Experimental Research on Impacts of Dimensionality on Clustering Algorithms

Hai-Dong Meng, Jin-Hui Ma
School of Information Engineering
Inner Mongolia University of Science and Technology
Baotou 014010, China
haidongm@imust.edu.cn

Guan-Dong Xu
School of Engineering and Science
Victoria University
72 Papillon Pde, Tarneit, Vic 3029, Australia

***Abstract*—Experiments are carried out on datasets with different dimensions selected from UCI datasets by using two classical clustering algorithms. The results of the experiments indicate that when the dimensionality of the real dataset is less than or equal to 30, the clustering algorithms based on distance are effective. For high-dimensional datasets--dimensionality is greater than 30, the clustering algorithms are of weaknesses, even if we use dimension reduction methods, such as Principal Component Analysis (PCA).**

Keywords-clustering algorithm; dimensionality; high-dimensional dataset; validity

I. INTRODUCTION

Cluster analysis is playing a more important role in resources evaluation and geoscientific data processing, and is also an important research topic in the field of data mining. Clustering high-dimensional datasets is a challenge of cluster analysis in the applications of resources evaluation and geoscientific data processing. As dimensionality increases, many types of data analysis becomes significantly harder, the data become increasingly sparse in the data space that it occupies, and the definitions of density and the distance between data points, which are critical for clustering, become less meaningful. As a result, many clustering algorithms have trouble with high-dimensional data--reduced clustering accuracy and poor quality clusters [1]. This is because when the dimensionality increases, usually only a small number of dimensions are relevant to certain clusters, but data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. Moreover, when dimensionality increases, data usually become increasingly sparse because the data points are likely to be located in different dimensional subspaces. When the data become really sparse, data points located in different dimensions can be considered as all equally distanced, and the distance measure, which is essential for cluster analysis, becomes meaningless.

To overcome this difficulty, the general (common) methods are feature (or attribute) transformation and feature (or attribute) selection techniques. Feature transformation methods, such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), transform the data into a smaller space while generally preserving the original relative distance between objects. They summarize data by creating linear combinations of the attributes, and may discover hidden structures in the data. Another way of tackling the curse of dimensionality is to try to remove some of the dimensions. Attribute subset selection is commonly used for data reduction by removing irrelevant or redundant dimensions (or attributes). Given a set of attributes, attribute subset selection finds the subset of attributes that are most relevant to the data mining task [2].

Subspace clustering is an extension to attribute subset selection that has shown its strength at high-dimensional clustering. It is based on the observation that different subspaces may contain different, meaningful clusters. Subspace clustering searches for groups of clusters within different subspaces of the same data set. The problem becomes how to find such subspace clusters effectively and efficiently [3]. Other clustering methods, such as grid-based method [4], graph-based method [5] and SNN method [6], and parallel method [7], are also proposed to analyze high-dimensional datasets.

However, what are the characteristics of the variation of sparsity with dimensionality? If we use the methods of feature transformation and feature selection techniques, can we obtain valid results? In these aspects, it is necessary to take a deeper level research.

The rest of this paper is organized as follows. We give the definitions of maximal distance, average distance and clustering accuracy, and the experimental datasets in section 2. The experimental analyses and performance evaluation of the clustering algorithms are given in section 3. Section 4 concludes with a summary.

II. CLUSTERING ALGORITHMS AND DATASETS

A. Clustering Algorithms

The traditional clustering algorithms can be divided into the following five categories: (1) partitioning methods, (2) hierarchical methods, (3) density-based methods, (4) grid-based methods, and (5) model-based methods. The K-means algorithm first chooses K initial centroids, where K is a user-specified parameter, namely the number of clusters desired. Each data point is then assigned to the closest centroid, and each collection of data points assigned to a centroid is a cluster.

The centroid of each cluster is then updated based on the data points assigned to the cluster. The algorithm repeats the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same [1].

Hierarchical clustering techniques are commonly used clustering methods. The methods complete the division of classification by a series of steps, and actually produce a nested set of clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters [2].

In this paper, we carry out the experiments on the datasets with different dimensionality by using a partitioning method, i.e., K-means algorithm and a hierarchical method, i.e., the agglomerative hierarchical clustering algorithm.

