IIT INDORE

11/09/2019

PROJECT PROPOSAL

Optimizing Machine Learning algorithms in constrained environments of an IoT node

Submitted by-

1) Seemandhar Jain
(170001046)

Submitted To-

Dr. Kapil Ahuja
Associate Professor in C.S.E.

Siddharth Gupta (M.S. (Res.))
Rohit Agrawal (PhD)

1. Introduction

Anomaly/Outlier detection is the process of finding abnormal data points in datasets or data streams. Anomaly detection finds its application in various fields like network intrusion detection, fraud detection, fault detection, etc. There are many anomaly detection algorithms available in the literature but most of these algorithms require setting of some parameters which significantly affect the performance of the algorithm. These parameters are generally set by hit-and-trial, hence performance is compromised with default or random values. In this project, we propose a optimizing algorithm for anomaly detection based on machine learning algorithms, algorithm for anomaly detection, though the approach used in the proposed solution is different. The algorithm is implemented on Python for scalability and thus the solution can handle big data as well and provides fast results. Experiments will going to be conducted on various datasets, and the results show that the proposed solution is much accurate than the standard algorithms of anomaly detection

2. Needs/Problems

A.) Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are called anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains. Noise removal is driven by the need to remove unwanted objects before data analysis is performed (Ye and Chen, 2001).

Given the various challenges, the anomaly detection problem, in its most general form, is not easy to solve (Garcia-Teodoro et al., 2009). In fact, most of the existing anomaly detection techniques solve a specific formulation of the problem. The formulation is induced by various factors such as the nature of the data, availability of labeled data, type of anomalies to be detected, etc. Often these factors are determined by the application domain in which the

anomalies need to be detected. Accordingly, researchers have adopted concepts from diverse disciplines, such as statistics, machine learning, data mining, information theory, and spectral theory and have applied them to specific problem formulations.

3. Objectives

Some approximation techniques used in this application are

1. LOF algorithm
2. Rounding Data and Dynamic Programming
3. Deterministic rounding of linear programs
4. Rounding sampling and randomized rounding of linear programs
5. Randomized rounding of semi-definite programs
6. BAAD: A self-optimizng algorithm for anomaly detection
7. Cuts and metrics
8. Further uses of the techniques

4. Timetable

	Description of Work	Start and End Dates
Phase One	Getting Ready and Collection of Data	11/9/2019 to 1/10/2019
Phase Two	Implementing the diff. algorithms and comparing them. Finding the results.	2/10/2019 to --/--/----
Phase Three	Final evaluation And Submission.	--/--/----

Seemandhar Jain, Project Candidate-1

Date: _____

Dr. Kapil Ahuja, Project Supervisor

Date: _____

References:

1. D. M. Hawkins, Identification of outliers, Vol. 11, Springer, 1980.
2. G. Beni and J. Wang, “Swarm intelligence in cellular robotic systems”, in: Robots and Biological Systems: Towards a New Bionics?, Springer, Heidelberg, 1993, pp. 703-712.
3. V. K.Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah and B. Saha, “Apache hadoop yarn: Yet another resource negotiator”, in: Proceedings of the 4th annual Symposium on Cloud Computing (p. 5). ACM, 2013.
4. M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, and A. Ghodsi, “Apache Spark: A unified engine for big data processing”, in: Communications of the ACM, 2016, pp.56-65.
5. R. Bryson, S. Kenwright, D. Cox, M. Ellsworth and D. Haimes, “Visually exploring gigabyte data sets in real time”, Commun. ACM, 1999.
6. J.W. Tukey, Exploratory data analysis, 1977.
7. A. Koufakou, J. Secretan, J. Reeder, K. Cardona and M. Georgiopoulos, “Fast parallel outlier detection for categorical datasets using MapReduce”, in: Proceedings of IEEE International Joint Conference on Neural Networks, 2008, pp. 3298–3304.