MOEA for Clustering: Comparison of Mutation Operators

Oliver Kirkland
School of Computing Sciences
University of East Anglia
Norwich, UK
O.Kirkland@uea.ac.uk

Beatriz de la Iglesia
School of Computing Sciences
University of East Anglia
Norwich, UK
B.Iglesia@uea.ac.uk

ABSTRACT

Clustering is an important task in data mining. However, there are numerous conflicting measurements of what a good clustering solution is. Therefore, clustering is a task that is suitable for a Multi-Objective Evolutionary Algorithm. Mutation operators for these algorithms can be designed to explore a diverse range of solutions or focus upon individual solution quality. We propose using a hybrid technique that generates a wide range of solutions and then improves them with respect to the data. We create an experimental set-up to assess mutation operators with respect to Pareto front quality. Using this set-up we find that mutation operators that mutate solutions with respect to the data perform better but hybrid mutation techniques show promise.

Keywords

Clustering, Multi-Objective Evolutionary Algorithms, Mutation Operators, Adaptive Techniques, Experimental Comparison

1. INTRODUCTION

Numerous clustering algorithms have been proposed. All of the algorithms try to minimise or maximise some property of the clustering solution. Many Cluster Quality Measures (CQM) have been proposed [11, 16]; to date there has been no consensus on what measures are best.

In an MO environment, the CQMs become the objectives for an MO Clustering algorithm. In this work we use Multi-Objective Evolutionary Algorithms to solve clustering problems. In this context, is not only important to obtain good individual clusterings but also a good Pareto front, showing good spread and convergence.

2. MO CLUSTERING ALGORITHM

The algorithm we use is built upon NSGA-II [3], it is similar to one we have presented previously [6]. We use a Centroid Based Real Encoding (CBRE) as the representation, a uniform crossover operator and, for the fitness functions, we use a measure of heterogeneity and the Connectivity measure [2]. A CBRE is a set of prototype cluster centroids. A varied number of prototypes are randomly copied from the data set to form initial solutions. We investigate the performance of three mutation operators, they are as follows:

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

The Randomness Mutation (RM) operator comprises three tasks which are performed with an equal probability. First, the number of prototypes may be decreased. For this, the closest pair of prototypes are found and the prototype that has the next closest neighbour is removed from the solution. Second, the number of prototypes may be increased. A new prototype is drawn from the data set by selecting the object that is furthest away from any prototype. Third, the prototypes may be modified. For each dimension of each prototype there is a 0.05 chance that it may be modified [10] by adding a negative or positive value scaled to that dimension of the data set [1].

The K-Means Like Mutation (KMLM) is an iteration of K-Means can be used as a mutation operator by recalculating the cluster prototypes [7].

The Hybrid Mutation (HM) operator combines the two previous operators with a linearly varying probability of application. Initially, RM is used to explore the solution space; later in the search, KMLM is used to refine solutions.

3. EXPERIMENTATION

Two clustering solutions may be compared to each other using the Rand Index [13]. If one of the clusterings is the preferred clustering solution, the value of the Rand Index is an indication of the quality of the clustering with respect to the known solution.

For comparing the quality of Pareto fronts generated by MO approaches, a number of measures can be used. The volume of the objective space covered by a Pareto front may be used as an indication of quality. Fronts that dominate a high volume are regarded as better [15].

A Pareto front can be said to be good if the solutions that form it are evenly spread [8, 4], that is, solutions are not clumped around local optima. To assess this we use the Spread, or diversity, measure [12, 3].

If the solutions within a Pareto front are among the optimal front, the front is considered good. This can be measured with the Generational Distance (GD) [5], which measures the Euclidean distance between each solution and its nearest neighbour in the optimal front. However, a front may only cover some of the optimal front. The Inverted Generational Distance (IGD) [9] is an alternative that gives extra weight to extreme solutions.

We use the information entropy of the population as a measure of the diversity of a given Pareto front [14]. Diverse fronts are desirable as they contain solutions that are not similar to each other.

As a benchmark we run 100 instances of the K-Means

algorithm in parallel for 100 iterations. Therefore, the solutions present at the first iterations are equivalent to the first generation of a genetic algorithm and so on which allows us to construct Pareto fronts to observe changes as the algorithms progress. Each run of NSGA-II starts with identical populations to ensure differences in performance are not due to the generation of the start populations. We also supply the same centroids to the K-Means instances. We use six popular benchmark data sets drawn from the UCI Machine Learning Repository[1] and three synthetic datasets. For each of dataset we perform 100 executions and calculate the arithmetic mean of the measures of assessment for each generation of execution.

