Malaya Journal of Matematik, Vol. 8, No. 4, 2144-2148, 2020

https://doi.org/10.26637/MJM0804/0139

Anomaly detection using machine learning techniques

Chellammal Suriyanarayanan[1]* and Saranya Kunasekaran[2]

Abstract

Anomaly represents deviation from the normal behavior of an event. Detection of anomaly provides means to take appropriate countermeasures in various domains. Examples include detection of fraudulent transaction in banking or financial domain, detection of cyber-attacks in networking environment, detection of abnormal behavior of vital signs of patient in healthcare domain. Also, detection of anomalies with respect to time of arrival of data is a crucial in deciding the accomplishment of successful countermeasures. Selection of suitable algorithm or method for detection of anomaly is also equally important for successful detection of anomalies. In this paper it is proposed to compare the performance of two different algorithms, namely, Isolation Forest (unsupervised) and Random Forest (supervised) by varying the operating parameters of the algorithms. Experiment is carried out using benchmark dataset that belongs to healthcare domain. The data is preprocessed for missing values and then detection accuracy of algorithm is analyzed with respect to number of records. Results are discussed.

Keywords

Anomaly detection, Random Forest, Isolation Forest, batch processing, healthcare.

[1,2] *Department of Computer Science, Bharathidasan University Constituent Arts and Science College, Affiliated to Bharathidasan University, Navalurkuttapattu, Tiruchirappalli, Tamil Nadu, India.*

***Corresponding author**: [1]drschellammal@gmail.com; [2]saranyasekar19@gmail.com

Article History: Received **24** September **2020**; Accepted **18** November **2020**

Contents

1. Introduction

Due to the advancement of hardware and software developments, most of the organizations naturally tend to generate boundless data. Sensors and devices are integrated with Internet of Things (IoT) in various applications such as weather forecasting, traffic control, smart phones, agriculture, house control devices., generate data which evolves in size. The characteristics of data generated from various sensors and applications are likely to be heterogeneous in formats. In addition, it may occur with speed also. Another important aspect of data is it may contain anomalies. An anomaly is the deviation from a normal behavior of an event. Anomalies in data are required to be handled very carefully as they are indicators for taking important decision. For example, detecting anomaly in the heart rate of a person who got admitted in say Intensive Care Unit is very important in deciding the appropriate treatment at right time. Consider another example. The presence of an abnormal data in a meteorological data gathered by meteorological department can be used to predict the forthcoming cyclone. So, detection of anomalies in various domains is a potential area of research.

Anomaly detection plays a crucial role in healthcare domain where the data is more likely to change with respect to time. For example, for an individual who got admitted into an Intensive Care Unit (ICU), the heart rate is being monitored once in a second. So, detecting anomalies in real time becomes very essential to take appropriate countermeasures. But, before putting an anomaly detection technique for real time into practice, it is a prerequisite that one should identify suitable technique for anomaly detection. In view of the above need, in this work, a comparative analysis is performed with two different algorithms, namely, Isolation forest, an unsuper-

vised algorithm and Random Forest, a supervised algorithm. The analysis is done using benchmark dataset collected from (https://archive.ics.uci.edu/ml/datasets/PAMAP2 + Physical + Activity + Monitoring). The dataset is preprocessed for missing values and the accuracy of anomaly detection of the algorithms are studies by varying the operational parameters of the algorithms. Recommendations are arrived based experimentation.

The rest of the paper is organized as follows. In section II, review on related literature is given. In section III, the approach for proposed comparative analysis is described. In Section IV, results are discussed. Section V concludes the work with future extension.

