International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-2, Issue-1, March 2013

Outlier Detection Techniques over Streaming Data in Data Mining: A Research Perspective

Prakash Chandore, Prashant Chatur

Abstract— Data mining is extensively studied field of research area; where most of the work is emphasized over knowledge discovery. Data stream mining is active research area of data mining. A data stream is a massive sequence of data elements continuously generated at a rapid rate. In streaming huge amount of data continuously inserted and queried such data has very large database. Streaming data analysis has recently attracted attention over data stream rather than mining large data sets in data mining community. Outlier Detection as branch of data mining has many applications in data stream analysis and requires more attention. Finding and removing outlier over data stream is very important aspect in data mining. Detecting outlier and analyzing data stream for large dataset we can consider two main groups where one group refers to data stream and data mining techniques and second group refers to different efficient algorithm to mine data stream. Detecting outliers and analyzing large data sets can lead to discovery of unexpected knowledge in area such as fraud detection, telecommunication, web logs, and web document and click stream, etc. In this paper we try to clarify problem with detecting outlier over Dynamic data stream and specific techniques used for detecting outlier over streaming data in data mining.

Index Terms—about four key words or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

Data mining is extensively studied field of research area; Extraction of interesting non-trivial, hidden and potentially useful patterns or knowledge from huge amount of data where most of the work is emphasized over knowledge discovery. With the advances and proliferation of information technologies and applications number of databases, as well as their dimension and complexity grows rapidly. However, there are a lot of problem exists in huge database such as data redundancy, missing data, skewed data, erroneous data etc. One of the major problem in data mining research increase in dimensionality of data gives rise to a number of new computational challenge not only due to increase in number of attributes.

In recent years, we have observed that enormous research activity actuated by the explosion of data collected and transferred in the format of streams. Outlier detection in data streams can be useful in many fields such as analysis and monitoring of network traffic data, web log, sensor networks and financial transactions, web click streams etc.

Outlier Detection over streaming data is active research area from data mining that aims to detect object which have different behavior, exceptional than normal object [1]. Depending upon different application domains these abnormal patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants. Identifying outliers within data lead to the discovery of useful and meaningful knowledge or improve data analysis for further discover within numerous applications domains, it also helps to avoid a wrong conclusion.

A lot of work outlier for detection has been done in data mining community using conventional methods. These conventional methods are more suitable over static data set. Such methods can be used for streaming data but these methods are not able to process data with single pass. As dimensionality increase traditional method takes high computing time and cannot provide an efficient result over analysis of streaming data.

Effective outlier detection requires the construction of a model that accurately represents the data. Over the years, a large number of techniques have been developed for building such models for outlier and anomaly detection. However, real world data sets, data stream and environments present a range of difficulties that limit the effectiveness of these techniques it also depend on domain.

The contribution of this review can be organized as follows. Section II provides a review of outlier detection techniques. Section III discuss about categorical classification of techniques. Section IV Advances in outlier detection technique is provided. Section V discusses about the drawbacks of some outlier detection techniques. Finally Section VI provides a conclusion for reviewed outlier detection techniques.

II. FUNDAMENTALS OUTLIER DETECTION OVER STREAMING DATA

This section depicts fundamentals of outlier detection over streaming data in data mining, including definitions of outliers, motivation of outlier detection, and challenges of outlier detection over streaming data with data mining perspective.

A. Outlier

Outlier is also called as anomaly due to behavior of object with respect to other data elements. Term outlier originates from Statistics [2]. Numerous definitions have been proposed for outlier in data mining a variety of definitions depending on the particular method outlier detection techniques are based upon to solutions identify outliers in a specific type of data set exist [5]. We consider following two definitions as a classical definitions for outlier in data mining perspective. Fist definition as *"An outlier is an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"*[3]. Second definition proposed by Barnett and Lewis that *"an outlier is an observation which appears to be inconsistent with the remainder of that set of data"* [4].

Manuscript received on March, 2013.

Prakash Chandore, Department of Computer Science & Engineering, Govt. College of Engineering Amravati , Amravati, India.

Dr. Prashant Chatur , Department of Computer Science & Engineering, Govt. College of Engineering Amravati , Amravati, India.

B. Motivation of Outlier Detection in Streaming Data

Data mining, predictive modeling, cluster analysis and association analysis these are prior fundamental task of data mining [6]. In comparison with fundamental task of data mining outlier detection is can be considered fundamental task in data analysis i.e., mining useful and interesting information from a large amount of data [1].

