ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Multi-step strategy for mortality assessment in cardiovascular risk patients with imbalanced data

Fernando Mateo[1], Emilio Soria-Olivas[2], Marcelino Martínez-Sober[2], María Téllez-Plaza[1], Juan Gómez-Sanchis[2] and Josep Redón[1] *

1- Instituto de Investigacion Sanitaria (INCLIVA)
Avda. Menéndez Pelayo 4, 46010 Valencia - Spain

2- University of Valencia - Department of Electronics Engineering
Avda. Universidades s/n, 46100 Burjassot - Spain

Abstract. The assessment of mortality in patients with cardiovascular disease (CVD) risk factors is typically a challenging task given the large amount of collected variables and the imbalance between classes. This is the case of the ESCARVAL-RISK dataset, a large cardiovascular follow-up record spanning 4 years. This study intends to give insight into: a) the performance of variable selection methods, b) the best class balancing method and c) choosing an adequate classifier to predict mortality. We conclude that combining ADASYN with SVM classifiers without and with AUC score-based feature selection, and RUSBoost combined with boosting tree ensembles are the most suitable methodologies among the tested.

1 Introduction

The identification of patients with increased risk of mortality based on health indicators constitutes a typical scenario of imbalanced class distributions, i.e. when there exists a majority (negative) class that dominates over a minority (or positive) class with abnormal or outstanding information. Fraud detection, anomaly detection and network intrusion are other typical examples.

This paper deals with the classification of patients with cardiovascular disease (CVD) risk factors and their mortality when the collected data are imbalanced. CVDs are the first cause of mortality and disease burden worldwide. Identifying susceptible individuals allows to implement high-risk preventive strategies.

In the literature it is possible to find several approaches to heart disease risk assessment as a function of risk factors evaluated from clinical records and their relevance, using machine learning methods [1, 2], multiple regression models [3, 4] and G-estimations [5], and other that deal with how to cope with imbalanced datasets for medical diagnosis [6]. There are also studies on the estimation of the outcome of cardiac rehabilitation in patients by balancing data and applying machine learning techniques [7].

*This work has been supported by GRANT 15_SIMBAD-SORIA-TELLEZ-2015: SIMBAD: *Sistema basado en datos como ayuda a la decisión clínica*, programme VLC-Bioclinic. Special thanks must be given to Escarval Study Group: D. Orozco-Beltrán, V. Gil-Guillén, J. Navarro-Pérez, V. Pallarés, F. Valls, A. Fernández, C. Sanchis, J. M. Martín-Moreno, G. Sanz, A. Domínguez-Lucas, M. Téllez-Plaza and J. Redón.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

The goal of this paper is, in the first place, to process the data and extract information and dependencies between variables to determine the most relevant ones according to well-known selection criteria. The next step is to evaluate several class balancing methods. Finally, several classifiers will be tested and the best overall performer will be selected.

The paper is divided as follows: Section 2 describes the origin of the data, Section 3 presents the proposed methodology to perform all the steps of the classification process, Section 4 discusses which are the most relevant results and determines the optimal methodology for the prediction and finally Section 5 draws some concluding remarks.

2 The ESCARVAL-RISK dataset

The data used in this study correspond to a very complete sample of patients receiving healthcare at the Valencian Health Agency [8], with at least one of the following CVD risk factors: hypertension, diabetes mellitus and/or dyslipidemia. After preprocessing and removal of samples with critical missing data, the final dataset contains $N = 54{,}678$ samples, $d = 109$ input variables and a binary output variable (mortality), which is positive for less than 2% of the participants.

3 Methodology

The three-step strategy to be followed is represented in Fig. 1. Feature selection gives insight into the actual relevant risk factors and greatly alleviates the computational cost of the learning algorithms. Class balancing methods typically oversample the minority class or undersample the majority class. It has proven useful to improve predictions depending on the chosen classifier but it sometimes leads to over generalization [9] when synthetic samples are created, and specially when the minority class samples are too scarce to adequately represent its distribution. Support Vector Machine (SVM) classifiers and Random Forests (RF) tend to profit from class balancing while Logistic Regression models are not the best choice [10]. After this step, several state-of-the-art classifiers will be evaluated, according to their general performance with both classes, but emphasizing their performance with the positive class as in medical diagnosis it is necessary to be conservative and minimize the false negatives.

