Hierarchical Anomaly Detection in Distributed Large-scale Sensor Networks

V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, B. Maglaris
Network Management & Optimal Design Laboratory (NETMODE),
School of Electrical & Computer Engineering
National Technical University of Athens (NTUA)
9 Iroon Polytechniou str. Zografou 157 73, Athens, Greece

Abstract

In this paper, an anomaly detection approach that fuses data gathered from different nodes in a distributed wireless sensor network is proposed and evaluated. The emphasis of this work is placed on the data integrity and accuracy problem caused by compromised or malfunctioning nodes. One of the key features of the proposed approach is that it provides an integrated methodology of taking into consideration and combining effectively 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 areas to detect correlated anomalies/attacks that involve multiple groups of nodes. The efficiency and effectiveness of the proposed approach is demonstrated for a real use case that utilizes meteorological data collected from a distributed set of sensor nodes.

1. Introduction

By integrating sensing, signal processing, and communications functions, a sensor network provides a natural platform for hierarchical and efficient information processing. It allows information to be processed at different levels of abstraction, ranging from detailed microscopic examination of specific targets to a macroscopic view of the aggregate behavior of targets. Usually the sensors are used to measure and/or monitor some parameters that may vary with place and time. Therefore a large number of sensors are required in order to obtain samples of these parameters at different locations and times. Furthermore, these sensors are networked in order to facilitate the transmission/dissemination of the measured/monitored parameters to some collector sites where the information is further processed for decision making purposes [1].

Unlike traditional wireless networks, in which the communication is person-to-person and the contents of conversations are irrelative to each other, in sensor networks, the data in the neighboring nodes are considered highly correlated since the observed objects in physical world are highly correlated as well [2] . Due to the critical nature of several applications of sensor networks, data integrity and accuracy problems that may be caused by compromised or malfunctioning nodes are of high research and practical importance [3].

Towards that direction, in this paper, we propose and evaluate an anomaly detection approach that fuses data gathered from different nodes in a distributed wireless sensor network. One of the key features of the proposed approach is that it provides an integrated methodology of taking into consideration and combining effectively 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 areas to detect correlated anomalies/attacks that involve multiple groups of nodes.

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 norm) discover the existence of faulty sensors, detect potential network attacks, and filter suspicious reports throughout the overall decision making process.

The remaining of the paper is organized as follows. In section 2 we present some relevant background information and related work. The proposed data fusion and anomaly detection technique is introduced and described in detail in section 3, while in section 4 we present some numerical results regarding the performance and operational effectiveness of our proposed anomaly detection approach for a real use case that utilizes meteorological data collected from a distributed set of sensor nodes.

2. Background Information

In monitoring sensor networks, data coming from many different streams of the sensor nodes have to be examined dynamically and combined into normal

Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC'06)
0-7695-2588-1/06 $20.00 © 2006 IEEE

patterns in order to detect potential anomalies. At the same time, to achieve cost-effectiveness and small size, in general the individual sensor nodes present several limitations, such as limited energy and memory resources, communication bandwidth and processing capabilities. Therefore, to minimize the processing and communication overhead of the sensors one must process as much of the data as possible in a decentralized fashion, so as to avoid unnecessary communication and computation effort.

Earlier related work reported in the literature has focused on detecting deviations in data patterns among the sensors. In [4] the authors proposed a sliding-window based spatio-temporal correlation analysis called "Abnormal Relationships Test (ART)" to detect outliers in the collected data. In [5] the authors described a technique for online identification of outliers in readings collected by individual wireless sensors, and attempted to extend this technique to an entire network of sensors, taking into consideration the distributed processing of events. Our approach focuses on the efficient detection of outliers throughout a sensor network in a distributed manner, and is based on the use of Principal Component Analysis (PCA) [6]. PCA has been shown to provide very efficient ways of modeling the spatio-temporal data correlations, and its basic principles have been used for anomaly detection purposes in several fields. For instance, in [7] and [8], techniques based on the use of PCA have been proposed for intrusion detection and network traffic anomaly detection, respectively.

3. Data Fusion and Anomaly Detection Approach

3.1. System Model and Architecture

We envision a sensor network paradigm with several heterogeneous sensor nodes, where each node may have different capabilities and execute different functions. For example, some nodes may have larger battery capacity and more powerful processing capability, some nodes may aggregate and relay data, while some others may only execute the sensing function and do not relay data for other nodes.

