rule applied to the ensemble of outputs decides the input data class assignment.

Extreme learning machines (ELM) [10, 11] were proposed as a very fast training algorithm for single-layer feedforward neural networks (SLFN). The ELM avoid gradient descent of the hidden layer weights by performing a random sampling, equivalent to a random subspace projection. The training problem reduces to the estimation of the output weights by linear least squares resolution of the network response minimizing the classification error, often solved by the Moore–Penrose generalized pseudo-inverse. Randomization of hidden layer weights introduce training instability which has been tackled in many ways.

Naive Bayes methods are based on the naive assumption that the components of the pattern vectors are statistically independent, so that the posterior probability of the class can be approximated by a product decomposition of the likelihood of individual features. The Gaussian naive Bayes (which we use in our experiments) assumes that the likelihood follows a Gaussian distribution, where the mean and standard deviation of each feature are estimated from the sample.

4.2 Bagging ensemble method

As shown in Table 1, our dataset is highly imbalanced (IR of 28.16); thus, we need to overcome the bias toward the majority class. Since our dataset has more than 96,000 negative samples, undersampling the majority class may achieve good results, while the risk of discarding crucial information during undersampling is low. We have found that oversampling methods, as SMOTE [7] or ADASYN [9], perform better in low imbalance ratios. Moreover, we experimentally found that the random generation of samples involving the qualitative variable that specifies the case of the admission gives very bad results. We think that oversampling qualitative or categorical variables is an open issue.

Our method combines a class balancing preprocessing technique (random undersampling) with bootstrap aggregating, also known as bagging. Bagging [5] consists in creating bootstrapped replicas of the original dataset with replacement (i.e., different copies of the same instance can be found in the same bag), so that different classifiers are trained on each replica. Originally each new dataset or bag maintained the size of the original dataset. Nevertheless, underbagging and overbagging strategies embed a resampling process, so