Distributed Maximum Likelihood Principal Component Analysis for Wireless Sensor Network Data

I. Nandhini[1],
PG Scholar,
Department of Computer Science and Engineering,
Coimbatore Institute of Technology,
Coimbatore.

Dr. K. Amshakala[2],
Assistant Professor,
Department of Computer Science and Engineering,
Coimbatore Institute of Technology,
Coimbatore.

***Abstract*:- Principal Component Analysis is a data dimensionality reduction technique well suited for processing data from sensor networks and also is an effective anomaly detection technique. In an environment where the anomalies are present in the dataset, the derived principal components can be misleading by the anomalies .The anomalies can be detected by analyzing the data collected from the wireless sensor network across the environment. In this paper, a distributed maximum likelihood PCA algorithm is proposed that is more efficient in finding the principal components from the data containing anomalies. The algorithm uses maximum likelihood functions to find principal components locally and then compares it with principal components computed across the network to identify the anomalies.**

Keywords— Principal Component Analysis, Distributed Maximum likelihood PCA, Anomalies.

1 INTRODUCTION

Efficient in-network data processing is a key factor for enabling wireless sensor networks (WSN) to extract useful information and an increasing amount of research has been devoted to the development of data processing techniques. Wireless sensors have limited resource constraints in terms of energy, network data throughput and computational power.

1.1 Motivation

Anomaly detection, also known as outlier detection, is a machine learning problem. An anomaly is defined by Barnett and Lewis as "an observation (or subset of observations) which appears to be inconsistent with the remainder of the data''[1] Anomaly detection aims to identify data that do not conform to the patterns exhibited by the data set.[2] Methods often use an unsupervised one-class classification approach. The problem thus has two important characteristics, the data are not labeled and there is a class imbalance in the training set where the number of normal data significantly exceeds the number of anomaly data. The nature of sensor, peer to peer and adhoc wireless networks requires a distributed learning approach, as it is infeasible to communicate all data to a centralized node for computation. There are several reasons why data might be in different physical locations. It is too costly to transfer the data to one physical location. Examples include limited energy resources, such as in Wireless Sensor Network (WSN)s, and limited time resources, such as in network intrusion(anomalies).

1.2 Contribution

In this paper, a distributed anomaly detection scheme based on the principal component analysis (PCA) and the maximum likelihood function is proposed. The approach addresses the challenge of performing anomaly detection in a wireless sensor networks.

- A distributed version of PCA based upon maximum likelihood function. This improves on the performance of classical PCA.
- A detailed evaluation of anomaly detection in a distributed environment is provided.

2 RELATED WORKS

There are two approaches to learning in a distributed environment. The first assumes a structure to the network, i.e. partially distributed learning and another approach is to make no assumptions on the structure of the network, i.e. fully distributed.[3]Large numbers of limited resource sensor nodes operate autonomously to collaborate and manage the wireless networks, through which critical raw data are collected and transmitted to the end users/decision makers. WSNs have been used in critical application scenarios, such as enemy target monitoring and tracking and fire detection system WSNs can be susceptible to anomalies due to cheap unreliable hardware and software, and unfavorable operating environment that can affect the network communication. These anomalies must be detected as they can cause failures in the network and hence affect quality of collected data.[4]Principal component analysis (PCA) is one of the most widely used multivariate techniques in statistics. It is commonly used to reduce the dimensionality of data in order to examine its covariance/correlation structure of a set of variables which gives optimal solution to the problem. PCA [5] is a spectral decomposition technique that has been shown to perform well as an anomaly detector (e.g.,[6], [7]). There are several methods that have been used to construct the

principal components (PC)s in a partially distributed environment, for example fusing data [6], and constructing PCs at a cluster head [8]. [9] Power iteration method in PCA, which closely matches the approach proposed in this article. Their algorithm aims at computing the principal eigenvector of the adjacency matrix of a peer-to-peer network, which leads to the ranking of the network components, in a way similar to the page rank algorithm. The underlying hypotheses of the distributed system are interestingly close to ours (each component has only access to one dimension the problem, and can communicate with components whose data is related).

