Active Semi-Supervised Clustering

Bc. Jakub Švehla
Supervisor: Ing. Tomáš Borovička

Department of Theoretical Computer Science
Faculty of Information Technology
Czech Technical University in Prague

June 14, 2018

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Motivation

- Clustering is a widely used data mining technique.
- It is strongly problem dependent and subjective.
  - Guide the clustering using domain knowledge.
- Getting the domain knowledge is expensive and time consuming.
  - Ask only the questions that you think will help the most.

Motivational example

Goals

1. Review and theoretically describe state of the art active semi-supervised methods.
2. Use or implement at least three methods and experimentally compare their performance.
3. Propose directions for further improvements of reviewed methods.

Types of Background Knowledge

- Partially labeled data
- Cluster-level constraints
- Instance-level (pairwise) constraints
    - Must-link constraints
    - Cannot-link constraints

Must-link constraints

Example

Cannot-link constraints

Example

Semi-Supervised Clustering using Pairwise Constraints

- Constraint-based methods
  - COP-KMeans
  - Pairwise Constrainted K-Means (PCK-Means)
- Metric-based methods
  - Metric K-Means (MK-Means)
- Combination of constraint- and metric-based methods
  - Metric Pairwise Constrainted K-Means (MPCK-Means)

Active Learning of Pairwise Constraints

- Active learning prior to clustering
  - Explore and consolidate
  - Min-Max
- Iterative active learning
  - Normalized point-based uncertainty (NPU) method

Implementation

- Implementation of all reviewed methods
- Works with *scikit-learn* machine learning library
- Will be published as a package

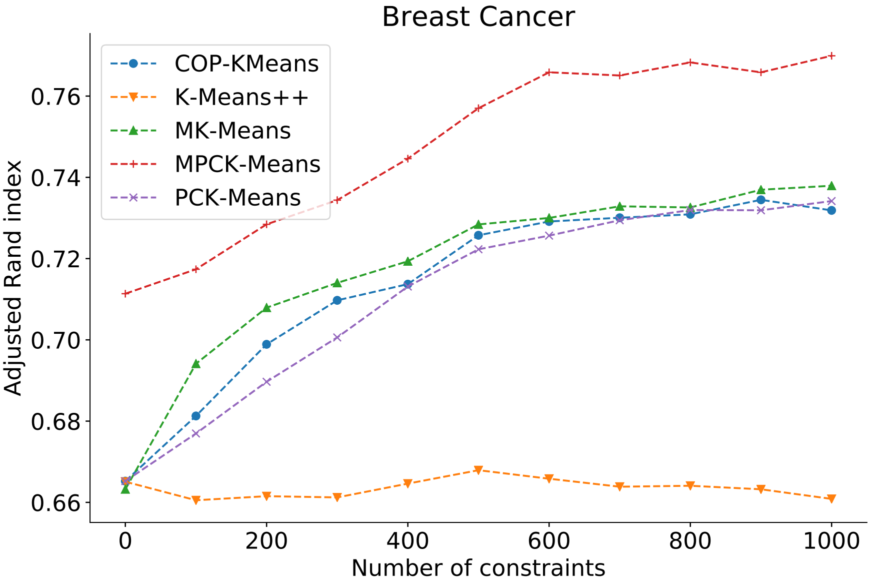
Implemented methods

- Semi-supervised clustering
  - COP-KMeans
  - PCK-Means
  - MK-Means
  - MPCK-Means (single diagonal matrix)
  - MPCK-Means (multiple full matrices)
- Active learning of pairwise constraints
  - Explore & Consolidate
  - Min-Max
  - Normalized point-based uncertainty method

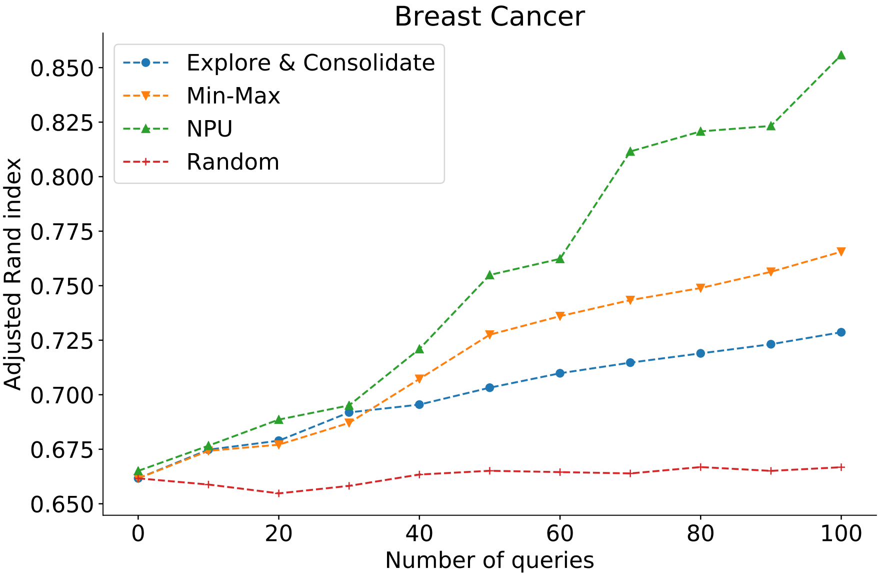
Experiments

- Experimentally evaluate influence of increasing background knowledge on the results of the methods.
    - Generated learning curves using cross-validation.
- 6 data sets of various sizes, dimensions and numbers of classes.
- Compare the resulting clustering to the known labels.

Experiments with semi-supervised clustering methods



Experiments with active learning methods



Experiment results

- Both semi-supervised clustering and active learning considerably outperform unsupervised clustering and random selection, respectively.
- None of the reviewed methods was superior to the other methods on the testing data sets.

Future work

- Consensus clustering
- Automated estimation of the number of clusters
- Querying the oracle in batches in the iterative active learning

Conclusion

- Reviewed and implemented 8 active semi-supervised methods.
- Experimentally evaluated and compared all methods.
- Proposed directions for further improvements of reviewed methods.

Remarks

Remark

Proč jste se nepokusil simulovat chování algoritmů i s *noisy constraints*?

Remarks

Remark

Jaké algoritmy lze použít, pokud mám v datech nečíselné hodnoty (např. *nominal features*)?

References

- Oldest Army Jeep, Retrieved from https://www.militarytimes.com/news/your-military/2015/12/13/oldest-army-jeep-finally-gets-some-tlc/
- Mustang GT Convertible, Retrieved from http://bestluxurycars.us/gallery/amazing-2014-mustang-gt-convertible-with-mustang-gt-convertible-gotta-have-it-green/
- Ferrari 488 GTB, Retrieved from https://www.autocar.co.uk/car-review/ferrari/488-gtb
- Lamborghini Huracán Spyder, Retrieved from http://money.cnn.com/2015/09/14/autos/lamborghini-huracan-spyder/index.html