2008

SubCOID: an Attempt to Explore Cluster-Outlier Iterative Detection Approach to Multi-Dimensional Data Analysis in Subspace

Yong Shi
Kennesaw State University, yshi5@kennesaw.edu

Recommended Citation

Yong Shi. 2008. SubCOID: an attempt to explore cluster-outlier iterative detection approach to multi-dimensional data analysis in subspace. In Proceedings of the 46th Annual Southeast Regional Conference on XX (ACM-SE 46). ACM, New York, NY, USA, 132-135.

SubCOID: An Attempt to Explore Cluster-Outlier Iterative Detection Approach to Multi-Dimensional Data Analysis in Subspace

Yong Shi
Department of Computer Science and Information Systems
Kennesaw State University
1000 Chastain Road
Kennesaw, GA 30144
yshi5@kennesaw.edu

ABSTRACT

Many data mining algorithms focus on clustering methods. There are also a lot of approaches designed for outlier detection. We observe that, in many situations, clusters and outliers are concepts whose meanings are inseparable to each other, especially for those data sets with noise. Clusters and outliers should be treated as the concepts of the same importance in data analysis. In our previous work [22] we proposed a cluster-outlier iterative detection algorithm in full data space. However, in high dimensional spaces, for a given cluster or outlier, not all dimensions may be relevant to it. In this paper we extend our work in subspace area, tending to detect the clusters and outliers in another perspective for noisy data. Each cluster is associated with its own subset of dimensions, so is each outlier. The partition, subsets of dimensions and qualities of clusters are detected and adjusted according to the intra-relationship within clusters and the inter-relationship between clusters and outliers, and vice versa. This process is performed iteratively until a certain termination condition is reached. This data processing algorithm can be applied in many fields such as pattern recognition, data clustering and signal processing.

1. INTRODUCTION

The generation of multi-dimensional data has proceeded at an explosive rate in many disciplines with the advance of modern technology. Many new clustering, outlier detection and cluster evaluation approaches are presented in the last a few years. Nowadays a lot of real data sets are noisy, which makes it more difficult to design algorithms to process them efficiently and effectively.

We observe that, in many situations, clusters and outliers are concepts whose meanings are inseparable to each other, especially for those data sets with noise. Thus, it is necessary to treat clusters and outliers as concepts of the same importance in data analysis.

ACM-SE '08, March 28–29, 2008, Auburn, AL, USA.

Based on this observation, in previous work [22], we present a cluster-outlier iterative detection algorithm for noisy multi-dimensional data set in which clusters are detected and adjusted according to relationships between clusters and outliers.

However, clustering and outlier detection approaches are not always efficient and effective when applied in full data space. It is well acknowledged that in the real world a large proportion of data has irrelevant features which may cause a reduction in the accuracy of some algorithms. In this paper, we propose a new approach *SubCOID*, tending to explore cluster-outlier iterative detection approaches in subspace. In our approach, each cluster is associated with its own subset of dimensions, so is each outlier. We first find some initial (rough) sets of clusters and outliers. Based on the initial sets, we gradually improve the clustering and outlier detection results. In each iteration, the partition, subsets of dimensions and compactness of each cluster are modified and adjusted based on intra-relationship among clusters and the inter-relationship between clusters and outliers. The subset of dimensions and quality rank each outlier is associated with are modified and adjusted based on relationship among outliers and the inter-relationship between clusters and outliers.

The remainder of this paper is organized as follows. Section 2 introduces the related work of clustering, outlier detection and cluster evaluation. Section 3 presents the formalization and definitions of the problem. Section 4 describes the subspace cluster-outlier iterative detection (*SubCOID*) algorithm.

2. RELATED WORK

More and more large quantities of multi-dimensional data need to be clustered and analyzed. Cluster analysis is used to identify homogeneous and well-separated groups of objects in data sets. It plays an important role in many fields of business and science. The basic steps in the development of a clustering process can be summarized as [9] data cleaning,

feature selection, application of a clustering algorithm, validation of results, and interpretation of the results. Among these steps, the clustering algorithm and validation of the results are especially critical, and many methods have been proposed in the literature for these two steps. Existing clustering algorithms can be broadly classified into four types: partitioning [13, 15, 18], hierarchical [25, 10, 11], grid-based [23, 21, 3], and density-based [8, 12, 4] algorithms.