Huang et al. [10] propose a distributed anomaly detection method that focuses on volume anomalies, unusual traffic load levels caused by worms, distributed denial of service attacks and so on. It is well-known that PCA is extremely fragile in the presence of anomalies in the training data set and even a small number of anomalies can significantly alter the subspace generated [11], [12]. Various techniques have been proposed in order to overcome this issue. Multivariate trimming [13], [14] aims to remove the outliers before deriving the PCs from the clean training data set. [15] Propose a distributed anomaly detection approach for WSNs which only assumes that the network is strongly connected. Each node has a local data set with the aim of computing the set of the global top-k anomalies, the scheme is generic in that it is suitable for all density-based methods, except Local Outlier Factor (LOF). A fully-distributed consensus-based approach for PCA is proposed by Macua et al. [16]. The network-wide covariance matrix is estimated through the use of a consensus averaging (CA) algorithm and an exchange of p × p matrices. PCA is then performed on each node. The algorithm is shown to have guaranteed convergence using only communication with neighboring nodes.[17] A probabilistic faulty detector in wireless sensor networks to improve the battery life of sensor nodes which minimizes additional power burden to sensor batteries.[18] The MLPCA dealing with incorporation of correlated measurement errors in PCA using maximum likelihood estimation. [19] PPCA demonstrate how the principal axes of a set of observed data vectors may be determined through maximum-likelihood estimation of parameters in a latent variable model closely related to factor analysis with iteratively given principal component subspaces. [20] Proposes novel based PCA for detecting anomalies in unsupervised training datasets by the measure of the difference of anomaly from normal instances in principal components.

3 DISTRIBUTED MAXIMUM LIKELIHOOD PCA

In order to define global anomalies in a local data set, it is a requirement to build a classifier on a local node that has been built using information from the local dataset in a network. In order to perform this, by anomaly detecting approach MLPCA. This technique performs superior in the presence of anomalies in the training dataset.

3.1 Maximum Likelihood

Maximum likelihood is the estimation of probability of occurrence of likelihood which is maximum in a dataset.

$$L(\theta, x) = \prod_{i=1}^{n} f(x_i, \theta)$$

Parameter θ would be the value of θ that maximizes the probability that is Likelihood of data.

3.2 Proposed system

Before look into the pseudo code and formal analysis of algorithm. Let us first gain some preprocessing work of given dataset because the sensors absorbs temperature and voltage data's which are inversely propositional so the scale value of both data is different. Convert temperature values and voltage values into HADS scaled form. Export the results and estimate the maximum likelihood function, find PCA for the ML. Detect the anomalies from PCA values. Calculate precision and accuracy for performance assessment provides which sensor was fault. Compare the mean average PCA and maximum likelihood PCA which gives DMLPCA was better and highly suits for the given environment. The pseudo code is given below.

Pseudo code for DMLPCA

Input: Temperature, Voltage

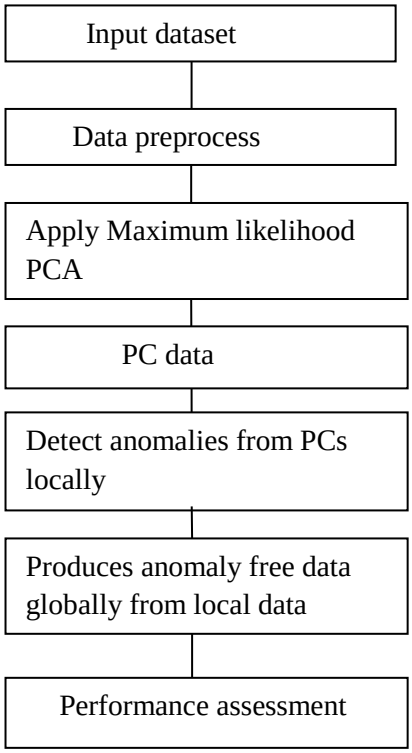
Output: X, Y

1. Preprocessing the values of Temperature and Voltage into scaled form
2. Export the result from step1
3. Compute Maximum likelihood
4. Compute PCA from the likelihood
5. Detect Anomalies from PCA values
6. Do byzantine agreement
7. Calculate precision and accuracy for performance assessment
8. Compare with classical PCA along with obtained mean

Fig 1.1 Pseudo code for DMLPCA

The flow for the Distributed maximum likelihood principal component analysis eventually performs same process described in the above pseudo code. At first it gets the input data and convert them into scaled form, (i.e.,HADS) apply the maximum likelihood function to the dataset, and compute PCA from the likelihood function data. Finally detect the anomalies from the PCs locally. Aggregate all the local anomalies to detect the global anomaly. The evaluations of the dataset, description of dataset and performance assessment are given below.