that bags are balanced by means of undersampling or oversampling techniques. To classify an unseen instance, the output predictions of the weak classifiers are collected performing a majority vote in order to produce the joint ensemble prediction. The purpose of this combination is to create a model to classify imbalanced data, improving the generalization capacity without sacrificing overall accuracy. As shown in Fig. 2, our approach

Table 1 Characteristics of the dataset

Variable	All patients n = 99858	Readmitted n = 3425	Not readmitted n = 96433
Age, mean (SD)	41.0 (22.4)	36.1 (22.9)	41.2 (22.4)
Male sex (%)	44956 (45.0)	1624 (1.6)	43332 (43.4)
Daytime (%)	69321 (69.4)	2171 (2.2)	67150 (67.2)
Evaluation, mean (SD)	5.0 (3.3)	4.8 (3.5)	5.0 (3.3)
Fragility idx, mean (SD)	0.0 (2.5)	0.0 (2.3)	0.0 (2.5)
Triage (%)			
I	182 (0.2)	2 (0.0)	180 (0.2)
II	12694 (12.7)	317 (0.3)	12377 (12.4)
III	77813 (77.9)	2718 (2.7)	75095 (75.2)
IV	9131 (9.1)	387 (0.4)	8744 (8.8)
V	38 (0.0)	1 (0.0)	37 (0.0)
Pathology (%)			
Gineco-obstetrics	236 (0.2)	6 (0.0)	230 (0.2)
General medicine	77192 (77.3)	2458 (2.5)	74734 (74.8)
Pediatrics	7094 (7.1)	563 (0.6)	6531 (6.5)
Traumatology	15336 (15.4)	398 (0.4)	14938 (15.0)
Destination (%)			
External center	3372 (3.4)	116 (0.1)	3256 (3.3)
Home	71999 (72.1)	2703 (2.7)	69296 (69.4)
Hospital	14700 (14.7)	61 (0.1)	14639 (14.7)
Left without being seen	9787 (9.8)	545 (0.5)	9242 (9.3)
Reason for consultation (%)			
Cephalea	6421 (6.4)	192 (0.2)	6229 (6.2)
Pain—abdomen gen.	9861 (9.9)	404 (0.4)	9457 (9.5)
Pain—epigastrium	3177 (3.2)	143 (0.1)	3034 (3.0)
Pain—lumbar	2964 (3.0)	107 (0.1)	2857 (2.9)
Pain—foot	2909 (2.9)	92 (0.1)	2817 (2.8)
General malaise	3027 (3.0)	78 (0.1)	2949 (3.0)
Other	10867 (10.9)	374 (0.4)	10493 (10.5)
...			
Saturation, mean (SD)	96.6 (9.6)	96.2 (12.1)	96.6 (9.5)
Tad, mean (SD)	74.1 (22.3)	67.6 (29.4)	74.3 (21.9)
Tas, mean (SD)	125.8 (35.9)	114.5 (48.8)	126.2 (35.3)
Temperature, mean (SD)	35.9 (4.5)	35.5 (5.9)	35.9 (4.4)
Heart rate, mean (SD)	87.2 (22.3)	92.7 (29.1)	87.0 (22.0)
Breath rate, mean (SD)	17.0 (5.6)	15.1 (7.6)	17.0 (5.5)1
Prevision (%)			
2	5943 (6.0)	180 (0.2)	5763 (5.8)
5	3641 (3.6)	108 (0.1)	3533 (3.5)
6	27903 (27.9)	1022 (1.0)	26881 (26.9)
9	11060 (11.1)	432 (0.4)	10628 (10.6)
18	44464 (44.5)	1468 (1.5)	42996 (43.1)
35	1011 (1.0)	30 (0.0)	981 (1.0)
37	1103 (1.1)	33 (0.0)	1070 (1.1)
48	2074 (2.1)	70 (0.1)	2004 (2.0)
...			

