"An Introduction to Quality Assessment in Data Mining"

– Tutorial –

Maria Halkidi Michalis Vazirgiannis

Dept. of Informatics
Athens Univ. of Economics & Business,
Email: {mhalk, mvazirg}@aueb.gr
http://www.db-net.aueb.gr

OUTLINE

- Introduction to Quality Assessment in DM
  - Data Preprocessing & Quality Assessment
- Classification Interestingness measures
- Cluster Validity
- Association Rules Interestingness Measures

Introduction

- A data mining system could generate under different conditions thousands or million of patterns. Then questions arise for their quality:
  - **which of the extracted patterns are interesting ?**
  - **which of them represent knowledge?**

- A pattern is interesting:
  - if it is easily understood, valid, potentially useful and novel.
  - if it validates a hypothesis that a user seeks to confirm.

- An interesting pattern represents knowledge.

- The quality of patterns depends on:
  - the quality of the analysed data and
  - the quality of data mining results.

Introduction - Quality in Data Mining

- The **Quality in Data Mining** corresponds to
  - the representation of the knowledge included in the analysed data
  - Algorithms tuning - Selection of a suitable algorithm for a specific data analysis task
  - Selection of the most interesting patterns from the set of extracted patterns or the patterns that best fits the analysed data.

What is Quality for the DM tasks?

- **Quality in Classification**
  - Ability of the designed classification model to correctly classify new data samples.
  - Ability of an algorithm to define classification models with high accuracy
  - Interestingness of the patterns extracted during the classification process
- **Quality in Clustering**
  - How well the defined clustering scheme fits our data set
  - The number of groups into which the analysed data can be partitioned
- **Quality in Association Rules**
  - Interestingness of the extracted rules
  - The proportion of data that the extracted rules represent

Introduction

The role of Data Preprocessing in Quality Assessment

- **Data pre-processing** is a major step in the whole KDD process
- **Data pre-processing** techniques applied prior to data mining step could help to improve the quality of analysed data and consequently of the data mining results.

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Major Tasks in Data Preprocessing

- **Data cleaning**, which can be applied to remove noise and correct inconsistencies in the data.
- **Data transformation**. It could be applied to improve the accuracy and efficiency of mining algorithms involving distance measurements.
- **Data reduction**. It is applied to reduce the data size by aggregating, eliminating redundant features.

Quality Assessment Methods for DM Tasks

- The number of patterns generated during the data mining process is very large but only few of these patterns are likely to be of any interest to the domain expert analyzing the data.
- Many of the patterns are either irrelevant or obvious and not provide new knowledge.
- Patterns in data can be represented in many different forms including: classification rules, association rules, clusters.
- Techniques for evaluating the relevance and usefulness of discovered patterns are required.
- These techniques are broadly referred to as
  - Interestingness measures in case of classification or association rules applications
  - Cluster validity indices (or measures) in case of clustering.

Quality Assessment Methods for DM Tasks

- Classification
  - Classifiers accuracy techniques and related measures
    - Accuracy is one of the most important and widely used quality criteria in the classification process
    - Accuracy

Evaluation of a classifier(classification model)

Comparison of different classification algorithms

  - Classification Rules Interestingness Measures
    - Evaluation of the classification results

Quality Assessment Methods for DM Tasks

- Clustering
  - Cluster Validity approaches
    - Evaluation of clustering results
    - Selection of the partitioning that best fits the considered data
- Association Rules
  - Association Rules Interestingness Measures
    - Selection of interesting rules, rules that are representative of a data set.

Classification Quality Assessment

Classification

- The goal in classification process is to induce a model that can be used to classify future data items whose classification is unknown
- Classification is based on:
  - A well-defined set of classes and
  - A training set of pre-classified examples.
- The **knowledge** produced during the classification process can be extracted and represented in the form of rules.

Evaluation of Classification methods

- **Classification methods** can be compared and evaluated based on the following criteria:
  - Classification model accuracy: The ability of the classification model to correctly predict the class into which new or previously unseen data are classified.
  - Speed: It refers to the computation costs in building and using the model.
  - Robustness: The ability of the model to handle noisy or data with missing values and make correct predictions.
  - Scalability: The method ability to construct the classification model efficiently given large amounts of data.
  - Interpretability: It refers to the level of understanding that the constructed model provides.

Classification Model Accuracy

- The accuracy of a classification model (classifier)
  - allows one to evaluate how accurately the designed model will classify future data ?
  - helps to the comparison of different classifiers

Techniques for assessing classifier's accuracy

- Hold-out method.
  - The data set is randomly partitioned into a training set and a test set.
  - The training data are used to define the classification model
  - Its accuracy is estimated based on the test data.
- k-fold cross-validation.
  - The initial data are portioned into k subsets, “folds”.
  - Training and testing are iteratively performed k times.
  - The accuracy is estimated as

$$\sum_{i=1}^{k} \mathrm{num}(\mathrm{correct\_classified}_i) \Big/ \mathrm{total\_samples}$$

Techniques for assessing classifier's accuracy

- Bootstrapping.
  - It is k-fold cross validation with k set to the number of initial samples.
  - It samples the training instances uniformly with replacement and leave-one-out.
  - Let S ={S1,…, Sk}
  - For r=1,…,k
    - Define the training set, T, as the set of k-1 samples randomly selected from S
    - Train the classifier on T
    - Test the classifier on the remaining set.

Alternative to the accuracy measures (I)

- The estimation of an accuracy rate based on training data may mislead us about the quality of the derived classifier.
- **Why?**
  - Let a classifier, *f*, be trained to classify a set of data as "positive" or "negative".
  - A high accuracy could be result of the *f*'s ability to recognize negative samples.
  - It gives no indication about the ability of *f* to recognize positive and negative samples.
- Alternative to the accuracy measures have been proposed

Alternative to the accuracy measures (II)

- Sensitivity assesses how well the classifier can recognize positive samples

$$\boxed{\text{Sensitivity} = \frac{\text{true\_positive}}{\text{positive}}}$$

- Specificity measures how well the classifier can recognize negative samples.

$$\boxed{\text{Specificity} = \frac{\text{true\_negative}}{\text{negative}}}$$

Alternative to the accuracy measures (III)

- Precision assesses the percentage of samples classified as positive that are actually positive

$$\mathrm{Precision} = \frac{\mathrm{true\_positive}}{(\mathrm{false\_positive} + \mathrm{true\_positive})}$$

- Accuracy can be defined as a function of sensitivity and specificity

$$\mathrm{Accuracy} = \mathrm{Sensitivity} \cdot \frac{\mathrm{positive}}{\mathrm{positive} + \mathrm{negative}} + \mathrm{Specificity} \cdot \frac{\mathrm{negative}}{\mathrm{positive} + \mathrm{negative}}$$

Comparison of classification algorithms

"Given two classification algorithms A and B and a dataset S which algorithm will produce more accurate classifiers when trained on the same dataset ?"

Approaches based on statistical tests have been proposed to answer the above question.

McNemar's test (I)

- Let S the available set of data, which is divided into a training set R, and a test set T.
- Let two algorithms A and B trained on the training set and the result is the definition of two classifiers f_A and f_B.
- We test these classifiers on T
  - for each example $x \in T$ we record how it was classified and construct the following contingency table:

Number of examples misclassified by both classifiers (n_{00}).	Number of examples misclassified by f_A but not by f_B (n_{01})
Number of examples misclassified by f_B but not by f_A(n_{10})	Number of examples misclassified neither by f_A nor by f_B(n_{11})

- The two algorithms should have the same error rate under the null hypothesis.

McNemar's test (II)

- McNemar's test is based on a χ^2 test for good-ness-of-fit
- It compares the distribution of counts expected under null hypothesis to the observed counts. The expected counts under the null hypothesis are:

n_{00}	$(n_{01+} n_{10})/2$
$(n_{01+} n_{10})/2$	n_{11}

- We consider the following statistic $$s = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

- If the null hypothesis,Ho, is correct, then

  $P\ (s > x^2_{1,\,0.95}\ ) < 0.05.$

If $|s| > x^2_{1,\,0.95}$
reject Ho

➔ the two algorithms have different performance

A test for the difference of two proportions (I)

- A statistical test that is based on
  - measuring the difference between the error rate of algorithm A and the error rate of algorithm B
- **$p_A = (n_{00} + n_{01})/n$** → proportion of test examples incorrectly classified by algorithm A and
- **$p_B = (n_{00} + n_{10})/n$** → proportion of test examples incorrectly classified by algorithm B.
- **Assumption:**
  - when algorithm A classifies an example x ∈ T, the probability of misclassification is p_A.
  - the number of misclassifications of *n* test examples is a binomial random variable with
    - mean np_A and
    - variance $p_A(1-p_A)n$.

A test for the difference of two proportions (II)

- if p_A and p_B are independent then
  - p_A-p_B can be viewed as normally distributed
- Under the null hypothesis, Ho, it will have a mean of zero and a standard deviation error of

$$se = \sqrt{\frac{2p(1-\frac{p_A+p_B}{2})}{n}}$$

- Based on the above analysis, we obtain the statistic

$$z = \frac{p_A - p_B}{\sqrt{2p(1-p)/n}}$$

which has a standard normal distribution.

- if $|z|>Z_{0.975}$ =1.96 then
  - ✓Ho is rejected

A test for the difference of two proportions (III)

- **Drawbacks**
  - The probabilities p_A and p_B are measured on the same test set and thus they are not independent.
  - The test does not measure variation due to the choice of the training set or the internal variation of the learning algorithm.
  - it measures the performance of the algorithms on training sets of size significantly smaller than the whole data set.

The resampled paired *t* test (I)

- The test conducts a series of 30 trials, Tr_i.
- ∀ Tr_i, i=1,…30
  - the available sample S is randomly divided into a training set R and a test set T.
- The algorithms A and B are both trained on R and the resulting classifiers are tested on T.
- Let p^i_A and p^i_B → observed proportion of test examples misclassified by algorithm A and B respectively during the *i*th trial.

The resampled paired *t* test (II)

- Let the 30 differences

$$p^{(i)} = p_A^{(i)} - p_B^{(i)}$$

be drawn independently from a normal distribution.

- Then we can apply Student's *t* test by computing the statistic

$$t = \frac{\bar{p} \cdot \sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}\left(p^{(i)} - \bar{p}\right)^2}{n-1}}} \quad \text{where} \quad \bar{p} = \frac{1}{n}\sum_{i=1}^{n} p^{(i)}$$

- Under Ho, the statistic *t* has a *t* distribution with n-1 degrees of freedom.
- Then for 30 trials: if $|t| > t_{29,\,0.975} = 2.045$.

  Ho could be rejected

The resampled paired *t* test (III)

- Drawbacks
  - Since p_A and p_B are not independent, the differences $p^{(i)}$ will not have a normal distribution.
  - The $p^{(i)}$s are not independent, because the test and training sets in the trials overlap.

Interestingness Measures of Classification Rules

- Different approaches may result in different sets of patterns (classification rules).
- It is important to evaluate the discovered patterns identifying these ones that are valid and provide new knowledge.
- Techniques that aim at this goal are broadly referred to as interestingness measures.
- The interestingness of the patterns that discovered by a classification approach could also be considered as another quality criterion.

