International Journal of Applied Research & Studies ISSN 2278 – 9480

Review Paper

Outlier Mining in Large Data Set: An Efficient Unsupervised Approach

Authors

[1] R.Prabahari*, [2] Dr.V.Thiagarasu, [3] P.Narendran

Address For correspondence:

[1] Assistant Professor, Dept. of Information Technology, Gobi Arts & Science College, Gobichettipalayam
[2, 3]Assoc. Professor, Department of Computer Science, Gobi Arts & Science College, Gobichettipalaaym

Abstract— Clustering can handle the unsupervised patterns such as observations, data items, or feature vectors into groups and also clustering used for outlier detection, where outliers are values that are far away from any cluster. An outlier in a dataset is an observation that is considerably different from the reminders as if it is generated by different mechanism. Mining for outlier is an important data mining research and there are various approaches for detecting outliers such as statistical based approaches, distance based approaches, cluster based approaches, density based approaches etc., the approaches in density based clustering outlier mining, are used to capture outliers. The clusters which are formed based on the cluster elements such as similarity, data set and parameters.

Keywords- *Cluster analysis, Unsupervised learning, Outlier Mining Approaches, Density based clustering.*

I. INTRODUCTION

Data objects or elements that are entirely different from others or inconsistent in comparison to other data elements referred Outliers [figure 1]. Outlier data do not comply with the general behavior of the database or data model, and exhibit deviant and aberrant behavior. It could also be viewed as the process of clustering, but with the difference that clusters look out for the objects or records that have the least similarity and different behavior compared to the rest of the data. Mining for outliers is an important data mining research with numerous applications, including credit card fraud detection[12], identifying computer network intrusions[4,13],detecting employers with poor injury histories[5],discovery of criminal activities in e-commerce, weather prediction, marketing and customer segmentation. Two key phases of outlier mining are identifying the inconsistent data in the large input database and the extraction of the expected number of outliers or deviant data points. The commonly used approaches are statistical based approaches, distance based approaches, cluster based approaches and density based approaches.

The Statistical model starts out with a distribution or probability model for the given data set and then looks out for deviation from the considered model. Distance based technique carries forward the concept used in clustering, with the modified objective of grouping or looking out for data points that lie in far distances. In unsupervised learning, outliers that are considered as noise and, because they can severely affect the results of clustering, are removed from analysis. Clustering based methods that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters[7].Density based outlier detection is closely related to distance based outlier detection since density is usually defined in terms of distance. One common approach is to define as reciprocal of the average distance to the k nearest neighbors. If the distance is small, the density is high, and vice versa [9]. Some of the above mentioned models are further discussed below.

II. STATISTICAL- BASED APPROACHES

The statistical approach to outlier analysis assumes a probability distribution or model for the given data set then perform a satisfiability test, also referred to as discordancy test to determine if data objects comply with the assumed model or not. Objects that buck the model are treated as outliers. The model requires that the data distribution, knowledge of distribution parameters, etc., be known to start with. It has its base test of goodness and fitness, where a hypothesis is accepted or rejected based on computed statistical parameters [7]. A probabilistic model can be either a prior given or automatically constructed by given data. Having constructed the probabilistic model, one sets the problem of determining whether a particular object of the data belongs to the probabilistic model or it was generated in accordance with some other distribution law. If the object does not suit the probabilistic model, it is considered to be an outlier.

prabahari_r@rediffmail.com * Corresponding Author Email-Id

Probabilistic models are constructed with the use of standard probability distributions and their combinations [9].

A. Smart Sifter

Another approach to detecting outliers by statistical method is implemented in the Smart Sifter algorithm [9]. The basic idea of this algorithm is to construct a probabilistic data model based on observations. In this case, only the model, rather than the entire dataset, is stored. The objects are processed successively, and the model learns while processing each data object. A data object is considered to be an outlier if the model changes considerably after processing it. For these purposes, a special metrics, the outlier factor, is introduced to measure changes in the probabilistic model after adding a new element [9].

