ROBUST FACTOR BASED ANOMALY DETECTION IN HIERARCHICAL WIRELESS SENSOR NETWORKS

Barakkath Nisha U. [1], UmaMaheswari N. [2], Venkatesh R. [3], Yasir Abdullah R. [4]

[1, 2, 3] PSNA College of Engg& Technology, Department of Computer Science
[4] Sri Subramaniya Colleges of Engg& Technology, Department of Computer Science

ABSTRACT: To improve data reliability, accuracy and to make effective and correct decisions using data collected from the wireless sensor network, it is necessary to detect the inconsistent data (outlier) caused by compromised or malfunctioning nodes. Data aggregation is augmented to eliminate the outlier data in the sensor network by multivariate analysis technique such as factor analysis and mahalanobis distance. Factor analysis is a way to fit a model to multivariate data to estimate the interdependence. In a factor analysis model, the measured variables depend on a smaller number of unobserved (latent) factors. The mahalanobis distance is used to determine the similarity of a set of values from an unknown samp1e to a set of values measured from a collection of known samples. Combined with factor analysis, Mahalanobis distance is extended to examine whether a given vector is an outlier from a model identified by factors based on factor analysis. In this paper to ensure accuracy during the aggregation process, factor analysis and mahalanobis distance methodologies are used and the data inconsistency is optimized. The performance graph shows that the Factor analysis and mahalanobis distance detects outlier better than the principal component analysis and subspace methods.

KEYWORDS: Aggregation, Distance Measure, Factors, Outlier, Sensor Network.

1. INTRODUCTION

The Wireless Sensor Networks (WSNs) are autonomous networks consisting of a large number of sensor nodes. WSNs consist of hundreds or even thousands of resource constrained nodes. Sensor is a small, lightweight device which measures the environment of physical parameters such as temperature, pressure, relative humidity etc. Sensor nodes are applied to perform measurements of some physical phenomena [Aky02]. Due to the hostile operation of sensor networks the sensor nodes are at risk to attacks, leading to the injection of false data into the network. This makes the data received at base station inconsistent with the observed phenomena or data. Thus the data integrity and accuracy becomes a major issue [CES04]. WSNs are powerful paradigm for many applications such as target tracking in battlefields, habitat monitoring in forest, environmental control and emergency response system. It is also employed in several real time monitoring systems like weather monitoring [Sze04], environmental monitoring, vehicle tracking etc. Due to the critical nature of such applications, data integrity and accuracy problems become major issues in WSNs. To keep data reliability and accuracy high and be able to make effective and correct decisions using data collected by WSNs, the problem of erroneous data due to the existence of either compromised or faulty nodes in the network becomes of higher importance when data aggregation occurs.

In WSNs neighbor nodes sends data to the aggregator nodes. Processing, caching and filtration are done by aggregator nodes to get meaningful information and thus by resending to sink nodes. Raw data are collected by aggregators from the remaining nodes in the cluster and summarizing is done by them to make the data usable. Unnecessary transmissions are eliminated from multiple sensors by aggregating the data and sending accurate information to the base station is called data aggregation in case of WSN's. Redundancies and faults are eliminated using aggregation to improve the level of accuracy [WDA09]. Thus performing aggregation can eliminate the redundancy and erroneous in data and there by improving the level of accuracy. Outliers are the metrics used to calculate the erroneous data, which deviates from the normal pattern of any sensor's data in WSN's.

Erroneous data termed as outliers in WSNs, are those measurements which deviate significantly from the normal pattern of sensor data. A key issue to extend WSNs lifetime is reducing energy consumption. One of the most important energy saving mechanism is to exploit data aggregation. Elimination of unnecessary packets transmission is done by data aggregators with the help of filtering. Data aggregation reduces the number of data packet transmitted and thus saves energy. But at the same time, it also brings the largest data latency because data from different sources may have to be held back at an aggregator node in order to be aggregated with data coming from other sources.

The remainder of this paper is organized as follows. In section 2, we discuss related works on data aggregation and outlier detection techniques. Section 3 describes the basic techniques involved in multivariate outlier detection. Section 4 elaborates the outlier detection process. In Section 5 we discuss the detail of experimental results. Section 6 concludes the paper.

