1-1-2010

A Novel Subspace Outlier Detection Approach in High Dimensional Data Sets

Jinsong Leng
Edith Cowan University

A Novel Subspace Outlier Detection Approach in High Dimensional Data Sets

Jinsong Leng
School of Computer and Security,
Edith Cowan University,WA, Australia
j.leng@ecu.edu.au,

Abstract—**Many real applications are required to detect outliers in high dimensional data sets. The major difficulty of mining outliers lies on the fact that outliers are often embedded in subspaces. No efficient methods are available in general for subspace-based outlier detection. Most existing subspace-based outlier detection methods identify outliers by searching for abnormal sparse density units in subspaces. In this paper, we present a novel approach for finding outliers in the 'interesting' subspaces. The interesting subspaces are strongly correlated with `good' clusters. This approach aims to group the meaningful subspaces and then identify outliers in the projected subspaces. In doing so, an extension to the subspace-based clustering algorithm is proposed so as to find the 'good' subspaces, and then outliers are identified in the projected subspaces using some classical outlier detection techniques such as distance-based and density-based algorithms. Comprehensive case studies are conducted using various types of subspace clustering and outlier detection algorithms. The experimental results demonstrate that the proposed method can detect outliers effectively and efficiently in high dimensional data sets.**

Keywords- Data Mining, Subspace Clustering, Outlier Detection, Dimensional Reduction

I. INTRODUCTION

Finding outliers is a challenging data mining task, especially for high dimensional data sets. The notion of outliers can be defined from different perspectives. Hawkins [5] defines an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Another definition is given by Barnett and Lewis in [2]: "An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of that dataset".

Normally, classical outlier detection techniques include distance-based, density-based, and distribution-based methods. The pioneer work by Knorr and Ng formalized the notion of outliers in terms of distance [6]. An outlier is defined as: "An object O in a dataset T is a DB(p, D)-outlier if at least a fraction q of the other objects in dataset T lies greater than distance D from O ". This definition can identify 'global' outliers effectively, but cannot detect 'local' outliers if the data set consists of clusters of diverse density. There are two parameters involved, i.e., the fraction *p* and the distance D. These parameters can have effects on the performance of the detection techniques. Another simple distance-based outlier definition is given in [7]: "Given an input data set with N points, parameters n and k, a point p is an outlier if there are no more than n -1 other points p' in the data set such that D^k (p') ≤ D^k (p), where D^k (p) denotes the distance of point p from its k^{th} nearest neighbor". This definition has only one parameter: the number of neighbors k. It ranks potential outliers based on the distance (D^k) of a point from its k^{th} nearest neighbor. The top N points with the maximum values D^k are considered as outliers.

Distance-based outlier detection methods rank outliers globally, but they cannot distinguish outliers from data points with diverse density. To overcome this problem, the local outlier factor [3] method mines outliers that deviate from their belong-to clusters, and ranks the outlier degree of data samples on the basis of the density of its local neighborhood.

Breunig et al. [3] proposed a local density-based outlier-detection method to identify local outliers (LOF) based on the local density of a sample's neighborhood. In [3], the LOF is introduced for each sample in the data set, indicating its degree of outlier-ness. The LOF of an object is calculated using the number of its nearest neighbors MinPts. The LOF of an object p represents the degree of outlierness. The LOF algorithm may not be effective with respect to density when its neighbors are sparse [8]. LOF cannot also find the potential outliers when their neighbors have similar densities.

Aggarwal and Yu [1] proposed a subspace outlier detection approach. The approach assumes that data points are based on certain statistical distribution, so potential outliers are those that the density of the data in lower dimensional projections is abnormally lower than average. This is a grid-based method that it first quantizes the object space into a finite number of cells that form a grid structure, and then performs mining algorithms on the grid structure. The search process starts from one-dimensional projections and grows up to higher dimensionality gradually. In this algorithm, the sparsity coefficient is used as the measure criteria, and the evolutionary computation is used as the search strategy to avoid intensive computation. The sparsity coefficient of a given projection is calculated according to its normal distribution. Then, the significance of the dimensions is evaluated in terms of the sparsity coefficient. In this aspect, the problem turns to find the subset of dimensions with the most negative sparsity coefficients.

To address the problems described above, this paper presents a novel approach by identifying outliers in the interesting subspaces. The interesting subspaces are found using some subspace-based clustering algorithms, and outliers are identified using classical outlier mining algorithms.

