Outlier Detection Techniques and Cleaning of Data for Wireless Sensor Networks: A Survey

[1]Vipnesh Jha, [2]Om Veer Singh Yadav

[1]Dept. of CSE, TMU Moradabad, UP, India

[2]Dept. of CSE, Apex Institute of Technology, Rampur, UP, India

Abstract

Pattern recognition is the scientific discipline where the goal is the classification of objects into a number of categories or classes. Pattern recognition is an integral part in most sensing networks built for outlier detection.

The significant deviations from the pattern of sensed data are considered as outliers in wireless sensor networks. These outliers include noise, errors, and malicious attack on the network. This affects the performance of the wireless sensor networks. Mostly the nature of sensor data is multivariate but it may be univariate also. Because of this, the traditional techniques are not directly applicable to wireless sensor networks. This contribution overviews existing outlier detection techniques developed for wireless sensor networks. It also presents a outlier detection technique framework to be used as a guideline to select a technique for outlier detection suitable for application based on the characteristics, such as, data type, outlier type and outlier degree.

Keyword

Wireless sensor networks, Pattern recognition, Outlier, Framework of outlier detection techniques.

I. Introduction

Wireless sensor systems enable fault tolerant monitoring and control of a variety of appli-cations. Due to the large number of sensor nodes that may be deployed and the long required system lifetimes, replacing the battery is not solution. Sensor systems must exploit the minimal possible energy while operating over a wide range of operating scenarios. This paper overviews the key technologies required for low-energy distributed microsensors. These include power responsive computation communication component technology, low-energy signaling and networking, system partitioning considering computation and communication trade-offs, and a power aware software infrastructure. In Wireless senor networks the use of sensors may result some errors or in many ways. These errors are known as outliers. The outliers affect the performance of the network in many ways like energy loss in sensing outliers. A wireless sensor network (WSN) typically consists of a large number of small, low-cost sensor nodes. The sensor nodes have sensing, processing and wireless communication capabilities. Each node is generally equipped with a small microcontroller, a wireless radio transceiver a power source and many sensors, such as, light, heat, pressure, temperature, humidity, sound and vibration. The WSN has a different area of applications, such as, in business purposes, in military, in environment, industries and some personal uses also. In many of these applications, real-time data mining of sensor data to promptly make intelligent decisions is essential (Ma et al., 2004).Data collected and measured by wireless sensor networks is normally untrustworthy and insecure. Errors, missing values, duplicate and replicate data, noise wrong entries and inconsistent data may affect the quality of the data and also the performance of WSN. The fluctuation of electricity plays a major role in energy loss of sensors. This may damage the sensors and whole network. The sensors have memory, computational capacity, energy and communication bandwidth. The limited resource and capability make data generated by sensor nodes untrustworthy and imprecise, especially when battery power is exhausted. Thus, the probability of generating erroneous data will grow rapidly (Subra-maniam et al., 2006). Conversely operations of sensor nodes are frequently vulnerable to environmental effects. The idea of large-scale and high density wireless sensor network is to randomly organize a large number of sensor nodes (up to hundreds or even thousands of nodes) in insensitive and unattended environments. It is predictable that in such environments some sensor nodes malfunction, which may result in noisy, faulty, missing and unnecessary data. Furthermore, sensor nodes are vulnerable to malicious attacks, such as, denial of service attacks, black hole attacks and eavesdropping (Perrig et al., 2004), in which data generation and processing will be manipulated by adversaries. These factors lead to unreliability of sensor data, which further influence quality and performance of raw data and aggregated results. Since in actual events it is extremely important to ensure the reliability and accuracy of sensor data because actual events like earthquakes, chemical spilling and fire cannot be accurately detected using inaccurate and incomplete data. Before the decision-making process. Outliers are the basic and main reason of influencing the quality and performance of data. This paper provides a comparison and classification of outlier detection techniques used in wireless sensor networks.

II. About Outliers

A. Outlier

An outlier is an observation of data that deviates from others observations so much that it arouses suspicious that was generated by a different mechanism. In clustering outliers are considered as noise, and error values. So we can say that outliers are those patterns that deviate from the normal pattern of sensed data. Outliers are of two types.

B. Local Outliers

An object in a data set is a local outlier if its density significantly deviates from the local area in which it occurs. Contextual outliers are a generalization of local outliers, a notion introduced in density-based outlier analysis approaches.