Rule-Interest Function (Piatetsky-Shapiro)

- It is used to quantify the correlation between attributes in a classification rule.
- It is suitable only for the single classification rules
- Let a rule X→ Y, the rule-interest function is given by the equation:

$$\boxed{RI = |X \cap Y| - \frac{|X||Y|}{\mathrm{N}}}$$

where :

- ✓ N is the total number of data points (or tuples of a database),
- ✓ |X| and |Y| are the number of tuples satisfying conditions X and Y respectively.
- ✓ $|X \cap Y|$ is the number of tuples satisfying X →Y and
- ✓ |X||Y|/N is the number of tuples expected if X and Y were independent (i.e., not associated).

Rule-Interest Function (Piatetsky-Shapiro)

- Depending on the values of RI we could evaluate the usefulness and interestingness of the extracted classification rules. Thus, if:
  - RI=0, then X and Y are statistically independent and the rule is not interesting.
  - RI>0 (RI<0), then X is positively (negatively) correlated to Y. The significance of the correlation between X and Y can be determined using chi-square test.

- Those rules, which do not exceed a pre-determined minimum significance threshold, are determined to be the most interesting.

Smyth and Goodman's J-Measure

- The ***J-measure*** is a measure for probabilistic classification rules and
- It is used to find the best rules relating discrete-valued attributes.
- A probabilistic classification rule is a logical implication

  X → **Y** with some probability p,

  - the left- and right-hand sides correspond to a single attribute.
  - The right-hand side is restricted to simple single-valued assignment expression while the left-hand-side may be a conjunction of simple expressions.

Smyth and Goodman's J-Measure

- The J-measure is given by the equation

$$J(X;Y) = p(Y)\left[p(X/Y)\log\left(\frac{p(X/Y)}{p(X)}\right) + (1 - p(X/Y))\log\left(\frac{1-p(X/Y)}{1-p(X)}\right)\right]$$

where p(Y), p(X) and p(X/Y) are the probabilities of occurrence of Y, X and X given Y, respectively.

- High values of J(X;Y) are desirable, but are not necessarily associated with the best rule.
- Why?
  - ✓ Rare conditions may be associated with the highest values for J(X; Y) but the resulting rule is insufficiently general to provide any new information.
- Further analysis is required in which the accuracy of a rule is traded for some level of generality or goodness-of-fit.

Gago and Bento's Distance Metric

- It measures the distance between classification rules
- It determines the rules that provide the highest coverage for the given data.

$$D(R_i,R_j)=\begin{cases}\dfrac{DA(R_i,R_j)+2DV(R_i,R_j)-2EV(R_i,R_j)}{N(R_i)+N(R_j)}, & \mathrm{NO}(\mathrm{R}_i,\mathrm{R}_j)=0\\ 2, & \text{otherwise}\end{cases}$$

- **DA(R_i, R_j)** → number of attributes in R_i and not in R_j + the number of attributes in R_j not in R_i,
- **DV(R_i, R_j)** → number of attributes in R_i and R_j that have slightly overlapping values (overlap < 66%),
- **EV(R_i, R_j)** → number of attributes in R_i and R_j that have overlapping values (overlap >66%)
- **N(R_i), N(R_j)** → number of attributes in R_i and R_j, respectively and
- **NO(R_i, R_j)** → number of attributes in R_i and R_j with non-overlapping values.

- $-1 <= D(R_i, R_j) <= 1$ or $D(R_i, R_j) = 2$.
- The rules with the highest average distance to the other rules are considered to be most interesting.

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

Cluster Validity Approaches

What is Clustering?

- Clustering aims at
  - grouping a set of data objects into clusters
  - identifying interesting distributions and patterns in underlying data

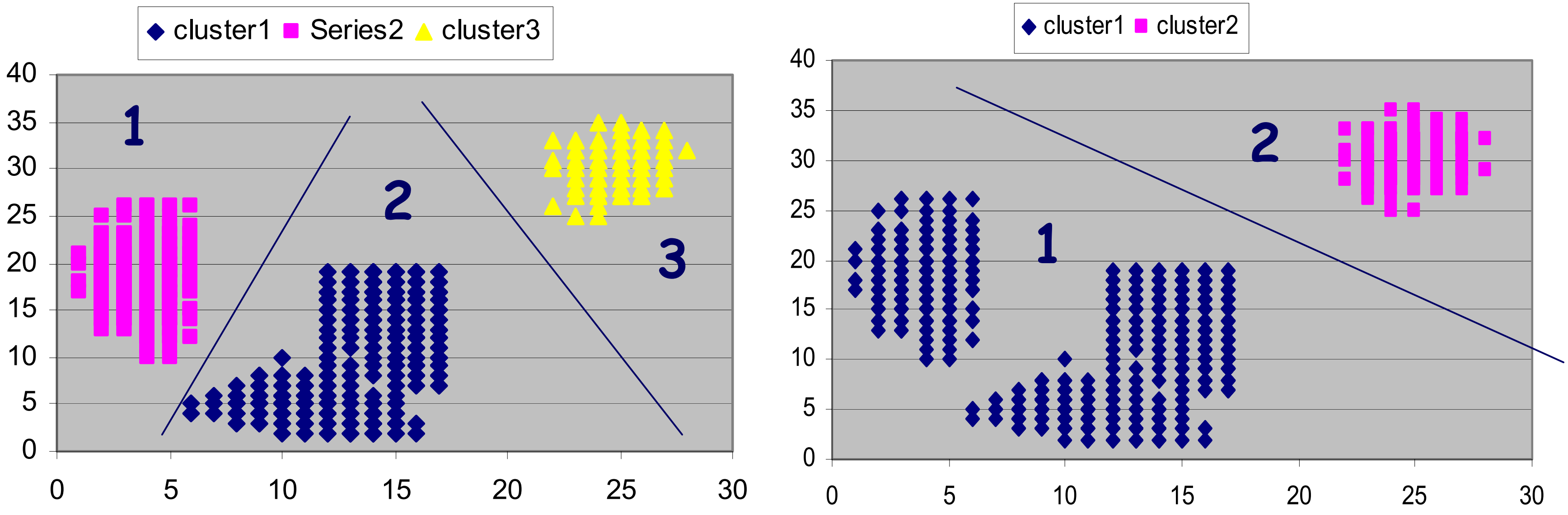
- Clustering is perceived as an unsupervised learning procedure
  - *no predefined classes* and *no examples* that indicates desirable relations among the data

Cluster Validity- Problem Specification

A problem we face in clustering is to decide the optimal number of clusters that fits a data set.

The various clustering algorithms behave in a different way depending on:

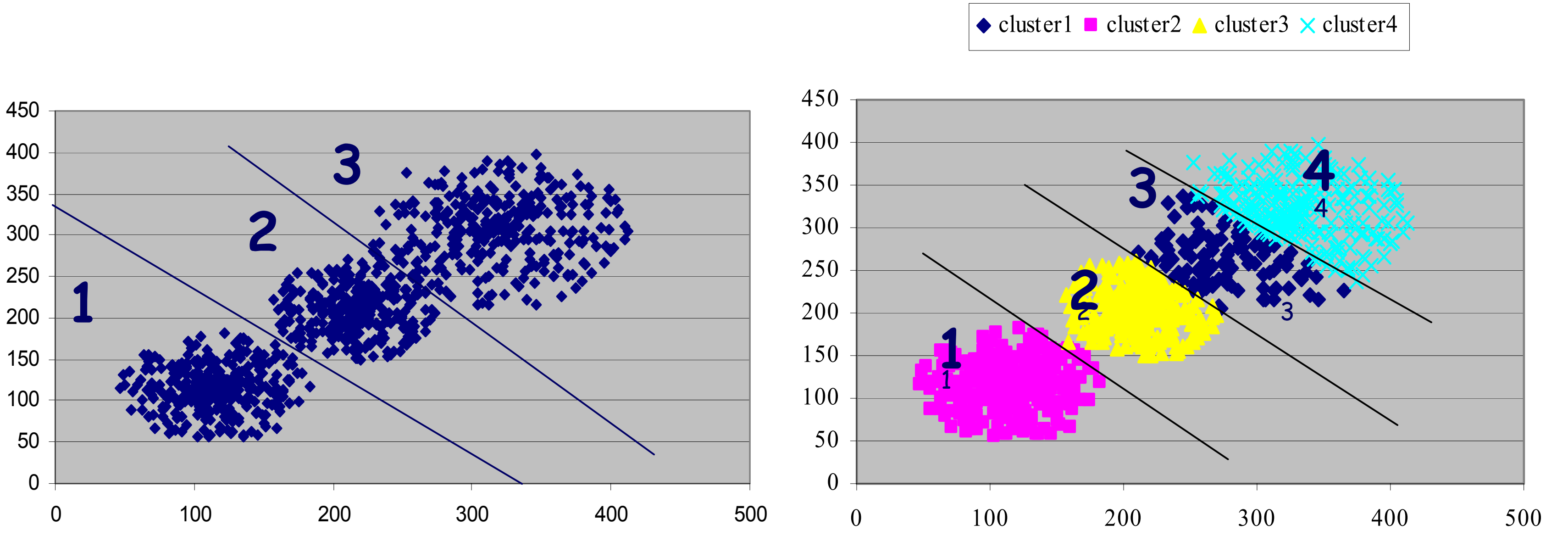
- the features of the data set (geometry and density distribution of clusters)
- the input parameters values



Eps=2, Nps=4

Eps=6, Nps=4

DBSCAN

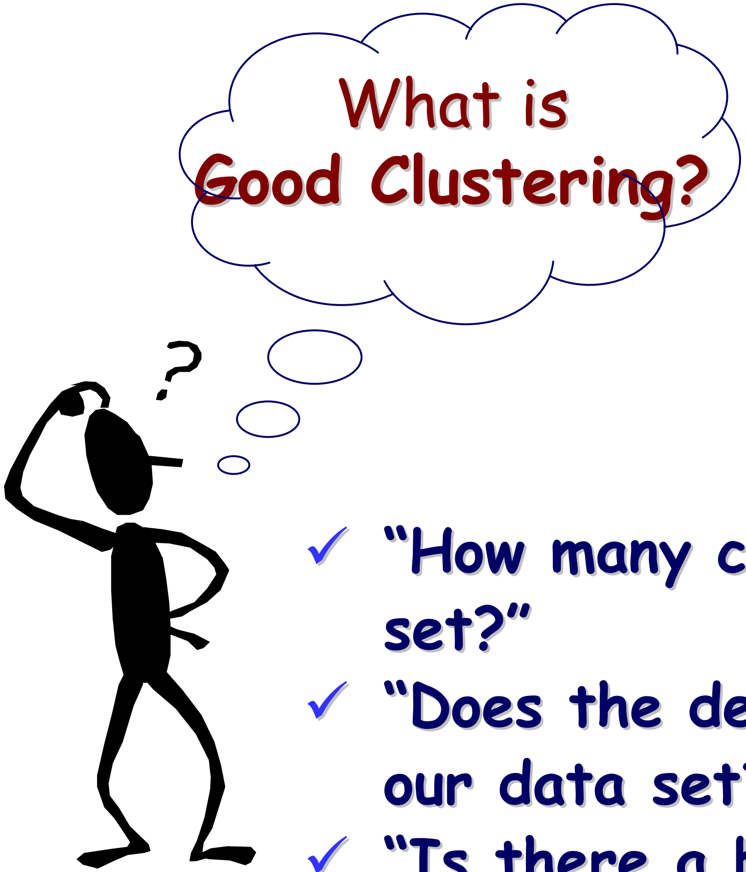


(a)

(b)

K-Means

The Cluster Validity Problem



What is
Good Clustering?