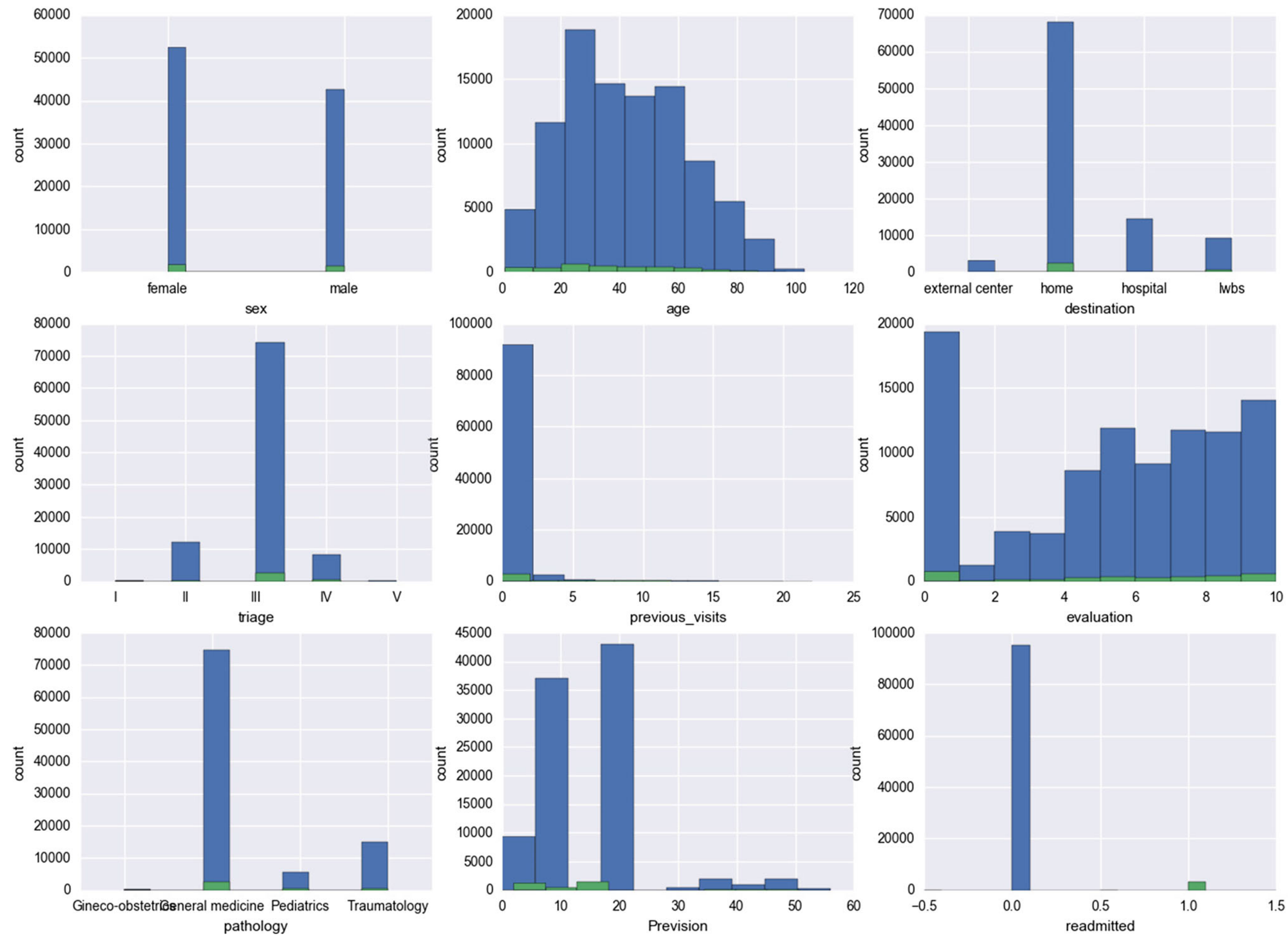


Fig. 1 Distribution of readmission class among different attributes (readmission in green, blue normal admissions) (color figure online)

consists in applying a balancing preprocess to each subset obtained from the bootstrap. Following, a ensemble classifier is built, combining ELM and decision tree classifiers using soft voting as combination strategy. The black-box nature of ELMs (and ensemble methods in general) is combined with the comprehensibility of a decision tree. Some works [14] have combined ELM with DT due to its interpretable ability as 'IF-THEN'-like rule generator.

4.3 Evaluation metrics

The evaluation metrics that we have used are: accuracy, recall, specificity and area under ROC curve (AUC), defined as follows (Table 2):

Accuracy In binary classification, accuracy is defined as the proportion of true results among the total population:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{1}$$

where TN is a true negative, TP is a true positive, FN is a false negative, and FP is a false positive. In heavily skewed datasets it is not very meaningful because a simple strategy such as assigning each test sample to the majority class provides high accuracy.

Recall aka sensitivity is a classification performance measure defined as the proportion of correctly classified positives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Recall provides more informative about the success on the target class.

Specificity is defined as the proportion of negatives that are correctly identified as such:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

AUC The area under ROC curve (AUC) shows the trade-off between the sensitivity or TP_{rate} and FP_{rate} ($1 -$ specificity):

$$\text{AUC} = \frac{1 + \text{TP}_{\text{rate}} - \text{FP}_{\text{rate}}}{2} \tag{4}$$

