Recognisation of Outlier using Distance based method for Large Scale Database

Madhav Bokare
Research Scholar
Priyadarshini Institute of Engineering & Technology, Nagpur.
bokaremadhav@yahoo.com

V.M Thakare
Professor
Sant Gadge Baba Amravati university, Amravati.
vilthakare@yahoo.co.in

Abstract—This paper studies the difficulties of outlier detection on inexact data. We study the normal instances for each uncertain object using the instances of objects with analogous properties. Outlier detection is a significant research problem in data mining that goals to determine valuable abnormal and irregular patterns hidden in vast data sets. Most existing outlier detection approaches only deal with static data with comparatively low dimensionality. Newly, outlier detection for high-dimensional stream data turn into a new emergent research problem. A key remark that inspires this research is that outliers in high-dimensional data are predictable outliers, i.e., they are embedded in lower dimensional subspaces. Detecting projected outliers from high-dimensional stream data is a very stimulating task for numerous reasons. The paper shows the detailed study of outlier detection algorithms and its results also.

Keywords— Masking, Medoid. Univariate.

*****

I. INTRODUCTION

Uncertainty is characteristic in data composed in various applications, such as sensor networks, marketing research, and community science. Uncertain data poses significant challenges for data analytic tasks. In this paper, we inspect an important data mining problem – outlier detection – on uncertain data.

Example 1 (Motivation – product evaluation):

A digital camera builder carries an entire series of products. Each product has some inherent properties, such as pixel density, sensor type, and aperture variety. The similarity between two products can be measured by the proximity of their inherent properties.By analyzing customer assessments on separate products, the manufacturer can study the market and regulate its policies on marketing and product development. One product may receive numerous assessments which may vary to one another. The reviews of the product are not certain and can be modeled as an uncertain object, where each evaluation is stared as an instance. Similar products are predictable to obtain similar evaluation grades from customers. For a product, some evaluations may be very different from the majority. Those outlier evaluations need to be examined. They may be interesting if they capture some issues in infrequent scenarios or reflect opinions of specific user groups. They may be excluded from analysis if they are noise, such as spam reviews. Some products may receive evaluations very different from similar products. Those outlier products are chiefly interesting since they may provide important hints on customer/market interests. Clearly, outlier detection on uncertain data at both the instance level and the object level is almost useful.

Example 2 (Motivation – sensor data):

As a lengthily adopted approach for environment surveillance, sensors are deployed to cover part of interest. Each sensor keeps broadcasting its readings to monitor several sensing measures at its place. Each sensor has some inherent properties, such as its spatial location. The spatial distance between two sensors can be measured. In many applications, the target sensing measure such as temperature can be regarded steady in a short period such as 30minutes. However, the factual value of temperature cannot be precisely obtained due to limitations of measuring equipment. Instead, sensor readings are collected in order to estimate the true temperature. So the true value of the temperature at a certain location is an uncertain object, where multiple readings collected from a sensor at this location are the instances of this uncertain object. Often, it is expected that the mark sensing measures at nearby locations are similar, so are the readings of the sensors. For a sensor, some readings may be very different from the true values of the target sensing measure due to factors like dynamic errors, drifts, and noise. It is important to clean those outlier readings to improve the correctness of the sensor. Moreover, some sensors may deviate significantly from their neighbors. Such outlier sensors may be caused by malfunctioning sensor units or sensors at specific locations such as a deep hole on the ground. Again, detecting outliers from uncertain data at both the instance level and the object level is meaningful [1].

