A General Anomaly Detection Framework for Internet of Things

Zibin Zheng
Dept. of Computer Science & Engineering
The Chinese University of Hong Kong
Hong Kong, China
zbzheng@cse.cuhk.edu.hk

Jian Wang, Ziyu Zhu
Next Generation Web
IBM China Research Lab
Beijing, China
{wangwj, zhuziyu}@cn.ibm.com

***Abstract*—This paper proposes a general and extensible anomaly detection framework for Internet of Things (IoT) systems. Case study using real-world traffic data is presented to demonstrate advantages of the proposed framework.**

I. Introduction

Internet of Things (IoT) is a trend taking shape in the information technology industry that could significantly change our lives in the coming decades via connecting physical things. These physical objects, like human senses for a human brain, will take an active part in the Internet and provide vivid information about themselves and their surroundings to dedicated computing facilities.

In IoT systems, data quality management is a critical technology to provide high-quality and trusted data to business-level analysis, optimization and decision making. Anomaly detection techniques are widely employed to remove noises and inaccurate data in order to improve data quality. Anomaly detection refers to the problem of finding patterns in data that do not conform to the expected behavior [1]. In IoT systems, anomalies might be induced in the data for a variety of reasons, such as resource constraints of sensors, environmental affects, communication problems, malicious attacks, etc [2]. Anomaly detection for IoT data is a challenging task, because: (1) defining a normal region that encompasses every possible normal behavior is very difficult; (2) the exact notion of an anomaly is different for different IoT systems in different application domains; (3) IoT systems usually include a lot of distributed heterogeneous sensors, mobile entities, and require dynamic system reconfiguration/upgrade; and (4) the data generated by sensors are usually continuous and periodic, temporal-spatial correlative, and with lots of noise.

Since there are a lot of IoT application domains (e.g., smart traffic, smart power grid, smart home, etc.) with their own characteristics, traditional approaches usually design different anomaly detection algorithms for these IoT systems case by case, which is time consuming. In this paper, we present a general and reusable anomaly detection framework for IoT systems, so that customized anomaly detection solutions can be generated efficiently by doing some configurations. Our framework decouples temporal, spatial and other domain specify data correlations via defining *features*, so that traditional point anomaly detection algorithms can be easily employed for different IoT systems. The main contributions of this paper includes: (1) propose a general and extensible framework for IoT systems; and (2) case study using real-world smart traffic IoT data. In the rest of this paper, Section 2 will introduce the system architecture; Section 3 will present implementation and case study; and Section 4 will conclude the paper.

II. System Architecture

Figure 1 shows our anomaly detection framework for IoT systems, which includes the following modules:

Feature extraction: Our framework decouples temporal, spatial and other data correlations via defining new features. By adding features to the original input data, enriched data is obtained. Various reusable domain specify features are defined and feature selection guidelines are provided for different IoT systems.

Anomaly detection: By adding features to the original data, the problem is transferred to be point anomaly detection, where data correlations are represented by features and no need to be considered by detection algorithms. A customizable algorithm lib is established, which includes different types of anomaly detection algorithms (e.g., nearest neighbor-based anomaly detection, classification, clustering, statistical anomaly detection, spectral anomaly detection, etc.). Best practice and guidelines are provided for algorithm selection and configuration.

Relearning: After providing the cleaned data to the business system, we collect feedbacks to further enhance the detection accuracy of detection algorithms and enrich our lib of best practice and domain-specified knowledge.

III. Implementation and Case Study

Our framework is implemented by Java. We use the smart traffic IoT domain as an example for case study. Real-world IoT data from Kunming city of China is used. The data include 3 weeks traffic monitoring data from more

