proceedings

MDPI

Proceedings

Regression Tree Based Explanation for Anomaly Detection Algorithm [†]

Iñigo López-Riobóo Botana [1,*], Carlos Eiras-Franco [2] and Amparo Alonso-Betanzos [2]

[1] Research group LIDIA, Universidade da Coruña, Campus Elviña, 15071 A Coruña, Spain
[2] CITIC Research Center, Universidade da Coruña, 15071 A Coruña, Spain; carlos.eiras.franco@udc.es (C.E.-F.); ciamparo@udc.es (A.A.-B.)
[*] Correspondence: inigo.lopezrioboo.botana@udc.es

[†] Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 18 August 2020

check for updates

Abstract: This work presents EADMNC (Explainable Anomaly Detection on Mixed Numerical and Categorical spaces), a novel approach to address explanation using an anomaly detection algorithm, ADMNC, which provides accurate detections on mixed numerical and categorical input spaces. Our improved algorithm leverages the formulation of the ADMNC model to offer pre-hoc explainability based on CART (Classification and Regression Trees). The explanation is presented as a segmentation of the input data into homogeneous groups that can be described with a few variables, offering supervisors novel information for justifications. To prove scalability and interpretability, we list experimental results on real-world large datasets focusing on network intrusion detection domain.

Keywords: XAI; CART; anomaly detection; scalability; distributed computing; Apache Spark

1. Introduction

Anomaly Detection is an old discipline that has become relevant in situations in which datasets are huge and contain unexpected events carrying important information. These methods have found applications in fields such as network intrusion detection, and surveillance, among others. Several machine learning models are available [1,2], but despite being capable of offering very effective detection, most of these algorithms are unable to provide justifications about their outputs. The lack of explanation is one of the most important shortcomings of Machine Learning at present [3]. The European Union cites XAI (Explainable Artificial Intelligence) in its Ethics Guidelines for Trustworthy AI [4].

This work extends the ADMNC algorithm [5], an anomaly detection algorithm developed by our research group, with a new layer that opens the ADMNC black box by offering pre-hoc explainability. Regression decision trees are used to segment input data into homogeneous groups that can be described with a few variables. The objective is to provide a helpful and intuitive description of anomalous data, thus offering information to make informed decisions.

2. Methodology

The original ADMNC algorithm [5] is a method for large-scale offline learning to obtain a model of normal data that is then used to detect anomalies. The model used to obtain the pre-hoc explanation will consist of a grouping of the input patterns attending to their numerical variables. Clusters will be defined as the leaf nodes of a shallow decision tree [6]. Each pattern will be assigned its ADMNC estimator [5]. This estimator will then be approximated with a simple regression model, learned using the Apache Spark MLLib implementation of CART. Variance gives us an idea about how homogeneous the estimators for elements in a tree node are. Successive divisions turn nodes into more specific groups that contain similar elements. This balance between cluster homogeneity and explanation quality, given by the depth of each path, allows us to choose the level of detail for explanations.

We define the clustering $Cl(D)$ over dataset D as a set of m clusters $Cl_i \ \forall i \in [1, m]$ that contains every element in D. The *weighted variance* (WV) of a $Cl(D)$ is defined as:

$$WV(Cl(D)) = \frac{\sum\limits_{i \in 1..m} ({\sigma^2}_{Cl_i})|Cl_i|}{|D|}. \tag{1}$$

The weighted variance of a clustering measures how homogeneous its components are. This measure is complemented with another measure that indicates the number of input variables employed to characterize each cluster Cl_i. As a result, the *quality*, $\mathcal{Q}$ of a clustering is defined as:

$$\mathcal{Q}(CL(D)) = -WV(Cl(D)) - \lambda \sum_{Cl_i \in Cl(D)} NV(Cl_i), \tag{2}$$

where $NV(Cl_i)$ represents the number of variables needed to describe cluster Cl_i and λ is a hyperparameter that allows the supervisor to balance the accuracy and interpretability [6] of the whole clustering. This quality measure is always negative and the goal of the algorithm is maximizing its value to approach 0. Maximizing this measure will ensure that the groups obtained are as homogeneous as possible and that they are explained using as few of the input variables as possible.

This method is carried out in two steps: (1) a full N level tree is built using the well-known CART algorithm. (2) This full tree is pruned to optimize the quality measure. Those node splits that decrease variance but also decrease quality are discarded, yielding a simpler tree that maximizes quality. The main features that lead data to be anomalous can be obtained as the path to anomalous clusters.

3. Experimental Results

To assess the validity of our approach, we considered two large datasets focusing on the network intrusion detection domain, KDDCup99 [5] and ISCXIDS 2012. For each resulting clustering, we measured its quality $\mathcal{Q}$ and weighted variance. We also included the number of clusters and the number of variables employed for both the full and pruned tree. These results are listed in Table 1. We set hyperparameter λ accordingly with pruning effort. This value can be modified by the supervisor, assigning more or less importance to interpretability in comparison to predictive power. Area under ROC (Receiver Operating Characteristic) curve is provided as fitness measure for anomaly detection, making five repetitions of each experiment. An example of explanatory tree is shown in Figure 1.

Table 1. Area under ROC curve (AUC) and explanatory tree metrics. Before pruning (Full, *F*) and after pruning (Pruned, *P*), considering hyperparameter λ, OV (Overall variance), $\mathcal{Q}$ (quality), WV (weighted variance), #*Cl* (number of clusters) and NV (number of variables to reach all clusters).

