IJITKM Volume 7 • Number 1 • December 2013 pp. 87-91 (ISSN 0973-4414)

Outlier Detection Techniques for Wireless Sensor Networks using Clustering and Support Vector Machines

Deepak Sinwar[#], Dr. Sudesh Kumar*
[#]Assistant Professor, *Associate Professor
Department of Computer Science & Engineering
BRCM College of Engineering & Technology, Bahal, Bhiwani
E-mail: deepak.sinwar@gmail.com, sudeshjakhar@gmail.com

ABSTRACT: Wireless Sensor Networks are vulnerable to many types of security attacks, including false data injection, data forgery, and eaves dropping. Sensor nodes can be compromised by intruders, and the compromised nodes can distort data integrity by injecting false data. The transmission of false data depletes the constrained battery power and degrades the band-width utilization. This paper examines various techniques to detect false nodes i.e. outliers. The main emphasis is to detect outliers on the basis of distance measures. Thus clustering and support vector machines are used as a basis. At the end we are able to answer various questions regarding outliers and their detection from wireless sensor networks.

Keywords: Outlier detection, sensor networks, fault tolerance, clustering.

1. INTRODUCTION

Wireless sensor networks (WSNs) have been widely used in various applications including those related to personal, industrial, business and military domains [1]. Many of these applications utilize real-time sensor data collected by WSNs to monitor the surrounding environment and detect time-critical events occurred in the physical world. Data collected by WSNs are often unreliable and inaccurate due to the following reasons: (i) the low cost and low quality sensor nodes have stringent resource constraints such as energy (battery power), memory, computational capacity, and communication bandwidth; (ii) operation of sensor nodes which are randomly deployed in a large area (and often with high density) are frequently susceptible to harsh and unattended environmental effects; (iii) sensor nodes are vulnerable to malicious attacks such as denial of service attacks, black hole attacks and eavesdropping. Sensor networks not only provide with real time data but also detect time-critical events which produce deviation from expected data results. The data provided by sensor networks are often unreliable. Data quality is affected by many factors like errors, missing values and compromised or malfunctioning nodes [4]. To keep the data quality and reliability high and be able to make effective and correct decisions using data collected by WSNs, it is essential to identify erroneous data as well as potential events and malicious attacks occurred in the network. Outliers in WSNs are those measurements that significantly deviate from the normal pattern of the sensed data. On the other hand advances in sensor technology and wireless communication have enabled deployment of low-cost and low-power sensor nodes that are integrated with sensing, processing, and wireless communication capabilities [2]. A geosensor network consists of a large number of these sensor nodes distributed in a large area to collaboratively monitor phenomena of interest. The monitored geographic space may vary in size and can range from small-scale room-sized spaces to highly complex dynamics of ecosystem regions. One important task of a typical sensor network is to monitor, detect, and report the occurrences of interesting events (e.g. forest fire, chemical spills, etc.) with the presence of faulty sensor measurements. These events usually span some geographic region and in many application scenarios the detection of the event boundary may become more important than the detection of the entire event region [3]. A good example is the timely estimation of the possible reach of the contamination in a surveillance network monitoring the transportation of chemical spills in soil. On the other hand, individual sensor reading is not reliable. Filtering out faulty readings and transmitting only the boundary information to the base station can save energy. In this paper we target the problem of identifying faulty sensors and detecting event boundaries in sensor networks with faulty sensors.

The rest of the paper is organized as follows. Background work will be outlined in section 2, whereas section 3 will describe basic clustering based outlier detection methods. Support Vector Machine based outlier detection approaches are briefed in section 4. Section 5 concludes with future work suggestions.