B. Experimental datasets

Datasets used in the experiment are from the UCI database, and the datasets include Iris [8], Wine [9], Wisconsin Diagnostic Breast Cancer [10], SPECT Heart [11] and Libras Movement [12]. Detailed description of datasets is in table 1.

TABLE I. THE DATASETS FOR EXPERIMENTS

Name of Dataset	Number of Tuple	Number of attribute	Number of category (Cluster)
Iris	150	4	3
Wine	178	13	3
Wisconsin Diagnostic Breast Cancer	569	30	2
SPECT Heart	269	44	2
Libras Movement	360	90	15

C. Definitions

In this paper, in order to research the characteristics of sparsity and the impacts of dimensionality on the accuracy of the clustering algorithms, some definitions are given as below: as follows:

(1) Maximal distance: Assume dataset D has n data objects and each data object has d attributes (dimensionality), i.e. $X_i = \{x_k, k = 1, \ldots, d\}, i = 1, \ldots, n$. Maximal distance between data objects is defined as:

$$Dist_{Max} = Max[(\sum_{k=1}^{d}(x_{ik} - x_{jk})^2)^{1/2}, i \neq j]_1 \quad (1)$$

(2) Average distance: The average distance between data objects in dataset D is:

$$Dist_{Aver} = (\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d}(x_{ik} - x_{jk})^2)^{1/2} \quad (2)$$

(3) Accuracy: Assume that there are k clusters in dataset D, i.e., $C_i (i = 1, \ldots, k)$, and $O_{ip}\left(p = 1, \ldots, m_p\right)$ indicates the objects in C_i. After clustering the dataset D, we find k clusters $C_i^{'}\left(i = 1, \ldots, k\right)$ arithmetically, and $O_{ip}^{'}\left(p = 1, \ldots, m_p^{'}\right)$ indicates objects in cluster $C_i^{'}$. The clustering accuracy is defined as follows:

$$Accuracy = \frac{\sum_{i=1}^{k} Max[|C_1 \cap C_i'|, |C_2 \cap C_i'|, \ldots, |C_k \cap C_i'|]}{|D|} \quad (3)$$

Here, $\left|C_k \cap C_i^{'}\right|$ is the number of objects that belong to C_i and $C_i^{'}$ simultaneously, and $|D|$ is the number of objects in the dataset D.

III. EXPERIMENTAL RESULTS AND ANALYSES

A. Impact on Sparsity

According to Formula (1) and (2) defined above, the maximal distance and average distance between data objects will increase with increasing dimensionality d. But, it's necessary to study the varying characteristics of the distances with dimensionality for real datasets, so as to understand the impacts of dimensionality on clustering algorithms. We use Libras Movement dataset in the UCI datasets, and calculate its maximal distance and average distance between data objects with the increase of dimensionality. Libras Movement dataset has 90 dimensions (attributes). The results are shown in Figure 1 and Figure 2 respectively. As we can see from the Figures, Maximal distance and Average distance between data objects in the dataset increase when the dimensionality increases, as shown in Figure 1 and Figure 2. When the dimensionality is less than 30, the increases of the maximal distance and average distance are relative rapid; when dimensionality is greater than 30, the increase of the maximal distance and average distance are relative slow, or linear. The curves has an inflection point, i.e., dimensionality=30. When the variations of the maximal distance and average distance are greater, this indicates that the variation of distance between data objects is greater. Therefore, we can deduce that when the dimensionality is less than 30, the clustering algorithms based on distance or density are effective.

The results also indicate that the dataset becomes sparse in high-dimensional data space, and validities of clustering algorithms based on distance between data objects become uncertain. At the same time, the algorithms based on density and density-reachable have to redefine the density, in order to obtain acceptable results.

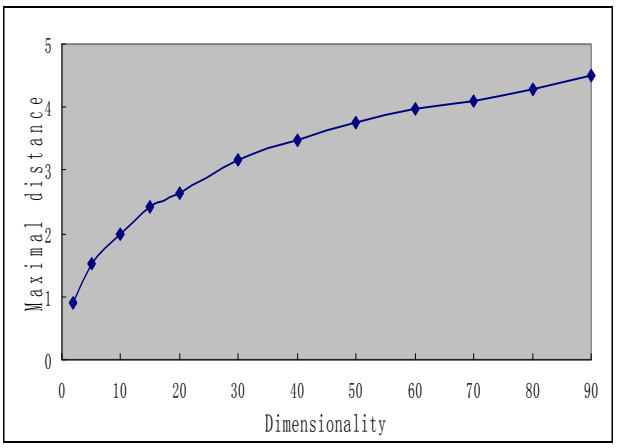


Figure 1. Variation of maximal distance with dimensionality.

