MOGCLA: A Multi-Objective Genetic Clustering Algorithm for Large Data Analysis

Héctor D. Menéndez
University College London
Gower Street, London, WC1E 6BT
United Kingdom
h.menendez@ucl.ac.uk

David Camacho
Universidad Autónoma de Madrid
C/ Tomás y Valiente 11
Madrid, Spain
david.camacho@uam.es

ABSTRACT

Automatic Manifold identification in large datasets is currently a challenging problem in Machine Learning. This process consists on separating a large dataset blindly, according to the form defined by the data instances in the space. Data is discriminated in groups described by their form. These approaches are usually focused on continuity-based methods where the manifold respects a continuity criterion. Currently, Map-Reduce and online clustering techniques try to deal with the discrimination process, but there are a few algorithms which can deal with the manifold extraction process. This work pretends to face this problem from a Genetic-based approach. We have designed a new algorithm, named MOGCLA, which performs a search in two levels combining Map-Reduce and Multi-Objective Optimization. We have compare the new algorithm against other clustering.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]: Heuristic methods

Keywords

MOGCLA; Genetic Algorithms; Multi-objective; Clustering; Manifold; Large Data Analysis

1. INTRODUCTION

This work has been focused on the automatic manifold identification problem. This problem consists on grouping the data according to the form they define in the search space. The automatic process is usually blind, for that reason, there are several clustering methodologies which have been tried to deal with it [3]. It is focused on combining Multi-Objective Genetic Algorithms (MOGA) [4] and Map-Reduce clustering [5] to improve the manifold identification process. The idea is to divide the clustering search in two levels: the first is a micro search which summarizes the information of the original search space using the Voronoi regions defined by a clustering algorithm, and the second step uses a multi-objective approach to join the regions discriminating the manifold structure.

In this work, we have compared this new approach with the classical K-means version of Map-Reduce [5] and the combination of Spectral Clustering and the Nyström method [1], in order to evaluate the performance of the new method comparing with synthetic datasets.

2. THE MOGCLA ALGORITHM

This section describes the Multi-Objective Genetic Graph-based Streaming Clustering Algorithm (MOGCLA). This algorithm combines a clustering version of K-means implemented in Map-Reduce and the Multi-Objective Genetic Graph-based Clustering algorithm [2] (MOGGC) in order to reduce the computationally effort and improve the scalability for large datasets. It has been divided in two levels: a micro search and a macro search.

The micro search chooses the Voronoi Regions that the algorithm is going to use during the clustering process. These regions are chosen using a Map-Reduce version of K-means [5]. The algorithm defines the regions optimizing the centroid positions. This information is sent to the next level in order to improve the clustering decision. Using only the centroid position we are able to reduce the genetic search, reducing the chromosomes size and the computationally effort during the macro search.

The macro search joins previous regions in order to create good clusters. These approaches are usually used in online clustering algorithms. Using the information of the centroids, we group the centroids in bigger clusters according to their continuity.

The manifold reconstruction process is based on the information extracted during the first step. The regions form a graph, they are used to represent the nodes, while their similarities are used to represent the edges. This topology is used to define the final clusters using a graph-cut methodology. In this work, the MOGGC algorithm [2] has been chosen for this task.

3. EXPERIMENTAL RESULTS

The synthetic datasets have been generated using the R package mlbench[1] which allows to generate big synthetic

[1]http://cran.r-project.org/web/packages/mlbench/mlbench.pdf

GECCO'15 Companion, July 11–15, 2015, Madrid, Spain.
ACM 978-1-4503-3488-4/15/07.
http://dx.doi.org/10.1145/2739480.2764689

Data	MOGCLA	Big K-means	SC+Nyström
Ca	**100%** ± 0.04	▼65.5% ± 0.05	98.6% ± 0.17
Cu	91.4% ± 0.08	▼87.2% ± 0.09	▲99.7% ± 0.17
Hy	**100%** ± 0.00	▼81.6% ± 0.12	▼76.7% ± 0.16
Sh	**100%** ± 0.04	**100%** ± 0.17	▼68.7% ± 0.44
Si	**100%** ± 0.00	**100%** ± 0.15	**100%** ± 0.00
Sm	**100%** ± 0.05	▼63.5% ± 0.18	▼75.0% ± 0.17
Sp	**100%** ± 0.10	▼50.0% ± 0.00	**100%** ± 0.10
SpN	**76.9%** ± 0.18	▼59.2% ± 0.00	▼59.6% ± 0.04