2. Literature Review

Machine learning algorithms are commonly used for detecting anomalies in data sets. In [3], how Support Vector Machine (SVM) and Random Forest (RF) are used to detect intrusions in industrial networks by analyzing the abnormal behavior in network traffic. In addition, from the above work, the authors showed that Random Forest algorithm outperforms the SVM in terms of accuracy and computation time. In another work [4], the authors have detected anomalies over streaming data by integrating Isolation Forest and Mondrian Forest, called iMondrian forest. The performance of the integrated algorithm in detecting anomalies is found to be better than Isolation Forest. In [5], fraud detection using various techniques such as Random Forest, Neural Autoencoder and Isolation Forest have been implemented. The authors found that the Random Forest algorithm gives better performance when compared with that of the other two algorithms. In another research work [6], malignant activities in the IoT based network are detected with better accuracy and less false rate using Random Forest. In [7] it is discussed that machine learning techniques such as Random Forest are efficient in detecting anomalies in network. In addition, the works highlights about the need for optimizing the operating parameters of algorithms the number of features. In this work also, Random Forest is found to give better accuracy.

3. Proposed Work

In this work, it is proposed to perform a comparative analysis between Random Forest and Isolation Forest in detecting anomalies in a dataset that belongs to healthcare domain. In contrast to several research works viz., [4-9] where the applicability of Random Forest for anomaly detection in network behaviour is discussed, in this work the algorithm is applied to healthcare data, more specifically, heart rate data. As mentioned earlier, before applying the algorithms for real time detection, suitable technique for anomaly detection is to be found out. From literature, it is clear that machine learning algorithms are capable of detecting anomalies with sufficient accuracy. An approach as given in Fig. 1 is proposed to perform a comparative analysis between two predominantly used machine learning algorithms(Random Forest and Isolation Forest) for their applicability in detection of anomalies in heart rate data.

3.1 Data set and Pre processing

Dataset consisting of 319352 records has been collected from UCI Repository. It represents the data acquired from individuals using 3 Colibri wireless IMUs (Inertial Measurement Units) sensors which are plated at different locations on human body namely individual's wrist, chest and ankle. These sensors are used to monitor the heartrate (bpm) of the individual based on different activities viz., lying, sitting, walking, running, cycling, etc. Each activity is identified with unique number. It acquires the heart rate along with time stamp. Each record consists of timestamp, activity ID, heartrate and class labels which defines the data is anomalous or not. Description about the dataset is given in Table 1. From the dataset which consists of 319352 records, it is found that, only 29189 records are filled with heartrate values and remaining records contain null values. The null values are replaced by the average of heart beat.

3.2 Anomaly detection techniques

Brief overview about the two algorithms is given.

Random Forest

Random Forest is an ensemble classifier used for both classification and regression task. Basically, Random Forest is a collection of decision trees and it contains one root node and several internal split nodes [11]. It applies majority voting to combine the outcomes of all trees. Usually, decision trees are trained with bagging method which is a combination of learning models which increases the overall result. In this approach, the features are randomly selected in each decision split. Random Forest is appropriate for high dimensional data model as it handles missing values, continuous, categorical and binary data [12]. This algorithm can be used in used in many applications such as banking for fraudulent detection, healthcare for analysis of patient's medical history and businesses for customer satisfaction as it supports larger dataset, correctly predicted and robust to outliers [13].

Isolation Forest

Isolation Forest is an unsupervised learning algorithm which isolates anomalies rather than separating normal points. The main idea behind this approach is that anomalies are separated in the tree by a Random Forest so the depth of the node can determine by the anomaly score of a point [4]. In order to isolate the data point, the method recursively generates partitions on the sample by randomly selecting an attribute and then randomly select a split value for the attribute between the minimum and maximum values allowed for that attribute. The recursion takes place on smaller and smaller subsets of the data until single data points are isolated or certain depth limit is reached. Anomalous points are isolated quickly as it has far shorter depths whereas normal data points reach far deeper into the tree [14]. In this approach, anomalies can be detected using path lengths or anomaly scores. This algorithm

