# A Comprehensive Study of Challenges and Approaches for Clustering High Dimensional Data

Neelam Singh
Graphic Era University
Dehradun

Neha Garg
Graphic Era University
Dehradun

Janmejay Pant
Graphic Era University
Dehradun

## ABSTRACT
Clustering is one of the most effective methods for summarizing and analyzing datasets that are collection of data objects similar or dissimilar in nature. Clustering aims at finding groups, or clusters, of objects with similar attributes. Most clustering methods work efficiently for low dimensional data since distance measures are used to find dissimilarities between objects. High dimensional data, however, may contain attributes which are not required for defining clusters and irrelevant dimension may produce noise and will hide the clusters that are required to be created. The discovery of groups of objects that are highly similar within some subsets of relevant attributes becomes an important but challenging task. In this paper we provide a short introduction to various approaches and challenges for high-dimensional data clustering.

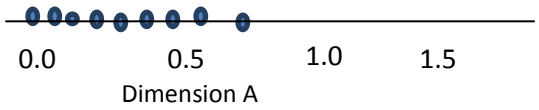
## Keywords
Clustering, high dimensional data, summarizing, analyzing, clusters

## 1. INTRODUCTION
With the proliferation of internet of things there is a formidable growth in the volume of information available on the Internet and also there is a ubiquity of data collection. Due to improved data acquisition techniques, low cost of data storage, organizations and researchers are investing a huge amount and interests on developing effective methods for analyzing and summarizing data. Document (text) data analysis requires more sophisticated techniques than numerical data analysis, which uses statistics and machine learning.

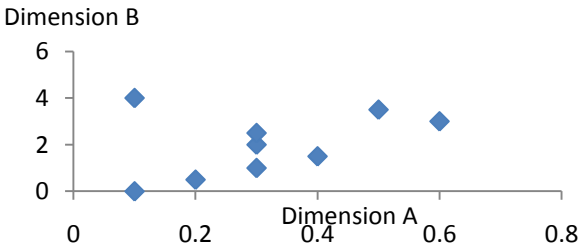
Clustering has proven to be one of the most effective methods for analyzing datasets containing large number of objects with plentiful attributes. Clustering groups, or make clusters, of objects with similar attributes. A cluster is defined as a subset of objects of similar attribute and objects which are dissimilar to the objects in other cluster. Clustering is a data segmentation technique where results are determined by partitioning the data sets according to similarity between pairs of objects. Clustering algorithms have proven to be successful for low-dimensional data, where the number of attributes is less and can be represented in two or three dimensions. But often the data collected for research contains multiple dimension, is sparse and highly skewed, known as high dimensional data. Finding clusters in high dimensional data often poses challenges and require more sophisticated techniques. In all cases, the approaches to clustering high dimensional data must deal with the "curse of dimensionality" [1]. In this paper we would like to describe the challenges faced in analysing high dimensional data and the clustering techniques which tries to resolve these threats.

## 2. THE CHALLENGE OF CLUSTERING HIGH DIMENSIONAL DATA
Real-world datasets have very high dimensional feature space and is highly sparse. It becomes difficult to generate meaningful results from such redundant and sparse data through traditional clustering algorithm. This is due to the fact that when dimensionality increases, data becomes sparse since data points are located at different dimensional subspaces. Thus it requires greater computational power to compute clusters as distance measure to find similarities between data objects become meaningless and often noise becomes prevalent and masks the real cluster to be discovered. The various challenges posed by high dimensional data are described in the next sections.

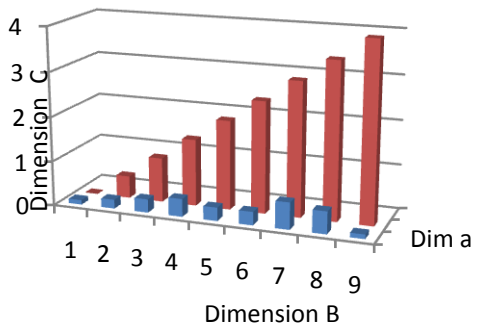
### 2.1 The Curse of Dimensionality
As described by Bellman dimensionality curse is one of the major problems faced by high dimensional data. In high dimensional space the points are more scattered or sparse and all points are almost equidistant from each other. Clustering approaches become ineffective to analyse the data due to this. The phenomena of the curse of the dimensionality become more prominent when points are moved from one dimensional space to a higher dimensional space as we can see in Figure 1. Thus to form or search clusters using traditional clustering algorithms becomes very difficult.




(a) Objects in 1 dimension




(b) Objects in 2-dimension




**(c) Objects in 3-dimension**
**Figure 1: As the dimensionality goes higher, points in the space are more spread out.**

## 2.2 Incompetence of Distance Measures

Clustering algorithms mainly finds dissimilarity between objects using distance functions like Euclidean distance, Manhattan distance, cosine distance, proximity measure etc. In low dimensional space distance measure seems to be more relevant to cluster objects but in high dimensional space , all pairs of points tend to be more or less equidistant from one another and therefore the discrimination of the nearest and farthest point in particular becomes meaningless. This can be expressed by the following equation [2] :

$$\lim_{d\to\infty} \frac{dist_{max} - dist_{min}}{dist_{min}} \to 0$$

## 2.3 Presence of Noise and Outliers

Datasets collected from real time applications like microarray data, pattern recognition often contains noise pertaining to measurement or because of the stochastic nature of systems involved in the process.

