Outliers Detection using Subspace Method: A Survey

Supriya Garule
M.E. Student,
Computer Science Department
JSCOE, Hadapsar, Pune, India

Sharmila M. Shinde
Faculty
Computer Science Department
JSCOE, Hadapsar, Pune, India

ABSTRACT
Outliers detection is currently very active area of research in data set mining community. Outliers detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the database. In this paper, we present a survey of outliers detection techniques using subspace method. The survey will not only cover the high dimensional datasets but also review the more recent developments that deal with more complex outliers detection problems in high-dimensional dataset.

Keywords
Data Mining, Outliers Detection, High-dimensional Datasets, Subspace, Outliers Ranking.

1. INTRODUCTION
Traditional outliers detection techniques operate on structured data such as corporate databases; they are not suitable for high-dimensional. So the need arise for developing new algorithms which is 1)Effective treatment of high dimensionality. 2) Interpret ability of results 3) Scalability and usability. 4) Find true outliers.

In high dimensionality, the data becomes sparse and all pairs of data points become almost equidistant from one another. From a density perspective, all regions become almost equally sparse in full dimensionality. Thus the extreme value deviations based on the distances in full dimensionality is meaningless. The reason for this behavior is that many dimensions may be very noisy, and they may show similar pairwise behavior in terms of the addition of the dimension-specific distances.

Subspace outliers detection is the best solution to the Curse of Dimensionality in outliers detection. Subspace outliers detection is a relatively new sub-field of outliers detection. For very high dimensional data, subspace clustering algorithms attempt to find subspaces underlying the data and then find clusters in each such subspace. The final clustering is obtained by combining the clusters in each subspace. Basically, the individual representations in each subspace will remove the overall clutter, and lead to clear view of the data which will be easily detect the hidden outliers

2. OUTLIER DETECTION IN HIGH DIMENSIONAS
The main aim of outlier detection is to uncover the "different mechanism".High Dimensional Approaches Genetic Algorithm,Improved Clustering methods[6], Evolutionary algorithms[3] , Subspace Outliers Detection [10] are some high dimensional techniques. Now a days data recording is more easy and less costly. Digitalization of the every thing is main source of high dimensional data so the data sets usually have many features.

2.1 Outliers Detection Challenges in High-Dimensional Data
Detecting outliers without saying why they are outliers is not very useful in high-D due to many features (or dimensions) are involved in a high-dimensional data set E.g., which subspaces that manifest the outliers or an assessment regarding the "outliers-ness" of the objects. Data in high-D spaces are often sparse. The distance between objects becomes heavily dominated by noise as the dimensionality increases. Adaptive to the subspaces signifying the outliers. Capturing the local behavior of data is one of the challenge,Need scalable data with respect to dimensionality of subspaces increases exponentially that is nothing but the **Curse of dimensionality.**

3. DIMENSIONALITY REDUCTION
Variables (attributes) where carefully evaluated if they are relevant for the analysis task.Data sets usually contain only a few number of relevant dimensions. Dimensionality reduction achieved by extracting 'important' features from the dataset. It is desirable to avoid the "curse of dimensionality" in high dimensional data and concentrate on relevant features of data to reduce computational Cost and Time.

4. SUBSPACE OUTLIER DETECTION METHODS
Main aim of the subspace outliers detection methods are finding outliers in relevant subspaces that are not outliers in the full-dimensional space (where they are covered by "irrelevant" attributes). To identify which subspace is relevant there are many techniques developed. Here overview of some latest and most important techniques is given.

4.1 Projected Outliers with Grids
First approach for high-dimensional (subspace) outliers detection is given in Aggarwal and Yu [2001].This approach resembles a grid-based subspace clustering approach but not searching dense but sparse grid cells. Report objects contained within sparse grid cells as outliers evolutionary search for those grid cells (Apriori-like search not possible, complete search not feasible) problems with this approach is increasing dimensionality, the expected value of a grid cell quickly becomes too low to find significantly sparse grid cells.

4.2 HOS-Miner
Zhang et al. [2004] identify the subspaces in which a given point is an outlier define the outlying degree of a point w.r.t. a certain space (or possibly a subspace) s in terms of the sum of distances to the k nearest neighbors in this (sub-)spaces. Problems of HOS-Miner is fixed threshold to discern outliers w.r.t. their score OD in subspaces of different dimensionality. These scores are rather incomparable.The monotonicity must not be fulfilled for true subspace outliers (since it would imply that the outlier can be found trivially in the full-dimensional space). Systematic search for the subspace with the highest

score is like data-snooping bias.