- ✓ “How many clusters are there in the data set?”
- ✓ “Does the defined clustering scheme fits our data set?”
- ✓ “Is there a better clustering possible?

Fundamental concepts of cluster validity

The procedure of evaluating the results of a clustering algorithm is known under the term **cluster validity.**

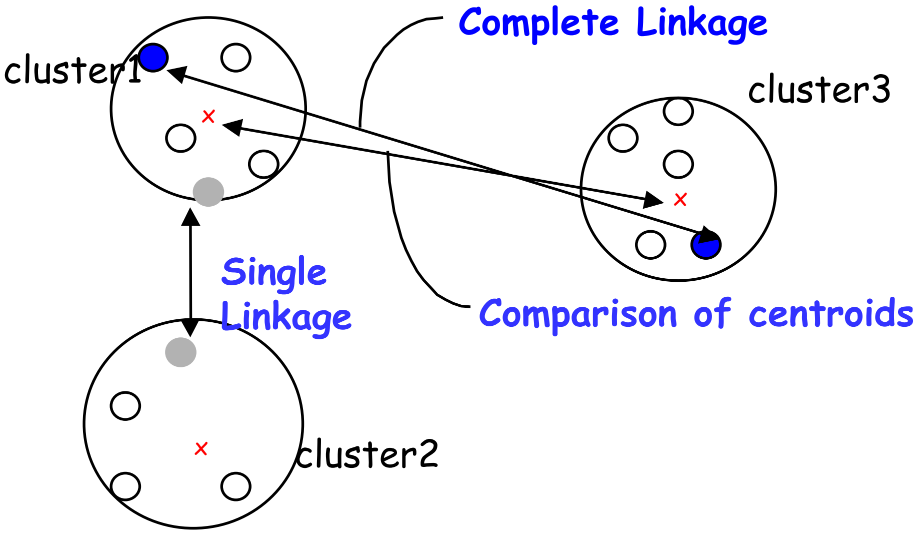
Three approaches to investigate cluster validity :

- **External criteria.** The results of a clustering algorithm are evaluated based on a pre-specified structure, which reflects our intuition about the clustering structure of the data set.
- **Internal criteria.** The results of a clustering algorithm are evaluated in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix).
- **Relative criteria.** The basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values.

Clustering Quality Criteria

- high intra-cluster similarity
  - Variance

- low inter-cluster similarity
  - Single Linkage
  - Complete Linkage
  - Comparison of centroids



Cluster validity approaches

External and Internal approaches

- ✓ based on statistical tests → high computational cost
- ✓ indices related to these approaches aim at measuring the degree to which a data set confirms an a-priori specified scheme.

Relative approaches

- ✓ finding the best clustering scheme that a clustering algorithm
  can define under certain assumptions and parameters.

External Criteria

The basic idea is to test whether the points of the data set are randomly structured or not.

This analysis is based on the ***Null Hypothesis, Ho***, expressed as a statement of random structure of a dataset.

Based on the external criteria we can work in two different ways:

- **Comparison of clustering structure C with partitioning *P***
- **Comparison of proximity matrix P with partitioning *P***

Comparison of *C* with partition *P*

Consider $\mathcal{C}$ ={C_1… C_m} is a clustering structure of a data set X and $\mathcal{P}$ ={P_1… P_s} is a defined partition of the data.

We refer to a pair of points ($\mathbf{x}_v$, $\mathbf{x}_u$) from the data set using the following terms:

- **SS**: if both points belong to the same cluster of the clustering structure **C** and to the same group of partition *P*.
- **SD**: if points belong to the same cluster of **C** and to different groups of *P*.
- **DS**: if points belong to different clusters of **C** and to the same group of *P*.
- **DD**: if both points belong to different clusters of **C** and to different groups of *P*.

Comparison of *C* with partition *P*

We can define the following indices to measure the degree of similarity between *C* and *P*:

- **Rand Statistic:** R = (a + d) / M,
- **Jaccard Coefficient:** J = a / (a + b + c),

✓ **a**, **b**, **c** and **d** are the number of SS, SD, DS and DD pairs respectively

✓ **a + b + c + d = M** which is the maximum number of all pairs in the data set,

✓ **M=N(N-1)/2** where N is the total number of pairs of points in the data set

External Validity Indices

- **Folkes and Mallows index:**

$$FM = a / \sqrt{m_1 m_2} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

where $m_1 = a / (a + b)$, $m_2 = a / (a + c)$.

- **Huberts Γ statistic:**

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j)\, Y(i,j)$$

high values of indices indicate great similarity between ***C*** and ***P***

- **Normalized Γ statistic**

$$\bar{\Gamma} = [(1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X(i,j) - \mu_X)(Y(i,j) - \mu_Y)] \Big/ \sigma_X \sigma_Y$$

where μ_x, μ_y, σ_x, σ_y are the respective means and variances of X, Y matrices.

Evaluation Procedure based on External Criteria

Let a data set X and $\mathcal{C}=\{C_1\ldots C_m\}$ be a clustering structure of X as defined by a clustering algorithm.

$\mathcal{P}=\{P_1\ldots P_s\}$ is a defined partition of the data, where $m \neq s$.

- **For i = 1 to r**
  - Generate a data set X_i with N vectors (points) in the area of X.
  - Assign each vector $y_{j,i}$ of X_i to the group that $x_j \in X$ belongs, according to the partition $\mathcal{P}$.
  - Run the same clustering algorithm used to produce structure $\mathcal{C}$, for each X_i, and let $\mathcal{C}_i$ the resulting clustering structure.
  - Compute $q(\mathcal{C}_i)$ value of the defined index q for $\mathcal{P}$ and $\mathcal{C}_i$.

  End For

- **Create** the plot of the r validity index values, $q(\mathcal{C}_i)$ (that computed into the for loop).
  - ✓Compare validity index value, let q, to the $q(\mathcal{C}_i)$ values, let q_i.
  - ✓The indices **R, J, FM, Γ** defined previously are used as the q index mentioned in the above procedure.

Comparison of P (proximity matrix) with partition *P*

Partition *P* can be considered as a mapping

$$g: X \rightarrow \{1 \ldots n_c\}.$$

Assuming matrix

$$Y: Y(i, j) = \begin{cases} 1, \text{ if } g(x_i) \neq g(x_j) \text{ and} \\ 0, \text{ otherwise}\}, \ i, j = 1 \ldots N, \end{cases}$$

- We compute Γ (or normalized Γ) statistic using the proximity matrix P and the matrix Y.

Index value → an indication of the two matrices' similarity.

- To proceed with the evaluation procedure we use the Monte Carlo techniques as mentioned above.

"*Generate*" → step of the procedure we generate the corresponding mappings g_i for every generated X_i data set.
"*Compute*" → step we compute the matrix Y_i, for each X_i in order to find the Γ_i corresponding statistic index.

Internal Criteria

We evaluate the clustering result of an algorithm using only quantities and features inherent to the dataset.

Two cases to which we apply internal criteria of cluster validity depending on the clustering structure:

a) **hierarchy of clustering schemes, and**

b) **single clustering scheme.**

Validating hierarchy of clustering schemes

Cophenetic matrix, P_c → represent the hierarchy diagram that produced by a hierarchical algorithm

Cophenetic Correlation Coefficient: A statistical index to measure the degree of similarity between P_c and P (proximity matrix)

$$\mathrm{CPCC} = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}\, c_{ij} - \mu_P \mu_C}{\sqrt{\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2\right]\left[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2\right]}}, \quad -1 \le \mathrm{CPCC} \le 1$$

where M=N·(N-1)/2 and N is the number of points in a dataset

$$\mu_P = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j),\ \mu_C = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P_c(i,j)$$

Validating a single clustering scheme

The goal here is to find the degree of agreement between

- a given clustering scheme **C**, consisting of **k** clusters, and
- the proximity matrix P.

The defined indices for this approach are

- ✓ **Hubert's Γ statistic**

  or

- ✓ **normalized Γ statistic.**

To compute the indices we use:

- A matrix is defined as

$$Y(i, j) = \begin{cases} 1\text{ , if } x_i \text{ and } x_j \text{ belong to different clusters, } i, j = 1,\dots, N. \\ 0\text{ , otherwise} \end{cases}$$

- Monte Carlo techniques is the way to test the random hypothesis in a given data set

Relative Criteria

The fundamental idea is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion.

The problem can be stated as follows:

> **"Let P the set of parameters associated with a specific clustering algorithm (e.g. the number of clusters nc). Among the clustering schemes C_i, i=1,..,n_c, defined by a specific algorithm, for different values of the parameters in P, choose the one that best fits the data set."**

Relative Criteria (continue…)

There are two approaches for defining the best clustering depending on the behaviour of q with respect to n_c.

The validity index

- ✓ does not exhibit an increasing or decreasing trend as the number of clusters increases

↓

we seek the maximum (minimum) of index in its plot with respect to n_c

- ✓ increase (or decrease) as the number of clusters increases

↓

we search for the values of n_c at which a significant local change in value of the index occurs.

Note: the absence of a knee may be an indication that the data set possesses no clustering structure.

Cluster Validity Indices
--Crisp clustering --

- **The modified Hubert Γ statistic**

$$\Gamma = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j)\cdot Q(i,j)$$

where

- M=N(N-1)/2,
- P is the proximity matrix of the data set and
- Q is an NXN matrix whose (i, j) element is equal to the distance between the representative points (v_{ci}, v_{cj})of the clusters where the objects x_i and x_j belong.

In the plot of normalized Γ versus n_c, the number of clusters at which a significant increase of normalized Γ occurs

Indication of the number of clusters that underlie the data

Cluster Validity Indices --Crisp clustering --

- Dunn index

$$D_{nc} = \min_{i=1,\ldots,nc}\left\{\min_{j=i+1,\ldots,nc}\left(\frac{d(c_i, c_j)}{\max_{k=1,\ldots,nc} diam(c_k)}\right)\right\}$$

✓ the dissimilarity function between two clusters c_i and c_j defined as

$$d(c_i, c_j) = \min_{x\in c_i,\, y\in c_j} d(\mathbf{x}, \mathbf{y}),$$

✓ diam(c) is the diameter of a cluster , which may be considered as a measure of dispersion of the clusters.

$$diam(C) = \max_{x,\, y\in C} d(x, y)$$

Cluster Validity Indices
--Crisp clustering --

Best clustering scheme → $d(c_i, c_j)$ ↑ & diam(c) ↓

The maximum in the plot of D_{nc} versus the number of clusters can be an indication of the number of clusters that fits the data.