More ever due to rapid stream evolution, data element property can change over time, here we aren't able to visit data second time. Data stream is likely to river, it means continuous data flow in and out. This introduces to main problem first as only one scan is possible to process data points and secondly due time constraints data is considered as an evolutionary stream. In order to deal with problem of processing streaming data efficient outlier detection method need to be used. Also it required more attention from data mining community.

The Constraints of dataset and the nature of data make design of an appropriate outlier detection technique more challenging. Traditional outlier detection techniques might not be suitable for handing dynamic nature of data. Some of the following aspects are shown in table which motivates us that we require efficient outlier detection method over streaming data.

TABLE I
PARAMETRIC CONTEXT OF STREAMING DATA.

Parameter	Context
Resource constraints.	Dynamic nature of data causese traditional methods to have a high computation cost for evaluation of outlier. Data is passing with a time constraints in distributed enviourment only one pass of scan is possible so we required much memory for data analysis and storage. So it needed that Minimize the time consumption while using a reasonable amount of memory for storage and computational tasks.
High Computation Cost	Earlier work for outlier detection is carried out mostly based on statistical methods that require a high computational cost for overall data stream.
Distributed streaming data.	One of the major problem exisitng approach that handling dynmic change in data where it diffcult to identify prior distribution of data steram. Direct computation of probabilities is difficult [29].Do not meet the requirement of handling distributed stream data.
Identifying outlier sources	Difficult to identify what has caused an outlier in streaming data due to the resource constraints and dynamic nature of data. Finding source of outliers within data is one of the diifcult task and it varies with respective to application domain.
Uncertain data or missing data	If prior distibutuion is not known then it is diffcult to formulate outlier detection model. Due to misssing valuses we may lead to wrong decision.
High Dimensional Data Stream	Due to high dimensional proerty we required a huge storage space and also processing complexity is high.

C. Challenges of Outlier Detection in Streaming Data

Mining useful and interesting information from a large amount of data is prior task in data analysis [1]. Designing an appropriate outlier detection technique is more challenging with respective context of streaming data and nature of data. With reference to following challenges traditional outlier detection methods might be not able to detect outliers over streaming data.

- *Distributed Streaming Data.*

Data coming from distributed environment may dynamically change. In such situation distribution of streaming data may not known. Due to dynamic nature direct computation of probability is difficult [7]. Traditional methods are offline in manner do not able to handle distributed streaming data.

- *Massive Data Processing.*

Data stream are having massive amount of data. Due to property of dynamic change of data it difficult to process data within single scan. In such situation prior challenge to traditional outlier detection is to provide a high detection rate. It is observed that traditional outlier detection techniques do not scale well to process large amount of distributed data streams in an online manner.

- *Limited Computation Resources.*

In many application domain, high computation power and the other measuring property such consumption of available memory at hand does not measure up the massive amount of streaming data. For example in single day Google provide search for 150 million query search, Telstra generated 15 million call records. Traditional outlier detection methods are not able to handle such requirement for streaming data. Stream mining algorithms shall learn fast and consume little memory resources.

- *Uncertain and Missing Data.*

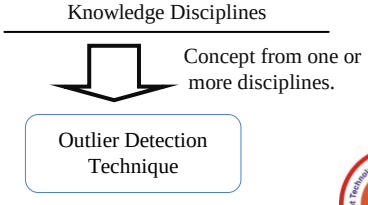
In most application domain that we don't have a sufficient data for operations. Such situation arouse due to uncertain and missing data. If we don't have sufficient information for data that may lead us for wrong decision. We required a method to manage such uncertain data and missing data within data stream.

- *High Dimensional Data Stream.*

High dimensional data stream contains a tremendous huge amount of data. Such massive amount data contains a large data with high dimensions with data complexity. For example wireless sensor network data, web logs, Google search, etc. Traditional methods are not suitable over high dimensional data as they required very high computation cost for processing data.

III. OUTLIER DETECTION OVER STREAMING DATA

Outlier detection is a primary step in many data-mining applications. It refers to the problem of finding patterns in data that do not conform to expected normal behavior or anomalous behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Following figure shows that it depend over application domain.



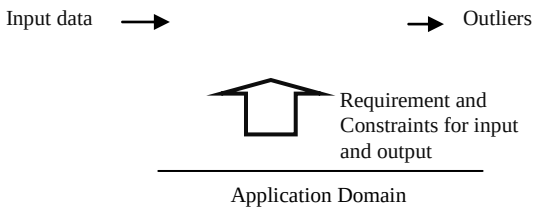


Fig. 1 A general design of an outlier detection technique.

As illustrated in Figure 1, any outlier detection technique has following major ingredients.

1. Nature of data, nature of outliers, and other constraints and assumptions that collectively constitute the problem formulation.
2. Application domain in which the technique is applied. Some of the techniques are developed in a more generic fashion others directly target a particular application domain.
3. The concept and ideas can be applied from one or more knowledge domains.

In data mining outlier detection has been extensively studied in past decade. Still most of the existing research work has been focused over static data set. Detecting outliers over data stream is active research area in recent years. Data are continuously coming in a streaming environment with a very fast rate and changing data distribution (change of data distribution is known as concept drift) [8], and thus, any fixed data distribution is not adequate to capture the dynamic behavior of data streams.

A. Evaluation of the Outlier:

Evaluation of detected outlier is an important task in the data analysis from data mining community. Numerous measures have been proposed for outlier evaluation such as *detection rate, false alarm rate and ROC curves.* No of outliers which are correctly classified such meaningful information provided by detection rate while no of outliers are misclassified with respective normal data is called a false alarm rate.

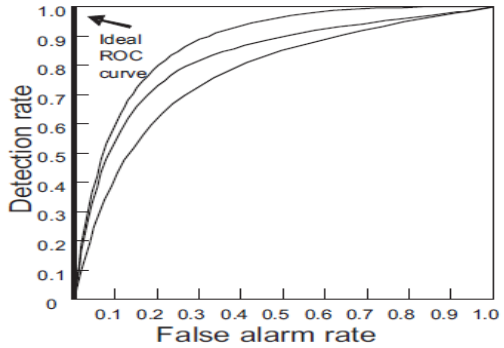


Fig. 2. ROC curves for different detection techniques [9].

A receiver operating characteristic (ROC) curves [9] most widely used measure to represent the trade-off between the detection rate and false alarm rate. Corresponding outlier detection technique measured as larger the area under ROC curve the performance is better as shown in fig 2. Outlier detection technique is measured as efficient if it is low computation cost which is also known as a time and space complexity.

IV. VARIOUS OUTLIER DETECTION TECHNIQUES

Outlier detection is varies in accordance with different entities in different domains. Outlier detection terminology refers to task of finding outliers as per behavior of data and distribution of data. Specific solution is formulated for outliers with reference to above point such formulation of the outlier detection is depend upon various factors such as input data type and distribution, availability of data and resource constraints introduced by application domain. Following outlier detection techniques widely used over streaming data.

A. Statistical Outlier Detection.

Statistical outlier detection techniques formulate the model using distribution of data point available for processing. Detection model is formulated to fit the data with reference to distribution of data. In this approach outliers are detected by using a standard probability distribution to fit the model or depth based approach also used for outlier. A Gaussian mixture model was proposed by Yamanishi et. al.[10]. Where each data point is given a formulated score and data point which have a high score declared as outlier. Detecting outlier based on the general pattern within data points was proposed by [11] where it combines a Gaussian mixture model and supervised method. This approach is can also referred as parametric approach.

Depth based outlier detection [12] is one the variant of statistical outlier detection where each data object of dataset represented by a n-d space having a assigned depth. These data points are organized into convex hull layers according to assigned depth and outlier is formulated on the basis of shallow depth values.

However these techniques are generally suited to quantitative real-valued data sets or quantitative ordinal data distributions.

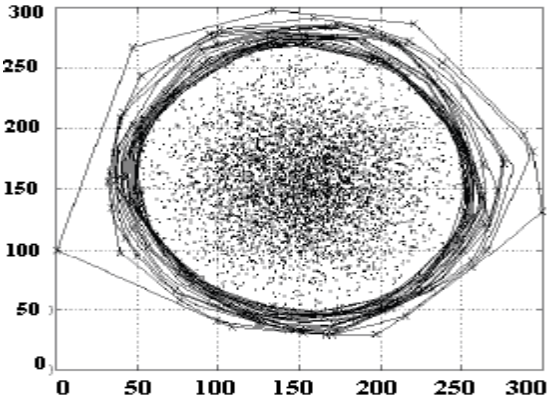


Fig. 3 data objects in convex hull layer. [58]

These models are not suitable for high dimensional data set. As dimensionality of dataset increases convex hull getting harder to make out such problem is recognized as “Curse of Dimensionality”.

