Scalable robust clustering method for large and sparse data

Tommi Kärkkäinen
University of Jyväskylä, Finland

Clustering is the most common unsupervised, descriptive analysis technique to reveal hidden patterns and profiles from a dataset. There exists large number of different clustering algorithms, but approaches that specifically address clustering of sparse datasets are still scarce, even if real world datasets are many times characterized by missing values with unknown sparsity pattern. Typical approaches in the knowledge discovery process is to either completely omit the observations with missing values or use some imputation method to fill in the holes of data. However, the throw data away approach does not utilize all possible data and the imputation necessarily introduces assumptions about the unknown density of the data. Moreover, by the well-known curse-of-dimensionality results, such assumptions are no more valid in the high dimensional spaces. The purpose of this presentation is to describe and summarize a line of research that addresses the sparse clustering problems with the available data strategy and robust prototypes. The strategy allows one to utilize all available data without any additional assumptions. The actual prototype-based clustering algorithm, the k-spatial medians, relies on the computation of a robust prototype as cluster centroid, again argumenting on non-Gaussian within-cluster error in comparison to the classical k-means method. As with any prototype-based algorithm, the initialization step of the locally improving relocation algorithm has an important role and should be designed to handle the sparse data. Such an approach is proposed and the scalability of a distributed implementation of the whole algorithm is tested with openly available large and sparse datasets.

Biography

Tommi Kärkkäinen has completed his PhD at University of Jyväskylä in 1995 and worked as a Full Professor in the Faculty of Information Technology since 2002. He has been and is serving in many positions of administration and responsibility at the faculty and the university level. He has published over 150 research papers, led dozens of R&D projects, and supervised over 20 PhD theses.

tommi.karkkainen@jyu.fi

Notes: