Pearson Correlation Based Outlier Detection in Spatial-Temporal Data of IoT Networks

M. Veera Brahmam, S. Gopikrishnan, K. Raja Sravan Kumar, and M. Seshu Bhavani

Abstract Outliers are values that change abruptly, either as a result of harsh climate (event) (or) as a result of sensor failure (error) [1]. Since sensor nodes are exposed to outside environment, they are more likely to generate defective data as a result of harse climate or sensor failure. Because more sensors are battery-powered, they may produce erroneous data if the batteries run out (error). Outlier detection techniques that have been used in the past have failed to distinguish between errors and events. This paper used spatial and temporal correlations to detect outliers in IoT sensors. Analysis of spatio-temporal data using Entropy-based and pearson correlation approaches is proposed in this paper.

Keywords Anomaly · Outlier · Spatial-temporal correlations · Pearson correlation · IoT sensors

1 Introduction

Physical objects or "things" that allowing electronics, software, sensors, and network access, to capture and share data is called "internet of things." This enables objects to be sensed and monitored from a far, allowing for direct integration of the physical and virtual worlds [9]. In the Internet of Things, a thing is a physical entity with sensors that interacts with the real world to perform specific tasks through a network [3]. Sensor data is typically time series data or a data stream in the Internet of Things. The data that listed in time order is Time series data, where as continuous generation of huge amounts of data is data stream [13].

The detection of anomalies from these large quantities of sensor data is one of the most difficult challenges in IoT data management because Sensors used in IoT are resource constrained. i.e., they have limited power, limited processing capability, limited memory [4]. Since the sensors are battery operated, they will be invalid when the batteries run out and sensors are susceptible to produce faulty data due to harse environment or sensor malfunctioning. So it is useful to identify errors and avoid

M. Veera Brahmam · S. Gopikrishnan (✉) · K. Raja Sravan Kumar · M. Seshu Bhavani
School of Computer Science and Engineering, VIT-AP University, Amaravati, India

J. S. Raj et al. (eds.), *Innovative Data Communication Technologies and Application*,
Lecture Notes on Data Engineering and Communications Technologies 96,
https://doi.org/10.1007/978-981-16-7167-8_75

deliver the erroneous data to cloud for this reason some amount of energy is saved energy saving is important to improve network life in IoT [7] spatial and temporal correlations are used to detect Outliers in this paper and both types of outliers are detected. IoT sensor network is frequently used to detect vital hidden information regarding events. Furthermore, events to be considered as outliers because they are not similar to normal sample readings [1]. The motivation for proposing this paper is how to separate errors from events. But in this paper, we have not given any details about how to separate errors from events. Here we given the details about how to detect outliers based on spatial and temporal correlations.

1.1 Categories of Outliers

Outliers are classified into three types as follows:
Global/point outlier: A data point is a global anomaly in a data set if it varies significantly from the other data points [5]. As a result, a proper assessment of divergence is required to distinguish between normal and anomalous points. The majority of anomaly detection techniques focus on detecting this type of anomaly because it is considered the easiest to detect. Intrusion detection and trading transaction auditing systems [5, 16] are only two examples of global anomaly detection applications.
Contextual Outliers: A data point is called a contextual outlier in a particular data set if it significantly differs in the specified context [11]. Because contextual anomalies are dependent on a particular context, they are also known as conditional anomalies. As a result, the context must be defined as part of the problem definition in order to identify contextual anomalies. The characteristics of the data objects in consideration are categorized into two categories of contextual anomaly detection.

- Contextual attributes: The context of the object is defined by these features. Context may refer to both time and place.
- Behavioral attributes: These attributes describe the object's characteristics.
- Collective anomaly: If a subset of data objects in a data set differs significantly from the entire data set, the objects as a whole generate a collective anomaly. It's possible that none of the data objects are outliers [14]. There are many uses for collective anomaly detection [6]. For example, a stock transaction between two parties is considered usual, but a large number of transactions of the same stock between small parties in a short period of time is considered a collective anomaly, as it could indicate market tampering.

