A SUPERVISED TECHNIQUE TO ENSURE DATA RELIABILITY IN WIRELESS SENSOR NETWORKS

N. CHITRADEVI1, V. PALANISAMY2, K. BASKARAN3 AND D. ASWINI4

1Associate Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore
2Principal, INFO Institute of Engineering, Coimbatore
3Assistant Professor, Department of Computer Science and Engineering, Government College of Technology, Coimbatore
4Department of Computer Science and Engineering, PG Scholar, Kumaraguru College of Technology, Coimbatore

Abstract: In Wireless Sensor Network (WSN), sensors at different locations can generate streaming data, which can be analyzed in real-time to identify events of interest. WSN usually have limited energy and transmission capacity, which cannot match the transmission of a large number of data collected by sensor nodes. So, it is necessary to perform in-network data aggregation in the WSN which is performed by aggregator node. Since, the nodes in WSN are vulnerable to malicious attackers and physical impairment; the data collected in WSNs may be unreliable. So, in this paper, we propose an efficient technique to detect the unreliable data. Earlier, A Principal Component Analysis (PCA) is used to compress the data by reducing the number of dimensions. But as a drawback, it is not robust to outliers. Hence, if the input data is corrupted, an arbitrarily wrong representation is obtained. To overcome this problem, we propose a Robust PCA which is augmented with Minimum Covariance Determinant (MCD), highly robust estimator. In the proposed approach the distributed nature of sensor data is modelled using the sound statistical technique PCA and MCD is employed to design a noise-free data model. The performance of proposed approach is evaluated and compared with previous approaches and found that our approach is effective and efficient.

Keywords: WSN, PCA, and MCD.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are comprised of many tiny low-cost battery-powered sensors in a small area. The sensors detect physical or environmental variations and then transmit the detection results to other sensors or a base station. WSNs are resource constraint: limited battery power supply, bandwidth for communication, processing speed, and memory space. One possible way of achieving maximum utilization of power is applying data aggregation on sensor data. Usually, processing data consumes much less power than transmitting data in wireless medium. Hence we go for a mechanism in which the packet size is reduced to considerable level and that leads to data aggregation. So, it is effective to apply data aggregation before transmitting data for reducing power consumption by a sensor node. Although aggregation markedly lowers the traffic between node and the base station, the sensor node is critical and vulnerable to malicious attacks, which affect various operations of the network such as routing, data aggregation, voting, and fair resource allocation in various ways such as sybil attacks or denial of service attacks. So, the data integrity and accuracy problems that may be caused by compromised or malfunctioning nodes are of high research and practical importance. Towards this direction, in this paper, we propose and evaluate an anomaly detection approach that fuses data gathered from different nodes in a distributed WSN and also provides data reliability. In earlier approach, A Principal Component Analysis (PCA) is used to compress the data by reducing the number of dimensions. But as a drawback, it is not robust to outliers. Hence, if the input data is corrupted, an arbitrarily wrong representation is obtained. To overcome this problem, we propose a Robust PCA which is augmented with Minimum Covariance Determinant (MCD), which removes the outliers. Thus the data reliability and accuracy is achieved. One of the key features of the proposed approach is that it provides an integrated methodology for effectively combining correlated sensor data, in a distributed fashion, in order to reveal anomalies that span through a number of neighboring sensors. Furthermore, it allows the integration of results from neighboring
network regions to detect correlated anomalies/attacks that involve multiple groups of sensors. Such an approach can be used in principle to identify an abnormal situation in measurements (e.g., cases where the values of the measured or monitored parameters may deviate significantly from the normal) discover the existence of faulty sensors, detect potential network attacks, and filter suspicious reports throughout the overall decision making process.

The remainder of this paper is organized as follows. In Section 2, we present some related work and in Section 3, we present the system model and the corresponding architecture used throughout this paper. In Section 4, the proposed model process is presented in detail. The corresponding performance of our proposed approach is evaluated in Section 5, while Section 6 concludes this paper.

