1. **A description of the problem to be solved**

Clustering

According to the course Mining of Massive Datasets textbook, “Clustering is the process of examining a collection of “points,” and grouping the points into “clusters” according to some distance measure.”

Clustering is an unsupervised Machine Learning technique used to group which automatically groups unlabeled data. It looks for similar patterns in the dataset such as shape, size, color, etc. Later, it divides them in presence or absence of similar patterns.

We can describe clustering as an attempt to understand the structure of a cloud of data points.

To use a clustering algorithm, the dataset should be a collection of points that belongs to some space. The space is a universal set of points, from where the points can be drawn.

Chart, scatter chart

Description automatically generated

**Figure 1**

The most common uses of clustering are:

* Market Segmentation
* Statistical data analysis
* Social network analysis
* Image segmentation
* Anomaly detection, etc.
* Recommendation systems (Amazon, Netflix)

Clustering Strategies

* Hierarchical: It is a type of clustering, that starts with a single point cluster, and moves to merge with another cluster, until the desired number of clusters are formed.

Clusters are combined based on their “closeness,” using one of many possible definitions of “close.”

* + Agglomerative
  + Divisive
* Point Assignment: Points are considered in some order, and each one is assigned to the cluster into which it best fits. This process is normally preceded by a short phase in which initial clusters are estimated. Variations allow occasional combining or splitting of clusters, or may allow points to be unassigned if they are outliers (points too far from any of the current clusters)

These algorithms work with data in high dimensions.

The main goal of the clustering algorithms is to group points from a data set, with a notion of distance between points, in a way that:

Members of a cluster are close when the look like each other

Members of different clusters are dissimilar

Usually points are in a high-dimensional space

Similarity is defined using a distance measure like Euclidean, Cosine or Jaccard

Diagram

Description automatically generated

Figure 2

The Curse of Dimensionality

It is the phenomenon that occurs when handling high dimensional data. It manifests when the distance of the points is equal from one another, or when two vectors are orthogonal.

1. **A description of the algorithm**

**CURE - Clustering Using Representatives**

The CURE algorithm is a hierarchical based clustering technique, that adopts a middle ground between the centroid based and the all-point extremes.

This algorithm is used for identifying spherical and non-spherical clusters. Allows clusters to take any shape, in other words the CURE algorithm assumes Euclidean space and identifies clusters of any space.

It can be good for discovering groups and identifying interesting distributions in the underlying data.

CURE uses a collection of representative points for efficiently handling the clusters and eliminating the outliers.

1. **An analysis of the algorithm's performance:**

**Why does it work?**

The main point of the CURE algorithm is to create partitions starting with a random sample from the data set. This is performed recursively, allowing to remove outliers, create the partial clusters and finally labeling the data.

Algorithm Steps

Diagram

Description automatically generated

Figure 3

Initialization

* Take a small random sample and cluster it to the main memory using hierarchical clustering
* Select a small set of points from each cluster to be representative points. The points will be selected as far as possible from one another.
* Move the partition of representative points to a fixed fraction of the distance between its location and the centroid of its cluster. This fraction can be 20% or 30% of the distance.
* After the initialization completes, cluster the remaining points and create an output cluster.
* Finally, merge two clusters if they representative points sufficiently close, defining a distance threshold. Repeat this step with all the clusters until there are no more close points.

**What is its cost?**

The cost of using hierarchical clustering could be high because this type of algorithms requires to compute distances between pairs of clusters. For example, let’s think about a basic hierarchical clustering algorithm:

The first step takes O(n^2) time.

Subsequent steps take proportional time (n − 1)^2, (n− 2)^2,…

The algorithm ends up being cubic O(n^3).

This could limit the use of the algorithm only for a small data set.

Using random samples and partitions help to improve the efficiency of the CURE algorithm.   
The cure algorithm is designed using a combination of partition based and hierarchical algorithms.

**Pros and cons?**

* The CURE algorithm is capable to identify arbitrary shape clusters.
* The algorithm is robust in the presence of outliers.
* It is suitable for handling large data sets.
* CURE is not good for noise handling.
* Given the quantity of the data it can handle the algorithm is slow
* The algorithm can’t be applied to large databases because of its high execution time

1. **A concrete example of how the algorithm works:**    - Choose a suitable dataset or generate one  
       - Implement the algorithm  
       - Show the results obtained from the algorithm