2. RELATED WORK

Most of research efforts presented in the literature have discussed the problem of developing of efficient data aggregation mainly for energy savings and minimization in sensor networks [BS07, JSF06]. Prior work has rarely been concerned with the delay of aggregation and accuracy together. In [TH05], 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. Each node stores the tests in a moving window, so any neighbour that starts reporting deviating values that exceed a predefined threshold is marked. Detection is achieved with collaboration between the nodes so as to isolate the problematic nodes.

The authors in [CP07, C+06, C+10b] have adopted principal component analysis (PCA) for dimension reduction and SPE to perform fault detection in the residual space. Since SPE is sensitive to modeling errors, it could increase the false alarm rate. However, that technique requires the complete knowledge of the density distribution function of the collected data. The authors in [C+10a] design a resilient aggregation using a robust estimation obtained from data distribution based on a statistical method.

The proposed approach affords an incorporated technique of taking into consideration and merging effectively correlated sensor data, in a distributed hierarchical network, in order to disclose outliers from the large dataset.

3. SYSTEM MODEL AND ARCHITECTURE

A sensor network is usually represented by a network graph. An algorithm that correlates metrics from neighbouring sensors is considered, to detect the nodes containing anomalies in the corresponding network graph. In order to decentralize the detection algorithm, the sensor network is divided into groups of sensors. Each group is controlled by a single coordinator. Coordinator is the responsible for processing the sensed data from the normal nodes in the group. Coordinators do the role of aggregator which aggregates and transmit data to the base station. This aggregation process reduces the energy consumption in order to avoid unwanted data transmission.

Figure 1 shows a cluster based sensor network organization. The cluster heads can communicate with the sink directly via long range transmissions or multi hopping through other cluster heads. In cluster based approach, sensor nodes is divided into towards the base station through a specific node called Cluster Head (CH) taking care about the data aggregation from ambient nodes. All nodes in the specific cluster send its measured data to assigned CH performing the summarization and aggregation tasks and transmit an aggregated message to another CH or directly to the base station.

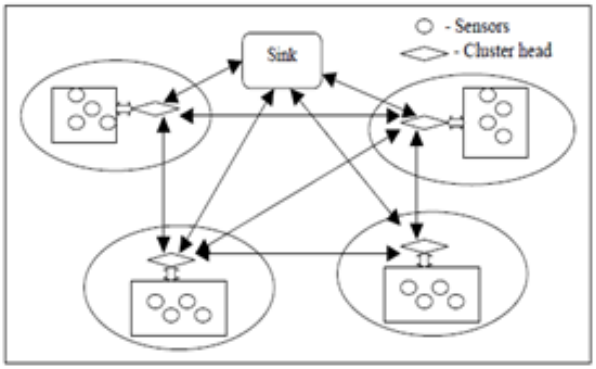

Figure 1. Cluster based sensor Network

4. MULTIVARIATE OUTLIER DETECTION

The objective of the multivariate outlier detection algorithm is to provide an efficient and effective methodology of combining data of heterogeneous monitors that spread throughout the network and analyzing interdependencies between variables in the data set. Outlier detection is possible only when multivariate analysis is performed, and the interactions among different variables are compared within the class of data [HCB00]. In this paper, we propose factor analysis as a multivariate technique to identify the linear relationships among the observed variables and model the normal pattern of sensor data. The technique is augmented with mahalanobis distance measure to detect and isolate the inconsistent data.

4.1 Factor Analysis

Factor Analysis (FA) technique is used to identify the correlated relationship among the set of variables. The hidden model helps to analyze the multivariate data. It is used to expose the dimensions of a set of variables and larger number of variables to a smaller number of factors [JW98, AB06]. Factor analysis deals with a set of interpretations obtained from a given sample, analyzing the set of data with their correlations and evaluate the deviation compared with previous samples.

Factor analysis is decompositional in nature in that it identifies the underlying relationships that exist within a set of variables. Factor analysis creates groups of metric variables (interval or ratio scaled) called factors. A factor is an underlying quality found to be characteristic of the original variables. Two types of factors exist. Common factors have effects

shared in common with more than one observed variable. Unique factors have effects that are unique to a specific variable. Being a branch of multivariate analysis, Factor Analysis offers a conceptual framework within which many disparate methods can be unified and a base from which new methods can be developed.

