Parameterless clustering by dynamic tree-cutting

Simon Lehmann Knudsen
simkn15@student.sdu.dk

ECTS: 10
04/09-2017 - 31/01-2018
5th semester in Computer Science

31/01-2018

Contents

1 Abstract

1.1 Background

1.2 Results

1.3 Conclusion

2 Introduction

In ancient history it was important to realize that many objects shared certain properties like being edible or poisonous. Since it could be a matter of living or dieing. Categorizing is something most people, if not all, are doing daily without paying too much attention to it. Categorizing, or classifying, people as cute, clever, marriage-potential, etc..

Two key learnings methods are: (Does not fit here)

- **Supervised Learning**: The machine learning task of inferring a function $f : X \rightarrow Y$ given a set of labeled training data $\{\langle x_i, y_i \rangle\}$.
- **Unsupervised Learning**: The task of building a model of X explaining its structure and inherit properties.

In short the difference is that supervised learning has a training data given a priori, where there is none in unsupervised learning. Thus, making supervised learning more subjective. Clustering is an unsupervised learning method.

2.1 What is Clustering?

Clustering, or cluster analysis, is a way of grouping a set of objects such that a cluster (group) of objects is more similar to each other than those of another cluster. Clustering can help with the description of patterns of similarities and differences in a data set. Similarity is highly subjective depending on the scenario. How a cluster is recognized is not entirely clear when displayed in the plane. One intuitive approach is assessing the relative distances between points. Figure 1 shows three types of shapes for clusters. Looking at the middle one, assessing the distance between points would not necessarily give two clusters. Taking the left cluster, shape of the letter **C**. The top most right point and the bottom most right point would have a high distance insinuating low similarity. Visually the two points are in the same cluster. Figure 2 shows an example without any 'natural' clusters. Visual examination does not leave any clues on the number of clusters. This scenario can be the geographical locations of houses in a town, where it makes sense to divide the houses into postal districts. Clustering is used in many different fields like market research, astronomy, weather classification, bioinformatics and genetics, etc..

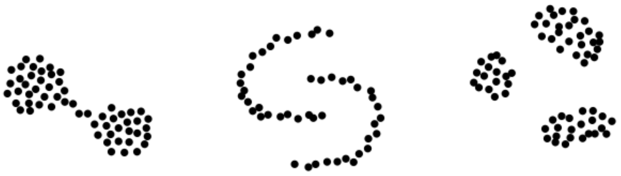


Figure 1: Different shapes of clusters

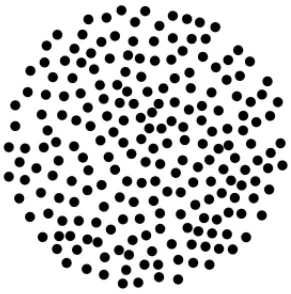
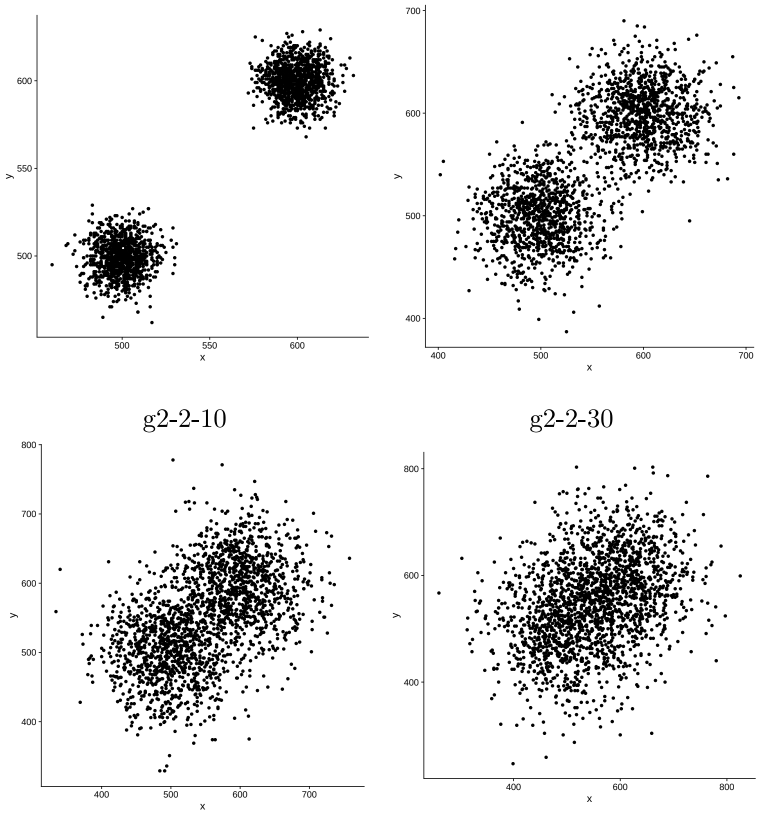