B. Distance Based Outlier Detection.

Currently, so-called distance-based methods or outlier detection, as typical *non-parametric* methods identify outliers based on the measure of full dimensional distance between a point and its nearest neighbor in the data set. Basic model of distance based outlier detection as shown in figure.

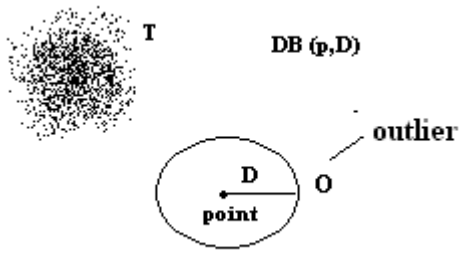


Fig. 4 Basic model of distance based method.

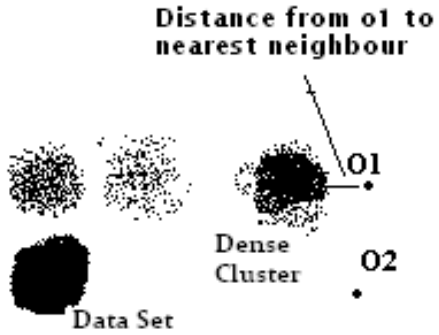


Fig. 5. Data set and dense cluster with outliers.

Earlier distance based method was introduced in [13] where outlier is detected as "*An object O in a dataset T is a DB (p,D)-outlier if at least fraction p of the objects in T lies at a distance greater than D from O"*. Where parameter p is threshold value a constraint on outlier from normal data points.

This approach is further extended in [14] with prior consideration on distance of a point from its k the nearest neighbor. Where top k point are declared as a outliers. This approach alternatively proposed by Angiulli and Pizzuti [15] on the basis of outlier factor. Each data point is assigned formulated outlier factor computed as sun of distance from its k nearest neighbors. Linear time is used for detecting outliers in [16] where data set get randomized for efficient search space. Recently we witnessed that a non parametric unsupervised based methods used for outlier detection which was proposed by a branch et al [17]. To address the uncertainty, temporal relation and transiency present within data distance based outlier detection for data stream method proposed (DBOD-DS) with the help of continuously adaptive data distribution function [20].

C. Deviation Based Outlier Detection.

Data set in which data elements are scattered as like a sparse matrix, such scarcity within data creates confusion over data analysis. Due to scattered form of data points are get deviated from normal points such points are declared as outliers. Sequential problem approach was proposed in [18] where outliers are identified by using normal features of data points and deviated features of data. To deal with time series constraint oriented data, Jagadish et al proposed a histogram based approach [19]. While considering this method not suitable for streaming data. So we observed that finding deviates over streaming data in distributed environment and over multivariate data is left as open.

D. Density Based Outlier Detection

This method uses density distribution of data points within data set. To decide outlier using density based a data point is taken into consideration where its density is compared with neighbor's density. This comparison is considered as score for outlier detection. Data points which are having a low density are considered as an outlier. The idea of density based local outlier using comparison with density of local neighborhood was introduced by Breuing et al. [24]. In this approach an outliers are measured by using a local outlier factor (LOF), which is ratio of local density of this point and the local density of its nearest neighbor. Data point whose LOF value is high is declared as outlier.

Papadimitriouet et al [30] proposed a Local correlation Integral (LOCI) based method which uses Multi Granularity Deviation Factor (MDEF) as a measure that how the neighborhood count of a particular data element compares with that of the values in its sampling neighborhood.

E. Clustering Based Outlier Detection.

Cluster analysis is popular unsupervised techniques to group similar data instances into clusters. The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. The assumed behavior of outliers is that they either do not belong to any cluster, or belong to very small clusters, or are forced to belong a cluster where they are very different from other members [21]**.** Concept local based outlier proposed using cluster based approach Z. He et al. [22]. A mining based method used for detecting outliers over categorical data using hyper graph model [23]. Previous algorithms were used for a single dimensional data. Aggarwal and Yu [24] proposed a approach for high dimensional data but it has a high computational cost.

F. Sliding Window Based Outlier Detection.

No of outlier detection techniques are get proposed over streaming data which uses a sliding window based concept. Where different multi pass algorithms are used for detecting outliers within streaming data [25][26]. This method is not efficient because some outliers may be considered as a inliers in other window. Major problem is that sometimes outlier point may get classified as a inlier [26]. Overall view regarding with the sliding window based outlier detection that we required to choose window size accurately. Choice of sliding window is independent on data point used for implementation which gives poor result over outlier detection.

