Volume 3, No. 1, January 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

NEW AND FAST OUTLIER DETECTION SCHEME IN WSN: NFODS-WSN

Vipnesh Jha*[1], Sumit kumar srivastava[2]
M. Tech. (CSE) Student & Faculty of Computer Science & Engineering, Apex Institute of Technology, Rampur, (U.P), INDIA[1]
M. Tech. (CSE) Student & Faculty of Computer Science & Engineering, D.B.I.T, Dehradun, (U.K.), INDIA[2]
vipneshjha@gmail.com[1], sumitsri15.007@gmail.com[2]

Abstract: - In this work, we focused on the problem of outlier detection in wireless sensor networks. Outlier detection techniques generally focus the developer's or user's context into the interesting events, or unexpected results in the network which has very low probability of occurrence. In this paper, we have proposed a model that is based on the approximation of the sensor data distribution. We processed and evaluated our proposed scheme with a set of experiments with datasets which is taken from Intel Berkeley research lab. The experimental evaluation shows that our algorithm can achieve very high precision and recall rates for identifying outliers.

Keywords: - wireless sensor networks, energy, security, privacy.

INTRODUCTION

As we know that sensors has become very tiny, low cost, power efficient and have very short span of life time because of low capacity inbuilt batteries. These sensor nodes are multifunctional and communicate un tethered to the short distances. So we must have to use them efficiently as much as we can with minimal wastage of energy. Moreover, these sensor nodes are deployed randomly and hazardously on a physical world for both in Civil and military applications. In such type of scenario, efficient monitoring of the physical phenomenon and detecting interesting events becomes a big issue. Along with these, detecting unwanted, uninteresting or faulty events makes the application more robust, efficient and accurate. Now there are new processor technologies and methodologies for sensor nodes and wireless communication. These technology or methodology actually enabled us to deploy these low cost, small sensors into various real world scenarios. A sensor network consists of huge number of sensors, collecting and communicating the sensor measurements of observing physical world. We must take into consideration various characteristics of these data streams generally called sensor data streams like imprecision, uncertainty and dynamism. ***The various characteristics of a data streams are discussed below:***

Uncertainty: Data streams generated by sensor readings are discrete in nature, for a continuous physical phenomenon. These samples of data describe a state of physical world at a particular time instant.

a. Data collection
b. Communication of data
c. Computation

Inter- and Intra-stream correlation:

Data samples have temporal correlation among and within data streams, since these data streams are only temporal observations of the physical world. Suppose sensors are deployed in traffic monitoring system and they are detecting the vehicle position say at time T0 and T1 it's obvious that these two readings are correlated through the vehicle velocity and time difference (T1-T0).

Sensitive to energy consumption:

A sensor has inbuilt low power battery within it and since the size of sensors is a big question of concern here, hence it usually get discharged by unnecessary transmission. So when and how frequently data samples are transmitted, this question now becomes a big issue while determining the lifetime of sensors in wireless sensor network.

Motivation: A noise (outlier) is most likely to occur while gathering the data from sensor networks since these sensor readings are discrete, uncertain and transient by nature. The noise in the network might be due to influence of external environment, or other sensors nearby it or may be due to abnormal behavior of particular sensors itself. Gathering the sensor's readings and deriving useful information from these becomes more challenging task because of

a. Sort life time of sensors,
b. Limited CPU capability and
c. Limited Network bandwidth.

These raw sensor readings having spurious data or noisy data are typically transmitted to a central base station that provides an interface to query and provide appropriate answer. Since the lifetime of sensors are very sort in time, and to prevent the network from overloaded network traffic, its totally absurd to transmit all the readings to the central base station.

Problem statement: A sensor network produces huge amount of data rapidly in the form of various streams hence these values became uncertain, discrete and unfaithful for our application. And also in these streams there might be some relation among other streams or might be within the particular stream itself. And also sensors have to work in unattended environment since it becomes isolated or no human intervention after deployment of those tiny sensors. Processing of these data now becomes much more tough and tedious since we have to take into consideration all these features of sensors and all these operating scenario in our work.