II. FUNDAMENTALS OF OUTLIER DETECTION

1. What is an Outlier?

The term outlier, also known as anomaly, initially stems from the field of statistics [3]. The two traditional definitions of outliers are: (Hawkins [4]): “an outlier is an observation, which diverges so much from other observations as to stimulate doubts that it was generated by a dissimilar mechanism”. (Barnett and Lewis [5]): “an outlier is an observation (or subset of observations) which seems to be unpredictable with the remainder of that set of data”. In addition, a diversity of definitions depending on the specific method outlier detection techniques are based upon exist [6]. Each of these definitions signify the solutions to recognize outliers in a specific type of data set. In WSNs, outliers can be defined as, “those measurements that meaningfully deviate from the usual pattern of sensed data” [7]. This definition is founded on the fact that in WSN sensor nodes are allocated to monitor the physical world and thus a pattern representing the normal behavior of

sensed data may exist. Potential sources of outliers in data composed by WSNs include noise and blunders, actual events, and malicious attacks. Noisy data as well as mistaken data should be removed or corrected if possible as noise is a random error deprived of any real implication that intensely affects the data analysis [8]. Outliers produced by other sources need to be recognized as they may contain important information about events that are of great interest to the researchers.

2. Motivation of Outlier Detection

Outlier detection also known as anomaly detection or deviation detection, is one of the fundamental tasks of data mining along with predictive modelling, cluster analysis and association analysis [8]. Compared with these other three tasks, outlier detection is the closest to the initial motivation behind data mining, i.e., mining useful and interesting information from a large amount of data [9]. Outlier detection has been widely researched in various disciplines such as statistics, data mining, machine leaning, information theory, and spectral decomposition [7]. Also, it has been widely applied to numerous applications domains such as fraud detection, network intrusion, performance analysis, weather prediction, etc. [7].

III. Multivariate Outlier Detection

In many cases multivariable observations cannot be detected as outliers when each variable is considered independently. Outlier detection is possible only when multivariate analysis is performed, and the interactions among different variables are compared within the class of data. A simple example can be seen in Figure 1.1, which presents data points having two measures on a two-dimensional space. The lower left observation is clearly a multivariate outlier but not a univariate one. When considering each measure separately with respect to the spread of values along the x and y axes, we can is seen that they fall close to the center of the univariate distributions. Thus, the test for outliers must take into account the relationships between the two variables, which in this case appear abnormal.

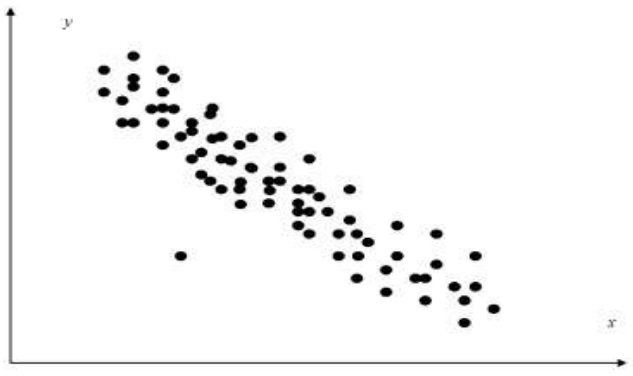

Figure 1 A Two-Dimensional Space with one Outlying Observation (Lower Left Corner).

Data sets with multiple outliers or clusters of outliers are subject to *masking* and *swamping* effects. Although not mathematically rigorous, the following definitions from (Acuna and Rodriguez, 2004) give an intuitive understanding for these effects (for other definitions see (Hawkins, 1980; Iglewics and Martinez, 1982; Davies and Gather, 1993; Barnett and Lewis, 1994)) [10].

Masking effect

It is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

Swamping effect

It is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers.