Let X be the observable random vector with m variables X_1, X_2,.. ..,X_n that have mean μ and covariance matrix Σ. The factor model postulates that X is linearly dependent on a few unobservable random variables F_1, F_2... F_p, called common factors, and m additional sources of variation e_1, e_2... e_m, called errors, or specific factors. The factor analysis model is as follows:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_m - \mu_m &= l_{m1}F_1 + l_{m2}F_2 + \cdots + l_{mp}F_p + \varepsilon_m \end{aligned}$$

or, in matrix notation

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \mathbf{e}$$

The coefficient l_{ij}, is called the loading of the ith variable on the jth factor, so the matrix L is the matrix of factor loadings. The ith specific factor ε_i is associated only with the ith .response X_i. The p deviations $X_1 - \mu_1, X_2 - \mu_2, \cdots, X_m - \mu_m$ are expressed in terms of m+p random variables F_1, F_2... F_p, e_1, e_2... e_m, which are unobservable. If assume the unobservable random vectors F and e satisfy the following conditions:

$$E[\mathbf{F}] = 0,\ Cov(\mathbf{F}) = E[\mathbf{F}\mathbf{F}^T] = \mathbf{I}$$

$$E[\mathbf{e}] = 0,\ Cov(\mathbf{e}) = E[\mathbf{e}\mathbf{e}^T] = ? = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \psi_m \end{bmatrix}$$

$$Cov(\mathbf{e}, \mathbf{F}) = E[\mathbf{e}\mathbf{F}^T] = 0$$

The orthogonal factor model implies a covariance structure for X. From the model

$$\begin{aligned} (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T &= (\mathbf{LF} + \mathbf{e})(\mathbf{LF} + \mathbf{e})^T \\ &= (\mathbf{LF} + \mathbf{e})((\mathbf{LF})^T + \mathbf{e}^T) \\ &= \mathbf{LF}(\mathbf{LF})^T + \mathbf{e}(\mathbf{LF})^T + \mathbf{LF}\,\mathbf{e}^T + \mathbf{e}\mathbf{e}^T \end{aligned}$$

So that,

$$\begin{aligned} \Sigma = Cov(\mathbf{X}) &= E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})T \\ &= \mathbf{L}E(\mathbf{F}\mathbf{F}^T)\mathbf{L}^T + E(\mathbf{e}\mathbf{F}^T)\mathbf{L}^T + \mathbf{L}E(\mathbf{F}\mathbf{e}^T) + E(\mathbf{e}\mathbf{e}^T) \\ &= \mathbf{L}\mathbf{L}^T + ? \end{aligned}$$

From the covariance calculation,

$(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}^T = (\mathbf{LF} + \mathbf{e})\mathbf{F}^T = \mathbf{LF}\mathbf{F}^T + \mathbf{e}\mathbf{F}^T$, so

$Cov(\mathbf{X}, \mathbf{F}) = E(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}^T = \mathbf{L}E(\mathbf{F}\mathbf{F}^T) + E(\mathbf{e}\mathbf{F}^T) = \mathbf{L}$.

Thus, we can get covariance structure for the orthogonal factor model:

1. $Cov(\mathbf{X}) = \mathbf{L}\mathbf{L}^T + ?$ or
   $\mathrm{Var}(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i$
   $Cov(X_i, X_k) = l_{i1}l_{k1} + \cdots + l_{im}l_{km}$
2. $Cov(\mathbf{X}, \mathbf{F}) = \mathbf{L}$ or
   $Cov(X_i, F_j) = l_{ij}$

The portion of the variance of the ith variable contributed by the m common factors is called the ith communality.

4.1.1 Methods of parameter estimation

The principal component factor analysis of the sample covariance matrix S is specified in terms of its eigenvalue-eigenvector pairs $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_m$. where, $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \cdots, (\hat{\lambda}_m, \hat{\mathbf{e}}_m)$

Let $p < m$ be the number of common factors. Then the matrix of estimated factor loadings $\{\mathbf{l_{ij}}\}$ is given by

$$\hat{\mathbf{L}} = [\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2, \cdots, \sqrt{\hat{\lambda}_p}\hat{\mathbf{e}}_p]$$

the estimated specific variance are provided by the diagonal elements of the matrix $\mathbf{S} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T$ so

$$\hat{?} = \begin{bmatrix} \hat{\psi}_1 & 0 & \cdots & 0 \\ 0 & \hat{\psi}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\psi}_m \end{bmatrix} \quad \text{with} \quad \hat{\psi}_i = s_{ii} - \sum_{j=1}^{p} \hat{l}_{ij}^2$$

