Emergency Department Readmission Risk Prediction: A case study in Chile

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

[1] Vicomtech-IK4 Research Centre, Mikeletegi Pasealekua 57, 20009 San Sebastian, Spain
aartetxe@vicomtech.org
[2] Computation Intelligence Group, Basque University (UPV/EHU) P. Manuel Lardizabal 1, 20018 San Sebastian, Spain
[3] Business Intelligence Research Center (CEINE), Industrial Engineering Department, University of Chile, Beauche 851, Santiago 8370456, Chile

Abstract. Short time readmission prediction in Emergency Departments (ED) is a valuable tool to improve both the ED management and the healthcare quality. It helps identifying patients requiring further post-discharge attention as well as reducing healthcare costs. As in many other medical domains, patient readmission data is heavily imbalanced, i.e. the minority class is very infrequent, which is a challenge for the construction of accurate predictors using machine learning tools. We have carried computational experiments on a dataset composed of ED admission records spanning more than 100000 patients in 3 years, with a highly imbalanced distribution. We employed various approaches for dealing with this highly imbalanced dataset in combination with different classification algorithms and compared their predictive power for the estimation of the ED readmission probability within 72 hour after discharge. Results show that random undersampling and Bagging (RUS-Bagging) in combination with Random Forest achieves the best results in terms of Area Under ROC Curve (AUC).

Keywords: readmission risk, imbalanced data, classification, bagging

1 Introduction

In hospitals inside public and private healthcare systems, there is a growing concern on the quality and sustainability of the service. The readmission events, defined as the recurrent visits of a patient in a time span smaller that a given threshold, has become one of the quality measures, both regarding patient attention and economical factors. In some countries, insurance companies have set a time threshold below which they decline to answer for the cost of the patient care, and the hospital must assume it. Therefore, the prediction and prevention of these events is becoming economically critical for some institutions. In other countries, healthcare quality is the primary concern, so that preventing readmissions is a measure of improved patient attention. Readmission predictors are

built by machine learning techniques, as specific two-class classifiers. A specific issue building these predictors from data is that the readmission events are much less frequent than normal admissions, i.e. the datasets are class imbalanced.

In supervised classification, data imbalance occurs 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. In healthcare, as well as in other fields (e.g. fraud detection or fault diagnosis), instances of the minority class are outnumbered by the negative instances. Also, the minority class is the target class to be predicted because it is related to the highest cost/reward events. Most classification algorithms assume equal a priori probability for all the classes, so when this premise is violated the resulting classifier is biased towards the majority class. The resulting classifier has a higher predictive accuracy over the majority class, but poorer predictive accuracy over the minority class.

The degree 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 classifier performance deteriorates even with modest class imbalance in the training data [11].

Although imbalanced data classes have been recognized as one of the key problems in the field of data mining [14], it is not usually taken into account in the literature of readmission risk prediction, despite some authors [2] have encountered class imbalance problems when building their predictive models. Some works such as [12, 15, 1] point out the existence of the class imbalance problem and propose methods to circumvent it. Nevertheless, only simple preprocessing approaches such as oversampling and under sampling are considered. Recent works [8, 10] in the field of disease risk prediction have attacked the problem of class imbalance using different preprocesing and ensemble techniques such as SMOTE or RUSBoost among others.

The main contributions of this paper are:

- A methodology proposal for overcoming the class imbalance problem based on RUSBagging
- An experimental study using real-world data where we compare the performance of different methods

The paper is organized as follows. In Section 2 we present our dataset as well as the methodological approach followed in order to build our models. Next, we describe the evaluation methodology and the experimental results. In Section 4 we discuss the conclusions and future work.

2 Materials and Methods

2.1 Experimental dataset

We used a pseudonymised dataset composed of 99858 admission records recorded between January 2013 and April 2016 in the Hospital José Joaquín Aguirre of

the Universidad de Chile, which is part of the public health system of Chile. The variables recorded in the dataset are divided into three main groups: i) Sociodemographic and administrative data, ii) Health status iii) Reasons for consultation or diagnoses made at admission. Records with missing values are discarded for this study. Table 1 shows the characteristics of the dataset and the distribution of 72-hour readmissions among different variables[1].

2.2 Data pre-processing

Data was provided in a large ASCII text file containing 156120 admission records corresponding to 102534 different patient identities. After parsing the data, we built a dataset combining admission and patient-related data. Next, we cleaned the data by removing inconsistent and missing samples. Missing values where imputed using the arithmetic mean for continuous variables and the mode for categorical 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 meaning was set to readmitted/not readmitted. Those patients returning to the ED within 72 hours after being discharged where considered readmitted, otherwise they were considered not readmitted.
Notice that a patient returning the very first day after discharge and another one returning the third day are both considered as readmitted. On the other hand, a patient returning the 73rd hour from discharge is considered as not readmitted.