A topology-aware 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 we divide the sensor network into groups. The division may be done either statically when the network is deployed, or the network may be dynamically rearranged periodically, if the environment changes. In any case, we assume that the division of the network into subgroups of nodes is based on the correlation coefficient tests among the nodes. The correlation coefficient $R_{X,Y}$ between two data sets X and Y is given by:

$$R_{X,Y} = \frac{Cov(X,Y)}{S_X \cdot S_Y} \quad (1)$$

where Cov(X,Y) denotes the covariance between data sets X and Y, while S_X and S_Y are sample variances of X and Y respectively.

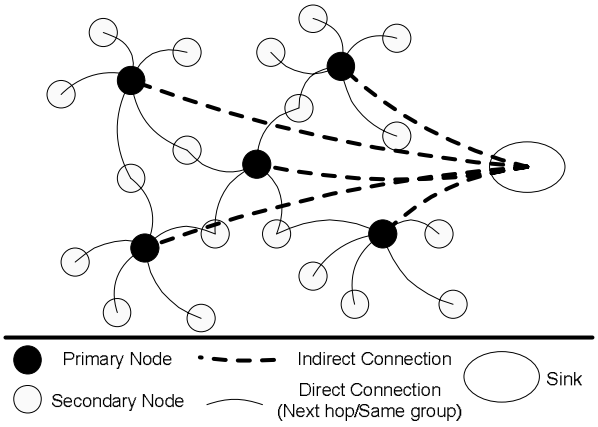

Figure 1: Sensor Network Topology and Architecture

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 may actually be more powerful and sophisticated. Neighboring nodes are expected to read data that are correlated. The primary node creates a group by querying all the nodes that are visible, for their recent readings. All nodes whose readings on the specific queried metric are above a predefined threshold enter the group. The outcome of this procedure is that the groups consist of nodes with interrelated readings. It should be noted that the various groups do not need to have mutually exclusive members, and therefore it is possible that a variant number of common secondary nodes may exist. Testing node readings in more than one group increases the probability of detecting an anomaly.

Each primary node obtains sensor readings from the nodes in its group and performs localized real time analysis, as described in detail later in this section. In general every network node collects data with reference to one or more metrics that describe the

environmental parameters that the node monitors. If the metrics of nodes are correlated, then the procedure may be expanded to take into consideration this correlation. In order to better model and represent this, we create a set of virtual nodes for every real node, with each virtual node corresponding to a different metric.

The aggregated values and information about the anomalies are forwarded to the sink. Information about the anomalies may include the number and identification of nodes that reported faulty readings and the deviation of readings compared to their standard deviation. The sink knowing the topology of the network may decide whether the anomaly spreads within multiple groups, and whether it follows one or more network paths. This process could provide useful information for a large number of applications. For example, there may be an extreme natural phenomenon that affects the sensor readings in an area, or there could be a virus or a malicious node moving and compromising or affecting a large number of dispersed nodes. In either case the group(s) that reported greater deviations, indicate the path of the detected anomalies as well as the current position of the source of the anomaly.

3.2. Anomaly Detection Process

The anomaly detection procedure may be divided into two different parts, as displayed in figure 2: the offline analysis, that creates a model of the normal condition of the monitored parameters, and the real time analysis that detects anomalies by comparing the current (actual) with the modeled one. The input of the offline analysis is the correlation matrix (the diagonal matrix containing the correlation coefficients of all the monitored metrics) of a sampled data set. During the offline analysis, PCA is applied on this data set and then the first few most important derived Principal Components (PCs) are selected. The number of the selected PCs depends on the 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 Subspace Method [10]. 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.

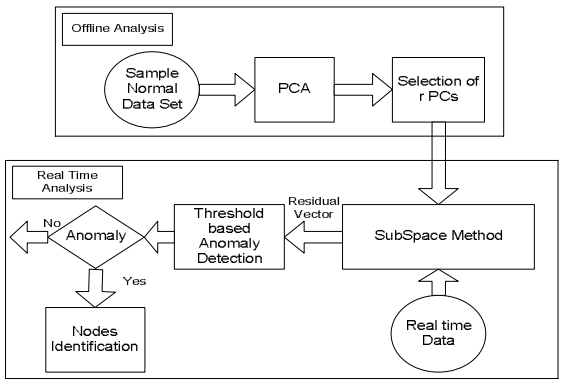

Figure 2: High-Level Methodology Representation

The goal of the Subspace Method is to divide current data in two different spaces: one containing readings that are considered normal and resemble to the modeled data patterns and one containing the residual. In general, anomalies tend to result in great variations in the residual, since they present different characteristics. During the real time analysis, the current data vector is projected into two different subspaces, with the use of the PCs estimated in the offline analysis (Subspace Method). 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. In the following subsections we provide a detailed description of each one of the components involved in this overall approach.