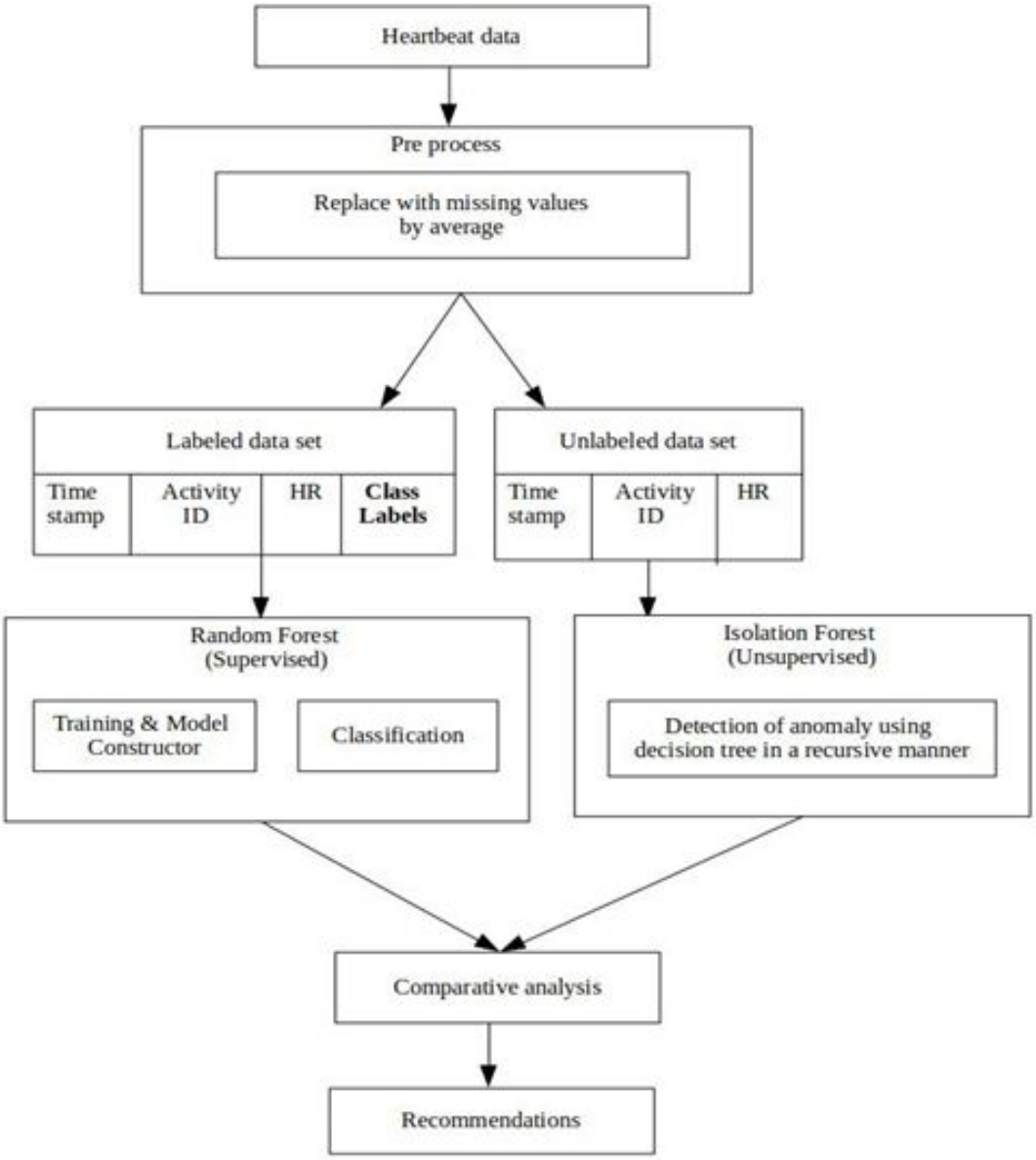

Figure 1. Schematic View of Proposed approach

can be applied in various use cases such as banking, healthcare, networking, etc. Isolation Forest can be work with high dimensional data and scalable in nature.