C. Global Outliers

Global outlier detection uses the whole data set as the context. Global outlier detection can be regarded as a special case of contextual outlier detection where the set of contextual attributes is empty. Contextual outlier analysis provides flexibility to users in that one can examine outlier in different contexts which can be highly desirable in many applications.

D. Need of outlier detection in WSN

Outlier is the very well known term in the field of data mining. It is used for finding errors, noise, missing values, inconsistent data, or duplicate data. These errors may affect the quality of data.

This results in the performance reduction of the system. The same happens with wireless sensor networks. Because WSNs is a network of sensors, the sensors have low cost and low energy. If the energy of sensors goes waste than the performance, of network lacks. So to improve the quality and performance, the outlier detection is very important in several real life applications, such as, environmental monitoring, health and medical monitoring, industrial monitoring, surveillance monitors and target tracking.

E. Problems in outlier detection in WSN

There are many problems in detection of outliers in WSNs. These can be summarized as follows:-

1. High communication cost
2. Modeling normal objects and outliers effectively
3. Application specific outlier detection
4. Identifying outlier source
5. Distributed data
6. Communication failures frequently
7. Dynamic network topology

III. Pattern Recognition And Its Process

Pattern recognition is the scientific discipline whose goal is the classification into a number of categories or classes,. Pattern recognition is an integral part in most machine intelligent system built for decision making. Pattern recognition is the discipline of building machines to perform perceptual tasks which human can perform easily like recognizing faces, voice, identify the spices of flowers and many more. Some pattern recognition tasks are everyday tasks, such as, speech recognition but some are not. A pattern is opposite of chaos; it is an entity vaguely defined, that could be given a name as finger print, handwritten word, human face, speech signal and many more.

Identification of pattern as a member of category (or class) we already know, or we are familiar with. It can be done in two ways.

1. Classification or categories to classify the objects in predefined classes.
2. Clustering (learning categories) to classify the objects in class those are not known. It is done by grouping objects.

A. Recognition Process

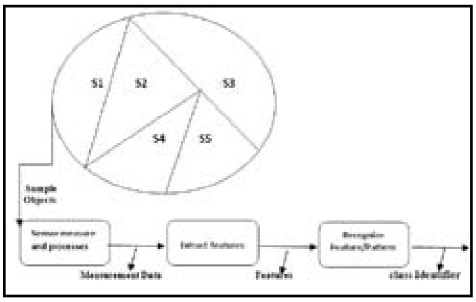


Fig. 1: The recognition/classification process

To recognize a sample object, consider the sub populations $S_1 \ldots S_4$ of a population P of non identical objects. An objects attributes are sensed or measured to give a pattern vector that is transformed into a reduced set of features, and object is recognized from its features by the recognizer. A feature extractor T transform the pattern vector m= $(m_1,m_2,\ldots\ldots,m_p)$ into a feature vector x=$(x_1,x_2,\ldots\ldots,x_n)$=T(m). A pattern recognizer is a system to which a feature vector is given as input and which operators on the feature vector to produce an output associated class to which the object belongs. Each individual object is an atomic class (i.e., has no sub classes), so that recognition includes identification.

Let P be the given population of non-identical objects where each object is represented an n-dimensional pattern vector of measurements. Suppose, P is partitioned into unknown equivalence classes $S_1,\ldots,S_k$. The classification problem is to decide whether or not multiple feature vectors belong to the same or different equivalence classes. The recognition problem is to decide whether or not any given sample feature vector is equivalent to a prototype or template vector that represents a class. This process of pattern recognition is concrete, i.e., it is based on measurements of physical attributes. On the other hand, abstracts pattern recognition is based upon the attributes and mental models. Here our concentration is more upon concrete rather than abstract.

