DETECTION OF ANOMALIES FROM DATASET USING DISTRIBUTED METHODS

S. E. Pawar and Agwan Priyanka R.
Dept. of I.T., University of Pune, Sangamner, Maharashtra, India
M.E. I.T., Dept. of I.T., University of Pune, Sangamner, Maharashtra, India

ABSTRACT

Detection of anomalies has been an ergonomic way to attention in many application and research area. This work includes with the detection of network intrusion, malware detection, biological data, image processing etc. The related researches were based on pair wise distance among all points in dataset. This approach is based on the concept of anomaly detection solving set, which is a small subset of the data set that can be also used for predicting unusual anomalies too. The algorithm-work with the parallel computation in order to meet two basic requirements-

1] The decrement in the execution time with respect to the centralized version.

2] Providing ability to handle distributed dataset.

Co-operating parallel task is a resultant of the overall computation, which is the formal goal to be achieved rather than presenting the correctness of the result the detection of the anomaly gives high performance in the propose work. The previous result shows that the as the no. of nodes increases the execution time increases. The later goal is completed to the execution of parallel task to using only on a portion of the entire dataset, so that the proposed schema is been used over distributed dataset. While solving the distance based anomaly detection task in the distributed scenario the main vitae of the proposed schema is the computation of anomaly detection solving set of the overall data set of the similar quality from those computed by the corresponding centralized schema. In this paper the techniques used are parallel and distributed mining.

KEYWORDS: *Data Mining, Parallel and Distributed Data Mining, Anomaly Detection.*

I. INTRODUCTION

Anomaly detection is also known as outlier detection or anomaly detection. Anomaly detection is the similar task to data mining whose main aim is to separate observation which are to a greater extend dissimilar from the remaining data. The proposed task has many practical applicative base in several areas such as fraud detection, intrusion detection, image processing, data cleaning and many other [1]. An outlier or anomaly is an observation which deviates to much from the other observation as to arouse suspicions that it was generated by a different mechanisms. Anomaly detection helps to isolate each information as normalized or exceptional. In the distance based approach distance to its nearest neighbor provides the outcome of an object as an anomaly. Distance based approach differ with the distance measure as it is been define, generally, consider an object with dataset linked with the weight or score which is belief based or a function of its kth nearest neighbor distance. This gradually increases the discrimination of the object from its neighbour[1].

Top nth outlier based anomaly in dataset is defined as an object associated with weight equal to and not smaller than the nth largest weight. Have the weight of the dataset object is the calculate sum of the distance from the given object to its k-nearest neighbour[1].

The extended work of anomaly detection leads to spatial detection. Spatial anomaly detection are those detection which seems to be inconsistent with respect to their neighbors but not differ from the entire crowd. Detecting spatial anomaly has an applicative domain such as in transportation, ecology, climatology and in location based services. Here the considered spatial dataset is a collection of objects as roads, houses, traffic sensors .Attributes changes according to the anomalies like here in spatial anomaly detection the attributes are the dimensions which are of two types spatial and the other is non-spatial. Spatial attributes include location, shape and other geometric properties. Spatial neighborhoods are determined according to the distance or adjacency. Non spatial attributes include traffic-sensors, manufacturer, owner, age and measurements readings[7].

II. ORGANIZATION OF MANUSCRIPT

Section 1 gives the Introduction of the Distributed Strategies for Mining Outliers in Large Data sets system architecture: It gives brief introduction about need of the system, architecture of the topic which measures the performance.
Section 3 gives the information about the basic concepts which are require for the main proposed system. It gives detailed introduction about the Data mining, Parallel and Distributed data mining etc. In the 4th section gives Some important concepts that we are using to analysis of existing system and comparison of existing system and proposed system. Section 5 gives the brief idea about the proposed system.

III. BACKGROUND

Data mining is a process of extraction of previously unknown and useful information [such as some pattern, constraints, knowledge rules and regulates] from large data base stored in the different format data can be store in image, file, video, audio, tabular format. Data mining is also isolation of knowledgeable information from the data base. It also refers to extraction of hidden predictive information from large data base. Itisnan intelligent method to extract the data patterns [5].