2. BACKGROUND WORK

Wireless Sensor Networks are vulnerable to many types of security attacks, including false data injection, data forgery, and eaves dropping. Sensor nodes can be compromised by intruders, and the compromised nodes can distort data integrity by injecting false data. The transmission of false data depletes the constrained battery power and degrades the band-width utilization. False data can be injected by compromised sensor nodes in various ways, including data aggregation and relaying. Because data aggregation is essential to reduce data redundancy and/or to improve data accuracy, false data detection is critical to the provision of data integrity and efficient utilization of battery power and bandwidth. In addition to false data detection, data confidentiality is required by many sensor network applications to provide safeguard against eavesdropping. Outlier detection techniques designed for WSNs can be categorized into statistical-based, nearest neighbor-based, clustering based,

classification-based, and spectral decomposition-based approaches [9] Classification-based approaches are important systematic approaches in the data mining and machine learning community. Classification-based techniques learn a classification model using a set of data instances in the training phase and classify an unseen instance into one of the learned (normal/outlier) class in the testing phase. SVM-based techniques are from family of classification-based approaches and have the following three main advantages: (a) have a simple geometric interpretation; (b) provide an optimum solution for classification by maximizing the margin of the decision boundary; (c) avoid the problem of the curse of dimensionality. The fact that in many WSNs applications, pre-classified normal/anomalous data is neither always available nor easy to obtain implies that unsupervised classification techniques suit the WSNs the best. Therefore, several unsupervised (one-class) SVM-based outlier detection techniques have been proposed [10], which model the normal pattern of the unlabelled data while automatically ignore the anomalies existed in the training data. The main idea of one-class SVM-based outlier detection approaches is that data measurements collected from the original space (input space) are first mapped to a higher dimensional space (feature space) using a non-linear function $\varphi(x)$. Then a decision boundary of normal data is found, which encompasses the majority of the data measurements in the feature space. Those falling outside the boundary are classified as anomalous.

Sheng et al. [11] developed a framework for the discovery of k-nearest-neighbor based outliers: points whose distance to their k-nn exceeds a fixed threshold or the top n points with respect to the distance to their k-nns. Each sensor maintains a histogram-type summary of pertinent information over a sliding window of its data points. The sink node collects these summaries and queries the network for any additional information needed to correctly determine the outliers over the whole network. The use of summaries allows less communication than a naive, centralized approach. Their approach differs from ours in several ways. First, they only detect outliers over one dimensional data, and difficulty of building compact, multi-dimensional histograms will hinder any extension beyond that. Second, they only consider the two k-nn-based outlier definitions described above, while our approach encompasses these and more. Thirdly, their approach only applies in settings where spatial proximity is unimportant while our approach can, if needed, to accommodate spatial proximity ("semi-local" outlier detection). Subramaniam et al. [12] require the sensors to maintain a tree communication topology and compute outliers using an estimate of the underlying probability distribution from which the data arises. Such an estimate is computed by each sensor maintaining a random sample of its data observations. Our approach differs in at least four ways. First, ours does not make any assumptions about the communication topology (e.g., that it is a tree), save that it is connected. Second, ours computes outliers with respect to all of the data observations at each sensor, not a sample. Third, ours can smoothly take into account spatial proximity among the sensors ("semi-local" outliers) while Subramaniam et. al do not focus on this task. Fourth, our approach is designed to smoothly adjust to changes in the underlying network topology while theirs requires that the underlying communication tree be reestablished by other means before the algorithm can resume operation. Janakiram et al. [13] developed a framework based on a Bayesian Belief Network (BBN) that has been constructed over the WSN (and distributed to each sensor). Using this, each sensor can estimate the likelihood of an observed tuple and, therefore, detect outliers, yet it is not clear to what extent the BBN construction phase can be carried out in-network. Moreover, the authors do not discuss the problem of updating the BBN given network/data change. In contrast, our processing is entirely in-network and smoothly adjusts to changes in data/network. Zhuang and Chen [14] use a wavelet-based technique for correcting large isolated spikes from single sensor data streams. A dynamic time warping (DTW) distance-based technique is also used to identify more steady intervals of erroneous sensor data by comparing the data streams of spatially close sensors assumed to produce similar streams. To reduce energy consumption, anomalous data streams are not transmitted to the base station. Our method is similar in that it is in-network. However, Zhuang and Chen's use of DTWis tightly integrated with a minimum hop count routing algorithm, which makes the approach more restrictive than ours. Rajasegarar et al. [15] describe an approach that is based on distributed non-parametric anomaly detection and requires sensors to maintain a tree communication network topology. Here, each sensor clusters its sampled measurements using a fixed-width clustering algorithm, then extracts statistics of the clusters (i.e., the centroid and number of contained data vectors), and then sends them its parent node. The parent uses its children's cluster statistics to form amerged cluster. The parent then transmits that cluster to its own parent. This process continues recursively until the base station receives all clusters, after which it will perform anomaly detection to identify all outliers. While this approach supports energy-efficiency by distributing the clustering operation throughout the network, anomaly detection is only performed at the base station. Our approach differs in that it distributes the anomaly detection process itself throughout the network, quickly enabling nodes to identify outliers and autonomously make further data processing decisions. Nor does our approach rely on the use and maintenance of a routing tree; hence, it smoothly adjusts to changes in the underlying network topology. Adam et al. [16] address the issue of accounting for spatially neighboring peers when detecting outliers in sensor networks. However, they assume the sensor datasets are centralized and the outlier processing is carried out at the central processing node. They do not consider the problem of carrying out the outlier detection in-network as we do. Palpanas et al. [17] propose a technique for distributed deviation detection using a network hierarchy of low- and high-capacity sensors that are differentiated with respect to processing

