Enhanced clustering

Gianluca Bortoli
DISI - University of Trento
Student id: 179816
gianluca.bortoli@studenti.unitn.it

Martin Brugnara
DISI - University of Trento
Student id: 182904
martin.brugnara@unitn.it

ABSTRACT

Keywords

Big data, Data mining, Clustering, Streaming, Parallel computation

1. INTRODUCTION

The clustering problem received a lot of attention in the data mining [3], statistics [5, 1] and machine learning literatures. Furthermore it has been exploited in many other disciplines.

Generally speaking, the problem consists in grouping together data items that are "similar" to each other. The concept of similarity varies a lot in the different contexts it is applied. For example the Euclidean distance (L2) can be used when dealing with continuous values, or the Jaccard similarity when dealing with generic sets of elements. Nevertheless, clustering can be also viewed as identifying the dense regions of the probability density of the data source [2].

The literature suggests two approaches: iterative and hierarchical algorithms. The first strategy usually needs some parameters to be set and known in advance. For example the *K-means*, which is one of the most popular and adopted algorithm, requires the number of cluster to be found (K). The latter can be implemented both in a top down (divisive) or a bottom up (agglomerative) manner. Initially, the divisive algorithm treats all data as a single big cluster and later splits it until every object is separated [4]. On the contrary, the agglomerative starts considering each "element" as a *singleton* (a cluster composed of one element). Next, the most similar clusters are collapsed together until only one big cluster remains. Implicitly the merging order defines a clear hierarchy among the intermediate representations.

2. RELATED WORK

3. PROBLEM DEFINITION

4. SOLUTION

5. CONCLUSIONS AND FUTURE WORK

6. REFERENCES

[1] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[2] P. S. Bradley, U. M. Fayyad, C. Reina, et al. Scaling clustering algorithms to large databases. In *KDD*, pages 9–15, 1998.

[3] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park, 1996.

[4] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[5] H. C. Tijms. *Stochastic models: an algorithmic approach*, volume 303. John Wiley & Sons Inc, 1994.