Outlier Preprocessing in Wireless Sensor Networks: A Two-layered Ellipse Approach

Ibrahim Khamis and Zeyar Aung
Institute Center for Smart and Sustainable Systems (iSmart)
Department of Computer Science and Electrical Engineering
Masdar Institute of Science and Technology
Masdar City, P. O. Box 54224
Abu Dhabi, United Arab Emirates
Emails: {ikhamis, zaung}@masdar.ac.ae

***Abstract*—Sensor nodes in wireless sensor networks have limited energy resources and this hinders the dissemination of the gathered data to a central location. This stimulated our research to make use of the limited computational capabilities of these sensor nodes to build a normal model of the data gathered. Hence by having the normal model, anomalies can then be detected and forwarded to a central location. This process is done locally in the sensor nodes and hence reduces the power consumption used in transmitting all the data. Our algorithm is an enhanced version of the Data Capture Anomalies Detection Algorithm; which is used to compute a local model of the normal data in wireless sensor networks. In this paper the Data Capture Anomalies Detection is used to partition the data and then send all the data to a central server for data classification, building on the Data Capture Anomalies Detection method and in order to classify the partitioned data; our algorithm Two-Layered Data Capture Anomalies Detection sends anomalies (2%) as well as roughly (2% or 4%) of normal data for further data processing and classification purposes. Experimental results on synthetic data show that Two-Layered Data Capture Anomalies Detection is able to provide promising results.**

Keywords—wireless sensor network; anomaly detection; elliptical boundary estimation; two-layered ellipse approach; distributed data mining.

I. INTRODUCTION

Wireless sensor network (WSN) refers to a group of spatially dispersed and dedicated sensors for monitoring and recording the physical conditions of the environment and organizing the collected data at a central location. WSNs measure environmental conditions like temperature, sound, pollution levels, humidity, wind speed and direction, pressure, etc. (definition given in [18]). WSNs are widely used in areas such as manufacturing industry [15], military [8], environmental monitoring [1], smart power grids [4], smart buildings/homes [9], and many other applications that require distributed location-aware data sensing [2].

The advantage of using WSNs is that they are cheaper and more practical then wired networks. However WSNs are vulnerable to intruders and faults [4]. In general, WSNs need to be data mined to find anomalies as efficient as possible and then send these data for the base station or the central location for further processing.

In order to find a balance among the three desirable factors of speed, accuracy, and low energy consumption for outlier preprocessing and detection, a Two-Layered Data Capture Anomaly Detection (TLDCAD) is proposed. The algorithm is based on the Ellipsoidal Data Capture Anomaly Detection (DCAD) method [14] where it collects the data then calculates the mean and the covariance matrix of the data. The ellipse is then drawn from the covariance matrix by using the Eigen structure of the covariance matrix.

The anomalies then are detected by setting a level set on the ellipse's boundary to cover at most 98% of the data and any data exceeds this level is considered to be anomalies and captured and send for further processing. In TLDCAD, in addition to that level, we set another level at 94% or 96% and capture an additional layer of data points between new level and the original 98% level, label them as normal, and send these data also for classification purposes. Note that the new level (96% or 94%) is based on the inverse of chi squared statistics with probability of 98% and degree of freedom of 2.

Below are some of the motivations behind our method:

- Reducing the power consumption in resource constrained devices like wireless sensors.
- Reducing the communication's noise level by preprocessing inside the sensor node then send a reduced sampled output to the server for further classification and exploration.
- A new approach of data preprocessing, by providing sampled data using two ellipses.
- Producing more balanced data sets with around 50% or 25% outliers instead of just 2% outliers. (Some classifiers do not work well with extremely unbalanced data sets like a typical Support Vector Machine – SVM.)

II. BACKGROUND AND RELATED WORKS

Wireless Sensor Network (WSN) is a network that consists of number of nodes; each one is connected to other nodes wirelessly in the network. WSN are feasible solutions in situations where it is difficult, or costly or even impractical to implement wired networks [14].