In general, the noise present in real applications often hides the clusters to be selected from clustering algorithm and the problem is worsened in high dimensional data, where the number of errors increases linearly with dimensionality. Noise-tolerance in clustering is very important to understand the real cluster structures in the datasets. However, distinguishing noise from accurate and relevant values is hard and searching for noise-tolerant clusters is even harder since in order to identify real cluster large number of clusters are considered.

## 2.3 Local Feature Relevance

Clustering is classified as unsupervised learning, where most of the algorithms require clusters to be flat or hierarchical partitions. Thus, an object is not allowed to exist in multiple clusters (at the same level). However, high dimensional data provides much richer information regarding each object than low dimensional data. An object might have one subset of attributes belonging to one subset of cluster and at the same time will also be similar to a different subset of objects under another set of attributes. Therefore, an object may be a member of multiple clusters. However, multi-cluster membership is restricted by traditional clustering algorithms where disjoint clusters are made.

## 2.4 Attribute Selection

One of the major problems faced by clustering algorithm is to select appropriate and relevant attribute for making clusters. Since in high dimensional datasets the objects are very sparse it becomes very difficult to select important attributes along with no or less redundancy. Any optimization problem becomes more difficult and hard to tackle as the number of variables or attributes increases thus making traditional algorithm ineffective to produce correct results.

# 3. HIGH DIMENSIONAL DATA CLUSTERING APPROACHES

Due to high dimension of data space or feature space clustering approaches are required to be more refined and generalized as the computational cost of traditional clustering algorithms increases with increase in dimension of the data. High dimensional data clustering approaches can be classified into two types namely dimensionality reduction and subspace clustering.

## 3.1 Dimensionality Reduction or Dimension Reduction

It is the process of reducing the number of random variables under consideration, and is addressed through two popular approaches namely, feature selection and feature extraction .

- **Feature Selection** (Variable Selection) – tries to construct a new feature space by transforming the original feature space into lower dimension. Feature selection filters out meaningful attributes from original data. Feature selection methods attempt to select a proper subset of features that best satisfies a relevant function or evaluation criterion. These are classified as Wrappers, Filters and Embedded [3].
- **Feature extraction or feature transformation-** transforms the data in the high-dimensional space to a space of fewer dimensions while generally preserving the novel relative distance between objects. The data transformation may be linear or non-linear. However these techniques do not actually remove any of the original attributes from analysis which may mask the real clusters, even after transformation. The popular algorithms used to transform the data are PCA (Principal component analysis) and SVD (Singular Value Decomposition).

## 3.2 Subspace Clustering

- Subspace clustering algorithms is a refinement of feature selection algorithm where the relevant subspaces for each cluster are selected independently from the given dataset. Subspace clustering can be under two perspectives i.e. top down or iterative subspace clustering algorithms and bottom up or grid based approaches [4] [5].

## 3.2.1. Subspace Clustering Approaches

Some of the popular approaches for subspace clustering are-

- Grid-based Subspace Clustering
- Projection-based Subspace Clustering
- Bipartitioning based Subspace Clustering
- Pattern-based Subspace Clustering

**a. Grid-based Subspace Clustering:** Existing subspace clustering algorithm often assumes a metric space, such as Euclidean space. Therefore, many clustering algorithms are grid-based. One of the pioneering subspace clustering is CLIQUE [6], which was followed by ENCLUS [7]. To approximate the density of the data points, CLIQUE partitions the data space using a uniform grid and counts the data points

that lie inside each cell of the grids. This is accomplished by partitioning each dimension into the same number of equal length intervals.

**b. Projection-based Subspace Clustering:** The projection-based algorithms generate clusters that are partitions of the dataset. These partitions best classify the set of points that are embedded in lower dimensional subspaces given some objective functions. Instead of projecting all the points into the same subspace, the algorithm allows each cluster to have a different subspace with variable dimensionality.

**c. Bipartitioning-based Subspace Clustering :**
Co-clustering is a branch of subspace clustering methods that usually generates partitions along both rows and columns simultaneously, which is the reminiscent of the k-means algorithms.

One of the pioneering co-clustering algorithms based on information theory was proposed by Dhillon et.al. [8].

**d. Pattern-based Subspace Clustering:** These algorithms finds pattern of interest in a subset based on some conditions [9]. One of the algorithm proposed, in this category aims at finding remarkable patterns by taking into consideration a subset of genes under a subset of conditions, by Cheng et al. [10]

## 3.3. Projected Clustering

In Projected clustering each point is assigned to a specific cluster, but clusters may exist in different subspaces. The general approach is to use a distinctive distance function along with a consistent clustering algorithm [11].

