Automated data clustering

Guided Unsupervised Search

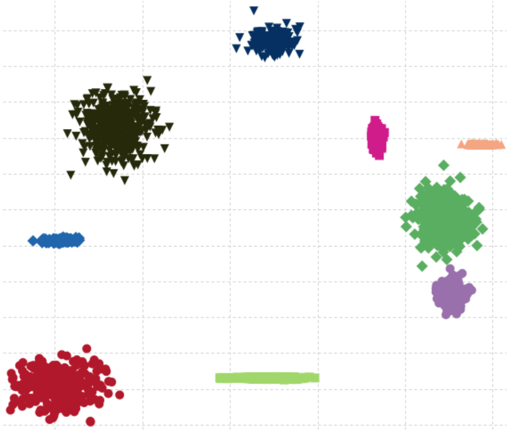
Tomas Barton
tomas.barton@fit.cvut.cz

01000110 Fakulta
01001001 Informačních
01010100 Technologií

May 23, 2019– Prague

Cluster analysis

1. Group similar items into same clusters and dissimilar into different clusters
2. Finds clusters in high-density regions



Clustering

Definition

Clustering is the organization of data points info a finite set of categories by abstracting the underlying structure of the data

– Hartigan JA (1975) Clustering Algorithms

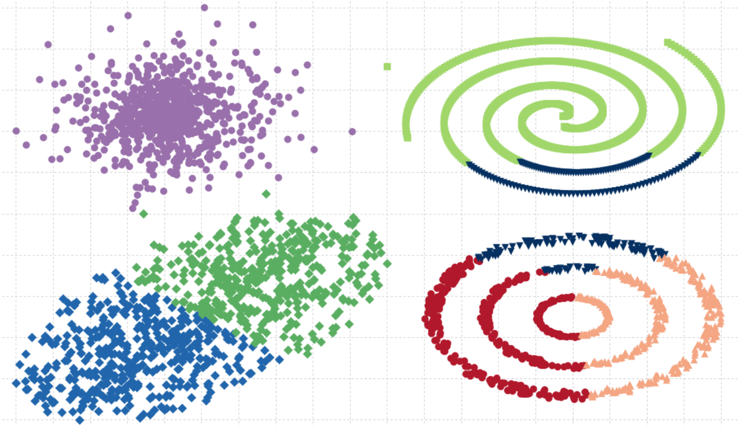
Clustering algorithms

There are many clustering algorithms:

- k-means
- Hierarchical clustering
- DBSCAN
- CLARANS
- Markov clustering
- Affinity propagation
- x-means
- Spectral clustering
- Self Organizing Maps
- Fanny
- Transitivity clustering
- CLUTO
- clusterdp
- Chinese Whispers
- Fast Community
- ... and many others

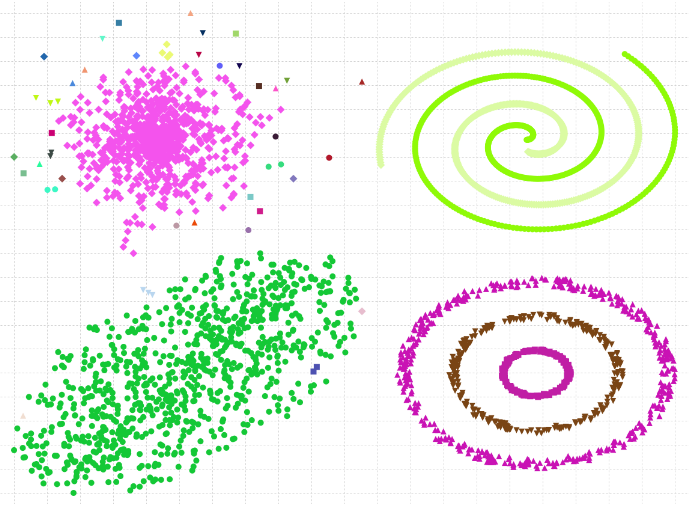
k-means clustering

- most algorithms optimize single objective
- e.g. minimize square distance inside a cluster
- fast, but inaccurate