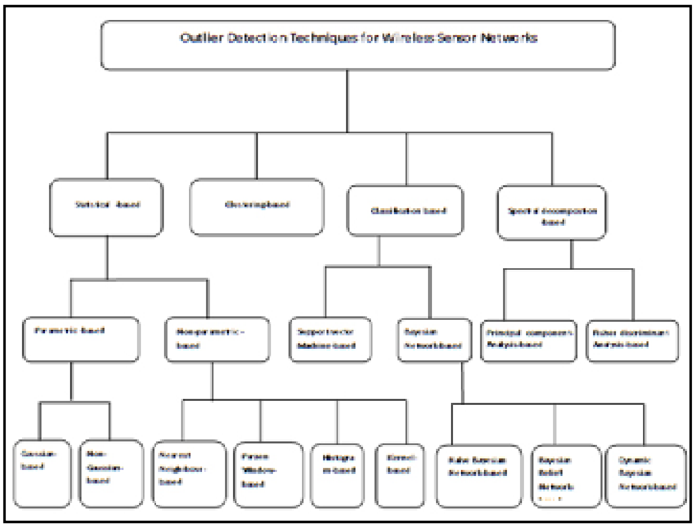

Outlier Detection Techniques for Wireless Sensor Networks
Statistical based
Clustering based
Classification based
Spectral decomposition based
Parametric based
Non-parametric-based
Support vector Machine based
Bayesian Network based
Principal component Analysis based
Fisher discriminant Analysis based
Gaussian based
Non-Gaussian-based
Nearest Neighbour-based
Parzen Window-based
Histogram based
Kernel based
Naive Bayesian Network based
Bayesian Belief Network based
Dynamic Bayesian Network based

Fig. 2 : Modified framework Zhang et.al.(2008) of outlier detection techniques for WSNs

IV. Techniques Of Outlier Detection For WSNs

This section provides classification of outlier detection techniques for WSNs based on discipline from which the ideas are taken. In this section, we also provide the short description of techniques used for outlier detection.

A. Statistical based approach

These approaches are the oldest approaches used for outlier detection. Statistical approaches are model-based technique. In this assumptions or estimation is based on probability distribution. This estimation captures the distribution of data and evaluate data instances with respect to how well they fit the model. Among the various frameworks in which pattern recognition has been traditionally formulated, the statistical approach is most extensively used. In this, each pattern is represented in terms of its features and goal is to identify the particular reason to which it belongs. In the statistical decision approach, the decision boundaries are determined by probability distribution of the patterns belonging to each class. Statistical based approaches are classified or categories into parametric and non-parametric based on how the probability distribution is built.

1. Parametric based approaches

In the parametric form of class-conditional type probability densities is given and then required parameters can be easily computed. This can be achieved by maximum likelihood method that finds the parameters value that is best supported by the data. Bayesian estimation is an alternative method can find the parameter by considering them as random variables with known prior density. Bayesian method is discussed for gaussian data. Parametric approaches are further categories into gaussian-based model and non-gaussian-based model.

(i). Gaussian based model

In gaussian-based model, the data are assumed to be normally distributed (Chandola et al. 2007).This technique relies on the spatial-temporal correlations of sensor data and uses two statistical tests to locally detect outliers. The drawback of this technique is that it only deals with one dimensional outlier data and too much memory is required for a node to store old values.

(ii). Non-Gaussian based model

In non-gaussian model the data are not normally distributed. This technique is based on symmetric α stable(S α S) distribution to find out outliers in form of impulsive noise. This technique uses spatial-temporal correlation of sensor data to locally detect outliers. It reduces the communication cost due to local transmission. The (S α S) distribution is not suitable for sensor data and cluster based-model. It is also susceptible to dynamic changes of network topology.

2. Non-parametric based approaches

Sometimes it is not necessary that we have known parametric family. So the parametric approach discussed above is not sufficient. In this technique the class conditional probability distributions are non-parametric. In this estimation p (x | ωi) is found to each point x. It is also possible to estimate the posterior probabilities P (ωi| x) directly. Non-parametric methods uses two kind of approaches: one is based on estimating the density first and then used for classification, second is based on choosing category directly. This approach is further categorized into four parts

(i). Nearest neighbor based

The nearest neighbor classification rule consist of finding k nearest training samples to test the point x and classifying x to the class which is most frequent in those k nearest neighbor of x. in the two class case, following is error estimate that holds in the case of unlimited number of training samples.

$E\ (_{\alpha bayes}) \le E(_{\alpha knn}) \le E\ (_{\alpha bayes}) + \sqrt{2}E\ (_{\alpha nn})/k$

From this error estimation, it states that this rule is optimal when $k\ \infty$.This technique is flexible in respect to multiple existing distance based outlier detection technique. This technique does not adopt any network structure so that every node uses broadcast to communicate with other sensor node.

(ii). Parzen Window Based

The parzen window is to estimate the densities by temporarily assuming that the region is a d-dimensional hypercube which encloses k-samples. The volume V_n of a hypercube, if h_n is the lengthof an edge of it,is

$V_n = h_n^d$

We can obtain an analytical expression for kn , the no. of samples falling in the hypercube, than the window function can be given as

$$\Psi(u) = \begin{cases} 1 & for\ |u_i| \le 1/2, i = 1,2,\ldots\ldots.d \\ 0 & \text{otherwise} \end{cases}$$

The problem with this technique is that it is also very costly for communication. This technique is different from nearest neighbour method in such a way that in parzen window method $V_n = 1/\sqrt{}_n$ while in k- nearest neighbor method $k_n = \sqrt{}_n$.

