

Clustering large data set: an applied comparative study

Bocci Laura, Mingo Isabella

Dipartimento di Sociologia e Comunicazione

Sapienza Università di Roma

laura.bocci@uniroma1.it, isabella.mingo@uniroma1.it

Abstract: The aim of this work is to compare different strategies to cluster large data sets. In particular, the performance of the classical *K*-means algorithm and two strategies which combine different clustering procedures in a sequential way, are investigated through the analysis of a real-life data set consisting of approximately 1.5 million units.

Keywords: large data set, mixed clustering, multiobjective genetic algorithm.

1. Introduction

Clustering is often useful for analyzing and summarizing information within large data sets. The growing size of data sets and databases has led to increased demand for good clustering methods for analysis and compression, while at the same time introducing constraints in terms of memory usage and computation time. Direct application of classical clustering technique to large data sets is often prohibitively expensive in terms of computer time and memory. Instead, extensions to large data sets usually rely on analyzing one or more samples of the data, and vary in how the sample-based results are used to derive a partition for all of the data. Extensive studies dealing with comparative analysis of different clustering methods suggest that there is no general strategy which works equally well in different problem domains. However, it has been found that it is usually beneficial to run schemes that are simpler and execute them several times rather than using schemes that are very complex but need to be run only once. The aim of this work is to analyze different strategies to cluster large data sets. In particular, the performance of a classical partitioning technique and two schemes which combine different clustering procedures in a sequential way are investigated through the study of a real-life data set. Firstly, we take into account the well known *K*-means algorithm which was directly applied to the entire data set to search for the appropriate cluster centers and the corresponding partition of units. Secondly, a scheme, called *mixed clustering strategy*, which implements both a partitioning technique and a hierarchical method is taken into account. Finally, since genetic algorithms (GA) perform well in problems with high computational complexity such as clustering and partitioning problems, especially if the number of groups is unknown, we propose a strategy, which will be called *multiobjective GA based clustering strategy*, that implements the *K*-means algorithm along with a genetic algorithm that optimizes two different functions.

The paper is organized as follows. Section 2 is devoted to the description of the above mentioned clustering strategies. The results of the comparative analysis, dealing with an application to a large real-life data set, are illustrated in section 3.

2. Clustering strategies

In this section we outline the two clustering strategies used in the analysis, the mixed clustering strategy and the multiobjective GA based clustering strategy, which combine different clustering approaches.

Mixed clustering strategy

The mixed clustering strategy, proposed by Lebart et al. (2004) and implemented in the package Spad 5.6, combines the method of clustering around moving centers and an ascending hierarchical clustering. In the first stage the procedure uses the algorithm of moving centers to perform several partitions (called base partitions) starting with several different sets of centers. The aim is to find out a partition of n objects into a large number G of stable groups by cross-tabulating the base partitions. Therefore, the stable groups are identified by the sets of units that are always assigned to the same cluster in each of the base partitions. The second stage consists in applying to the G centers of the stable clusters, a hierarchical classification method. The dendrogram is built according to Ward's aggregation criterion which has the advantage of accounting for the size of the elements to classify as a weight in the classification of loss of variance through aggregation. The final partition of the population is defined by cutting the dendrogram at a suitable level identifying a smaller number K ($K < G$) of clusters. At the third stage, to improve the partition obtained by the hierarchical procedure, a so called consolidation procedure is performed. It consists of applying the method of clustering around moving centers to the entire population searching for K clusters and using as starting points the centers of the partition identified by cutting the dendrogram.

Multiobjective GA (MOGA) based clustering strategy

This clustering scheme combines the K -means algorithm and the multiobjective genetic clustering technique, which simultaneously optimizes more than one objective function for automatically partitioning data set. In multiobjective optimization search is performed over a number of, often conflicting, objective functions giving rise to a set of optimal solutions, instead of one optimal solution, known as Pareto-optimal solution. Pareto optimality usually admits a set of solutions called *non-dominated* solutions. In our study we apply first the K -means algorithm to the entire population to search for a large number G of small homogeneous clusters. Then, only the centers of those clusters resulting from the previous step undergo the multiobjective genetic algorithm. Therefore, each center represents a unit to cluster and enters in the analysis along with a weight (mass) corresponding to the number of original units belonging to the group it represents. The entire mass of the subpopulation consisting of center-units is the total number of units. Real-coded multiobjective genetic algorithms is applied at this stage to the subpopulation of center-units in order to determine the appropriate cluster centers and the corresponding membership matrix defining a partition of the units into K ($K < G$) clusters. Non-Dominated Sorting Genetic Algorithm II (NSGA-II) proposed by Bandyopadhyay et al. (2007) as clustering procedure, is used. The Xie-Beni (XB) index and FCM measure are taken as the two objective functions that need to be simultaneously optimized. Since NSGCA-II is applied to the data set formed by the G centers obtained from K -means algorithm, XB and FCM indices are modified to take into account the weight of each center-unit to cluster. The XB index is defined as

$XB = W / n \times sep$ where $W = \sum_{k=1}^K \sum_{i=1}^G u_{ik}^2 p_i d^2(\mathbf{x}_i, \mathbf{c}_k)$ is the within-clusters deviance in which the squared Euclidean distance $d^2(\mathbf{x}_i, \mathbf{c}_k)$ between unit $\mathbf{x}_i$ and center $\mathbf{c}_k$ is weighted by the

mass p_i of $\mathbf{x}_i$ and u_{ik} is the membership value of unit i to cluster k ; $n = \sum_{i=1}^G p_i$ and $sep = \min_{k \neq h} \{d^2(\mathbf{c}_k, \mathbf{c}_h)\}$ is the minimum separation of the clusters. The FCM measure is defined as $FCM = W$, having chosen the fuzzy exponent equal to 2. The evolutionary operators implemented by NSGA-II are the crowded binary tournament selection, conventional crossover and mutation. Moreover, the elitist strategy is implemented. For a description of the different genetic processes refer to Deb et al. (2002). The near-Pareto-optimal chromosomes of the last generation provide the different solutions to the clustering problem. As the multiobjective genetic algorithm generates a set of Pareto optimal solutions, the solution producing the best PBM (or J) index (Bandyopadhyay et al., 2007) is chosen. Therefore, the centers encoded in this optimal chromosome are extracted and each original unit is assigned to the group with the nearest centroid in terms of squared Euclidean distance.