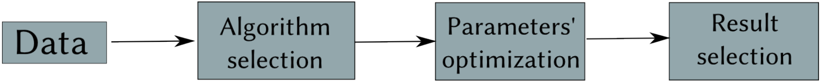
Single-Link clustering

- capable of discovering arbitrary shaped clusters
- but too sensitive to noise



Problems with clustering

1. Too many existing algorithms
2. Absence of “correct” objective function
3. Difficult to compare results
4. Too many parameters to optimize



Clustering valiadation

- Ball-Hall
- TraceW
- AIC
- Caliński-Harabasz
- Dunn index
- Gamma
- Tau
- McClain-Rao
- C-index
- BIC
- Ratkowsky-Lance
- Davies and Bouldin
- Silhouette
- Krzanowski-Lai
- Xie-Beni
- Banfield-Raftery
- GDI
- Ray-Turi
- SD index
- S_Dbw
- PBM
- Overall deviation
- Connectivity
- Compactness
- and many others …

Clustering validation

Most metrics considers following criteria:

$$f(\mathbb{C}) = \frac{\sum \text{distances in a cluster}}{\sum \text{distances between clusters}}$$

Clustering validation

Most metrics considers following criteria:

$$f(\mathbb{C}) = \frac{\sum \text{distances in a cluster}}{\sum \text{distances between clusters}}$$

Other concepts:

- variance-covariance
- entropy
- disconcordant pairs

Clustering objectives

C-index

$$f_{\text{c-index}}(\mathbb{C}) = \frac{S_w - S_{min}}{S_{max} - S_{min}}$$

where

- S_w is the sum of the within cluster distances
- S_{min} is the sum of the N_w smallest distances between all the pairs of points in the entire dataset. There are N_t such pairs
- S_{max} is the sum of the N_w larges distances between all the pairs of points in the entire dataset

Clustering objectives

Davies-Bouldin

Davies-Bouldin indexs combines two measures, one related to dispersion and the other to the separation between different clusters

$$f_{\mathrm{DB}}(\mathbb{C}) = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\left(\frac{\bar{d}_i + \bar{d}_j}{d(\mathbf{c}_i, \mathbf{c}_j)}\right)$$

where $d(\mathbf{c}_i, \mathbf{c}_j)$ corresponds to the distance between the center of clusters C_i and C_j, $\bar{d}_i$ is the average within-group distance for cluster C_i.

$$\bar{d}_i = \frac{1}{|C_i|}\sum_{l=1}^{|C_i|} d(\mathbf{x}_i(l), \bar{\mathbf{x}}_i)$$

No evaluation objective can outperform all others in all scenarios.

Clustering Evaluation

On clustering evaluation criteria

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.