The implications of the Dunn index are:

- the considerable amount of time required for its computation,
- the sensitive to the presence of noise in datasets, since noise is likely to increase the values of diam(c)

Cluster Validity Indices
--Crisp clustering --

➢The Davies-Bouldin (DB) index

A similarity measure R_{ij} between the clusters C_i and C_j is defined based on

- ✓a measure of dispersion of a cluster C_i and
- ✓a dissimilarity measure between two clusters d_{ij}.

The R_{ij} index is defined to satisfy the following conditions:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- if $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
- if $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
- if $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$.

A simple choice for R_{ij} that satisfies the above conditions is:

$$R_{ij} = (s_i + s_j)/d_{ij}$$

Cluster Validity Indices
--Crisp clustering --

The DB index is defined as

$$DB_{n_c} = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i$$

$$R_i = \max_{\substack{i=1,\ldots,n_c \\ i \neq j}} R_{ij}, \; i = 1,\ldots, n_c$$

☑ DB_{nc} is the average similarity between each cluster c_i, i=1, …, nc and its most similar one.

☑ It is desirable for the clusters to have the minimum possible similarity to each other

+

☑ The DB_{nc} index exhibits no trends with respect to the number of clusters

we seek the minimum value of DB_{nc} in its plot versus the number of clusters.

Cluster Validity Indices
--Crisp clustering --

- RMSSDT, SPR, RS, CD (*Hierarchical Clustering Algorithms*)

These four indices can be applied to each step of a *hierarchical* clustering algorithm and they are known as:

- Root-mean-square standard deviation (RMSSTD) of the new cluster
- Semi-partial R-squared (SPR)
- R-squared (RS)
- Distance between two clusters (CD)

They have to be used simultaneously to determine the number of clusters existing in our data set.

Cluster Validity Indices
--Crisp clustering --

- RMSSDT (*Hierarchical Clustering Algorithms*)

RMSSTD is the square root of the attributes variances used in the clustering process.

- ✘ It measures the homogeneity of the formed clusters at each step of the hierarchical algorithm.
- ✘ The RMSSTD of a cluster should be as small as possible.
- ✘ If the values of RMSSTD are higher at one step than the ones of the previous step, we have an indication that the new clustering scheme is not homogenous.

Cluster Validity Indices
--Crisp clustering --

- SPR (*Hierarchical Clustering Algorithms*)

We define the term *Sum of Squares* as

$$SS = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Also we use the following symbolisms:

- ✳ SS_w referring to the within cluster sum of squares,
- ✳ SS_b referring to the between clusters sum of squares
- ✳ SS_t referring to the total sum of squares, of the whole data set.

***SPR* = (SS_w of the new cluster - the sum of SS_w of clusters joined to obtain the new cluster) / SS_t for the whole data set.**

This index measures the loss of homogeneity after merging the two clusters of a single algorithm step.

- ✓ SPR= 0 → the new cluster is obtained by merging two perfectly homogeneous clusters.
- ✓ SPR > 0 → the new cluster is obtained by merging two heterogeneous clusters.

Cluster Validity Indices
--Crisp clustering --

- RS (Hierarchical Clustering Algorithms)

$$RS = SS_b / SS_t.$$

❖ SS_b is a measure of difference between groups.

$$SS_t = SS_b + SS_w$$

❖ RS may be considered as a measure of
  - the degree of difference between clusters
  - the degree of homogeneity between clusters.

✓ RS = 0 → **no difference exists among clusters**

✓ RS = 1 → **there is significant difference among clusters.**

Cluster Validity Indices
--Crisp clustering --

- **CD (Hierarchical Clustering Algorithms)**

The *CD* index measures the distance between the two clusters that are merged in a given step.

- **Centroid hierarchical clustering**
  - CD is the distance between the centers of the clusters.
- **Single linkage**
  - CD measures the minimum Euclidean distance between all possible pairs of points
- **Complete linkage**
  - CD is the maximum Euclidean distance between all pairs of data points.

Cluster Validity Indices
--Crisp clustering --

- RMSSTD & RS (Non- Hierarchical Clustering Algorithms)

$$RMSSTD = \left[\frac{\sum\limits_{\substack{i=1\ldots nc \\ j=1\ldots v}} \sum\limits_{k=1}^{n_{ij}} \left(x_k - \overline{x}_k\right)^2}{\sum\limits_{\substack{i=1\ldots nc \\ j=1\ldots v}} \left(n_{ij} - 1\right)} \right]^{\frac{1}{2}}$$

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t}$$

Find the optimal Clustering

✓Run the algorithm a number of times for different number of clusters each time.

✓Plot the respective graphs of the validity indices vs number of clusters

✓Search for the significant "knee" in these graphs.

Optimal clustering for our data set → number of clusters at which the "knee" is observed

SD Validity Index
--Crisp clustering --

- SD Index

★ Variance of data set

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \overline{x}^p\right)^2$$

, where $\overline{X} = \frac{1}{\mathrm{n}}\sum_{k=1}^{n} x_k, \forall x_k \in X$

★ Variance of cluster i.

$$\sigma_{v_i}^p = \frac{1}{n_i}\sum_{k=1}^{n}\left(x_k^p - v_i^p\right)^2$$

★ Average scattering for clusters.

$$Scat(c) = \frac{\frac{1}{c}\sum_{i=1}^{c}\left\|\sigma(v_i)\right\|}{\left\|\sigma(X)\right\|}$$

★ Total separation between clusters.

$$Dis(c) = \frac{D_{\max}}{D_{\min}}\sum_{k=1}^{c}\left(\sum_{z=1}^{c}\left\|v_k - v_z\right\|\right)^{-1}$$

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

SD Index Definition

$$\boxed{SD(c) = a \cdot Scat(c) + Dis(c)}$$

$a = Dis(c_{max})$, where c_{max} is the maximum number of input clusters.

Scat ↘ & Dis ↘ ⟷ Optimal Clustering

- *SD* proposes an optimal number of clusters almost irrespectively of c_{max}.
- *SD* handle properly convex clusters. The same applies to all the aforementioned indices.

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

S_Dbw: A validity index based on Scattering and Density between clusters

Objective : Definition of a relative algorithm-independent validity index, for assessing the quality of partitioning for each set of the input values.

Main features of the proposed approach

Validity index S_Dbw. Based on the features of the clusters:

- ✓ evaluates the resulting clustering schemes as defined by the algorithm under consideration.
- ✓ selects for each algorithm the optimal set of input parameters with regards to the specific data set.

S_Dbw Definition

Let **D={v$_i$| i=1,..., c}** a partitioning of a data set S into *c* clusters where v$_i$ is the center of i cluster as it results from applying a clustering algorithm *alg$_j$* to S.

Let **stdev** the average standard deviation of clusters defined as:

$$\text{stdev} = \frac{1}{c}\sqrt{\sum_{i=1}^{c}\left\|\sigma(v_i)\right\|}$$

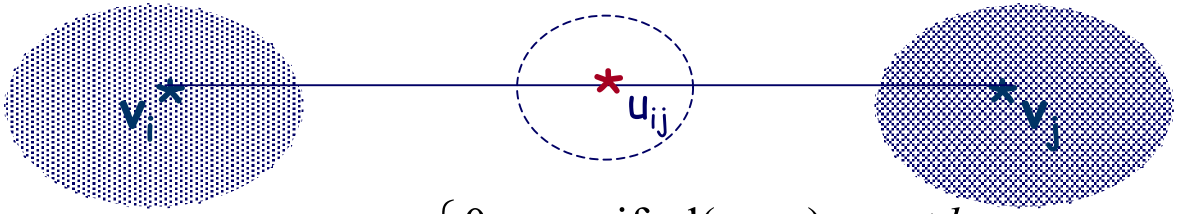
S_Dbw definition: Inter-cluster Density (ID).

- It evaluates the average density in the region among clusters in relation with the density of the clusters

$$\boxed{Dens\_bw(c)=\frac{1}{c\bullet(c-1)}\sum_{i=1}^{c}\left(\sum_{\substack{j=1\\ i\neq j}}^{c}\frac{density(u_{ij})}{\max\{density(v_i),density(v_j)\}}\right),}$$

$$density(u\ )=\sum_{l=1}^{n_{ij}} f(x_l,u\ ),$$

where n_{ij} = number of tuples that belong to the clusters $\mathrm{c_i}$ and $\mathrm{c_j}$, $i.e., x_l\in c_i\cup c_j\subseteq S$



$$f(x,u)=\begin{cases}0, & \text{if d(x, u)} > stdev\\ 1, & \text{otherwise}\end{cases}$$

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

S_Dbw definition: Intra-cluster variance

- Average scattering for clusters

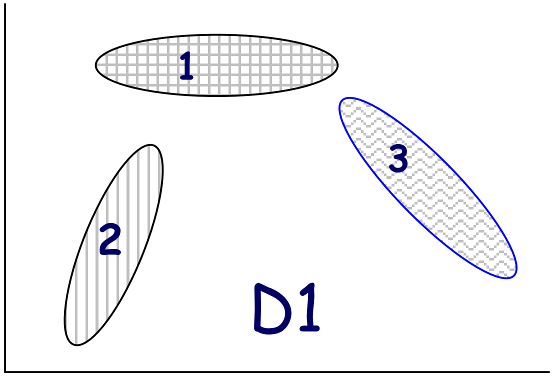
$$Scat(c) = \frac{\frac{1}{c}\sum_{i=1}^{c}\left\|\sigma(v_i)\right\|}{\left\|\sigma(X)\right\|}$$

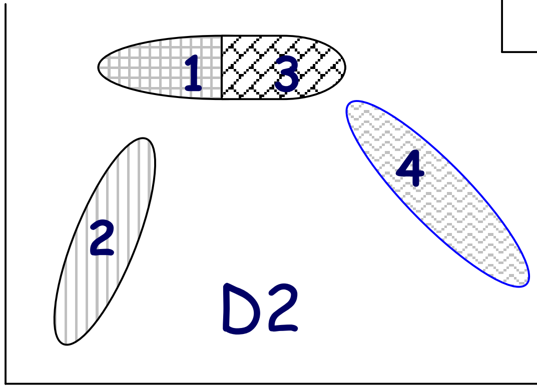
where $$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \overline{x}^p\right)^2$$

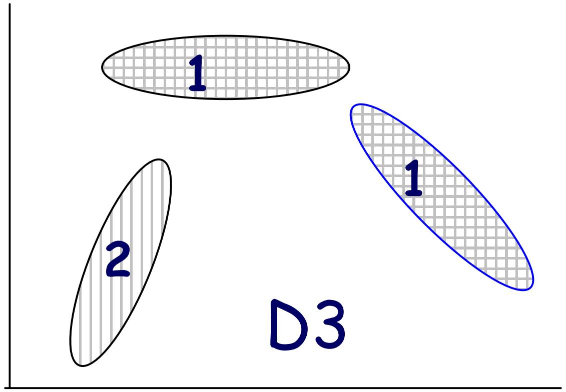
where $\overline{x}^p$ is the p $_{th}$ dimension of $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall x_k \in X$

$$\sigma_{v_i}^p = \sum_{k=1}^{n_i}\left(x_k^p - v_i^p\right)^2 \Big/ n_i$$

S_Dbw(c) = Scat(c) + Dens_bw(c)