IV. Outlier Detection Models

Feature Selection in Outlier Detection:

It is notoriously difficult to perform feature selection in outlier detection because of the unsupervised nature of the outlier detection problem. Unlike classification, in which labels can be used as guiding posts, it is difficult to learn how features relate to the (unobserved) ground truth in unsupervised outlier detection. Nevertheless, a common way of measuring the non-uniformity of a set of univariate points x1 . . . xN is the Kurtosis measure. The first step is to compute the mean μ and standard deviation σ of this set of values and standardize the data to zero mean and unit variance as follows:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Note that the mean value of the squares of z_i is always 1 because of how z_i is defined. The Kurtosis measure computes the mean value of the fourth power of z_i:

$$K(z_1 \dots z_N) = \frac{\sum_{i=1}^{N} z_i^4}{N}$$

Feature distributions that are very non-uniform show a high level of Kurtosis. For example, when the data contains a few extreme values, the Kurtosis measure will increase because of the use of the fourth power. Kurtosis measures are often used [367] in the context of subspace outlier detection methods, in which outliers are explored in lower dimensional projections of the data.

Extreme-Value Analysis

The most basic form of outlier detection is extreme-value analysis of 1-dimensional data. These are very specific types of outliers in which it is assumed that the values that are either too large or too small are outliers. Such special kinds of outliers are also important in many application-specific scenarios. The key is to determine the statistical tails of the underlying distribution. As illustrated earlier in Figure 1.3, the nature of the tails may vary considerably depending upon the underlying data distribution. The normal distribution is the easiest to analyze, because most statistical tests (such as the Z-value test) can be interpreted directly in terms of probabilities of significance. Nevertheless, even for arbitrary distributions, such tests provide a good heuristic idea of the outlier scores of data points, even when they cannot be interpreted statistically. The problem of determining the tails of distributions has been widely studied in the statistics literature.

Probabilistic and Statistical Models

In probabilistic and statistical models, the data is modeled in the form of a closed-form probability distribution, and the parameters of this model are learned. Thus, the key assumption here is about the specific choice of the data distribution with which the modeling is performed. For example, a Gaussian mixture model assumes that the data is the output of a generative process in which each point belongs to one of k Gaussian clusters. The parameters of these Gaussian distributions are learned with the use of an expectation-maximization (EM) algorithm on the observed data so that the probability (or likelihood) of the process generating the data is as large as possible. A key output of this method is the membership probability of the data points to the different clusters, as well as the density-based fit to the modeled distribution. This provides a natural way to model the outliers, because data points that have very low fit values may be considered outliers. In practice, the logarithms of these fit values are used as the outlier scores because of the better propensity of the outliers to appear as extreme values with the use of log-fits.

A drawback of probabilistic models is that they try to fit the data to a particular kind of distribution, which may sometimes not be appropriate. Furthermore, as the number of model parameters increases, over-fitting becomes more common. In such cases, the outliers may fit the underlying model of normal data. Many parametric models are also harder to interpret in terms of intentional knowledge, especially when the parameters of the model cannot be intuitively presented to an analyst in terms of underlying attributes. This can defeat one of the important purposes of anomaly detection, which is to provide diagnostic understanding of the abnormal data generative process.

Linear Models

These methods model the data along lower-dimensional subspaces with the use of linear correlations. For example, in the case of Figure 1.2, the data is aligned along a 1-dimensional line in a 2- dimensional space. The optimal line that passes through these points is determined with the use of regression analysis. Typically, a least-squares fit is used to determine the optimal lower-dimensional hyperplane. The distances of the data points from this hyperplane are used to quantify the outlier scores because they quantify the deviations from the model of normal data. Extreme-value

analysis can be applied on these scores in order to determine the outliers.

For example, in the 2-dimensional example of Figure 1.4, a linear model of the data points {(xi, yi), i ∈ {1 . . . N} in terms of two coefficients a and b may be created as follows:

$$y_i = a \cdot x_i + b + \epsilon_i \quad \forall i \in \{1 \ldots N\}$$

Here, i represents the residual, which is the modeling error. The coefficients a and b need to be learned from the data to minimize the least-squares error, which is denoted by $\sum_{i=1}^{N} \epsilon_i^2$. .This is a convex non-linear programming problem whose solution can be obtained in closed form. The squared residuals provide the outlier scores. One can use extreme-value analysis to identify the unusually large deviations, which should be considered outliers.