In contrast to global or contextual anomaly detection, we must consider behavior of individual objects and the behavior of groups of objects in collective anomaly detection. As a result, prior understanding of the relationships between data objects, such as To find collective anomalies, distance or similarity measurements are required.

2 Related Study

Analysis of spatio-temporal data using Entropy-based and pearson correlation approaches is proposed in this paper. Correlation is a similarity measure which is used to how well one value related to another value. The study of smart city data with only a few studies examining the relationships, is a new field of study, between various measurement data in the context of cities [8, 12]. Pearson correlations were used in all of these studies to analyze the data.

Pearson correlation, on the other hand, is well-known to have some flaws.when the data has non-linear distributions, Pearson correlation does not find out the dependence between two or more variables. Although some of the statistical options, Mutual information, Distance correlation and correlation ratio are not fully correct, they may provide a better indication of data dependency or serve as a supplement to the Pearson correlation.

2.1 Measures of Dependence Between Variables

There are many statistical methods for determining the interdependence of two or more variables. Pearson correlation coefficient is the first and most common measure [15]. If data has a Gaussian distribution, this measure is most subtle to direct correlations between two or more variables. The data in the field of smart cities is frequently intermittent, and the data distribution is not always Gaussian, particularly in short-term windows. As a result, Gaussian methods aren't always appropriate for Examining and handling multivariate data [2]. Other correlation coefficients, which are extra strong than the Pearson correlation, have been developed to overcome this constraint. This research paper chose mutual information in particular since it can measure heterogeneous relationships. Furthermore, this mutual information measure makes no assumptions about the data's statistical distribution. As a result, it's a non-parametric approach.

- correlation of readings: correlation explains how well one or more readings related to each other. We can use Pearson's correlation coefficient or mutual information to compute the correlation(similarity) between two readings.
- Similarity of sensor readings: If the correlation value is greater than or equal to the threshold (T), then two sensor readings are said to similar. The user establishes a threshold value (T) based on sensor readings.
- Temporal correlation: If the current reading of one sensor node has similarity with its last reading, then we say those two readings are temporally correlated.
- Spatial correlation: If the current readings of one sensor node and the current readings of the other sensor node taken at same time are identical, then the two sensor nodes have spatial correlation.

3 Pearson Correlation

Correlation is a metric for determining how closely two sets of data are related. The Pearson Correlation is the most widely used statistician's correlation measure. Its full name is Pearson Product Moment Correlation (PPMC). It is a linear representation of the relationship between two sets of data. The range of Pearson correlation coefficient absolute values are from -1 to $+1$. Here, lower values indicating lower variable dependence and higher absolute values indicating higher variable dependence. For two random variables x and y, the Pearson coefficient is expressed as Eq. (1)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1}$$

where,

- σ_x—variable x standard deviation
- σ_y—variable y standard deviation
- σ_{xy}—covariance of variable x and variable y.

If, the correlation coefficient of 1 means that with every positive increase in one variable, a certain proportion of the other increases as well. Shoe sizes, for example, rise in (nearly) exact proportion to the length of the foot. If the correlation coefficient of -1 means that for every positive increase in one variable, a certain proportion of the other decreases negatively. If the correlation coefficient of 0 indicates that they are not related. The strength of the relationship is determined by the correlation coefficient's absolute value. The greater the number, the higher the relationship. For instance, $|-0.85| = 0.85$ has a stronger relationship than 0.75.

Figure 1 depicts a situation where Pearson's correlation fails to find out the dependence. Since Pearson correlation find out linear correlations between jointly normally distributed variables, these types of cases arise. If two variables do not adopt a linear correlation, the Pearson Coefficient may fail to identify a dependency.