3.3 Experimentation

Experiment is carried out in a laptop having CPU, intel CORE i3 processor with 1.70 GHz, RAM with 8GB, Ubuntu 16.10 operating system. Code has been developed using Python 3.7.6 with Pandas Machine Learning library. After preprocessing, The labeled dataset is applied over Random Forest, whereas, the unlabeled dataset is applied for Isolation Forest as it is unsupervised. The accuracy is found by varying the batch size of records. Batch sizes are constructed from 100 to 319352 with different sizes namely, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 40000, 80000, 120000, 160000, 200000, 240000, 280000, 319352. The algorithm has been applied over each batch to find the accuracy.

The experimental results are evaluated using accuracy, precision, recall and F score. Accuracy, Precision, Recall and Fscore are calculated using the following formulae

$$\text{A ccuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F}- \text{ score} = \frac{2 \times (\text{ Recall } \times \text{ Precision })}{(\text{ Recall } + \text{ Precision })}$$

Table 1. Description about dataset and No. of Records

Parameters	Timestamp (s)	Activity ID & Description	Heartrate (bpm)	Class labels & No. of Records
1. Timestamp 2. Activity ID 3. Heartrate 4. Class labels	Starts at 10.03 Ends at 3203.54	9 - Watching TV 11 - Car driving 18 - Folding laundry 19 - House cleaning 0 - Others	Minimum rate - 76 Maximum rate - 141	0 - Normal (9407) 1 - Anomaly (309945)

The accuracy, precision, recall and F-score obtained by Random Forest is given in Table 2.

Table 2. Accuracy, Precision, Recall and F-score of anomaly detection using Random Forest by varying batch size

Batch size	Accuracy	Precision	Recall	F score
100	1.0	1.0	1.0	1.0
200	1.0	1.0	1.0	1.0
300	1.0	1.0	1.0	1.0
400	1.0	1.0	1.0	1.0
500	1.0	1.0	1.0	1.0
600	1.0	1.0	1.0	1.0
700	1.0	1.0	1.0	1.0
800	1.0	1.0	1.0	1.0
900	1.0	1.0	1.0	1.0
1000	1.0	1.0	1.0	1.0
40000	1.0	1.0	1.0	1.0
80000	1.0	1.0	1.0	1.0
120000	1.0	1.0	1.0	1.0
160000	1.0	1.0	1.0	1.0
200000	1.0	1.0	1.0	1.0
240000	1.0	1.0	1.0	1.0
280000	1.0	1.0	1.0	1.0
319352	1.0	1.0	1.0	1.0

Similarly, the accuracy of Isolation Forest is computed using the following formula

$$\text{A ccuracy} = \frac{100 \times \text{ M odel detected anomalies}}{\text{A ctual anomalies present}}.$$

The accuracy obtained by Isolation Forest is represented in Table 3 as follows.

Table 3. Accuracy of anomaly detection using Isolation Forest by varying batch size

Batch size	IF Accuracy (%)
100	92.0
200	81.0
300	99.7
400	91.5
500	99.2
600	93.8
700	95.1
800	86.6
900	93.7
1000	76.9
40000	99.5
80000	84.9
120000	71.8
160000	71.1
200000	91.2
240000	87.1
280000	96.5
319352	88.6

4. Discussion

The accuracy obtained using Random Forest and Isolation Forest for different number of records is shown in Fig. 2. From Table 2, Table 3 and Fig. 2, it is understood that, Random Forest gives higher accuracy than Isolation Forest for various batch size. Further, the precision, recall and F-score of Random Forest also found to be good. Thus, Random Forest is found to outperform Isolation Forest with the dataset considered.