power and communication range. Here, low-capacity sensors aim to detect local outliers while high-capacity sensors detect more spatially dispersed outliers using an aggregation of low-capacity sensors' data. Kernel density estimators are used to model the distribution of data values reported by sensors, and distance-based detection techniques are used for identifying outliers. The authors present no formal evaluation of the proposed technique. Our approach differs in that it does not rely on a hierarchy of device capabilities. Radivojac et al. [18] address the process of sensors learning data distributions from class-imbalanced data. Here, sensors send data points to a central base station that generates a classification model from class-imbalanced data (i.e., having abundant negative samples and few positives) and the total cost of detection and classification (e.g., costs of transmitting false positives and false negatives). In contrast, our framework operates in-network.

3. OUTLIER DETECTION

Outlier is defined as an observation that deviates too much from other observations. The identification of outliers can lead to the discovery of useful and meaningful knowledge and has a number of practical applications in areas such as transportation, ecology, public safety, public health, climatology, and location based services [7]. Jingke Xi has presented outliers into two categories viz. classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approach analyzes outlier based on spatial dataset, which can be grouped into space based approach, graph-based approach. Thirdly, they conclude some advances in outlier detection recently.

3.1 Clustering based Outlier Detection

Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In statistics, an outlier is an observation that is numerically distant from the rest of the data. Grubbs defined an outlier as: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high kurtosis and that one should be very cautious in using tools or intuitions that assume a normal distribution.

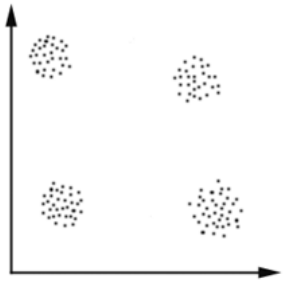

Figure 1: Clustering of data

Causes: Outliers can have many anomalous causes. A physical apparatus for taking measurements may have suffered a transient malfunction. There may have been an error in data transmission or transcription. Outliers arise due to changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. A sample may have been contaminated with elements from outside the population being examined. Alternatively, an outlier could be the result of a flaw in the assumed theory, calling for further investigation by the researcher. Additionally, the pathological appearance of outliers of a certain form appears in a variety of datasets, indicating that the causative mechanism for the data might differ at the extreme end (King effect).

There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outlier detection can identify system faults and fraud before they escalate with potentially catastrophic consequences. The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, drawn from the full gamut of computer science and statistics. There are three fundamental approaches to the problem of outlier detection:

Type 1 - Determine the outliers with no prior knowledge of the data. This is essentially a learning approach analogous to unsupervised clustering. The approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers.

Type 2 - Model both normality and abnormality. This approach is analogous to supervised classification and requires pre-labeled data, tagged as normal or abnormal.

Type 3 - Model only normality (or in a few cases model abnormality). This is analogous to a semi-supervised recognition or detection task. It may be considered semi-supervised as the normal class is taught but the algorithm learns to recognize abnormality.