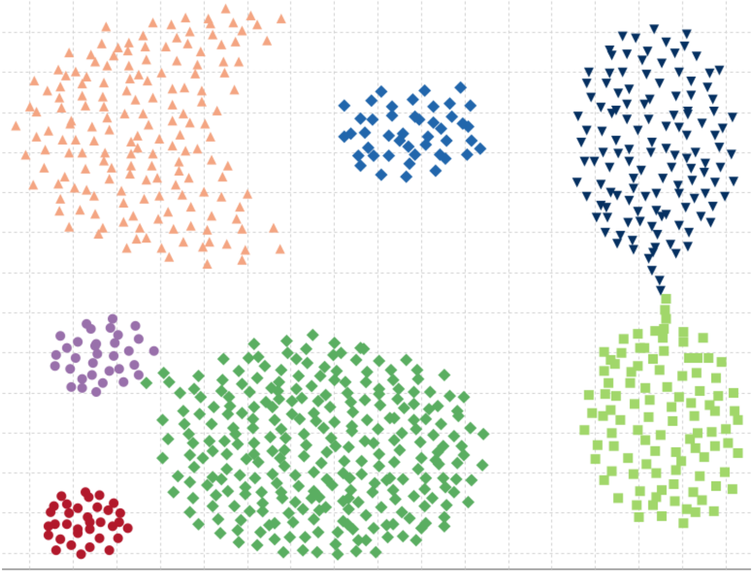
– Jain and Dubes, 1988

Problems with clustering evaluation

1. Unstable
2. Data biased
3. Some minimized other maximized
4. Unbounded definition range

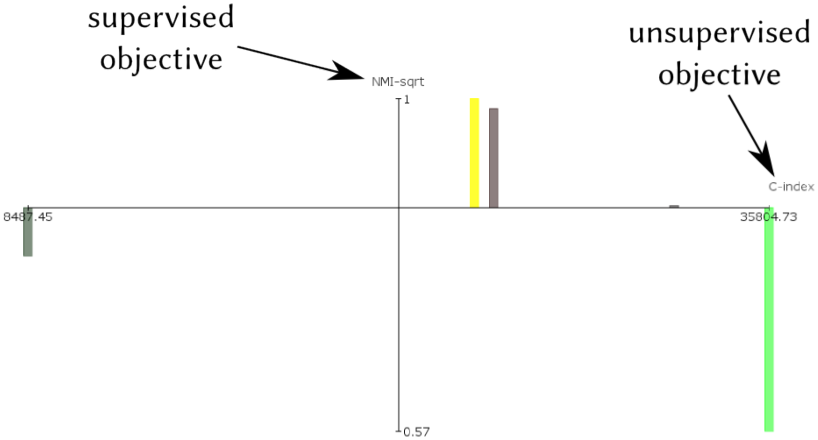
Clustering Ranking

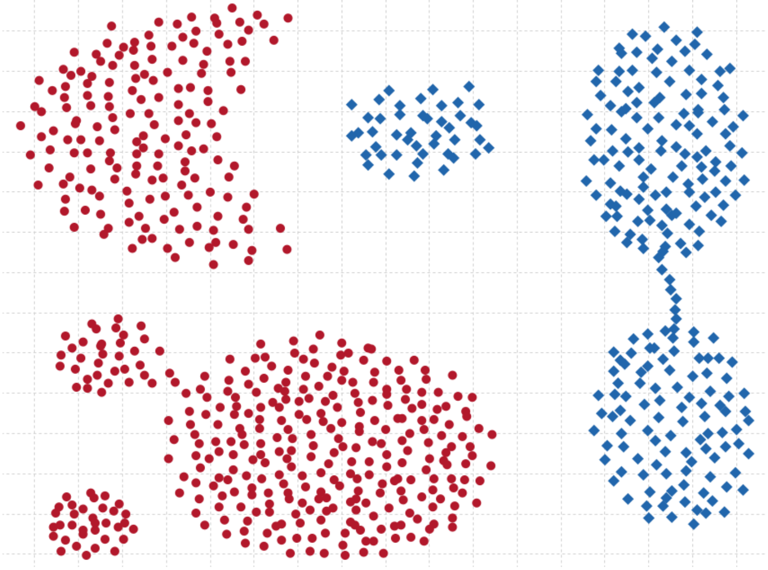
- Given a set $\mathbb{R}$ of clustering solution $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_\pi\}$ created from the same dataset
- We use a supervised function as reference
  $f_{supervised}(\mathbb{R}) \rightarrow \tau_{\text{sup}} = \texttt{rank}\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_\pi\}$
- And an unsupervised function
  $g_{unsupervised}(\mathbb{R}) \rightarrow \tau_{\text{unsup}} = \texttt{rank}\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_\pi\}$



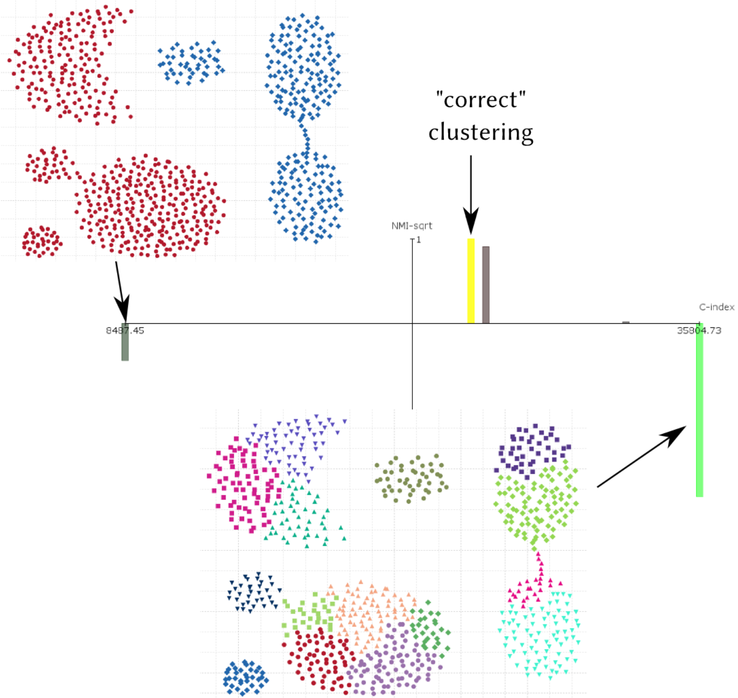
- `aggregation` **dataset – 7 clusters**

