ISSN:0975-9646

Sreevidya S S et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 8153-8156

A Survey on Outlier Detection Methods

Sreevidya S S

Dept of Computer Science, Sarabhai Institute of Science and Technology,

Vellanad, Thiruvananthapuram, India

***Abstract*- Outlier detection is an active area for research in data set mining community. Finding outliers from a collection of patterns is a very well-known problem in data mining. Outlier Detection as a branch of data mining has many applications in data stream analysis and requires more attention. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the data set. 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, etc. This paper focuses to clarify the problem with detecting outlier over data stream and specific techniques used for detecting outlier over streaming data in data mining. Also this study is focusing on outlier detection techniques and recent research on outlier analysis.**

***Keywords*-Outliers, data mining, data stream, fraud detection**

I. INTRODUCTION

Data mining extracts hidden and useful information from the data. Valid, previously unknown, useful and high quality knowledge is discovered by data mining. Outlier detection is an important task in data mining. Outlier detection has many important applications and deserves more attention from data mining community. Outlier detection is an important branch in data pre-processing and data mining, as this stage is required in elaboration and mining of data coming from many application fields such as industrial processes, transportation, ecology, public safety, climatology. Outliers are data which can be considered anomalous due to several causes. Outlier detection techniques are used, for instance, to minimize the influence of outliers in the final model to develop, or as a preliminary pre-processing stage before the information conveyed by a signal is elaborated. On the other hand in many applications, such as network intrusion, medical diagnosis or fraud detection, outliers are more interesting than the common samples and outliers detection techniques are used to search for them.

In the presence of outliers many data mining and machine learning algorithms and techniques for statistical analysis may not work well. Accurate and efficient removal of outliers may greatly enhance the performance of statistical and data mining algorithms and techniques. Data cleaning is the process of detecting and eliminating outliers as a pre-processing step. As can be seen, different domains have different reasons for discovering outliers. They may be noise that we want to remove. Detecting outliers has important applications in data cleaning. Also outlier detection is used in the mining of abnormal points for fraud detection, intrusion detection, stock market analysis, network sensors, and marketing.

II. OUTLIERS

Informally, an outlier can be defined as any data value that seems to be out of place with respect to the rest of data. Numerous definitions have been proposed for outlier in data mining, such as outlier is an outlying observation, or outlier is one that appears to markedly from other members of the sample in which it occurs. Another definition for outlier is, it's an observation that deviates so much from other observations.

III. VARIOUS OUTLIER DETECTION TECHNIQUES

Outlier detection is varies in accordance with different entities in different domains. Formulation of 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 uses certain kind of statistical distribution and computes the parameters by assuming all data points have been generated by statistical distribution. In this approach outliers are points that have a low probability to be generated by the overall distribution

Statistical outlier detection technique is also known as parametric approach. This technique is formulated by using the distribution of data point available for processing. Detection model is formulated to fit the data with reference to distribution of data. A Gaussian mixture model was proposed by Yamanishi et. al.[1]. 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 [2] where it combines a Gaussian mixture model and supervised method

Depth based outlier detection [3] is one of the variant of statistical outlier detection. Depth based outlier detection search outliers at the border of the data space bur independent of statistical distributions. These techniques are generally suited quantitative real-valued data sets or quantitative ordinal data distributions. In this approach each data object of dataset represented by an 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. Outliers are objects on outer layers. These models are not suitable for high dimensional data set.

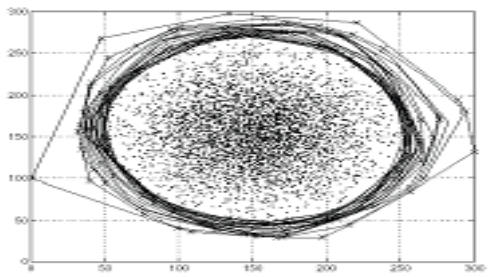

Fig 1: Picture taken from [Johnson et al. 1998]

B. Deviation Based Outlier Detection

In this method a set of data points is given (local group or global set). In this method outliers are detect as points that do not fit to the general characteristics of that set. So the variance of the set is minimized when removing the outliers.

Data elements scattered as like a sparse matrix in data set. This will creates confusion over data analysis. Some points are get deviated from normal points in scattered form, such points are declared as outliers. Sequential problem approach was proposed in [4] 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 [4]. While considering this method not suitable for streaming data. So it is observed that finding deviates over streaming data in distributed environment and over multivariate data is left as open.

C. Distance Based Outlier Detection

Distance based outlier detection technique judge a point based on the distance(s) to its neighbors. Basic model of distance based outlier detection as shown in figure.

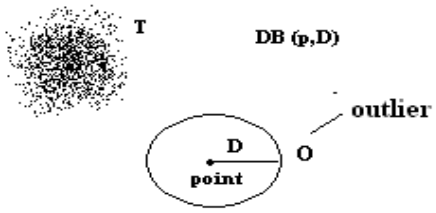

Fig. 2: Basic model of distance based method.

Explicit distance-based approaches are based on the well known nearest-neighbor principle. Ng and Knorr [5] propose a well-defined distance metric to detect outliers. They define outlier as the object which is greater in distance to its neighbors. The basic algorithm, the nested loop (NL) algorithm, calculates the distance between each pair of objects and then set as outliers those that are far from most objects. The NL algorithm has quadratic complexity, with respect to the number of objects, making it unsuitable for mining very large databases such as those found in government audit data, clinical trials data, and network data.

The distance-based outlier method was presented in Knorr and Ng (1998) as: “An object O in a dataset T is a DB (p,D)-outlier if at least fraction p of the objects in T lie at a distance greater than D from O”. The parameter *p* used here is the minimum fraction of objects that must lie outside an outlier's D-neighborhood.