II. RELATED WORK

In [9]–[10], the authors presented several attack scenarios that exploit the weaknesses of WSNs. The scale of deployments of WSNs requires careful decisions and tradeoffs among various security measures. The authors discussed these issues and considered mechanisms to achieve a higher level of security and reliability in these networks. In [11], the authors presented a statistical en-route filtering mechanism to detect and drop false reports during the forwarding process. Assuming that an event may be detected by multiple sensors, each of the detecting sensors generates a keyed message authentication code (MAC), and multiple MACs are attached to the event report. In [12], the problem of faulty or malicious nodes is formalized as how to construct a dominating tree to cover all the neighbors of the suspect and give the lower bound of the message complexity. Earlier related work reported in the literature has focused on detecting deviations in data patterns among the sensors. In [13], the authors presented a framework for random sampling mechanisms and interactive proofs to check that the values returned by aggregators are good approximations of the true values, even when the aggregators and a fraction of the sensor nodes are corrupted. However, the described tests serve only as a proof of concept, as they consist of simple algorithms such as finding the minimum, maximum, and median values. In [14], a spatial-temporal correlation analysis called “abnormal relationships test” (ART) is proposed, to detect outliers in the collected data. This method is based on correlation coefficient tests between neighboring nodes. In [15], the authors describe a technique for online identification of outliers in readings collected by individual wireless sensors, and attempt to extend this technique to an entire network of sensors, taking into consideration the distributed processing of events. However, that technique requires the complete knowledge of the density distribution function of the collected data. In [8], the approach focused on the efficient detection of outliers throughout a sensor network in a distributed manner, and is based on the use of PCA [4].

But the above works are not robust to outliers. Our approach makes PCA as Robust PCA by combining MCD with PCA, where MCD is a robust estimator of location and scatter. The goal of PCA is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. When an anomaly occurs, the system detects the path containing the anomaly and necessary action is taken. The MCD algorithm [3] uses several time-saving techniques which make it available as a routine tool to analyze data sets with large n, and to detect deviating substructures in them. The proposed approach is useful in several fields like, online event detection, intrusion detection and network anomaly detection.

III. SYSTEM ARCHITECTURE

We envision a sensor network paradigm with several heterogeneous sensor nodes, where each node may have different capabilities and execute different functions. A sensor network is usually represented by a network graph. An algorithm that correlates metrics from neighboring sensors is considered, to detect the node(s) containing anomalies in the corresponding network graph. In order to decentralize the detection algorithm, the sensor network is divided into groups of sensors. The division may be done either statically when the network is originally deployed, or the network may be dynamically rearranged periodically, if the environment changes. In any case, we consider that the division of the network into subgroups of nodes is based on correlation tests among the nodes.

Figure 1 presents the sensor network topology and architecture under consideration. The creation of groups is based on the interrelation in the corresponding readings of the sensors. In every group we assume that there is a primary node, which usually
is equipped with more processing capability, power, sophistication, and intelligence. As the data in the neighbouring nodes are considered highly correlated, the data processing and aggregation might utilize the spatial correlation of local nodes and dramatically decrease the amount of information to be transmitted. Estimation of the data correlation properties based on the previous and current collected data, can be used to locally optimize the data compression and aggregation in the subsequent data gathering phases [2]. The outcome of the grouping procedure is that the groups consist of nodes with interrelated readings. It should be also noted here that the various groups do not need to have mutually exclusive members as shown in Fig. 2, the existence of a number of common secondary nodes is actually desired in many cases in order to improve the detection effectiveness. Each primary node obtains sensor readings from the nodes in its group and may perform localized real-time analysis. In general, every network node collects data with reference to one or more metrics that describe the specific parameters that the node monitors.

IV. PROPOSED MODEL

The objective of this paper is to provide an efficient and effective methodology of fusing and combining data of heterogeneous monitors that spread through out the network, in order to provide a generalized framework, capable of detecting a wide range of classes of anomalies, such as the ones created randomly by faulty nodes or others that result from coordinated compromised nodes. It provides data reliability. In our work, this is achieved by using PCA with MCD.