(ii). Histogram Based

This technique is used for identifying global outliers in the applications of sensor networks. In this technique, we use histogram data rather than collecting raw data for processing. In this method, the sink uses histogram information to extract data distribution from the network and filters out the non-outliers. Outliers can be identified by recollecting more histogram information from the network. The drawback of this technique is that the information of histogram is taken many times.

(iii). Kernel based

This technique is used for identification of outliers in streaming sensor data. This technique requires data distribution and uses kernel density function for the distribution of sensor data. A value is considered as an outlier if the no. of values being its neighborhood is less than a user specified threshold. The main problem of this technique is its high dependency on the defined threshold.

B. Clustering -based Approach

Clustering is a very well known term in data mining. Its object is to group data points into clusters so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Cluster analysis is the term applied to a group of analysis that seek to divide a set of objects into members of homogeneous group or clusters , when there is no prior information about the group . These techniques suffer from the choice of an appropriate parameter of cluster width. Hierarchical and partitioning clustering are basically used techniques of cluster analysis.

C. Classification-based Approach

Classification is also a very important term in data mining and machine learning. This technique also does not require the prior knowledge about the data set. Classification based approach is categories into two parts.

1. Support Vector Machine-based Approach

A SVM is a binary classifier. It abstracts a decision boundary in multi dimensional space using an appropriate sub set of the training set of vectors; the elements of this sub set are the support vectors. Geometrically, support vectors are those training patterns that are closest to the decision boundary.SVM are useful to understand several associated concepts including linear discriminant functions and neural networks. This technique identifies outliers from the data measurements collected after a long time window and is not performed in real time.

2. Bayesian Network Based approach

Bayesian network-based approaches are based on probabilistic analysis. This approach is categorized into three parts.

(i). Naive- Bayesian Network model

A naive bayes classifier is a simple probabilistic classifier based on applying bayes theorem where every feature is assumed to be class conditionally independent. Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variables. This assumption is called class conditional independence. It is made to simplify the computation and in his sense, it is considered to be naïve.

(ii). Bayesian Belief Network

A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic dependencies. Formerly, Bayesian network are directed acyclic graphs whose nodes represents variables and whose arcs encode conditional dependencies between the variables. This technique provides better results in comparison to naive bayes classifier.

(iii). Dynamic Bayesian Network Based approach

Bayesian networks that model sequences of variables (for example speech signal and protein sequences) are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams. This technique can handle several data streams at once.

D. Spectral Decomposition Based Approach

The aim of this approach is finding normal modes of behavior in the data by using principal components.
This approach is categorized in two parts.

1. Principle Component Analysis

PCA involves a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much as of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.PCA finds the most accurate data representation in a lower dimensional space. The data is projected in the direction of maximum variance. The data which violates this is considered as outliers.

2. Fisher Linear Discriminant Analysis Based

Fisher‘s linear Discriminant projects high-dimensional data onto a line and performs classification in this space. If there are two classes, the projection maximizes the distance between the means and minimizes the variance within each class. Fisher’s criterion which is maximized over all linear projection V can be defined as

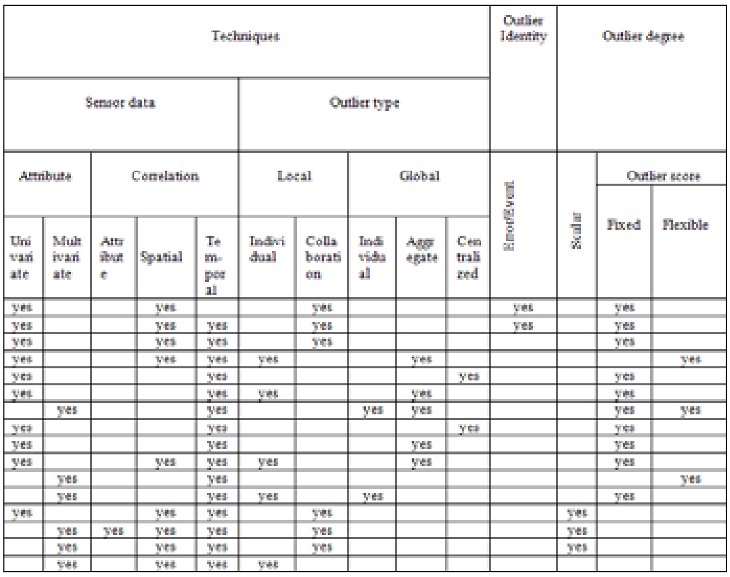
$$J(v) = \frac{|\mathrm{mean}_1 - \mathrm{mean}_2|}{S_1^2 + S_2^2}$$

Where mean1 and mean2 represent the mean of class 1 patterns and class2 patterns respectively, and S12 and S22 is propositional to variance. Like PCA, the data are protected in the directions of maximum variance. And the data which violates, are considered as outlier.