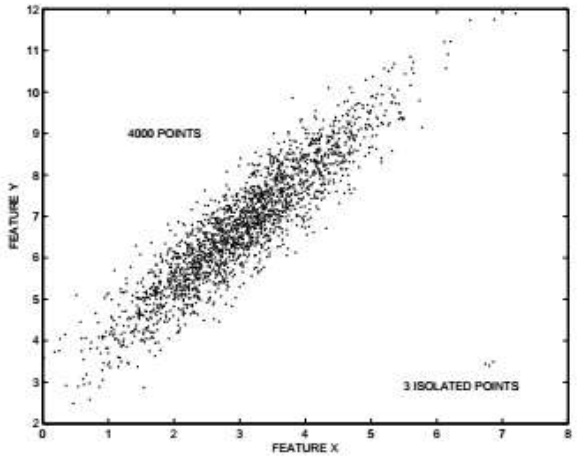

Figure 2 Small groups of anomalies can be a challenge to density-based methods

Spectral Models

Many of the matrix decomposition methods such as PCA are also used in the context of graphs and networks. The main difference is in how the matrix is created for decomposition. Some variations of these methods, which are used in certain types of data such as graphs and networks, are also referred to as spectral models. Spectral methods are used commonly for clustering graph data sets, and are often used in order to identify anomalous changes in temporal sequences of graphs. Spectral methods are closely related to matrix factorization, which can also be used in such settings.

Proximity-Based Models

The idea in proximity-based methods is to model outliers as points that are isolated from the remaining data on the basis of similarity or distance functions. Proximity-based methods are among the most popular class of methods used in outlier

analysis. Proximity-based methods may be applied in one of three ways, which are clustering methods, density-based methods and nearest-neighbor methods. In clustering and other density-based methods, the dense regions in the data are found directly, and outliers are defined as those points that do not lie in these dense regions. Alternatively, one might define outliers as points that are located far away from the dense regions. The main difference between clustering and density-based methods is that clustering methods segment the data points, whereas the density-based methods such as histograms segment the data space. This is because the goal in the latter case is to estimate the density of test points in the data space, which is best achieved by space segmentation.

High-Dimensional Outlier Detection

The high-dimensional case is particularly challenging for outlier detection. The reason for this behavior is that many dimensions may be noisy and irrelevant for anomaly detection, which might also increase the propensity for pairwise distances to become more similar. The key point here is that irrelevant attributes have a dilution effect on the accuracy of distance computations and therefore the resulting outlier scores might also be inaccurate. When using distance-based algorithms to score outliers, one often observes the effect of weakly correlated and irrelevant attributes in the concentration of distances. In high-dimensional space, the data becomes increasingly sparse, and all pairs of data points become almost equidistant from one another. As a result, the outlier scores become less distinguishable from one another. In such cases, outliers are best emphasized in a lower-dimensional local subspace of relevant attributes. This approach is referred to as subspace outlier detection, which is an important class of algorithms in the field of outlier analysis. The assumption in subspace outlier detection is that outliers are often hidden in the unusual local behaviour of low-dimensional subspaces, and this deviant behaviour is masked by full-dimensional analysis.

V. METHODOLOGIES

K-Means Clustering Algorithm.

K-means is a prototype-based, simple partition clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice. This algorithm consist of two separate phases: the first phase is to select k centres randomly, where the value of k is fixed in advance. The next phase is to assign each data object to the nearest centre. Euclidean distance is generally considered to determine the distance between each data object and the cluster centres. When all the data objects are included in some clusters, recalculation is done on the average of the clusters. This iterative process continues repeatedly until the criterion function become minimum [12].