For example, the PreDeCon algorithm checks which attributes seem to support a clustering for each point, and adjust the distance function such that dimensions with low variance are amplified in the distance function [12].

## 3.4. Hybrid Clustering Algorithms

Algorithms that do not aim at uniquely assigning each data point to a cluster or at finding all clusters in all subspaces are called hybrid algorithms. Some hybrid algorithms offer the user an optional functionality of a pure projected clustering algorithm. Others aim at computing only the subspaces of potential interest rather than the final clusters. For e.g. DOC, MINECLUS Usually, hybrid methods that report clusters allow overlapping clusters, but do not aim at computing all clusters in all subspaces.

Not all the algorithms above generate the complete set of patterns. Some take a greedy approach of finding one maximal pattern at a time [13]. These algorithms often carry a polynomial time complexity with regard to the number of objects and the number of attributes for searching one cluster. Such algorithms may not identify a globally optimal solution and they may miss many important subspace clusters as well.

# 4. CONCLUSION

Clustering real-world data sets often categorised as high dimensional data is often hindered by the "curse of dimensionality". Clusters are embedded in subspaces since some features of the original space may be irrelevant. So most clustering algorithms fail to generate meaningful results for high dimensional data due to sparse nature of the data space. Research in data mining and related disciplines such as statistics, pattern recognition, machine learning, and also applied sciences like, for example, bioinformatics, has led to the emergence of a large variety of clustering techniques which can also address the specialized problem of clustering high-dimensional data. New approaches to this problem are proposed in numerous conferences and journals every year.
However, it is a well-accepted opinion that there is no generalised clustering technique suitable to all problems and universally applicable to any random datasets. The aim of the task of data analysis affects the choice of the clustering algorithm and also the analysis of the results of the clustering process.

# REFERENCES

[1] R.Bellman, "Dynamic Programming". Princeton University Press, 1957.

[2] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft, (1998), "When is 'nearest neighbor' meaningful?" In Proceedings of 7th International Conference on Database Theory (ICDT-1999), Jerusalem, Israel, pp. 217-235, (1999).

[3] Yiu-ming Cheung, Hong Jia, "Unsupervised Feature Selection with Feature Clustering", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012. doi 10.1109/WI-IAT.2012.259

[4] Hans-Peter Kriegel, Peer Kröger, Matthias Renz, Sebastian Wurst, "A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data", *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM)* (Washington, DC: IEEE Computer Society): 205–257, 2005. doi:10.1109/ICDM.2005.5, ISBN 0-7695-2278-5

[5] L.Parsons, E.Haque, and H. Liu, "Subspace clustering for high dimensional data: a review". SIGKDD Explorations, 6(1):90–105. 2004.

[6] R.Agrawal, J.Gehrke, D.Gunopulos and P.Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications". In SIGMOD Conference, pages 94–105. 1998.

[7] C.-H.Cheng, A.W .Fu, and Y.Zhang, "Entropy-based subspace clustering for mining numerical data". In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 84–93, New York, NY, USA. ACM Press. 1999

[8] I.S.Dhillon, S.Mallela, and D. S Modha, "Information-theoretic co-clustering". In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 89–98, New York, NY, USA. ACM Press.2003

[9] Hans-Peter Kriegel, Peer Kröger,, Arthur Zimek , "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering", *ACM Transactions on Knowledge Discovery from Data* (New York, NY: ACM) **3** (1): 1–58, 2009. doi:10.1145/1497577.1497578

[10] Y.Cheng, and G. M. Church, "Biclustering of expression data". In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pages 93–103. AAAI Press. 2000

[11] Charu C.Aggarwal, Joel L.Wolf, Philip S .Yu, Cecilia Procopiuc, Jong Soo Park, "Fast algorithms for projected clustering", *ACM SIGMOD Record* (New York, NY: ACM) **28** (2): 61–72, 1999. doi:10.1145/304181.304188

[12] Christian Böhm, Karin Kailing, Hans-Peter Kriegel, Peer Kröger, "Density Connected Clustering with Local Subspace Preferences", *Data Mining, IEEE International Conference on* (Los Alamitos, CA, USA: IEEE Computer Society): 24–34, 2004. doi:10.1109/ICDM.2004.10087, ISBN 0-7695-2142-8

[13] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan, "Automatic Subspace Clustering of High Dimensional Data", *Data Mining and Knowledge Discovery* (Springer Netherlands) 11 (1): 5–33, 2005. doi:10.1007/s10618-005-1396-1

[14] A. Zimek, "Clustering High-Dimensional Data**",** In C. C. Aggarwal, C. K. Reddy (ed.): Data Clustering: Algorithms and Applications, CRC Press: 201–230, 2013.

[15] E. Ntoutsi, A. Zimek, T. Palpanas, P. Kröger, H.-P. Kriegel, "Density-based Projected Clustering over High Dimensional Data Streams**",** In Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA: 987–998, 2012.