3.3. Combining Performance Metrics

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 [6],[9]. The extracted non-correlated components are called Principal Components (PCs) and are estimated from the eigenvectors of the covariance matrix of the original variables.

Let the original data $\mathbf{x}$ be an n×p data matrix of n observations on each of p variables (x_1, x_2,…, x_p), and let $\mathbf{S}$ be a p×p sample covariance matrix of x_1, x_2, …,x_p. If (λ_1, $\mathbf{e_1}$), (λ_2, $\mathbf{e_2}$),…, (λ_p, $\mathbf{e_p}$) are the p (eigenvalue, eigenvector) pairs of matrix $\mathbf{S}$, then the i-th PC is given by:

$$\mathbf{z_i} = \mathbf{e_i}^T(\mathbf{x}-\mathbf{x_m}) \quad (2)$$

Where $\lambda 1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, $\mathbf{e_i}^T$ is the i-th transposed eigenvector, and $\mathbf{x_m}$ is the mean of $\mathbf{x}$.

If k is the number of nodes and l is the number of different metrics collected in each node, then the virtual nodes that represent the variables of our data set x, are p=l × k. Each row of this matrix is an observation of the monitored network in a specific time bin, whereas each column contains the time series for each virtual node. The output of PCA is a p×p matrix of PCs. The next step, within the overall process, is to select the number r (r<p) of PCs required to capture the percentage of variance that the system needs to model normal data.

Traditionally, PCA is based on the extraction of the eigenvectors from the covariance matrix of a sample data set. Our method achieves to overcome the problem of different scales by utilizing the correlation matrix instead of the covariance matrix. The problem of using PCA based on covariance matrices is that PCs are sensitive to large differences between the variances of the elements of x. Therefore in order to alleviate the problem of scale dependence of PCA, we used standardization of the virtual node variables. Each Principal Component can also be defined as:

$$\mathbf{z_i} = \mathbf{es_i}^T \mathbf{x}^* \quad (3)$$

where $\mathbf{es_i}^T$ is the i-th transposed eigenvector of the correlation matrix, and $\mathbf{x}^*$ consists of standardized variables. The goal in adopting such an approach is to find the principal components of a standardized version $\mathbf{x}^*$ of $\mathbf{x}$, where $\mathbf{x}^*$ has j-th element $x_j/\sigma_{jj}^{1/2}$, j=1,2…p, x_j is the j-th element of $\mathbf{x}$, and σ_{jj} is the variance of x_j. Then the covariance matrix of $\mathbf{x}^*$ is the correlation matrix of $\mathbf{x}$, and the PCs of $\mathbf{x}^*$ can be extracted based on expression (2).

One of the main tasks in all PCA-based anomaly detection approaches is the choice of the number of PCs required to capture the percentage of variance desired. In our case, we need to determine the most suitable value of the number r of the PCs required for the application of the subspace method. One of the most common criterions for choosing r is the cumulative percentage of total variation [6]. An alternative rule, which is specific to the use of correlation matrices as in our case, is based on the size of variances of PCs. The main idea behind this rule is that if all elements of $\mathbf{x}$ are independent, then the PCs are the same as the original variables and they should all have variances equal to 1. Therefore, any PC with variance less than 1 contains less information than any of the original variables, and as a result it is not worth retaining. For instance, if the data set contains a group of variables with large within-group correlation, then there will be only one PC associated with this group whose variance is greater than 1. Thus, the rule will generally retain only one PC associated with that group. This criterion, that in its simplest form is called Kaiser's rule [6], is also chosen to be used in our environment, while extensive experiments have demonstrated its suitability and applicability.

