ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering &Technology (IJARCET)
Volume 2, Issue 9, September 2013

Outlier Detection Using K-Mean and Hybrid Distance Technique on Multi-Dimensional Data Set

Shruti Aggarwal, Janpreet Singh

Abstract - Outlier Detection is a major issue in data mining. Outliers are the containments that divert from the other objects. Outlier detection is used to make the data knowledgeable, and easy to understand. There are many type of databases used now days, and many of them contains anomaly objects, detection or removal of these objects is known as outlier detection. In the proposed work outliers are detected by partitioning the dataset with the clustering method that is the K – Mean method using the Mean of Euclidean and Manhattan distance and then find out the outlier with the Hybrid technique that is the mean of the Euclidean and Manhattan Distance. The proposed work is highly efficient in detection of outliers and produces much efficient outliers by using the real bench marked data sets: Iris dataset and Pima Indian Diabetes data set.

Index Terms - Outliers, Euclidean Distance, Manhattan Distance, Hybrid Technique

I. INTRODUCTION

Data Mining is the task of extracting useful knowledge from a collection of data bases or data warehouses, nowadays data is stored in various formats such as documents, images, audio, videos, scientific data, etc. [1]. It is also the process of discovering relationships within data. Identified relationships can be used for scientific discovery, business decision making, or data profiling. There are many applications going on these days that use the techniques to detect outliers from the databases.

1.1 Data Mining Functionalities

Data mining functionalities [13] are used to specify the kind of patterns to be found in data mining tasks. There are various types of databases and information repositories on which data mining can be performed. There are different data mining functionalities such as,

A. Concept/Class Description: Characterization and Discrimination
B. Classification and Prediction
C. Cluster Analysis
D. Evolution and Deviation Analysis
E. Outlier Analysis.

1.2 Clustering

Clustering is process of grouping the similar data into groups that are different from each other. Clustering is an unsupervised classification technique, which means that it does not have any prior knowledge of its data and results before classifying the data [2]. Clustering is used to improve the efficiency of the result by making groups of the data. So to cluster the data means specifying the data objects to a specific cluster which has similar objects or a group of objects.

1.3 Outlier Detection

Outliers are the objects that are not same as the other objects in the cluster or in the database. So to detect or remove anomaly objects from the dataset outlier detection technique are used, such as Density based method, Distance Based method, Clustering Based method, Graph Based method etc. Outliers Detection is used in many applications, such as credit card fraud detection, discovery of criminal activities in electronic commerce, weather prediction, marketing and customer segmentation. [3].

1.3.1 Applications of Outlier Detection

1. Credit Card Fraud Detection
2. Cyber-Intrusion Detection
3. Medical Anomaly Detection
4. Industrial Damage Detection
5. Textual Anomaly Detection
6. Image Processing
7. Motion segmentation
8. Structural defect detection
9. Pharmaceutical research
10. Loan application processing

II. Literature Survey

Outlier detection is an extremely important task in a wide variety of application domains. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data or which are far away from

- ***Shruti Aggarwal,*** *Assistant Professor, Dept. of CSE, Sri Guru Granth Sahib World Universi-ty, Fatehgarh Sahib, Punjab, India.*
- ***Janpreet Singh,*** *M.Tech Research Scholar, Dept. of CSE, Sri Guru Granth Sahib World University, Fatehgarh Sahib,*

their cluster centroids [4]. There are many methods to detect outliers from the dataset and they are as follow:

2.1 Partitioning Based Method

Partitioning based technique is a well known method to cluster the dataset into group of similar objects. Some of the well known techniques are 1. Centroid - Based Technique and 2. Object - Based Technique. Both of them are mainly used techniques because they are very effective in finding the clusters

Centroid-Based Technique mainly uses the K – Mean algorithm. The algorithm takes the input parameter, K, and partitions a set of N objects into K clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low.

S. D. Pachgade et.al [4] have used the simple k – mean method to partition the data into group of different clusters and then get the threshold value from the user and find the outliers using the Euclidean distance method above the user defined threshold value and showed the result with the real data set that how efficient it is in finding the outliers.

Ville Hautamaki et al. have presented the new method named Outlier Removal Clustering [5] that uses the clustering technique to find out the outlier by first removing the far vectors of data from the grouped dataset then analyzing the remaining cluster for outlier removal and recalculating the k – mean for grouping the vectors of the dataset, thus k – mean is used for efficient clustering techniques and for outlier removal method with Euclidean distance based methods.

2.2 Graph Based Method

Pang-Ning Tan et al. have proposed a new graph-based algorithm, called Outrank [6], for detecting outlying objects. In this method, a matrix is constructed using the similarity between objects and used as the adjacency matrix of the graph representation. The heart of this approach is the Markov model that is built upon this graph, which assigns an outlier score to each object.

Some of graph-based methods have been proposed in which Outlier Detection using In Degree Number (ODIN) with KNN algorithm. Outlier's defined using k nearest neighbor (KNN) graph is a weighted directed graph, as shown in Figure 2.1 in which every vertex represents a single vector, and the edges correspond to pointers to neighbor vectors. Every vertex has exactly k edges to the k nearest vectors according to a given distance function [7].

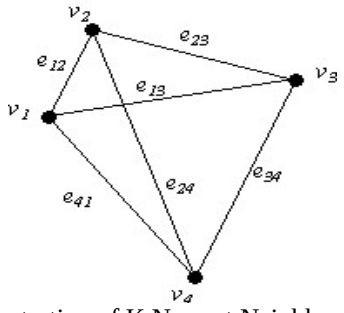


Fig. 2.1: Illustration of K Nearest Neighbour Graph [7]

2.3 Distance - Based Method

Gustavo H. Orair et.al [8] Thus Explicit distance-based approaches, based on the well known nearest-neighbor principle, were first proposed by Ng and Knorr and employ a well-defined distance metric to detect outliers, that is, the greater is the distance of the object to its neighbors, the more likely it is an outlier. The basic algorithm for such distance-based algorithms, it the nested loop algorithm, which calculates the distance between each pair of objects and then set as outliers to those that are far from the objects or from the centroid value.

S. D. Pachgade et.al [4] have used Euclidean distance method for finding outliers firstly the k- mean method is proposed to partition the data into k number of clusters and after that Euclidean method is used to detect the abnormal values form the dataset that are far away from the centroid value of the k clusters. Thus the points that have max distance from the threshold value are considered as outliers.

$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Fig. 2.2: Equation of Euclidean Distance [13]

Gerhard Munz et.al [9] have used the distance method for finding the outlier from the network traffic firstly the K-mean method is used to cluster the vector points as by the k number of clusters then the Improved distance is being used to find out the network traffic that is being abnormal data. Firstly it considers the far vectors and removes them from the dataset and then considers the dense clustered vectors and find out the outliers from them.

James M. Coughlan et.al [10] have proposed a method for outlier detection using the Bayesian model and Manhattan based technique with visual scenes are based on a Manhattan three dimensional grid which imposes regularities on the image statistics. For many images, these estimates are good approximations to the viewer orientation. These estimates also make it easy to detect outlier structures which are unaligned to the grid. In their proposed method firstly they apply Bayesian method for the detection of the x, y, and z planes and there pixels that are

on or off then they apply Manhattan based technique on the image to detect the outliers from the image.

$$d = \sum_{i=1}^{n} |X_i - Y_i|$$

Fig. 2.3: Equation of Manhattan Distance Algorithm [13]

2.4 Density - Based Method

This method assigns each object a degree to be an outlier. This degree is called the local outlier factor (LOF) [11] of an object. It is local in that, the degree depends on how isolated the object is with respect to the surrounding neighborhood. In LOF algorithm, outliers are data objects with high LOF values whereas data objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high potential of being outlier. DBSCAN and its extension, OPTICS, are typical density-based methods that grow clusters according to a density-based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions [13].

III. Proposed Method

In the approached method first the clustering algorithm is used that uses the K – mean algorithm, to partition the dataset into a set of clusters, by finding the cluster center with mean of the Euclidean and Manhattan technique and partition the dataset into user defined clusters. Then after clustering the dataset outliers are to found from the each cluster by reusing the mean of Euclidean and Manhattan techniques and finding out the outlier from each cluster. There after calculate the mean of all the outliers that are found from each clusters and by comparing the each outlier with the mean of the previous outliers, real outlier will be seprated and that are the real outliers that are different from the other cluster data.

The basic structure of the proposed method is as follows:

Input: Data set D = {d1, d2.....dn}, where di=data points of dataset D, n= no of data points, Cluster centre C = {c1, c2 ...ck), where cj=cluster centre, k = no of cluster centers, dn = data point of k cluster.

Step 1: Randomly select k data object from array of dataset D as initial cluster centers.

Step 2: Repeat step 3 to step 5 till no new cluster centers are found or it reaches to the maximum limit of the iteration where the max count value is being set.

Step 3: Calculate the distance with the mean of the Euclidean and Manhattan 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.

Step 5: Calculate the distance of each data points dn and the k cluster centers cj with mean of Euclidean and Manhattan Distance.

Step 6: Assign that dn point to an array ai that contains the outliers of all the k clusters.

Step 7: Repeat the Steps 5-6 till no new outlier is founded or until the distance criteria met.

Step 8: Calculate the mean mi of all data point dj of ai collection of outliers.

Step 9: Calculate the distance of each data point's dj with mi.

Step 10: If the calculated distance is less than the mi then the data points stay in the previous cluster.

Step 11: Else, the data point will be considered as real outlier.

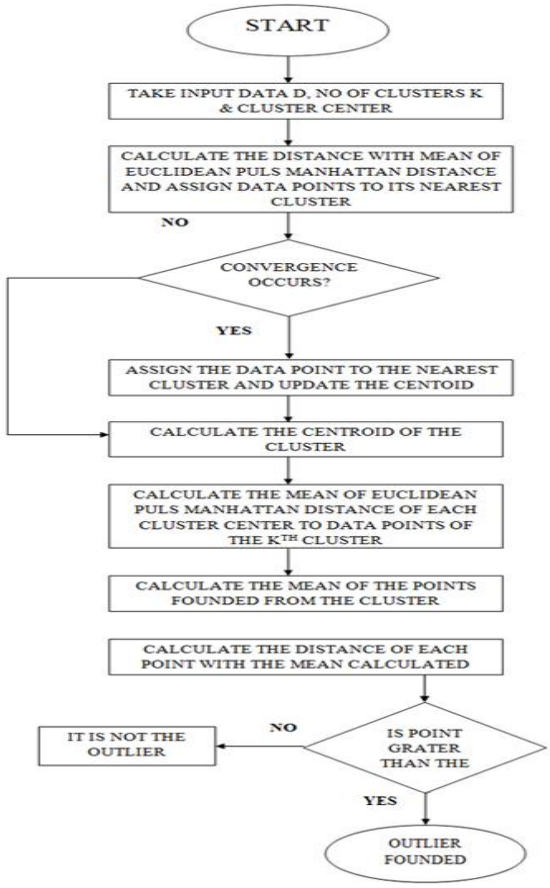


Fig. 3.1: Flow Diagram of Hybrid Technique

IV. Results and Discussion

This section will present the effectiveness of the proposed approach when applied on real bench marked data sets: Iris data set with 4 dimensions, Pima Indian Diabetes data set with 9 dimensions. Experimentation starts with well-known Iris data set [12]. The Iris data set has three classes of Iris Plants: Iris setosa, Iris versicolor and Iris viginica and has four variables (dimensions).

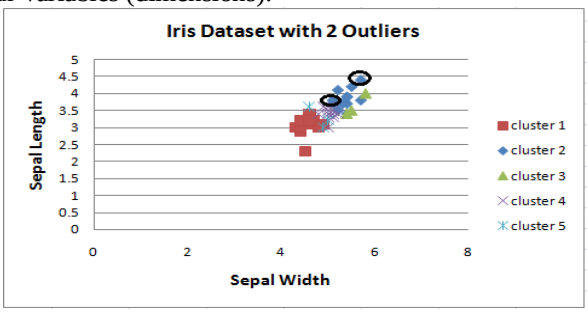
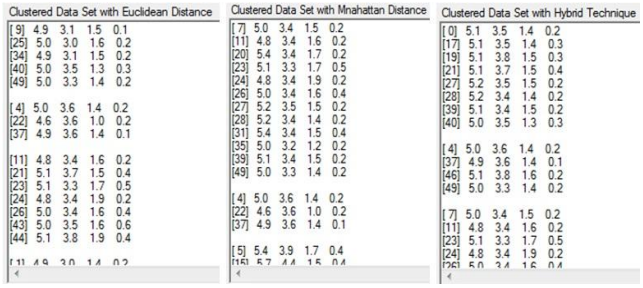


Fig. 4.1: Iris Dataset with five clusters and two outliers circled

Iris Data set is real world dataset with a four attributes which are the measurements of an eye of a person with a name sepal length, sepal width, petal length, petal width. Here in the above Fig. 4.1 data set is clustered into five clusters and 2 outliers are founded