V. Conclusion

In this paper we discussed the problem of outliers in wireless sensor networks. A modified frame work of outlier detection technique is given and discussed. We also discussed a classification and comparison of outlier detection techniques based on some criterion. We discussed the drawback of each outlier detection technique, knowing these drawbacks we can easily overcome these shortcomings. We discussed about WSN and its characteristics. We can further work on each technique discussed in modified framework in future.

Table 1: Classification and Comparison of general outlier detection techniques for WSNs (Yang Zhang et al. 2008)



Techniques										Outlier Identity	Outlier degree		
Sensor data					Outlier type								
Attribute		Correlation			Local		Global			Error/Event	Scalar	Outlier score	
Univariate	Multivariate	Attribute	Spatial	Temporal	Individual	Collaboration	Individual	Aggregate	Centralized			Fixed	Flexible
yes			yes			yes				yes		yes	
yes			yes	yes		yes				yes		yes	
yes			yes	yes		yes						yes	
yes			yes	yes	yes			yes					yes
yes				yes					yes			yes	
yes				yes	yes			yes				yes	
	yes			yes			yes	yes				yes	yes
yes				yes					yes			yes	
yes				yes				yes				yes	
yes			yes	yes	yes			yes				yes	
	yes			yes									yes
	yes			yes	yes		yes					yes	
yes			yes	yes		yes					yes		
	yes	yes	yes	yes		yes					yes		
	yes		yes	yes		yes					yes		
	yes		yes	yes	yes								

References

[1] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., Cayirci, E., "Wireless sensor net-works: a survey", Computer Networks, Vol. 38, No. 4, pp. 393-422, 2002.

[2] Barnett, V., Lewis, T.,"Outliers in statistical data", New York: John Wiley Sons, 1994.

[3] Bhuse, V., Gupta, A., "Anomaly intrusion detection in wireless sensor networks", Journal of High Speed Networks, Vol. 15, No. 1, pp. 33-51, 2006.

[4] C.M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2009.

[5] Chandola, V., Banerjee, A., Kumar, V., "Outlier detection: a survey", Technical Report, University of Minnesota, 2007.

[6] Chen, J., Kher, S., Somani, A., "Distributed fault detection of wireless sensor networks",Proceedings of the 2006 workshop on dependability issues in wireless ad hoc networks and sensor networks, pp. 65-72, 2006.

[7] Ding, M., Chen, D., Xing, K., Cheng, X.,"Localized fault-tolerant event boundary detection in sensor networks", Proceedings of IEEE Conference of Computer and Communications Societies, pp. 902- 913, 2005.

[8] Gaber, M. M.,"Data Stream Processing in Sensor Networks", Springer, 2007.

[9] Han, J., Kamber, M.,"Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2006.

[10] Hawkins, D. M., "Identication of outliers", London: Chapman and Hall, 1980.

[11] Hodge, V. J., Austin, J.,"A survey of outlier detection methodologies', Articial Intelligence Review, Vol. 22, pp. 85-126, 2003.

[12] Jeffery, S. R., Alonso, G., Franklin, M. J., Hong, W., Widom, J.,"Declarative support for sensor data cleaning", International Conference on Pervasive Computing, pp. 83-100, 2006.

[13] Keinosuke Fukunaga,"Introduction to Statistical Pattern Recognition", Acadmic Press.

[14] Knorr, E., Ng, R.,"Algorithms for mining distance-based outliers in large data sets",International Journal of Very Large Data Bases, pp. 392-403, 1998.

[15] Krishnamachari, B., Iyengar, S.,"Distributed Bayesian algorithms for fault-tolerant event region detection in wireless sensor networks", IEEE Transactions on Computers, Vol. 53,No. 3, pp. 241- 250, 2004.