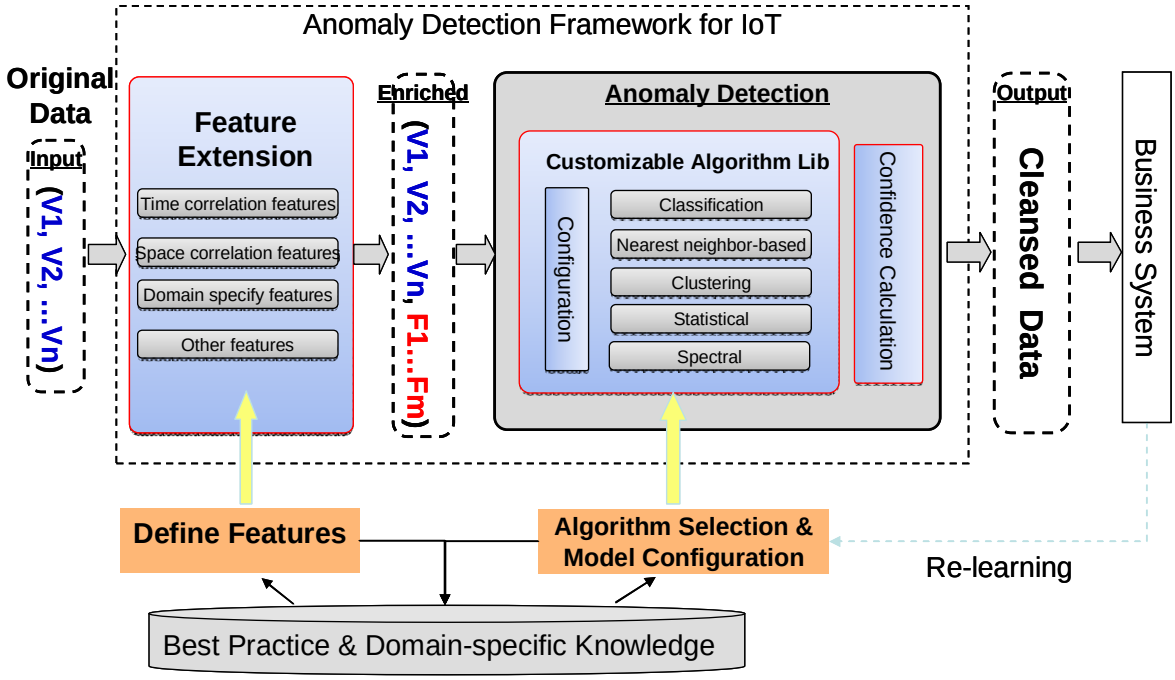

Figure 1. Anomaly Detection Framework for IoT

Table I
CASE STUDY

Original Data					Features		
Time	SID	Volume	Occupation	Speed	F1	F2	F3
19:00	6	330	2.59	57.55	Null	Null	Null
20:00	6	466	4.03	53.21	136	1.44	-4.34
21:00	6	467	4.14	53.35	1	0.11	0.14
22:00	6	237	1.87	55.68	-230	-2.27	2.33
23:00	6	165	1.25	53.94	-72	-0.62	-1.74
00:00	6	466	4.17	53.68	301	2.92	-0.26
00:01	6	155	1.87	55.68	-311	-2.3	2

than 100 inductive loop sensors in roads, which includes 5 attributes, i.e., time, sensor ID (SID), volume, occupation, and speed. *Volume* is the number of detected vehicles in one hour. *Occupation* is the time percentage that the lane is occupied by vehicles in one hour. *Speed* is the average speed of vehicles in unit of kilometers per hour. Our target is to detect abnormal values provided by sensors. We use one-class Support Vector Machine (SVM) as the anomaly detection algorithm, which is an unsupervised algorithm that estimates anomalies in a dataset.

Table I shows the original input data from sensor 6 (SID = 6). In row 6 of the table, the value of volume is 466. This large volume value is abnormal at the time of midnight. When using one-class SVM to process the data, we can not find this outlier. Because (1) 466 is normal within all input data (e.g., similar to values of row 2 and 3), and (2) one-class SVM does not consider temporal correlation of the input data. To address this problem, we add three features (F1 to F3) for this smart traffic IoT data. These features represent the change between the current value and the value of previous time slot. As shown in Table I, F1 of row 6 is 301 (466-165=301), which is large compared with other F1 values. By adding features and employed the enriched data as input to the one-class SVM, row 6 can be detected as anomaly. Due to page limitation, we use this simplest feature for demonstration purposes. In our framework, there are more features handling various types of data correlations (e.g., historical values at the same time periods, spatial correlations between different sensors, correlations between different attributes, etc.). Instead of designing new anomaly detection algorithms for each IoT systems, using the framework, the domain experts can define domain-specified features and reuse existing common features to construct customized anomaly detection algorithms efficiently.

IV. CONCLUSION AND FUTURE WORK

This paper presents a general anomaly detection framework for IoT systems. In the future, we will conduct more investigations on differentiating between events and fault data, which are both shown as abnormal data.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[2] Y. Zhang, N. Meratnia, and P. Havinga, “Outlier detection techniques for wireless sensor networks: A survey,” *Journal of IEEE Communications Survey & Tutorials*, vol. 12, no. 2, 2010.