4.1. Comparing Hybrid method with Euclidean and Manhattan distance techniques on Iris Dataset

4.1.1. Clustered Dataset

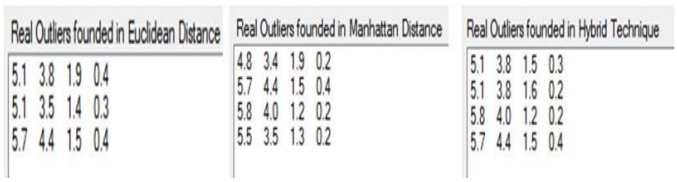


Euclidean Manhattan Hybrid

Fig. 4.2: Clustering analysis of Euclidean, Manhattan and Hybrid on Iris Dataset

In Figure 4.2 the Clusters of Iris dataset are different with each method, here 7 clusters are made and each technique makes different clusters.

4.1.2. Outliers Detected



Euclidean Manhattan Hybrid

Fig. 4.3: Comparative analysis of Outliers Found in Iris Dataset

4.1.3. Accuracy Measure of the methods

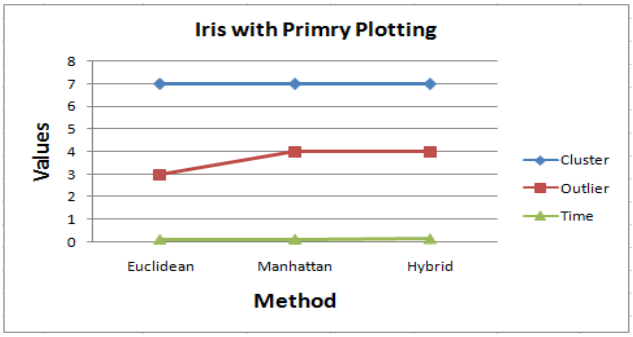


Fig. 4.4 (A): Accuracy Measure of Cluster, Outlier and Time on Iris Dataset

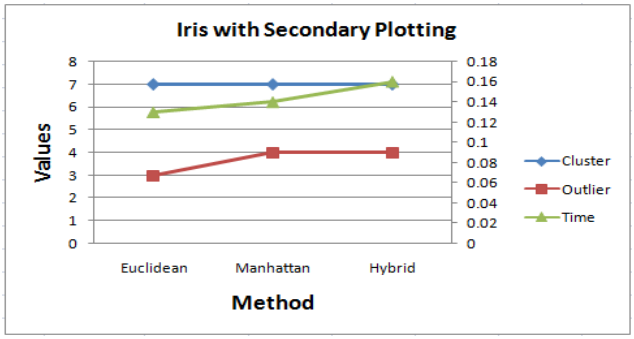


Fig. 4.4 (B): Accuracy Measure of Cluster, Outlier and Time on Iris Dataset.

Second dataset is Pima Indian Diabetes Dataset which has 9 attributes and 60 instances and partitioned into 10 clusters and outlier are founded on behalf of the clusters.

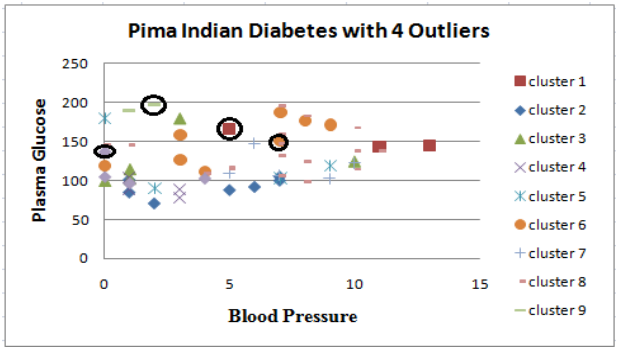


Fig. 4.5: Pima Indian Diabetes Dataset with 10 clusters and 4 outliers circled

Pima Indian Diabetes Data set with 10 clusters is presented with different colors, and 4 Outliers are found.

4.2. Comparing Hybrid method with Euclidean and Manhattan distance techniques on Pima Indian Diabetes.

4.2.1. Clustered Dataset

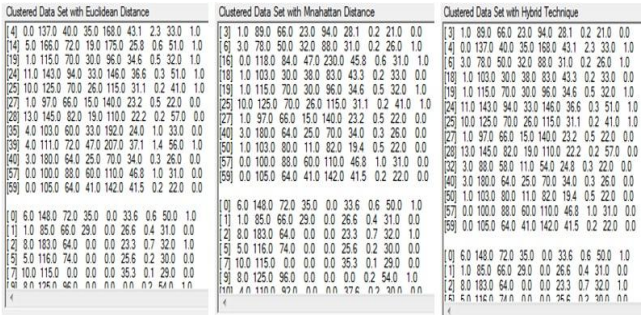


Fig. 4.6: Clustering Analysis of Hybrid, Euclidean, and Manhattan on Diabetes Dataset

In Figure 4.6 the Clusters of Pima Indian Diabetes Dataset are different from each other, here 4 clusters are made and each cluster is different from other cluster.