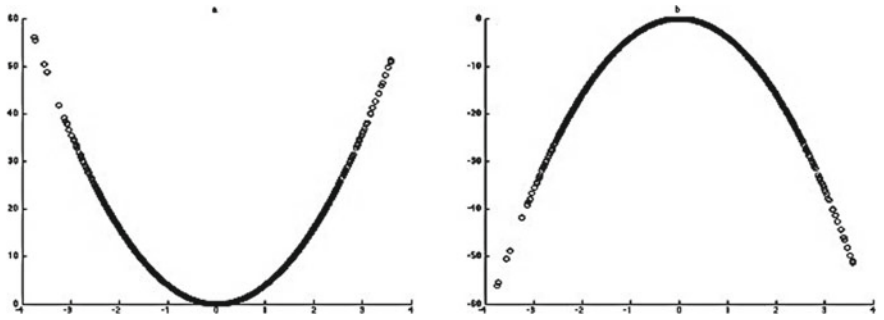

Fig. 1 Examples of Pearson's failure to detect non-linear correlations

As shown in Fig. 1, even though variables are correlated, Pearson coefficient fails to detect the dependency. For this reason, this research suggesting an entropy-based method mutual information.

4 Mutual Information

Mutual information is one of many terms that describe how much information one random variable can provide about another. It is when one random variable's uncertainty is reduced as a result of learning about another. When one random variable's uncertainty decreases after seeing another, it's said to have high mutual information. Low mutual information denotes a rapid decrease in the uncertainty about one random variable after witnessing another. We must first define entropy, which is represented by $H(A)$, and then conditional entropy, which is represented by $H(A|B)$, to define mutual information. A measure of uncertainty in its information content is called the entropy of a random variable.

Consider two random variables: A, which corresponds to the number on a roulette wheel, and B, which corresponds to the number on a fair 6-sided die. The entropy of A is greater than that of the entropy of B. In addition to the numbers 1 through 6, the roulette wheel will have values ranging from 7 to 36. In certain cases, it is less predictable. If a random variable A is defined by a probability distribution $P(A)$ and has values in a set $A = \{a_1, a_2, \ldots, a_n\}$, then the entropy of the random variable is written as Eq. (2).

$$H(A) = -\sum_{a \epsilon A} P(a) \log P(a) \tag{2}$$

Here, $P(a)$ is probability distribution of a. We may also write this as $H(P(a)) \equiv H(P) \equiv H(A)$. Joint entropy is the entropy of a multi-valued random variable or a joint probability distribution. The joint entropy of a distribution of people described by hair color A and eye color B, where C can take four different values from a set A and B can take three different values from a set B, for example, may be of interest. We can write their joint entropy as follows if $P(A; B)$ denotes the joint probability distribution of hair and eye color as Eqs. (3) and (4).

$$H(A, B) = -\sum_{a \epsilon A} \sum_{b \epsilon B} P(a, b) \log P(a, b) \tag{3}$$

$$H(A; B) \equiv H(P(A; B)) \equiv -\sum_{a \epsilon A} \sum_{b \epsilon B} P(a, b) \log P(a, b) \tag{4}$$

The average uncertainty about variable A after observing a second random variable B is called Conditional entropy $H(A \mid B)$, and it can be expressed as

$$H(A \mid B) = \sum_{b \epsilon B} P(b) \left[- \sum_{a \epsilon A} P(a \mid b) \log P(a \mid b) \right] \tag{5}$$

Here, $P(a \mid b) = P(a, b)/P(b)$ is the conditional probability of a given b. The reduction in uncertainty about variable A after observing variable B is known as Mutual information and is written as Eq. (6).

$$I(A, B) = H(A) - H(A|B) = H(A) + H(B) - H(A, B) \tag{6}$$

From Fig. 2, we can easily conclude that Mutual information is the interconnection of uncertainity of the two variables, and it is expressed in Eq. (4). The processing difficulty of Pearson correlation is lower than that of Mutual information. This is one of the drawbacks of mutual information. However, a portion of the analysis of correlation can be performed on the data collection nodes since most current sensor nodes can process data [12]. Mutual information values are always positive and range from zero to the smallest entropy value for a variable: $I(A, B) \in [0, \min(H(A), H(B)]$. $I(A, B) = 0$ signifies no correlation between A and B, values near zero denote low correlation, while values near $\min(H(A), H(B)$ denote significant correlation.

