Sources: 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

December 30, 2017

Contents

1 Articles

- A gold standard set of mechanistically diverse enzyme superfamilies.
  Shoshana D Brown*, John A Gerlt†, Jennifer L Seffernick‡ and Patricia C Babbitt§.
  Article downloaded from SDU library and located in articles folder.

- Comprehensive cluster analysis with Transitivity Clustering.
  Tobias Wittkop, Dorothea Emig, Anke Truss, Mario Albrecht, Sebastian Böcker & Jan Baumbach
  Article downloaded from SDU library and located in articles folder.

- Comparing the performance of biomedical clustering methods.
  Christian Wiwie1, Jan Baumbach & Richard Röttger
  Article downloaded from sdu library and located in articles folder.

- Large scale clustering of protein sequences with FORCE -A layout based heuristic for weighted cluster editing.
  Tobias Wittkop, Jan Baumbach, Francisco P Lobo and Sven Rahmann.
  Article downloaded from unsupervised learning and located in articles folder.

- Partitioning biological data with transitivity clustering
  Tobias Wittkop, Dorothea Emig, Sita Lange, Sven Rahmann, Mario Albrecht, John H Morris, Sebastian Böcker, Jens Stoye & Jan Baumbach.
  Article downloaded from unsupervised learning "TransClust Supplement".

- On Clustering Validation Techniques
  Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis
  Article received from Richard and located in articles folder.

- Internal versus External cluster validation indexes
  Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz.
  Article received from Richard and located in articles folder.

- Estimating the number of clusters in a data set via the gap statistic
  Robert Tibshirani, Guenther Walther and Trevor Hastie
  Article downloaded from: https://web.stanford.edu/ hastie/Papers/gap.pdf

2 Notes

2.1 On Clustering Validation Techniques

2.1.1 Introduction

- Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters (Guha et al., 1998).

- In the clustering process, there are no predefined classes and no examples thatwould show what kind of desirable relations should be valid among the data that is why it is perceived as an unsupervised process (Berry and Linoff, 1996).

- On the other hand, classification is a procedure of assigning a data item to a predefined set of categories (Fayyad et al., 1996).

Clustering produces initial categories in which values of a data set are classified during the classification process. The basic steps to develop clustering process:

- Feature selection
- Clustering algorithm

  Proximity measure

  Clustering criterion
- Validation of the results: The final partition of data requires some kind of evaluation in most applications (Rezaee et al., 1998)
- Interpretation of the results

Clustering algorithms can be classified according to:

- The type of data input to the algorithm.
- The clustering criterion defining the similarity between data points.
- The theory and fundamental concepts on which clustering analysis techniques are based (e.g. fuzzy theory, statistics).

Thus according to the method adopted to define clusters, the algorithms can be broadly classified into the following types (Jain et al., 1999):

- Partitional clustering.
- Hierarchical clustering.
- Density clustering.
- Grid-based clustering.

For each of above categories there is a wealth of subtypes and different algorithms for finding the clusters. Thus, according to the type of variables allowed in the data set can be categorized into (Guha et al., 1999; Huang et al., 1997; Rezaee et al., 1998):

- Statistical.
- Conceptual: which are used to cluster categorical data. They cluster objects according to the concepts they carry.
- Fuzzy clustering.
- Crisp clustering: considers non-overlapping partitions meaning that a data point either belongs to a class or not. Most of the clustering algorithms result in crisp clusters, and thus can be categorized in crisp clustering.
- Kohonen net clustering.

Input parameters of an algorithm can be number of clusters, density of clusters, etc.. Attempt to define the best partitioning of a data set for the given parameters. Thus, they do not necessarily give the "best" partitioning. Since clustering algorithms discover clusters, which are not known a priori, the final partitions require some sort of evaluation in most applications (Rezaee et al., 1998).

2.1.2 Clustering algorithms

- Partitional algorithms

  K-Means: Optimization of an objective function that is described by the equation: $E = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i)$.

  PAM (Partitioning Around Medoids)

  CLARA (Clustering Large Applications)

  CLARANS (Clustering Large Applications based on Randomized Search)

  K-prototypes, K-mode are based on K-Means, but aims at clustering categorical data.

- Hierarchical algorithms (Theodoridis and Koutroubas, 1999):

  Agglomerative algorithms: Decreasing number of clusters.

  Divisive algorithms: Increasing number of clusters.

  BIRCH (Zhang et al., 1996)

  CURE (Guha et al., 1998)

  ROCK (Guha et al., 1999)

- Density-based algorithms

  DBSCAN

  DENCLUE

- Grid-based algorithms

  STING (Statistical Information Grid-based method)

  WaveCluster

- Fuzzy clustering

  Fuzzy C-Means (FCM)

  EM (Expectation Maximization)

2.1.3 Cluster validity assessment

- Problem specification

  Defining the optimal number of clusters

- Fundamental concepts of cluster validity

  Three approaches: external criteria, internal criteria and relative criteria.

  External criteria: Monde Carlo

  Compactness

  Separation: Single linkage, complete linkage, comparison of centroids.

2.2 Internal versus External cluster validation indexes

- The purpose of clustering is to determine the intrinsic grouping in a set of unlabeled data, where the objects in each group are indistinguishable under some criterion of similarity. Clustering is an unsupervised classification process fundamental to data mining (one of the most important tasks in data analysis).

- *Compactness*: This measures closeness of cluster elements. A common measure of compactness is variance.
  Separability: This indicates how distinct two clusters are. It computes the distance between two different clusters. The distance between representative objects of two clusters is a good example. This measure has been widely used due to its computational efficiency and effectiveness for hypersphereshaped clusters.

- In recent times, many indexes have been proposed in the literature, which are used to measure the fitness of the partitions produced by clustering algorithm.

- The Dunn index, DB measures, SD index, S_Dbw index, CS index, BIC index. (Check if all these are internal!)

- External measures include Entropy, Purity, NMIMeasure and F-Measure.

Since clustering algorithms discover clusters, which are not known a priori, the final partitions of a data set requires some sort of evaluation in most applications (Rezaee et al.,1998).

Clustering is an unsupervised classification process fundamental to data mining. Data mining is one of the most important tasks in data analysis. Applications in several fields like bioinformatics[14], web data analysis [13], text mining [17] and scientific data exploration [1]. Clustering refers to unsupervised learning, where it has no a priori information about the data. A clustering algorithm depends on input parameters. k-means require a number of clusters (k) to be created. The question remains, what is the optimal number of clusters? Cluster validity indexes are a important factor as a means to give a solution [7]. Many methods have been proposed without any a priori information.

- External Criteria: Evaluate the result with respect to a pre-specified structure.

  External validation: Based on previous knowledge about data.

- Internal Criteria: Evaluate the result with respect a information intrinsic to the data alone.

  Internal validation: Based on the information intrinsic to the data alone.

Most clustering algorithms are dependent on the characteristics of the dataset and the input parameters. Incorrect parameters may lead to clusters that deviate from the optimum solution. Usually clustering validity indexes are defined by combining `compactness` and `separability`.

Compactness