Visualization of objectives

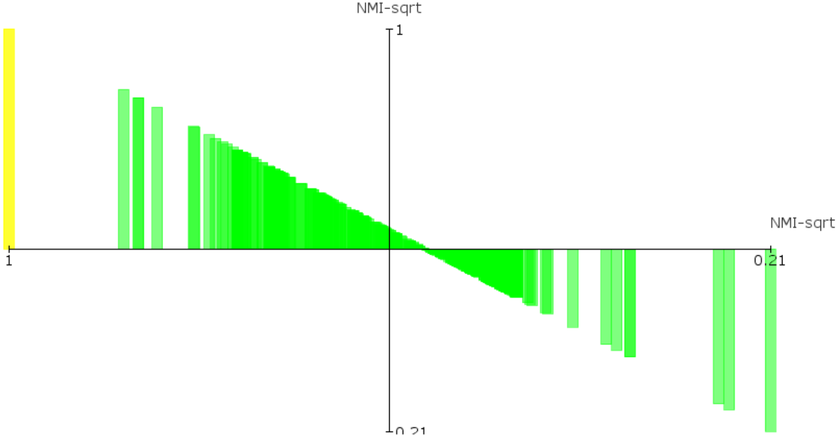




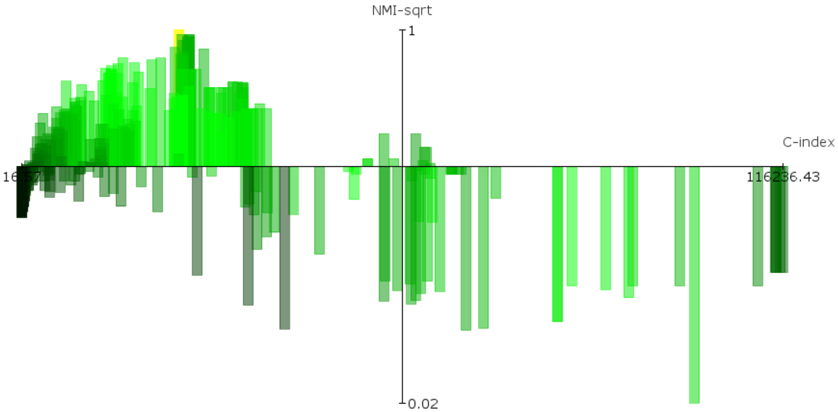
- Over-optimized clustering (highest C-index)