Mutual information is more suitable for larger datasets. It will give more accurate results for larger data sets compared to pearson correlation [10] and the variance of mutual information is higher for datasets with a small number of samples. When compared to Pearson Correlation, this method is more accurate. As a consequence, if we want more precise results for broader datasets, we should use mutual information rather than pearson correlation. The need for a sample size of minimum to find correlations is one of the issues when working with correlations. If the observations are inadequate in number, the correlation algorithms can produce false negatives or false positives. Sample size is decided by the type of data and infrastructure used

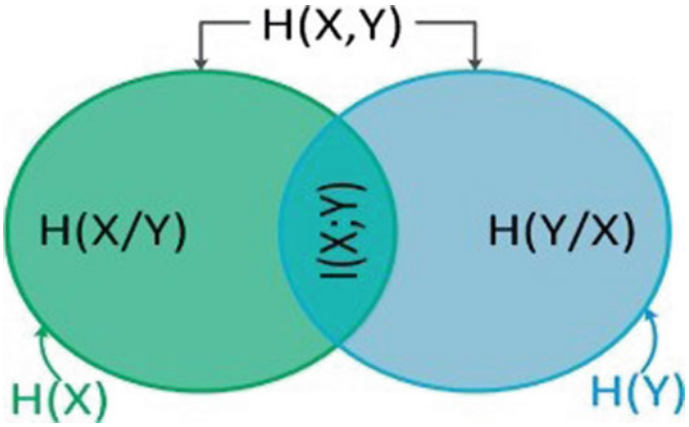

Fig. 2 The relationship between two correlated signals' distinct entropy notions

to obtain it. As a result, every type of data must be prepared in order to choose the appropriate amount of observations. Some background information, such as how the data delivery can be affected by a grasp of daily events can affect data delivery is also useful.

5 Results Analysis

5.1 Evaluation

This research uses Pearson Correlation and mutual information methods to detect outliers with the help of spatial and temporal correlations. First user sets up a threshold value based on the sensor generated readings. The correlation value of current reading and previous reading of sensor node, which will be computed by using Pearson Correlation or mutual information of current reading and previous reading, is greater than or equal to threshold value, then those readings are valid readings. Otherwise they are invalid. i.e., there is no similarity between those readings. When the current reading of one sensor node differs from its previous reading, it is called an outlier.