(A) Proposed Scheme

The proposed scheme involves two stages of operation, as shown in Figure 3: the offline analysis that creates a model of the normal pattern of the monitored parameters, and the real-time analysis that detects anomalies by comparing the current with the modeled ones. The input of the offline analysis is the correlation matrix of a sampled data set. During the offline analysis, MCD is applied on the data set, where the outliers are removed and output is given to PCA, and then the first few most important derived principal components (PCs) are selected. The number of the selected PCs depends on the sensor network and the number of virtual nodes, and it represents the number of PCs required for capturing the percentage of variance that the system needs to model its normal status. The output of the offline analysis is the PCs to be used in the next stage. Since this procedure is computationally heavy, it must be carried out only when there is a significant change in one or more of the correlation coefficients. A feasible solution is to use a sliding window containing the last readings and re-estimate the PCs only when the deviation in one or more correlation coefficients exceeds a threshold. Many multivariate techniques applicable to anomaly detection problems are based upon the concept of distance. The Mahalanobis Distance (MD) is a well known multivariate distance metric defined as the distance of a vector from the centroid in the multidimensional space, defined by the correlated independent variables. In general, anomalies tend to result in great variations in the residual, since they
present different characteristics. When there are only a few multivariate outliers, examine each individually to determine how they differ from the centroid. When there are many outliers we want to describe how the outliers as a group differ from the centroid – just compare the means of the outliers with the means of the total data set. When an anomaly occurs, the residual vector presents great variation in some of its variables and the system detects the path containing the anomaly by selecting these variables.

\[ Y = Y_{\text{norm}} + Y_{\text{res}} \]  

Such \( Y_{\text{norm}} \) corresponds to modeled (normal) data and \( Y_{\text{res}} \) to the residual. We form \( Y_{\text{norm}} \) by projecting \( y \) onto the normal subspace \( S \), and we form \( Y_{\text{res}} \) by projecting onto the abnormal subspace \( S \). To accomplish this, we arrange the set of PCs corresponding to the normal subspace \((v_1, v_2, \ldots, v_r)\) as columns of a matrix \( P \) of size \( p \times r \), where \( r \) denotes the number of normal axes. Following this approach, \( Y_{\text{norm}} \) and \( Y_{\text{res}} \) may be rewritten as follows:

\[ Y_{\text{norm}} = PP^T y = Cy \] and \[ Y_{\text{res}} = (I - PP^T)y - C \]

where matrix \( C = PP^T \) represents the linear operator that performs projection onto the normal subspace \( S \), and \( C \) likewise projects onto the anomaly subspace \( S \). Thus, \( Y_{\text{norm}} \) contains the modeled (normal) data, while \( Y_{\text{res}} \) contains the residual. In general, the occurrence of an anomaly tends to result in a large change to \( Y_{\text{res}} \). A change in variable correlation will increase the projection of \( y \) to the subspace \( S \). Within such a framework a typical statistic for detecting abnormal conditions is the squared prediction error (SPE) [7].

\[ \text{SPE} = ||Y_{\text{res}}||^2 = ||e||^2 \]  

When an anomaly occurs, the SPE exceeds the normal thresholds and the system detects the set of sensors containing the anomaly, by selecting the variables that contribute mostly to the large change of the SPE. This may be realized by selecting the virtual nodes in the residual vector whose variation is significantly larger than the corresponding one under normal conditions.

**(C) Minimum Covariance Determinant**

The MCD is a robust method in the sense that the estimates are not unduly influenced by outliers in the data, even if there are many outliers. Due to the MCD’s robustness, we can detect outliers by their large robust distances. The latter are defined like the usual Mahalanobis distance, but based on the MCD location estimate and scatter matrix (instead of the nonrobust sample mean and covariance matrix). The FASTMCD algorithm [3] uses several time-saving techniques which make it available as a routine tool to analyze data sets with large \( n \), and to detect deviating substructures in them. An important feature of the FASTMCD algorithm is that it allows for exact fit situations, i.e. when more than \( h \) observations lie on a (hyper) plane. Then the program still yields the MCD location and scatter matrix, the latter being singular,
as well as the equation of the hyper plane. The MCD objective is to find h observations (out of n) whose classical covariance matrix has the lowest determinant. The MCD estimator of location is then the average of those h points and the MCD estimate of scatter is their covariance matrix. The minimum (and default) h = (n + nvariables + 1)/2 so the algorithm is effective when less than (n + nvar + 1) / 2 variables are outliers. The Fast MCD core idea is as follows:

1. The default h is [(n+p+1)/2], but the user may choose any integer h with [(n+p+1)/2] ≥ h ≤ n.
2. If h = n then the MCD location estimate \( T \) is the average of the whole data set, and MCD scatter estimate \( S \) is its covariance matrix. Report these and stop.
3. If p = 1 (univariate data) compute the MCD estimate \((T, S)\) by the exact algorithm of Rousseeuw and Leory.
4. From here on, h < n and p ≥ 2. If n is small then Repeat (say) m times:
   - Construct an initial h-subset \( H_1 \) starting from a random \((p + 1)\) subset.
   - Carry out two C-steps [3].
   For the 10 results with lowest det \( (S) \):
   - Carry out C-steps until convergence
   Report the solution \((T, S)\) with lowest det\( (S) \)
5. If n is larger then
   Construct up to five disjoint random subsets of size \( n_{sub} \).
   Inside each subset, repeat \( m/k \) times:
   - Construct an initial subset \( H_i \) of size \( h_{sub} = [(h/n)] \)
   - Carry out two C-Steps, using \( n_{sub} \) and \( h_{sub} \)
   - Keep the 10 best results \((T_{sub}, S_{sub})\)
   Pool the subsets, yielding the merged set
   In the merged set, repeat for each of the 50 solutions \((T_{sub}, S_{sub})\)
   - Carry out two C-setps, using \( n_{merged} \) and \( h_{merged} = [n_{merged} (h/n)] \)
   - Keep the 10 best results \((T_{merged}, S_{merged})\)
   In the full data set, repeat for the best results
   - Take several C-steps, using n and h
   - Keep the best final result \((T_{full}, S_{full})\)

Thus MCD based PCA is a robust version of multivariate analysis.

V. PERFORMANCE EVALUATION

This section evaluates the performance of the proposed mechanism numerically by considering the Detection Rate (DR) and False Alarm Rate (FAR) as performance metrics.

The performance is monitored for varied percentage of outliers found in the data. RPCA uses Principal components (PCs) to compress the data. The selection of PC is explained in Figure 4 & 5.

Figure 4 helps to know about the importance of different Principal Components. From the figure we can conclude the component 1 is the vital one. Figure 5 shows the relationship between Eigen values and Principal components. The PC having Eigen value >=1 will be considered as important component. In our paper, we consider the cumulative percent variance captured by the first few (r) PCs that accounts for greater than 80% of the variability of the data as shown in Figure 6.
The relation between the outliers and detection rate is shown in Figure 7. Detection rate is the ratio between the number of correctly detected anomalies and the total number of anomalies. From the graph, we can find that proposed approach provide 100 percent detection rate until 40% of nodes connected in network are found to be faulty or compromised while existing approach shows an inclined detection. The performance of both approaches deteriorates when outliers dominates normal data. Even in such cases it could be found that RPCA provides a detection rate of 0.6 while the detection rate drastically reduces to 0.3 for the classical PCA.

Figure 8 shows the relation between Outlier and the False Alarm Rate. FAR is the ratio of the number of data records from normal class that are misclassified as anomalies to the total number of data records from normal class. From the graph, we could infer that FAR of proposed approach is always greater than the classical PCA. Thus it is observed that our RPCA is efficient though level of contamination increases to half of the network size.

The performance of proposed method is analyzed with Receiver Operating Characteristics (ROC) curve, where the quality goal is to maximize the detection rate and minimize the false alarm rate (false positives). The ROC comparison was done for both classical and robust PCA approach for different scenarios involving single and multiple faulty nodes where magnitude of data alteration varies between .05 to 90 percent of original data. Figure 9 clearly shows that the proposed robust PCA approach is effective in detecting the inconsistent data as the detection rate of proposed method is consistently higher than classical PCA method and it was found that the proposed method offers an average detection rate of 90.9% in comparison to existing which offers a detection rate of 70.5% when the network contains 30 percent contamination.

Figure 10 shows the detection rate and false alarm rate of proposed approach for different data alteration rate in varied outlier percentage for a fixed cluster size.
From the graph, we can find that the proposed approach provides optimistic (100%) detection when there is data contamination or signal to noise ratio is high. When the data is altered in low magnitude the system fails to detect and offers a low detection rate. From the graph it is evident that the proposed approach supports an average detection rate of 72.5% and false alarm rate of 40.5% even when nearly half of network size (40%) fails.

VI. CONCLUSION

In this paper, we proposed an anomaly detection approach that aggregates data gathered from different nodes in a distributed WSN. We proposed a robust PCA which is augmented with MCD. The proposed robust procedure localizes the inconsistent data within data used for model construction and isolates it. Thus it improves the model prediction accuracy thereby achieves data reliability. We compared the performance of our approach with the classical PCA. Experimental research shows that our approach gives better performance for contaminated data even when nearly 50% of nodes involve in inserting false data, thus the result is more accurate and reliable.

References


