VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Data Analysis 3

Feature selection, Representative-based Clustering, Hierarchical Clustering

Jan Platoš

September 29, 2021

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Unsupervised learning

Unsupervised learning

Unsupervised machine learning is the machine learning task of inferring a function that describes the structure of ”unlabeled” data (i.e. data that has not been classified or categorized).

Clustering

Clustering

Given a set of data points, partition them into groups containing very similar points.

- Possible applications:
  - Data summarization
  - Customer Segmentation
  - Social network analysis
  - Preprocessing data for other algorithms (classification, outliers detection, etc.)

Feature Selection for Clustering

- How to select only the features that are important?
- How to measure the ability of the feature to cluster objects?
- Two main approaches exists:
  - Filter models
  - Wrapper models

Feature Selection for Clustering - Filter models

Filter models

- Definition of measures that may evaluate the quality of a feature or a feature combination

Term Strength

- Suitable for Text data or other sparse documents.
- A conditional probability that a selected term appear in a document Y when it appear in a document X.
- The documents pairs are randomly sampled from similar documents.

$$Term\ strength = P(t \in Y | t \in X)$$

Feature Selection for Clustering - Filter models

Predictive Attribute Dependence

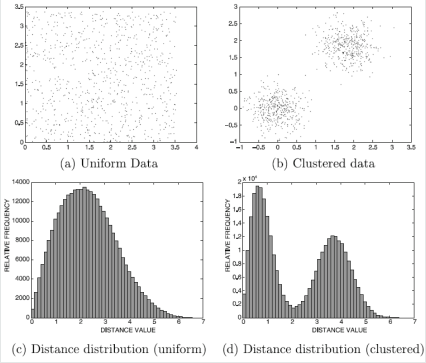
- Correlated features should end in better results than uncorrelated.
- Set of features should predict the value of another correlated feature.
- A regression or classification algorithm is used as a predictor.

Predictive Attribute Dependence - Principle

- For each feature:
  - Use other features to predict the value of the selected feature.
  - Compute the accuracy of prediction and use it as relevance measure.
- All relevant features are used for clustering.

Feature Selection for Clustering - Filter models

Entropy



- Highly clustered data reflects some characteristics in the underlying distance distribution.
- The goal is to quantify the shape of the distance distribution on a given subset of features and to pick the most suitable subset.
- A systematic way to search possible combination of features.
- A algorithm for entropy measure have to be defined (greedy, ...).

Feature Selection for Clustering - Filter models

Entropy

- For k-features a k-dimensional space is defined.
- Define m multidimensional grid regions over this space.
- The p_i is a fraction of points in a region i.

$$E = \sum_{i=1}^{m} p_i \log(p_i)$$

- Distribution with poor clustering behavior results in high entropy.
- Alternatively an entropy over a distance distribution may be computed.

Feature Selection for Clustering - Filter models

Hopkins Statistics

- Evaluates the clustering tendency of the whole datasets/selected features (greedy approach may be used).
- D is a dataset, R is a random sample of r points from D, sample S of r randomly generated points.
- The $a_1, a_2, \ldots, a_r$ are distances from points from R to the nearest neighbors from D and $b_1, \ldots, b_r$ are distances from points from S to the nearest neighbors from D.

$$H = \frac{\sum_{i=1}^{r} b_i}{\sum_{i=1}^{r} (a_i + b_i)}$$

- H is in the range $(0, 1)$.
- Uniformly distributed data will have $H = 0.5$, H close to 1 means clustered data.
- Due to random sampling, each tests will produce different results.
- Confidence test have to be used for statistical importance.

Feature Selection for Clustering - Wrapper models

Wrapper models

- Uses a cluster validity criterion for feature subset evaluation.
- The are highly imperfect.
- Search space of features is exponentially related to the dimensionality.
- The principle is sensitive to the choice of the validity criterion.
- Simpler methodology:
  - Cluster points according the selected feature subset.
  - Assign labels to the points according to the cluster the points belongs to.
  - Use supervised criterion to measure the quality of each feature.