"correct"
clustering
NMI-sqrt
1
C-index
8487.45
35804.73

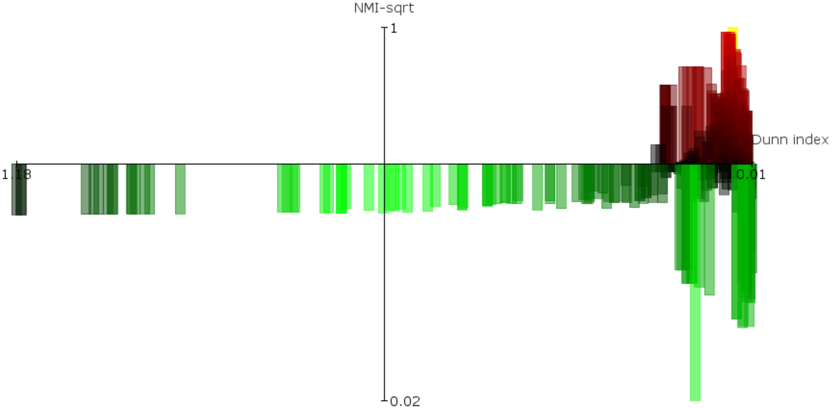
Ideal objective



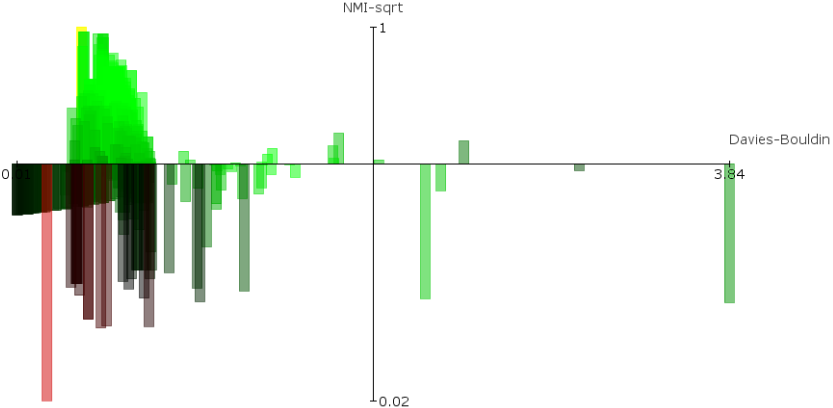
C-index



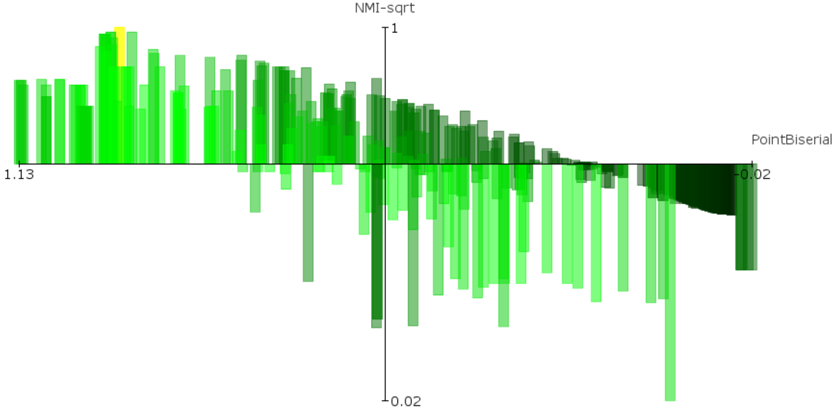
Dunn



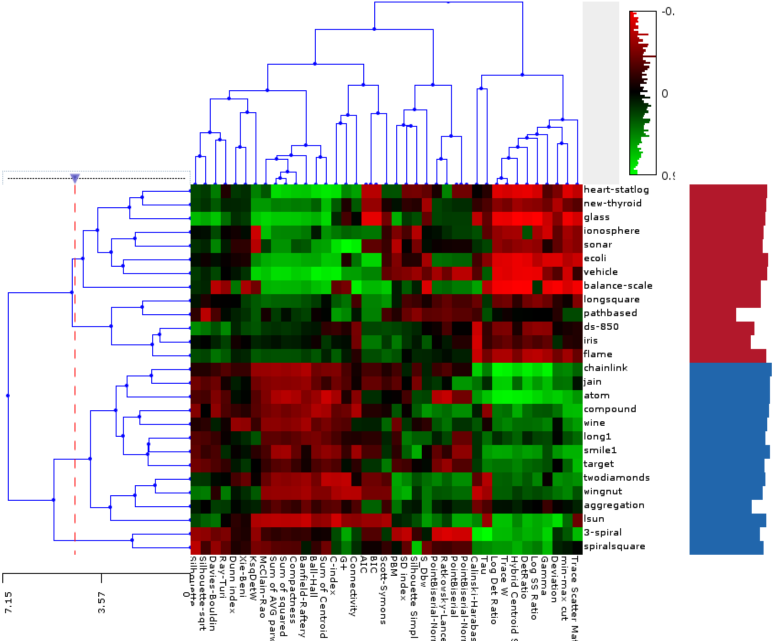
Davies-Bouldin



Point-Bi serial



Clustering correlations between sortings



Combinations of evaluation metrics

How to improve current state of single evaluation criterion?

Combinations of evaluation metrics

How to improve current state of single evaluation criterion?

- Select best performing criteria
- Combine them using ensemble approach

Combinations of evaluation metrics

How to improve current state of single evaluation criterion?