5. Conclusion and Future Work

Anomaly detection leads to potential applications such as credit card fraud detection and intruder detection in many domains. In health care domain, anomaly detection plays a vital role in taking decisions for timely diagnosis and treatment. In this work, the performance of Random Forest and Isolation Forest have been analyzed for their accuracy in detecting anomalies by varying the data size. With the experimentation carried out, it is suggested that Random Forest give better accuracy than the Isolation Forest. The reason is that, Random Forest is guided with labels. In the forthcoming work, it is

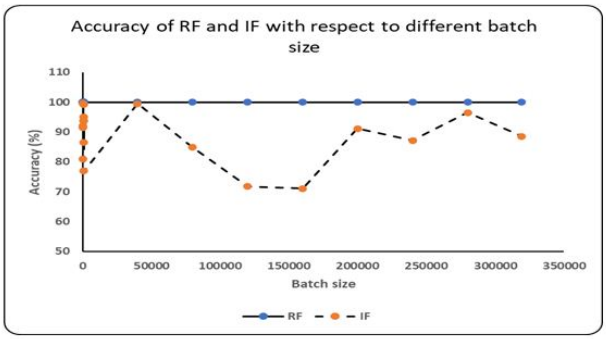

Figure 2. Comparison of RF and IF based on accuracy

planned to employ Random Forest for detecting anomalies in heart rate at real time using big data platform, Apache Kafka.

References

[1] Nisha P. Shetty, Jayashree Shetty, Rohil Narula and Kushagra Tandona, Comparison study of machine learning classifiers to detect anomalies, *International Journal of Electrical and Computer Engineering (IJECE),* 10(5)(2020), 5445-5452.

[2] Luise Pufahl and Mathias Weske, Requirements Framework for Batch Processing in Business Processes, *International Conference on Exploring Modeling Methods for Systems Analysis and Design*, 2017.

[3] Simon D. Duque Anton, Sapna Sinha and Hans Dieter Schotten, *Anomaly-based Intrusion Detection in Industrial Data with SVM and Random Forests,* 27th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 1(2019).

[4] Haoran Ma, Benyamin Ghojogh, Maria N. Samad, Dongyu Zheng, Mark Crowley, *Isolation Mondrian Forest for Batch and Online Anomaly Detection*, IEEE International Conference on Systems, Man, and Cybernetics (SMC), (2020).

[5] Kathrin Melcher, Fraud Detection Using Random Forest, Neural Autoencoder, and Isolation Forest Techniques, AI, *ML & Data Engineering*, 2019.

[6] Rifkie Primartha and Bayu Adhi Tama, *Anomaly Detection using Random Forest: A Performance Revisited*, 2017 International Conference on Data and Software Engineering (IcoDSE), (2018).

[7] Rashmi H Roplekar and Prof. N. V. Buradkar, Survey of Random Forest Based Network Anomaly Detection Systems, *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 6(12)(2017), 95-98.

[8] Lekha R. Nair and Sujala D. Shetty, Research in Big Data and Analytics: An Overview, *International Journal of Computer Applications*, 108(14)(2014), 19-23.

[9] Mansi Shah and Vatika Tayal, Future of Big Data beyond Batch Processing, *International Journal for Scientific Research & Development (IJSRD)*, 3(01)(2015), 217-220.

[10] P. Sai Pranavi, H. D. Sheethal, Sharanya S Kumar, Sonika Kariappa and B. H. Swathi, Analysis of Vehicle Insurance Data to Detect Fraud using Machine Learning, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 8(7)(2020), 2033-2038.

[11] Vrushali Y Kulkarni and Pradeep K Sinha, Effective Learning and Classification using Random Forest Algorithm, *International Journal of Engineering and Innovative Technology (IJEIT)*, 3(11)(2014), 267-273.

[12] Leo Breiman, Random Forests, *Machine Learning*, 45(2001), 5–32.

[13] Sahand Hariri, Matias Carrasco Kind and Robert J. Brunner, *Extended Isolation Forest*, 3(2020), 1–10.

★★★★★★★★★
ISSN(P):2319 – 3786
Malaya Journal of Matematik
ISSN(O):2321 – 5666
★★★★★★★★★