International Conference on Intelligent Technologies & Science - 2021
(ICITS-2021)

Outlier Detection of WSN Application using ANN

Brijesh Yadav[1], Ravi Prakash Chaturvedi[2], Sameer Asthana[3]
Computer Science Engineering,
United College of Engiineering & Research, Greater Noida., India
brijeshmailon@gmail.com, rpchaturvedi51@gmail.com, sam.asthana@gmail.com

Abstract: The dynamic environment of network and roughness of the working condition are also responsible to generate inaccuracy in measurements. In this paper, an approach for outliers detection based entropy value of received sensor voltages is applied using ANN prediction model .The algorithm development and analysis involves a real time database generated on 14 sets of MICA2 wireless sensor kit with anomaly inserted by real time motion based intrusion in the lab by volunteers from Intel Berkeley lab. On each sensor data pair segmentation is applied by fixed window size in order get large outliers' measurements training dataset. The analysis demonstrates the measurement accuracy in detection of number of outliers that its 86%. Moreover, the algorithm also provides an analysis in terms of impact of variation in learning types and number of nodes in the ANN prediction model. This work is helpful in the application in the situations where high amount of noise or distortions are present. The outlier part from distorted data can be figured out and recollected to enhance application accuracy.

Keywords: Anomaly Detection, Entropy, Outliers Detection, Wireless Sensor Networks.

1. Introduction

The complex and dynamic characteristics of WSNs have made them vulnerable to anomalies. Anomalies are observations that do not correspond to a well defined notion of normal behaviours. [1].In WSNs, anomalies can occur in the nodes, networks, transmission channels and application data and can be caused by systematic errors, random errors and malicious attacks. For instance, WSNs may be deployed in a hostile and inaccessible location, maintenance on the network components is impossible. These nodes usually operate unattended over a long period of time until the battery depleted. Node failure can cause the networks to be unavailable. The networks are also susceptible to systematic hardware failure, random hardware and communication errors, and malicious attacks.

Anomalies in WSNs can be [2] classified into three broad categories, Node, Network and Data anomaly. Node Anomalies occur due to fault at single node. Main reason behind this anomaly is battery issue, i.e. battery failure or depletion. The node fault occurs due to deployment of nodes in harsh environment Unlike node anomalies, the network Anomalies can occur at group of nodes. These are mainly communication related problem. The sensor nodes communicate with each other and if that communication is interrupted due to some reasons then network anomaly occurs. Malicious attacks such DOS, sinkhole, black hole, selective forwarding and wormhole attacks contributes to the occurrence of network anomalies. Data Anomaly occurs when there are some irregularities are present in the sensed data. Some security breaches can also lead to anomalous data. Data anomalies are of three types, viz. Temporal, Spatial and Spatial temporal. Temporal anomaly at a single node location due to changes in data values over time. Spatial anomaly at a single node location due to comparison with neighbouring nodes. Spatiotemporal anomaly detected through a number of node location due to changes in data value over time and space.

2. Related Work

Outlier detection techniques designed for WSNs can be categorized into statistical based, nearest neighbour-based, clustering based, classification-based, and spectral decomposition-based approaches [3] 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.

In centralized detection, the anomaly detection is performed at the base station. WSNs collect information from the sensor nodes and send it to the base station to be processed and analyzed. The anomaly detection techniques can utilize this information to detect any missing data or data anomalies collected. [7] used an ant colony based intrusion detection mechanism that could keep track of the intruder trails is presented. This technique can work in conjunction with conventional machine learning based intrusion detection techniques to secure sensor networks. This work tracks the paths of intrusion after anomalies are detected. [13] described 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.

In distributed approach, the detection agent is installed in every node. It monitors the behaviour of neighbouring node within its transmission range locally to detect any abnormal behaviour. To perform a real time anomaly detection, some rule based detection techniques are used in a node. Node listens promiscuously to neighbouring nodes within its transmission range to collect data necessary for anomaly detection. The collected data will be analyzed to detect any deviation from normal behaviour using neighbouring historical data stored in the memory. Once anomalies have been detected, an alarm is sends to alert the base station or neighbouring nodes.[4] presented Intrusion Detection Systems (IDS) for a sensor network that is based on the network activities (e.g., number of success and failure of authentications). The system compares event data with signature records to find harmful attacks from an intruder. [5] applied the detection system in a cluster-based sensor network very much like the developed system in this dissertation. This type of detection system can only identify the anomalies that it has seen before. However, this research is interested in detecting anomalies in unknown environments, in which there are no abnormal prototypes available for the system to learn.[6] presented intrusion detection schemes that build a model of normal traffic behaviour, and then use this model of normal traffic to detect abnormal traffic patterns. Their approaches are able to detect attacks that have not been previously seen. [8] presented an outlier detection algorithm based on Bayesian Belief Networks (BBN). The system is also able to estimate missing values in the sensor data. The BBNs are able to capture the relationship between the attributes of sensor nodes as well as spatial temporal correlations that exist among the sensor nodes.[9] developed a framework for the discovery of k-nearest-neighbour 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).