Data mining requires access to data.
1] Data can be access from
2] Data can be access from data ware house
3] Data can be access from flat file/spreadsheet

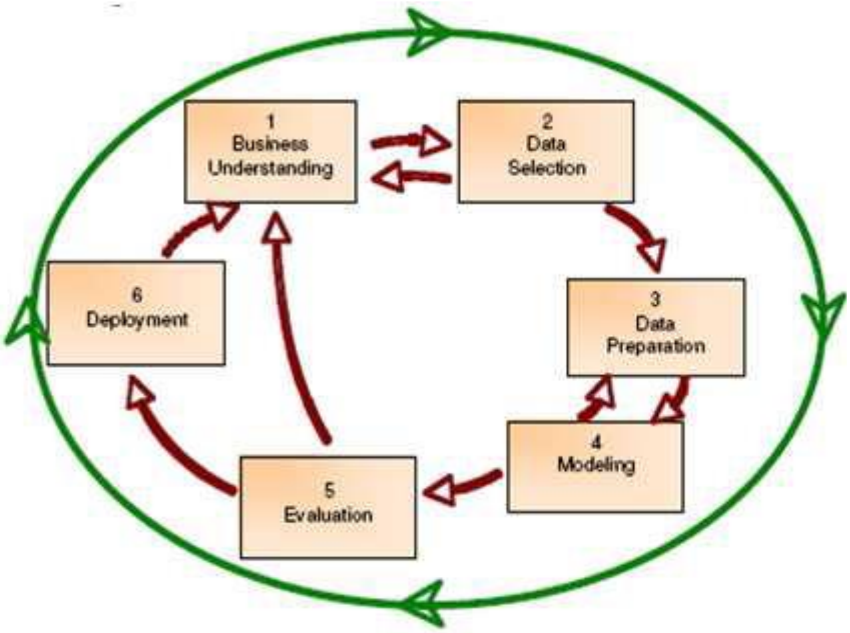


Fig1. Life cycle of Data Mining

There are 6 stages in the Life cycle of data mining
Stage 1- Business understanding- It focuses on identifying the problem we are trying to solve by using the DM. In this stage the business objectives are defined.
Stage 2-Data selection- There is need of the collecting all the data together needed for the data mining process by data load process from the particular dataset to the data ware house (DW).

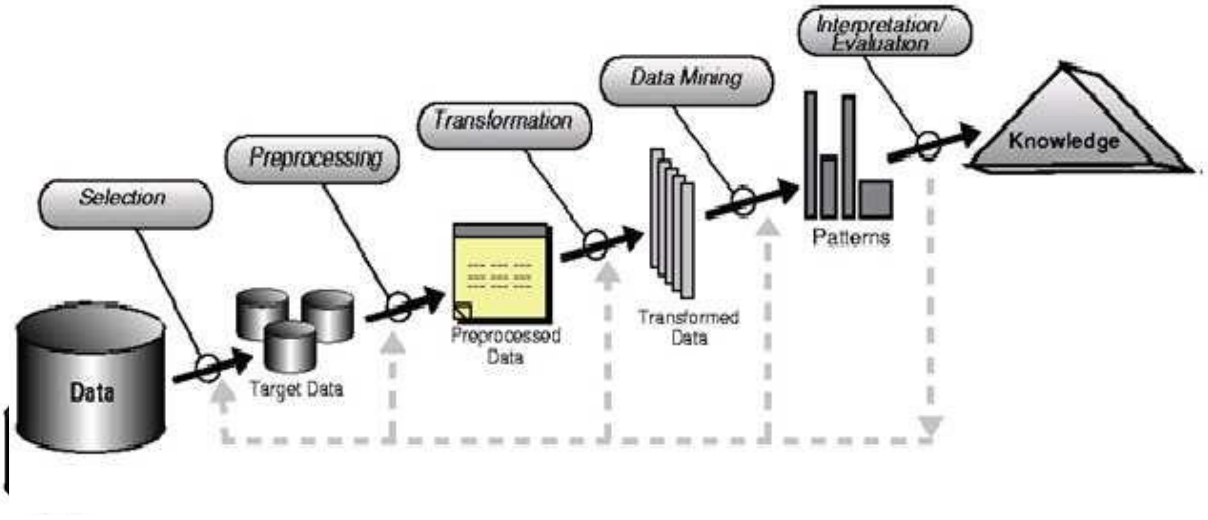
Stage 3-Data Preparation-Data preparation involves the construction of new attributes, grouping continuous data in ranges and the data is transformed in the same format or in the same kind of range the data placed.