Representative-based Algorithms

- Simplest clustering algorithm.
- Based on the distance/similarity measure between points.
- Clusters are created in a single step.
- Hierarchical relationships do not exist among different clusters.
- The representatives are computed or selected from the cluster.
- A distance function is used for defined representatives to assign points into closest representative/cluster.

Representative-based Algorithms

- A number of clusters k is usually defined by the user.
- A dataset D contains n data points $X_1, \ldots, X_n$.
- The goal is to determine k representatives $Y_1, \ldots, Y_k$ that minimizes function O:

$$O = \sum_{i=1}^{n} \left[min_j Dist(X_i, Y_i) \right]$$

- i.e. the sum of the distances of the different data points to their closest representatives needs to be minimized.

Representative-based Algorithms

- The position of representatives and points assignment is not known a priori.
- An Iterative approach is used to solve the problem of representative/point assignment.
- General approach for representative-based algorithms:
  1. Initialize k representatives (random sampling or other method)
  2. Assign each data point to its closest representative using distance function $Dist(\cdot, \cdot)$
  3. Form a clusters from points assigned to each of the representative.
  4. Determine the optimal representative for each cluster C_j using minimization of a local objective function $\sum_{X_i \in C_j} [Dist(X_i, Y_j)]$

Representative-based Algorithms

Algorithm 1: GenericRepresentative(Database: D, Number of Rpresentatives: k)

1 **begin**
2 Initialize representative set S;
3 **repeat**
4 Create clusters $(C_1, \ldots, C_k)$ by assigning each point in D to closest representative in S using the distance function $Dist(\cdot, \cdot)$;
5 Recreate set S by determining one representative Y_j for each C_j that minimizes $\sum_{X_i \in C_j} Dist(X_i, Y_j)$;
6 **until** *convergence*;
7 **return** $(C_1, \ldots, C_k)$
8 **end**

Definition of the distance function

- The distance function $Dist(X, Y)$ defines the behavior of the algorithm.
- The general definition is a L_p-norm.

$$Dist(X, Y) = \|X - Y\|_p^p = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

- Another possibility is a cosine measure.

$$\cos(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^{d} (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$

Definition of the distance function

- Mahalanobis distance is a measure that takes into account a statistical distribution along each dimension

$$Maha(X, Y) = \sqrt{(X - Y)\Sigma^{-1}(X - Y)^T}$$

- Where Σ is a covariance matrix where Σ_{ij} is a variance between i-th and j-th dimension.

$$\Sigma = \begin{bmatrix} E\left[(X_1 - \mu_1)(X_1 - \mu_1)\right] & E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] & \cdots & E\left[(X_1 - \mu_1)(X_d - \mu_d)\right] \\ \vdots & & \ddots & \vdots \\ E\left[(X_d - \mu_d)(X_1 - \mu_1)\right] & E\left[(X_d - \mu_d)(X_2 - \mu_2)\right] & \cdots & E\left[(X_d - \mu_d)(X_d - \mu_d)\right] \end{bmatrix}$$

- μ_i is expected value of the dimension i.

Representative-based Algorithms - K-medians Algorithm

K-medians algorithm

- Uses an Manhattan distance for measuring distance between data points and representatives.

$$Dist(X, Y) = \|X - Y\|_1^1 = \sum_{i=1}^{d} |x_i - y_i|$$

- The optimal representatives for each cluster is a **median along each dimension in a cluster**.
- The k-medians algorithm is more robust than k-means: median is not as sensitive to outliers as the mean.

Representative-based Algorithms - K-means Algorithm

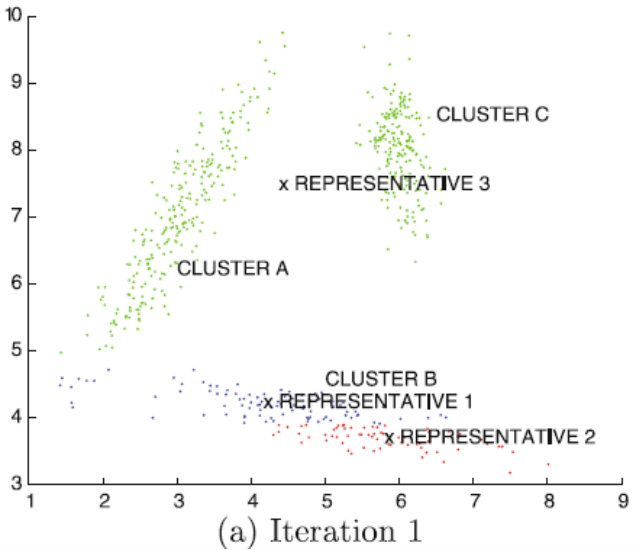
K-means algorithm

- Uses an Euclidean distance for measuring distance between data points and representatives.