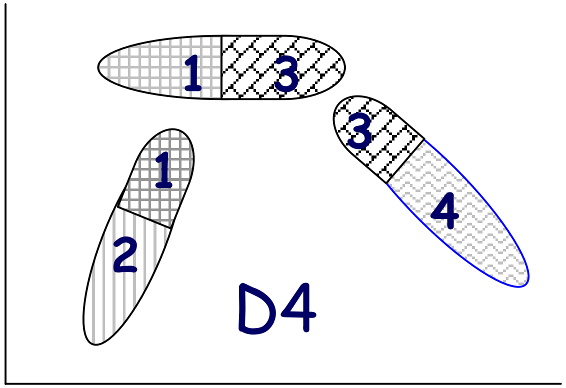


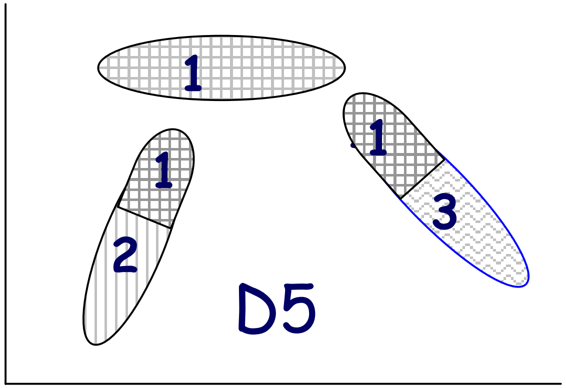




Scat ↑ & Dens_bw ~

Scat ~↘ & Dens_bw ↑





Scat ↑ & Dens_bw ↑

Scat ↑ & Dens_bw ↑

A new cluster validity Index – CDbw approach

Objective: Definition of a relative algorithm-independent validity index, for assessing the quality of partitioning for each set of the input values.

Main features of the proposed approach:

- evaluates the resulting clustering schemes as defined by the algorithm under consideration.
- selects for each algorithm the optimal set of input parameters with regards to the specific data set.

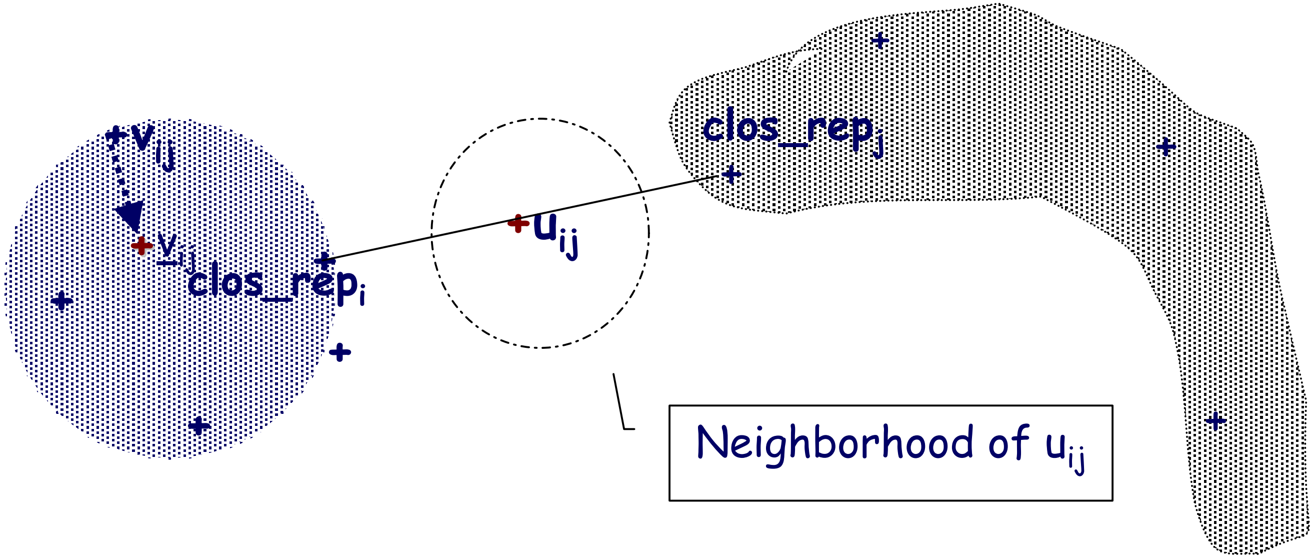
CDbw: Compose Density between and within clusters

CDbw is formalized based on:

- multi-representative points to represent the clusters defined by an algorithm. The result is a better description of the clusters' structure than this achieved by others approaches, which consider a single center point.
- clusters' compactness (in terms of intra-cluster density), and
- clusters' separation (combining the distance between clusters and the inter-cluster density).

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

Validity Index Definition



Let **D={V$_1$,…, V$_c$}** a partitioning of a data set S into *c* convex clusters

- V_i= {v_{i1},…, v_{ir} | r=number of representatives per cluster}
- v_{ij} is the jth representative of cluster i as it results from applying a clustering algorithm to S.

$$\boxed{CDbw(c) = Intra\_dens(c) \cdot Sep(c)}$$

Cbw Definition: Inter-cluster Density

It evaluates the average density in the region among clusters. The goal is the density in the area among clusters to be significant low.

$$\mathrm{Inter\_dens}(c) = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ i \neq j}}^{c} \left(\frac{d(\mathrm{clos\_rep}_i, \mathrm{clos\_rep}_j)}{\mathrm{stdev}_i + \mathrm{stdev}_j} \cdot \mathrm{density}(u_{ij}) \right),$$

$$c > 1, c \neq n$$

where

$$\mathrm{density}(u_{ij}) = \frac{\sum_{l=1}^{n_i+n_j} f(x_l, u_{ij})}{n_i + n_j}, \qquad f(x, u_{ij}) = \begin{cases} 0, & \text{if } d(x, u_{ij}) > (\mathrm{stdev}_i + \mathrm{stdev}_j)/2 \\ 1, & \text{otherwise} \end{cases}$$

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

Cbw Definition: Clusters' Separation

- It evaluates the separation of clusters taking into account both the distances between the closest clusters and the Inter-cluster density.
- The goal is the distances among clusters to be high while the density in the area among them to be low.

$$\mathrm{Sep}(c) = \frac{\sum_{i=1}^{c} \min_{\substack{j=1,\ldots,c\\ i\neq j}} \{d(\mathrm{clos\_rep}_i, \mathrm{clos\_rep}_j)\}}{1 + \mathrm{Inter\_dens}(c)}, c > 1$$

Cbw Definition: Intra-cluster Density

- Shrinked representatives: Shrink the initial representatives towards the center of clusters, $\underline{v}_{ij}$.
- The average density within clusters is defined as the percentage of points that belong to the neighborhood of $\underline{v}_{ij}$.

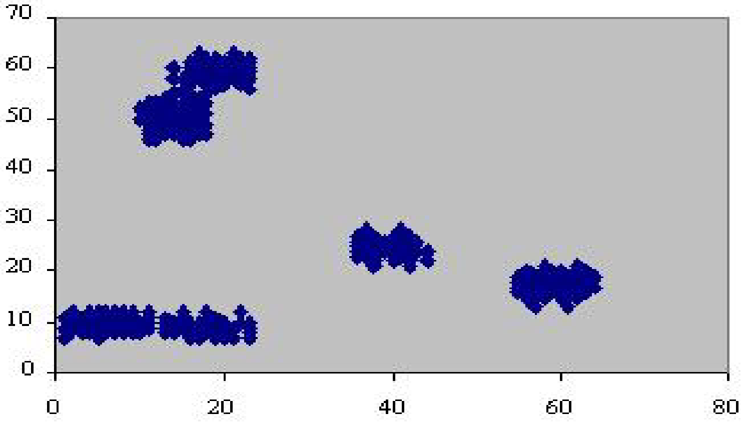
$$\text{Intra\_dens}(c) = \frac{1}{c \cdot r} \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\text{density}(\underline{v}_{ij})}{\text{stdev}}, \; c > 1,\; c \neq n$$

$$\text{density}(\underline{v}_{ij}) = \frac{\sum_{l=1}^{n_i} f(x_l, \underline{v}_{ij})}{n_i}, \qquad f(x, \underline{v}_{ij}) = \begin{cases} 0, & \text{if } d(x, \underline{v}_{ij}) > \text{stdev} \\ 1, & \text{otherwise} \end{cases}$$