Rather than searching for outliers in sparse grids, we attempt to find the projected dimensions with strong correlation. Normally, clusters lie in the projection with high densi-

978-1-4244-7224-6 /10/$26.00 © 2010 IEEE

ty. The mathematical root is that the subspaces can be measured by the correlation criteria among data samples. In doing so, we use subspace-based clustering to find the interesting subspaces. After that, the outliers embedded in the interesting subspaces are detected by using distance-based or density-based outlier detection techniques. This approach is able to provide a promising result over high-dimensional data sets, and also can avoid the intensive computation load as compared with other subspace outlier detection methods.

This paper makes the contributions as follows: A novel approach is proposed for finding outliers in the interesting subspaces with tight clusters. The proposed approach takes advantage of some existing techniques, i.e., subspace clustering, distance-based and density-based outlier detection methods. Comprehensive case studies have been conducted with various types of high dimensional data sets to demonstrate the effectiveness of the proposed approach.

The rest of the paper is organized as follows: In Section II, we detail the problems and gives some definitions. Section III introduces the interesting subspaces and identifies them using subspace clustering methods. The algorithm for mining outliers in subspaces is described in Section IV. The experimental results are presented and discussed in Section V. Finally, Section VI concludes the paper.

II. PROBLEM FORMULATION

The existing subspace outlier mining algorithms focus on the identification of abnormal, low-dense projections. These algorithms are not able to determine the degree of correlation among dimensions, and hence no evidence is available about the correlation relationship among dimensions. The existing subspace outlier mining algorithms ignore some classical outlier mining methods, for example, distance-based and LOF (local outlier factor) [3] algorithms, which are able to identify outliers very effectively at lower dimensions. To address the issue of identification of meaningful outliers, we first find the interesting subspaces with tight clusters and with abnormal distributions. Next, we score the outlier-ness in the projected subspaces using existing classical outlier mining algorithms. The fundamental problem is that what kind of criteria can be used to find the interesting subset of dimensions and to further rank the outliers obtained from those projections.

Our approach is different from the existing subspace outlier detection approaches. To the best of our knowledge, there are no similar approaches using classical outlier mining algorithms in high-dimensional data sets. The correlated dimensions can be found on the basis of major distribution of data samples in subspaces. Subspace clustering is a good approach that can find correlated dimensions while not inferring any causal relationship [4]. Since the local feature correlation of dimensions can be determined by the feature of data points among dimensions, subspace clustering methods are a better choice to find the correlated dimensions.

Usually, a matrix is used to represent a data set, in which the columns represent the dimensions or attributes and the rows indicate the objects. Suppose that matrix D with n rows and m columns is used to represent a data set. It can thus be presented as $D = (X, A)$ where:

- X is a set of data objects, $X = \{X_1, ..., Xn\}$;
- A is a set of dimensions, $A = \{A_1, \ldots, A_m\}$;

Definition 1 (z-scores: $z^k(a_i)$ of a data point a_i) $z^k(a_i)$ is the normalization of $d^k(a_i)$ of a data point a_i in each subspace, indicated as $d^k(a_i)/\sigma$.

Definition 2 ($d^k(x_i)$ of a data point x_i) The k^{th} –distance $d^k(x_i)$ of a data point x_i is its k^{th} -nearest neighbor.

Given a value of k, the outlier-ness of data points in D are ranked in terms of the kth -distance of data points. In order to rank the outliers across different subspaces, the z-scores of data points in each subspace can be normalized by the standard derivation σ as $dk(xi) / \sigma$.

Definition 3 ($LOF^k(x_i)$ of a data point x_i) Given a value of k, $LOF^k(x_i)$ of a data point x_i is the local outlierfactor of its k nearest neighbors.

Similarly, we normalize the z-scores of data points in-each subspace using $LOF^k(x_i) / \sigma$.

Definition 4 (Top N outliers) The top N outliers are the N data points in D with the highest z-scores in the full and all interesting subspaces.

III. MINING INTERESTING SUBSPACES

To analyze the correlation among the dimensions of a data set, we introduce the entropy and joint entropy measures. Given a discrete variable X, entropy H(X) describes the uncertainty about the value of X. If X consists of several events x, whereby each occurs with the probability p_x, then the entropy of X is given by:

$$H(X) = -\sum_{x} p_x \log_2(p_x) \quad (1)$$

Definition 5 *The mutual information I(X; Y) is defined as:*

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \quad (2)$$

Mutual information is an important indicator to reveal the non-linear correlation relationship between variables X and Y. Mutual information indicates the amount of uncertainty remaining about X after Y is known, which is equivalent to the amount of uncertainty in X, minus the amount of uncertainty in X which remains after Y is known. Entropy indicates the uncertainty of variables. We can use entropy and mutual information of variables as the measurement criteria to find the correlated dimensions in high-dimensional data sets.

