Skript zur Vorlesung
Datenbanksysteme I
im Wintersemester 2019/20

Kapitel 11: Clustering

Vorlesung: Christian Böhm
Übungen: Dominik Mautz, Collin Leiber

http://dmm.dbs.ifi.lmu.de
Motivation

- Big data sets are collected in databases
- Manual analysis is no more feasible
Big Data


• “The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus […]”

• “Data have swept into every industry and business function and are now an important factor of production, alongside labor and capital”
  – Potential Revenue in US Healthcare: > $300 Million
  – Potential Revenue in public sector of EU: > €100 Million

• “There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.”
Big Data

• Data Mining is obviously an important technology to cope with Big Data

• Caution: “Big Data” does not only mean “big”

  => Three V’s (the three V’s characterizing big data)

  – Volume    Many objects but also huge representations of single objects
  – Velocity  Data arriving in fast data streams
  – Variety   Not only one type of data, but different types, semi- or unstructured
A Paradigm Shift in Science?

- Some 1,000 years ago, science was empirical (describing natural phenomena)
- Last few hundred years, science was theoretical (Models, generalizations)
- Last few decades, science became computational (data intensive)
  - Computational methods for simulation
  - Automatic data generation, high-throughput methods, …
- Data Science
Definition KDD

[Fayyad, Piatetsky-Shapiro & Smyth 1996]

„Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying patterns in data which are
• valid
• novel
• potentially useful
• and ultimately understandable“
Feature Vectors Associated to Objects

• Objects of an application are often complex
• It is the task of the KDD expert to define or select suitable features which are relevant for the distinction between various objects

Example: CAD-drawings:

Possible features:
• height $h$
• width $w$
• Curvature parameters $(a, b, c)$
Feature Vectors Associated to Objects

- In a statistical context, we call the features often *variables*.
- The selected features form a *feature vector*.
- The feature space is often high-dimensional (in our example 5-D). 

\[ ax^2 + bx + c \]

\( (h, w, a, b, c) \)
Further Examples of Features

Image Databases: Color Histograms

Genetic Databases: Level of Gene Expression

Text-/Document-DBs: Frequency of terms

The feature-based approach facilitates a uniform methodology for a great variety of applications
# Levels of Measurement

<table>
<thead>
<tr>
<th>Nominal (Categorical)</th>
<th>Ordinal</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Properties:</strong></td>
<td><strong>Properties:</strong></td>
<td><strong>Properties:</strong></td>
</tr>
<tr>
<td>We can only determine if two values are equal or not. No „better“ and „worse“, no directions. Features with 2 possible values are called <em>dichotome</em></td>
<td>We have a ordering relation (like „better“, „worse“) among the values but not a uniform distance.</td>
<td>Differences and proportions can be determined. Values can be discrete or continuous.</td>
</tr>
<tr>
<td><strong>Examples:</strong></td>
<td><strong>Examples:</strong></td>
<td><strong>Examples:</strong></td>
</tr>
<tr>
<td>Gender (dichotome)</td>
<td>Quality grade (A/B/C)</td>
<td>Weight (continuous)</td>
</tr>
<tr>
<td>Eye/Hair Color</td>
<td>Age class (child, teen, adult, senior)</td>
<td>Number of sales (discrete)</td>
</tr>
<tr>
<td>Healthy/sick (dichotome)</td>
<td>Questionnaire answer: (completely agree,…)</td>
<td>Age (contin. or discrete)</td>
</tr>
</tbody>
</table>
Similarity Queries

• Specify query-object $q \in \text{DB}$ and…
  
  – … search threshold-based ($\varepsilon$) for similar $o$. – Range-Query
  
  \[ \text{RQ}(q, \varepsilon) = \{ o \in \text{DB} \mid \delta(q, o) \leq \varepsilon \} \]

  – … search for the $k$ most similar objects – Nearest Neighbor

  \[ \text{NN}(q, k) \subseteq \text{DB} \text{ having at least } k \text{ objects, such that} \]

  \[ \forall o \in \text{NN}(q, k), p \in \text{DB-NN}(q, k) : \delta(q, o) < \delta(q, p) \]
Similarity of Objects

Euklidean distance (L$_2$):
\[ \delta_2 = ((p_1 - q_1)^2 + (p_2 - q_2)^2 + ...)^{1/2} \]

Manhattan-Distance (L$_1$):
\[ \delta_1 = |p_1 - q_1| + |p_2 - q_2| + ... \]

Maximum-Distance (L$_\infty$):
\[ \delta_\infty = \max \{|p_1 - q_1|, |p_2 - q_2|, ...\} \]