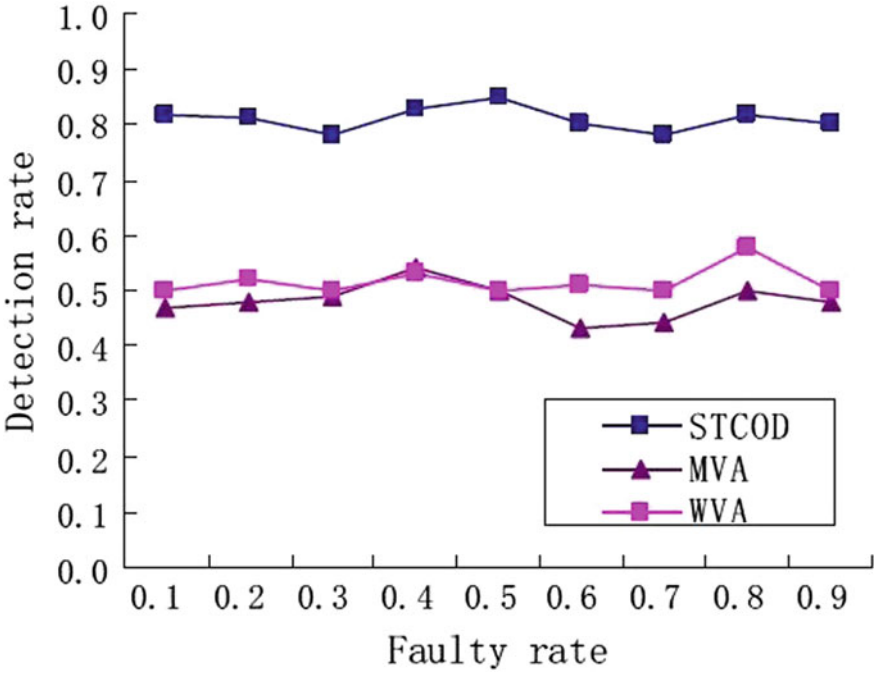

Fig. 3 Detection rate

If readings are invalid, then the sensor node sends these Readings to its neighbors to confirm that the produced readings are invalid or not. Then, neighbors compare their own readings with received readings and sends message to sensor node regarding the comparison. Based on all its neighbors replies, sensor node justify that the generated readings are valid or not.

5.2 Results

To assess the efficiency of experiment results, this study defines detection rate and false detection rate. Here, this research adopt one data set and then it adds some outliers and faulty data into the data set. Data set is labeled X, and the defective data set in X is labeled Y. After running the algorithm, the research will filter out a series of faulty readings denoted as Y'. It can define detection rate as $\frac{|Y \cap Y'|}{|Y|}$ and false Detection rate as $\frac{|Y \cup Y'| - |Y \cap Y'|}{|X|}$.

Detection rate of Spatial-temporal correlation based outlier detection (STCOD) algorithm is compared to detection rate of Majority voting (MVA) and detection rate of Weight based voting (WVA) algorithms in the first experiment. Figure 3 depicts the outcome, with the X-axis representing the faulty data rate and the Y-axis representing

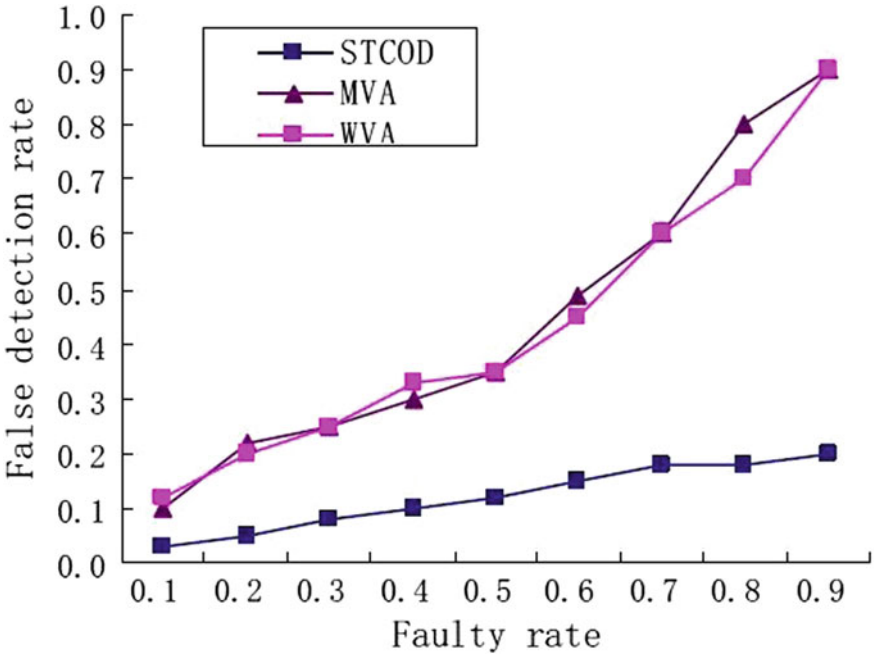

Fig. 4 False detection rate

the detection rate. Approximately, 80% of faulty readings can be detected by STCOD, while only 50% of faulty readings can be detected by MVA and WVA, as can be seen. The other two voting algorithms seem to be outperformed by STCOD.

The second experiment compares STCOD's false detection rate to that of MVA and WVA. The result is shown in Fig. 4, with the X-axis representing the rate of defective data and the Y-axis representing the rate of false detection. Since there are more faulty sensors, the false detection rate of three algorithms increases in tandem with the faulty rate. When the defective rate is high, MVA and WVA have a hard time correctly distinguishing faulty data from outliers. STCOD, on the other hand, has a strong showing.

6 Conclusion

This paper suggests an effective method to detect anomalies/outliers on the basis of spatial and temporal correlations. This paper proposed Pearson Correlation and mutual information to detect similarity on the basis of Spatial and temporal correlations. Our future work will focus on how to separate errors from events and how to undeliver erroneous data to cloud there by saving energy in IoT sensor network.

References

1. T.B. Dang, D.T. Le, T.D. Nguyen, M. Kim, H. Choo, Monotone split and conquer for anomaly detection in IoT sensory data. IEEE Int. Things J. (2021)
2. Y. Djenouri, A. Belhadi, G. Srivastava, U. Ghosh, P. Chatterjee, J.C.W. Lin, Fast and accurate deep learning framework for secure fault diagnosis in the industrial internet of things. IEEE Int. Things J. (2021)
3. D. ElMenshawy, W. Helmy, Detection techniques of data anomalies in IoT: a literature survey. Technology **9**(12), 794–807 (2018)
4. S. Gopikrishnan, P. Priakanth, G. Srivastava, DEDC: sustainable data communication for cognitive radio sensors in the internet of things. Sustain. Comput. Inf. Syst. **29**, 100471 (2021)
5. D.J. Hand, Principles of data mining. Drug Saf. **30**(7), 621–622 (2007)
6. K. Jitkajornwanich, N. Pant, M. Fouladgar, R. Elmasri, A survey on spatial, temporal, and spatio-temporal database research and an original example of relevant applications using SQL ecosystem and deep learning. J. Inf. Telecommun. **4**(4), 524–559 (2020)
7. R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar, B. Qureshi, An overview of IoT sensor data processing, fusion, and analysis techniques. Sensors **20**(21), 6076 (2020)
8. N. Lathia, D. Quercia, J. Crowcroft, The hidden image of the city: sensing community well-being from urban mobility, in *International Conference on Pervasive Computing* (Springer, 2012), pp. 91–98
9. N. Nesa, T. Ghosh, I. Banerjee, Outlier detection in sensed data using statistical learning models for IoT, in *2018 IEEE Wireless Communications and Networking Conference (WCNC)* (IEEE, 2018), pp. 1–6
10. J. Numata, O. Ebenhöh, E.W. Knapp, Measuring correlations in metabolomic networks with mutual information, in *Genome Informatics 2008: Genome Informatics Series*, vol. 20 (World Scientific, 2008), pp. 112–122

11. I. Portugal, P. Alencar, D. Cowan, A framework for spatial-temporal trajectory cluster analysis based on dynamic relationships. IEEE Access **8**, 169775–169793 (2020)
12. D. Quercia, D. Ò. Séaghdha, J. Crowcroft, Talk of the city: our tweets, our community happiness, in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6 (2012)
13. S. Ray, Y. Jin, A. Raychowdhury, The changing computing paradigm with internet of things: a tutorial introduction. IEEE Des. Test **33**(2), 76–96 (2016)
14. M. Safaei, S. Asadi, M. Driss, W. Boulila, A. Alsaeedi, H. Chizari, R. Abdullah, M. Safaei, A systematic literature review on outlier detection in wireless sensor networks. Symmetry **12**(3), 328 (2020)
15. J.M.T. Wu, L. Sun, G. Srivastava, J.C.W. Lin, A novel synergetic LSTM-GA stock trading suggestion system in internet of things. Mob. Inf. Syst. (2021)
16. Z. Yue, W. Sun, P. Li, M.U. Rehman, X. Yang, Internet of things: architecture, technology and key problems in implementation, in *2015 8th International Congress on Image and Signal Processing (CISP)* (IEEE, 2015), pp. 1298–1302