4. RESULTS

In the majority of cases the volume dominated increases as the algorithms progresses. KMLM increases at the quickest rate but in most cases HM and KMLM converge to similar values. KMLM is often slightly higher. K -Means improves initially but then stabilises. RM also stops improving early on and in some cases the dominated area decreases.

KMLM shows continuous improvement of GD. Initially we observed that K-Means performs similarly to KMLM, but it does not continue to improve with more generations. RM showed the worst performance. Using IGD we found that the worst performance was that of K-Means and RM which are very similar. The performance of HM and KMLM are less similar when measured with IGD. These results imply that for some data sets KMLM is more effective at locating the extreme ends of the optimal front.

Solutions generated using RM become spread quickly and stay this way. In the majority of cases HM and KMLM are less spread and sometimes behave erratically indicating significant changes within the population during execution.

The starting populations are initially diverse when measured by Entropy, the implementations of NSGA-II immediately become significantly less diverse before quickly regaining diversity before becoming stable. RM and HM made less diverse populations than the solutions using by KMLM.

The average similarity of the clustering solutions generated from the solutions in the Pareto fronts to the intended solution varies significantly. Variation is related to the data set more than any other factor. As the algorithms progress solutions found using RM often become worse, while those found using KMLM and HM generally improve.

5. CONCLUSIONS

We saw that KMLM performs best for Pareto front quality by a number of measures. However, we have also seen that these solutions are not always evenly spread and that the similarity to the clustering solutions that we desired varies from data set to data set. We postulate that KMLM is performing some form of local search on specific solutions which leads to faster convergence to the optimal Pareto front.

We have also found that using RM produced poorer Pareto fronts that contain a large quantity of unique and varied solutions that are evenly spaced. RM appears to offer some advantages in terms of diversity.

HM did not represent a good improvement with respect to KMLM. A better operator or combination of operators that consistently delivers diverse and improved Pareto fronts needs to be found. We expect that some form of adaptive mechanism which switches the emphasis of the search from the quality of the Pareto front to the quality of individual solutions or to the diversity in the population may present advantages for this problem.

[1] `http://archive.ics.uci.edu/ml/datasets.html`

6. REFERENCES

[1] S. Bandyopadhyay and U. Maulik. An evolutionary technique based on k-means algorithm for optimal clustering in $\mathbb{R}^n$. *Information Sciences*, 146:221–237, 2002.

[2] E. Chen and F. Wang. Dynamic clustering using multi-objective evolutionary algorithm. *Computational Intelligence and Security*, pages 73–80, 2005.

[3] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182, 2002.

[4] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A. Fast, and E. Algorithm. NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2), 2002.

[5] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. Scalable test problems for evolutionary multiobjective optimization. *EMO*, pages 105–145, 2005.

[6] O. Kirkland, V. Rayward-Smith, and B. de la Iglesia. A novel multi-objective genetic algorithm for clustering. *IDEAL 2011*, pages 317–326, 2011.

[7] K. Krishna and N. Murty. Genetic K-means algorithm. *IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics*, 29(3):433–439, 1999.

[8] S. Lee, P. von Allmen, W. Fink, A. Petropoulos, and R. Terrile. Comparison of multi-objective genetic algorithms in optimizing q-law low-thrust orbit transfers. In *GECCO Conference Late-breaking Paper, Washington, DC*, 2005.

[9] H. Li and Q. Zhang. Multiobjective optimization problems with complicated pareto sets, MOEA/D and NSGA-II. *IEEE Trans. Evol. Comput.*, 13(2):284–302, 2009.

[10] U. Maulik and S. Bandyopadhyay. Genetic algorithm-based clustering technique. *Pattern recognition*, 33(9):1455–1465, 2000.

[11] G. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[12] A. Nebro, F. Luna, E. Alba, B. Dorronsoro, J. Durillo, and A. Beham. Abyss: Adapting scatter search to multiobjective optimization. *IEEE Trans. Evol. Comput.*, 12(4):439–457, 2008.

[13] W. Rand. Objective criteria for the evaluation of clustering methods. *JASA*, 66(336):846–850, 1971.

[14] A. Reynolds and B. de la Iglesia. Managing population diversity through the use of weighted objectives and modified dominance: An example from data mining. In *IEEE MCDM*, pages 99–106. IEEE, 2007.

[15] A. Reynolds and B. De la Iglesia. A multi-objective grasp for partial classification. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 13(3):227–243, 2009.

[16] L. Vendramin, R. Campello, and E. Hruschka. On the comparison of relative clustering validity criteria. In *Proc. SIAM Internat. Conf. on Data Mining, Sparks, USA*, volume 733–744, 2009.