$$Dist(X, Y) = \|X - Y\|_2^2 = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

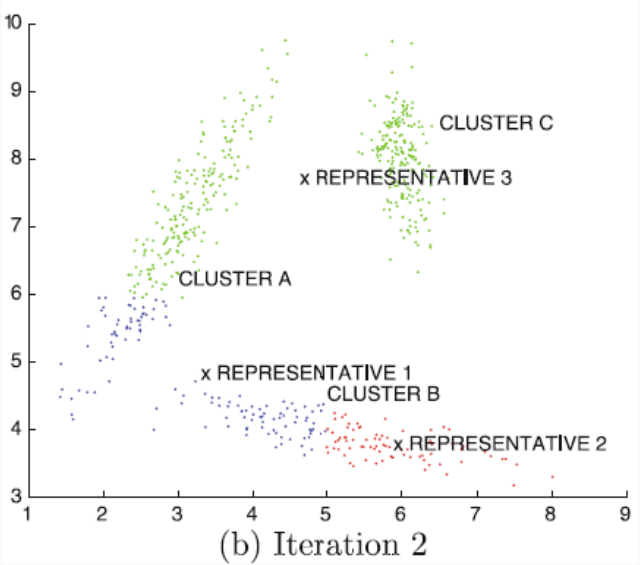
- The optimal representatives for each cluster is a mean.
- The k-means does not work well when the clusters are not spherical.
- A kernel variant of this algorithm is possible.