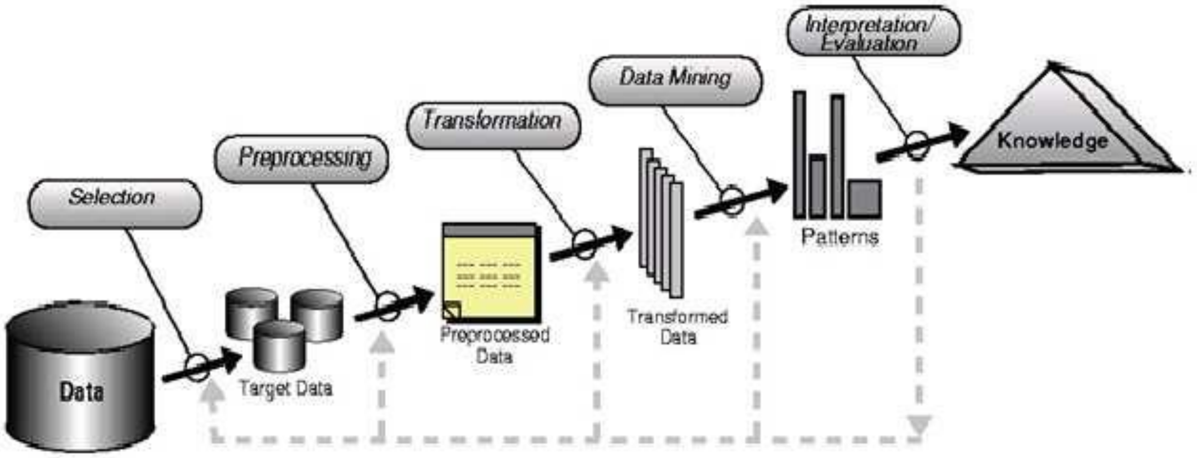
Stage 4-Modeling-It provides data visualization capabilities to discover patterns in any collection of information. In this stage the data gets mined. Here actually data mining process ends and data is ready to use.

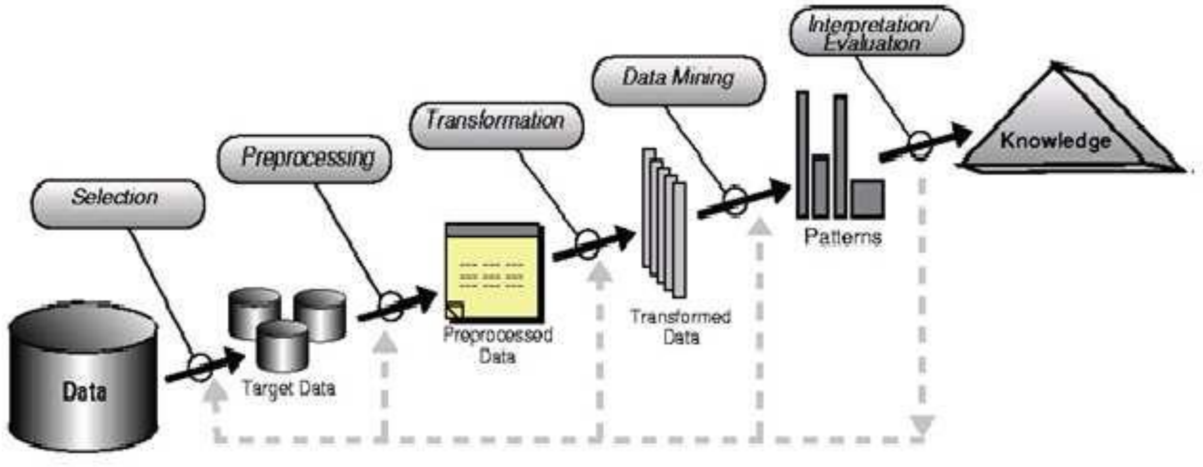
Stage 5-Evaluation- It helps in understanding and analyzing business complexities and its nature so that decision make knowledge is updated as a resultant.

Stage 6- Deployment- It is an automated stage that deals with the testing and feedback.

Data mining is usually showed as a step towards the process of knowledge discovery. The knowledge data discovery (KDD) process includes the steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation.