[16] Lazarevic, A., Ozgur, A., Ertoz, L., Srivastava, J., Kumar, V.,"A comparative study of anomaly detection schemes in network intrusion detection", SIAM Conference on Data Mining, 2003.

[17] Luo, X., Dong, M., Huang, Y., "On distributed fault-tolerant detection in wireless sensor networks", IEEE Transactions on Computers, Vol. 55, No. 1, pp. 58-70, 2006.

[18] Ma, X., Yang, D., Tang, S., Luo, Q., Zhang, D., Li, S. "Online mining in sensor networks", IFIP international conference on network and parallel computing, Vol. 3222, pp. 544-550, 2004.

[19] Markos, M., Singh, S.,"Novelty detection: a review-part 1: statistical approaches", Signal Processing, Vol. 83, pp. 2481-2497, 2003.

[20] Papadimitriou, S., Kitagawa, H., Gibbons, P. B., Faloutsos, C., "LOCI: fast outlier detection using the local correlation integral", International Conference on Data Engineering,pp. 315-326, 2003.

[21] Perrig, A., Stankovic, J., Wagner, D.,"Security in wireless sensor networks", CACM, Vol. 47, No. 6, pp. 53-57, 2004.

[22] Rajasegarar, S., Leckie, C., Palaniswami, M., Bezdek, J. C., "Distributed anomaly detection in wireless sensor networks", Proceedings of IEEE ICCS, 2006.

[23] Rajasegarar, S., Leckie, C., Palaniswami, M., Bezdek, J. C.,"Quarter sphere based distributed anomaly detection in wireless sensor networks", Proceedings of IEEE International Conference on Communications, pp. 3864-3869, 2007.

[24] Ramaswamy, S., Rastogi, R., Shim, K.,"E±cient algorithms for mining outliers from large data sets", ACM Special Interest Group on Management of Data, pp. 427-438, 2000.

[25] Richard Duda, Peter E. Hart, David G. Stock,"Pattern classification", 2nd edition, John Wiley, 2006.

[26] S. Theodoridis, K. Koutroumbas, "pattern Recognition", 4th edition, Academic Press.

[27] Yang Zhang, Nirvana Meratnia, Paul Havinga,"Outlier detection Techniques for wireless sensor networks: A survey", pp .11-20, 2008.

[28] Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., Gunopulos, D.,"Online outlier detection in sensor data using non-parametric models", Journal of Very Large Data Bases.

[29] Sun, P.,"Outlier detection in high dimensional, spatial and sequential data sets", Doctoral dissertation, University of Sydney, Sydney, 2006.

[30] Tan, P. N.,"Knowledge Discovery from Sensor Data", Sensors.

[31] Tan, P. N., Steinback. M., Kumar, V.,"Introduction to data mining', Addison Wesley, 2006.

[32] Zhang, K., Shi, S., Gao, H., Li, J.,"Unsupervised outlier detection in sensor networks using aggregation tree", Proceedings of ADMA, 2007.

[33] Zhang, Y., Meratnia, N., Havinga, P. J. M.,"A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets", Technical Report, University of Twente, 2007.

[34] Zhuang, Y., Chen, L.,"In-Network outlier cleaning for data collection in sensor networks",Proceedings of VLDB, 2006.

[35] Zoumboulakis, M., Roussos, G.,"Escalation: complex event detection in wireless sensor networks", Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 270-285.

Vipnesh Jha received his B.Tech degree in Computer Science and Engineering from Bharat Institute of Technology Meerut affiliated to UPTU Lucknow, India, in 2007 and pursuing M.Tech in Computer Science and Engineering from TMU Moradabad, India. He worked in Steria India Ltd as a Software Engineer. Currently he is working with Department of Computer Science and Engineering in Apex institute of technology as a lecturer. His research interests include Pattern Recognition, artificial Intelligence data mining and digital image processing. At present, He is engaged in Support vector Machines (SVM) based technique for outlier detection.

Om Veer Singh Yadav received his B.Tech degree in Computer Science & Engineering from Dehradun Institute of Technology, Dehradun affiliated to UPTU Lucknow, India in 2006 and pursuing M.Tech degree in Information Technology from Karnataka State Open University (KSOU). He worked as Teaching Personnel with Computer Science Department in College of Technology G.B. Pant University of Agriculture and Technology, Pantnagar. Currently he is working with Department of Computer Science & Engineering in Apex Institute of Technology as a lecturer. His research interests include Digital Image Processing, Pattern Recognition and Network Security. At present, He is engaged in Support Vector Machines (SVM) based technique for outlier detection.