3.4. Subspace-based Anomaly Detection

After having acquired the PCs and determined the number of PCs that will be retained, a normalized sample vector can be decomposed into two portions, as follows:

$$y = ynorm + yres \quad (4)$$

such that ynorm corresponds to modeled (normal) data and yres to the residual. We form ynorm by projecting y onto the normal subspace S, and we form yres by projecting y onto the abnormal subspace ˜S. To accomplish this, we arrange the set of principal components corresponding to the normal subspace (v1, v2, ..., vr) as columns of a matrix P of size p×r where r denotes the number of normal axes. Following this approach ynorm and yres can be rewritten as follows:

$$ynorm = PPT\, y = Cy \text{ and } yres = (I - PPT)y = \tilde{C}y \quad (5)$$

where matrix C = PPT represents the linear operator that performs projection onto the normal subspace S, and ˜C likewise projects onto the anomaly subspace ˜S. Thus, ynorm contains the modeled (normal) data while yres contains the residual. In general, the occurrence of an anomaly tends to result in a large change to yres. 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) [11]:

$$SPE \equiv \|y_{res}\|^2 = \left\|\tilde{C}\, y\right\|^2 \quad (6)$$

When an anomaly occurs the SPE exceeds the normal thresholds and the system detects the set of nodes 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.

4. Performance Evaluation

In this section the performance and operational effectiveness of our proposed anomaly detection approach is evaluated, for a real use case that utilizes meteorological data collected from a distributed set of sensor nodes. The data contain meteorological readings such as wind speed, air temperature, dew point

temperature and humidity from various neighboring ground stations in the island of Crete in Greece. In order to better demonstrate the improvements that can be achieved by the proposed methodology, we also present some comparative results of our approach against the corresponding ones that could be achieved by another correlation-based methodology that has been presented in the literature, namely the Abnormal Relationships Test (ART) [4] .

4.1. Numerical Results

In order to better evaluate the effectiveness of the anomaly detection algorithm three performance metrics of interest were utilized: detection probability Pd, false alarm probability Pf and miss probability Pm. Here the detection probability is defined as the probability that the abnormal data is being detected and recognized. The false alarm probability is defined as the probability that the normal data are being classified as anomalies, while the miss probability is defined as the probability that anomaly data is failed to be recognized. A successful anomaly detection algorithm should achieve high Pd, low Pf and low Pm. Since Pd+Pm=1, we usually evaluate the detection algorithm performance by the detection probability and the false alarm probability. In this paper, we use the Receiver Operating Characteristic curve to visualize the trade-off between the detection and false alarms probability.

The analysis presented in this section is based on an extensive set of real collected temperature readings, where anomalies were inserted randomly in the corresponding data set. The readings are in time bins of one hour during a single day. The network was divided into groups of sensors reporting readings with correlation larger than 90%, based on the correlation test described in subsection 3.1. The division resulted into four groups with about ten nodes per group. The anomalies were inserted randomly in one or more nodes per group each time, and their magnitude varied from 4% to 12% of the original value.

The corresponding numerical results are depicted in the ROC curves shown in the following Figures 3, 4 and 5. Specifically, Figure 3 displays the detection probability vs. the false alarm probability for two different cases: local detection (L.D.) versus global detection (G.D.), of an anomaly that occurs within a specific group. The two different cases refer to two different ways of applying our proposed anomaly detection approach. The global detection refers to the centralized application of our proposed anomaly detection algorithm directly on the total number of sensor nodes (i.e. without group division), while the local detection refers to the decentralized application of our approach, separately for each specific group of nodes, as described earlier in the paper. In each case, we estimate the achievable effectiveness for different anomaly magnitudes. For each curve the point at the upper left corner represents the optimal detection, with high detection probability and low false alarm probability. From this figure we confirm, that the detection performance improves significantly by the localized/distributed application of our approach.

Figure 4 displays some comparative numerical results for both the proposed approach and the ART approach, in a group of correlated nodes. The anomaly is randomly generated in one node of the group each time. Again, in each case, we estimate the effectiveness for different anomaly magnitudes. As observed from this figure, ART approach fails to detect the anomaly, unless it becomes significantly large, therefore limiting its application and effectiveness only to cases that a large number of nodes reports faulty values.