Communalities are estimated as

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{ip}^2$$

The principal component factor analysis can also be applied to the sample correlation matrix R. For the principal component solution, the estimated loadings

for a given factor do not change as the number of factors is increased. For example, if $p=1,\ \hat{L}=[\sqrt{\hat{\lambda}_1}\hat{e}_1]$ and if p=2 $\hat{L}=[\sqrt{\hat{\lambda}_1}\hat{e}_1,\sqrt{\hat{\lambda}_2}\hat{e}_2]$, where $(\sqrt{\hat{\lambda}_1},\hat{e}_1)$ and $(\sqrt{\hat{\lambda}_2},\hat{e}_2)$ are the first two eigen value-eigenvector pairs for S (or R). The principal component method satisfies

$$\Sigma = LL^T + 0 = LL^T$$

Factor scores are estimates of values for the unobserved random factor vector F_j, j= l... n. That is, factor scores

$\hat{f}_j$ = estimate of the values f_j attained by F_j Since the unobserved quantities f_j and e_j, outnumber the observed X_j, some approaches to estimating factor values have been proposed including the weighted least squares method and the regression method.

$$\hat{f}_j = (\hat{L}^T\hat{L})^{-1}\hat{L}^T(x_j - \bar{x})$$

where,

$\bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j$ is the sample mean.

4.1.2 Shaping the quantity of factors

The number of factors can be determined based on the following approaches.

a) Kaiser decisive factor: The Kaiser decisive factor is the default method to detect number of factors. It eliminates all components with eigen values under 1.0. But it is not suitable when used as the so single threshold value is used for factor estimation.

b) Screen scheme: The screen scheme plots the factors as the X axis and the corresponding eigen value as the Y-axis. The eigen values drop when the curve makes an elbow toward less steep decline, screen test says to drop all supplementary factors after the elbow.

c) Variance explained principle: Common rule for keeping factors to report for 90% of the variation. In this principle correlation based factor dropping method is followed. Threshold is set based on the dependent variable. A very small factor can have a large correlation with the dependent variable, in which case it should not be dropped. In this paper, numbers of factors are selected from factor analysis by using variance explained criteria.

4.2 Mahalanobis Distance

Many multivariate techniques applicable to anomaly detection problems are based upon the concept of distance. The mahalanobis distance 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. If the independent variables are uncorrelated, it is the same as the simple Euclidean distance. Thus, this measure provides an indication of whether or not an observation is an outlier with respect to the independent variable values [JW98]. It is a useful way of determining the similarity of a set of values from an unknown sample to a set of values measured from a collection of known samples. When there are two groups with x_i and x_j which follows multivariate normal distribution, then, Mahalanobis distance is given by the following formula

$$d_{ij} = \left((x_i - x_j)^T S^{-1}(x_i - x_j)\right)^{\frac{1}{2}}$$

S is a correlation coefficient between x_i and x_j data sets

Algorithm for Distance Calculation
(i) Determine the unknown sample scores from factor analysis (x_i)
(ii) Calculate the centroid value of each variable
(iii) Calculate correlation *s* and mean (x_j) of the factor score.
(iv) Finally, calculate the mahalanobis distance d_{ij} from the center of the data.

5. OUTLIER DETECTION PROCESS

In this paper, we propose combined mahalanobis distance with factor analysis to detect outlier data in WSN. Our proposed technique enables the aggregator in network to identify the new arriving data measurements from its members as normal or abnormal. Using the naturally existing correlation among the sensor readings, the aggregator can efficiently detect the local outliers.

The algorithm for implementing the proposed outlier detection as follows:

Algorithm for Outlier Detection
Input: Sample vector v representing a test event, threshold for mahalanobis distance. ***Output:*** Decision on v, •Compute the factor scores of v •Compute mahalanobis distance(d1) of v •If d1>d, v is an abnormal event, otherwise v is a normal event. Where, d is the distance of known event.

The conceptual model of the proposed system is illustrated in Figure 2. The algorithm is implemented in two phases: the data modeling phase, that creates a model of the normal condition of the monitored parameters, and the outlier detection phase, that detects anomalies by comparing the actual data with the modeled one. During the data modeling phase, FA is applied on the correlation matrix of the sample data set and the first k most important factors are selected.