Table 1: Median and Standard Deviation accuracy results of the application of the algorithms to the synthetic datasets. Values in bold shows the best results. ▲symbol represents those cases there the algorithm is significantly better than MOGCLA and ▼represents the opposite.

datasets with a topological structure. The datasets -all composed by 50000 instances- that have been generated are the following: Cassini, Cuboids, Hypercube, Shapes, Simplex, Smiley, Spirals with and without noise.

Table 1 show the results for the algorithms. As we can appreciate, MOGCLA obtains generally good results, but it is important to analyse the algorithm according to each dataset in order to identify its weaknesses.

In the case of **Cassini** (Ca) dataset, MOGCLA obtains the best Median results (100%). SC+N obtains good results (98.6%) while Big K-means obtains similar and worse results than the previous algorithms (around the 66%). According to the Wilcoxon test, MOGCLA is statistically better than K-means.

For **Cuboids** (Cu) dataset, SC+N obtains the maximum results (99.7%). MOGCLA obtains worse results than the previous algorithm (91.4%). K-means obtains the worst results (87.2%).

Hypercube (Hy) results show that MOGCLA is able to discriminate the clusters perfectly. The rest of the algorithms have more problems to discriminate the cluster distribution.

In the case of **Shapes** (Sh), all the algorithms obtain the maximum Median results (100% of Mediam) but the SC+N algorithm (68.7%). However, MOGCLA is the most stable algorithm according to the standard deviation (0.04).

Simplex (Si) is easy for all the algorithms. However, only MOGCLA and SC+N keep these results in all the iterations.

Smiley (Sm) shows that MOGCLA is the only algorithm which discriminates the manifolds perfectly. The rest of the algorithms have problems discriminating the manifolds.

The **Spiral** (Sp) dataset tests how the algorithms can deal with continuity datasets without noise. In this case, Spectral and MOGCLA obtain the best results (100.0%). K-means obtains the worst possible results.

The **SpiralN** (SpN) dataset introduces noise to the previous one. In this case, the results are similar than in the Spiral case, however, MOGCLA obtains better results than SC+N. This should be a consequence of the data structure which is difficult to discriminate. When the number of instances is increased, the noise also increases and SC+N and K-means are not able to deal with it. The Wilcoxon test show that all the algorithms obtain significantly worse results than MOGCLA.

4. CONCLUSIONS

This work proposes MOGCLA, a new MOGA algorithm, that performs a manifold identification process in two levels: a micro search which summarizes the original search space using a centroid-based clustering algorithm combined with a Map-Reduce architecture, and a macro search which joins the Voronoi regions identified during the micro search in other to discriminate the manifolds. It obtains good results in a bigger search space. The experiments show that the new algorithm obtains competitive results that are better than the classical algorithms, and has a similar (or better) clustering results than previous obtained using SC and Nyström.

5. ACKNOWLEDGMENTS

This work has been partly supported by: Spanish Ministry of Science and Education under projects TIN2010-19872 and TIN2014-56494-C4-4-P, Comunidad Autonoma de Madrid under project CIBERDINE S2013/ICE-3095 and Savier an Airbus Defense & Space project (FUAM-076914 and FUAM-076915).

6. REFERENCES

[1] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(2):214–225, 2004.

[2] H. D. Menéndez, D. F. Barrero, and D. Camacho. A multi-objective genetic graph-based clustering algorithm with memory optimization. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 3174–3181. IEEE, 2013.

[3] H. D. Menéndez, D. F. Barrero, and D. Camacho. A genetic graph-based aprroach for partitional clustering. *International Journal of Neural Systems*, 24(03):1430008, 2014.

[4] T. Murata and H. Ishibuchi. Moga: Multi-objective genetic algorithms. In *Evolutionary Computation, 1995., IEEE International Conference on*, volume 1, page 289. IEEE, 1995.

[5] W. Zhao, H. Ma, and Q. He. Parallel k-means clustering based on mapreduce. In *Cloud Computing*, pages 674–679. Springer, 2009.