Outlier detection is concerned with discovering the exceptional behaviors of certain objects. It is an important branch in the field of data mining and in some sense it is at least as significant as cluster detection. There are numerous studies on outlier detection. D. Yu etc. [24] proposed an outlier detection approach termed *FindOut* as a by-product of WaveCluster [21] which removes the clusters from the original data and thus identifies the outliers. E. M. Knorr etc. [16] detected a distance-based outlier which is a data point with a certain percentage of the objects in the data set having a distance of more than d_{min} away from it. S. Ramaswamy etc. [19] further extended it based on the distance of a data point from its k^{th} nearest neighbor and identified the top n points with largest k^{th} nearest neighbor distances as outliers. M. M. Breunig etc. [5] introduced the concept of *local outlier* and defined *local outlier factor* (LOF) of a data point as a degree of how isolated the data point is with respect to the surrounding neighborhood. Aggarwal etc. [2] considered the problem of outlier detection in subspace to overcome dimensionality curse.

High dimensional data sets continue to pose a challenge to clustering algorithms at a very fundamental level. One of the well known techniques for improving the data analysis performance is the method of dimension reduction[3, 1, 20] in which data is transformed to a lower dimensional space while preserving the major information it carries, so that further processing can be simplified without compromising the quality of the final results. Dimension reduction is often used in clustering, classification, and many other machine learning and data mining applications.

There are some previous work on detecting clusters and outliers in subspace [1]. However, PROCLUS [1] does not explore the interactivity between clusters and outliers. Also PROCLUS favors spherical clusters, which limits its application for the real data with clusters of arbitrary shapes.

Our approach is different from the previous clustering and outlier detection methods in that we tried to detect and adjust the set of clusters and outliers according to the intra-relationship in the set of clusters and the set of outliers, as well as the inter-relationship between clusters and outliers. Furthermore, our algorithm is performed in subspace, rather than in full data space.

There are several criteria for quantifying the similarity (dissimilarity) of the clusters. ROCK[11] measures the similarity of two clusters by comparing the aggregate interconnectivity of two clusters. Chameleon [14] measures the similarity of two clusters based on a dynamic model.

Many approaches [6, 17] have been proposed for evaluating the results of a clustering algorithm. These clustering validity measurements evaluate clustering algorithms by measuring the overall quality of the clusters.

3. PROBLEM DEFINITION

In order to describe our approach we shall introduce a few notation and definitions. Let n denote the total number of data points and d be the dimensionality of the data space. Let the input d-dimensional dataset be $\mathbf{X}$

$$\mathbf{X} = \{\vec{X_1}, \vec{X_2}, ..., \vec{X_n}\},$$

which is normalized to be within the hypercube $[0, 1]^d \subset \mathbb{R}^d$. Each data point $\vec{X_i}$ is a d-dimensional vector:

$$\vec{X_i} = [x_{i1}, x_{i2}, ..., x_{id}]. \quad (1)$$

We assume the current number of clusters is k_c, and the current number of outliers is k_o. The set of clusters is $\mathcal{C} = \{C_1, C_2, ..., C_{k_c}\}$, and the set of outliers is $\mathcal{O} = \{O_1, O_2, ..., O_{k_o}\}$. For a given cluster C_i, i=1,...,k_c, its associated subspace is s_{c_i}. For a given outlier O_j, j=1,...,k_o, its associated subspace is s_{o_j}.

We use $d_s(X_1, X_2)$ to represent the distance between two data points X_1 and X_2 under a certain distance metric. In a high dimensional space the data are usually sparse, and widely used distance metric such as Euclidean distance may not work well as dimensionality goes higher. The L_p norm is widely used in the research work of distance measurement. L_p: $d(X_1, X_2) = (\sum_{i=1}^{d} |X_{1i} - X_{2i}|^p)^{1/p}$. In our previous work, we prefer $L_{0.1}$ to L_2 metric.

For two data points X_1 and X_2, under $L_{0.1}$ norm, their distance under a d-dimensional data space is: $L_{0.1}$: $d(X_1, X_2) = (\sum_{i=1}^{d} |X_{1i} - X_{2i}|^{0.1})^{10}$.

However, since in our approach we focus on working on clustering and outlier detection in individual subsets of subspaces, it's crucial that the distance measurements in different subspaces are fair to each other. Hence we modified the $L_{0.1}$ norm slightly as $L'_{0.1}$. For two data points X_1 and X_2, under $L'_{0.1}$ norm, their distance under a d-dimensional data space is:

$d(X_1, X_2) = (\frac{\sum_{i=1}^{d} |X_{1i} - X_{2i}|^{0.1}}{d})^{10}$. $L'_{0.1}$ norm erases the difference caused by the different set of dimensions involved in the distance metrics.

Suppose the distance is calculated in subspace s, we denote it as: $d_s(X_1, X_2)$.

In our previous work, we proposed some concepts regarding to the diversities between clusters, cluster-outlier pairs and outliers. However, they are not applicable regarding to subspace problem. First of all, each cluster/outlier now has its own associated subsets of dimensions, instead of the usual full data space. Thus Compactness of a cluster should be changed since it was original defined in full data space. We should also significantly redefine the diversities between clusters, cluster-outlier pair and outliers.