- Select best performing criteria
- Combine them using ensemble approach

1. Score based
2. Rank based
3. Multi-Objective sorting

Score based

Evaluation Ensembles

- Score normalization is needed
- Convert minimization to maximization – e.g. by flipping values around their mean

Strategies (Vendramin L. at al. 2013):

1. *Mean* arithmetic mean
2. *Harmonic Mean* penalize worst performing clusterings with a low score in at least one criterion
3. *Mean-2* remove most discrepant values
4. *Median* The median of the evaluation scores

Rank based

Evaluation Ensembles

Borda count method

- Classical voting scheme
- Can be adapted to minimization or to maximization of criteria
- Corresponds to mean of ranks
- Alternatively could be computed as median of ranks

Rank based

Evaluation Ensembles

Footrule

- Computes distance between two rankings

$$\text{Footrule}(\mathbb{R}) = \arg\min_{\pi} \left(\sum_{\tau \in \mathbb{R}} d(\tau, \pi) \right)$$

Distance between rankings:

$$d(\tau_1, \tau_2) = \sum_{i=1}^{|\tau|} |\tau_1(i) - \tau_2(i)|$$

Rank based

Evaluation Ensembles

Inconsistency

- Relative contribution is based on tendency to agree with the rest of the pool
- Inconsistency for given f_i criterion:

$$\text{Inconsistency}(\tau_{\mathrm{f_i}}) = \sum_{j=1}^{|\tau_{fi}|} \left(\tau_{fi}(j) - \mu(j)\right)^2$$

Weight for each ranked list:

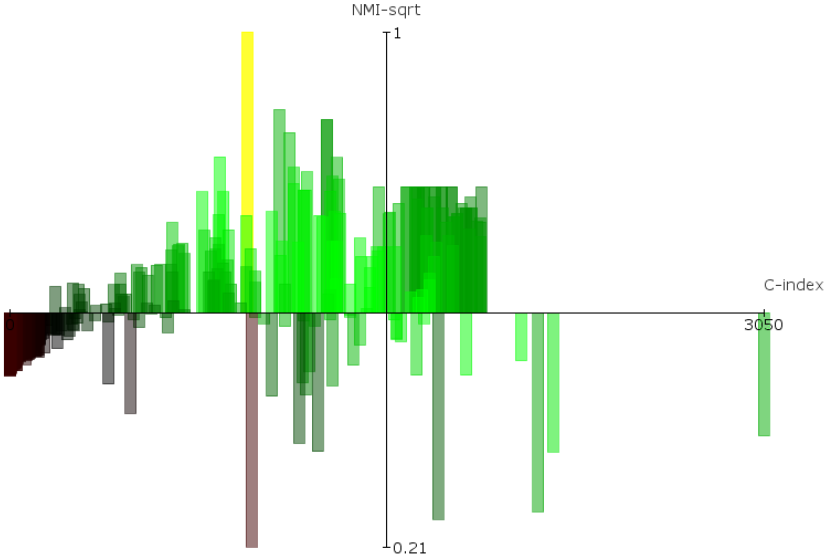
$$W(\tau_{fi}) = \frac{\text{Inconsistency}(\tau_{\mathrm{f_i}})}{\sum_{j=1}^{|\tau|} \text{Inconsistency}(\tau_{\mathrm{f_j}})}$$

Evaluation Ensembles

Problems

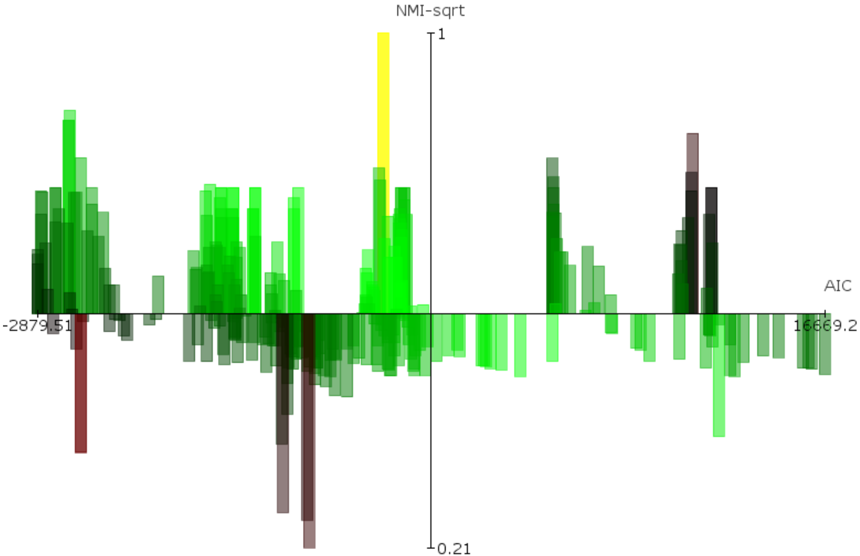
- Criteria needs to be carefully selected
- Improvement only over the weakest member of the ensemble