2.3 Evaluation metrics

The evaluation metrics that we have used are: sensitivity, specificity, accuracy and Area Under ROC Curve (AUC), defined as follows:

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

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

  where TN is a true negative, TP a true positive, FN is a false negative and FP a false positive. In heavily umbalanced datasets it is not very meaningful because a simple strategy such as assigining each test sample to the majority class provides high accuracy.
- Sensitivity. Sensitivity is a classification performance measure defined as the proportion of correctly classified positives:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{2}$$

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

[1] Most common categorical values are only shown

Table 1. Characteristics of the dataset

Variable	All patients n=99858	Readmitted n=3425	Not readmitted n=96433	p-value
age, mean (SD)	41.0 (22.4)	36.1 (22.9)	41.2 (22.4)	<0.001
male sex (%)	44956 (45.0)	1624 (1.6)	43332 (43.4)	0.004
daytime (%)	69321 (69.4)	2171 (2.2)	67150 (67.2)	<0.001
evaluation, mean (SD)	5.0 (3.3)	4.8 (3.5)	5.0 (3.3)	0.040
fragility idx, mean (SD)	0.0 (2.5)	0.0 (2.3)	0.0 (2.5)	0.991
triage (%)				<0.001
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 (%)				<0.001
Gineco-obstetrics	236 (0.2)	6 (0.0)	230 (0.2)	
General medicine	77192 (77.3)	2458 (2.5)	74734 (74.8)	
Pedaitrics	7094 (7.1)	563 (0.6)	6531 (6.5)	
Traumatology	15336 (15.4)	398 (0.4)	14938 (15.0)	
destination (%)				<0.001
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 (%)				<0.001
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)	<0.001
tad, mean (SD)	74.1 (22.3)	67.6 (29.4)	74.3 (21.9)	<0.001
tas, mean (SD)	125.8 (35.9)	114.5 (48.8)	126.2 (35.3)	<0.001
temperature, mean (SD)	35.9 (4.5)	35.5 (5.9)	35.9 (4.4)	<0.001
heart rate, mean (SD)	87.2 (22.3)	92.7 (29.1)	87.0 (22.0)	<0.001
breath rate, mean (SD)	17.0 (5.6)	15.1 (7.6)	17.0 (5.5)	<0.001
Prevision (%)				0.408
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)	
...				

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

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

– AUC. The Area Under ROC Curve (AUC) shows the trade-off between the sensitivity or TP_{rate} and FP_{rate} (1 - specificity):

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{4}$$

where the True Positive rate is equal to the Sensitivity and the False Positive rate is defined as $FP_{rate} = \frac{\Sigma FP}{\Sigma FP + \Sigma TN}$.

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)

2.4 Learning from Imbalanced Data

The main issue of learning from imbalanced datasets is that classification learning algorithms are often biased towards the majority class and hence, there is a higher misclassification rate of the minority class instances (which is usually the most interesting ones from the practical point of view). Figure 1 depicts a taxonomy of methods developed to deal with class imbalance[9] where three main techniques are identified, namely *preprocessing*, *cost-sensitive learning* and *ensemble* techniques. We give a quick overview of the different strategies.

Preprocessing

Methods following this strategy carry out resampling of the original dataset in order to change the class distribution. Resampling techniques can be divided into three groups: i) *Undersampling techniques*, consisting on deleting instances of the majority class, ii) *Oversampling techniques*, that replicate or create new instances of the minority class, such as the Synthetic Minority Over-sampling Technique (SMOTE) [4], and iii) *Hybrid techniques*, those that combine both resampling techniques.

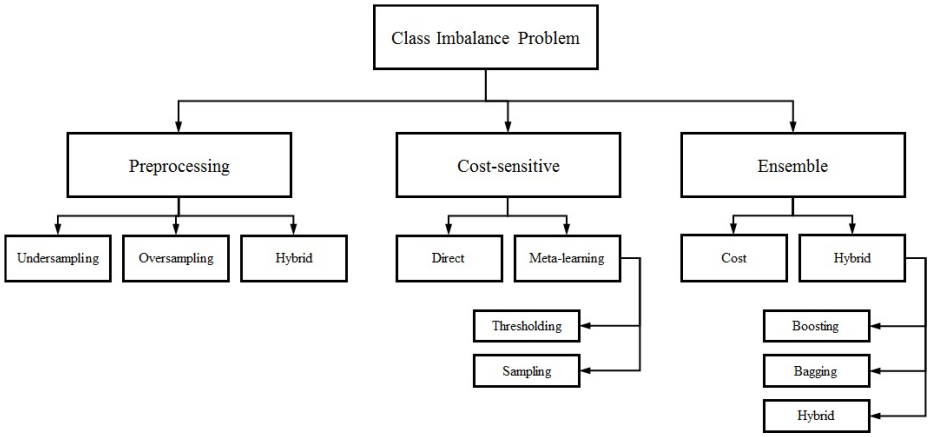

Fig. 1. Taxonomy of Class imbalance problem addressing techniques as proposed in [9]

Cost-sensitive learning

The strategy followed by cost-sensitive learning methods is to assign different cost values to each class misclassifications, so that the bias towards the majority class is balanced by the lower cost of misclassifications. A cost matrix is build assigning cost values to the entries of the confussion matrix giving (see Table 2). The usual approach is to heavily penalize misclassifications of the minority class. They are categorized into the following groups:

- Direct methods, that introduce the misclassification cost within the classification algorithm.
- Meta-learning, where the algorithm itself is not modified. Instead, a preprocessing (or postprocessing) mechanism is introduced to handle the costs. Meta-learning methodologies can be divided into two categories, namely *thresholding* and *sampling*.

Ensemble classifiers

Ensemble methods rely on the idea that the combination of many "weak" classifiers can improve the performance of a single classifier [6]. They are divided in two groups, namely *cost-sensitive* ensembles and *data and algorithmic* approaches.

- Cost-sensitive ensemble techniques, are analogous to cost-sensitive methods mentioned earlier, although in this case, the cost minimization is undertaken by the boosting algorithm.
- Data and algorithmic approaches, which embed a data preprocessing technique in an ensemble algorithm. Depending on the ensemble algorithm they use, three groups are identified: i) Boosting, ii) Bagging and iii) Hybrid.

Bagging [3] 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 data-set or bag mantained the size of the original data-set. 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. In this group we find, among others, algorithms like SMOTEBoost [5] or UnderBagging [13] which embed undersampling within the ensemble algorithm. We propose RUSBagging which carries out a random undersampling for each bag generated in the ensemble creation. An individual weak classifier is trained from the data in each bag.

3 Experimental results

In this section we present the results obtained when trying to predict the readmission risk before 72 hours over the dataset presented in the previous section. We have tested two data balancing methods: random undersampling (RUS) and random undersampling embedded in a bagging approach. We used the following well-known classification algorithms, implemented in the open source machine learning library scikit-learn[4]:

1. Decision Tree (DT), setting Gini impurity as splitting criterion
2. Random Forest (RF), setting Gini impurity as splitting criterion and number of estimators=10

The models were evaluated using 10-fold cross-validation, performing 10 independent executions. Accuracy, specificity, sensitivity and AUC were calculated for each execution, so average and standard deviation were computed. In order to statistically compare results we employed an Analysis of Variance (ANOVA) approach.
The following data balancing approaches were compared: i) Original dataset with its imbalanced class distribution, ii) Undersampling with random undersampling and iii) RUSBagging. Table 3 shows the average accuracy, sensitivity, specificity and AUC along with its respective standard deviation, for each method and classifier.

3.1 Comparison of classifiers

According to the results shown in Table 3, both classification algorithms, Random Forest achieve significantly better results (p¡0.001) than Decision Trees looking at the AUC. Though DT performs better in the original dataset (anyhow both classifiers perform poorly), when preprocessing and ensemble approaches

[4] http://scikit-learn.org/

Table 3. Mean (± standard deviation) of performance metrics for each data balance method and classifier model configuration

method	classifier	accuracy	specificity	sensitivity	AUC
None	DT	.9293 ± .0006	.9599 ± .0006	.0673 ± .0030	.5136 ± .0017
	RF	.9655 ± .0001	.9997 ± .0001	.0012 ± .0003	.5005 ± .0002
RUS	DT	.5578 ± .002	.5574 ± .002	.5674 ± .012	.5624 ± .005
	RF	.6622 ± .0016	.6676 ± .0018	.5086 ± .0096	.5881 ± .0043
RUSBagging	DT	.6530 ± .0011	.6576 ± .0012	.5244 ± .0079	.5910 ± .0037
	RF	.7679 ± .0014	.7796 ± .0015	.4359 ± .0041	.6078 ± .0020

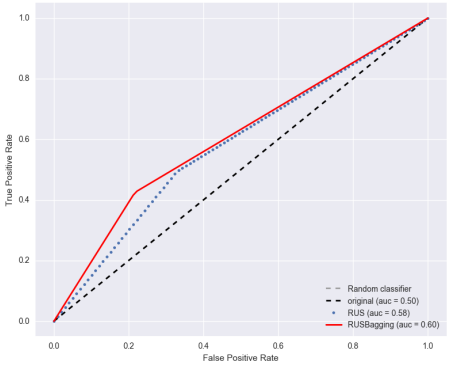

Fig. 2. ROC curve for DT using undersampling, RUSBagging and original

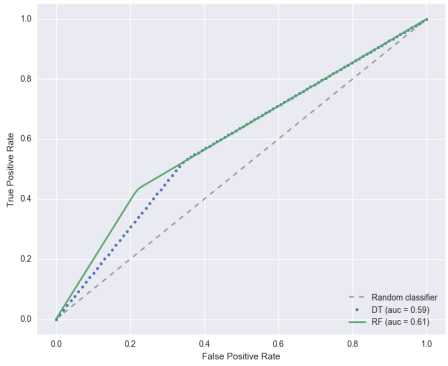

Fig. 3. ROC curve for DT and RF algorithms using RUSBagging method

are utilized RF performs much better. As shown in Figure 3, the AUC is significatively greater for RF when RUSBagging is used, however, sensitivity is sacrificed if compared with DT. Overall, results are poor, however they compare well with the state of the art in readmission prediction. In a recent review [7], most studies reported performances measured by AUC near 0.5, with some outliers achieving a maximum of 0.7.

3.2 The effect of preprocessing and ensemble methods

Several conclusions can be extracted from the results shown in Table 3.

- The models trained without modifying the original class distribution were clearly biased towards the majority class. Although accuracy scores were high (>90%), specificity was close 100% while sensitivity tended to zero. Thus, according to the AUC scores, models performed similar or just slightly better than a random classifier.
- Using random undersampling for class balancing had a direct effect in the performance of the resulting model. Results show that both DT and RF get better AUC scores, 0.56 and 0.58 respectively, and sensitivity increases considerably. However, as could be expected, both accuracy and specificity tend to decrease.

– RUSBagging, which embeds random undersampling within a bootstrap aggregating algorithm, outperforms both previous methodologies. According to the AUC scores, the combination of RUSBagging and Random Forest shows the best performance with a mean of 0.60.
– The performance of the models considering the AUC metric, suggests poor discrimination ability. Nevertheless, a systematic review on risk prediction models for hospital readmission documented similar AUC scores (ranging from 0.50 to 0.70) in most of the studies [7].

4 Conclusions and future work

In this paper we have presented the results of readmission prediction based on a real dataset from a hospital in Santiago, Chile. To overcome the class imbalance problem we propose an approach called RUSBagging, that carries out random undersampling for each bag in a bagging ensemble training.

Results show that RUSBagging in combination with Random Forest significantly improves predictive performance in the context of a highly imbalanced dataset. Nevertheless, our model has shown limited predictive ability for clinical purposes, what seems to be related with the inherent dfficulties and limitations of the readmission risk prediction problem. We have attacked one major issue (data imbalance) but others such as the appriate selection and measurement of varaibles remain untouched in this paper. In order to validate the usefulness of our presented approach, we plan to gather and include additional baseline status and administrative data, to perform a prospective study. Future work will also include an extension of our comparative study including new methodologies and classifiers.

References

1. Artetxe, A., Beristain, A., Graña, M., Besga, A.: Predicting 30-day emergency readmission risk. In: International Conference on EUropean Transnational Education. pp. 3–12. Springer (2016)
2. Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., Bardsley, M.: Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (parr-30). BMJ open 2(4), e001667 (2012)
3. Breiman, L.: Bagging predictors. Machine learning 24(2), 123–140 (1996)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357 (2002)
5. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 107–119. Springer (2003)
6. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(4), 463–484 (2012)

7. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., Kripalani, S.: Risk prediction models for hospital readmission: a systematic review. Jama 306(15), 1688–1698 (2011)
8. Khalilia, M., Chakraborty, S., Popescu, M.: Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making 11(1), 1 (2011)
9. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113–141 (2013)
10. Mateo, F., Soria-Olivas, E., Martınez-Sober, M., Téllez-Plaza, M., Gómez-Sanchis, J., Redón, J.: 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 (2016)
11. Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks 21(2), 427–436 (2008)
12. Meadem, N., Verbiest, N., Zolfaghar, K., Agarwal, J., Chin, S.C., Roy, S.B.: 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) (2013)
13. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on. pp. 324–331. IEEE (2009)
14. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making 5(04), 597–604 (2006)
15. Zheng, B., Zhang, J., Yoon, S.W., Lam, S.S., Khasawneh, M., Poranki, S.: Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Systems with Applications 42(20), 7110–7120 (2015)