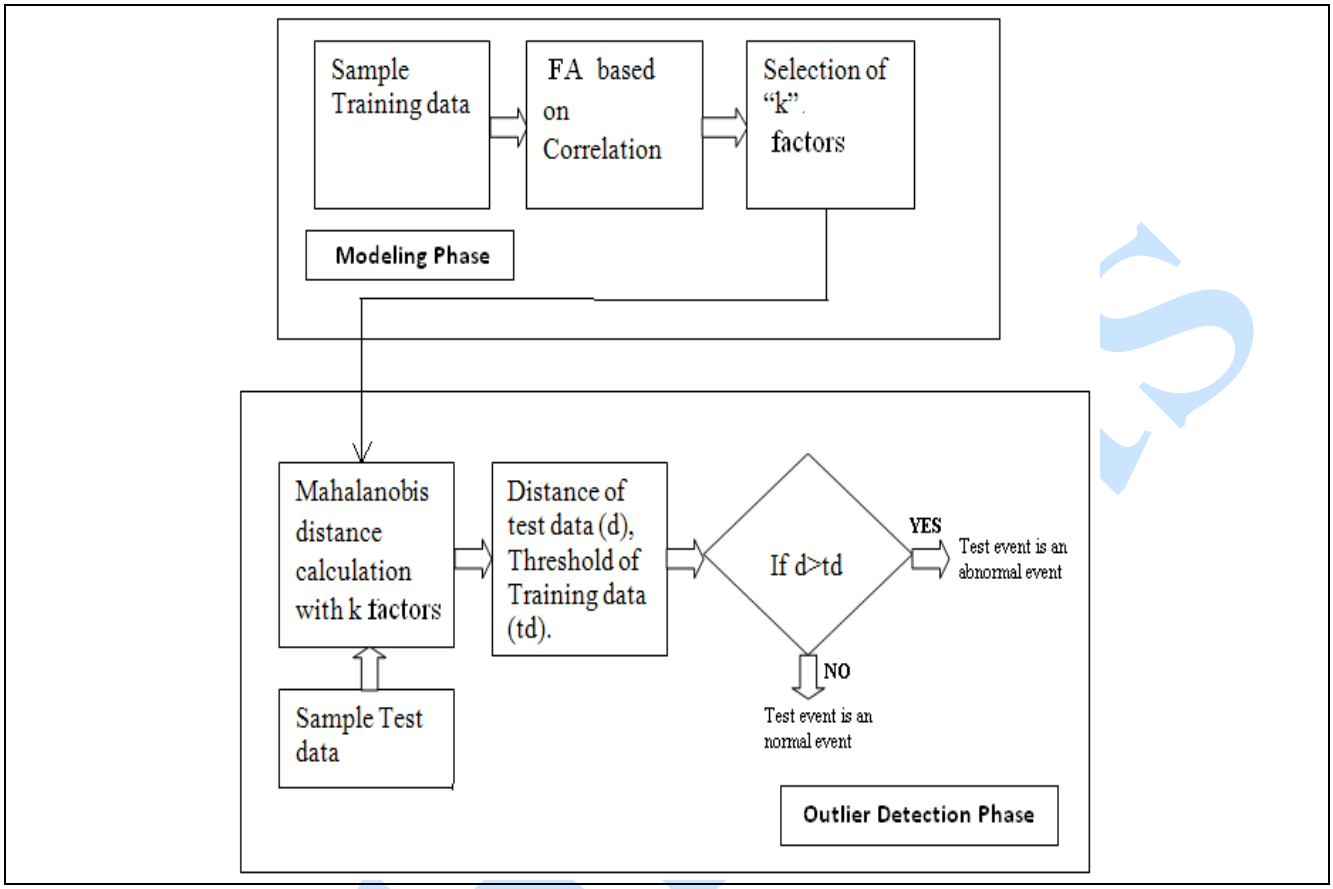

Figure 2. Conceptual Model

6. RESULTS AND DISCUSSIONS

This section specifies the performance evaluation of our technique compared to the PCA with subspace approach in [CP07] .We conduct experiments on the real data gathered from a deployment of WSN in the Intel Berkeley Research Laboratory [***]. The protocol is simulated in Matlab and considers a cluster as shown in Figure 3. The real data are collected from a closed neighbourhood from a WSN deployed in the Intel Berkeley Research Laboratory as shown in Figure 3. IBRL data set obtained from 54 sensor nodes, namely node ID from 1 to 54, during the four hours period collected on 1st March 2004 during the time interval 00:00am to 03:59am. We consider three features namely temperature, humidity and voltage. The closed neighbourhood contains the node 35 and its 5 spatially neighbouring nodes, namely nodes 1, 2, 3, 33, 35. The network recorded temperature, humidity, light and voltage measurements at 31 seconds intervals.

We consider a simulation setup, where N sensors are deployed over a particular region to monitor a specified parameter. We assume the sensors communicate in multi-hop fashion. The sample was generated by multivariate dependency. We have kept o_1 as the number of compromised nodes at a time and C_1 as the corruption rate that defines the rate at which an adversary makes the data alteration.