4.2.2. Outliers Detected

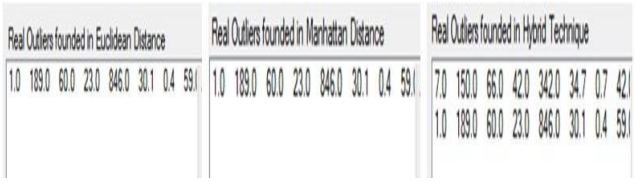


Fig. 4.7 Comparative analysis of Outliers Found in Pima Indian Diabetes Dataset

In the above Figure 4.7 Hybrid technique finds 2 outliers where as Euclidean technique and Manhattan finds only 1 outlier.

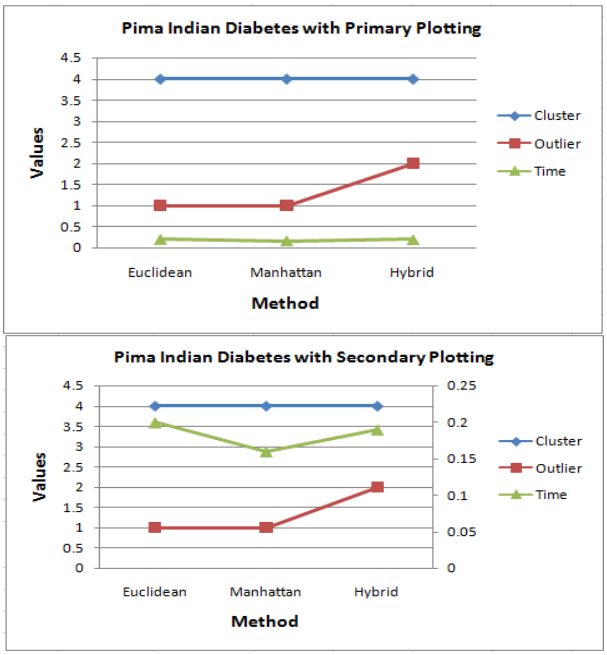


Fig. 4.8: Comparative Study of Cluster, Outlier and Time on Pima India Diabetes Dataset

4.3. Comparative Analysis of Euclidean and Hybrid Technique

Table 1: Comparative analysis of Euclidean and Hybrid Technique

Dataset	Euclidean Distance			Hybrid Technique		
	No. of Clusters	Time Taken	No. of Outliers	No. of Clusters	Time Taken	No. of Outliers
Iris	7	.13	3	7	0.16	4
Pima Indian Diabetes	4	0.20	1	4	019	2
Breast Cancer	9	0.30	4	9	0.25	6

4.4. Comparative Analysis of Manhattan and Hybrid Technique

Table 2: Comparative analysis of Manhattan and Hybrid Technique

Dataset	Manhattan Distance			Hybrid Technique		
	No. of Clusters	Time Taken	No. of Outliers	No. of Clusters	Time Taken	No. of Outliers
Iris	7	0.14	4	7	0.16	4
Pima Indian Diabetes	4	016	1	4	019	2
Breast Cancer	9	0.21	5	9	0.25	6

.