4. NOTATIONS AND NETWORK MODEL

Assume that N sensors are uniformly deployed in a b×b squared field. A sensor's reading is faulty (abnormal) if it deviates significantly from other readings of neighboring sensors [3]. Sensors with faulty readings are called faulty sensors. Generally we use S to denote the set of all the

sensors in the field and R denote the radio range of the sensors. Let x_i denote the reading of the sensor S_i. Instead of a 0-1 binary variable, x_i is assumed to represent the actual reading of a factor or variable, such as temperature, light, sound, the number of occurrences of some phenomenon, and so on. For example, a rogue node that continues to inject messages to the network or drop all relay messages in DOS attack is a misbehaving node. Therefore, x_i can be continuous or discrete. A faulty sensor can be viewed as a special event which contains only one point, i.e., the sensor itself. Each sensor can compute its physical position through either GPS or some GPS-less techniques.

4.1 Support Vector Machine based Technique

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. SVM-based techniques are from family of classification-based approaches and have the following three main advantages: (i) have a simple geometric interpretation; (ii) provide an optimum solution for classification by maximizing the margin of the decision boundary; (iii) avoid the problem of the curse of dimensionality [19]. The main idea of one-class SVM-based outlier detection approaches is that data measurements collected from the original space (input space) are first mapped to a higher dimensional space (feature space) using a non-linear function. Then a decision boundary of normal data is found, which encompasses the majority of the data measurements in the feature space. Those falling outside the boundary are classified as anomalous.

4.1.1 Hyperellipsoidal SVM VS Hyperspherical SVM

Hyperspherical SVM assumes that the target sample points are distributed around the center of mass in an ideal spherical manner. However, if the data distribution is non-spherical, using a spherical boundary to fit the data will increase the false alarm rate and reduces the detection rate. This is because many superfluous outliers are mistakenly considered in the boundary and consequently outliers are classified as normal. On the contrary, the hyperellipsoidal SVM is able to best capture multivariate data structures by considering not only the distance from the center of mass but also the data distribution trend, where the latter is learned by building the covariance matrix of the sample points. This feature can be used well for geosensor data, where multivariate attributes may induce certain correlation, e.g., the readings of humidity sensors are negatively correlated to the readings of temperature sensors. A hyper ellipsoidal boundary is used to enclose the majority of the data vectors in the feature space as shown in figure 2.

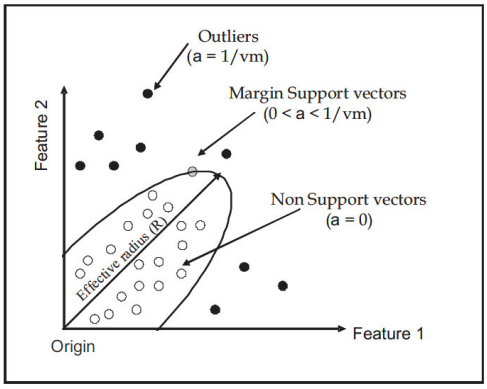

Figure 2: Geometry of the hyper ellipsoidal formulation of one-class SVM [19]

Another method on the basis of measurement collected from each node has been developed by Zhang et al. [2]. They assume that wireless sensor nodes are time synchronized and densely deployed in a homogeneous geosensor network, where sensor data tends to be correlated in both time and space. A sensor sub-network consists of n sensor nodes S1; S2; : : : Sn, which are within radio transmission range of each other. This means that each node has n-1 spatially neighboring nodes in the sub-network. At each time interval ti, each sensor node in the sub-network measures a data vector. The identification of anomalous node can be detected by means of local data processing at each node. In addition to near real- time identification of outliers, increasing data quality, and reducing communication overhead, this local processing also has the advantage of coping with (possibly) large scale of the geosensor network. Thus using an ellipsoidal boundary to enclose geosensor data aims to increase outlier detection accuracy and reduce the false alarm rate. However, as a tradeoff, the hyperellipsoidal SVM has more computational and memory usage cost than the hyperspherical SVM. To correctly select the most appropriate outlier detection technique, we believe that having some understanding about data distribution and correlation among sensor data is crucial.

5. Conclusion and Future Work

This paper has examined clustering and support vector machine based methods for detecting outliers from wireless sensor networks. The main focus was on detecting outliers on the basis of distance measures. There exist a lot of other methods to detect outliers from WSN's. The work can be extended by combining distance measures along with other methods in a real environment.