There are many studies about outlier detection. For example, Jiang and Yang made clusters as a unit and find the outliers clusters as a unit [6]. In this case the whole cluster becomes an outlier. Lee *et al.* proposed a novel work for trajectory outlier detection [7]. The abnormal trajectory among other trajectories becomes an outlier. Moreover, Menold *et al.* inferred that data point is compared to the median of the past and present value and the result is outlier if it exceeds certain threshold [11]. This show the implementation of the temporal data (data related to time). On the other hand some people are concerned with the privacy issue of outlier detection. In Challagalla *et al.* [1], detection of outliers threats some organizations and raises their concerns about the privacy of the analyzed data. For that reason it is important to incorporate some sort of privacy protection in the outlier detection technique.

There are many approaches to classify outlier detection. Zhang *et al.* [17] divided outlier detection in WSNs to three branches in terms of the outlier sources, the first is Fault detection in WSNs and it deals with noise and errors, the second division is event detection in WSNs which deal with events. The last division is intrusion detection in WSNs and this one handles the malicious attacks.

Qu *et al.* [13] gives the categorization of the outlier detection methods into 5 main groups: distribution-based, depth-based, clustering, distance-based, and density-based. The widely used ones are density-based and distance-based. Li and Kitagawa [10] observed that the distance-based method is one of the most common and simplest methods for outlier detection.

Xi has provided the following classification for outlier detection methods [15]. In this classification he divided the outlier detection algorithms to three main categories. The first main category is classic outlier which in turn is divided to four sub categories; statistical based, distance based, deviation based, and density based approaches. The second main category is the spatial outlier and this is just a modification of the classic based approach by taking into account the spatial attributes of the data. Spatial attributes are the attributes that relate to location. The third outlier detection main category implicitly stated by Xi is the "recent advances" in outlier detection. In this category there are two sub categories; high dimension based approach and SVM based approach.

Janssens *et al.* [5] also compared some method from Machine Learning (ML) and Knowledge Discovery in Databases (KDD). The ML techniques used are SVM and Parzen Windows, and the KDD techniques used are heuristic local-density estimation methods such as LOF and LOCI. They used the one class classification framework. He selects this framework to be able to use AUC (Area under the Curve) which is a famous performance measurement tool. Janssens *et al.* found that Support Vector Domain Description (SVDD) is one of the best performing methods [5]. Moreover, the authors of [17] proposed taxonomy for outlier detection techniques. The main categories are statistical based which is further subdivided to parametric and non parametric, nearest neighbor based, clustering based, classification based, and finally the spectral decomposition based. The parametric based is divided to Gaussian based and non Gaussian based. The non parametric based is divided to kernel based and histogram based. The classification based is divided to Support vector machine based and Bayesian network-based. The Bayesian network based is subdivided again to naïve Bayesian network based, Bayesian believe network based, and dynamic Bayesian network based. The spectral decomposition is subdivided to the principle component analysis only.

III. DEFINITIONS AND NOTATIONS

In this section, we introduce our notations and formulas that are used.

The definitions and notations used are mainly adopted from [12]. The data are represented by $X_k = \{x_1, x_2, \dots, x_k\}$ where x_1 is the first sample and it is a $d \times 1$ vector in $\Re^d$. Each element in the vector represents a measurement attribute of interest such as temperature and relative humidity. The sample mean m_k of X_k can be calculated using Equation 1 and the sample covariance S_k can be calculate by Equation 2.

$$m_k = \frac{1}{k}\sum_{j=1}^{k} x_j \tag{1}$$

$$S_k = \frac{1}{k-1}\sum_{j=1}^{k}(x_j - m_k)(x_j - m_k)^T \tag{2}$$

The hyperellipsoid with effective radios t centered at m_k with covariance matrix S_k is defined as:

$$e_k(m_k, S_k^{-1}; t) = \left\{ x \in \Re^d \middle| (x_j - m_k)^T S_k^{-1}(x_j - m_k) \le t^2 \right\} \tag{3}$$

Comment 1: $(x_j - m_k)^T S_k^{-1}(x_j - m_k)$ is the Mahalonobis distance from m_k to x_j and S_k^{-1} is the characteristic matrix of e_k.

The boundary of hyperellipsoid e_k is defined as:

$$\delta_{e_k}(m_k, S_k^{-1}; t) = \{ x \in \Re^d | (x_j - m_k)^T S_k^{-1}(x_j - m_k) = t^2 \} \tag{4}$$

Comment 2: using $t^2 = (\chi_d^2)_p^{-1}$ with $p = 0.98$ results in a hyper ellipsoidal boundary that covers at least 98% of the data under the assumption that the data are in normal distribution [12].