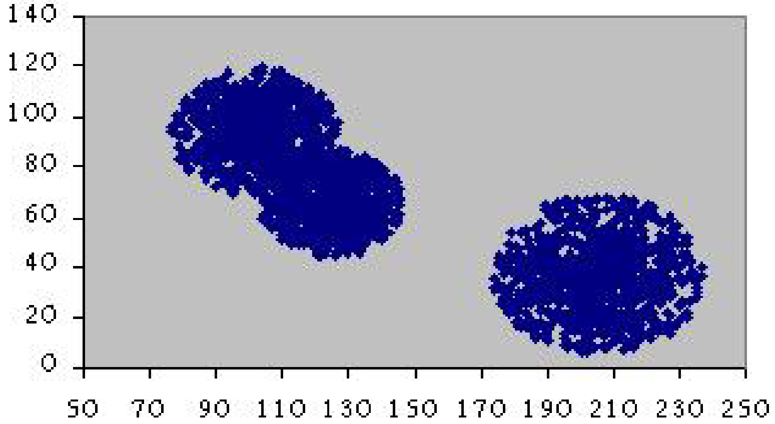
Experimental Study

Cluster Validity

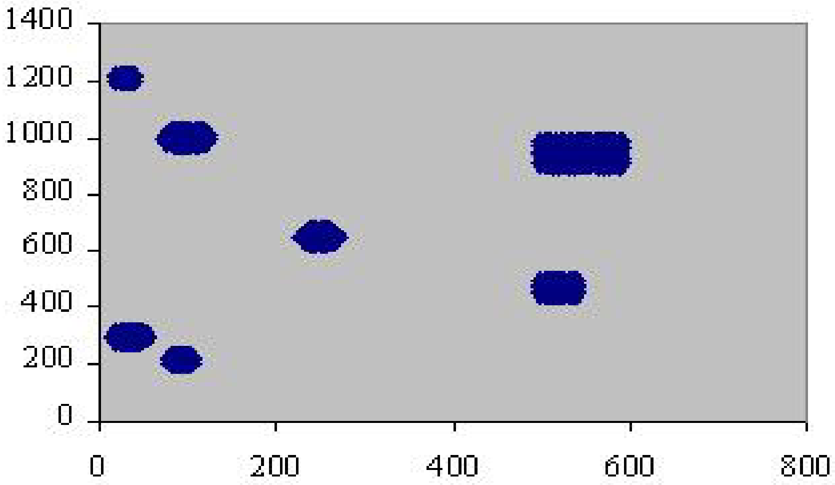
Comparison of Cluster validity Indices



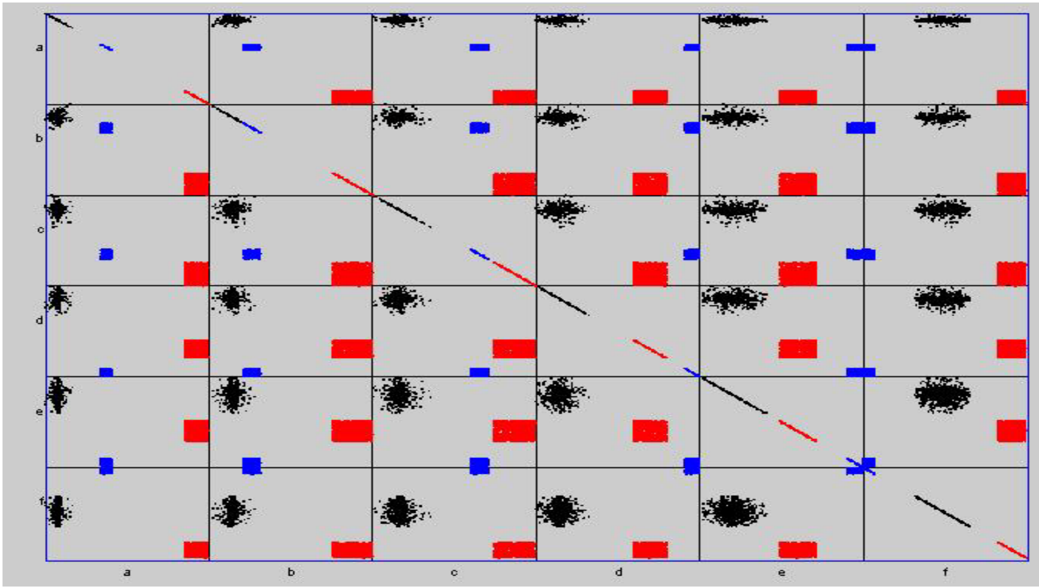
DataSet1



DataSet2



DataSet3



ND_Set

Optimal partitioning as proposed by validity indices

	DataSet1	DataSet2	DataSet3	Nd_Set
	Optimal number of clusters			
RS, RMSSTD	3	2	5	3
DB	6	3	7	3
SD	4	3	6	3
S_Dbw	4	2	7	3

Case Study - Cluster Analysis of Epidemiological Data

- The data are collected from the hospitals based on the daily isolations of the Microbiology laboratory.
- The data *used for analysis* refers to the resistances of Sau organism isolated from a hospital to a set of antibiotics.

Sau-AVM Descriptive Statistics

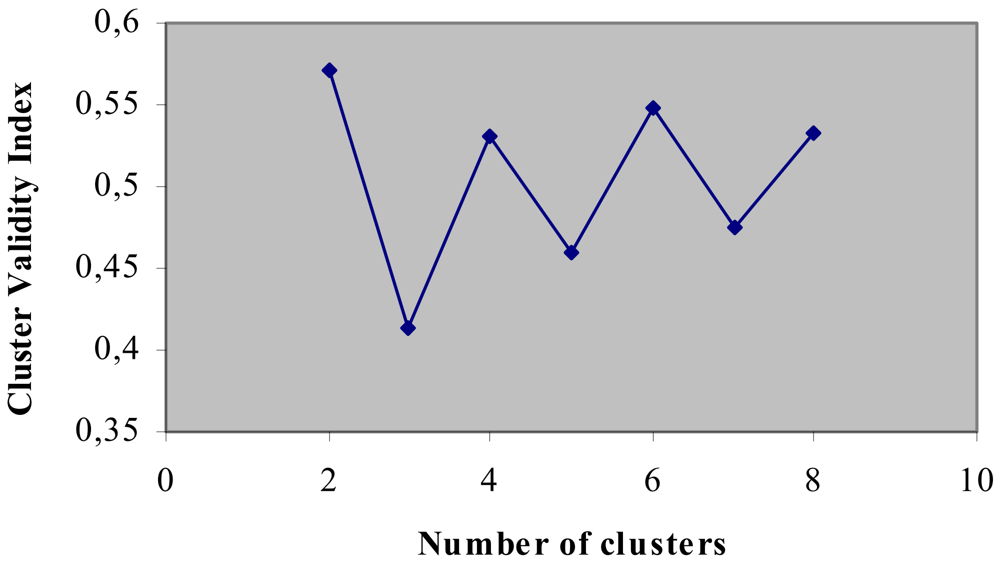
	N	Minimum	Maximum	Sum	Mean	Std. Deviation
VAN	907	11,00	28,00	17046,00	18,7938	1,9385
OXA	888	6,00	32,00	11671,00	13,1430	6,9921
GEN	898	6,00	35,00	14990,00	16,6927	8,0066
Valid N (listwise)	879					

Total Num of Rows: 908

Statistics of the Sau organism datasets with respect to VAN, OXA, GEN

Cluster Analysis of Epidemiological data

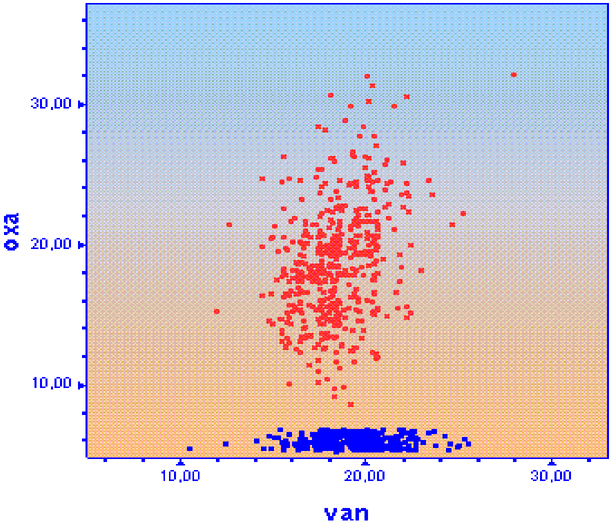
- Goal of study → Identify significant groups in the data regarding the Sau organisms resistance to VAN and OXA,
- The "Average Linkage" algorithm (hierarchical algorithm) is used to find partitions in the dataset.
- A dendrogram is defined, each level of which corresponds to a different partitioning of the dataset.
- **"Which of the defined partitioning fits the data?".**
- A cluster validity approach is adopted to evaluate the clustering algorithm results and select the one that best fits our data.
- Considering the results of clustering algorithm for 2 to 8 clusters. A set of seven different partitionings are defined. Then the value of cluster validity CDbw is calculated.



Resistance of Sau organisms to OXA and VAN

sau org clustering

VAN, OXA



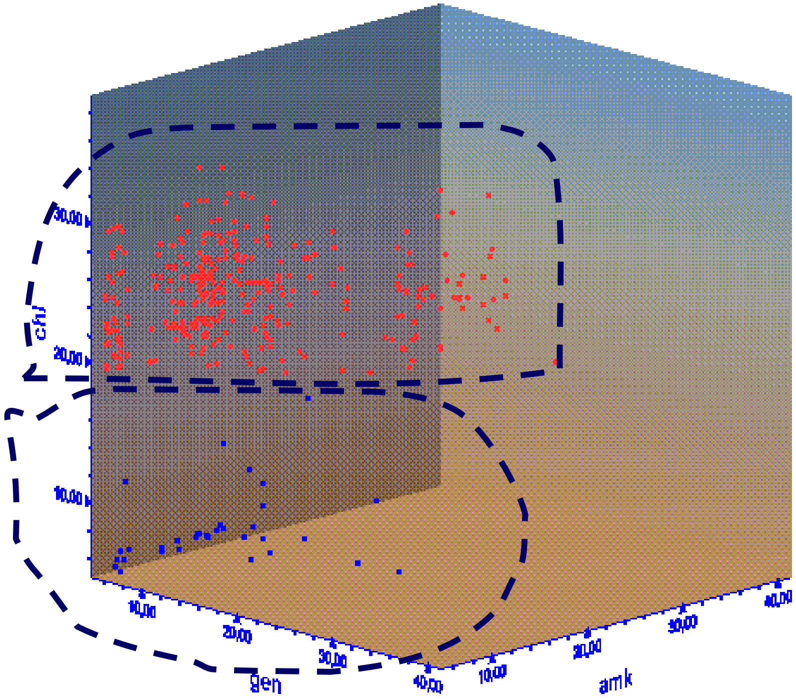
Cloud is jittered

Jittering adjusts the display of points that would fall exactly on top of one another if no adjustment were made. **Jitter all scale variables.** Adds a small amount of random noise to any scale data, whether or not the points are coincident. Categorical data are not jittered.

AVM Hospital

sau org clustering

GEN, AMK, CHL



Average Linkage (Between Groups)

- 🟥 1
- 🟦 2

Cloud is jittered

Partitioning of the resistances of Sau OXA resistant organisms to GEN, AMK and CHL into two clusters defined by Average Linkage

OXA resistant Sau orgs
AVM Hospital

	1			2			Total		
	Mean	N	Std.	Mean	N	Std.	Mean	N	Std.
GEN	10,4768	323	6,7102	8,4412	34	4,1208	10,2829	357	6,5313
AMK	15,2074	323	4,9208	14,0294	34	5,1315	15,0952	357	4,9460
CHL	24,3344	323	2,8956	7,1176	34	2,4217	22,6947	357	5,8087

Validity Indices for Fuzzy Clustering

Fuzzy Clustering Validity Indices

The objective is to seek clustering schemes where most of the vectors of the dataset exhibit high degree of membership in one cluster.

A **fuzzy clustering** is defined by

- ✓ a matrix $U=[u_{ij}]$, where u_{ij} denotes the degree of membership of the vector x_i in the j cluster.
- ✓ a set of the cluster representatives.

To evaluate clustering schemes

- we define **validity index**, q, and
- we plot the **q** versus **number of clusters**.

If **q** exhibits a trend with respect to **the number of clusters**, we seek a **significant knee** of decrease (or increase) in the plot of q.

Validity Indices involving only the membership values

- Partition coefficient

$$\boxed{PC = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{nc} u_{ij}^{2}}$$

The PC index values range in [1/nc, 1], where nc is the number of clusters.

- If PC →1 indicates crisp clustering
- If PC =1/nc indicates the fuzzy clustering or there is no clustering tendency in the considered dataset or the clustering algorithm failed to reveal it.

Validity Indices involving only the membership values

- Partition entropy

$$\boxed{PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{nc} u_{ij}\cdot\log_a\left(u_{ij}\right)}$$

The index is computed for $nc > 1$ and PE $\in [0, \log_a nc]$.

- If PE → 0, indicates crisp clustering
- If PE→ $\log_a nc$ indicates absence of any clustering structure in the dataset or inability of the algorithm to extract it

Validity Indices involving only the membership values

👎 Drawbacks

- ✓ their monotonous dependency on the number of clusters. Thus, we seek significant knees of increase (for PC) or decrease (for PE) in plot of the indices versus the number of clusters,
- ✓ their sensitivity to the fuzzifier, m. More specifically, as m→1 the indices give the same values for all values of nc. On the other hand when m→ ∞, both PC and PE exhibit significant knee at *nc=2*,
- ✓ the lack of direct connection to the geometry of the data [Dave96], since they do not use the data itself.

Indices involving the membership values and the dataset.