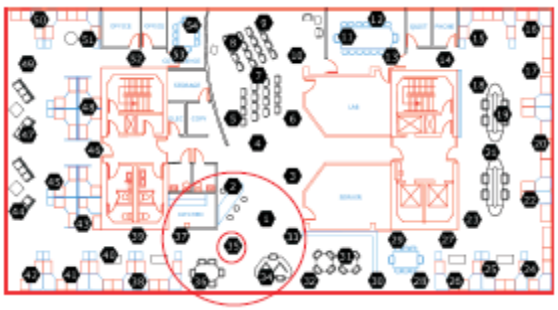

Figure 3. Sensor nodes deployed in Intel Berkeley Research Lab

The outlier was simulated by a function which alters the measurements in sample according to the corruption rate C_1 .

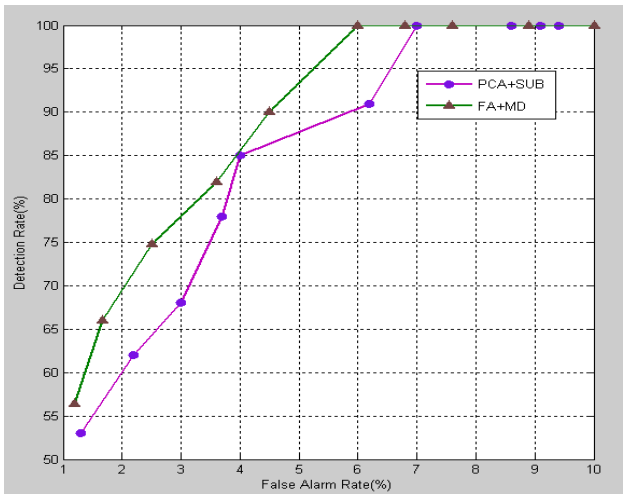

Figure 4. Factor Analysis Vs PCA approach

To obtain the model we have made 50 simulation runs for zero o_1 and C_1. We evaluated the performance using two metrics namely detection rate and false alarm rate. Receiver Operating Characteristics (ROC) curve is used to visualize the tradeoff between the detection and false alarm rate. Figure 4 compares the detection rate of the factor analysis with mahalanobis distance to the PCA with subspace approach for multiple outliers. In Figure 4, the x-axis represents the faulty data rate and y-axis represents detection rate for different cluster size C. From Figure 4, we can see that Factor analysis based approach is effective in detecting the inconsistent data as the detection probability rapidly increases to a high value (90%). While PCA approach offer only an average of 85% detection rate. Figure 5 shows the detection rate of factor analysis for different data alteration rate in different outlier percentage OL for a fixed cluster size.

From the Figure 5, it is clear that it is effective in detecting outliers even when half of the cluster nodes in a cluster are faulty and found to maintain an average detection rate of 90%.