Definition 1: single point anomaly with respect to e_k is defined as any data point that is outside the boundary. x is anomaly for:

$$e_k \Leftrightarrow (x_j - m_k)^T S_k^{-1} (x_j - m_k) > t^2 \quad (5)$$

Definition 2: 2% single points normal with respect to e_k is defined as any data point that lies between the corresponding values of t^2 by setting $p > 0.96$ and $p \leq 0.98$.

Definition 3: 4% single points normal with respect to e_k is defined as any data point that lies between the corresponding values of t^2 by setting $p > 0.94$ and $p \leq 0.98$.

Definition 4: 98% single points normal with respect to e_k is defined as any data point that lies between the corresponding values of t^2 by setting $p > 0.00$ and $p \leq 0.98$.

IV. METHODOLOGY

The main methods used are adding an additional layer to DCAD to get the TLDCAD and then label the output data with two labels (classes): "anomalous" and "normal". Then, we send the data to a classifier Support Vector Machine (SVM) or Artificial Neural Networks (ANN) and then compare the output data from the DCAD and TLDCAD to draw the final conclusion. The flowchart in Fig. 1 summarizes the methodology used in this paper.

V. EVALUATION

A. Datasets

The synthetic data are generated using the following settings where Σ is the covariance matrix and μ is the mean for the synthetic data sets as described in [12].

$$\Sigma = \begin{pmatrix} 0.6797 & 0.1669 \\ 0.1669 & 0.7891 \end{pmatrix}$$

$$\mu = (5, 5)$$

The synthetic data are two dimensional. In reality, each dimension represents one attribute of the sensed data. For example, the X-axis may represents "temperature" and Y-axis may represent "humidity".

We used two data sets with 5,000 and 50,000 two-dimensional data points respectively.

B. Classifiers

The testing on the output data from the ellipse samples are evaluated on two main classifiers: Support Vector Machine (SVM) and Artificial Neural Network (ANN).

Fig. 2 is a display of how the data are generated and how they are preprocessed using the ellipses for the SVM classifier. (That is for the ANN classifier is virtually the same.)

C. Results

Tables I and II show some promising and even competitive results from our approach TLDCAD in comparison with the approach of classifying all the data points from the DCAD. Note that how the F1 measure value increases with the increase of the normal layered data.

For example the average value for the SVM cross validation was 0.963 for 2% TLDCAD and increased to 0.989 which is very competitive to the DCAD 98% normal output which has the value for F1 measure of 0.998. That is by using TLDCAD with 4% normal data we can obtain very competitive accuracy measures and using only 6% of the data instead of using 100% of the data in order to have just 1% of accuracy increase.

The main advantage of TLDCAD over DCAD in a classification context; is the TLDCAD runtime efficiency, which is especially important in WSNs in order to save power resources. The much reduced running times of TLDCAD over DCAD on a standard PC for various experimental setups can be observed in Tables I and II.

VI. CONCLUSION AND FUTURE WORKS

In This work, the DCAD algorithm is used in this paper to partition the data and then sends all the data for classification purposes. Building on this work we proposed a TLDCAD algorithm to send reduced amount of data than the data obtained by the DCAD. The output of the two algorithms is compared; the results obtained show promising results for TLDCAD. The current work is conducted using synthetic datasets; however the research in the process of gathering renewable energy projects data to test the TLDCAD algorithm on real data from renewable projects.

Our future research direction involves information security since we preprocess the data inside the node and send sample of the data which avoid the noise while sending all the data. In addition TLDAD helps for security and privacy proposes when part and not all the data are communicated.

ACKNOWLEDGMENT

This research was sponsored by the Government of Abu Dhabi, United Arab Emirates through its funding of Masdar Institute of Science and Technology's research project on "Monitoring and Optimization of Renewable Energy Generation using Wireless Sensor Data Analytics."

REFERENCES

[1] M. A. Azim, F. M. Kiaie, and M. H. Ahmed, "Environmental forest monitoring using wireless sensor networks," in *Wireless Sensor Networks: Current Status and Future Trends*, CRC Press, 2012, pp. 61-78.