3. Methodology

Finally use these ensemble model as unseen point, integrate all outputs of individual models. Use random subspace ensembles (Subspace) for better accuracy.

3.1 Artificial Neural Network (ANN)

This algorithm is based on non linear algebric classifiers scheme. The class label of a new data is equal to the class of the number of nodes found using specific network formulae. Mean square error and regression is applied for distance measure to get the better network by updating parameters [18]. Here is step described that are used in ANN algorithm:

1. Determine parameter K = number of nodes
2. Calculate the error between the query-instance value and all the training output
3. Sort the network with minimum distance to actual output value
4. Gather the category of the best networks at different nodes
5. Use simple majority of the category of networks as the prediction value of the testing instance

3.2 Load & initialization the variables

a) Each sample has time gap t = 0.5 sec., segment length =50 samples .For 14 sensors data has 14x13=182 sensor pair id and each sensor pair record has 3127 sample after pre-processing of total 3600 sample record 30minutes [19].

b) Sensor data matrix $Z_{182x3127}$ and motion data vector $_{1 x 3127}$ (which is as outlier [0 or 1] dataset) stored in the excel sheet.

c)Take the different distance type as:
distname=['cityblock','chebychev','mahalanobis','minkowski' ,'euclidean','seuclidean','spearman','cosine', 'hamming','jaccard'];

Step 1: Select sensor pair ids:

Pair id is id any two sensor communicate with each other. It is selected randomly to create a training data set.

Step 2: Data segmentation:

a) Dataset is divided into segment of length 50 samples thus total segment are 62 and it is stored in another matrix.

b) Motion dataset which is known as outlier data is also segmented into 62 segments of 50 samples.

c) Each pair is broken into segments such that total segment per sensor pair =floor [(total sample)/(segment length length)]=62.

d) Calculate the entropy and number of outlier of each segment of Z. Number of outlier is named as outlier level.

Step 3: Segment entropy Evaluation:

a) After segmentation the entropy of each segment is evaluated which is input of ANN predictor algo. The function used for entropy calculation is wentropy(dataseg(i,:),'Shannon').

b) Each segment is selected iteratively and entropy is evaluated.

c) Save the entropy of all segment as variable 'Entrpall'.

d) Outlier data is also taken segment wise and outlier are summed and saved as outlier level.

f) Total pairid =182 for segment length =50 we have 62 segment and we take random pair id of 10 pair id thus 620 data recorded are generated and considered as training dataset.

4. Result And Discussion

WSN monitors physical parameters changes in parameters are due to various reasons. Outliers represent unusual readings in sensors due to e.g. sensors fault, a change in some monitored parameter property, obstacle or communication faults in sensors, etc. As a result readouts are different from others in common ambient conditions such that they follow a different distribution. Outlier or anomalies detection specifies abnormal behaviour in data. A basic application of WSN is detection in large areas related to environmental change (temperature, atmospheric pressure or the received signal which are different from which are received in past. Anomaly detection becomes more challenging than conventional detection due to less knowledge of the signal that is to be detected [16],[17]. Various approaches are proposed for outlier detection. This paper follows entropy estimation of data segments. The aim of this analysis is to verify the outlier detection methods by ANN classifier using data entropy as a parameter:

- Estimate PDF of a data by data-split technique
- calculate entropy of the data using PDF
- use entropy as metric to detect outlier by using ANN
- To investigate different distance metrics and numbers of neighbours on ANN classification accuracy.