G. Auto Regression Based Outlier Detection.

This Technique is mostly used for outlier detection over time series data [4]. Auto-regression is also adopted for some outlier detection over streaming data [27], [28]. Outlier is detected using an estimated model and metric which computed based on comparisons. If measured metric is crossing the limit or not fall within a cutoff limit then it is declared as an outlier. Most of the auto regression based techniques use same above methodology. Overall efficiency of such methods depends upon the model chosen and cutoff limit. Streaming data are dynamic in nature where data pattern frequently changes, so it is difficult to select an appropriate model for data streams [29].

V. DISCUSSION

Effective outlier detection requires the construction of a model that accurately represents the data. Over the years, a large number of techniques have been developed for building such models for outlier and anomaly detection. To present effectiveness for outlier detection that should able to handling following weakness with respective outlier detection technique.

A. Outlier Detection Approaches.

Statistical Outlier Detection: Statistical method of construct data distribution model. Based on model it declares point as outlier. But it observed that it is applicable for single dimension. Parametric assumption also does not hold good on distributional data set.

Depth based Outlier Detection: This method is belongs to statistical outlier detection but it is independent of data distribution. Here data points are organized in convex layers as shown in figure 3 which causes curse of dimensionality.

Distance based Outlier Detection: This approach use a notion of distance of data point within data distribution and by using threshold value it decides outlier within data .It suffers from detecting a local outlier within multi categorical data or diverse density data. It suffers from high computational cost for high dimensional data set.

Density based Outlier Detection: This approach uses a density estimation of data and the data element which has alow density is declared as outlier. Selecting boundaries for outlier is difficult task. We require lower bound and upper bound for selection of parameter.

Cluster based Outlier Detection: This approach uses a cluster based technique for detecting outlier where it finds closely related objects. Object which does not belong to any cluster or belongs to a small cluster is declared as outlier. A major limitation of clustering-based approaches to outlier detection is that they require multiple passes to process the data set. Outlier detection also highly depends upon type of clustering used.

Sliding window based Outlier Detection: This method uses a sliding window for detecting outlier with the help of multi pass algorithm. One of problem with this method is to select sliding window properly. It does not capture all data element within a data stream which also causes a poor result.

Auto-Regression based Outlier Detection: This technique uses a similar approach like statistical method. It formulates a model based on data distribution and uses a measure to declare a data point as outlier. Efficiency of this method depends upon the model and measured limit used for outlier detection. While it less efficient when data distribution is dynamic in nature.

VI. CONCLUSION

In This paper we provide a review of outlier detection methods over streaming data with data mining perspective. Based on review we conclude that most of the outlier detection research focuses over algorithms which require a special background and notion of finding outlier also varies from domain to domain. We observed that efficiency of outlier detection method is highly dependent upon data distribution and type data. For instance that statistical technique uses a data distribution and model, while some techniques require a prior knowledge about data. We observed that assumption based method can work quite well if prior assumption made about data is correct. Detecting Outliers over streaming data is important research problem in data mining community. Detecting outlier is important because it contains useful information which may lead for further research in domain. We observed that individual methods are not efficient over streaming data. In such case if prior information about data is not known then better to use combine approach for outlier detection.

REFERENCES

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

[2] V. Hodge and J. Austin, A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, Vol. 22, pp. 85-126, 2003

[3] D.M. Hawkins, *Identification of Outliers*, London: Chapman and Hall, 1980.

[4] V. Barnett and T. Lewis, *Outliers in Statistical Data*, New York: John Wiley Sons, 1994.

[5] Y. Zhang, N. Meratnia, and P.J.M. Havinga, A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets,*Technical Report*, University of Twente, 2007.

[6] P.N. Tan, M. Steinback, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.

[7] M. M. Gaber, Data Stream Processing in Sensor N etworks. In J. Gama and M. M. Gaber, *Learning from Data Streams Processing Techniques in Sensor Network*, pp. 41-48. Springer Berlin Heidelberg, 2007.

[8] Jiang, N. and Gruenwald, L. 2006. Research issues in Data Stream Association Rule Mining. ACM SIGMOD RECORD, Volume 35, Issue 1. Pages 14 -19.

[9] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, AComparative Study of Anomaly Detection Schemes in Network Intrusion Detection, *SIAM Conference on Data Mining*, 2003.

[10] K. Yamanishi et al, 2004. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *In Proceedings of Data Min. Knowledge Discovery.* Vol. 8, No. 3, pp 275-300.