Definition 1: *For a cluster C_i, let s_{c_i} be its associated subspace, let $MST(C)$ be a minimum spanning tree on the dense cells of the minimal subgraph containing C_i. The* ***internal***

connecting distance (ICD) *of C_i, denoted as $ICD(C_i)$, is defined as the length of a longest edge of $MST(C_i)$. The* ***external connecting distance (ECD)*** *of C_i, denoted as $ECD(C_i)$, is defined as the length of a shortest edge connecting the centers of C_i and other clusters. The* ***compactness*** *of C_i, denoted as $Compactness(C_i)$, is defined as*

$$Compactness(C_i) = \frac{ECD(C_i)}{ICD(C_i)}. \quad (2)$$

In the following we use $CPT(C_i)$ to denote $Compactness(C_i)$.

Definition 2: *The diversity between a cluster C and an outlier O is defined as:*

$$D_1(C,O) = \frac{w_1 \cdot d_{min}(O,C) + w_2 \cdot d_{avr}(O,C)}{1 + |Cos\theta|} \quad (3)$$

where $w_1 = \frac{1}{CPT(C)+1}$, $w_2 = \frac{CPT(C)}{CPT(C)+1}$, $d_{avr}(O,C) = d_s(O,m_c)$, $d_{min}(O,C) = max(d_s(O,m_c) - r_{max}, 0)$ where $s = s_c \cap s_o$, r_{max} is the maximum distance of the data points in C from its centroid, and θ is the angle between the eigenvector of cluster C corresponding to the largest eigenvalue and the vector connecting m_c and O. The criteria for setting the weights w_1 and w_2 are similar to those in [7].

Definition 3: *The diversity between two clusters C_1 and C_2 is defined as:*

$$D_2(C_1,C_2) = \frac{d_s(C_1,C_2) * (1 + |Cos\theta|)}{CPT(C_1) + CPT(C_2)} \quad (4)$$

where $d_s(C_1,C_2)$ can be either the average distance between the two clusters or the minimum distance between them in subspace $s = s_{c_1} \cap s_{c_2}$. Here we simply apply the former one $d(m_{C_1}, m_{C_2})$. θ is the angle between the two eigenvectors corresponding to the two largest eigenvalues of C_1 and C_2, respectively. The larger the value of $D_2(C_1,C_2)$ is, the larger diversity the clusters C_1 and C_2 have to each other.

Definition 4: *The diversity between two outliers O_1 and O_2 is defined as:*

$$D_3(O_1,O_2) = d_s(O_1,O_2) \quad (5)$$

where $s = s_{o_1} \cap s_{o_2}$

Definition 5: *We measure the quality of a cluster C as:*

$$Q_c(C) = \frac{\frac{\sum_{C' \in \mathcal{C}, C' \neq C} D_2(C,C')}{k_c - 1} + \frac{\sum_{O \in \mathcal{O}} D_1(C,O)}{k_o}}{CPT(C)} \quad (6)$$

The larger $Q_c(C)$ is, the better quality cluster C has.

Similarly, the quality of an outlier O is reflected not only by the diversity between it and clusters, but also by the diversity between it and other outliers. The farther distances it has from other outliers and clusters, the better quality it should obtain.

Definition 6: *We measure the quality of an outlier O as:*

$$Q_o(O) = \frac{\sum_{O' \in \mathcal{O}, O' \neq O} D_3(O,O')}{k_o - 1} + \frac{\sum_{C \in \mathcal{C}} D_1(C,O)}{k_c} \quad (7)$$

The larger $Q_o(O)$ is, the better quality outlier O has.

4. ALGORITHM

The main goal of the *SubCOID* algorithm is to mine the optimal set of clusters and outliers for the input data set in cluster/outlier associated subspaces. As we mentioned in the previous sections, in our approach, for a given multi-dimensional data, clusters and outliers associated with individual subsets of dimensions are detected, adjusted and improved iteratively. Clusters and outliers are closely related and they affect each other in a certain way. The relationship between clusters with different subsets of dimensions are complicated, so are that of outliers and that of cluster-outlier pairs. The basic idea of our algorithm is that clusters are detected and adjusted according to the intra-relationship within clusters and the inter-relationship between clusters and outliers in subspace, and vice versa. The adjustment and modification of the clusters and outliers are performed iteratively until a certain termination condition is reached. This analysis approach for multi-dimensional data can be applied in many fields such as pattern recognition, data clustering and signal processing. The overall pseudocodes for the algorithm is given in Figure 1.