This method is applied to actual measured data that incorporates literature survey on the use of anomaly detection in sensor networks along with MATLAB simulation on sensor pair data. Validation of ANN technique on the recorded signals found at: www-personal.umich.edu/~kksreddy/rssdata.html. The data is generated under an experiment conducted at the University of Michigan. The 4th level of the EECS office block has the site of the research. Mica2 platform has been used for this experiment, in this fourteen sensor nodes arbitrarily deployed inside and outside a lab space. Broad casting is used for Wireless sensors network applications and the received signal strength (RSS) in terms of voltages called as received signal strength indicator circuit (RSSI).In this work the data is for pair of transmitting and receiving sensor nodes RSS value for a 30 minute period. Total 14 x 13 = 182 sensor pairs of RSS value measurements at sample time of 0.5 sec is collected to give 3191 samples of data. During the data recording the volunteer student‟s walked through the lab at random interval of times. It created anomaly patterns in the values RSSI. A web camera employed to record the walk through activity as ground truth resemblance. Experiment produced the 182 RSS sequence data array to support the model development task of detection of any intruders (anomalies). The original raw data is stored in the matrix of size 182 x 3191.Using webcam records manual record is made as value of 1 to indicate the presence of an intruder.

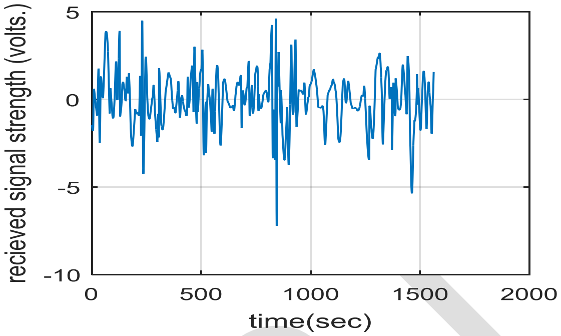

Fig. 1. Received signal strength voltage with respect to time (sec)

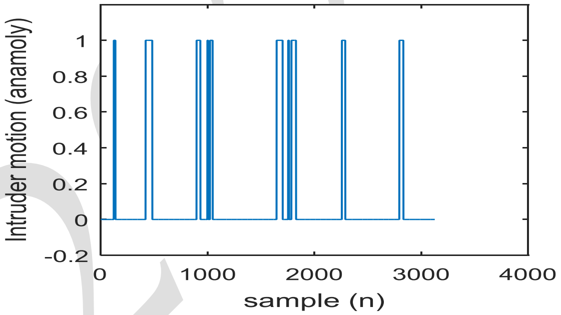

Fig. 2. Intruder motion (anamoly) representation with respect to time samples (n)

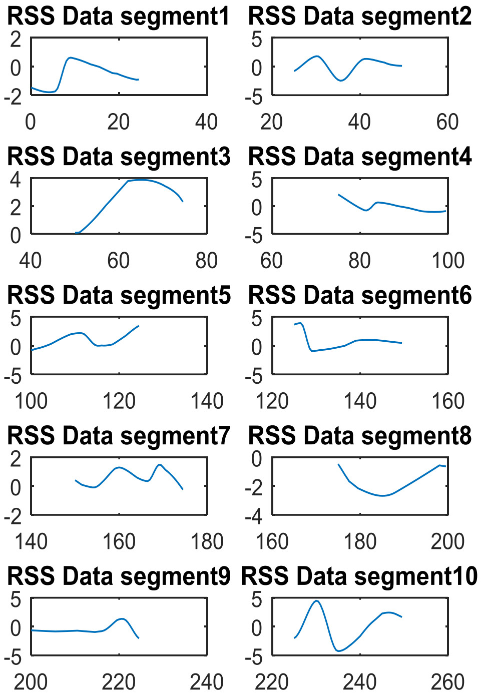

Fig. 3. RSS data segments plot for segment length of 50 samples.