Most natural measure of Dissimilarity

The individual dissimilarities of the features are summed up

Only the dissimilarity of the least similar feature is taken into account

Generalization L$_p$-Distance:
\[ \delta_p = (|p_1 - q_1|^p + |p_2 - q_2|^p + ...)^{1/p} \]
Adaptable Similarity Measures

**Weighted Euclidean distance:**
\[ \delta = (w_1(p_1 - q_1)^2 + w_2(p_2 - q_2)^2 + ...)^{1/2} \]

**Quadratic form distance:**
\[ \delta = ((p - q) M (p - q)^T)^{1/2} \]

Often the features have (heavily) varying value ranges:
Example: Feature F_1 \in [0.01 .. 0.05]
Feature F_2 \in [3.1 .. 22.2]
We need a high weight for F_1
(otherwise \( \delta \) would ignore F_1)

Sometimes we need a common weighting of different features to capture dependencies,
e.g. in color histograms to take color similarities into account

Some methods do not work with distance measures (where \( =0 \) means equality) but with positive similarity measures (\( =1 \) means equality)
Data Mining Tasks

Most important data mining tasks based on feature vectors:

- Classification
- Regression
- Clustering
- Outlier Detection

\[
\begin{align*}
\text{Supervised Learning} & : \text{Learn rules to predict a previously identified feature} \\
\text{Unsupervised Learning, Exploratory Analysis} & : \text{Learn some regularity/rules}
\end{align*}
\]

But there is a plethora of methods and tasks not based on feature vectors but directly working on text, sets, graphs etc.
Task:
Learn from previously classified *training data* the *rules*, to predict the class of new objects just based on their properties (features)

The result feature (class variable) is nominal (*categorical*)
Application: Newborn Screening

Blood sample of a newborn → Mass spectrometry → Metabolite spectrum

14 analysed amino acids:
- alanine
- arginine
- argininosuccinate
- citrulline
- glutamate
- glycine
- methionine
- phenylalanine
- pyroglutamate
- serine
- tyrosine
- valine
- leuzine+isoleuzine
- ornitine

[Baumgartner et al., Bioinformatics 20(17), 2004]
Application: Newborn Screening

Result:
- New diagnostic test
- Glutamine is a marker which was previously unknown
Regression

task:
Similar as classification, but the result feature to be predicted or estimated, is *numeric*
Clustering means: Decompose a set of objects (a set of feature vectors) into subsets (called clusters), such that

- the similarity of objects of the same cluster is maximized
- the similarity of objects of different clusters is minimized

Motivation: Different clusters represent different classes of objects

In contrast to classification: Number and meaning of the classes is unknown.
Application: Generation of Thematic Maps

Image of earth surface in 5 different color spectra

Value in band 2

Pixel \((x_1, y_1)\)

Pixel \((x_2, y_2)\)

Value in band 1

Cluster-Analysis

Retransform into \(xy\)-Coordinates

Color coding of Cluster-membership
Application: Tissue Classification

- Black: Ventricle + Background
- Blue: Tissue 1
- Green: Tissue 2
- Red: Tissue 3
- Dark red: Big vessels

**Result:** Automatic classification of cerebral tissue possible with dynamic CT.

[Baumgartner et al.: J. Digital Imaging 18(3), 2005]
Application: Gene expression clustering

Genes and conditions are hierarchically clustered (dendrogram). Simultaneous row and column clustering is called co-clustering.
Goals of Clustering

Challenges:
• Clusters of varying size, form, and density
• Hierarchical clusters
• Noise and outliers

=> We need different clustering algorithms
K-Means

• Goal
  – Partitioning into $k$ clusters such that a cost function (to measure the quality) is minimized
  – $k$ is a parameter of the method (specified by user).

• Locally optimizing method
  – Choose $k$ initial cluster representatives
  – Optimize these representatives iteratively
  – Assign each object to its closest or most probable representative
  – Repeat optimization and assignment until no more change (convergence)

• Types of cluster representants
  – Center (mean, centroid) of each cluster $\rightarrow$ k-means clustering
  – Most central data object assigned to cluster (medoid) $\rightarrow$ k-medoid clustering
  – Probability distribution of the cluster $\rightarrow$ expectation maximization

[Duda, Hart: Pattern Classification and Scene Analysis, 1973]
**K-Means**

**Idea of the algorithm**

- Algorithmus starts e.g. with randomly chosen objects as initial cluster representatives (many other initialization methods have been proposed).
- The algorithm is composed from two alternating steps:
  - Assignment of each point to its closest representative point
  - Recomputation of the cluster representative (center of its objects)
- Repeat the alternating steps until no more change (convergence)
K-Means

Properties of the algorithm

• Fast convergence to a local minimum of the objective function (Variance of the clusters, averaged over all clusters and dimensions)

\[ J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_{i}^{(j)} - c_j||^2 \]

• It is easy to see that
  – Assignment of points to clusters minimizes the objective function.
  – Re-determination of cluster centers minimizes the objective function.