Figure 2

There are many different tools to do cluster analysis, which all intend to find an optimal clustering depending on a set of criteria. Given a set of criteria and parameters, we can analyze the data and discover the clusters. The resulting clusters can all be either feasible or infeasible. In some circumstances overlapping clusters can provide acceptable solutions. When there are no clear separation of clusters, it can be hard to determine if the solution is acceptable or not.
Figure 3 shows similar datasets with decreasing separating and density of the clusters. `g2-2-70` does not have any clear separation between the clusters, and visually determining the number of clusters is no longer an option.



g2-2-10

g2-2-30

g2-2-50

g2-2-70

Figure 3: URL: https://cs.joensuu.fi/sipu/datasets/

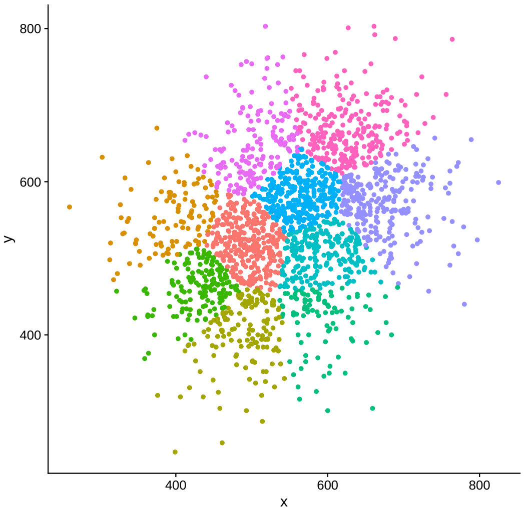


Figure 4: g2-2-70 clustering with 10 clusters

Clustering requires an understanding of the dataset, in order to select the proper features of the data to determine the similarity. E.g. the optimal clustering solution could be one with 10 clusters, shown in figure 4. A feature is some attribute which describes the data. Lets say the dataset was about the appearance of people. The features could be hair color, eye color, height, weight, etc.. The basic data for a cluster analysis starts with a $n \cdot p$ multivariate data matrix, `X`, which describes each object to be clustered. The entry x_{ij} gives the value of the jth variable on object i:

$$\mathtt{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{bmatrix}$$

Figure 5: Multivariate data matrix

Many clustering techniques begins by converting `X` into a symmetric $n \cdot n$ matrix. Both rows and columns represents the objects in the data set. The resulting matrix could be of similarities, dissimilarities or distances between all objects.

2.2 Clustering Approaches

There are many different clustering approaches. Each has its purpose dependent on the data. The three most notable approaches are

- **Partitional**: Requires a parameter k, which is the number of returned clusters while optimizing a criterion function, e.g. a distance function.
- **Hierarchical**: Proceeds by either merging smaller clusters into larger, or splitting larges cluster into smaller. Resulting in a dendrogram, which is a tree of clusters, showing the related clusters. The clustering solution is given by a cut on a desired level of the dendrogram.
- **Density-based**: Grouping neighboring objects into clusters based on a density condition.

With partitional clustering it can be hard to derive a proper k for the returned clusters. With low dimensional data plotting the data can be helpful to determine k, but gets increasingly difficult with high dimensional data. As a rule of thumb $1 \ll k \leq 10$. Calculating the F-ratio can possible give a good estimation for k. Density-based clustering is usually not a good approach if the dataset has a high variance of density between the clusters. Next is some algorithms under each approach

- **Reference: On Clustering Validation Techniques**
- Partitional algorithms

  K-Means

PAM (Partitioning Around Medoids)

CLARA (Clustering Large Applications)

CLARANS (Clustering Large Applications based on Randomized Search)

- Hierarchical algorithms:

  Agglomerative algorithms:

  Divisive algorithms:

  BIRCH (Zhang et al., 1996)

  CURE (Guha et al., 1998)

  ROCK (Guha et al., 1999)

- Density-based algorithms

  DBSCAN

  DENCLUE

  Transitivity Clustering (TC)

  A description of the algorithms can be found in the appendix section(Missing at the moment)

For this project the density-based algorithm Transitivity Clustering is used, and is incorporated into a hierarchical clustering approach.