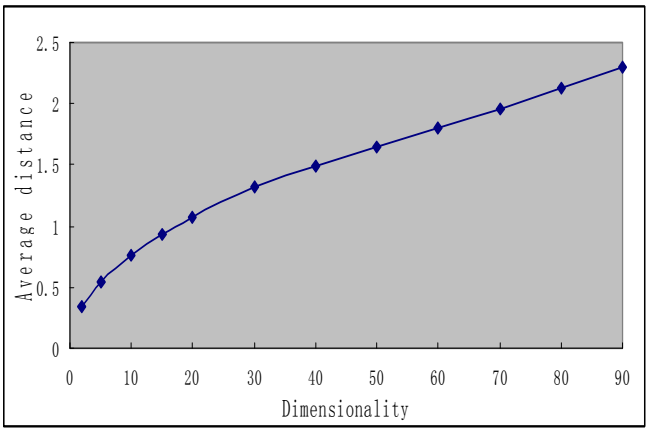


Figure 2. Variation of average distance with dimensionality.

B. *Impact on Clustering Accuracy*

In order to research the impacts of dimensionality on clustering accuracy, the five datasets with different dimensionality are handled with K-means and the agglomerative hierarchical clustering algorithm. The clustering results are shown in Figure 3. When the dimensionality is less than 30, the clustering algorithms have good performance, but when the dimensionality is greater than 30, the clustering accuracies decrease. Perhaps, the experimental results reveal the fact that if the dimensionality is less than 30, clustering algorithms, such as K-means and the agglomerative hierarchical clustering algorithm would have good performance, but they can not work well when dimensionality is greater than 30.

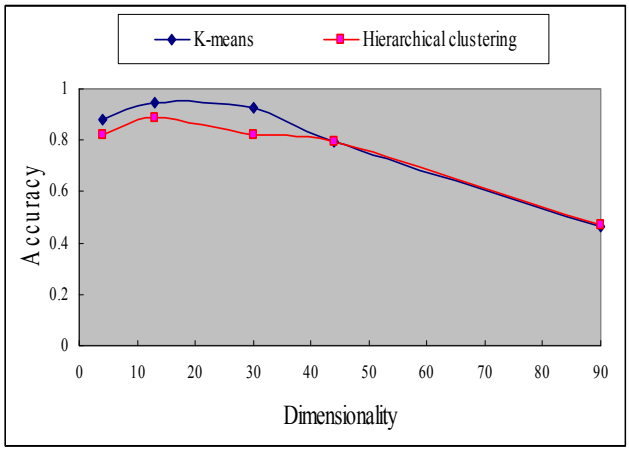


Figure 3. Variation of accuracy with dimensionality.

C. *Feature Reduction Experiments*

Wine dataset has 13 attributes. Principal Component Analysis (PCA) [13] transforms the dataset into a 3-dimensional data space, and removes the irrelevant or redundant attributes (or dimensions). In order to compare the clustering results, the original dataset and the transformed dataset are handled with K-means and the hierarchical clustering algorithm, as shown in Figure 4. The clustering results indicate that for the low-dimensional dataset, the clustering accuracies of the original dataset and the transformed dataset are almost the same. After dimension reduction, the clustering accuracy is improved, but the effect is not very clear.