4.3 OutRank

Müller et al. [2008] analyse the result of some(grid-based/density-based) subspace clustering algorithm. Clusters are more stable than outliers to identify in different subspaces .They avoids statistical bias. How often is the object recognized as part of a cluster and what is the dimensionality and size of the corrletated subspace clusters it is called outlierness. Thera are some problems with OutRank such as a strong redundancy in the clustering is implicitly assumed. Loke result biased towards (anti-)hubs.Outliers as just a side-product of density-based clustering can result in a large set of outliers.

4.4 SOD

SOD (subspace outlier detection) [Kriegel et al., 2009a] finds outliers in subspaces without an explicit clustering a reference set is possibly defining (implicitly) a subspace cluster (or a part of such a cluster). If the query point deviates considerably from the subspace of the reference set, it is a subspace outlier w.r.t. the corresponding subspace. subspace distance outlier score is not based on a decision like outlier vs. Inlier but a normalized, sort of point,.Some limitations are how to find a good reference set and normalization of scores is over-simplistic.

4.5 OUTRES [Müller et al., 2010]

Müller et al.,[2010] found a method OUTRES which assess deviations of each object in several subspaces simultaneously .They combine ranking of the objects according to their outlier scores in all 'relevant subspaces'. It requires comparable neighborhoods for each point to estimate densities and adjust for different number of dimensions of subspaces .Score in a single subspace is comparing the object's density to the average density of its neighborhood .Total score of an object is the product of all its scores in all relevant subspaces.There are some disadvantages of this method such as an apriori-like search strategy finds subspaces for each point, not outliers in the subspaces which is quite expensive approach.Worst-case exponential behavior in dimensionality score increses the cost. Time complexity becomes O(n3) for a database of n objects unless suitable data structures (e.g., precomputed neighborhoods) are used and due to the adaptation to different dimensionality of subspaces, data structure support is not trivial.

4.6 HighDOD

HighDOD (High-dimensional Distance-based OutlierDetection) [Nguyen et al., 2011] was introduced Motivation behind it was the sum of distances to the k nearest neighbors as the outlier score is monotonic over subspaces but a subspace search (as in HOS-Miner) is pointless as the maximum score will appear in the full-dimensional space. Modify the kNN-weight outlier score to use a normalized Lp norm.Pruning of subspaces is impossible as need to examine all subspaces up to a user-defined maximum dimensions m. It use a linear-time (O(n · m)) density estimation to generate outlier candidates they compute the nearest neighbors for. But examine all subspaces is like data-snooping. No normalization to adjust different variances in different dimensionality.

4.7 HiCS [Keller et al., 2012]

In high contrast subspaces (HiCS), core concept for subspaces with high contrast.The correlation among the attributes of a subspace.They aggregate the LOF scores for a single object over all "high contrast" subspaces. So instead of LOF, it could be better to use any other outlier measure.In these subspaces, outliers are not trivial (e.g., identifiable already in 1-dimensional subspaces) but deviate from the (although probably non-linear and complex) correlation trend exhibited by the majority of data in this subspace. But as combine LOF scores from subspaces of different dimensionality without score normalization causes Bias Scores problem causes the problem of Bias of Scores. Also combination of scores is rather naïve, could benefit from ensemble reasoning.To identify interesting subspaces will relate quite differently to different outlier ranking measures, their measure of interest is based on an implicit notion of density,.It may only be appropriate for density-based outlier scores however, this decoupling allows them to discuss the issue of subspace selection with great diligence as this is the focus of their study.

Table 1. Existing Outliers Detection Methods

Outliers Detection Methods	**Outliers Detection Ratio for HD Data**	**Suitable For High-dimensional Data**
Statistical based Methods	Low	NO
Distance based methods	Very Low	NO
Density based methods	Low	NO
Clustering based Methods	High	Not all (HD based Clustering detects)
Subspace methods	Very High	YES

5. PROPOSED APPROIVACHE

As there are various techniques related to find Outlier detection but no one give us true and exact outlier as outlier definition change with respective application and it use.So here we try to combine some techniques is such a way that we can get genuine outliers with less complexity and with minimum cost. Here we will first apply data pre processing techinique.In this we use data reduction so that we can extract only relevant attributes and pruning of non relevant attributes. Then we will use subspace method and clustering method in each subspace so that we can get outliers. Finally we will try to apply Outlier Ranking method in which we will find score of each outliers and then sort them according to the increasing score and the top ranking score outliers are nothing but the genuine outliers.