3.1 Feature selection

Subsets of features have been extracted according to the following criteria:
1. *Baseline variables* based on clinical heuristics (BAS). These variables have proven useful in previous studies to establish a prospective association of renal function with mortality[1] [8].

[1]Clinically meaningful variables to predict mortality and cardiovascular risk from renal function and, in particular, from estimated glomerular filtration rate (eGFR).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

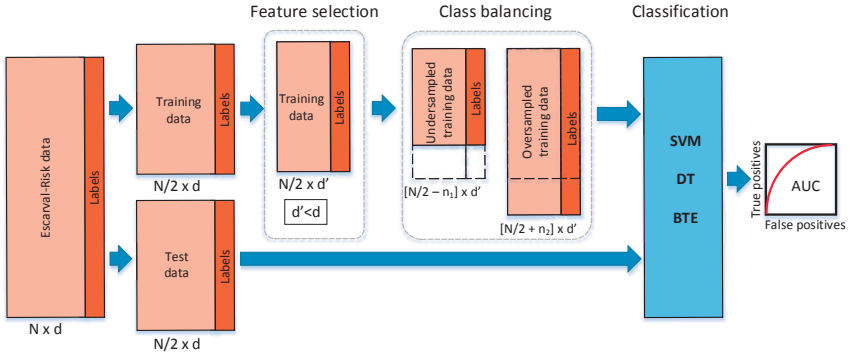


Fig. 1: Flow chart of the proposed multi-step strategy. After an initial hold-out partition of the data (50% of the samples for training/test), the first step is the feature selection on the training data. Next, class balancing methods are applied. Finally, the classification rate is evaluated on the test data.

2. *Mutual Information (MI)*: Ranking the input variables according to their MI with the output and using heuristics to choose a suitable threshold.
3. *Statistical Dependency (SD)*: Ranking the variables according to the maximization of the SD criterion, connected with the maximum joint probability of the input variables and the output, and selecting a suitable threshold.
4. *ReliefF (REL)*: Feature ranking according to a nearest neighbors formulation which modifies the weights of a feature depending on the differences between its closest instances belonging to the same or to different class.
5. *AUC score (AUC)*: Maximization of the AUC variation of the ROC associated with the positive class as an indicative of the joint influence in the sensitivity and specificity of the classifier.

3.2 Class balancing

Both oversampling and undersampling methods have been assessed for class balancing purposes:
1. *Synthetic Minority Over-sampling Technique (SMOTE)*: The minority class is oversampled by creating random "synthetic" examples [11].
2. *Adaptive synthetic sampling approach for imbalanced learning (ADASYN)*: The idea is to generate more synthetic data for the minority class examples that are harder to learn compared to those that are easier to learn [12].
3. *Random Undersampling Boosting (RUSBoost)*: RUSBoost combines random undersampling (which arbitrarily removes samples from the majority class) with the principles of the SMOTEBoost algorithm, i.e. it applies an intelligent oversampling technique, as compared to the randomness of SMOTE, which helps to balance the class distribution [13].

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Classifier	TPR (%)	TNR (%)	PPV (%)	NPV (%)	Error (%)	AUC
SVM	48.41	100	100	99.10	0.89	0.85
DT	33.97	99.13	40.61	98.85	1.99	0.67
BTE	23.14	100	100	98.67	1.32	0.95

Table 1: Classification results on the test set with all variables.