C-index (Iris datset)



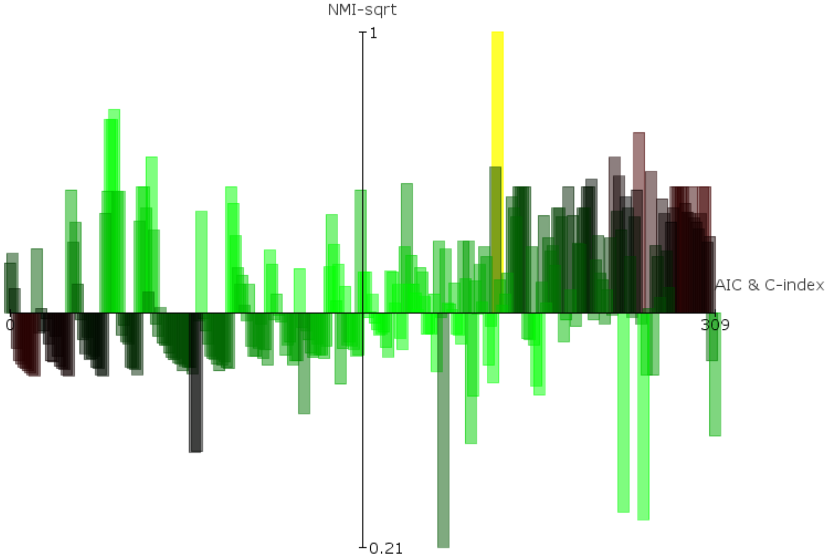
- correlation -0.81

AIC (Iris datset)



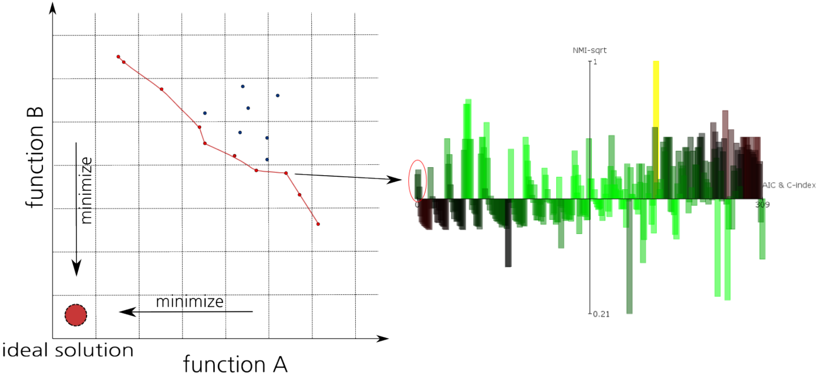
- correlation = 0.13

AIC & C-index (Iris datset)