2.3 Cluster validity

It is difficult to visually compare two clusterings if they differ in a few points, thus it is needed to calculate the quality of a clustering. Clustering is a unsupervised learning where are no training data to guide, so how do we measure the quality? There are many different methods, and which one to use can be different depending on the data. To validate a clustering there are two methods, internal measures and external measures. Internal measures are often based on

- Compactness: Measures how closely related the objects are in a cluster
- Separation: Measures how distinct or well-separated a cluster is from other clusters

Internal measures are e.g. Sum of Squares(SSQ), Silhouette Coefficient and Dunn Index. External measures use external information not present in the data. Normally a 'gold standard' is used. A gold standard is a clustering solution developed by experts, and are rarely available. Comparing against this ground truth is the way for determining the quality. External measures are e.g. F-measure and Jaccard coefficient. For this project F-measure is the used quality measure and will be further discussed later.

3 Background

3.1 The Dataset

3.2 Transitivity Clustering

Before going into any details about Transitivity Clustering(TC) we need some basic graph-theoretic definitions.

Definitions from 'Extension and Robustness of Transitivity Clustering for Protein...'

Definition 1 (`Undirected simple graph`)**.** An `undirected simple graph` $G = (V, E)$ consists of a set of nodes V and a set of edges $E \subseteq \binom{V}{2}$, where $\binom{V}{2}$ denotes the set of two-element subsets of V. The edges are undirected and contains no self-loops or multiple edges between two nodes. uv is an unordered par $\{u, v\} \in \binom{V}{2}$.

Definition 2 (`Transitive graph`)**.** An undirected simple graph $G = (V, E)$ is called `transitive`

$$\text{if for all triples } uvw \in \binom{V}{3}, uv \in E \text{ and } vw \in E \text{ implies } uw \in E.$$

Definition 3 (`Weighted Transitive Graph Projection Problem(WTGPP)`)**.** Given a set of objects `V`, a threshold $t \in \mathbb{R}$, and a pairwise similarity function sim: $\binom{V}{2} \to \mathbb{R}$, the graph G is defined as

$$G = (V, E);\ E = \left\{ uv \in \binom{V}{2} : \text{sim}(uv) > t \right\} \tag{1}$$

The WTGPP is the determination of a transitive graph $G' = (V, E')$ such that there exist no other transitive graph $G'' = (V, E'')$ with $\text{cost}(G \to G'') < \text{cost}(G \to G')$. The modification costs are defined as

$$\text{cost}(G \to G') := \underbrace{\sum_{uv \in E \setminus E'} |\text{sim}(uv) - t|}_{\text{deletion cost}} + \underbrace{\sum_{uv \in E' \setminus E} |\text{sim}(uv) - t|}_{\text{addition cost}} \tag{2}$$

Transitivity Clustering takes one parameter, *t*, which is the threshold for similarities. Following is the steps in Transitivity Clustering:
Reference: Comprehensive cluster analysis with Transitivity Clustering

1. Model the given pairwise similarity, from the similarity matrix, as a similarity graph, *G*. The nodes corresponds to the objects, with weighted edges as the similarity values.
2. Transform the similarity graph, *G*, into another graph, *G'*, by subtracting the threshold from the edge weights. Subsequently removing those edges with weights below zero, which is the deletion cost for equation 2.
3. Transform *G'* into a transitive graph, *G"*, with minimal cost. Thus, in this step we add all edges such that the graph is transitive, which is the addition cost for equation 2.

The resulting transitive graph, *G"*, is the clustering solution.

3.3 Hierarchical Clustering

Definition 4 (`Hierarchical Clustering(HC)`). **Reference from Unsupervised Learning slides: Hierarchical Clustering**
Builds a nested structural partition $C = \{C_1, \dots, C_k\}$ of V such that $\cup_{i=1}^{k} C_i = V$ and $C_i \neq \emptyset \ \forall i \in \{1, \dots, k\}$. $\forall$ pairs C_i, C_j where $i, j \in \{1, \dots, k\}, i \neq j$, exactly one of the following holds

- $C_i \cap C_j = \emptyset$
- $C_i \subset C_j$
- $C_j \subset C_i$

There are two forms of hierarchical Clustering, `agglomerative` and `divisive`. Agglomerate starts with n clusters, where n is the number of objects in the data set. Joining clusters until one cluster is remaining, where all n objects are members. Divisive is the opposite. Starting with one cluster with n objects. Splitting the clusters until all clusters are singletons. Thus, all n objects represents are cluster. Agglomerative is the most used, as divisive usually is a more expensive procedure. One important feature of hierarchical clustering is that a join or split of clusters are irrevocable, thus cannot be undone. Figure 6 shows an overview of agglomerative vs. divisive.