- Xie-Beni index

Let a fuzzy partition of the data set $X=\{x_j; j=1,.., n\}$ with $v_i(i=1,..., nc)$ the centers of each cluster and u_{ij} the membership of data point j belonging to cluster i.

Compactness of cluster i

$$\pi=(\sigma_i/n_i).$$

n_i : the number of point in cluster belonging to cluster i,
σ_i : variance of cluster i

Separation of the fuzzy partitions

$$d_{min} = \min||v_i - v_j||$$

$$\boxed{XB=\pi/N\cdot d_{min}}$$

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

Indices involving the membership values and the dataset.

- **Fukuyama-Sugeno index**

$$FS_m = \sum_{i=1}^{N} \sum_{j=1}^{n_c} u_{ij}^m \left(\left\| x_i - v_j \right\|_A^2 - \left\| v_j - v \right\|_A^2 \right)$$

  - v : the mean vector of X and
  - A :a positive symmetric matrix, when A=I, the above distance become the squared Euclidean distance.
  - small values for FS_m → compact and well-separated clusters

Indices involving the membership values and the dataset.

- Fuzzy covariance matrix of the j-th cluster

$$\Sigma_j = \frac{\sum_{i=1}^{N} u_{ij}^m (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^m}$$

- Fuzzy hyper volume of j-th cluster

$$V_j = |\Sigma_j|^{1/2}$$

- **Total fuzzy hyper volume**

$$FH = \sum_{j=1}^{nc} V_j$$

- **Average partition density**

$$PA = \frac{1}{nc} \sum_{j=1}^{nc} \frac{S_j}{V_j} \qquad where \qquad S_j = \sum_{x \in X_j} u_{ij}$$

M. Halkidi, M. Vazirgiannis, PKDD, August 2002

ASSOCIATION RULES

Interestingness Measures

Association Rules

- **Association rules** reveal underlying interactions between the attributes in the data set.
- These interactions can be presented in the form:

  A→ B

  where A, B refer to sets of attributes in underlying data.
- A and B are selected so as to be frequent item sets.
- a **frequent item set** is a set of attributes' values, which are found together in at least T records in a dataset (T is a user-defined threshold).

Coverage

- The coverage of an association rule is the proportion of cases in the data that have the attribute values or items specified on the Left Hand Side(LHS) of the rule.

$$\boxed{\text{Coverage} = n(LHS)/N = P(LHS)}$$

where N is the total number of cases under consideration.

- Coverage takes values in [0,1]
- if coverage → 1 then
  - the rule is considered as an important association rule.

Support

- The **support** of an association rule is the proportion of all cases in the dataset that satisfy a rule.

$$\boxed{\text{Support} = n(\text{LHS} \cap \text{RHS})/N}$$

- **Support** corresponds to the statistical significance of the rule
- a <u>high support</u> of the rule is an indication that a high number of tuples contains both LHS and RHS of this rule, i.e., the rule is representative of the considered data

Confidence

- The **confidence** of an association rule is the proportion of the cases covered by the LHS of the rule that are also covered by the RHS

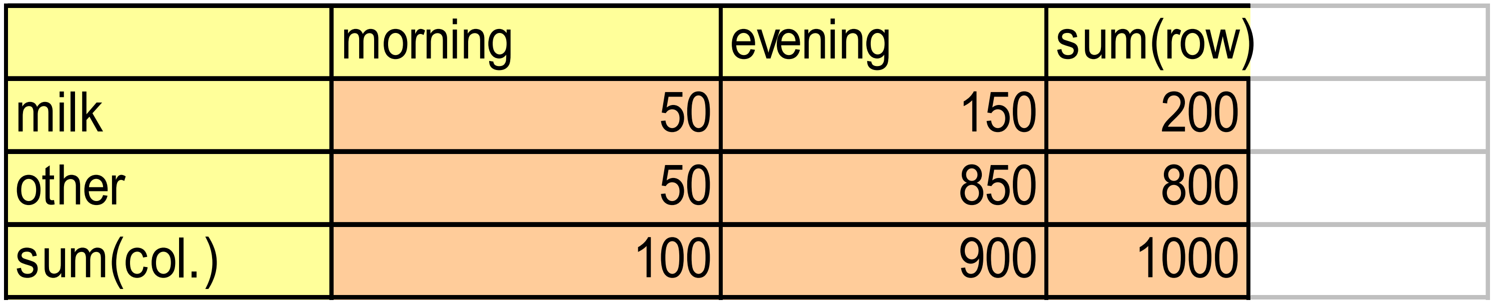
$$\boxed{\textbf{Confidence} = n(\text{RHS} \cap \text{LHS})/n(\text{LHS})}$$

where n(LHS) denotes the number of cases covered by LHS

- Confidence corresponds to the strength of a rule.
- It takes values in [0,1]
  - If confidence → 1
    - The rule is considered as important.

Example (I)

- Among 1000 transactions
  - 200 transactions contain milk,
  - 100 transactions take place early in the morning,
  - 50 transactions that contain milk took place early in the morning
- Let the rule
  - R: buy milk → morning
  - n(LHS) =200, n(RHS)=100, n(RHS∩LHS)=50



	morning	evening	sum(row)
milk	50	150	200
other	50	850	800
sum(col.)	100	900	1000

Example (II)

- **Coverage:** 200/1000 = 0.2.
- **Support:** 50/1000 = 0.05.
- **Confidence:** 50/200 = 0.25.

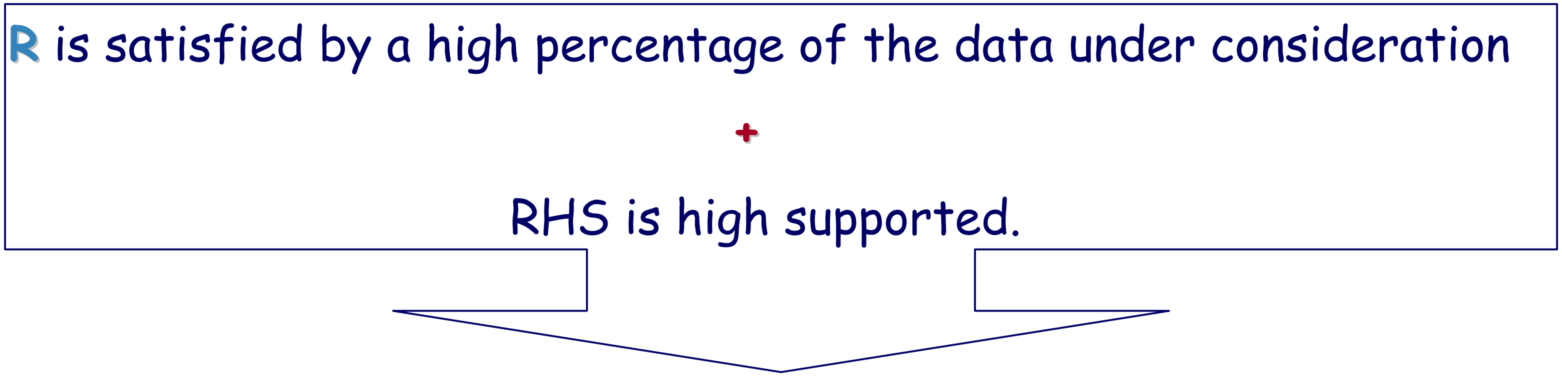
	morning	evening	sum(row)
milk	50	150	200
other	50	850	800
sum(col.)	100	900	1000

Criticism Confidence and Support (I)

- Let a rule **R : A+B → G** , **confidence = 85%** and **support (G) =90%.**
- Strength (R) is high → R is a significant rule.

However,

- **RHS (G)** represents the 90% of the studied data → a high proportion of the data contains G.
- there is a high probability **RHS (G)** to be satisfied by our data



R is satisfied by a high percentage of the data under consideration

+

RHS is high supported.

R may not make sense in making decisions or extracting general rule as regards the behaviour of the data.

Criticism to Support and Confidence (II)

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Rule	Support	Confidence
X=>Y	25%	50%
X=>Z	37,50%	75%

- Example:
  - support and confidence of X=>Z dominates
- We need a measure of events' dependence.
- Lift gives an indication of events correlation
  - X and Y: positively correlated (lift >1),
  - X and Z, negatively related (lift <1)

Itemset	**Support**	**lift**
X,Y	**25%**	**2**
X,Z	**37,50%**	**0,9**
Y,Z	**12,50%**	**0,57**

Lift (I)

The lift of an association rule is the confidence divided by the proportion of all cases that are covered by the RHS.

$$\boxed{\text{Lift} = \text{Confidence} / P(\text{RHS})}$$

- It is a measure of the importance of the association.
- As for the values of lift there are some conditions to be considered:
  - If lift → 1 then RHS and LHS are independent, which indicates that the rule is not important.
  - If lift → +∞ we have the following sub-cases:
    - If $RHS \subseteq LHS$ or $LHS \subseteq RHS$ then the rule is not important.
    - If $P(RHS) \rightarrow 0$ then the rule is not important.
    - If $P(RHS|LHS) \rightarrow 1$ then the rule is interesting.
  - If lift = 0 means that $P(RHS|LHS) = 0 \Leftrightarrow P(RHS \cap LHS) = 0$, which indicates that the rule is not important.

Lift (II)

- **Lift** gives an indication of rule significance, or how interesting is the rule.
  - It represents the predictive advantage a rule offers over simply guessing based on the frequency of the rule consequence (RHS).
  - It is an indication whether a rule could be considered as representative of the data so as to use it in the process of decision-making.

Leverage

The leverage of an association rule is the proportion of additional cases covered by both the LHS and RHS above those expected if the LHS and RHS were independent

$$\boxed{\text{Leverage} = P(RHS \mid LHS) - (P(LHS) * P(RHS))}$$

- Leverage takes values in [-1,1].
- if leverage <= 0, then
  - there is a strong independence between LHS and RHS.

  else if leverage → 1
  - indication of an important association rule

Example (III)

- **Lift:** p(RHS)=100/1000=0.1, Confidence =0.25
  - lift= 0.25/0.1 = 2.5.
- **Leverage**
  - P(LHS and RHS)= 50/1000 = 0.05.
  - The proportion of cases that would be expected to be covered by both LHS and RHS if LHS and RHS are independent is
    - P(LHS and RHS)=(200/1000) ∗ (100/1000) = 0.02.
  - The leverage = (0.05 - 0.02) = 0.03.

	morning	evening	sum(row)
milk	50	150	200
other	50	850	800
sum(col.)	100	900	1000

Gray and Orlowska's Interestingness

- Gray and Orlowska use the term interestingness to evaluate the strength of associations between sets of items in association rules.
- **Interestingness** contains a discriminator component that gives an indication of the independence of the antecedent (X) and consequent (Y). Interestingness is given by:

$$\boxed{I = \left(\left(\frac{P(X \cap Y)}{P(X) \times P(Y)} \right)^{k} - 1 \right) \times \left(P(X) \times P(Y) \right)^{m}}$$

- ✓ P(X∩Y) is the "confidence",
- ✓ P(X)×P(Y) is the "support",
- ✓ P(X∩Y) / P(X)×P(Y) is the discrimination,
- ✓ k and m are parameters to weight the relative importance of the discrimination and support components,

- Rules with higher values of interestingness are considered more interesting

Dong and Li's Interestingness

- **Interestingness** is used to evaluate the importance of an association rule by considering its unexpectedness in terms of other association rules in its neighbourhood.
- An *r-neighborhood* of a rule is given by the set:

$$N(R_0, r) = \{R \mid D(R, R_0) \le r, R \text{ a potential rule}\}$$

- The distance metric is given by the equation:

$$D(R_1, R_2) = \delta_1 \times \left|(X_1 \cup Y_1)\Theta(X_2 \cup Y_2)\right| + \delta_2 \times \left|X_1 \Theta X_2\right| + \delta_3 \times \left|Y_1 \Theta Y_2\right|$$

where $R_1=X_1 \rightarrow Y_1$, $R_2=X_2 \rightarrow Y_2$, δ_1, δ_2,δ_3 are parameters to weight the relative importance of all three terms, and Θ is an operator denoting the symmetric difference between X and Y (i.e. $(X-Y) \cup (Y-X)$).