[2] M. A. Azim, Z. Aung, W. Xiao, V. Khadkikar, and A. Jamalipour, "Localization in wireless sensor networks by constrained simultaneous perturbation stochastic approximation technique," in *Proceedings of the 6th IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, Gold Coast, Queensland, Australia, 2012, pp. 1-9.

[3] A. Challagalla, S. S. S. Dhiraj, D. V. L. N. Somayajulu, T. S. Mathew, S. Tiwari, and S. S. Ahmad, "Privacy preserving outlier detection using hierarchical clustering methods," in *Proceedings of the 34th IEEE Annual Computer Software and Applications Conference Workshops (COMPSACW*), Seoul, Korea, 2010, pp. 152-157.

[4] M. A. Faisal, Z. Aung, J. Williams, and A. Sanchez, "Securing advanced metering infrastructure using intrusion detection system with data stream mining," in *Proceedings of the 2012 Pacific Asia Workshop on Intelligence and Security Informatics (PAISI)*, Kuala Lumpur, Malaysia, 2012, pp. 96-111.

[5] J. H. M. Janssens, I. Flesch, and Eric O. Postma, "Outlier detection with one-class classifiers from ML and KDD," in *Proceedings of the 2009 International Conference on Machine Learning and Applications (ICMLA)*, Miami Beach, Florida, USA, 2009, pp. 147-153.

[6] S.-Y. Jiang and A.-M. Yang, "Framework of clustering-based outlier detection," in *Proceedings of the 6th international conference on Fuzzy Systems and Knowledge Discovery (FSKD) Volume 1*, Tianjin, China, 2009, pp. 475-479.

[7] J. G. Lee, J. Han, and X. Li, "Trajectory outlier detection: A partition-and-detect framework," in *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE)*, Cancun, Mexico, 2008, pp. 140-149.

[8] S. H. Lee, S. Lee, H. Song, and H.-S. Lee, "Wireless sensor network design for tactical military applications: Remote large-scale environments," in *Proceedings of the 2009 IEEE Conference on Military Communications (MILCOM)*, Boston, Massachusetts, USA, 2009, pp. 1-7.

[9] D. Li, Z. Aung, S. Sampalli, J. Williams, and A. Sanchez, "Privacy preservation scheme for multicast communications in smart buildings of the smart grid," *Smart Grid and Renewable Energy*, vol. 4, no. 4, 2013, pp. 313-324.

[10] Y. Li and H. Kitagawa, "Db-outlier detection by example in high dimensional datasets," in *Proceedings of the 2007 IEEE International Workshop on Databases for Next Generation Researchers (SWOD)*, Istanbul, Turkey, 2007, pp. 73–78.

[11] P. H. Menold, R. K. Pearson, and F. Allgower, "Online outlier detection and removal," in *Proceedings of the 7th Mediterranean Conference on Control and Automation (MED)*, Haifa, Israel, 1999, pp. 1110-1133.

[12] M. Moshtaghi, C. Leckie, S. Karunasekera, J. C. Bezdek, S. Rajasegarar, and M. Palaniswami, "Incremental elliptical boundary estimation for anomaly detection in wireless sensor networks," in *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM)*, Vancouver, British Columbia, Canada, 2011, pp. 467-476.

[13] J. Qu, "Outlier detection based on Voronoi Diagram," in *Proceedings of the 4th International Conference on Advanced Data Mining and Applications (ADMA)*, Chengdu, China, 2008, pp. 516-523.

[14] S. Rajasegarar, J. C. Bezdek, C. Leckie, and M. Palaniswami, "Elliptical anomalies in wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 6, no. 1, 2009, pp. 1-28.

[15] I. Silva, Luiz A. Guedes, P. Portugal, and F. Vasques, "Reliability and availability evaluation of wireless sensor networks for industrial applications," *Sensors*, vol. 12, no. 1, 2012, pp. 806-838.

[16] J. Xi, "Outlier detection algorithms in data mining," in *Proceedings of the 2nd International Symposium on Intelligent Information Technology Application (IITA)*, Shanghai, China, 2008, pp. 94-97.

[17] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, 2010, pp. 159-170.