Flow diagram for DMLPCA



4 EVALUATION

In this section, evaluations on real-world data are presented to illustrate the performance of DMPCA and distributed MLPCA. The evaluation environment is varied in order to examine the behavior of the proposed algorithm in a broad range of settings. All algorithms are implemented in C#.

4.1 Evaluation Environment

The elements considered in the evaluation are wireless sensor network topology and data sets.

4.2 Dataset description

Experiments were carried out using a set of five days of temperature readings obtained from a 54 Mica2Dot sensor deployment at the Intel research laboratory at Berkeley. The readings were originally sampled every thirty-one seconds. A preprocessing stage where data was discretized in thirty second intervals was applied to the dataset. After preprocessing, the dataset contained a trace of 14400 temperature and voltage readings from 52different sensors, randomly chosen without replacement from the appropriate class of the data set. The testing set consists of an equal number of normal and anomaly samples. To form the data sets in a distributed environment, an equal number of data instances is randomly distributed across the sensors. Temperature over the whole set of data ranged from about 0_C to 35_C and voltage ranged from about 0_ V to35_V.

4.2 Performance Assessment

To examine performance, false positive rate (FPR) is the ratio of false positives and the true positive rate (TPR) is the ratio of true positives are calculated to find precision and accuracy for anomaly measurements of both DMPCA and DMLPCA. To compare schemes, curves are generated by varying the number of faulty sensor nodes.

4.2.1 DMLPCA evaluation on real world data

In this section, the performance of DMLPCA is examined and compared with DMPCA anomaly detection methods. Real-world data sets are used to examine performance of DMLPCA. The algorithms, is used to determine the optimal value of the data.

In the local approach, the data are randomly distributed between the nodes in the network. Each node constructs a classifier from the data available on the sensor node. The same test data set is used across all nodes. For DMPCA mean of the local classifiers is noted and then the mean and standard deviation of the performance over the 10 iterations are recorded. For DMLPCA maximum likelihood of the local classifiers over the 10 iterations are recorded.

