ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

A Novel One-Class Classification Framework for Highly Imbalanced Binary Outcomes: the OC-Cat Approach

Federico Fassio[1], Jessica Leoni[2], Rebecca Fattore[3], Giovanni Scaglione[3], Giovanni De Capitani[3], Fabio Borgonovo[3], Claudia Conflitti[4], Daniele Zizzo[3], Antonio Gidaro[5], Maria Calloni[5], Francesco Casella[5], Chiara Cogliati[5], Andrea Gori[3,6], Antonella Foschi[3], Marta Colaneri[3,6], Valentina Breschi[7]

(1) Department of Public Health, Experimental and Forensic Medicine, Section of Biostatistics and Clinical Epidemiology, University of Pavia, Pavia, Italy
(2) Department of Electronics, Information and Bioengineering (DEIB) Politecnico di Milano, Milan, Italy
(3) Department of Infectious Diseases, Unit II, L. Sacco Hospital, ASST Fatebenefratelli Sacco, Milan, Italy
(4) National PhD Programme in One Health Approaches to Infectious Diseases and Life Science Research, Department of Public Health, Experimental and Forensic Medicine, University of Pavia, 27100 Pavia, Italy
(5) Division of Internal Medicine, Luigi Sacco Hospital, University of Milan, Milan, Italy
(6) Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy
(7) Department of Electrical Engineering, University of Technology, Eindhoven, Netherlands

these Authors contributed equally to this work

CORRESPONDING AUTHOR: *Fassio Federico, federico.fassio01@universitadipavia.it*

INTRODUCTION

Extremely rare events can challenge traditional classification models, which may exhibit reduced power in highly unbalanced datasets (i.e., when two or more target groups are unevenly represented). Moreover, this effect seems to be accentuated by the reduction of the sample size. Some of the easiest and intuitive methods proposed to handle unbalanced datasets, while still using a classical statistical models, are random under- or oversampling or hybrid methods[1]. Alternatively, other approaches have been proposed with different strategies, such as ensemble models (e.g. AdaBoost, XGBoost), or novelty detection models[2].

In medicine, this kind of scenario can occur when analysing catheter related/associated blood stream infections (CRBSI/CABSI), whose incidence usually remains <1/1000 catheter days[3], but could be higher in very frail patients[4]. Catheter insertion has a potential risk of complications and longer hospitalization: the use of decision-making algorithms is of great importance in order to avoid complications for these patients[5].

OBJECTIVES

The main purpose of our study is to adopt a novel anomaly detection model focused on binary/categorical covariates to predict risk of CRBSI/CABSI occurrence at baseline. To reach this result, we use a combined approach: features reduction, novelty detection algorithm and importance grid for model explainability.

METHODS

Data from hospital patients who received a vascular access device (VAD) placements at the University Hospital Luigi Sacco in Milan between January 2021 and January 2025 were analysed. All patients underwent central or peripheral catheterization in a non-ICU department. Parameters were collected at catheter insertion: age, sex, any major comorbidities, active intravenous drug usage, parenteral nutrition, regimen of hospitalization, transfer from the ICU, type of catheter, number of lumens, tunnel, exit site and number of placement attempts. All continuous variables were discretized into categorical format, yielding 29 Boolean and 2 categorical features.

DOI: 10.54103/2282-0930/29381

The designed framework (OC-Cat) combines:

1. a graph-search-based feature selection method;
2. a one-class soft classifier designed (based on characterization of patients who didn't incurred in catheter infection);
3. a feature ranking that clarifies the classifier's decisions by ordering features based on their unique role in identifying uninfected patients.

In details:

1. we assess the redundancy of each pair of features using the excess over independence metric[6]. Then, we design a undirected connected graph where each node represents a feature, and the edge weights reflect the excess over independence between feature pairs. From each node, we apply the Bellman-Ford algorithm[7] to find the shortest closed path. Among all paths, we select the one that best represents the original data based on the Bayesian Information Criterion (BIC). The features included in this optimal path constitute the final selected feature set;
2. to design the soft-classifier, we rely on the assumption that a higher occurrence of a specific feature combination in majority class records (uninfected) implies that each new instance with those values is less likely to be infected. The learning phase consists of estimating the probability for a majority-class record occurring, given the distribution of uninfected patients. The prediction phase, instead, consists of estimated the majority-class probability for a new record (based on its i th attribute combination) using a weighted inverse Hamming distance [8]. The weight increases with the record's frequency among uninfected patients;
3. accordingly, the method ranks features based on a tailored definition of importance, stating that a feature - or a features set - is more important if it consistently exhibits the same value in majority-class data. To achieve this, we build a tree where nodes represent subsets of features, and each step measures the contribution of each new feature in reducing the majority-class data entropy. Last, once exploring all feature combinations and identifying the path with minimal entropy, the algorithm reports the features ranking as the order in which features appear along the path: from the root (most important) to the leaf (least important).

To evaluate the framework performance in terms of one-class classification, we compared OC-Cat probability distribution with that obtained from Isolation Forest (iForest) and One-Class Support Vector Machine (OCSVM). For the analysis, dataset was split into training and test set (August 2023 as threshold: ~75% vs 25%).