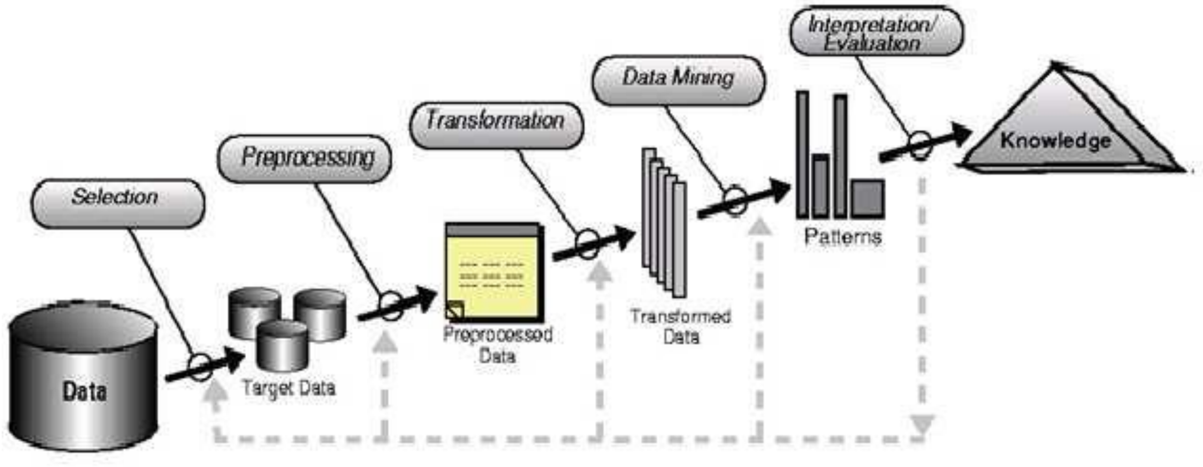


Fig 2. KDD Process

The quality of data plays vital role to detect results from the larger database. So the outlier or anomaly detection is main area of data mining research. Maximum of anomalies are occurs due to human errors. Data cleaning process does the removal of noise and data which is differ from the other data. Data integration refers to collect the information from different data sources such as database and datawarerhouse. Data selection is the process of retrieve the relevant data to the work. Data transformation is to transform the information in the proper format. Knowledge evaluation is to view and present the mined knowledge to the user. Association rule mining, classification, clustering, sequential pattern mining are the many data mining technique [2].

Let us focus on the parallel and distributed data mining-

Distributed data mining-

Databases in today's world are inherently distributed. There are many organizations that operate on global market that requires performing data mining [4] on distributed data sources and needing to integrate knowledge from the database. The organizations are geographically separated from the data sources. This inherent distribution of databases and large volume of information leads to increases the communication cost. Therefore, it is convenient that previous data mining model co-location of users, data and computational resources is inadequate when deal with distributed environment [4].

Distribution of users and computational resources are addresses the effect on the data mining process. Hence there are significance of distributed data mining over the centralized mining.

- The integration of the data and need to mine distributed subset of data of which is non-trivial and expensive.
- Need of data mining for performance and scalability bottle necks.
- Distributed data mining provides a framework to mine distributed data paying careful attention towards the distributed data and computing resources.

Parallel Data mining-

The enormity and high dimensionality of datasets typically available as input to the problem of association rule discovers it makes an ideal problem for solving multiple machines in parallel. The reasons are the CPU speed limitation and memory faced by single computer. So it is difficult to design efficient parallel algorithm to do the task. The reason is the many transaction databases are already available or they are distributed at multiple sites. Thus for the serial discovery the cost gets increases [9].

So the data parallelism approach comes. Parallelism is an effective approach for improving performance and achieving scalability. There are two parallel approaches such as task and data parallelism. Four major classes of parallel implementations are distinguished [4].

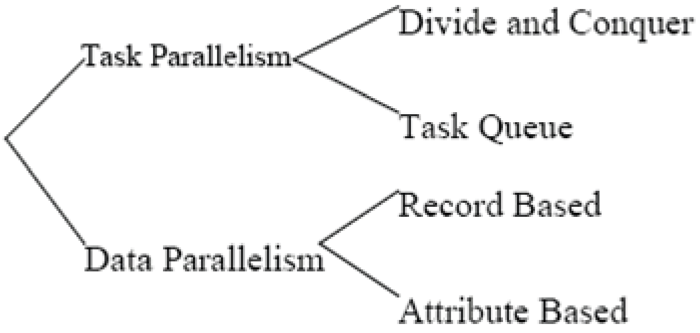


Fig 3. Methods of parallelism

IV. RELATED WORK

Anomaly detection (or outlier detection or microcluster detection) involves algorithms that try to spot various types of “strange” behaviours, which deviate “much” from the normal expected behavior. In general the problem is not well defined and the anomalousness of an entity heavily depends on the application that is of interest. The interest can be malware detection, network detection, biological data, image processing etc. The anomaly detection consists an important research direction in multiple fields [10].