PROPOSED APPROACH

Statistical modeling technique: A sensor network captures samples of data from real world physical phenomenon. In this type of organization the sensors are located at the network nodes. Let us assume that each sensor is measuring a single real valued attribute X_i, at each time instant. So considering this situation we have to model the set of attributes X_1, ..., X_n as an n-dimensional random variable X = (X_1, ..., X_n), i.e., we assume that the sensor readings are samples of the random variable X.

Kernel Density Estimation: As discussed above and in literature survey that in order to process the streaming sensor values, the first step is to produce a uniform random sample of streaming sensor data. However random sampling is the basic and simple statistical estimation technique for Density Estimation and we all are aware that distribution of random sample is defined by its Probability Density Function (PDF). And in Kernel Density Estimation all the game go just around the Kernel Functions, it's the sum over Kernel Functions which are centered at sample points of concerned data stream. Generally in kernel density estimation a kernel function depicts the way of distributing the weights in the area near to the values or points which is being processed. To determine the density estimation for whole data set we must have to combine all the kernel functions.

Methods for Weighting: An appropriate weighting scheme may prove a better accuracy and can increase the efficiency in kernel density estimation. The Arithmetic weighting method and exponential weighting method discussed in [12] are the two methods which will fulfill our requirement here.

Outlier Detection: To detect a point as an outlier we need to compute the deviation factor (DEVF)[18] and Normalized Deviation Factor (NDEVF) [18] for a particular point in the data stream. Deviation factor at radius r for a point y is the relative deviation of its local neighborhood density from the average local neighborhood density in its r-neighborhood i.e. (sampling neighborhood) [18]. According to this we need to calculate the count of neighbors and samples of neighbors [18] which are also called counting neighborhood and sampling neighborhood. These two values would be determined from the above discussed Kernel Density Estimation model which are (i) αr-neighbors of the observation y (counting neighborhood) and (ii) the total no of observations in the interval 2αr (sampling neighborhood).

RESULT AND ANALYSIS

To apply our proposed method, we need a dataset. And, we got the real data set from Intel Berkeley Research lab [20] and downloaded it. This dataset is freely available on their website [20] and contains information about data collected from 54 sensors deployed in the lab between February 28th and April 5th, 2004. This file includes a log of about 2.3 million readings collected from these sensors. The file is 34MB gzipped, and 150MB uncompressed. In our work, with the help of 30-40 lines of java code, we have derived column vector of single attributes for individual sensors. Each one is having approx. 50000 data samples. These vectors will be used as input dataset for density estimation and outlier detection components. The important features have been calculated in advance as it will be used further. We are describing below the table of features for sensor1 and sensor2 only.

Table: 1 Statistical characteristic for sensor1For sensor2

Dataset	Min	Max	Mean	Median	StdDev
Temperature	17.1954	122.1530	35.8824	22.1444	33.6511
Humidity	-4	50.7387	34.3193	38.6334	13.8804
Voltage	2.0065	2.7624	2.5196	2.5823	0.1690

Table 2: Statistical characteristics for sensor2

Dataset	Min	Max	Mean	Median	StdDev
Temperature	3.4068	122.1530	40.2018	22.4286	37.8656
Humidity	-3	50.5784	34.2987	40.1284	18.0268
Voltage	0.018	2.7244	2.4584	2.4850	0.1697

Moreover, we have implemented the following components:
(i) chain-sample, which maintains a running sample of the sensor readings in the window,
(ii) Variance estimator, which maintains a running estimate of the standard deviation of those values,
(iii) Kernel density estimator, which is used to approximate the data distributions,
(iv) DEVF based outlier detection algorithm.

We have applied the kernel density estimation techniques for sensor-1 and sensor-2. It approximated the density at various kernel points and given the points where we have to plot the density. We have plotted the density estimate for temperature, humidity and voltage for sensor-1.