3.3 Classifiers

The performance obtained from several types of classifiers has been compared:
1. *Support Vector Machine (SVM)*: The SVM binary classifier tries to solve the primal problem by finding the optimal separation hyperplane between the two classes in a higher dimensional dual space.
2. *Decision Tree (DT)*: Binary classification model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
3. *Boosting tree ensemble (BTE)*: A fitted ensemble of trees combining trained weak learner models and data on which these learners were trained. The ensemble prediction is made by aggregating predictions from its weak learners. Boosting involves incrementally building the ensemble by training each new model instance to emphasize the training instances that previous models misclassified.

Linear logistic regression models have not been included in the analysis as they do not benefit from the class balancing methods.

4 Results

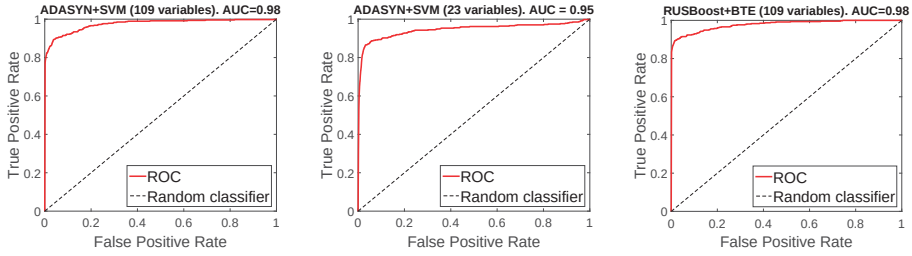
The data were randomly divided at 50%/50% ratio into training and test sets. Ten different seeds were used for robustness and reproducibility and the results were averaged. The objective is to determine the combination of methods that maximizes the area under the ROC curve (AUC score), maximizes the negative predictive value (NPV) rate and finally maximizes the true positive rate (TPR). Initially, the different classifiers have been evaluated on the data with all the variables. The test results are listed in Table 1. The best performing method is BTE in terms of AUC. In preliminary tests, RF were essayed as a bagging alternative to BTE. The results of both methods were similar when using all variables but the performance of RF with fewer features produced a much higher FN rate. SVM seems a potential candidate too because of the higher TPR in comparison with BTE. Therefore, BTE and SVM are the potential best classifiers.

The next step was to assess the diverse feature selection methods in combination with class balancing techniques and evaluate the classification performance with SVM and BTE. The candidate strategies were: SMOTE + SVM, ADASYN + SVM and RUSBoost + BTE, as RUSBoost's weak learner architecture favours the simpler tree ensembles rather than SVMs. Table 2 summarizes the results of these strategies with the different feature selection criteria. The best performing combinations of methods are ADASYN + SVM and RUSBoost + BTE with all the variables when basing the decision firstly on the maximum AUC score (0.98), secondly on the minimum amount of FNs (maximum NPV) and thirdly on the

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Method	FS	N° vars	TPR (%)	TNR (%)	PPV (%)	NPV (%)	AUC
SMOTE + SVM	-	109	67.94	99.97	97.56	99.44	0.86
	AUC	23	59.02	99.56	70.20	99.28	0.94
	REL	31	0	100	-	98.28	0.6
	BAS	16	0	100	-	98.28	0.55
	MI	20	46.28	98.91	42.66	99.06	0.77
	SD	25	46.92	99.03	45.85	99.07	0.79
ADASYN + SVM	-	109	73.89	99.94	95.87	99.54	**0.98**
	AUC	23	**75.58**	98.65	49.51	**99.57**	0.95
	REL	31	61.57	87.11	7.73	99.23	0.84
	BAS	16	63.91	85.91	7.37	99.27	0.84
	MI	20	65.39	97.00	27.67	99.38	0.85
	SD	25	65.61	97.26	29.54	99.38	0.84
RUSBoost + BTE	-	109	74.31	100	100	99.55	**0.98**
	AUC	23	62.85	99.61	73.82	99.35	0.94
	REL	31	55.20	95.93	19.20	99.19	0.87
	BAS	16	60.08	84.33	6.30	99.18	0.81
	MI	20	60.72	95.62	19.56	99.29	0.84
	SD	25	61.15	95.84	20.47	99.29	0.84

Table 2: Classification results on the test set with class balancing and feature selection (FS). The best results are highlighted in bold.



(a) (b) (c)

Fig. 2: ROC curves of the best performing methods on the test set: a) ADASYN + SVM with all variables, b) ADASYN + SVM with 23 selected variables using the maximum AUC influence criterion, c) RUSBoost + BTE with all variables.

amount of hits (maximum TPR). ADASYN + SVM with AUC score feature selection represents an interesting trade-off. Although conservative and not the best in terms of AUC (0.95), it achieves the highest NPV and TPR. Additionally, it benefits from a reduced computational cost. In Fig. 2 we represent the ROC curves of these three best strategies. It is worth to point out that the baseline feature selection based on renal function performs the worst.

5 Conclusions

A large record of patients with CVD risk factors and their associated mortality risk has been analyzed by means of a multi-step strategy that combines feature

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

selection, class balancing and classification. From the results of the analysis, we conclude that the best strategies based purely on the AUC and error rate correspond to skipping feature selection and directly apply class balancing and classification with ADASYN + SVM and RUSBoost + BTE (AUC = 0.98). However, the best trade-off considering a good AUC (0.95), optimal TPR and NPV and reduced computational cost is ADASYN + SVM with feature selection using the AUC score influence to rank the inputs.

References

[1] J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, H.-J. Dai, and C.-Y. Hsu. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. *BioMed Research International*, 2015(636371), 2015.

[2] L. J. Mena, E. E. Orozco, V. G. Felix, R. Ostos, J. Melgarejo, and G. E. Maestre. Machine learning approach to extract diagnostic and prognostic thresholds: Application in prognosis of cardiovascular mortality. *Comput. Math. Methods Med.*, 2012(750151), 2012.

[3] P. Jousilahti, J. Tuomilehto, E. Vartiainen, J. Pekkanen, and P. Puska. Body weight, cardiovascular risk factors, and coronary mortality. 15-year follow-up of middle-aged men and women in eastern Finland. *Circulation*, 93(7):1372–1397, 1996.

[4] P. Jousilahti, J. Tuomilehto, E. Vartiainen, J. Pekkanen, and P. Puska. Sex, age, cardiovascular risk factors, and coronary heart disease. a prospective follow-up study of 14786 middle-aged men and women in Finland. *Circulation*, 99(9):1165–1172, 1999.

[5] K. Tilling, J. A. C. Sterne, and M. Szklo. Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-estimation. *American Journal of Epidemiology*, 155(8):710–718, 2001.

[6] L. Mena and J. A. González. Machine learning for imbalanced datasets: Application in medical diagnostic. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS-2006)*, pages 574–579, 2006.

[7] A. Van, V. C. Gay, P. J. Kennedy, E. Barin, and P. Leijdekkers. Understanding risk factors in cardiac rehabilitation patients with random forests and decision trees. In *Proceedings of the 9th Australasian Data Mining Conference (AusDM'11)*, pages 11–22, 2011.

[8] V. Gil-Guillen, D. Orozco-Beltran, J. Redon, S. Pita-Fernández, J. Navarro-Pérez, V. Pallares, F. Valls, C. Fluixa, A. Fernandez, J. M. Martin-Moreno, M. Pascual de-la Torre, J. L. Trillo, R. Durazo-Arvizu, R. Cooper, M. Hermenegildo, and L. Rosado. Rationale and methods of the cardiometabolic valencian study (Escarval-Risk) for validation of risk scales in mediterranean patients with hypertension, diabetes or dyslipidemia. *BMC Public Health*, 10(717), 2010.

[9] V. López, A. Fernández, S. García, V. Palade, and F. Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[10] G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[12] H. He, Y. Bai, E. A. Garcia, and S. Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks (IJCNN-2008)*, pages 1322–1328, 2008.

[13] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics. Part A.*, 40(1):185–197, 2010.