[18] http://www.techopedia.com/definition/25651/wireless-sensor-network-wsn

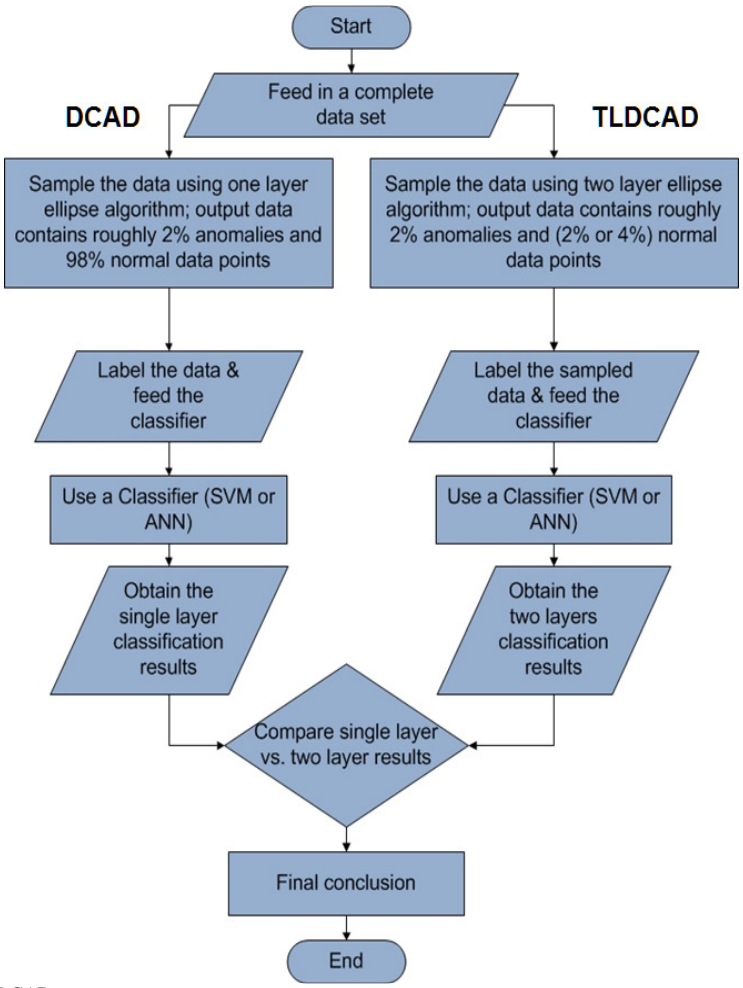


Fig. 1. Flowchart of DCAD vs TLDCAD.

TABLE I. AVERAGE RESULTS FOR 5,000 SYNTHETIC DATA POINTS

		TLDCAD: 2% normal vs. 2% anomalous	TLDCAD: 4% normal vs. 2% anomalous	DCAD: 98% normal vs. 2% anomalous
SVM with SMO	**Precision**	0.958763	0.994845	0.998775
	Recall	0.968750	0.984694	0.999591
	F1	0.963731	0.989744	0.999183
	Time (sec)	3.645893	3.780674	13.269996
ANN using 10 hidden layers	**Precision**	0.900000	0.969697	1.000000
	Recall	0.750000	0.941176	0.989276
	F1	0.818182	0.955224	0.994609
	Time (sec)	2.000225	2.985272	5.224756

TABLE II. AVERAGE RESULTS FOR 50,000 SYNTHETIC DATA POINTS

		TLDCAD: 2% normal vs. 2% anomalous	TLDCAD: 4% normal vs. 2% anomalous	DCAD: 98% normal vs. 2% anomalous
SVM with SMO	**Precision**	0.905149	0.999484	1.000000
	Recall	1.000000	0.961787	0.992368
	F1	0.950213	0.980273	0.996169
	Time (sec)	9.808853	12.681529	85.694093
ANN using 10 hidden layers	**Precision**	0.993243	0.996656	0.999728
	Recall	0.967105	0.980263	0.999728
	F1	0.980000	0.988391	0.999728
	Time (sec)	7.051017	11.293773	248.290314