• Thus the objective function is monotonic and bounded.

• Typically a small number of iterations (3-50) needed.

• To find the global optimum is more difficult (NP-hard in general)
  – Typical heuristic: Multiple (e.g. 10) runs with different initialisations of the starting points
Mining Interaction Patterns of Brain Regions

Parcellation into 90 anatomical regions.

fMRI data: Time Series of 3d volume images of the brain.

Each person is represented by a multivariate times series with $d = 90$ dimensions.

[Plant, Wohlschläger, Zherdin: ICDM 2009]
Clustering Multivariate Time Series
...by Interaction Patterns

\[
\text{orange} = \text{blue} + 3 \times \text{green}
\]

\[
\text{orange} = \text{blue} - \text{green}
\]
Interaction-based Cluster Notion

Cluster:

• set of *linear models* representing the dependency of each single Y dimension w.r.t. other dimensions X

\[ Y = X\beta + \varepsilon \]

orange = blue + 3* green + \varepsilon
blue = …
green = …

• set of objects.

orange = blue – green + \varepsilon
blue = …
green = …
Model Finding

Set of *linear models* representing the dependency of each single Y dimension w.r.t. *other dimensions* X

\[
\begin{align*}
Y_1 &= X_1 \ast \beta + \varepsilon \\
Y_2 &= X_2 \ast \beta + \varepsilon \\
& \quad \cdots \\
Y_D &= X_D \ast \beta + \varepsilon
\end{align*}
\]

Can be straightforward solved by multidimensional linear regression

But which dimensions X should be applied?

Usually not all \(d\) dimensions…
Greedy Stepwise Regression Controlled by BIC

First concatenate all objects of the cluster,

then greedily add and remove dimensions evaluating intermediate results with Bayesian Information Criterion (BIC).

\[ Y = X\beta + \varepsilon \] and \[ \beta = (X^TX)^{-1}(X^TY) \]

\[ BIC(M) = -2L_n(\hat{\beta}, \hat{\sigma}_{ML}^2) + \log(n)(\text{dim } \beta + 1) \]

\[ L_n(\hat{\beta}, \hat{\sigma}_{ML}^2) = -\frac{n}{2} - \frac{n}{2} \log \hat{\sigma}_{ML}^2 - \frac{n}{2} \log(2\pi) \]

\[ \hat{\sigma}_{ML}^2 = \frac{1}{n} \| Y - X\beta \|^2 \]
Algorithm Interaction K-means (IKM)

1) **Initialization:** Random partitioning into K equally sized clusters
2) Iterate the following steps until convergence:

   **Assignment:** Assign each object to that cluster to which it has the smallest sum of errors over all $d$ dimensions

   **Update:** Apply greedy-stepwise regression with BIC to all clusters.

**Major differences to standard K-means:**

- Similarity measure is the sum of errors of an object w.r.t. a set of models

- Cluster representative is not an object but a set of models describing characteristic interaction patterns shared by the objects within the cluster.

**Inherited from K-means:** Efficiency due to fast convergence;
Further improvement by aggregative pre-computing;
Results: Interaction patterns of brain regions

- resulting from clustering fMRI data with IKM.
- study on Somatoform Pain Disorder (pain without any clinical cause).
- Task fMRI: while in scanner the persons have been exposed to painful stimuli.

Right Amygdala (green) is interacting with different regions in patients and controls:
- controls: sensory areas (temporal, auditory)
- patients: frontal control areas.
only useful for this special fMRI application?

- also effective on synthetic and publicly available benchmark data from various domains.

- in comparison to standard K-means (Naive) and the state-of-the-art approach: Statistical Features Clustering (SF) (Wang et al., ICDM 2007)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>RI</th>
<th>IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>IKM</td>
<td>0.99</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>0.49</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td>0.72</td>
<td>2.38</td>
</tr>
<tr>
<td>DS2</td>
<td>IKM</td>
<td>0.56</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>0.48</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td>0.49</td>
<td>0.98</td>
</tr>
<tr>
<td>DS3</td>
<td>IKM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>0.61</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td>0.49</td>
<td>0.95</td>
</tr>
<tr>
<td>DS4</td>
<td>IKM</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td>0.98</td>
<td>0.2</td>
</tr>
<tr>
<td>DS5</td>
<td>IKM</td>
<td>0.88</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>0.79</td>
<td>2.36</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td>0.83</td>
<td>2.04</td>
</tr>
</tbody>
</table>