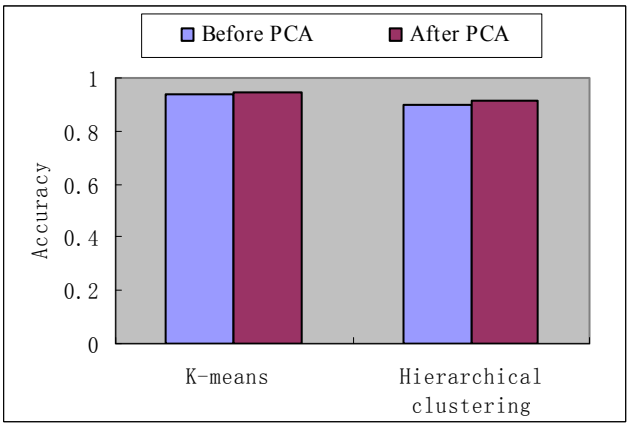
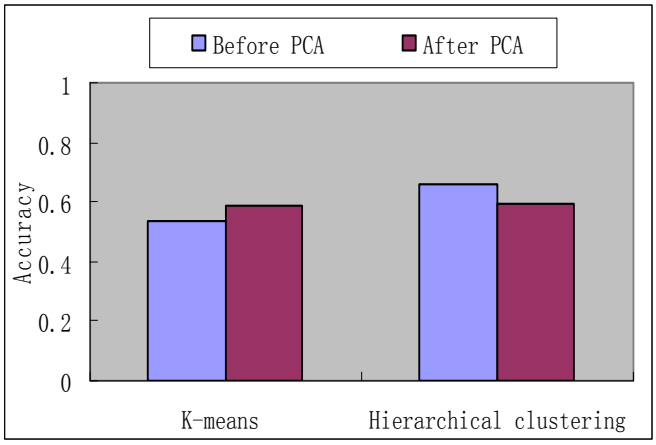


Figure 4. The clustering results of Wine dataset.

Libras Movement dataset has 90 dimensions, and is transformed into a 10-dimensional dataset by PCA dimension reduction. The clustering accuracies of the original dataset and the transformed dataset are relative low for both K-means and Hierarchical clustering algorithm, which indicates that: (1) the clustering algorithms are not effective for high-dimensional dataset; (2) feature (dimension) reduction is not always effective for clustering algorithms; (3) the dataset includes 15 clusters (groups) which are difficult to identify, as shown in Figure 5.



IV. CONCLUSIONS

The analyses of experimental results indicate that when the dimensionality of the real dataset is less than or equal to 30, the clustering algorithms based on distance or density are effective. For high-dimensional datasets (dimensionality is greater than 30), the clustering algorithms are of weaknesses, even if we use dimension reduction methods, such as PCA.

ACKNOWLEDGMENT

The materials are based on work supported by National Natural Science Foundation of China under Grant No. 40762003, National Chunhui Project under Grant No. Z2009-1-01041, and Natural Science Foundation of Inner Mongolia under Grant No. 200711020814.

REFERENCES

[1] Tan, P-N., Steinbach, M., and Kumar, V. Introduction to Data Mining. Pearson Education, Inc., 2006.

[2] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Piblishers, Second Edition, 2006.

[3] R. Agrawal, J. Gehrke, D.Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proc. of 1998 ACM-SIGMOD Intl. Conf. on Management of Data, pp. 94-105, Seattle, Washington, June 1998. ACM Press.

[4] A. Hinneburg and D. A. Keim. Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. In Proc. of the 25th VLDB Conf., pp. 506-517, Edinburgh, Scotland, UK, 1999.

[5] E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher. Hypergraph Based Clustering in High Dimensional Data Sets: A Summary of Results. IEEE Data Eng. Bulletin, 21(1): pp.15-22, 1998.

[6] L. Ertöz, M. Steinbach, and V. Kumar. A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In Workshop on Clustering High Dimensional Data and its Applications, Proc. of Text Mine'01, First SIMA Intl. Conf. on Data Mining, Chicago, IL, USA, 2001.

[7] H. Nagesh, S. Goil, and A. Choudhary. Parallel Algorithms for Clustering High-Dimensional Large-Scale Datasets. Data Mining for Scientific and Engineering Applications, pp. 335-356, Kluwer Academic Publishers, Dordrecht, Netherlands, October 2001.

[8] http://archive.ics.uci.edu/ml/datasets/Iris

[9] http://archive.ics.uci.edu/ml/datasets/Wine

[10] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[11] http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

[12] http://archive.ics.uci.edu/ml/datasets/Libras+Movement

[13] L. T. Jolliffe. Principal Component Analysis. Springer Verlag, 2nd edition, October 2002