4.2.2 Visualization on a real data Set

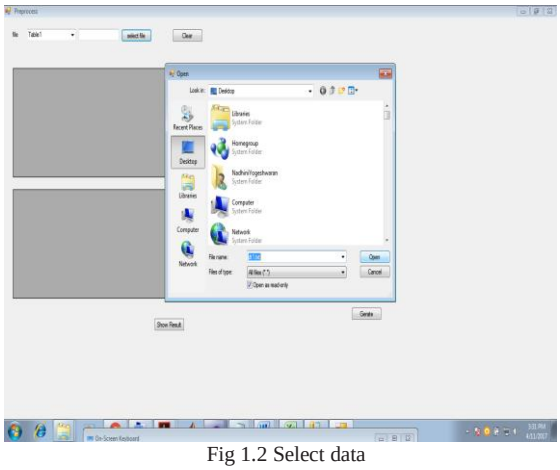


Fig 1.2 Select data

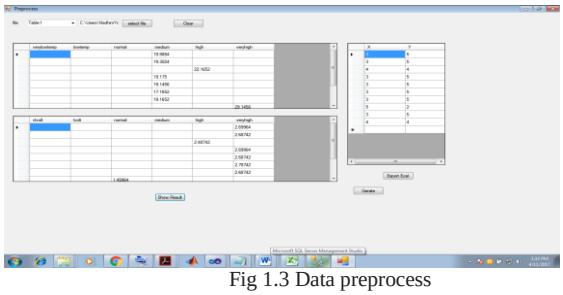


Fig 1.3 Data preprocess

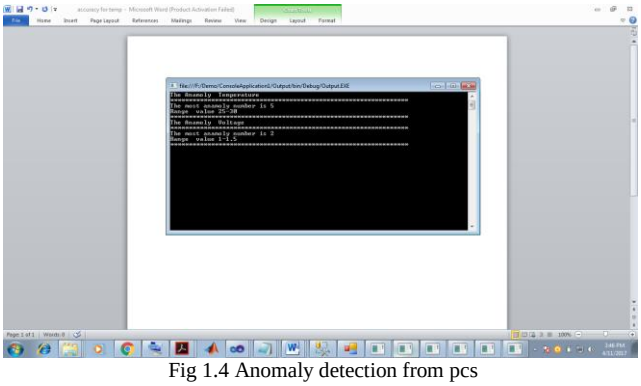


Fig 1.4 Anomaly detection from pcs

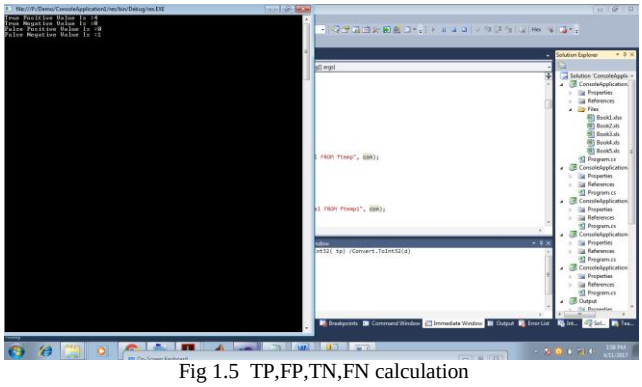


Fig 1.5 TP,FP,TN,FN calculation

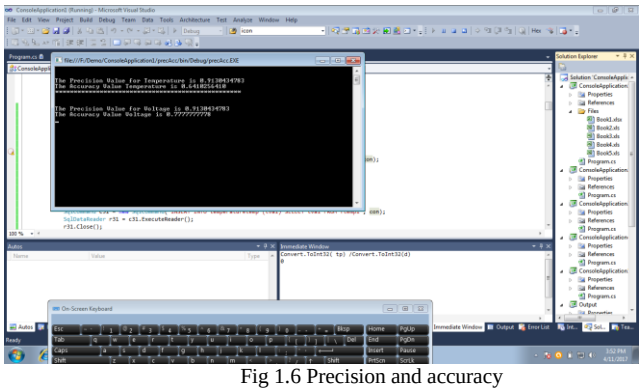
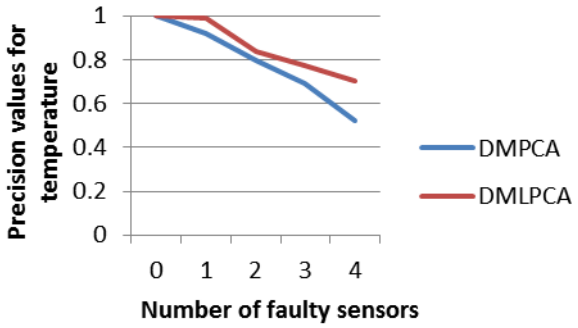
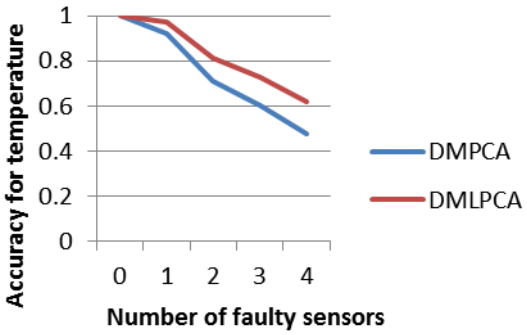
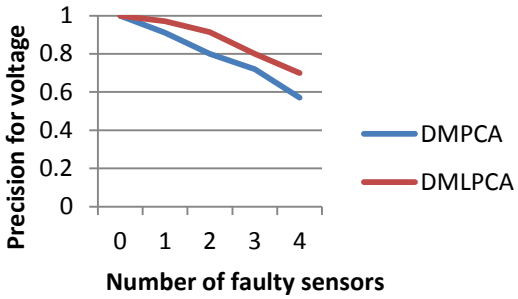


Fig 1.6 Precision and accuracy

4.3 Comparision of DMPCA and DMLPCA







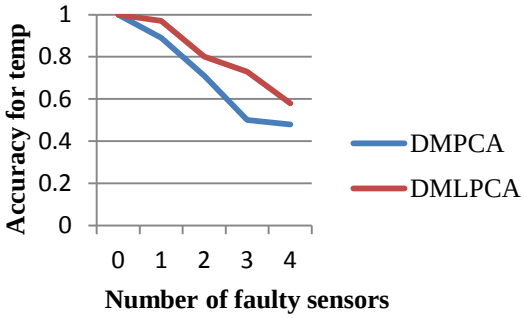


Fig 1.7 Comparison graph

5 CONCULSION

A robust PCA-based anomaly detection algorithm that operates in a distributed environment was proposed. Maximum likelihood PCA is able to determine the PCs more robustly in the presence of anomalies by constructing likelihood around the data that reduces the influences of anomalies in the training set. Evaluations on real-world data sets show the performance of DMLPCA exceeds that of DMPCA (Distributed Mean PCA) anomaly detection techniques. DMLPCA takes maximum likelihood function and follows byzantine agreement of N+1/2 non faulty nodes. The byzantine agreement problem requires all fault-free processors to agree on a common value, even if some components are corrupt. Real-world data set shows distributed algorithm is able to increase the performance of DMLPCA than DMPCA.