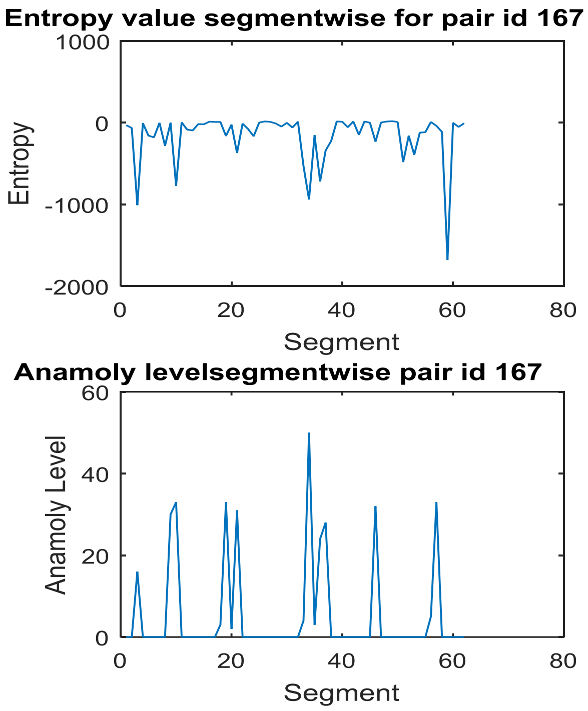

Fig. 4. a) Entropy value calculted for segment of sensor pairid 167, b)Anamoly level with respect to all segment for anonymous sensor pairid 167.

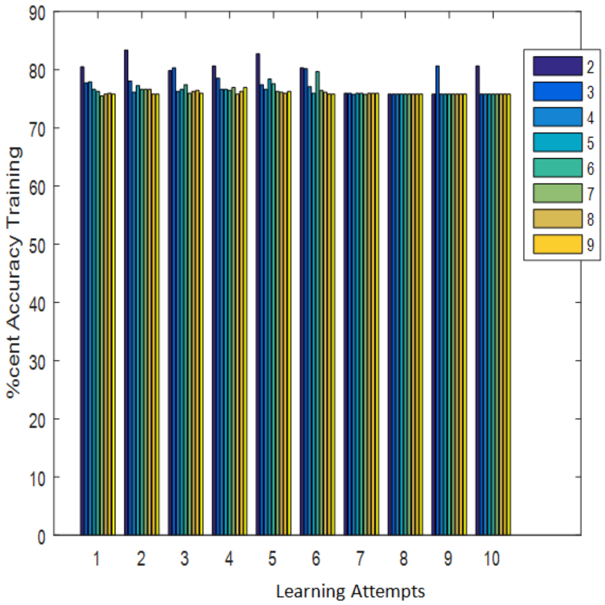

Fig. 5. Percent Training Accuracy at different learning attempts with different number of nodes.

Figure 5 the plot is shown for percent training accuracy for different times of training attempts (1 to 10) at different no of nodes varying from 2 to 9. It can be observed that at attempt "1" the training accuracy is highest at 2 nodes that gradually decrease as no of nodes and finally become constant for large no of nodes. Similar trend in is observed in testing attempts. The highest accuracy observed as 86% at number of nodes equals to 2.In all the cases training accuracy is found to be above than 75%.

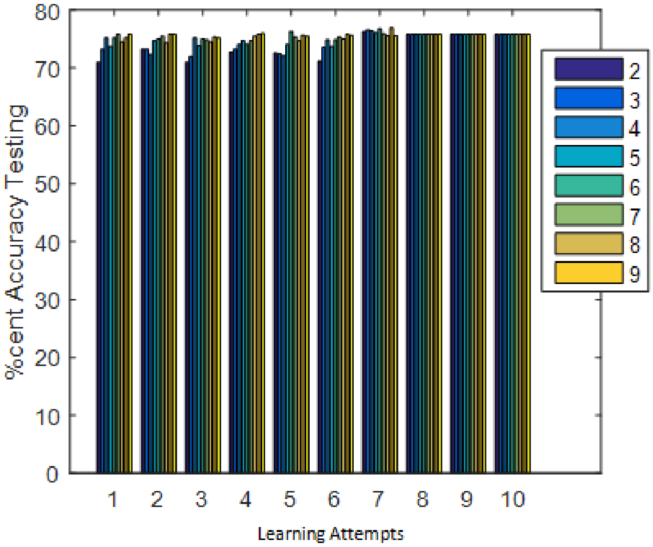

Fig. 6. Percent Testing Accuracy at different learning attempts with different number of nodes.

Figure 6 plot is shown for percent testing accuracy for learning attempts (1 to 10) at different number of nodes varying from 2 to 9. The testing accuracy is highest at 3 nodes and that gradually decrease as no of nodes and finaly become constant for large number of nodes. The highest accuracy observed as 78% at number of nodes equals to 3.In all the cases testing accuracy is found to be above than 70%.

