Multi-hierarchical clustering for exploring complex data relationships

Milton Pividori[1]

1 Section of Genetic Medicine, Department of Medicine, The University of Chicago.

Background

- Clustering have become an essential set of methods to understand the structure of the increasing amount of information produced today [1]. Traditional techniques are categorized in:

→ **Partitional**: produce a flat partition with k clusters; Examples: k-means, self-organizing maps.

→ **Hierarchical**: produce a tree structure that represents a set of clusters organized into a single and independent hierarchy.

- **Consensus clustering**, on the other hand, are a recent group of methods that produce higher quality partitions, outperforming classical methods [2].

→ First, an ensemble is created, that is, a set of several base data partitions.

→ Then, the ensemble is combined into a consensus partition, which outperforms ensemble members.

Motivation

- Big data is about **complexity**, not only size/volume.
- Traditional hierarchial approaches produce a **single** and **independent hierarchy**. However, complexity of data could contain many different, valid and **interconnected hierarchies**.
- Consensus methods enable **distributed computing**, are appealing for **privacy-sensitive** scenarios and address handling of **data heterogeneity**, what make them interesting for **big data**.

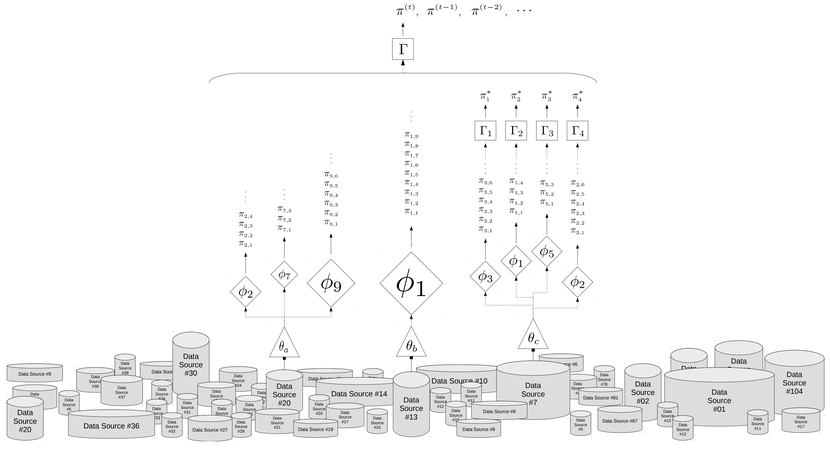
References

[1] Rui Xu and Don Wunsch. *Clustering*. Wiley-IEEE Press, 2009.

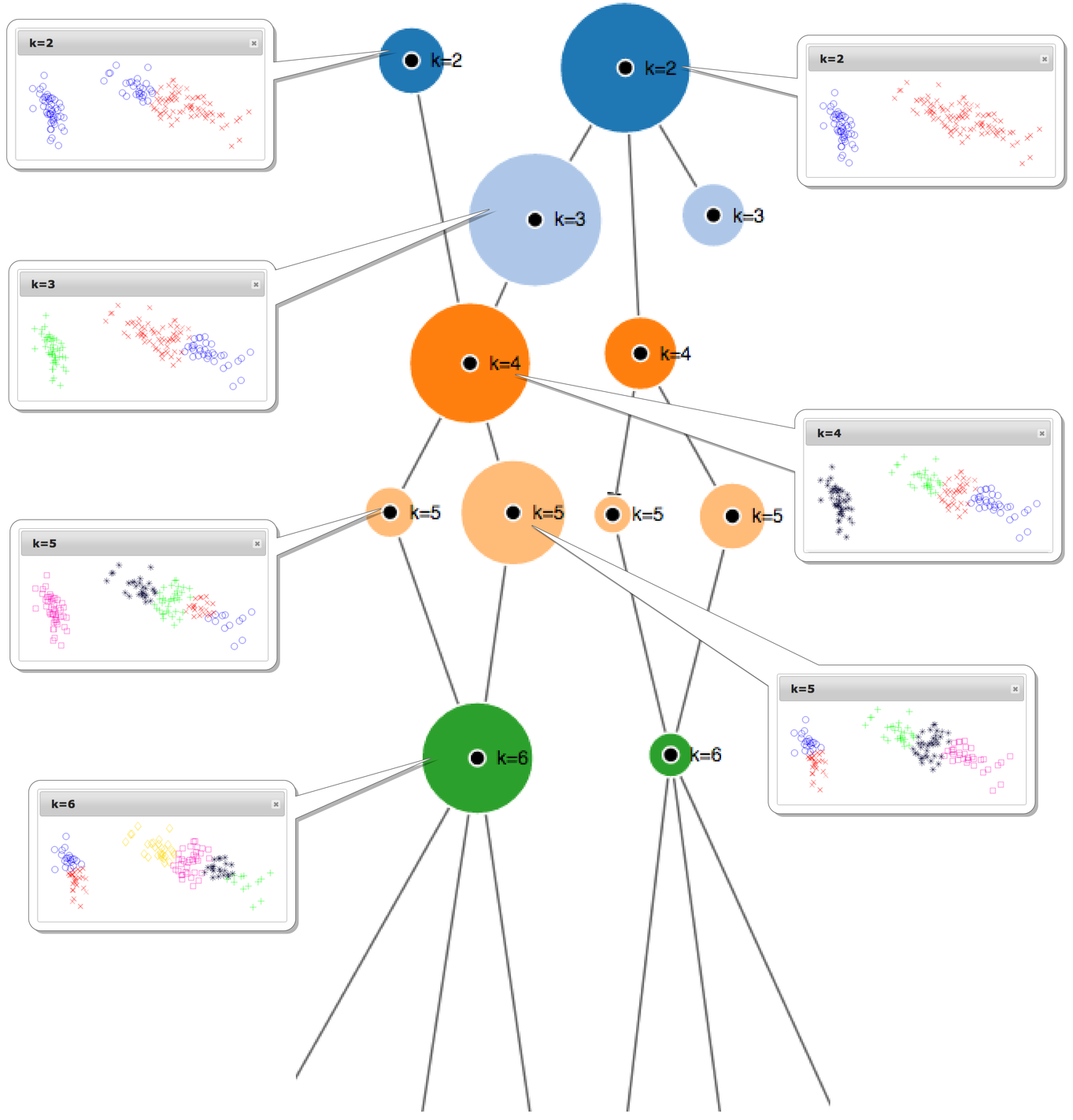
[2] Dong Huang, Jianhuang Lai, and Chang-Dong Wang. "Ensemble clustering using factor graph". In: *Pattern Recognition* 50 (2016), pp. 131–142.

Proposal

- First, leveraging consensus clustering methods, we propose a **scheme** for processing **large volumes of data**, which is able to handle **highly heterogeneous data sources** and enable decentralized processing of **privacy-sensitive** data.
- This scheme is employed to produce several combined partitions by varying **ensemble diversity**.
- This set of consensus solutions is analyzed using a clustering index that finds partitions that have a hierarchical relationship (their clusters are consistently divided or merged). This information is used to generate a graph unveiling **multiple interconnected hierarchies** in data.



Example of multi-hierarchical clustering applied on the Iris dataset



LaTeX TikZposter