B. Regression Analysis

Methods for detecting outliers based on the regression analysis are also classified among statistical methods. The regression analysis problem consists in finding a dependence of one random variable Y on another variable X. Specifically, the problem is formulated as that of examining the conditional probability distribution *Y /X*. In the framework of the first approach, the regression model is constructed with the use of all data, and then the objects with the greatest error are successively, or simultaneously, excluded from the model. This approach is called a reverse search. The second approach consists in constructing a model based on a part of data and, then adding new objects followed by the reconstruction of the model. Such a method is referred to as a direct search. Then, the model is extended through addition of most appropriate objects, which are the objects with the least deviations from the model constructed. The objects added to the model in the last turn are considered to be outliers. Basic disadvantages of the regression methods are that they greatly depend on the assumption about the error distribution and need a prior partition of variables into independent and dependent ones[9].

III. DISTANCE-BASED APPROACHES

Local distance-based algorithms also use distances between objects from the dataset being analyzed. It is based on the principle of measuring distances of objects and then classifying those objects whose distances exceed a specified threshold. The basic distance-based approach is that implemented in the *DB* (*p, D*) method [5]. A database D, an object O is termed an outlier when at least a fraction of the objects in D are at distance greater than D from O .All those objects that lie at a greater distance are interpreted a specific model O which is deviant from the fraction of objects F in the database D and are treated as outliers. In terms of neighborhood analysis, distance based method concentrate on identifying those objects in the database which do not have a sufficient amount of neighborhood objects in the database.

A. Index Based Algorithm

Index based algorithm uses multidimensional indexing structures, to search for neighbors of each object O within radius d around that object. Let M be the maximum number of objects within the dmin-around that object. Let M be the maximum number of objects within the dmin-neighborhood of an outlier. Therefore, once M+1 neighbors of object O are found, it is clear that O is not an outlier. This algorithm has a worst case complexity of O (n^2k), where n is the number of objects in the data set and k is the dimensionality. The index based algorithm scales well as k increases. However, this complexity evaluation takes only the search time into account, even though the task of building an index in itself can be computationally intensive[10].

B. Nested Loop Algorithm

The Nested-loop algorithm has the same computational complexity as the index based algorithm but avoids index structure construction and tries to minimize the number of I/Os. It divides the memory buffer space into two halves and the data set into several logical blocks. By carefully choosing the order in which blocks are loaded into each half, I/O efficiency can be achieved [11].

C. Cell Based Algorithm

In the cell based algorithm, the complexity of examining all pairs of objects is reduced through a preliminary partition of the space into cells and construction of estimates for distances between the objects. The nested-loop algorithm requires not less than $m - 2$ passes, where m is the number of blocks in the dataset [5].

IV. CLUSTER BASED APPROACHES

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 to a cluster where they are very different from other members. Clustering based outlier detection techniques have been enveloped which make use of the fact that outliers do not belong to any cluster since they are very few and different from the normal instances [6].

V. DENSITY BASED APPROACHES

Density-based local outliers based on the density in the local neighborhood. Each data point is assigned a local outlier factor (LOF) value, which is calculated by the ratio of the local density this point and the local density of its MinPts nearest neighbors. The single parameter MinPts of a point determines the number of its nearest neighbors in the local neighborhood. The LOF value indicates the degree of being an outlier depending on how isolated the point is with respect to the density of its local neighborhood. Points that have the largest LOF values are considered as outliers [14].

VI. DISCUSSION

Statistical based methods have been proposed during the earlier works on outlier detection and they are mostly applicable to datasets that have a single variable. These methods have significant drawbacks that limit their applicability. They require that the dataset follow some standard distribution. In Distance based methods, the outlier definition is based on a single global value of the given parameters. If the dataset has both dense and sparse regions, this can lead to problems. If the neighborhood is specified to be

large, then some outliers will not be detected. Clustering-based approaches are used to detect clusters and not outliers. Therefore, they may not be optimized for detecting outliers and do not require a prior knowledge of data distribution and exploit clustering techniques to filter efficiently and remove outliers in large data sets[7]. The advantage of the clustering-based approaches is that they do not have to be supervised. Density based methods is used to overcome the drawbacks of distance based methods. One drawback with density based methods is that require a parameter k-distance and the quality of result is sensitive to the selection of this parameter. Clustering objects of a database into meaningful subclasses is one of the major data mining methods [8].Among many types

of clustering algorithms, density based algorithm is more efficient in detecting the clusters with varied density.