5. Conclusion

In this article the results of evaluation are presented using data sets of actual wireless sensor to validate the acceptability and effectiveness of detection performance using entropy as parameter and ANN as predictor. The development of code is performed on MATLAB 2015a software. Outlier detection is challenging due to the lack of open access real time data. The algorithm discussed in this paper is based on time-series sensor data acquired from a wireless sensor nodes of Mica2 based network of half hour of recording of received signal strength[22]. The datasets has outlier due to intrusion caused by motion of volunteers in lab during recording session. In this way motion injects outliers in the datasets. We calculated entropy of the data segments then trained the ANN using 62 segments of 182 sensor pair records. Each segment length was of 50 samples of 25 second duration.

As a performance metrics training and testing accuracy is calculated. The highest accuracy is 86% is obtained and 78% for testing data. The novelty of this approach is that the number outliers that is level of anomaly is predicted and in other journals related to this field only predicts whether anomaly is present or not. It can be concluded that the ANN approach is faster and simpler than other methods. If proper selection of distance and number of neighbours is performed than using the entropy as a feature the anomaly detection can be performed with low complexity and higher accuracy.

References:

[1] Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. 41*, 15, 2009.
[2] Raja Jurdak, X. Rosalind Wang, Oliver Obst, and Philip Valencia,"Wireless Sensor Network Anomalies: Diagnosis and Detection Strategies", 2011.
[3] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Outlier Detection Techniques for Wireless Sensor Network" A Survey,Technical Report, University of Twente, 2008.
[4] Techateerawat, P. and Jennings, "A Energy efficiency of intrusion detection systems in wireless sensor networks", In Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology (WI- IATW), pages 227–230, Washington, DC, USA. IEEE Computer Society, 2006. [5] Hai, T. H., Khan, F. and nam Huh, E. "Hybrid intrusion detection system for wireless sensor networks", Computational Science and Its Applications ICCSA, 4706:383–396, 2007.
[6] Onat and Miri, Onat, I. and Miri, "An intrusion detection system for wireless sensor networks", In IEEE International Conference on Wireless And Mobile Computing, Networking and Communications, Los Alamitos, IEEE Computer Society Press, 2005.
[7] Banerjee, Banerjee S., Grosan C., and Abraham, "Ideas: intrusion detection based On emotional ants for sensors", In The 5th International Conference on Intelligent Systems Design and Applications (ISDA), pages 344–349, Wroclaw, Poland, 2005.
[8] Janakiram, D., Reddy, V., and Kumar A., "Outlier detection in wireless sensor Networks using Bayesian belief networks". In First International, Conference on Communication System Software and Middleware, pages 1–6, 2006.
[9] Sheng B, Li Q, Mao W, Jin W, "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, 2007. [10]Subramaniam S, Palpanas T, Papadopoulos D, KalogerakiV, Gunopulos D., "Online outlier detection in sensor data using non-parametric models" In Proceedings of ACM conference on very large databases (VLDB06), pp 187–198, 2006.
[11] Janakiram D, Reddy VA, Kumar AVUP, "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, 2006.
[12] Zhuang Y, Chen L "In-network outlier cleaning for data collection in sensor networks", In proceedings of the 1st international VLDB workshop on clean databases (CleanDB06), 2006.
[13] Rajasegarar S, Leckie C, Palaniswami M, Bezdek J, "Distributed anomaly detection in wireless sensor networks", In proceedings of the IEEE Singapore international conference on communication systems, pp 1–5, 2006.
[14] Satish S. Bhojannawar, Chetan M Bulla, Vishal M Danawade, "Anomaly Detection Techniques for Wireless Sensor Networks - A Survey", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, pp.3852-3857, 2013.
[15] R. Sharma, Anurag, " Detect Skin Defects by Modern Image Segmentation Approach, Volume 20, Issue 1, 2020.
[16] Anurag, R. Sharma, " Modern Trends on Image Segmentation for Data Analysis- A Review", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020.
[17] Young Soo Jang et. al., "Development of the cost-effective, miniaturized vein imaging system with enhanced noise reduction", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.6, November – December 2019.
[18] Irma T. Plata1, et. al., "Development and Testing of Embedded System for Smart Detection and Recognition of Witches' Broom Disease on Cassava Plants using Enhanced Viola-Jones and Template Matching Algorithm", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.6, Volume 8, No.5, September - October 2019.
[19] Anurag, R. Sharma, " Load Forecasting by using ANFIS", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020.
[20] R. Sharma, Anurag, " Load Forecasting using ANFIS A Review", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020.