Martin Hahmann, Markus Dumat, Dirk Habich and Wolfgang Lehner

Explorative Multi-View Clustering Using Frequent-Groupings

MultiClust 2012
Two sides of Multiple Clusterings

Out of many, one:
- integrate multiple solution into a single result
- ensemble clustering
- multi-source clustering

Out of one, many:
- generate different results from the same source
- alternative clustering
- subspace clustering

Our goal:
- bring both sides together using frequent groupings
Core Concept
Best of both worlds

Frequent Groupings Idea (MultiClust2011)

- starting point: ensemble clustering
- goal: find cluster assignment that agrees with majority of ensemble
- find objects frequently assigned to the same cluster

- parallel to the concept of frequent itemsets
- mapping to ensemble clustering:

**items**

**transaction**

**support**
Frequent Groupings

Frequent Grouping Graph

- Frequent Groupings $fg_1$: $x_1, x_3$ (4 times)
- Frequent Groupings $fg_2$: $x_1, x_3, x_4, x_5$ (3 times)
- Frequent Groupings $fg_3$: $x_1, x_2$ (4 times)
- Frequent Groupings $fg_4$: $x_4, x_5$ (4 times)
- Frequent Groupings $fg_5$: $x_6, x_8$ (4 times)
- Frequent Groupings $fg_6$: $x_4, x_5$ (4 times)
- Frequent Groupings $fg_7$: $x_4, x_5, x_6$ (3 times)
- Frequent Groupings $fg_8$: $x_7, x_8, x_9$ (4 times)
- Frequent Groupings $fg_9$: $x_7, x_8, x_9$ (4 times)
Result Extraction

- Alternative combinations

- Frequent groupings overlap

- Explorative Multi-View Clustering

- Maximize cluster size

- Maximize support
Scaled Up Scenario
Example Dataset

Manual processing not feasible

- 1500 objects, 10 clusterings, 69 clusters
- apply existing frequent itemset mining algorithms
- CARPENTER with minsupp= 0.5 → 72 frequent groupings
How to create alternative Solutions?

- assumption: every object is assigned to a cluster
- resembles exact-cover problem (np-hard)

First Approach

- Algorithm X by Donald E. Knuth
- implementation: Dancing Links

570,000 alternative clusterings
Result Extraction

Problems

- too many results to work with
- high similarity between alternatives
- smallest frequent groupings define minimum difference

Options for result limitation:

- filter small groupings
- compute similarity matrix

- top-k Dancing Links:

  Top-k: 5

  Clusterings: \{ (fg14, fg35, fg7), (fg10, fg7), (fg3, fg7) \}
Result Extraction

Second Approach: Exploring the graph
Result Extraction

Explorative Multi-View Clustering
Result Extraction

Explorative Multi-View Clustering
Result Extraction

Explorative Multi-View Clustering
Result Extraction

Database Technology Group

Explorative Multi-View Clustering
Result Extraction

Explorative Multi-View Clustering
- number of alternatives is predictable
- less or even the number of root nodes
- additional alternatives via root removal & recursion
Result Similarity

Database Technology Group

id : [support] : [size]

• A1
• A2
• A3

Explorative Multi-View Clustering
Result Similarity

Similarity measurement via intersection

\[
\frac{(105 + 208)}{620} = 0.58
\]
Future Work
Future Work

**Exploration so far**
- top-down traversal from roots \(\rightarrow\) alternatives through splitting clusters

**Other possibilities**
- bottom-up traversal from leaves \(\rightarrow\) alternatives through merging clusters
- starting level between root and leaves?
- combine bottom-up and top-down
Future Work

Further Graph Utilization

- examine branching for dissimilarity of alternatives
- for feedback on ensemble diversity/homogeneity
- relation between ensemble and alternatives
Conclusion

- frequent groupings combine alternative and ensemble-clustering
- offer alternative and robust results
- automated recombination and its problems
- user-centered exploration
Questions?