4.5. Accuracy Measure of Hybrid Method on Different Dataset

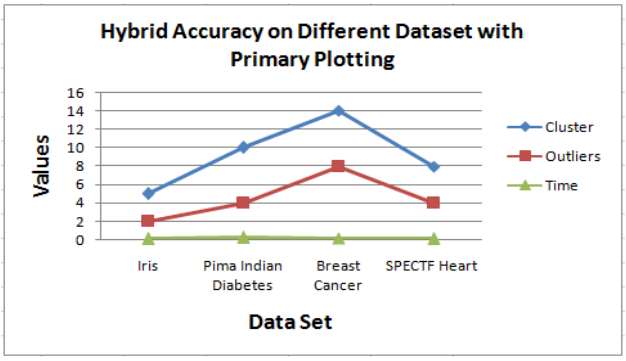


Fig. 4.9 (A): Accuracy measure on Hybrid Method

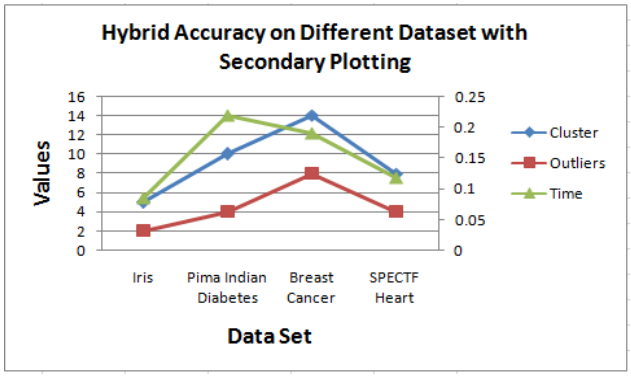


Fig. 4.9 (B): Accuracy measure on Hybrid Method

4.6. Comparative Analysis of Different Datasets

Table 3: Comparative analysis of different datasets

Dataset	No. of Items in Datasets		Proposed Algorithm		
	No of Attributes	No of Instances	Time taken(ms)	No. of Clusters	No. of Outlier Found
Iris	4	50	85	5	2
Pima Indian Diabetes	9	60	219	10	4
Breast Cancer	11	80	190	14	8

V. Conclusion

The proposed method first finds out the user defined number of clusters with the mean of Euclidean plus Manhattan Distance then outlier are detected from each cluster. After the detection of outliers find the mean of the outliers and compare it with the outliers. The outliers that are greater than the mean are considered as the real outliers.

1. The advantage of the clustering-based approaches is that they do not have to be supervised.
2. Clustering-based techniques are capable of being used in an incremental mode.
3. Hybrid method can cluster the data according to user need and find the outliers that differ from the other data in the dataset.
4. This approach makes those two problems solvable for less time, using the same process and functionality for both clustering and outlier identification.
5. The approached method can be widely used with the multidimensional datasets.

VI. Future Scope

Several interesting areas of future research have opened up from the work described in the thesis. In the future, more modifications can be made to the proposed method such as dimension reduction, reducing the number of iterations etc. Moreover, the different categories can also be used for outlier detection algorithms i.e. density-based, distribution based and hierarchy -based methods etc.

Further the approached method can be extended to the applicability of anomaly detection in different contexts such that with the combination of more than two methods, and can be broadens its application domains using with hybrid method.

VII. References

[1] Venkatadri.M, Dr. Lokanatha C. Reddy"*A Review on Data mining from Past to the Future*" International Journal of Computer Applications (0975 –8887) Volume 15, No.7, February 2011.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth" *From Data Mining to Knowledge Discovery in Databases*" American Association for Artificial Intelligence 0738-4602-1996.

[3] Moh"d Belal Al-Zoubi "*An Effective Clustering-Based Approach for Outlier Detection*" European Journal of Scientific Research Volume 28, No.2, 2009.

[4] Ms. S. D. Pachgade, Ms. S. S. Dhande "*Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach*" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, June 2012.

[5] Ville Hautamaki, Svetlana Cherednichenko, Ismo Karkkainen, Tomi Kinnunen, and Pasi Franti " *Improving K-Means by Outlier Removal*" Springer-Verlag Berlin Heidelberg 2005 LNCS 3540, pp. 978–987, 2005.

[6] H.D.K. Moonesinghe, Pang-Ning Tan" *Outlier Detection Using Random Walks*".

[7] T1-Svetlana Cherednichenko (2005)"*Outlier Detection in Clustering*" Master's Thesis, University of Joensuu 2005.

[8] Gustavo H. Orair, Carlos H. C. Teixeira, Ye Wang, Srinivasan Parthasarathy" *Distance Based Outlier Detection: Consolidation and Renewed Bearing*" Volume 3, No. 2, 2010 VLDB Endowment 21508097/10/09.

[9] Gerhard Munz, Sa Li, Georg Carle" *Traffic Anomaly Detection Using K-Means Clustering*".

[10] James M. Coughlan A.L. Yuille" *Manhattan World: Orientation and Outlier Detection by Bayesian Inference*".

[11] Dr. Shuchita Upadhyaya, Karanjit Singh "Nearest *Neighbour Based Outlier Detection Techniques"* ISSN: 2231-2803 International Journal of Computer Trends and Technology- Volume3, Issue2, 2012.

[12] UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[13] J. Han and M. Kamber. "*Data Mining: Concepts and Techniques"* (2nd Ed.). Morgan Kaufmann, San Francisco, CA, 2006.