The k-means algorithm works as follows:
Step 1: Randomly select k data object from dataset D as initial cluster centers.
Step 2: Repeat step 3 to step 5 till no new cluster centers are found.
Step 3: Calculate the distance between each data object di(1<=i<=n) and all k cluster centers cj(1<=j<=n) and assign data object di to the nearest cluster.
Step 4: For each cluster j(1<=j<=k), recalculate the cluster center

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. The definition can be narrowed to outlier detection can be thought of detection of anomalous data patterns in the data sets by pre learned techniques of the data sets got from the observation of different experimental condition The output of an outlier detection technique could be labeled patterns. Some of the outlier detection techniques also assign a score to a pattern based on the degree to which the pattern is considered an outlier. Such a score is referred to as outlier score [12].
Outlier detection can be broadly classified into three groups. First is the distance based outlier detection, it detects the outlier from the neighbourhood points. Second is the density based outlier detection, here it detects the local outlier from the neighbourhood based on the density or the no. of data points in the local neighbourhood. Third is the distribution based outlier detection, this approach is based on the finding outlier using some statistical model.

Density based outlier detection

Density-based methods have been developed for finding outliers in a spatial data. These methods can be grouped into two categories called multi-dimensional metric space-based methods and graph-based methods. In the first category, the definition of spatial neighborhood is based on Euclidean distance, while in graph-based spatial outlier detections the definition is based on graph connectivity. Distance based outlier detection. In Distance-based methods outlier is defined as an object that is at least dmin distance away from k percentage of objects in the dataset. The problem is then finding appropriate dmin and k such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge [11].

Outlier Finding Technique (OFT).

Outlier Finding Technique (OFT) is a hybridized form of both distance based and density based outlier finding technique. Here after cluster formation has taken place with the help of k-means clustering then we are left with the cluster of data points and the cluster centre.

VI. EXPERIMENT

In this concept we used the large data set for finding the outliers. The excel sheet contents are large set of values. Here we are accessing the excel sheet following way.

ISSN: 2321-8169
1001 – 1006

Figure 3 Button for Selecting Sheet

Figure 3 indicates the select file button, by using this button we can select or search file from system.

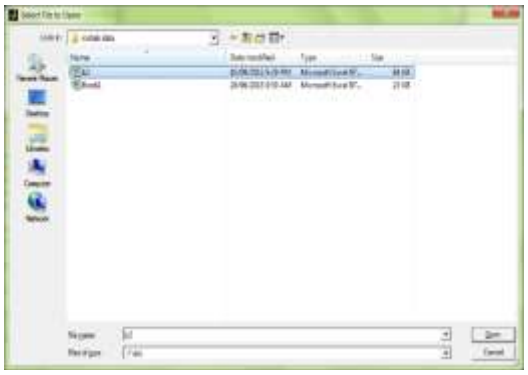

Figure 4 Select excel file

Figure 4 indicates that the browsing window from which we can select excel file.

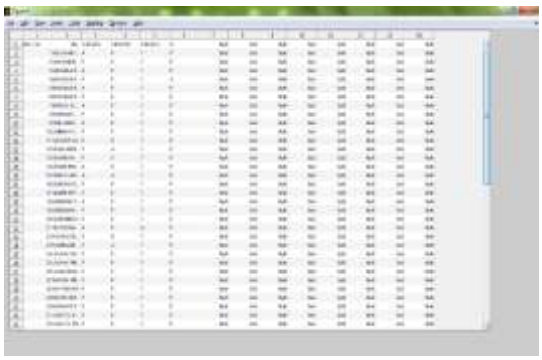

Figure 5 Selected Data File

VII. RESULT

Step 1: Selecting Excel File
Step 2: Selecting Data Set From Excel File
Step 3: Getting the Control Chart(CCT)
Step 4: Generating K-means outlier
Step 5: Generating K-medoid
Step6: Exit