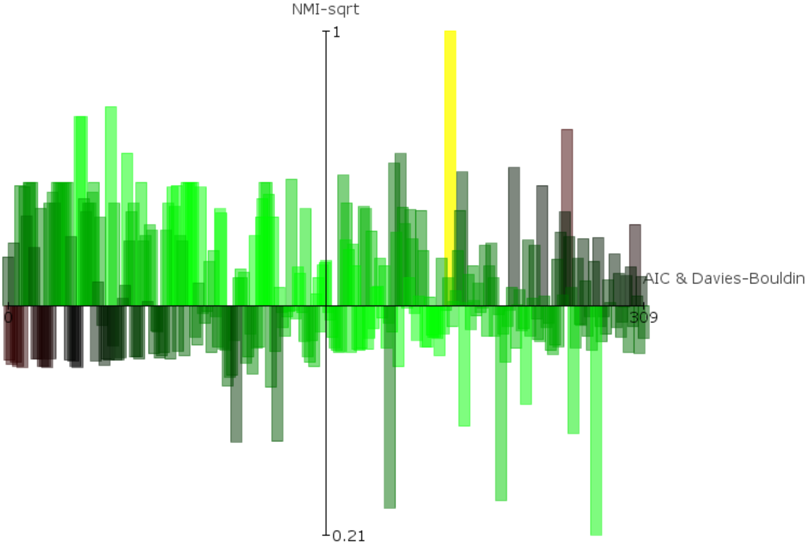


- correlation $= -0.47$

Pareto front projection

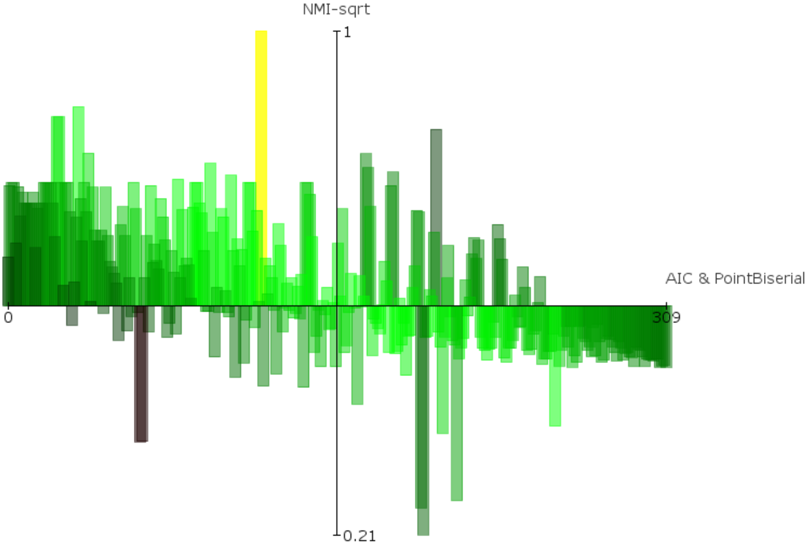


AIC & Davies-Bouldin (Iris datset)



- correlation = 0.12

AIC & Point BiSerial (Iris datset)



- correlation = 0.62

Meta-features

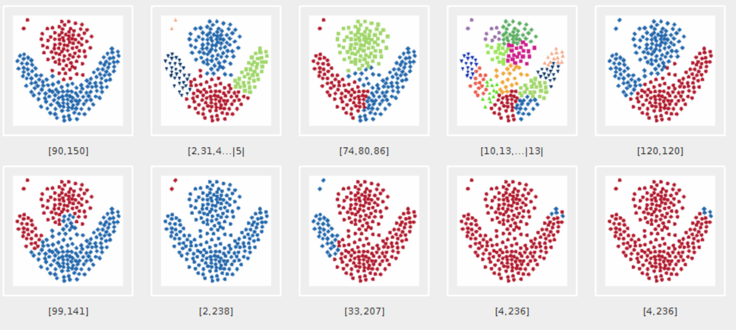
- $\log_2 N$ Input data size.
- $\log_2 D$ Number of attributes.
- **AV** – Average attribute variance (σ).
- **CV** – Coefficient of variation (CV) defined as the ratio of the standard deviation σ to the attribute mean.
- **CVQ1-4** Standard deviation of all attribute's first quartiles divided by their means.
- **SKEW** – The Pearson median skewness
- **KURT** – Kurtosis (min,max, mean, std).
- **KNN4** – Average distance to 4th nearest neighbor.
- **N2ER** – Node to edge ratio after k-NN graph bisection.
- **PCA** – Basic statistics of the principal component.

AutoML clustering

```
procedure AutomlClustering(dataset)
    extract meta-features
    choose ranking metric(s)
    landmarking - run fast templates
    find top-N templates based on meta-features
    rank clusterings
    while max. explored states not reached or time limit not reached do
        expand top performing templates
        remove worst solution from population
    end while
end procedure
```

AutoML exploration

- Goal is to be able to obtain diverse set of clusterings



Conclusion

- There are combinations of objectives that work in many cases, but are data dependent
- Evaluation ensembles needs to combine complementary objectives
- AutoML clustering heavily depends on training datasets and chosen objectives

Questions?

Thank you for your attention

tomas.barton@fit.cvut.cz