Representative-based Algorithms - K-means Algorithm



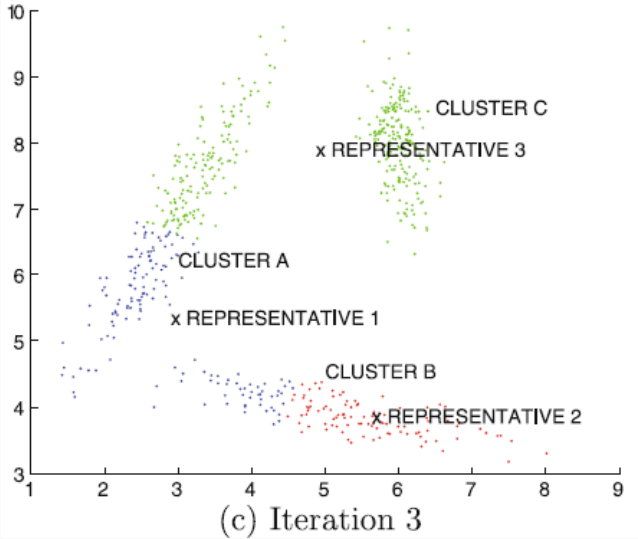
(a) Iteration 1

Representative-based Algorithms - K-means Algorithm



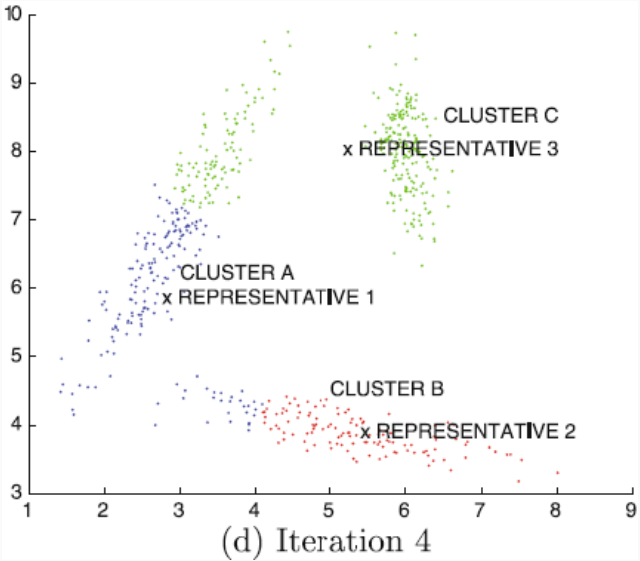
(b) Iteration 2

Representative-based Algorithms - K-means Algorithm



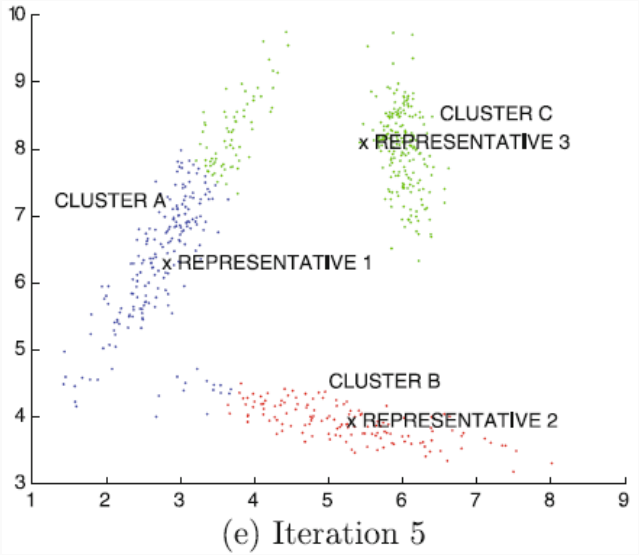
(c) Iteration 3

Representative-based Algorithms - K-means Algorithm



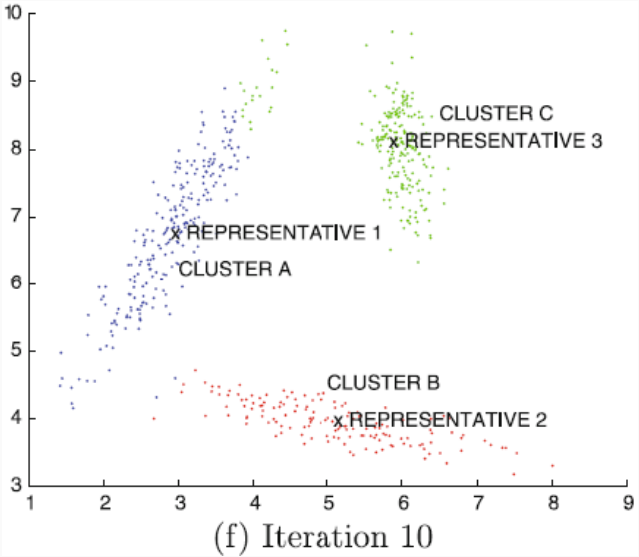
(d) Iteration 4

Representative-based Algorithms - K-means Algorithm



(e) Iteration 5

Representative-based Algorithms - K-means Algorithm



(f) Iteration 10

Representative-based Algorithms - K-medoids Algorithm

K-medoids algorithm

- The representatives are always selected from the dataset.
- Why:
  - For means: the representatives may lie outside the cluster in an empty region due to presence of outliers.
  - Sometimes it is difficult to compute the optimal central representative of a set of data points of a complex data type, e.g. set of time series of varying lengths.
- The *k*-medoid approach may be defined for any data type as long as a distance/similarity function is known.

Representative-based Algorithms - K-medoids Algorithm

K-medoids algorithm

- A set of representatives Y is randomly selected from dataset.
- Iterative improvement is performed on the set of representatives:
  - Exhaustive search (extremely expensive).
  - A set of pairs (X, Y) is randomly generated, X from dataset and Y from representatives, and the best pair is used for exchange.
- The k-medoids is slower than k-means.
- May be scalable implemented.

Representative-based Algorithms - Implementation Issues

Selection of k

- Very difficult automatically.
- Usually a large value is chosen.
- A post processing step may reduce number of clusters.

Initialization

- Algorithms are robust to the choice of the initialization step.
- Random generation of the representatives.
- Random sampling of the dataset.
- A centroids of m randomly chosen samples.

Questions?