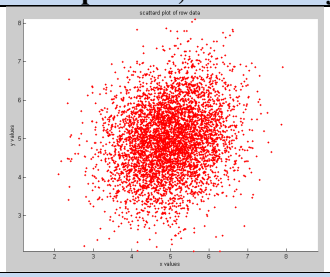
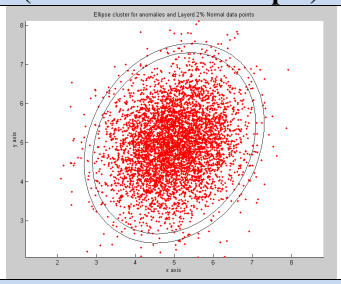
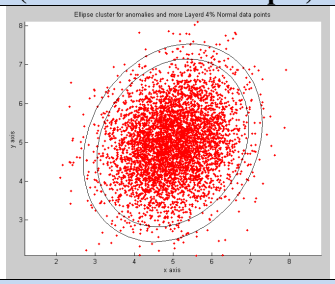
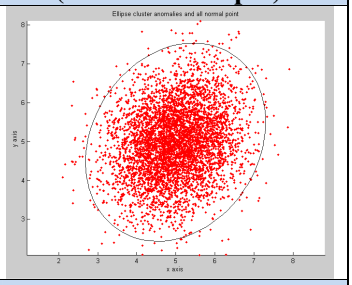
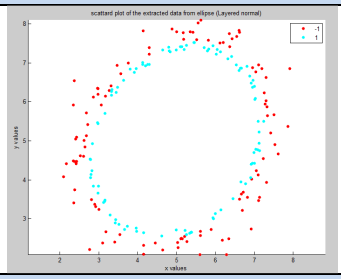
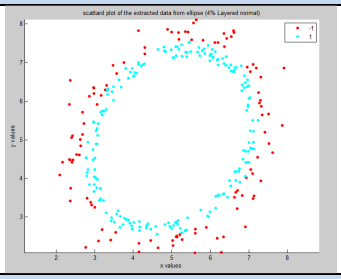
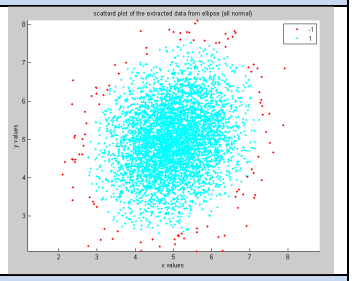
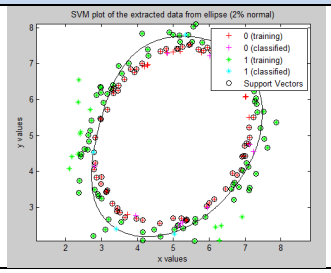
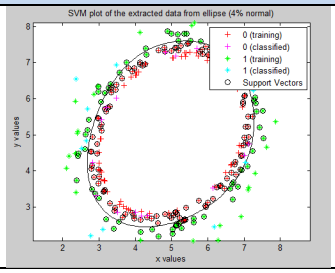
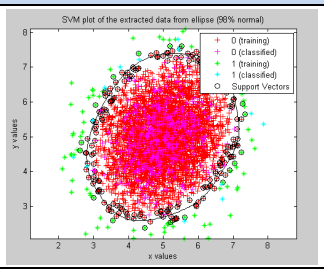
Step 1	Synthetic data generation: scatter plot of 5,000 normally distributed synthetic data		
			
Step 2	TLDCAD's output: 2% normal data (between the two ellipses) and 2% anomalous data (outside the outer ellipse)	TLDCAD's output: 4% normal data (between the two ellipses) and 2% anomalous data (outside the outer ellipse)	DCAD's output: 98% normal data (within the ellipse) and 2% anomalous data (outside the ellipse)
			
Step 3	Scatterd plot for 2% normal data vs. 2% anomalous data	Scatterd plot for 4% normal data vs. 2% anomalous data	Scatterd plot for 98% normal data vs. 2% anomalous data
			
Step 4	SVM output for 2% normal data vs. 2% anomalous data	SVM output for 4% normal data vs. 2% anomalous data	SVM output for 98% normal data vs. 2% anomalous data
			

Fig. 2. Demonstration of how TLDCAD (with 2% and 4% normal data) and DCAD work respectively for SVM classifier.