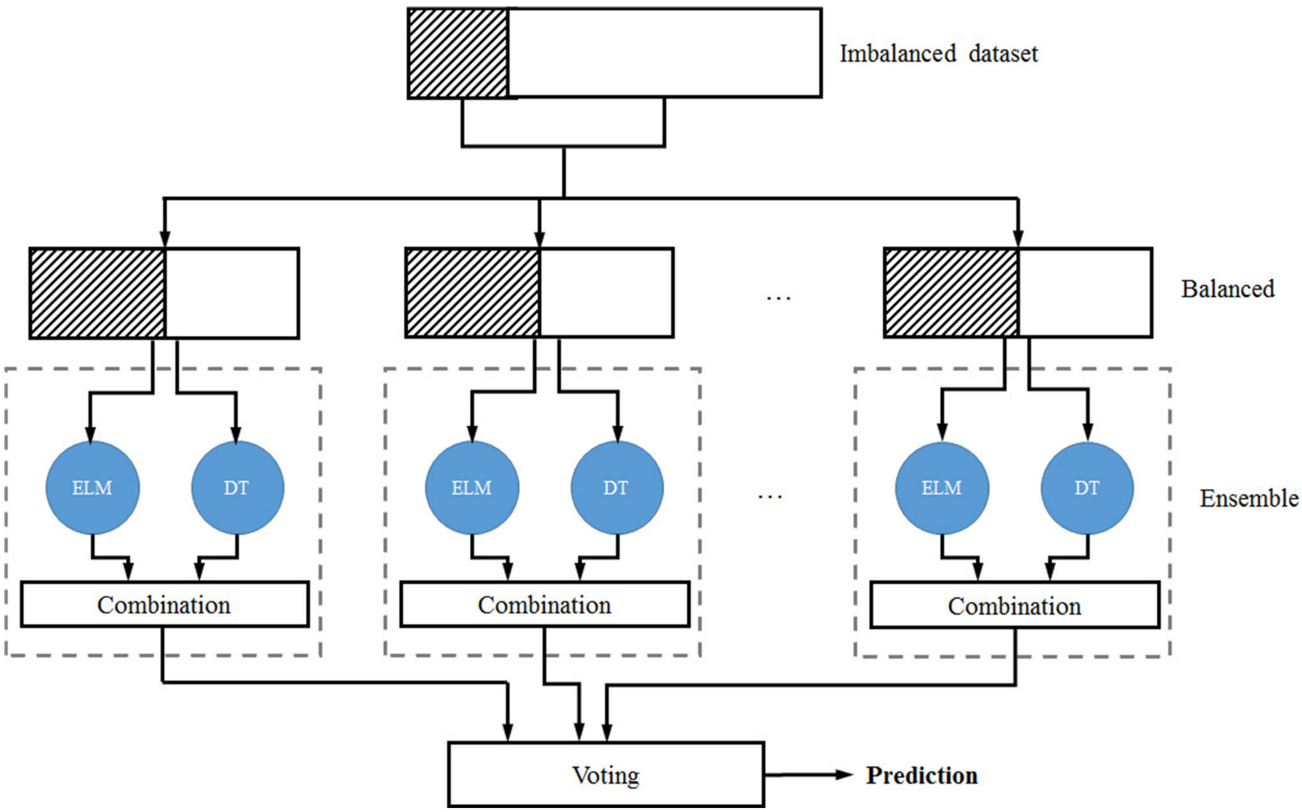


Fig. 2 Bagging ensemble method flowchart

Table 2 Confusion matrix for a binary classifier

	Predicted	
	Positive	Negative
Actual		
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

where the true positive rate is equal to the sensitivity and the false positive rate is defined as $FP_{rate} = \frac{FP}{FP+TN}$.

5 Results

In order to evaluate the effectiveness of our proposed approach, henceforth denoted bagging ensemble, we compare its performance with other well-known classifiers, namely naive Bayes, decision tree, random forest and extreme learning machine. We have evaluated each method using (1) the original data distribution, and (2) applying random undersampling (RUS) as a preprocessing technique to achieve a training dataset with balanced a priori class distribution. Our experiments were implemented using the open source machine learning library scikit-learn. All the evaluations were performed using fivefold cross-validation.

According to the results shown in Table 3, it is clear that class imbalance conditions overall performance of the model, regardless of the classifier we use. When original skewed data are employed, high accuracy scores (above 90% in all cases) and fairly poor recall scores are achieved. This behavior, sometimes referred as ‘accuracy paradox’, is caused by a high class imbalance that imposes a strong bias toward the majority (normal admission) class. When random undersampling is applied, accuracy decreases and recall increases due to the a priori class probability balancing. Tree-type algorithms (DT and RF) achieve better AUC scores when class balancing techniques are applied (increases of 3.6 and 6.8%, respectively). This improvement, on the other hand, does not occur when using naive Bayes and ELM, which perform similarly in both scenarios.