VII. DENSITY BASED CLUSTERING APPROACHES

Outlier detection is important in many fields and concept about outlier factor of object is extended to the case of cluster. Both Statistical and distance based outlier detection depend on the overall or "global" distribution of the given set of data points. Data are usually not uniformly distributed. This method encounters difficulties when analyzing data with rather different density distribution. This brings the notion of local outliers. Density-based approaches apply a local cluster criterion. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise). These regions may have an arbitrary shape and the points inside a region may be arbitrarily distributed. The clusters which are formed based on the density are easy to understand and it does not limit itself to the shapes of clusters. Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. An object is local outlier if it is outlying relative to its local neighborhood, with respect to the density of the neighborhood. This forms the basis of density based local outlier detection. Another one approach, it does not consider being an outlier as binary property. Instead, it assesses the degree to which an object is an outlier. This degree of "outlierness" is computed as the local outlier factor of an object. It is local in the sense that the degree depends on how isolated the object is with respect to the surrounding neighborhood. This approach can detect both global and local outliers. A cluster as dense regions of objects in the data space that are separated by regions of low density. DBSCAN grows clusters according to a density-based connectivity analysis. OPTICS [1] extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings. DENCLUE [6] clusters objects based on a set of density distribution functions. LOF [2] uses a more meaningful way to assign to each object a degree of being an outlier than to consider being an outlier as a binary property. LDBSCAN [3] combines the concepts of DBSCAN and LOF to discover clusters and outliers. There are two potential benefits of combining clustering and outlier detections are increasing precision and facilitate data understanding.

VIII. CONCLUSION

Outliers can be caused by measurement or execution error. Many data mining algorithm try to minimize the influence of outliers or eliminate them all together. This however could result in the loss of important hidden information because "one person's noise could be another person's signal". So the outlier detection is important task. Thus, detection can be commonly categorized into four approaches which are statistical approach, distance based approach, cluster based approach, and density based approach. Through this, the various outlier approaches are discussed based on the similarity and data set. These limitations are overcome by using the combination of cluster and density that is density based clustering approaches, for used in outlier detection in large data sets.

REFERENCES

[1] [1] Ankerst, M.; Breunig, M. M. Kriegel, H.-P. & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. In Proceedings of the 1999 ACM SIGMOD international conference on Management of data (SIGMOD '99). ACM, New York, NY, USA, 49-60.

[2] [2] Breunig, M. M. Kriegel, H.-P. ; Ng, R. T. & Sander, J. (2000), LOF: identifying density-based local outliers, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, p.93-104, May 15-18, 2000, Dallas, Texas, United States.

[3] [3] Duan, L. Xu, L. ; Guo, F. ; Lee, J. & Yan, B. (2007). A local-density based spatial clustering algorithm with noise. Inf. Syst. 32, 7 (November 2007), 978-986.

[4] [4] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Data Mining for Security Applications, 2002.

[5] [5] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: algorithms and applications.

[6] VLDB Journal: Very Large Databases, 8(3-4):237{253, 2000}

[7] [6] Hinneburg, A. & Keim, D. A. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, pp. 58-65.

[8] [7] Jiawei Han and Micheline Kamber "Data Mining: Concept and Techniques" second edition [8] Matheus C.J., Chan P.K., and Piatetsky-Shapiro G. 1993. "Systems for Knowledge Discovery in Databases*".IEEE Transactions on Knowledge and Data Engineering 5(6): 903-913.*

[9] [9] M. I. Petrovskiy "Outlier Detection Algorithms in Data Mining Systems**"** *Department of Computational Mathematics and Cybernetics, Moscow State University, Vorob'evy gory,Moscow, 119992* Received February 19, 2003.

[10] [10] N.P. Gopalan and B.Sivaselvan "Data Mining: Techniques and Trends" Eastern Economy Edition.

[11] [11] Pang-Ning Tan, Vipin Kumar,Michael Steinbach "Introduction To Data Mining"

[12]] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review (with discussion). Statistical Science, 17(3):235{255, 2002}.

[13] T. Lane and C. E. Brodley. Temporal sequence learning and data reduction for anomaly detection. ACM Transactions on Information and System Security, 2(3):295{331, 1999}.

[14] Yang Zhang, Nirvana Meratnia, Paul Havinga "A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets "Department of Computer Science, University of Twente, P.O.Box 217 7500AE, Enschede, the Netherlands.