6. BLOCK DIGRAM OF PROPOSED SYSTEM

Fig 1. Shows the overall working of the system. For the given dataset here we applied data reduction techniques so that all non relevant attributes minimizes or reduced. Its like pruning of attributes. Then use one of the subspace outliers detection technique to find out the Outliers from relevant data. Here analysis of the data done on the basis of subsets instead of full data analyze at once. So the profitability to get hidden or missing outliers became more. Calculate degree of outlierness of each outliers based on scoring function, then applied Outliers Ranking algorithm to sort them on increasing order.

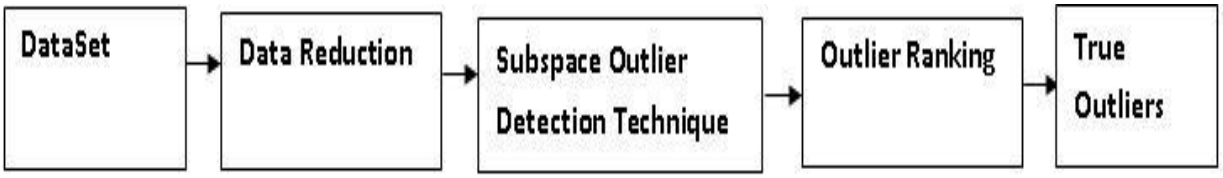

Fig 1: Block Diagram of the system

7. CONCLUSIONS

There are different outlier detection methods,but no one is complete and give the best solution. In this research paper we discuss about various outlier detection methods and their advantage and limitations as well. We also look at recent algorithms like Subspace Method to solve this problem. Here we proposed one system in which we will try to we combine techniques like Preprocessing of data, dimensionality reduction and subspace method ,Outlier ranking, which will surely get better solution to find hidden genuine with less complexity and cost.

8. ACKNOWLEDGMENTS

I am grateful to a number of individuals, specially I express my deep and sincere sense of indebtedness to Prof. S. M. Shinde Department of Computer Engineering for her valuable guidance, pain taking effort, constant encouragement and inspiration during each and every step of my paper.

9. REFERENCES

[1] E. Muller, I. Assent, U. Steinhausen, and T. Seidl,Outlier Ranking via Subspace Analysis in Multiple Views of the Data, in ICDM, 2012.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, in SIGMOD, 1998,pp. 94105.

[3] C. C. Aggarwal, Outlier Analysis, in Springer,2013, pp. 6172.

[4] M. L. Yiu and N. Mamoulis, Frequent-pattern based iterative projected clusterng, in ICDM, 2003, pp. 689692.

[5] K. Sequeira and M. Zaki, SCHISM: A new approach for interesting subspace mining, in ICDM, 2004, pp. 186193.

[6] Ji Zhang , Advancements of Outlier Detection: A Survey , in ICST Transactions on Scalable Information Systems January-March 2013, Volume 13 ,Issue 01-03.

[7] Assent, R. Krieger, E. Muller, and T. Seidl, INSCY: Indexing subspace clusters with in-process-removal of redundancy,in ICDM, 2008, pp. 719724.

[8] G. Moise and J. Sander, Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering,in KDD, 2008, pp. 533 541.

[9] E. Muller, I. Assent, S. Gunnemann, R. Krieger, and T. Seidl,Relevant Subspace Clustering: Mining the Most Interesting Non-redundant Concepts in High Dimensional Data, in ICDM, 2009, pp. 377386.

[10] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong, A Survey on Enhanced Subspace Clustering, DMKD, 2012.

[11] E. Muller, S. Gunnemann, I. Assent, and T. Seidl, Evaluating clustering in subspace projections of high dimensional data,PVLDB, vol. 2, no. 1, pp. 12701281,2009.

[12] C. C. Aggarwal and P. S. Yu, Outlier detection for high dimensional data, in SIGMOD, 2001, pp. 3746.

[13] H.-P. Kriegel, E. Schubert, A. Zimek, and P. Kroger, Outlier detection in axis-parallel subspaces of high dimensional data,in PAKDD, 2009, pp. 831838.

[14] E. Muller, M. Schiffer, and T. Seidl, Statistical selection of relevant subspace projections for outlier ranking, in ICDE,2011, pp. 434445.

[15] E. Muller, I. Assent, U. Steinhausen, and T. Seidl, OutRank:ranking outliers in high dimensional data, in ICDE Workshops,DBRank. IEEE, 2008, pp. 600603.

[16] P. Rousseeuw and A. Leroy, Robust Regression and Outlier Detection. Wiley,1987.

[17] V. Chandola, A. Banerjee, and A. Kumar, Anomaly detection:A survey, ACM Computing Surveys, Vol. 41, No.3, July 2009.