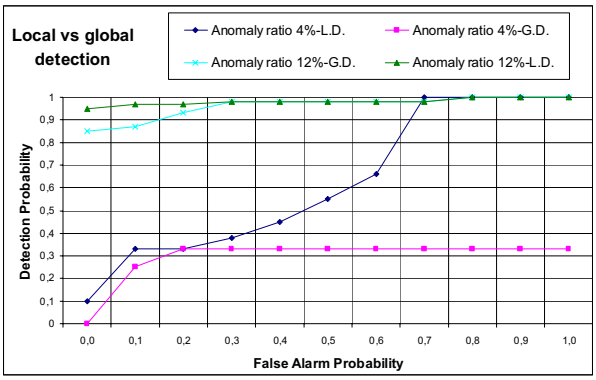

Figure 3: ROC curves for local vs. global detection

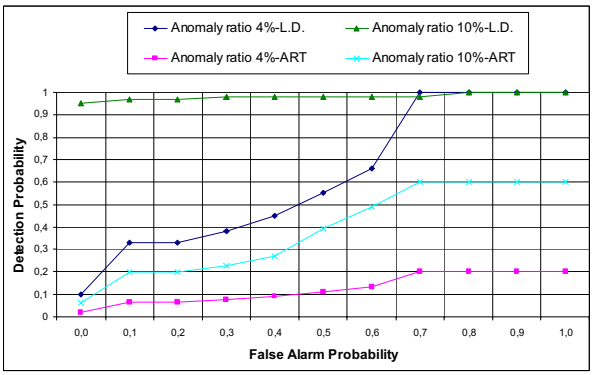

Figure 4: Comparative ROC curves for the proposed approach and ART approach

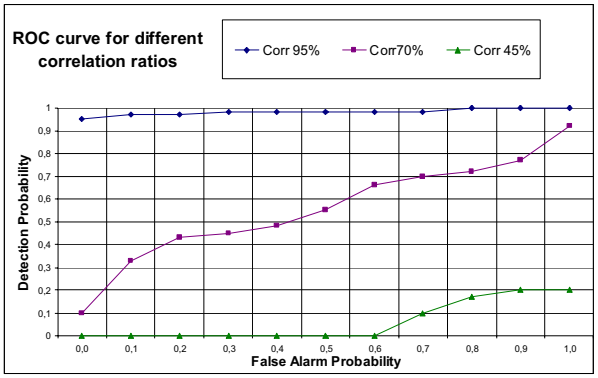

Figure 5: Correlation impact on detection

Finally in figure 5 we present the impact and importance of the correlation among nodes of the same group. Larger correlation among the nodes of the same group may require the creation of more and smaller groups in the sensor network and therefore requires a larger number of primary nodes, however increases the effectiveness and accuracy of the overall detection process.

10. References

[1] Symeon Papavassiliou and Jin Zhu, "Architecture and Modeling of Dynamic Wireless Sensor Networks", Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, CRC Press, 2004, pp. 15.1-15.13.

[2] Mehmet C. Vuran ,B. Akan, Ian F. Akyildiz, "Spatio-temporal correlation: theory and applications for wireless sensor networks", Computer Networks: The International Journal of Computer and Telecommunications Networking,Vol. 45, Issue 3, pp. 245 - 259 , 2004

[3] J. Newsome, E. Shi, D. Song, A. Perrig, "The SybilAttack in Sensor Networks: Analysis & Defenses", in Proceedings of the third international symposium on Information processing in sensor networks, pp. 259 - 268 ,2004

[4] Sapon Tanachaiwiwat and Ahmed Helmy, "Correlation analysis for alleviating effects of inserted data in wireless sensor networks", in Proceedings of Mobile and Ubiquitous Systems: Networking and Services, pp. 97- 108, 2005.

[5] T. Palpamas, et al, "Distributed Deviation Detection inSensor Networks," in Proceedings of ACM SIGMOD, Vol.32 Issue 4, pp. 77 - 82 ,2003

[6] I.T. Jollife. "Principal Component Analysis", Second Edition, Springer, 2002

[7] M. Oka, Y. Oyama, H. Abe, and K. Kato, "Anomaly Detection Using Layered Networks Based on Eigen Co-occurrence Matrix", in Proceedings of the Seventh International Symposium on Recent Advances in Intrusion Detection (RAID),pp. 223-237, 2004.

[8] A.Lakhina, M.Crovella, C.Diot. "Diagnosing network-wide traffic anomalies", in Proceedings of the conference on applications, technologies, architectures, and protocols for computer communications (SIGCOMM), pp. 219 - 230, 2004

[9] J. Edward Jackson. "A user's Guide to Principal Components", Wiley, 2003

[10] R. Dunia and S. J. Qin. "A Subspace Approach to Multidimensional Fault Identification and Reconstruction", American Institute of Chemical Engineers (AIChE) Journal, pp 1813–1831, 1998.

[11] J. E. Jackson and G. S. Mudholkar, "Control Procedures for Residuals Associated with Principal Component Analysis", Technometrics, pp. 341–349, 1979.

[12] Bhaskar Krishnamachari,Deborah Estrin,Stephen B. Wicker, "The Impact of Data Aggregation in Wireless Sensor Networks", in Proceedings of the 22nd International Conference on Distributed Computing Systems, pp. 575 - 578 , 2002