[11] K. Yamanishi and J. Takeuchi, 2001. Discovering outlier filtering rules from unlabeled data-combining a supervised learner with an unsupervised learner. *In Proceedings of KDD'01*, pp 389-394

[12] R. Nuts and P. Rousseeuw, 1996. Computing depth contours of bivariate point clouds. *Computational Statistics andData Analysis*, Vol 23, No 2, pp 153-168.

[13] Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, pp.211-222, September 1999.

[14] Ramaswamy S., Rastogi R., Kyuseok S.:Efficient Algorithms for Mining Outliers from Large Data Sets,*Proc. ACM SIDMOD Int. Conf. on Management of Data, 2000.*

[15] F. Angiulli and C. Pizzuti, 2002. Fast outlier detection in high dimensional spaces. In Procedings of PKDD'02, 2002.

[16] Bay S. D. and Schwabacher M., 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp 29-38.

[17] J. W. Branch et al, 2006, In-network outlier detection in wireless sensor networks, In 26th IEEE International Conference on Distributed Computing Systems (ICDCS'06), pp 49.

[18] H. V. Jagadish et al, 1999. Mining Deviants in a Time Series Database. *In Proceedings of 25 international Conference on Very Large Data Bases.* Edinburgh, Scotland, pp 102-113.

[19] H. V. Jagadish et al, 1999. Mining Deviants in a Time Series Database. *In Proceedings of 25th International Conference on Very Large Data Bases.* Edinburgh, Scotland, pp 102-113

[20] Sadik, S. and Gruenwald, L. 2010. DBOD-DS: Distance Based Outlier Detection for Data Stream. DEXA' 10.

[21] M. F. Jiang et al, 2001. Two-phase clustering process for outlier detection. *Pattern Recognition Letters*. Vol 22, No.6-7, pp 691-700.

[22] Z. He et al, 2003. Discovering cluster based local outliers. *Pattern Recognition Letters*. Vol 24, No. 9-10, pp 1641-1650.

[23] Wei et.al, 2002, Outlier detection integrating semantic knowledge. *In Proceedings of Third international Conference on Advances in*

Web-Age information Management. Lecture Notes in Computer Science, Springer- Verlag, London, Vol. 2419. pp 126-131.

[24] C. C. Aggarwal and P. S. Yu., 2001. Outlier detection for high dimensional data. In *Proc. 2001 ACM-SIGMOD Int.Conf. Management of Data (SIGMOD'01)*, pp37-46.

[25] Anguiulli, F. and Fassetti, F. 2007. Detecting Distance-Based Outliers in Streams of Data. CIKM' 07. Pages 811 - 820.

[26] Basu, S. and Meckesheimer, M. 2007. Automatic outlier detection for time series: an application to sensor data. Knowledge Information System. Pages 137 – 154.

[27] Curiac, D., Banias O., Dragan F., Volosencu C., and Dranga O. 2007. Malicious Node Detection in Wireless Sensor Networks Using an Autoregression Technique. ICNS' 07. Pages 83 – 88.

[28] Puttagunta, V. and Kalpakis, K. 2002. Adaptive Methods for Activity Monitoring of Streaming Data. ICMLA' 02, Pages 197-203.

[29] M. M. Gaber, Data Stream Processing in Sensor Networks. In J. Gamaand M. M. Gaber, *Learning from Data Streams Processing Techniquesin Sensor Network*, pp. 41-48. Springer Berlin Heidelberg, 2007.

Prakash Chandore has received his B. Tech degree in Computer Science and Engineering from Shri Guru Gobind Singhji Institute Of Engineering and Technology, Nanded, Maharashtra, India in 2010. Currently pursuing M.Tech with Computer Science and Engineering stream from Govt. College of Engineering Amravati, Maharashtra, India. His area of research includes Distributed Data mining, Data stream mining, Outlier Detection, High Performance Computing. At present he is working with Outlier Detection in High Dimensional Data Stream.

Dr. P N Chatur has received his M.E. degree in Electronics Engineering from Govt. College of Engineering Amravati, India and Ph.D degree from Amravati University. He Has Published twenty papers in national Conferences and Ten papers in international journals. His area of research includes Artificial Neural Network, Data Mining, Data Stream Mining and Cloud computing. Currently he is Head of Computer Science and Engineering Department at Govt. College of Engineering Amravati, Maharashtra India. At present he is engaged with large database mining analysis and stream mining.