This approach is further extended in [6]. In which they give a prior consideration on distance of a point from its k the nearest neighbor. Where top k point are declared as an outliers. This approach alternatively proposed by Angiulli and Pizzuti [7] on the basis of outlier factor. Each data point is assigned formulated outlier factor computed as sum of distance from its k nearest neighbors. For detecting outliers linear time is used in [8] 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 [9]. 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 [10].

D. Density Based Outlier Detection

This method compares the density around a point with its local neighbors densities. The relative density of a point compared to its neighbors is computed as an outlier score. Density based outlier detection method uses density distribution of data points within data set. The idea of density based local outlier using comparison with density of local neighborhood was introduced by Breuing et al. [11]. 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.

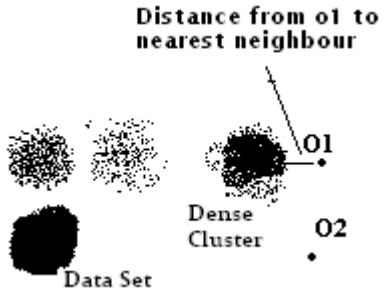

Fig. 3: Data set and dense cluster with outliers.

Papadimitriouet et al [12] 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. Clustering partitions the data into groups, in which similar objects are contained. The assumed behavior of outliers is that they either do not belong to any cluster, or are forced to belong a cluster where they are very different from other members or belong to very small clusters [12].

A variation regarding clustering is the use of a fuzzy model. Fuzzy clustering assigns a membership degree to each sample for each cluster. The most popular known fuzzy clustering algorithm is the *Fuzzy C means (FCM)*. The fuzzy C means is an unsupervised clustering algorithm due to Dunn (1974) and it is based on the minimization of an objective function which is defined as the weighted sum of squared error within groups, as described in the following equation:

$$J_m(U,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c} {u_{ik}}^m \left\| x_k - v_i \right\|^2 ,$$

where $V = (v1 , v2 ,..., vc)$ is the vector of the centers of the clusters, u_{ik} is the grade of membership of data $x_k \in X$ to the cluster *i*. When a stable condition is reached the iteration stops and a point is associated to the cluster for which the value of membership is maximal.

F. Sliding-Window Based Outlier Detection

Streaming data uses a sliding window concept and different multi pass algorithms are used for detecting outliers within streaming data ([13],[14]). Some outliers were considered as inliers in other window, so this method is not efficient. Major problem is that sometimes outlier point may get classified as a inliers [14]. Choosing accurate window size in sliding window based outlier detection is required. 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

Outlier detection over time series data is mostly done using auto regression based outlier detection technique [15]. Auto-regression is also adopted for some outlier detection over streaming data [16], [17]. 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. Streaming data are dynamic in nature where data pattern frequently changes, so it is difficult to select an appropriate model for data streams .

IV. RESULTS AND DISCUSSION

In data mining outlier detection is an important task. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. Construction of a model that accurately represents the data is required for effective outlier detection. 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. Statistical Based Outlier Detection

- Statistical method constructs data as distribution model.
- Based on this model it declares a point as outlier.
- It is applicable for single dimension.
- Parametric assumption does not hold well on distributional data set.
- Data distribution is fixed.
- Low flexibility (no mixture model).
- Designed as a global method.
- Outputs both label and score.

B. 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 which causes curse of dimensionality.
- The idea used is similar like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution.
- Convex hull computation is efficient only in 2D / 3D spaces.
- Originally outputs a label and it can be extended for scoring easily (take depth as scoring value).
- For outlier detection use a global reference set.

C. Deviation Based Outlier Detection

- Idea used in this method is similar like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution.
- Best first search or random sampling are applied.
- Applicable to any data type (depends on the definition of SF).
- Designed as a global method.
- Outputs a labeling.

D. 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.
- Operate on whole data.
- Cannot give number of clusters.
- Hence computation time will increase.
- Only one value is given as most expected outlier.
- 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.

E. Clustering-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.
- Outlier detection also highly depends upon type of clustering used.
- Groups data in to number of clusters
- Reduce the size of database, it will reduce computation time.
- To each cluster user can give certain radius to find outliers.

- A major limitation of clustering-based approaches to outlier detection is that they require multiple passes to process the data set.

F. 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.

G. 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

Table I shows the comparison of the above mentioned systems based outlier detected. The above defined system were evaluated using many datasets, the table1 only shows their maximum cases.

TABLE I:
Comparison of Outlier Detection Methods

Approach	**Outlier Detected (%)**
Statistical based outlier detection	78%
Depth based outlier detection	85%
Deviation based outlier detection	80%
Distance based outlier detection	84%
Clustering based outlier detection	88%
Sliding window based outlier detection	75%
Auto regression based outlier detection	75%

V. CONCLUSION

A large number of techniques have been proposed in outlier detection area, but most of them have some inherent limitations. Outlier detection over streaming data is an important research problem in data mining community. Detecting outlier is important because it contains useful information which may lead for further research in domain This paper provide a review of outlier detection methods over streaming data with data mining perspective. Based on the review we can conclude that most of the techniques used are focuses over algorithms. These require a special background and notion of finding outlier also varies from domain to domain. It is observed that efficiency of outlier detection method is highly dependent upon data distribution and type of data. Some techniques mentioned in this paper require a prior knowledge about data. For instance that statistical technique uses a data distribution and model. Also the assumption based method can work quite well if prior assumption made about data is correct. The 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.

ACKNOWLEDGEMENT

I sincerely express my gratitude to everyone who supported me to prepare this paper. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the work.

REFERENCES

[1] 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.

[2] 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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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

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

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

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

[16] 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.

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