Clustering High Dimensional Data

Abstract:

High dimensional data is increasingly common in many fields such as e-commerce applications ,bioinformatics,web logs,medical records,financial transactions and machine learning data sets (1) the curse of dimensionality and more crucial (2) The specificity of similarities between points in a high dimensional space diminish

Introduction:

High-dimensional data arise naturally in many domains, and have regularly presented a great challenge for traditional data mining techniques, both in terms of effectiveness and efficiency. Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty in distinguishing distances between data points. Instead of attempting to avoid the curse of dimensionality by observing a lower dimensional feature subspace, Cluster analysis divides data into groups (clusters) for the purposes of summarization or improved understanding. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, or as a means of data compression. While clustering has a long history and a large number of clustering techniques have been developed in statistics, pattern recognition, data mining, and other fields, significant challenges still remain. In this chapter we provide a short introduction to cluster analysis, and then focus on the challenge of clustering high dimensional data. We present a brief overview of several recent techniques, including a more detailed description of recent work of our own which uses a concept-based clustering approach.

Problem:

Clustering is a widely adopted data mining model that partitions data points into a set of groups, each of which is called a cluster. A data point has a shorter distance to points within the cluster than those outside the cluster. In a high dimensional space, for any point, its distance to its closest point and that to the farthest point tend to be similar

As the number of dimensions increase, many clustering techniques begin to suffer from the curse of dimensionality, degrading the quality of the results. In high dimensions, data becomes very sparse and distance measures become increasingly meaningless We cannot expect that one type of clustering approach will be suitable for all types of data or even for all high dimensional data.,

Methods to Handle High Dimensional Data sets

Hypergraph Partitioning

Hypergraph-based clustering is an approach to clustering in high dimensional spaces, which is based on hypergraphs. Hypergraphs are an extension of regular graphs, which relax the restriction that an edge can only join two vertices. Instead an edge can join many vertices. Hypergraphbased clustering consists of the following steps:

i. Define the condition for connecting several objects (each object is a vertex of the hypergraph) by a hyperedge.
ii. Define a measure for the strength or weight of a hyperedge.
iii. Use a graph-partitioning algorithm to partition the hypergraph into two parts in such a way that the weight of the hyperedges cut is minimized.
iv. Continue the partitioning until a fixed number of partitions are achieved, or until a new partition would produce a poor cluster, as measured by some fitness criteria.

The disadvantage of this hypergraph Partitioning is Even though it produce high quality partitions it has some limitations.particularly the number of partitions must be specified by specified by the users as this algorithm does not know when to stop recursive bisection

Sub Space Clustering

Subspace clustering methods will search for clusters in a particular projection of the data . These methods can ignore irrelevant attributes and also problem is known as Correlation clustering. Two-way clustering, or Co-Clustering or Biclustering are known as the special case of axis-parallel subspaces. In these methods the objects are clustered simultaneously as the feature matrix consisting of data objects as they are span in rows and . As in general subspace methods they usually do not work with arbitrary feature combinations. But this special case it deserves attention due to its applications in bioinformatics.

CLIQUE-Clustering in Quest , is the fundamental algorithm used for numerical attributes for subspace clustering. It starts with a unit elementary rectangular cell in a subspace. If the densities exceeds the given threshold value, those cell are will be retained . It applies a bottom-up approach for finding such units. First, it divides units into 1-dimensional equal units with equal-width bin intervals as grid. Threshold and bin intervals are the inputs for this algorithm [8]. It uses Apriori-Reasoning method as the step recursively from q-1-dimensional units to q-

dimensional units using self-join of q-1. The total subspaces are sorted based on their coverage. The subspaces which are less covered are pruned.

Hierarchical Based Clustering

Hierarchical clustering techniques is one of the clustering technique used for clustering of high dimensional data. Hierarchical clustering techniques works in two ways. One is Agglomerative (top-bottom) and another on is Divisive (bottom- top). In Agglomerative approach,initially starts with one object and successively merges the neighbour objects based on the distance (minimum, maximum and average) . The process is continuous until a desired cluster is shaped. In Divisive approach, starts with set of objects as single cluster and divides them into further clusters until desired number of clusters are shaped.

Future work:

Clustering is a powerful data exploration tool capable of uncovering previously unknown patterns in data. Often, users have little knowledge of the data prior to clustering analysis and are seeking to find some interesting relationships to explore further. Unfortunately, all clustering algorithms require that the user set some parameters and make some assumptions about the clusters to be discovered. Subspace clustering algorithms allow users to break the assumption that all of the clusters in a dataset are found in the same set of dimensions. There are many potential applications with high dimensional data where subspace clustering approaches could help to uncover patterns missed by current clustering approaches. Applications in bioinformatics and text mining are particularly relevant and present unique challenges to subspace clustering. As with any clustering techniques, finding meaningful and useful results depends on the selection of the appropriate technique and proper tuning of the algorithm via the input parameters. In order to do this, one must understand the dataset in a domain specific context in order to be able to best evaluate the results from various approaches. One must also understand the various strengths, weaknesses, and biases of the potential clustering algorithms.

REFERENCES

1. Yonggang Lu, Xiaoli Hou and Xurong Chen, A novel travel-time based similarity measure for hierarchical clustering, Neurocomputing, 173, (3), (2016).
2. Hu Yang and Xiaoqin Liu, Studies on the Clustering Algorithm for Analyzing Gene Expression Data with a Bidirectional Penalty, Journal of Computational Biology, 24, 7, (689), (2017)
3. Karen Sargsyan, Cédric Grauffel and Carmay Lim, How Molecular Size Impacts RMSD Applications in Molecular Dynamics Simulations, Journal of Chemical Theory and Computation, 13, 4, (1518), (2017).
4. Onapa Limwattanapibool and Somjit Arch-int, Determination of the appropriate parameters for K-means clustering using selection of region clusters based on density DBSCAN (SRCD-DBSCAN), Expert Systems, 34, 3, (2017)
5. Marco Livesu, A heat flow based relaxation scheme for n dimensional discrete hyper surfaces, Computers & Graphics, 71, (124), (2018).
6. Divya Pandove, Shivan Goel and Rinkl Rani, Systematic Review of Clustering High-Dimensional and Large Datasets, ACM Transactions on Knowledge Discovery from Data, 12, 2, (1), (2018)
7. Kavan Fatehi, Mohsen Rezvani, Mansoor Fateh and Mohammad-Reza Pajoohan, Subspace Clustering for High-Dimensional Data Using Cluster Structure Similarity, International Journal of Intelligent Information Technologies, 10.4018/IJIIT.2018070103, 14, 3, (38-55), (2018).
8. Parul Agarwal and Shikha Mehta, Subspace Clustering of High Dimensional Data Using Differential Evolution, Nature-Inspired Algorithms for Big Data Frameworks, 10.4018/978-1-5225-5852-1.ch003, (47-74)(2018)