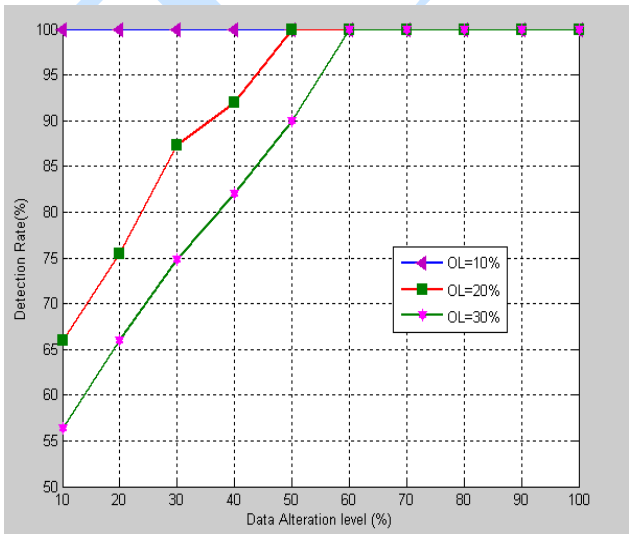

Figure 5. Detection rate for different outlier percentage

CONCLUSION

In this paper, we have proposed an outlier detection technique based on factor analysis and mahalanobis distance. Factor analysis describes the ordinary network behaviour by discovering the hidden structure of them. It also reduces large number of dataset into a smaller number of factors. The Mahalanobis distance is used to determine the "similarity" of network activities to the profiles. We compare the performance of our approach with the PCA and Subspace approach using real data of the Intel Berkeley Research Laboratory. Experimental research shows that our approach gives better performance for outlier data and does not require outlier free training data to design model. Based on comparative results that we obtained, we conclude that our proposed methodology improves significantly the anomaly detection capabilities, especially when compared against conventional principal component analysis based anomaly detection approach.

REFERENCES

[Aky02] **I. F. Akyildiz et al.** - *Wireless Sensor Networks: A Survey*, Computer Networks, vol. 38, no. 4, 2002, pp. 393–422.

[AB06] **Hair Anderson, Tatham Black** - *Multivariate Data Analysis*, Dorling Kindersley, Pearson Education, 2006.

[BS07] **S. Brown, C. J. Sreenan** - *A Study on Data Aggregation and Reliability in Managing Wireless Sensor Networks*", IEEE 2007.

[CES04] **David Culler, Deborah Estrin, Mani Srivastava** - *Overview of Sensor Network*, IEEE 2004.

[CP07] **Vassilis Chatzigiannakis, Symeon Papavassiliou** - *Diagnosing Anomalies and Identifying Faulty Nodes in Sensor Networks*, IEEE Sensors Journal, VOL.7, NO.5, MAY 2007.

[C+06] **V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, B. Maglaris** - *Hierarchical Anomaly Detection in Distributed Large-scale Sensor Networks*, IEEE 2006.

[C+10a] **N. Chitradevi et al.** - *Estimation based Efficient and Resilient Hierarchical In-*

Network Data Aggregation Scheme for Wireless Sensor Network, in International Journal of Futuristics Computer Applications, FCS Publisher, 25th February 2010

[C+10b] **N. Chitradevi** - *Outlier aware Data Aggregation in Distributed Wireless Sensor Network using Robust Principal Component Analysis*, Selected for IEEE sponsored International Conference ICCCN 2010, Chettinad College of Engineering and Technology, Karur, 31st May 2010.

[HCB00] **W. R. Heinzelman, A. Chandrakasan, H. Balakrishnan** - *Energy-Efficient Communication Protocol for Wireless Micro-sensor Networks*, in the Hawaii International Conference on System Science, Maui, Hawaii, 2000.

[JSF06] **Xu Jianbo, Zeng Siliang, Qu Fengjiao** - *A new In-network Data Aggregation Technology of Wireless Sensor Networks*, IEEE 2006

[JW98] **R. A. Johnson, D. W. Wichern** - *Applied Multivariate Statistical Analysis*, Upper Saddle River, NJ: Prentice Hall, 1998.

[Sze04] **R. Szewczyk et al**. - *Habitat monitoring with sensor networks*, CACM, vol.47, no.6, pp.34-40, 2004.

[TH05] **S. Tanachaiwiwat, A. Helmy** - *Correlation analysis for alleviating effects of inserted data in wireless sensor networks*, in Proc. Mobile and Ubiquitous Syst.: Networking Services, 2005, pp. 97–108.

[WDA09] **Mohamed Watfa, William Daher and Hisham Al Azar** - *A Sensor Network Data Aggregation Technique*, International Journal of Computer Theory and Engineering Vol. 1, No. 1, April 2009

[***] http://db.csail.mit.edu/labdata/labdata.html