The interest (mutual information) [4] is calculated by:

$$Interest(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(x_i) - H(X_1, \cdots, X_n) \quad (3)$$

Definition 6 (The interesting subspaces) *The interesting subspaces are those with high Interest and tight clusters.*

ENCLUS [4] uses entropy and interest (mutual information) to carry out the downward and upward pruning processes. This algorithm groups subspaces with strong correlation among dimensions. We use the entropy-based clustering algorithm to identify the interesting subspaces, and then distinguish outliers from the projected subspaces using distance-based or density-based algorithms. Accordingly, the

Definition 4 for defining the top *N* outliers is modified as below.

Definition 7 (Top N **outliers)** *Top* N *outliers are* N *data points in* D *with the highest* z*-scores in the interesting subspaces.*

IV. ALGORITHM

Outliers can be ranked in the aggregated view by combining outliers identified in the interesting subspaces. The interesting subspaces are ranked by the goodness of clustering. Based on the interesting subspaces, we are able to calculate *z*-scores in the limited subspaces using distance-based or density-based algorithms. We search for top N data points with the highest z-scores in the reported subspaces using an iterative procedure.

Algorithm 1 is used to mine top *N* outliers in interesting subspaces (MOIS). If the replaced rate (M / N) is smaller than δ for a couple of times, convergence of *R* is achieved, and MOIS is stopped. Algorithm 1 finds the minimal number of interesting subspaces, in order to obtain the consistent top N outliers. However, the reported interesting subspaces have good clustering, and may not consist of high percentage of outliers. Such subspaces may have effect on the precision of top N outliers. In this regard, it is required to further refine the interesting subspaces in terms of their shape factors.

Algorithm 1 MOIS: Mining top N outliers in interesting Subspaces

---

Input: a data set D, integer k, N, threshold ω, ε, and δ
Output: Top N outliers, and minimal number of interesting subspaces
1: Initialize a list O for top N outliers;
2: Initialize a list T for *z*-scores;
3: Calculate z-scores (Distance-based or LOF-based measure) in the full space, and add them with related indexes into T ;
4: Find top N objects in T and add them with related indexes into O ;
5: Call ENCLUS INT (D, ω, ε) [4] to find all interesting subspaces;
6: Rank the reported interesting subspaces in a list L;
7: **for** Each subspace in list L **do**
8: Copy O into a list O_1 ;
9: Calculate new z-scores in the subspace, and add them with related indexes into a list T_1 ;
10: Find top N objects in T_1, and add them into a list O_2 ;
11: Compare O_1 with O_2 , record the duplicate objects with the highest *z*-scores in a list T_2 ;
12: Remove the objects with indexes that exist in T_2 from O_1 with O_2;
13: Merge O_1 , O_2 , and T_2 into a list S ;
14: Sort *S* in descend order (based on *z*-scores);
15: Count the number M of objects (based on indexes) in O being replaced, and add M into a list R;
16: Clear the list of O ;
17: Add the top N objects from S into O ;
18: **if** R converges **then**
19: Break;
20: **end if**
21: Clear the lists of O_1 , O_2 , T_1 , T_2 and S ;
22: **end for**
23: Return O and R ;
24: Find top N outliers in O, and minimal number of interesting subspaces (equivalent to size of R).

We run the algorithm by comparing top *N* outliers of every subspace based on an iterative procedure.

We use the following terminologies to interpret the results:

a) True positive rate =
$$\frac{TurePositive(TP)}{TurePositive(IP) + FalseNegative(FN)}$$

b) False positive rate =
$$\frac{FalsePositive(FP)}{TureNegative(TN) + FalsePositive(FP)}$$

V. EXPERIMENTAL RESULTS

We evaluated the distance-based and LOF-based algorithm of MOIS over the Statlog data set. For the Sonar data set, we set ω=9.0, ε = 0.1, and interest gain = 0.8. Similarly, a performance metric was obtained by tuning the number of *N*. Based on the performance metric, the ROC curves were drawn for comparing the performances. Since the percentages of outliers in data sets are known in advances, we conduct the experiments with actual percentage of outliers.

A. Breast cancer Wisconsin (Diagnostic) Dataset

The Breast Cancer Wisconsin (Diagnostic) (BCWD) data set contains 569 data objects with 32 attributes. It has two classes: malignant and benign. We generated a new data set from BCWD, with 483 data samples (357 of benign and 26 of malignant). This experiment aimed to identify the samples of malignant as outliers in subspaces.

Now we defined the percentage of outliers as 6.8%, i.e., the number of outliers N was 26. We performed distance-based MOIS over the BCWD data set. The results are displayed in Table 1. It is clear that the results indicate that all subspaces and aggregated projections have better performance than that of the full space. In some subspaces, the outliers can be identified effectively, for example, subspace (0,3,23) represents the combination of subset of attributes (0, 3, 23) (starting from index 0), which results in very high hit rate and precision. Next, we performed the LOF-based algorithm MOIS over the BCWD data set. The results of seven subspaces are displayed in Table 2. The results also indicate that all subspaces and aggregated projections performed bet-

ter than that of the full space. The results are slightly different than those with distance-based MOIS.

