A Mathematical Assessment of the Isolation Random Forest Method for Anomaly Detection in Big Data.

Fernando Morales[1], Jorge Ramírez[1], and Edgar Ramos[1]

[1]Universidad Nacional de Colombia Sede Medellin

January 30, 2021

Abstract

We present the mathematical analysis of the Isolation Random Forest Method (IRF Method) for anomaly detection, introduced in {\sc F.˜T. Liu, K.˜M. Ting, Z.-H. Zhou:}, {\it Isolation-based anomaly detection}, TKDD 6 (2012) 3:1–3:39. We prove that the IRF space can be endowed with a probability induced by the Isolation Tree algorithm (iTree). In this setting, the convergence of the IRF method is proved, using the Law of Large Numbers. A couple of counterexamples are presented to show that the method is inconclusive and no certificate of quality can be given, when using it as a means to detect anomalies. Hence, an alternative version of the method is proposed whose mathematical foundation is fully justified. Furthermore, a criterion for choosing the number of sampled trees needed to guarantee confidence intervals of the numerical results is presented. Finally, numerical experiments are presented to compare the performance of the classic method with the proposed one.

Hosted file

`Isolation_Forest_Analysis.pdf` available at https://authorea.com/users/392691/articles/506466-a-mathematical-assessment-of-the-isolation-random-forest-method-for-anomaly-detection-in-big-data