Dataset			**AUC**	**Explanation**				
Name	**OV**	λ	$(\mu \pm \sigma)$	**Tree**	$\mathcal{Q}$	**WV**	**#*Cl***	**NV**
ISCXIDS 2012	0.105	10^{-4}	0.919 ± 0.02	*F*	−0.062	0.048	29	142
				P	−0.051	0.049	7	25
KDDCup99 - FULL	0.049	10^{-3}	0.758 ± 0.05	*F*	−0.147	0.011	28	136
				P	−0.032	0.012	6	20
KDDCup99 - SMTP	2.846	10^{-3}	0.980 ± 0.01	*F*	−0.105	3.630×10^{-9}	22	105
				P	−0.005	6.632×10^{-6}	3	5
KDDCup99 - HTTP	0.843	10^{-3}	0.992 ± 0.01	*F*	−0.898	0.831	15	67
				P	−0.842	0.837	3	5
KDDCup99 - 10	2.454	10^{-3}	0.966 ± 0.02	*F*	−1.320	1.227	20	93
				P	−1.247	1.228	6	20

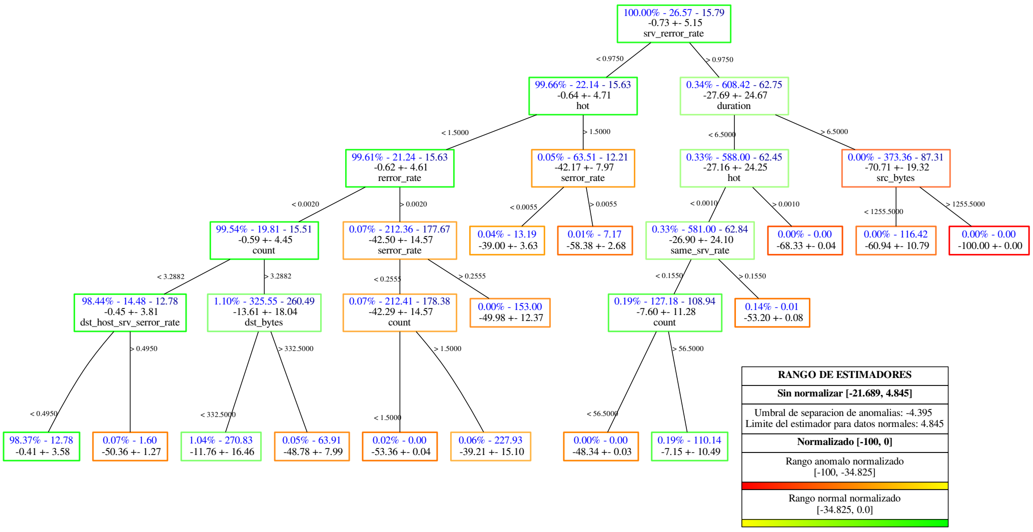

Figure 1. Explanatory tree after pruning ($\lambda = 10^{-3}$) using the KDDCup99-SMTP dataset. Named sequentially, reading from left to right, each node shows: the proportion of elements that it represents regarding the full dataset (shown in blue), overall variance (shown in blue), the weighted variance w.r.t children nodes (shown in dark blue) and mean and standard deviation for the subset of estimators. Further experimental results are given through supplementary materials reference.

4. Discussion and Conclusions

XAI is necessary to provide transparency to model predictions. It is a growing field of study that guarantees compliance with new European Union regulations. The proposed method allows us to examine differences between normal and anomalous data, potentially allowing the identification of generalization power, biases and formulation of hypothesis for abnormal data context.

In the future, we plan to add the categorical variables to the tree-based pre-hoc explanation. This will paint a more accurate picture of the input dataset. Another possible future research line is to improve explanations by introducing a previous dimensionality reduction step, as high dimensional data present redundant and irrelevant variables that produce bias and generalization errors.

Supplementary Materials: Pre-hoc regression trees are available online at https://www.dropbox.com/sh/m6lyn8zpss75sru/AADO_OFwzNwUTHD24vgJXhwma?dl=0

Funding: This research was partially funded by European Union ERDF funds, Ministerio de Ciencia e Innovación grant number PID2019-109238GB-C22, Xunta de Galicia through the accreditation of Centro Singular de Investigación 2016-2020, Ref. ED431G/01 and Grupos de Referencia Competitiva, Ref. GRC2014/035

Acknowledgments: We would like to thank CESGA for the use of their computing resources. Special recognition is given to the Spanish Ministerio de Educación for the predoctoral FPU funds, grant number FPU19/01457.

References

1. Liu, F.T.; Ting, K.; Zhou, Z. Isolation-Based Anomaly Detection. *TKDD* **2012**, *6*, 1–39.
2. Lu, Y.C.; Chen, F.; Wang, Y.; Lu, C.T. Discovering anomalies on mixed-type data using a generalized student-t based approach. *IEEE Trans. Knowl. Data Eng.* **2016**, *10*, 2582–2595.
3. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.
4. High Level Expert Group on Artificial Intelligence. Ethics Guidelines on Trustworthy Artificial Intelligence. 2019. Available online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed on 1 July 2020).

5. Eiras-Franco, C.; Martínez-Rego, D.; Guijarro-Berdiñas, B.; Alonso-Betanzos, A.; Bahamonde, A. Large Scale Anomaly Detection in Mixed Numerical and Categorical Input Spaces. *Inf. Sci.* **2019**, *487*, 115–127.
6. Eiras-Franco, C.; Guijarro-Berdiñas, B.; Alonso-Betanzos, A.; Bahamonde, A. A scalable decision-tree-based method to explain interactions in dyadic data. *Decis. Support Syst.* **2019**, *127*, 113141.