The area under the ROC curve (AUC) is the most widely used metric to evaluate readmission risk prediction in the literature. According to the results shown in Table 3, our bagging ensemble achieves the best score followed by random forest with random undersampling preprocessing. Figure 3 shows the ROC curves for different classifiers using random undersampling for data balancing. We can

Table 3 Comparison of different machine learning methods (mean ± standard deviation) measured by AUC, recall, specificity, and accuracy

Model		AUC	Recall	Specificity	Accuracy
Bagging ensemble	–	0.647 ± .01	0.474 ± .04	0.759 ± .00	0.736 ± .03
Naive Bayes	–	0.587 ± .01	0.145 ± .01	0.944 ± .00	0.917 ± .00
	rus	0.589 ± .01	0.211 ± .03	0.894 ± .00	0.869 ± .03
Decision tree	–	0.517 ± .01	0.071 ± .01	0.959 ± .00	0.929 ± .00
	rus	0.553 ± .00	0.470 ± .02	0.555 ± .00	0.647 ± .01
Random forest	–	0.559 ± .00	0.001 ± .00	0.999 ± .00	0.965 ± .00
	rus	0.627 ± .00	0.373 ± .01	0.665 ± .00	0.761 ± .01
ELM	–	0.546 ± .02	0.001 ± .00	0.999 ± .05	0.965 ± .00
	rus	0.551 ± .02	0.452 ± .1	0.626 ± 0.00	0.624 ± 0.09

rus random undersampling is applied

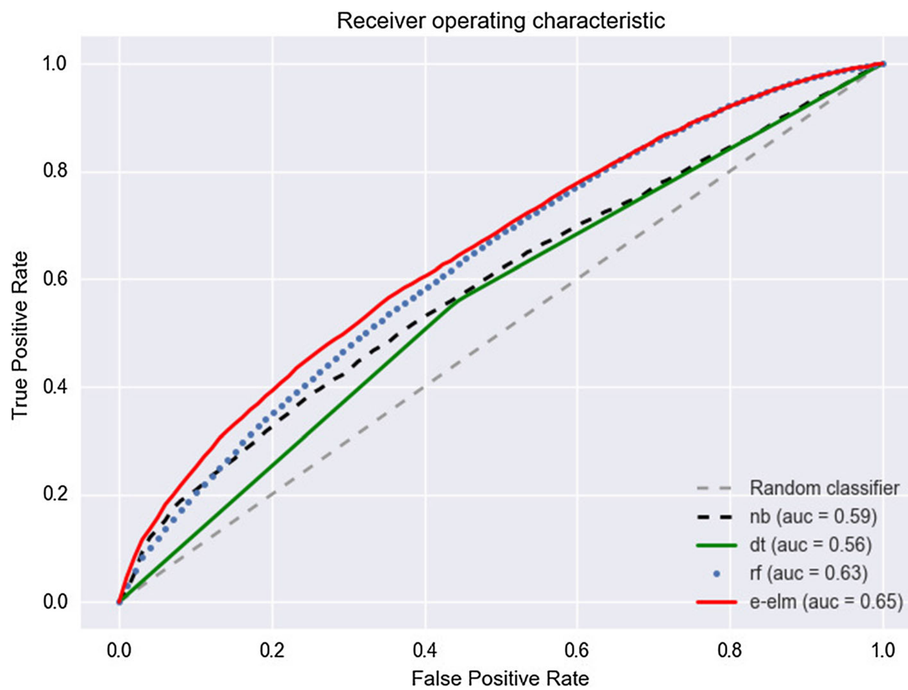


Fig. 3 Comparison of ROC curves for different methods with random undersampling

see that bagging ensemble (red) is the best performing method, followed by random forest (blue dots).

The individual classifiers with better sensitivity performance are DT and ELM with class balancing. This explains why bagging ensemble has the best sensitivity scores (47.4%). Random forest and naive Bayes, on the other hand, score poorly in comparison (37.3 and 21.1%, respectively).