Traditional method of anomaly detection is very time consuming task but recently there are emergence of different techniques such as parallel and distributed method for anomaly detection. Lozano and Acuna [3] presented a system which is based on a definition of distance-based anomalies similar with the one used here. It is a parallel version of Bay’s algorithm. Moreover, the method does not scale well in two out of the four experiment result presented. However this parallel version does not include the drawbacks of the centralized version. It is sensitive to the distribution of the dataset and to the order.

Hung and Cheung proposed a PENL [10] of the basic nested loop algorithm. It is also a parallel version. PENL is depends on a definition of anomaly employed in- A point for which less than n points are present within the certain distance d in the given input dataset which is called a distance based anomaly. The definition does not give the ranking of anomalies and requires to find out an appropriate value of the distance d. However, PENL requires the whole dataset is transferred among all the network nodes hence the PENL is not suitable for distributed mining [10].

Giannella, Kargupta, Borne and Dutta [9] gives algorithms for the top-n anomaly detection and the distributed computation of principal components. In that the anomaly are the patterns that deviate far from the correlation structure of the other data. The top-n anomalies are the entities having at most the n-th largest sum of squared values in a fixed number of lowest order principal elements, where each element is normal to its deviation. The above definition is not implies as well as implied by the definition employed in this system. For example if all clusters are located far from the mean of the data set, distance-based outliers close to the mean are not necessarily exceptional in the correlation structure. On the other hand, elements having large values in the first principal components need not have smaller weight than elements which deviate from the correlation structure in the low-order components.

Ghoting, Otey and Parthasarathy and Koufakou and Georgiopoulos [9] proposed their methods for distributed and data set with high dimensions. These strategies are based on definition of anomaly which are completely different from the definition explained here. In that instead of the use of distances they used the concept of support.

V. THE PROPOSED SYSTEM

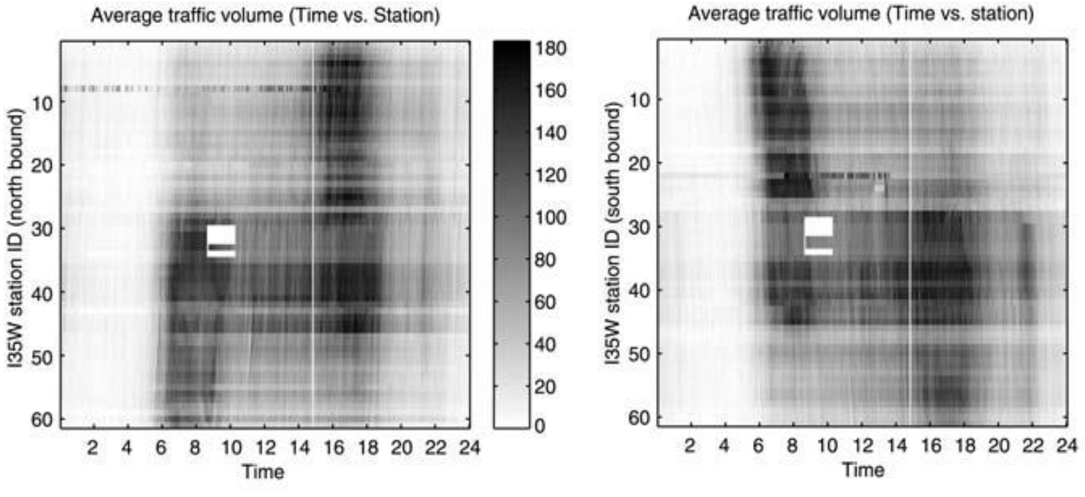
The design models give an analytical cost models for anomaly detection algorithms, and a comparison to an alternative for data clustering methods. Here we consider a applicative domain as "TRANSPORTATION TRAFFIC MANAGEMENT CENTER" where freeway operations group attains traffic measurements [7].