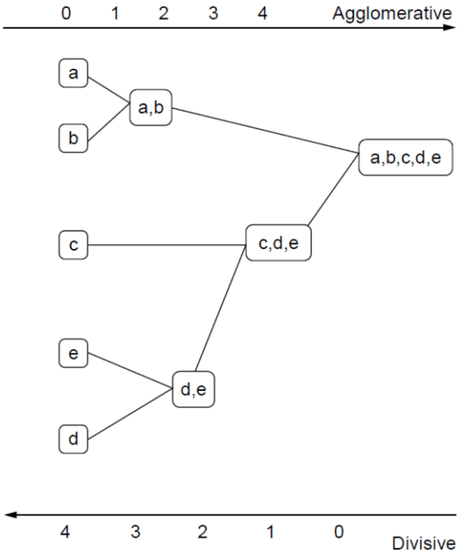


Figure 6: Agglomerative vs. Divisive: Reference Cluster Analysis, page 72

The steps of joins or splits is often showed as a `dendrogram`, which is viewed as a tree structure. The root of the tree is the cluster containing all n objects. Moving down the tree the nodes represents the clusters which was split from its parent. At the bottom of the tree we have the leafs, where the number of leafs represents all singleton clusters(the n objects). The tree can be cut a given height resulting in a clustering solution. Figure 7 shows a `dendrogram` of the tree structure of a HC. The horizontal axis shows all the objects in the data set.

The vertical axis shows the distances between objects and/or clusters. Objects 1 and 2 are joined to a cluster at height 2. These are joined with the remaining objects at height 5, resulting in one cluster holding all objects.

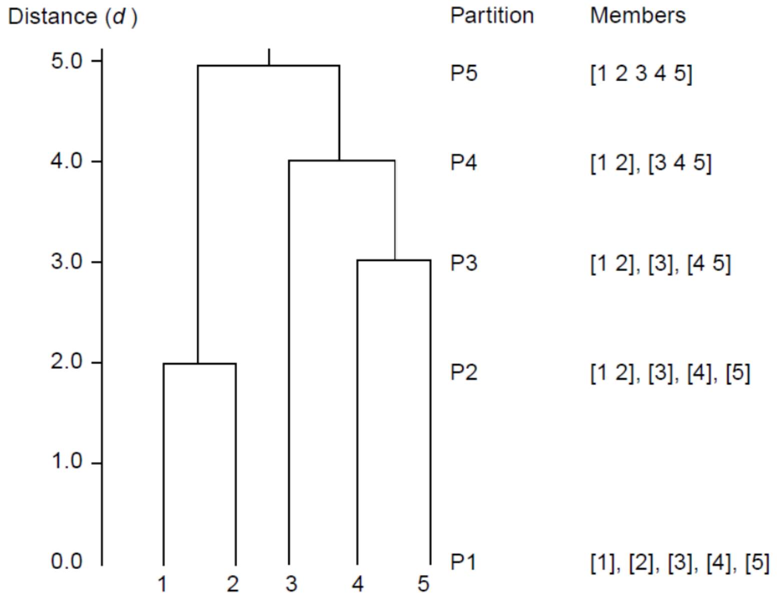


Figure 7: Dendrogram of a Hierarchical Clustering: Reference Cluster Analysis, page 75

In a typical hierarchical Clustering the size and number of clusters are given by a threshold, much like the one used in `TC`. Meaning that one iteration can possibly affect all clusters, which either increases or decreases the overall quality. In Figure 7 the tree could be cut between distance 3-4 to obtain a solution containing three clusters.

3.4 Cluster Validation - F-measure

Often it is crucial to measure the quality of a clustering result. Evaluating the best of two results can be difficult, if not impossible, when looking at plots of the clusterings. Thus, we need other methods to validate the results. `Cluster validity indexes` are a important factor to evaluate solutions. There are three categories of cluster validity indexes:

- External Criteria/Validation: Evaluate the result with respect to a pre-specified structure, such as a gold standard.
- Internal Criteria/Validation: Evaluate the result with respect to information intrinsic to the data alone.
- Relative Criteria: Choosing the best clustering scheme of a set of defined schemes, according to a pre-specified criterion.