When it comes to decision tree classifiers, in our preliminary experiments we have used the default configuration of the CART algorithm [22] implemented in scikit-learn. In that case, the maximum depth of the tree is not specified beforehand, so that tree's depth is set according to a certain termination criterion. In order to analyze the effect of the maximum tree depth in the overall performance of the model, we have evaluated several decision trees with different 'maximum tree depth' values. Figure 4 shows the AUC scores of decision tree classifiers trained using both original and balanced datasets. Both configurations achieve the best results at a depth of 5–10 and results get worse afterward, although trends are different. However, when we explore the behavior of recall scores, we find that classifiers trained with imbalanced dataset achieve poor results as shown in Fig. 5. Class balancing, on the other hand, improves classifiers' performance to a 55%.

In order to determinate the impact that the number of hidden units of the ELM has in the performance of our bagging ensemble, we have conducted a test consisting of measuring the recall scores of models with different hidden unit values. Figure 6 shows a peak at 30 hidden units and a plateau at around 150 units. According to the results, in our tests we have used 30 hidden units.

Fig. 4 AUC versus maximum DT depth

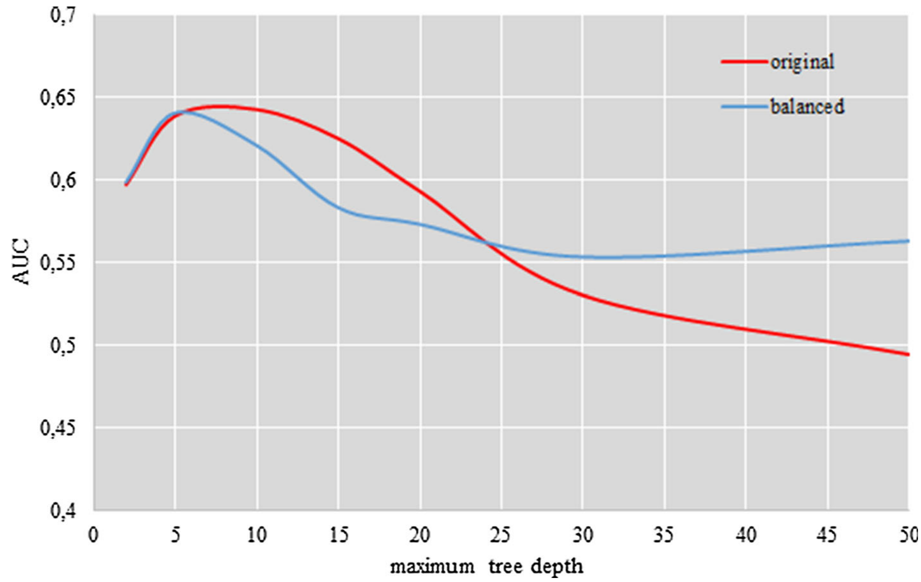
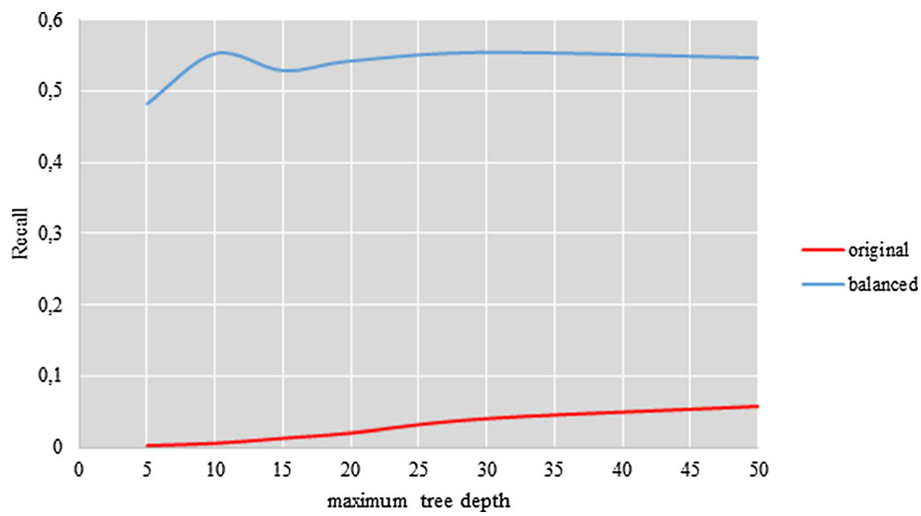