MODULES:

1] Around 900 stations are included in sensor network in which each station consist of one to four loop detector, which ultimately depend upon the number of lanes present. Sensors installed in the freeway system checks the volume of the traffic on the road. At each consecutive intervals the enrollment of increase or decrease in volume is send for operational purposes, such as ramp meter control or research on traffic modeling, to Traffic management system.

2] In the proposed applicative based system each station is attributed with locations that is they are spatial and the next attribute is related to measurements that is they could be non-spatial too. Spatial arrangement of stations here it is called as nodes and is considered under a directed graph. Consider tow nodes that is station S1 and S2 are placed at a Euclidean space and are joined by a directed edge, this edge indicates the segment of a lane or road for traffic to move. This graph is called as spatial graph where each station is specified with coordinates as its location for the attribute (e.g. Highway, mile point). Whereas the sensor-id and traffic measurements (e.g. volume, occupancy) are the attributes for non-spatial. The proposed system here is used to detect spatial anomaly in the form of location of nodes whose measurements are not consistent with their neighborhoods.

3] A collection of spatially referenced object constitutes the traffic sensors. The location of sensor is denoted by X as it is a spatial attribute and the traffic measurements is denoted as f(X) since it is a non-spatial attribute. The neighborhood is denoted as N(X), is a set of traffic sensors which are adjacent to X (sensor located). Neighborhood relation is based on directed edges hence the sensors on opposite sides are not considered to be neighbors even though their pair wise Euclidean distance is very less in the underlying spatial graph.



(a) I-35W north bound (b) I-35W south bound

Fig. 4 An Example of an outliers

4] S(x)=[f(X)-Ey]

VI. Conclusion and Future Scope

We have presented a distributed strategies for determining anomaly detection and the top-n distance based anomalies according to the definition. In this paper we have also proved that the centralized algorithms can be make larger or longer in space or time to work in distributed environment. The proposed solution was 1. To increase the speedup i.e. about to linear w.r.t. number of nodes in distributed environment. 2. System gives good result for increasing number of nodes w.r.t. the computation in the co-ordinator node as well as data transmission. There are two cases – 1) when data is located on distributed node, so by sending all information to a co-ordinator node can be neglected and increases the safety and decreases the performance. 2) When distributed computing power is available the maximum speedup gives guaranty to minimize exploitation of computing services and a good throughput [5].

The extreme system will be investigate spatial outliers with multiple non-spatial attributes, such as the combination of volume, occupancy and speed in the large data sets. The key challenge is to define general distance function in a multidimensional data space. We can also explore graphical method for spatial anomaly detection [1].

References

[1]. Fabrizio Angiulli, Stefano Basta, Stefano Lodi, and Claudio Sartori. ”Distributed Strategies for Mining Outliers in Large Data Sets.” IEEE TRANSACTIONS month 2012.

[2]. H. Kargupta and P. Chan, editors. ” Advances in Distributed and Parallel Knowledge Discovery”. AAAI/MIT Press, 2000.

[3]. A. Koufakou and M. Georgiopoulos. “A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes.” Data Min. Knowl. Discov, 2009

[4]. M. J. Zaki and C.-T. Ho, editors. “Large-Scale Parallel Data Mining,” volume 1759 of LNCS. Springer, 2000.

[5]. F. Angiulli and C. Pizzuti. “Outlier mining in large high dimensional data sets.” TKDE, 2(17):203215, February 2005.

[6]. Y. Tao, X. Xiao, and S. Zhou. “Mining distance-based outliers from large databases in any metric space.” In KDD, pages 394 403, 2006.

[7]. Shashi Shekhar, Chang- tien lu and Pusheng Zhang “A Unified Approach to Detecting Spatial Outliers.”

[8]. A. Ghoting, S. Parthasarathy, and M. E. Otey. “Fast mining of distance-based outliers in high-dimensional datasets.” Data Min. Knowl. Discov., 16(3):349364, 2008.

[9]. E. Lozano and E. Acu na.”Parallel algorithms for distance based and density-based outliers.” In ICDM, pages 729732, 2005.

[10]. S. Ramaswamy, R. Rastogi, and K. Shim. “Efficient algorithms for mining outliers from large data sets.” In SIGMOD, pages 427438, 2000.

Authors

Pawar Suvrna is the Head of Information Technology, Amrutvahnini collage of engineering. She has been doing research in databases, information systems, data mining.

Agwan Priyanka R. B.E. in Computer Engineering from Pune University, in 2012. She is currently pursuing her M.E. in Information Technology from University of Pune.