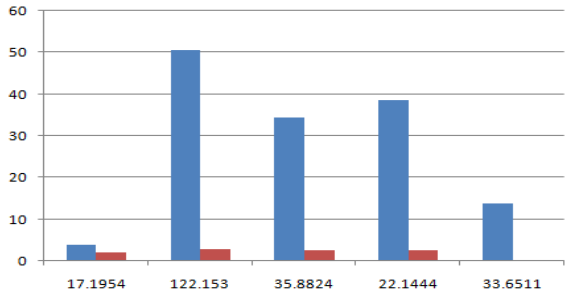

Figure. 1: performance evaluation based on the sensor1

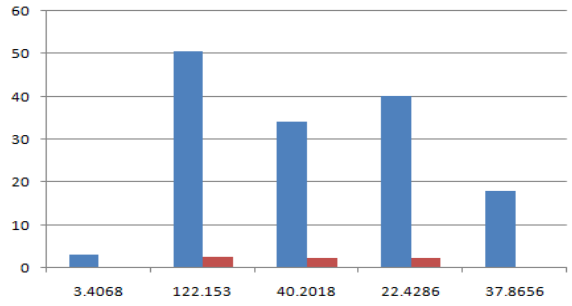

Figure. 2: performance evaluation based on the sensor2

CONCLUSION AND FUTURE WORK

In this work, we focused on the problem of outlier detection in wireless sensor networks. Outlier detection techniques generally focus the developer's or user's context into the interesting events, or unexpected results in the network which has very low probability of occurrence. Rather than working on the raw sensor readings at first, a statistical modeling technique transforms it into meaningful information which will yield effective output, hence offering a more reliable way to gain insight into the physical phenomena under observation. For the same, we have proposed a model that is based on the approximation of the sensor data distribution. Our approach takes into consideration various characteristics and features of streaming sensor data. We processed and evaluated our proposed scheme with a set of experiments with datasets which is taken from Intel Berkeley research lab. The experimental evaluation shows that our algorithm can achieve very high precision and recall rates for identifying outliers, and demonstrate the effectiveness of the proposed approach. As future work, we will be focusing on other density estimation techniques like orthogonal series expansion (wavelet density estimation). The basic idea of this method is to compute the distribution of measurement by estimating the coefficients of its fourier transform. Recent studies and works have shown that wavelet based density estimation techniques promises to be superior to others due to its local nature. Right now we are working for single attribute sensors but in future we will try to extend our method for multi-attribute sensors. And will focus on some other idea for outlier detection for multi-attribute sensors if required.

REFERENCES

[1]. A. Faradjian, J. Gehrke and P. Bonnet, "GADT: A Probability Space ADT for Representing and Querying the Physical World" ICDE, 2002.

[2]. Y. Rozanov, "Probability theory, random processes, and mathematical statistics," Kluwer Academic Publishers, 1995

[3]. C. Wright, "Applied Measurement Engineering: How to Design Effective Mechanical Measurement Systems," Prentice Hall, 1994

[4]. Elnahrawy, E., Nath, B.: Cleaning and Querying Noisy Sensors. In: Proc. of WSNA. (2003)

[5]. Liu, H., Hwang, S., Srivastava, J.: Probabilistic stream relational algebra: A data model for sensor data streams. Technical report, University of Minnesota (2004)

[6]. A. Mills, "Heat and mass transfer," Burr Ridge, 1995

[7]. Christoph Heinz Bernhard Seeger : Statistical Modeling of Sensor Data and its application to Outlier Detection in 2006

[8]. Fan Ye, Haiyun Luo, Jerry Cheng, Songwu Lu, and Lixia Zhang.A Two-Tier Data Dissemination Model for Large-Scale Wireless Sensor Networks. In MOBICOM, Atlanta, GA, USA, 2002.

[9]. S. Subramaniam T. Palpanas D. Papadopoulos, V. Kalogeraki, D. Gunopulos: Online Outlier Detection in Sensor Data Using Non-Parametric Models

[10]. Brian Babcock, Mayur Datar, and Rajeev Motwani. Sampling From a Moving Window Over Streaming Data. In SODA, 2002.

[11]. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall (1986)

[12]. Blohsfeld, B., Heinz, C., Seeger, B.: Maintaining Nonparametric Estimators over Data Streams. In: Proc. of BTW. (2005)

Short Bio Data for the Author

Mr. Vipnesh Jha is a M. Tech. student of Computer Science & Engineering at Teerthanker Mahaveer University, Moradabad. Presently he is working as lecturer at Department of CSE, Apex Institute of Technology, Rampur, (U.P), India

Sumit kumar srivastava is a M. Tech. student of Computer Science & Engineering at Teerthanker Mahaveer University, Moradabad. Presently he is working as Assistant professor. at Dev Bhoomi institution of Engineering & Technology Dehradun, Uttarakhand, INDIA.