Fig. 5 Recall versus maximum DT depth



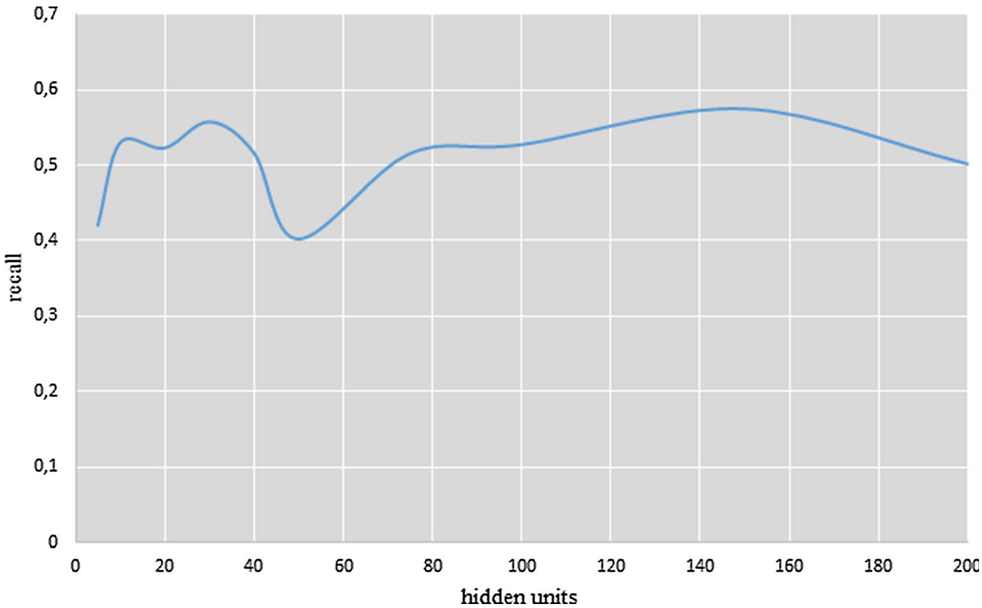
6 Conclusions

In this paper, we have presented a ensemble method that combines random undersampling with bootstrap aggregation (bagging) and an ensemble of extreme learning machine (ELM) and decision tree (DT) pairs for modeling heavily imbalanced datasets. We have presented the results of a comparative study where the performance of different classification approaches is evaluated and the effect of various parameters examined, in the context of short-time ED readmission risk prediction. Our experiments were carried out using a real clinical dataset from the Hospital Clinico Universidad de Chile.

Experiments have shown that our approach outperformed other well-known classification algorithms. The combination of two heterogeneous classifiers along with random undersampling and bagging produced a more robust classifier, achieving better overall results in terms of AUC and recall.

Class imbalance is a major problem in machine learning in general and in readmission risk prediction in particular, due to the bias it introduces toward the majority (negative) class. Our experiments demonstrate that random undersampling is an effective mechanism to overcome the class imbalance problem, according to the specific characteristics of our dataset. Although performance results are modest, we achieve AUC scores above 0.64, what is a state-of-the-art performance, according to the results presented in the review by Kansagara et al. [12]. Taking into account that our dataset is composed of pediatric and adult patients altogether, and the relatively low number of clinical indicators of the patient (such as comorbidities or treatments), our approach proves to perform reasonably well. As future work we plan to extract IF-THEN-like rules

Fig. 6 Recall versus number of hidden units in the ELM



from our model (based on the decision tree) in order to analyze them with the practitioners and get feedback for further experiments. We also plan to enhance the dataset by gathering further features extracted from the EHR. Feature subset selection techniques are to be utilized with the aim of reducing the complexity of the model as well as improving the predictive model's performance.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest

References

1. Arora S, Patel P, Lahewala S, Patel N, Patel NJ, Thakore K, Amin A, Tripathi B, Kumar V, Shah H, Shah M, Panaich S, Deshmukh A, Badheka A, Gidwani U, Gopalan R (2017) Etiologies, trends, and predictors of 30-day readmission in patients with heart failure. Am J Cardiol 119(5):760–769
2. Artetxe A, Ayerdi B, Graa M, Rios, S (2017) Using anticipative hybrid extreme rotation forest to predict emergency service readmission risk. J Comput Sci
3. Artetxe A, Beristain A, Graña M, Besga A (2016) Predicting 30-day emergency readmission risk. In: International conference on European transnational education, Springer, pp 3–12
4. Billings J, Blunt I, Steventon A, Georghiou T, Lewis G, Bardsley M (2012) Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (parr-30). BMJ Open 2(4):e001,667
5. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
8. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239
9. He H, Bai Y, Garcia EA, Li S (2008) Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: IEEE international joint conference on neural networks, 2008. IJCNN 2008, IEEE world congress on computational intelligence, IEEE, pp 1322–1328
10. Huang G, Huang GB, Song S, You K (2015) Trends in extreme learning machines: a review. Neural Netw 61:32–48
11. Huang GB, Wang DH, Lan Y (2011) Extreme learning machines: a survey. Int J Mach Learn Cybern 2(2):107–122
12. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S (2011) Risk prediction models for hospital readmission: a systematic review. JAMA 306(15):1688–1698
13. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inf Decis Mak 11(1):1
14. Lin SJ, Chang C, Hsu MF (2013) Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. Knowl Based Syst 39:214–223
15. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141
16. Mateo F, Soria-Olivas E, Martınez-Sober M, Téllez-Plaza M, Gómez-Sanchis J, Redón J (2016) Multi-step strategy for mortality assessment in cardiovascular risk patients with imbalanced data. In: European symposium on artificial neural networks, computational intelligence and machine learning
17. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD (2008) Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 21(2):427–436
18. Meadem N, Verbiest N, Zolfaghar K, Agarwal J, Chin SC, Roy SB (2013) Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients. In: Data mining and healthcare (DMH), at international conference on knowledge discovery and data mining (KDD)

19. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, Negahban SN, Krumholz HM (2016) Analysis of machine learning techniques for heart failure readmissions. Circ Cardiovasc Qual Outcomes 9:629–664
20. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
21. Shi X, Xu G, Shen F, Zhao J (2015) Solving the data imbalance problem of p300 detection via random under-sampling bagging SVMs. In: 2015 international joint conference on Neural networks (IJCNN), IEEE, pp 1–5
22. Steinberg D, Colla P (1995) Cart: tree-structured non-parametric data analysis. Salford Systems, San Diego
23. Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 23(04):687–719
24. Turgeman L, May JH (2016) A mixed-ensemble model for hospital readmission. Artif Intell Med 72:72–82
25. Urma D, Huang CC (2017) Interventions and strategies to reduce 30-day readmission rates. Hosp Med Clin 6(2):216–228
26. Wang B, Pineau J (2016) Online bagging and boosting for imbalanced data streams. IEEE Trans Knowl Data Eng 28(12):3353–3366
27. Yang Q, Wu X (2006) Ten challenging problems in data mining research. Int J Inf Technol Decis Mak 5(04):597–604
28. Yoon K, Kwek S (2007) A data reduction approach for resolving the imbalanced data issue in functional genomics. Neural Comput Appl 16(3):295–306
29. Young WA, Nykl SL, Weckman GR, Chelberg DM (2015) Using voronoi diagrams to improve classification performances when modeling imbalanced datasets. Neural Comput Appl 26(5):1041–1054
30. Zhang Y, Fu P, Liu W, Chen G (2014) Imbalanced data classification based on scaling kernel-based support vector machine. Neural Comput Appl 25(3):927–935
31. Zhang Z, Krawczyk B, Garcia S, Rosales-Perez A, Herrera F (2016) Empowering one-versus-one decomposition with ensemble learning for multi-class imbalanced data. Knowl Based Syst 106:251–263
32. Zheng B, Zhang J, Yoon SW, Lam SS, Khasawneh M, Poranki S (2015) Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Syst Appl 42(20):7110–7120