REFERENCES

[1] V. Barnett and T. Lewis, Outliers in Statistical Data. New York, NY, USA: Wiley, Apr. 1994.
[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Surveys, vol. 41, no. 3, pp. 15:1–15:58, Jul.2009.
[3] S. Rajasegarar, C. Leckie, and M. Palaniswami, “Hyperspherical cluster based distributed anomaly detection in wireless sensor networks,” J. Parallel Distrib. Comput., vol. 74, no. 1, pp. 1833–1847, 2014.
[4] Robert Reris and J. Paul Brooks “Principal Component Analysis and Optimization: A Tutorial” 14th INFORMS Computing Society Conference Richmond, Virginia, January 11{13, 2015pp. 212
[5] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” J. Educational Psychol., vol. 24, pp. 417–441, 1933.
[6] V. Chatzigiannakis and S. Papavassiliou, “Diagnosing anomalies and identifying faulty nodes in sensor networks,” IEEE Sens. J., vol. 7, no. 5, pp. 637–645, May 2007.

[7] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," ACM SIGCOMM Comput. Commun. Rev.,vol. 34, pp. 219–230, 2004.
[8] M. A. Livani and M. Abadi, "Distributed PCA-based anomaly detection in wireless sensor networks," in Proc. 5th Int. Conf. Internet Technol. Secured Trans., Nov. 2010, pp. 1–8.
[9] L. Huang, X. Nguyen, M. Garofalakis, M. I. Jordan, A. Joseph, and N. Taft, "In-network PCA and anomaly detection," in Proc. Adv. Neural Inform. Process. Syst., Dec. 2006, pp. 617–624.
[10] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in Proc. Adv. Neural Inform. Process. Syst., Dec. 2009, pp. 2080–2088.
[11] E. J. Cand_es, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" J. ACM, vol. 58, no. 3, pp. 1–37, May 2011.
[12] D. M. Titterington, "Estimation of correlation coefficients by ellipsoidal trimming," Appl. Statist., vol. 27, no. 3, pp. 227–234, 1978.
[13] R. Kwitt and U. Hofmann, "Robust methods for unsupervised PCA-based anomaly detection," in Proc. IEEE/IST Workshop Monitoring, Attack Detection Mitigation, Sep. 2006, pp. 1–3.
[14] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," Knowl. Inform. Syst., vol. 34, no. 1, pp. 23–54, 2013.
[15] S. Macua, P. Belanovic, and S. Zazo, "Consensus-based distributed principal component analysis in wireless sensor networks," in Proc. IEEE 11th Int. Workshop Signal Process. Adv. Wireless Commun., Jun. 2010, pp. 1–5.
[16] Yann-A¨el Le Borgne 1;?, Sylvain Raybaud2, and Gianluca Bontempi1 "Distributed Principal Component Analysis for Wireless Sensor Networks" Sensors 2008, 8, 4821-4850; DOI: 10.3390/s8084821
[17] Bill C.P. Lau a, Eden W.M. Maa, Tommy W.S. Chow a,b,⇑ "Probabilistic fault detector for Wireless Sensor Network" B.C.P. Lau et al. / Expert Systems with Applications 41 (2014) 3703–3711.
[18] Peter D. Wentzell,1 Darren T. Andrews,1 David C. Hamilton,2 klaas faber3 and Bruce R. KowalskI3 "Maximum likelihood principal component analysis" Journal of Chemometrics, Vol. **11**, 339–366 (1997).
[19] Michael E. Tipping Christopher M. Bishop "Probabilistic Principal Component Analysis" Journal of the Royal Statistical Society, Series *B*, **61**, *Part* 3, *pp.* 611–622.
[20] Intel Berkeley Research lab between February 28th and April 5th, 2004.