- EEG data (UCI)
- motion streams
- language processing (UCI)
Further Benefits of IKM

• Robust against noise objects and noise dimensions,

\[ \text{IC} \]

(a) Noise-Objects

(b) Noise-Dimensions

• Scalable,

\[ \text{IC} \]

(a) #dimensions/time

(b) length/time

(c) #objects/time

• and does not require all objects having time series of equal length.
Goals of Clustering

Challenges:
- Clusters of varying size, form, and density
- Hierarchical clusters
- Noise and outliers
  ➜ We need different clustering algorithms

K-Means can handle compact, spherical clusters like in (1)

For clusters with arbitrary shape like (3) we need a different clustering notion:
  - Density-Based Clustering
Density-based Clustering with DBSCAN

Idea: Clusters are areas of high object density which are separated by areas of lower object density.

A density-based cluster is a maximal set of density-connected objects.

[Ester et al. KDD 1996]
DBSCAN - Example

Start cluster expansion with an arbitrary core object; add objects within $\varepsilon$-range into seedList;

**While** the seed list is not empty:

  - Remove top element; set its cluster Id;
  - If it is a core object: add objects within $\varepsilon$-range to seed list as well.
Understanding the connectome of the brain

Basic anatomy of the brain:

Grey Matter: neuronal cell bodies
White Matter: myelinated axons
The brain is a highly efficient network!

But what are the nodes or functional units?

And what are the edges or major highways?

Why is this important to know?
• surgery planning (epilepsy, tumor),
• understanding brain development during adolescence and normal aging,
• understanding the onset and progression of neurodegenerative diseases like Alzheimer.

[Shao et al., ICDM Workshop 2010]
Visualizing the White Matter by diffusion tensor imaging (DTI)

Basic Principle

• movement of water molecules is restricted by white matter;
• in magnetic field moving molecules emit radiofrequency signals;
• DTI measures strength and direction of movement with 2 magnetic pulses coming from a specific direction called gradient: the first pulse labels the molecules, the second pulse reads out the displacement in a voxel in the gradient direction.

• Different gradient images need to be combined to capture the 3-d diffusion, 55 on our experimental data.
(1) **Combination**: Motion correction, co-registration

(2) **Fiber Tracking**

**Runge Kutta Method (4th order):**
- requires pre-defined seed and end region
- a fiber is modeled as a 3-d discrete curve which is drawn step by step
- select the next voxel by solving an ordinary differential equation involving the leading Eigenvector of the ellipsoid, the start and the end point

55 gradient images

Diffusion tensor field
256x256x70 -> 4,578,520 voxels

Thousands of fibers
What are the major highways?

More than 1,000 fibers only for the Corpus Callosum

Hundreds of thousands fibers in the brain

-> Fiber Clustering – suitable to deal with noise!
-> We need an effective and efficient similarity measure!
Evaluating similarity by 3-d fiber warping

Strength of DTW:
Optimal local alignment of timeseries to capture local similarity

Extending DTW to 3 dimensions:

\[ d(p_i, q_j) = |p_i^1 - q_j^1| + |p_i^2 - q_j^2| + |p_i^3 - q_j^3| \]

- Optimal Warping Path is determined using Quadratic programming as for DTW
- Avoiding that the fiber length overly dominates the similarity:
  Averaging all point-to-point distances along the optimal warping path.
Experiments – Similarity Measure

6 seed regions
For fiber tracking in the internal and external Capsules and the Corpus Callosum

Fibers grouped by medical experts into the corresponding 6 bundles and 3 outlying fibers

DTW
perfect

MPC
Outliers Assigned to cluster
(mean of closest pairwise distances)

Haussdorff
10 clusters
(maximum of longest pair-wise distances)
Results

Effective detection of clusters of different size and separation of noise \(\rightarrow\) DBSCAN is good!

Data Set 2: 973 fibers
What have we learned?

• Data Mining (Knowledge Discovery in Databases, KDD) is a central technology to cope with Big Data.
• Feature vectors are the most common objects used in data mining
• We distinguish between two philosophies
  – Supervised (attribute to be predicted is known)
  – Unsupervised (exploratory data analysis)
• Clustering is an unsupervised technique to group objects
  – Maximize intra-cluster similarity
  – Minimize between-cluster similarity
• There exists a large number of approaches with different properties:
  – Partitioning clustering like K-Means (spherical clusters)
  – Density-based clustering like DBSCAN (arbitrary shapes)