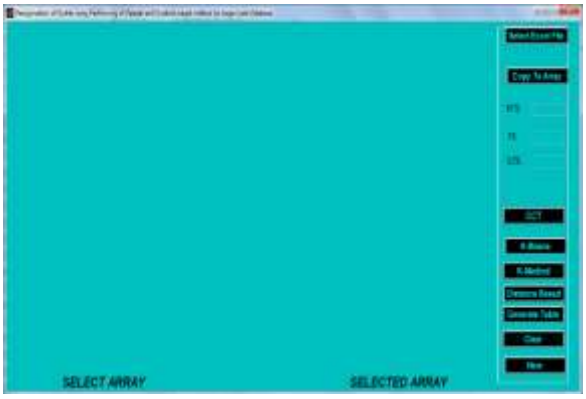

Figure 6 GUI

This GUI indicates the selection of excel sheet, CCT, K-Means, K-Medoid, Distance based and lastly generated tables.

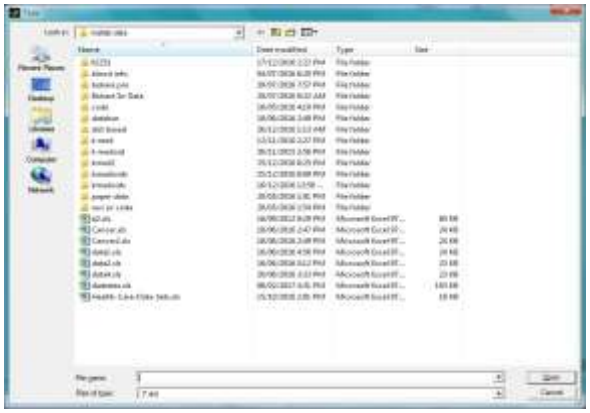

Figure 7 Selection of Sheet

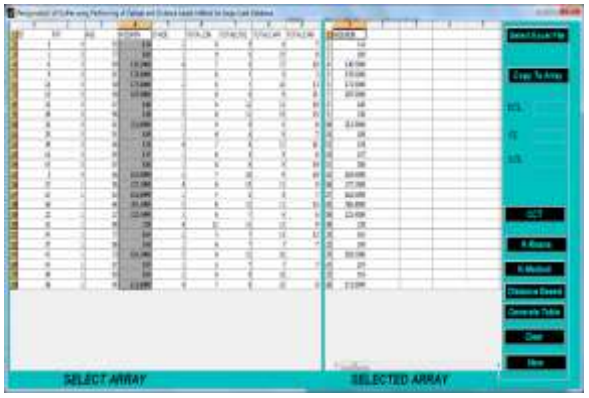

Figure 8 Highlighting Selected column

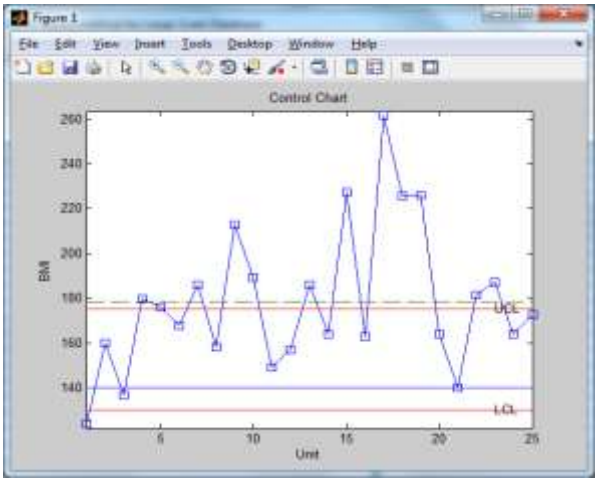

Figure 9 Output of CCT

 ISSN: 2321-8169
1001 – 1006

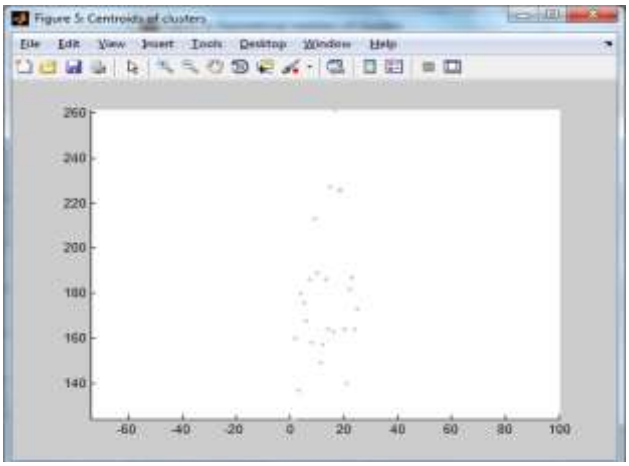

Figure 10 Output of Distance Based

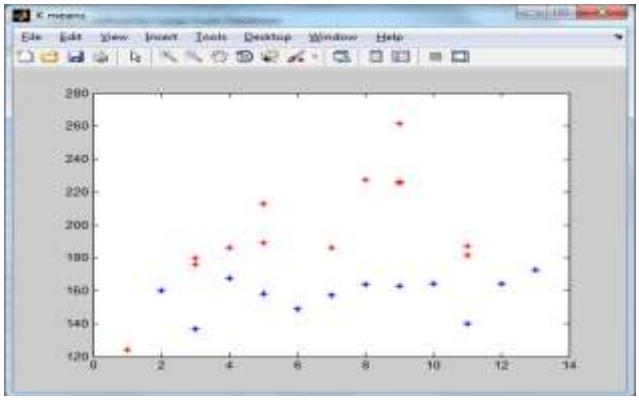

Figure 11 Output of K-Means

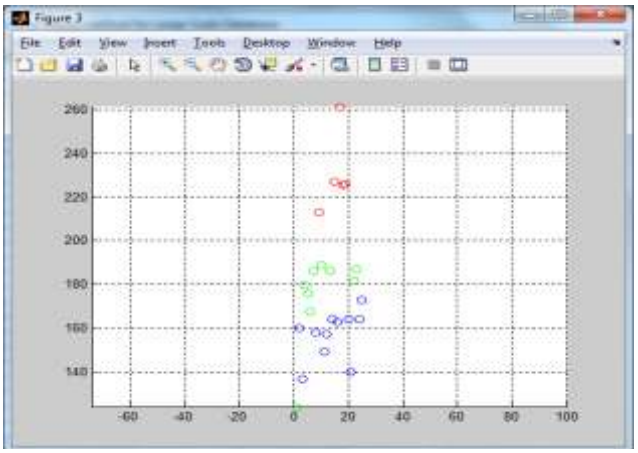

Figure 12 Output of K-Medoid

VIII. CONCLUSIONS

In this paper we have shown the outlier of sampled data set by applying CCT, the k-means, Distance Based and K-medoid.

ACKNOWLEDGMENT

We wish to thank Dr. S. B. Thorat, Director,ITM,and Nanded.

REFERENCES

[1] Outlier Detection on Uncertain Data: Objects, Instances, and Inferences, Bin Jiang, Jian Pei, 978-1-4244-8960-2/11/ ©2011 IEEE.

[2] Outlier Detection Techniques for Wireless Sensor Networks: A Survey, Yang Zhang, Nirvana Meratnia, and Paul Havinga, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 12, NO. 2, SECOND QUARTER 2010.

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

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

[5] V. Barnett and T. Lewis, Outliers in Statistical Data, NewYork: John Wiley Sons, 1994.

[6] 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.

[7] V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection: A Survey,Technical Report, University of Minnesota, 2007

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

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

[10] Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

[11] "Outlier Detection Using k-Nearest Neighbor Graph" by V. Hautamäki, I. Kärkkäinen and P. Fränti,In Proceedings of the International Conference on Pattern Recognition, Volume 3 pages 430 – 433, Cambridge, UK, August 2004.

[12] H.S.Behera, Abhishek Ghosh, Sipak ku. Mishra A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012, ISSN: 2277 128X