Dong and Li's Interestingness

- Two types of interestingness are:
  - Unexpected confidence. It is given by the following equation:

$$UCI = \begin{cases} 1, & \text{if } \left\| c(R_0) - ac(R_0, r) \right| - \mathrm{sc}(\mathrm{R}_0, \mathrm{r}) \right| > \mathrm{t}_1 \\ 0, & \text{otherwise} \end{cases}$$

- ✓ $c(R_0)$ is the confidence of R_0,
- ✓ $ac(R_0, r)$ average confidence
- ✓ $sc(R_0, r)$ are the standard deviation of the rules in the set $M \cap N(R_0, r) - \{R_0\}$
- ✓ (M is the set of rules satisfying the minimum support and confidence),
- ✓ t_1 is a threshold.

Dong and Li's Interestingness

- Isolated confidence.

$$II = \begin{cases} 1, & \text{if } \left|N(R_0, r)\right| - \left|M \cap N(R_0, r)\right| > \mathrm{t}_2 \\ 0, & \text{otherwise} \end{cases}$$

✓ $|N(R_0, r)|$ is the number of potential rules in an r-neighborhood,

✓ $|M \cap N(R_0, r)|$ is the number of rules generated from the neighborhood, and t_2 is a threshold.

Peculiarity

- It is used to determine the extent to which one data object differs from other similar data objects

$$\boxed{PF(x_i) = \sum_{j-1}^{n} \sqrt{N(x_i, x_j)}}$$

where x_i and x_j are attributes values, n is the number of different attribute values and $N(x_i, x_j)$ is the conceptual distance between x_i and x_j. The conceptual difference is given by:

$$N(x_i, x_j) = \left| x_i - x_j \right|$$

Conclusions

- Data Mining is mainly concerned with methodologies for extracting patterns from large data repositories
- A data mining system could generate under different conditions thousands or million of patterns,
- One of the main questions that arises "Which of the extracted patterns are interesting and which of them represent knowledge?"
- A pattern is interesting if it is easily understood, valid, potentially useful and novel.
- The *interestingness* of patterns depends both on the quality of the analysed data and the quality of data mining results
- Several techniques have been developed aiming at evaluating and preparing the data used as input in data mining process

Conclusions

- **Data pre-processing techniques** applied prior to mining could help to improve the quality of data and consequently of the data mining results. The most common pre-processing techniques are: i) Data cleaning, ii) Data transformation, iii) Data reduction.
- **Classification approaches** can be **compared** and **evaluated** based on the following criteria: i) Classification model accuracy, ii) Speed, iii) Robustness, iv) Scalability, v) Interpretability

Conclusions

- The **accuracy of a classification model** designed according to a set of training data is one of the most important and widely used criteria in the classification process. The most common techniques for assessing classifier accuracy are: *i*) Hold-out method, ii) k-fold cross-validation, iii) bootstrapping
- Different classification methods may produce different classification models trained on the same data set.
- A number of methods have been proposed to compare classification algorithms with respect to the accuracy of the defined models: i) McNemar's test, ii) A test for the difference of two proportions, iii) The resampled paired t test

Conclusions

- The interestingness of the classification patterns could also be considered as another quality criterion. Techniques that aim at this goal are broadly referred to as **interestingness measures**.
- Some representative measures for ranking the usefulness and utility of discovered classification patterns (i.e., classification rules) are: i) Rule-Interest Function, ii) Smyth and Goodman's J-Measure, iii) Gago and Bento's Distance Metric.

Conclusions

- The various clustering algorithms behave in a different way depending on the features of the data set , the input parameters values.
- The procedure of evaluating the results of a clustering algorithm is known under the term cluster validity. There are three approaches to investigate cluster validity based on: i) external, ii) internal and iii) relative criteria.
- A number of cluster validity indices have been proposed for both crisp and fuzzy clustering

Conclusions

- The **interestingness measures of association rules** could give an indication of the rules' importance and confidence.
- Some of the most known association rules interestingness measures are: Support, Confidence, Coverage, Leverage and Lift.
- Other also well-known approaches and measures for evaluating association rules are: Dong and Li's Interestingness , Gray and Orlowska's Interestingness, Peculiarity.

Thank you for your attention !

db-net group
http://www.db-net.aueb.gr

References

- Roberto J. Bayardo Jr, Rakesh Agrawal, Dimitrios Gunopoulos. "Constraint-based Rule Mining in Large, Dense Databases". Proc. Of the 15th ICDE, 1999
- Michael J. A. Berry, Gordon Linoff . *Data Mining Techniques For marketing, Sales and Customer Support*. John Willey & Sons, Inc, 1996.
- Bezdeck, J.C, Ehrlich, R., Full, W.. "FCM:Fuzzy C-Means Algorithm", *Computers and Geoscience,* 1984.
- Dave, R. N. . "Validating fuzzy partitions obtained through c-shells clustering", *Pattern Recognition Letters*, Vol .10, pp613-623, 1996.
- Davies, DL, Bouldin, D.W. "A cluster separation measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 1, No2, 1979.
- Thomas G. Dietterich. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms", Neural Computation, 10(7), 1998.
- Dunn, J. C. . "Well separated clusters and optimal fuzzy partitions", *J. Cybern.* Vol.4, pp. 95-104, 1974.
- Ester, M., Kriegel, H-P., Sander, J., Xu, X.. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of 2nd Int. Conf. On Knowledge Discovery and Data Mining,* Portland, pp. 226-23, 1996

References

- Fayyad, M. U., Piatesky-Shapiro, G., Smuth P., Uthurusamy, R.. Advances in Knowledge Discovery and Data Mining. AAAI Press, 1996
- P. Gago, C. Bentos. “A metric for selection of the most promising rules”. Proceedings of the 2nd European Conference on The Principles of Data Mining and Knowledge Discovery (PKDD’98). Nantes, France, September 1998.
- Gath I., Geva A.B. “Unsupervised optimal fuzzy clustering”, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 11(7), 1989.
- B. Gray, M. E. Orlowka. “Ccaiia: clustering categorical attributes into interesting association rules”. In Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD ’98). Melbourne, Australia, April 1998.
- Guha, S, Rastogi, R., Shim K.. "ROCK: A Robust Clustering Algorithm for Categorical Attributes", Published in the Proceedings of the IEEE Conference on Data Engineering, 1999.
- M. Gupta, and T. Yamakawa, (eds). "Fuzzy Logic and Knowledge Based Systems", Decision and Control (North Holland). 1988.
- R. J. Hilderman, H. J. Hamilton. “ Knowledge Discovery and Interstigness Measures: A Survey”, Technical Report CS 99-04, Dept of Computer Scinece, University of Regina, October 1999.

References

- Han, J., Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.
- Halkidi, M., Vazirgiannis, M., Batistakis, I.. "Quality scheme assessment in the clustering process", Proceedings of PKDD, Lyon, France, 2000.
- Halkidi M, Vazirgiannis M., "A data set oriented approach for clustering algorithm selection", Proceedings of PKDD, Freiburg, Germany, 2001
- M. Halkidi, M. Vazirgiannis, "Clustering Validity Assessment: Finding the optimal partitioning of a data set", to appear in the Proceedings of ICDM, California, USA, November 2001.
- M. Halkidi, Y. Batistakis, M. Vazirgiannis. "On Clustering Validation Techniques", Intelligent Information Systems Journal, 2001, Kluwer Publishers.
- Jain, A.K., Murty, M.N., Flyn, P.J.. "Data Clustering: A Review", ACM Computing Surveys, Vol.31, No3, 1999.
- Krishnapuram, R., Frigui, H., Nasraoui. O. "Quadratic shell clustering algorithms and the detection of second-degree curves", Pattern Recognition Letters, Vol. 14(7), 1993
- M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, A. I. Verkamo. "Finding interesting rules from large sets of discovered association rules". In Proceedings of the 3rd International Conference on Information and Knowledge Management. Gaitersburg, Maryland, 1994.

References

- H. Liu, W.Hsu, S. Chen. “Using general impressions to analyze discovered classification rules”. In proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD’97). Newport Beach, California, August 1997.
- MacQueen, J.B. "Some Methods for Classification and Analysis of Multivariate Observations", In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp281-297, 1967.
- Milligan, G.W. and Cooper, M.C.. "An Examination of Procedures for Determining the Number of Clusters in a Data Set", Psychometrika, Vol.50, pp 159-179, 1985.
- Rezaee, R, Lelieveldt, B.P.F., Reiber, J.H.C. "A new cluster validity index for the fuzzy c-mean", Pattern Recognition Letters, 19, pp. 237-246, 1998.
- Theodoridis, S., Koutroubas, K.. Pattern recognition, Academic Press, 1999.
- Xie, X. L, Beni, G.. "A Validity measure for Fuzzy Clustering", IEEE Transactions on Pattern Analysis and machine Intelligence, Vol.13, No4, 1991.

References

- Pal, N.R., Biswas, J.. “Cluster Validation using graph theoretic concepts”. Pattern Recognition, Vol. 30(6), 1997.
- Rezaee, R, Lelieveldt, B.P.F., Reiber, J.H.C. "A new cluster validity index for the fuzzy c-mean", Pattern Recognition Letters, 19, pp. 237-246, 1998.
- Snedecor G. W., Cochran W. G. Statistical Methods. Iowa State University Press, Ames, IA, 8th Edition.
- P. Smyth, R.M. Goodman. “Rule induction using information theory”. In Knowledge Discovery in Databases, AAAI/MIT Press,1991.
- Sharma, S.C.. Applied Multivariate Techniques. John Willwy & Sons, 1996.
- G. Piatetsky-Shapiro. “Discovery, analysis and presentation of strong rules. In Knowledge Discovery Databases. AAAI/MIT Press, 1991
- Smyth, P. "Clustering using Monte Carlo Cross-Validation". Proceedings of KDD Conference, 1996.
- Theodoridis, S., Koutroubas, K.. Pattern recognition, Academic Press, 1999.
- N. Zhong, Y. Yao, S. Ohsuga. “Peculiarity-oriented multi-database mining”. In Proceedings of the 3rd European Conference on the Principles of Data Mining and Knowledge Discovery, Czech Republic, September 1999.