We are using the brown data set, which has a gold standard. Thus, we can evaluate results with the external criteria. F-measure falls under this category,

and is the chosen quality measure.There are multiple versions of the F-measure. We will be using the $\mathrm{F}_1 - measure$, where the measures Recall and Precision is weighted equally. In order to understand the F-measure we need the following definitions:
$K = (K_1, \ldots, K_m)$ = Clustering result obtained from the algorithm. K_i is the ith cluster in K.
$G = (G_1, \ldots, G_l)$ = Gold standard clustering. G_j is the jth cluster in G.
n = amount of objects in the data set.
n_i = number of objects in cluster K_i.
n^j = number of objects in cluster C_j.
n_i^j = number of objects contained in $K_i \cap C_j$

Definition 5 (True Positive)**.** The number of common objects between cluster i and the compared gold standard cluster j.

$$\mathrm{TP} = |K_i \cup C_j| \tag{3}$$

Definition 6 (False Positive)**.** The number of objects in cluster i, which are not in the compared gold standard cluster j.

$$\mathrm{FP} = |K_i \backslash C_j| \tag{4}$$

Definition 7 (False Negative)**.** The number of objects that are not in cluster i, which are in the compared gold standard cluster j

$$\mathrm{FN} = |C_j \backslash K_i| \tag{5}$$

Definition 8 (Recall)**.**

$$\mathrm{Recall}(i,j) = \frac{n_{ij}}{n_i} = \frac{TP}{TP + FP} \tag{6}$$

Definition 9 (Precision)**.**

$$\mathrm{Precision}(i,j) = \frac{n_{ij}}{n_j} = \frac{TP}{TP + FN} \tag{7}$$

Definition 10 (F-measure for a cluster)**.** The F-measure of cluster j and class i is given by:

$$F(i,j) = 2 \cdot \frac{\mathrm{Recall}(i,j) \cdot \mathrm{Precision}(i,j)}{\mathrm{Precision}(i,j) + \mathrm{Recall}(i,j)} \tag{8}$$

Definition 11 (F-Measure for a clustering)**.** In order to obtain the F-measure for a clustering solution, we need to find the mean F-measure. The F-measure of a cluster j is multiplied by the amount of objects in the gold standard cluster which have most in common objects. Take the sum over all clusters. Divide by total amount of objects in the data set. Each cluster from the gold standard can only be referenced/mapped once.

$$\frac{\sum_{i=1}^{m} \text{F-measure}(K_i) \cdot n_i^j}{n} \tag{9}$$

A F-measure is between 0 and 1. A value near 1 indicates a good match with the gold standard(good clustering result). Values near 0 indicates a bad result. It is important to pay attention to the statement in Definition 11 saying that each cluster from the gold standard can only be referenced once. If, e.g. three clusters from K maps to the same cluster in C, two clusters will be neglected. The F-measure will be 0.0 for the neglected clusters, which negatively influences the quality. This scenario can happen in all clustering solutions, but the probability for multiple mappings on to the same gold standard cluster increases when $|K| > |C|$. In the opposite direction the scenario where $|K| < |C|$, also has a negative impact. With $|K| < |C|$ the clusters in K would be bigger than the cluster in C. Looking back at Definition 11, we see that the multiplication is done with the size of C_j. Therefore we can conclude that the amount of objects that K_i is bigger than C_j, $|K_i| - |C_j|$, has a smaller influence on the total F-measure than the remaining objects in K_i.

3.5 Multidimensional Scaling

3.6 GAP Statistics

4 Method

4.1 Assessing best clustering

4.2 Randomization approach 1

4.3 Randomization approach 2

4.4 Randomization approach 3

4.5 Randomization approach 4

4.6 Results for randomizations

5 Implementation : Keep it short!

6 Conclusion

7 Appendix

7.1 Clustering Approaches and Algorithms : Missing description of each algorithm

- **Reference: On Clustering Validation Techniques**
- Partitional algorithms

  K-Means

  PAM (Partitioning Around Medoids)

  CLARA (Clustering Large Applications)

  CLARANS (Clustering Large Applications based on Randomized Search)
- Hierarchical algorithms:

  Agglomerative algorithms:

  Divisive algorithms:

  BIRCH (Zhang et al., 1996)

  CURE (Guha et al., 1998)

  ROCK (Guha et al., 1999)
- Density-based algorithms

  DBSCAN

  DENCLUE

  Transitivity Clustering
- Grid-based algorithms

  STING (Statistical Information Grid-based method)

  WaveCluster
- Fuzzy clustering

  Fuzzy C-Means (FCM)

  EM (Expectation Maximization)

8 References