TABLE I. RESULTS OF DISTANCE-BASED MOIS OVER BCWD

Subspaces	True Classification					
	TP	FP	TN	FN	HR(%)	PS(%)
(20,22)	19	7	350	7	73.1	73.1
(0,3,20)	21	5	352	5	80.8	80.8
(3,20,23)	21	5	352	5	80.8	80.8
(0,3,23)	22	4	353	4	84.6	84.6
(0,3,22)	21	5	352	5	80.8	80.8
(0,22,23)	21	5	352	5	80.8	80.8
Full Space	11	15	342	15	42.3	42.3
Aggregation	14	12	345	12	53.8	53.8

TABLE II. RESULTS OF LOF-BASED MOIS OVER BCWD

Subspaces	True Classification					
	TP	FP	TN	FN	HR(%)	PS(%)
(20,22)	21	5	353	5	80.8	80.8
(0,3,20)	19	7	351	7	73.1	73.1
(3,20,23)	21	5	353	5	80.8	80.8
(0,3,23)	20	6	352	6	76.9	76.9
(0,3,22)	19	7	351	7	73.1	76
(0,22,23)	21	5	352	5	80.8	80.8
Full Space	15	11	347	11	57.7	57.7
Aggregation	20	6	352	6	76.9	76.9

B. Landsat Satellite Data Set

The original Landsat Satellite data set in Statlog consists of 6435 samples with 36 attributes. It has six classes. We generated a new test data set with 1839 samples. Class 5 (69 samples) was considered as outliers to be detected. By setting the percentage to 3.75%, the number of outliers was 69. The results with the distance-based and LOF-based

MOIS algorithms are detailed in Table 3 and Table 4, respectively. We can find that the performance of distance-based MOIS was better than that of LOF-based MOIS.

TABLE III. RESULTS OF DISTANCE-BASED MOIS OVER STATLOG

Subspaces	True Classification					
	TP	FP	TN	FN	HR(%)	PS(%)
(24,28)	8	61	1709	61	11.6	11.6
(1,2,3)	21	48	1719	48	30.4	30.4
(13,14,15)	22	47	1720	47	31.9	31.9
(3,7)	7	62	1705	62	10.1	10.1
(5,6,7)	19	50	1717	50	27.5	27.5
(21,22,23)	21	48	1719	48	30.4	30.4
Full Space	8	61	1706	61	11.6	11.6
Aggregation	12	57	1710	57	17.4	17.4

TABLE IV. RESULTS OF LOF-BASED MOIS OVER STATLOG

Subspaces	True Classification					
	TP	FP	TN	FN	HR(%)	PS(%)
(24,28)	8	61	1709	61	11.6	11.6
(1,2,3)	21	48	1719	48	30.4	30.4
(13,14,15)	22	47	1720	47	31.9	31.9
(3,7)	7	62	1705	62	10.1	10.1
(5,6,7)	19	50	1717	50	27.5	27.5
(21,22,23)	21	48	1719	48	30.4	30.4
Full Space	8	61	1706	61	11.6	11.6
Aggregation	12	57	1710	57	17.4	17.4

VI. CONCLUSION

This paper presents a novel approach for mining outliers in subspaces. There are two steps behind this method: 1).find the correlated subspaces using the entropy-based algorithm; and 2). identify outliers in the related subspaces using classical outlier detection methods. This paper describes the criteria for measuring the degree of correlation among dimensions. The results are more meaningful and interpretable than those of some direct subspace outlier mining methods. Future work includes formulating a criterion to identify the most interesting subspaces, and evaluate the outliers in the most interesting subspaces. Another direction of this work is to further investigate the groups in subspaces, and design a powerful visualization toolbox so as to provide interpretable solutions to the results.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu. Outlier Detection for High Dimensional Data. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pages 37–46, 2001.

[2] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley&Sons, 1994.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference On Management of Data*, pages 93–104, 2000.

[4] C. H. Cheng, A. W.-C. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 84–93, 1999.

[5] D. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.

[6] E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-based Outliers in Large Datasets. In *Proceedings of the 24th International Conference on Very Large Data Bases*, pages 392–403, 1998.

[7] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. *SIGMOD Record*, 29(2):427–438, 2000.

[8] J. Tang, Z. Chen, A. W.-C. Fu, and D. W.-L. Cheung. Enhancing Effectiveness of Outlier Detections for Low Density Patterns. In *PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 535–548, 2002.