3. Main results

In order to compare the above-mentioned clustering strategies for large data set, we have applied them on a real-life data set. We have examined the INPS administrative archive related to the special fund for self-employed workers, called para-subordinate, where the periodical payments made from company for its employees are recorded. The data set consists of about 1.5 million records ($n=1528865$) where each record represents a worker and the J variables are the result of specific aggregations, considered more suitable, of the original variables (Mingo, 2009). The variables refers to the workers' objective conditions: social and demographic characteristics, working characteristics.

A two-step sequential, tandem approach was adopted to perform the analysis. In the first step a low-dimensional representation of the variables was obtained via Multiple Correspondence Analysis. We have chosen to perform the cluster analysis in the space of the first five factors, that explain about 38% of inertia and 99.6% of revaluated inertia. In the second step, the three clustering strategies presented above were used to identify a set of relatively homogenous workers' groups on the basis of the low-dimensional data. For K -means algorithm the maximum number of iterations was fixed to be 200. Fixed the number of classes K ($K=5, \dots, 9$) to search for, the best solution in terms of objective function in 100 different runs of the K -means algorithm was retained to prevent the algorithm from falling in local optima due to the starting solutions. For mixed clustering strategy, in order to identify stable clusters, 4 different partitions around 10 different centers were performed. In this way, 4^{10} stable groups are potentially achievable. Since many of these were empty, the stable groups that undergo the hierarchical method were 281. Then, consolidation procedures were performed using as starting points the centers of different partitions with the number of clusters ranging from 5 through 9 identified by cutting the dendrogram at several levels.

The parameters of MOGA based clustering strategy were fixed as follows: 1) for K -means algorithm applied at the first stage, the number of clusters to search for was 500; 2) NSGA-II, which was applied at the second stage to a data set of 500 center-units, was implemented with number of generations=150, population size=100, crossover probability=0.8, mutation probability=0.01. NSGA-II was run by varying the number of clusters to search for from 5 to 9.

Performances of the clustering strategies were evaluated using the PBM index as well as the Calinski-Harabasz (VRC) (Calinski and Harabasz, 1974) and Davies-Bouldin (DB) (Davies and Bouldin, 1979) indexes (Table 1).

Table 1: *Validity index values of several clustering solutions*

Index	Strategy	NUMBER OF CLUSTERS				
		5	6	7	8	9
PBM	K-means	4.3959	5.7641	7.0831	6.6677	6.5378
	Mixed clustering	4.4010	5.7886	7.0855	6.6868	6.5648
	MOGA based clustering	4.3963	5.7644	5.4627	4.7711	4.5733
VRC	K-means	6.9003	7.7390	7.3870	7.2495	7.2858
	Mixed clustering	6.9004	7.7295	7.3772	7.2465	7.2824
	MOGA based clustering	6.9003	7.7390	7.3007	6.8391	6.2709
DB	K-means	1.0257	0.9564	1.0554	1.0656	1.0495
	Mixed clustering	1.0253	0.9470	1.0451	1.0605	1.0438
	MOGA based clustering	1.0257	0.9558	0.9862	1.1014	1.3375

VRC and DB index values suggest the partition in six clusters as the best partitioning solution for all the strategies. The comparison of the partitions obtained by each strategy points out that they achieve rather similar results - also confirmed by MRand index (Hubert and Arabie, 1985) which is always greater than 0.97 - leading to a grouping having substantive meanings. A brief interpretation of the 6 clusters identified by the mixed clustering strategy along with the related percentage of coverage of each group in every strategy is displayed in table 2. Despite global performances of each strategy are found not to differ significantly, both mixed and MOGA strategy perform equally favorably in terms of computation time than K-means algorithm.

Table 2: *Substantive meanings of clusters and coverage in each clustering strategy*

Clusters	Mixed	K-means	MOGA
1: Young people with insecure employment	30.8%	30.9%	31.7%
2: People with more than a job	12.4%	12.7%	11.2%
3: People with permanent insecure employment	18.6%	18.3%	18.9%
4: Qualified young adults between insecurity and flexibility	15.6%	15.6%	15.5%
5: Strong flexible workers	15.6%	15.5%	15.1%
6: Flexible Northern managers	7.0%	7.0%	7.6%

References

- Bandyopadhyay S., Maulik U., Mukhopadhyay A. (2007) Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 45 (5), 1506 - 1511.
- Calinski R.B., Harabasz J. (1974) A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3, 1 - 27.
- Davies D. L., Bouldin D. W. (1979) A Cluster Separation Measure, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1, 224 - 227.
- Deb K., Pratap A., Agarwal S., Meyarivan T. (2002) A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, *IEEE Trans. Evol. Comp.*, 6 (2), 182 - 197.
- Hubert L., Arabie P. (1985) Comparing partitions, *Journal of Classification*, 2, 193-218.
- Lebart L., Morineau A., Piron M. (2004) *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- Mingo I. (2009), *Concetti e quantità, percorsi di statistica sociale*, Bonanno Editore.