References

[1] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "An Outlier Detection Techniques for Wireless Sensor Networks using Unsupervised Quarter-Sphere Support Vector Machine", Fourth Intl Conf. on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2008

[2] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Hyperellipsoidal SVM-based Outlier Detection Technique for GeoSensor Networks", GeoSensor Networks, Third Intl. conf., GSN, Oxford, pp. 31-41, 2009.

[3] M. Ding, D. Chen, K.Xing and X.Cheng, "Localized Fault-Tolerant Event Boundary Detection in Sensor Networks", in: 24th Annual Joint Conf. of the IEEE Computer and Communications Societies, 902 - 913 vol. 2, 2005

[4] N.C. Devia, V. Palanisamyb, K. Baskaranc, and S.Prabeelada, "A Novel Distance for Clustering to Support Mixed Data Attributes and Promote Data Reliability and Network Lifetime in Large Scale Wireless Sensor Networks", in proceedings of Intl. Conf. on Communication Technology and System Design, pp. 679-677, 2011

[5] A. Fawzy, H.M. O. Mokhtar, O. Hegazy. "Outliers detection and classification in wireless sensor networks", Egyptian Informatics Journal, pp. 157–164, 2013

[6] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff and H. Kargupta, "In-network outlier detection in wireless sensor networks", Knowledge Information Systems, pp. 23-54, 2013

[7] J. Xi, "Outlier Detection Algorithms in Data Mining", IEEE IITA, 2008, PP. 94-97

[8] E. Muller, I. Assent, U. Steinhausen, T. Seidl, "OutRank: ranking outliers in high dimensional data", IEEE ICDE Workshop, 2008, pp. 600-603

[9] Y. Zhang, N. Meratnia, and P. J. M. Havinga, Outlier Detection Techniques for Wireless Sensor Network: A Survey, Technical Report, University of Twente, 2008.

[10] P. Laskov, C. Schafer, and I. Kotenko, Intrusion Detection in Unlabeled Data with Quarter Sphere Support Vector Machines, in Detection of Intrusions and Malware & Vulnerability Assessment, (Dortmund), 2004.

[11] Sheng B, Li Q, Mao W, Jin W (2007) Outlier detection in sensor networks. In: Proceedings of the 8th ACM international symposium on mobile and ad hoc networking and computing (MobiHoc), pp 219–228

[12] Subramaniam S, Palpanas T, Papadopoulos D, KalogerakiV, GunopulosD(2006) Online outlier detection in sensor data using non-parametric models. In: Proceedings of ACM conference on very large databases (VLDB06), pp 187–198

[13] Janakiram D, Reddy VA, Kumar AVUP (2006) Outlier detection in wireless sensor networks using Bayesian belief networks. In: Proceedings of IEEE conference on communication system software and middleware (Comsware06), pp 1–6

[14] Zhuang Y, Chen L (2006) In-network outlier cleaning for data collection in sensor networks. In: Proceedings of the 1st international VLDB workshop on clean databases (CleanDB06)

[15] Rajasegarar S, Leckie C, Palaniswami M, Bezdek J (2006) Distributed anomaly detection in wireless sensor networks. In: Proceedings of the IEEE Singapore international conference on communication systems, pp 1–5

[16] Adam N, Janeja V, Atluri V (2004) Neighborhood-based detection of anomalies in high dimensional spatio-temporal sensor datasets. In: Proceedings of ACM symposium on applied computing (SAC04), pp 576–583

[17] Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D (2003) Distributed deviation detection in sensor networks. In: ACM SIGMOD Record, pp 77–82

[18] Radivojac P, Korad U, Sivalingam KM, Obradovic Z (2003) Learning from class-imbalanced data in wireless sensor networks. In: Proceedings of the IEEE 58th vehicular technology conference, vol 5, pp 3030–3034

[19] Zhang, Y., Meratnia, N., Havinga, P.J.M.: An Online Outlier Detection Technique for Wireless Sensor Networks using Unsupervised Quarter-Sphere Support Vector Machine. In: 4th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 151-156. IEEE Press, Sydney (2008)