5. REFERENCES

[1] C. C. Aggarwal, C. Procopiuc, J. Wolf, P. Yu, and J. Park. Fast algorithms for projected clustering. In *Proceedings of the ACM SIGMOD CONFERENCE on Management of Data*, pages 61–72, Philadelphia, PA, 1999.

[2] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *SIGMOD Conference*, 2001.

[3] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 94–105, Seattle, WA, 1998.

[4] Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure. *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA*, pages 49–60, 1999.

[5] M. Breunig, H. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD CONFERENCE on Management of Data*, pages 93–104, Dallas, Texas, May 16-18 2000.

[6] Chi-Farn Chen, Jyh-Ming Lee . The Validity Measurement of Fuzzy C-Means Classifier for Remotely Sensed Images. In *Proc. ACRS 2001 - 22nd Asian Conference on Remote Sensing*, 2001.

[7] Dantong Yu and Aidong Zhang. *ClusterTree*: Integration of Cluster Representation and Nearest Neighbor Search for Large Datasets with High Dimensionality. *IEEE Transactions on Knowledge and Data Engineering(TKDE)*, 14(3), May/June 2003.

[8] M. Ester, K. H.-P., J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996.

[9] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[10] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In

Algorithm SubCOID (*k: No. of Clusters*)
Begin
1. Initialization Phase
Repeat
Const1 and Const2 are two proportion constants to k
Const1 > Const2;
RandomSize1 = Const1·k;
RandomSize2 = Const2·k;
RS_1 = random sample with the size of Randomsize1;
RS_2 = FindKMedoids(RS_1, RandomSize2);
{*Assign data points to medoids to form medoid-associated sets*}
E ← SubDispatchDataPoints(RS_2); {E is set of initial data division;}
{*Determine the characteristics of the medoid-associated sets, and the initial subspace for each set*}
{$\mathcal{C}$ and $\mathcal{O}$} ← SubClusterOrOutlier();
Until($|\mathcal{C}| \geq$ k)

2. Iterative Phase
{*Merge the clusters according to the input cluster number k*}
$\mathcal{C}$ ← MergeSubspaceCluster($\mathcal{C}$);
Repeat
{*Find the nearest cluster for each outlier*}
For each outlier o ∈ $\mathcal{O}$ do
Begin
Find its nearest cluster ∈ $\mathcal{C}$
End
Sort current set of clusters in ascending order based on their qualities;
Sort current set of outliers in ascending order based on their qualities;
{*Reorganize the structure of clusters and outliers*}
ExchangeSubspaceClusterAndOutlier();
$\mathcal{O}'$ is the set of outliers with worst qualities;
BDP is the set of boundary data points with worst qualities;
$U = \mathcal{O}' \cup$ BDP;
Until(U is stable or iteration number $\geq \Im$)
End.

Figure 1: Algorithm: SubCOID

Proceedings of the ACM SIGMOD conference on Management of Data, pages 73–84, Seattle, WA, 1998.

[11] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In *Proceedings of the IEEE Conference on Data Engineering*, 1999.

[12] A. Hinneburg and D. A.Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 58–65, New York, August 1998.

[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I,Statistics.*, 1967.

[14] G. Karypis, E.-H. S. Han, and V. K. NEWS. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[15] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[16] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th VLDB conference*, pages 392–403, New York, August 1998.

[17] Maria Halkidi, Michalis Vazirgiannis. A Data Set Oriented Approach for Clustering Algorithm Selection. In *PKDD*, 2001.

[18] R. T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. In *Proceedings of the 20th VLDB Conference*, pages 144–155, Santiago, Chile, 1994.

[19] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD CONFERENCE on Management of Data*, pages 427–438, Dallas, Texas, May 16-18 2000.

[20] T. Seidl and H. Kriegel. Optimal multi-step k-nearest neighbor search. In *Proceedings of the ACM SIGMOD conference on Management of Data*, pages 154–164, Seattle, WA, 1998.

[21] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *Proceedings of the 24th International Conference on Very Large Data Bases*, 1998.

[22] Y. Shi and A. Zhang. Towards exploring interactive relationship between clusters and outliers in multi-dimensional data analysis. In *International Conference on Data Engineering (ICDE)*, 2005.

[23] W. Wang, J. Yang, and R. Muntz. STING: A Statistical Information Grid Approach to Spatial Data Mining. In *Proceedings of the 23rd VLDB Conference*, pages 186–195, Athens, Greece, 1997.

[24] D. Yu, G. Sheikholeslami, and A. Zhang. Findout: Finding outliers in very large datasets. *The Knowledge and Information Systems (KAIS)*, (4), October 2000.

[25] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103–114, Montreal, Canada, 1996.