Searching for the Most Plausible Partition: an Evidential Reasoning Approach to Clustering

Orakanya Kanjanatarakul
Heudiasyc UMR 7253
Université de Technologie de Compiègne & CNRS
Sorbonne Universités, France
Faculty of Management Sciences
Chiang Mai Rajabhat University, Thailand
okanjana@utc.fr

Thierry Denoeux
Heudiasyc UMR 7253
Université de Technologie de Compiègne & CNRS
Sorbonne Universités, France
tdenoeux@utc.fr

Clustering can be seen as the search for a "good" partition of a set of n objects described either by attributes, or by a dissimilarity matrix. Usual approaches are based either on a geometric criterion, as in the k-means algorithm, or on a finite mixture model whose parameters are estimated using, e.g., the EM algorithm. Here, we propose a different view of partitional clustering, in which dissimilarities are seen as pieces of evidence and represented as belief functions on the set of all partitions of the dataset under study. Using a technique similar to the one used in [1] for the association problem, we show that the most plausible partition can be found for small n. We then propose a heuristic algorithm that can handle large datasets.

Formalization Let $\mathcal{O}$ denote a set of n objects and let $\mathcal{R}$ be the set of equivalence relations on $\mathcal{O}$ (this set is in one-to-one correspondence with the set of partitions). We assume the existence of a true equivalence relation R_0. Dissimilarities between objects are considered as items of evidence about R_0, which can be represented by mass function m_{ij} with three focal sets: the set $\mathcal{R}_{ij}$ of equivalence relations containing objects i and j, its complement $\neg\mathcal{R}_{ij}$, and $\mathcal{R}$, and corresponding masses $m_{ij}(\mathcal{R}_{ij}) = \alpha_{ij}$, $m_{ij}(\neg\mathcal{R}_{ij}) = \beta_{ij}$ and $m_{ij}(\mathcal{R}) = 1 - \alpha_{ij} - \beta_{ij}$. After combining these $n(n-1)/2$ mass functions by Dempster's rule, we get a mass function m on $\mathcal{R}$ with contour function pl defined by the following equation,

$$\ln pl(R) = C + \sum_{i<j} R_{ij} \ln \frac{1-\beta_{ij}}{1-\alpha_{ij}}, \qquad (1)$$

where C is a constant. The most plausible partition can thus be found exactly, for small n (until, say, $n \leq 100$) using a binary linear programming solver.

Hopfield Model To make the above approach feasible for large n, we need a heuristic optimization method. We show that a local maximum of $\ln pl(R)$ defined by (1) can be found by a Hopfield neural network model [2] with n neurons, in which each neuron can be in one of c states, where c is the desired number of clusters. The weight v_{ij} of the connection between neurons i and j is the coefficient of R_{ij} in (1). Starting from random initial states, the state of each neuron i is updated at asynchronous times, by finding k such that $\sum_{j\neq i} v_{ij}s_{jk}$ is maximum, where $s_{jk} = 1$ if neuron j is in state k, and $s_{jk} = 0$ otherwise. This algorithm is shown to converge to a global network state that corresponds to a local maximum of (1).

Results and Conclusions The above clustering algorithm was applied to several datasets with different characteristics, including: large numbers of objects and/or clusters, non-metric dissimilarities, and complex cluster shapes, showing good performances as compared to existing algorithms. The definition of constants α_{ij} and β_{ij} is problem-specific and is an important step for ensuring good performances of the method. The application of this approach to semi-supervised clustering is currently under study.

Keywords. Clustering, Dempster-Shafer theory, Evidence theory, belief functions, Hopfield network.

Acknowledgement This research was supported by the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" by the National Agency for Research (reference ANR-11-IDEX-0004-02).

References

[1] T. Denœux, N. E. Zoghby, V. Cherfaoui, and A. Jouglet. Optimal object association in the Dempster-Shafer framework. *IEEE Transactions on Cybernetics*, 44(22):2521–2531, 2014.

[2] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558, 1982.