RESULTS

Data from 2836 hospitalized patients with VADs were retrieved. After keeping only the first VAD placement for each patients, we considered 2275 subjects (1222 women and 1053 men between 18 to 101 years) Among them, 148 become infected: 62 patients developed a CRBSI, 80 a CABSI and 3 both. In the first step, our approach retained 16 out of 29 variables, which were then inserted in the novel model in the second step. Figure 1 displays the risk factor index distributions for the training and test sets of our model, iForest, and OCSVM, along with their respective ROC curves. Lastly, catheter insertion site (upper vs lower limb vs neck), biological sex, hypertension, Charlson Comorbidity index, neurological disease and diabetes resulted the first most characterizing feature.

CONCLUSION

Our model introduces a novel, integrated approach for both characterizing and forecasting outcomes under severe imbalance in the target variable. It outperformed the iForest and OCSVM models applied to categorical and Boolean variables in a specific clinical contest. We are currently conducting further analysis and refinements to optimize performance on both our internal and external datasets, enhancing the model's generalization.

REFERENCES

1. Hoens TR, Chawla NV. Imbalanced Datasets: From Sampling to Classifiers. Imbalanced Learning, John Wiley & Sons, Ltd; 2013, p. 43–59.
2. Pimentel MAF, Clifton DA, Clifton L et al. A review of novelty detection. Signal Processing 2014;99:215–49.
3. Dreesen M, Foulon V, Spriet I et al. Epidemiology of catheter-related infections in adult patients receiving home parenteral nutrition: a systematic review. Clin Nutr 2013;32:16–26.
4. Zhao VM, Griffith DP, Blumberg HM, et al. Characterization of Post-Hospital Infections in Adults Requiring Home Parenteral Nutrition. Nutrition 2013;29:52–9.
5. Catho G, Fortchantre L, Teixeira D, et al. Surveillance of catheter-associated bloodstream infections: development and validation of a fully automated algorithm. Antimicrob Resist Infect Control 2024;13:38.
6. Maron ME, Kuhns JL. On Relevance, Probabilistic Indexing and Information Retrieval. J ACM 1960;7:216–44.
7. Bellman R. On a routing problem. Quart Appl Math 1958;16:87–90.
8. Hamming RW. Error Detecting and Error Correcting Codes. Bell System Technical Journal 1950;29:147–60.
9. Robertson SE. The probability ranking principle in IR. Journal of Documentation 1977;33:294–304.

DOI: 10.54103/2282-0930/29381

ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII[th] SISMEC Congress - Vol. 20 Suppl. 1

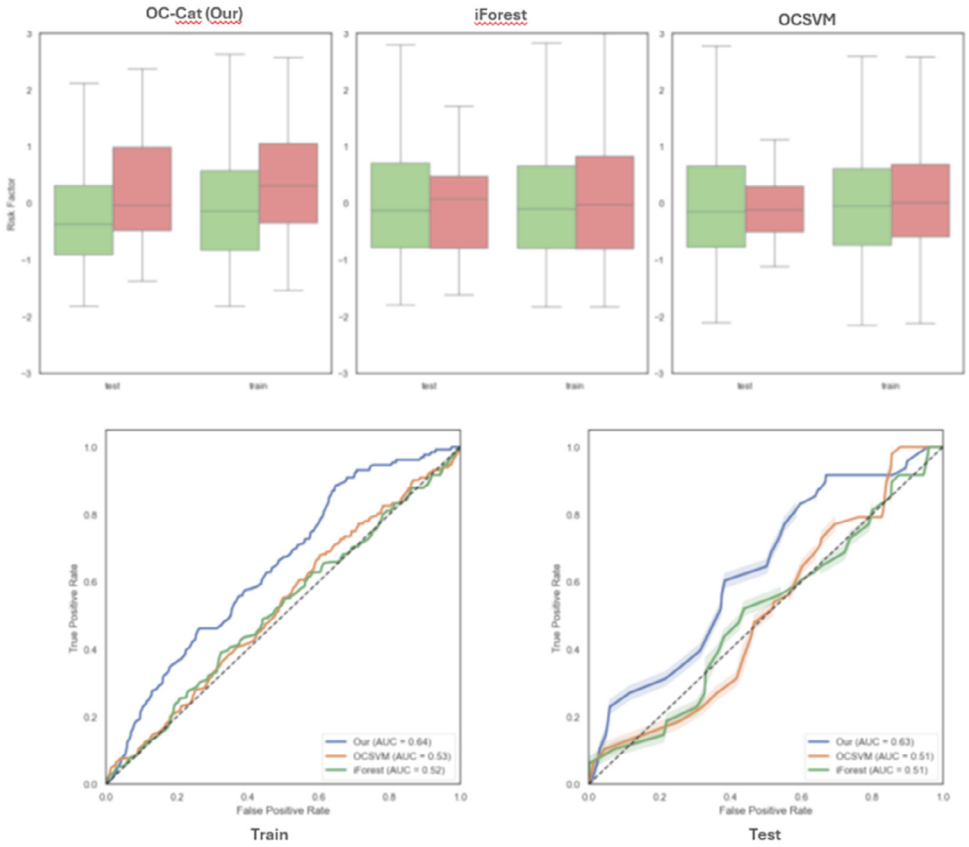

Figure 1. Risk distributions for the two patient classes (top of the figure): patients who didn't experience a CRBSI/CABSI infection during the follow-up period (green boxplot) vs patients who experienced a catheter infection (red boxplot). The distributions were calculated separately for the training set (left) and the test set (right). The same analyses, using the selected covariates, were conducted with two other models: Isolation Forest (iForest) and One-Class Support Vector Machine (OCSVM). In the lower part of the figure, the Receiver Operating Characteristic (ROC) curves for the three models are shown, separately for the training set (left) and the test set (right)

DOI: 10.54103/2282-0930/29381