Flat vs. Symbiotic Evolutionary Subspace Clusterings

Ali Vahdat
Faculty of Computer Science
Dalhousie University
Halifax, Canada, B3H 4R2
vahdat@dal.ca

Malcolm I. Heywood
Faculty of Computer Science
Dalhousie University
Halifax, Canada, B3H 4R2
mheywood@dal.ca

ABSTRACT

Subspace clustering coevolves the attribute space supporting clusters at the same time as parameterizing the cluster location and combination. Typically, a 'flat' representation is pursued in which individuals describe both the property of individual clusters as well as the combination of clusters used to define the overall solution; hereafter F-ESC. Conversely, a symbiotic approach was recently proposed in which candidate clusters and the combination of clusters are coevolved from independent populations; hereafter S-ESC. In this work a common framework is pursued in order for flat and symbiotic evolutionary subspace clustering to be compared directly. We show that F-ESC might match S-ESC results for data sets with high proportions of cluster support, however, the gap between the two algorithm increases as cluster support decreases.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering—*subspace clustering, evolutionary clustering*

General Terms

Algorithms, Experimentation, Performance

Keywords

Subspace Clustering, Symbiosis, Multi-Objective Optimization

1. INTRODUCTION

Application domains with large attribute spaces – such as genomics and text analysis – potentially benefit from both locating the specific attribute support for clusters as well as the cluster location. However, filter and wrapper approaches generally assume that the goal is to identify a single common attribute subspace. Conversely, subspace clustering algorithms attempt to identify both clusters and subspaces simultaneously, potentially resulting in unique attribute subspaces for each cluster [1].

Recently a symbiotic evolutionary multi-objective optimization – Symbiotic Evolutionary Subspace Clustering or S-ESC – was proposed [3]. Benchmarking against well known non-evolutionary subspace clustering algorithms (MINECLUS, PROCLUS, and STATPC) demonstrated its effectiveness. In this work we are interested in understanding how much of the performance from the S-ESC architecture is contributed by the symbiotic component of the framework.

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

2. FROM S-ESC TO F-ESC

The two-level representation of S-ESC is simplified (and condensed) into a flat single-level representation. Here, each individual encodes all k cluster centroids necessary for a partitioning, therefore the chromosome is composed of k cluster centroids laid out into a single string of integer pairs. The same limits are set for F-ESC with respect to minimum and maximum number of clusters in a data set and attributes per cluster. The range [2,20] is selected for both constraints, which means that a single individual's length can vary between 4 and 400 in length.

The second main difference between S-ESC and F-ESC is the use of a crossover operator replacing the single-level mutation operator utilized in S-ESC. The single-level mutation operator in S-ESC is responsible for removing / adding / swapping cluster centroids (symbionts) from / to / between clustering solutions (hosts). The crossover operator in F-ESC essentially performs the same modification between two flat individuals. The variation operator is a 2-point crossover operator which swaps one or more cluster centroids from parent a with one or more cluster centroids from parent b. There is also a repair mechanism to make sure the offspring satisfy the cluster limit constraints.

The remaining components and sub-components of F-ESC are the same as S-ESC [3]. Grid generation is the preprocessing component with its output being the genetic material used by evolutionary process. F-ESC employs the same EMO algorithm utilizing both compactness and connectivity objectives. The selection operator is a tournament process between four individuals where the best two individuals are retained as parents; thus elitist. The same mutation operators are implemented to remove and add attributes to a randomly selected cluster centroid within an individual with a third mutation operator modifying the 1-d centroid of a randomly selected attribute within a cluster centroid.

To account for robustness to data set cardinality, a subsampling process is used. Thus, some subset of exemplars are randomly selected anew at each generation and individuals' objectives are evaluated against this set instead of the whole data set. Once the evolutionary process provided a pool of solutions – the Pareto front – the procedure as suggested by [2] identifies the solution at the 'knee' as the champion solution.

3. RESULTS

In this section we present and compare results from S-ESC and F-ESC on five synthetically-generated data sets that were used for S-ESC evaluations in [3]. They range from 50 to 1000 dimensions, 1300 to 3800 instances, 4 to 10 clusters, 20% to 90% irrelevant attributes, 10 to 100 attributes per cluster and 3% to 43% coverage per cluster. The data generation routine is explained in [3]. In all cases performance will be measured relative to the single champion individual returned following the common post processing step of knee detection. Micro F-measure will then characterize solution quality. Such a measure is adopted as it penalizes under / over estimation of cluster count as well as measuring the purity of the clusters.

Rather than using S-ESC parameter values for F-ESC, experiments are performed to optimize evolutionary parameter setting for F-ESC; whereas the S-ESC parameterization follows from the previous study [3]. These parameters include crossover and mutation probabilities, and the population size. S-ESC does not use the crossover operator and relies only on mutation operators. The probability of mutation is 100% in S-ESC meaning that all individuals go through the mutation process. For F-ESC on the other hand we run four experiments to identify the best probabilities for crossover and mutation rates.[1] The best results were achieved when both crossover and mutation probabilities are set to 100%. For the sake of brevity we will not present any plots for this experiment.

With regards to population size, the situation is somewhat different. S-ESC uses two distinct populations in which the host population is a fixed size and the symbiont population size varies; whereas F-ESC assumes a single fixed size population. In order to characterize an appropriate F-ESC population size, S-ESC is applied to a sample of the benchmarking tasks and the variation in symbiont population size recorded. There is some variation between the profiles, however, the minimum and maximum bounds are very close. Symbiont population size for the 200-d data set varies between 284 and 392, while more complex data set, 800-d, utilizes between 297 and 436 symbionts. Similar trends can be observed for the other three data sets. In general symbiont population size never exceeded 4.5 times the host population size (100 in all experiments).

With this in mind, F-ESC benchmarking will assume three different values for population size: 100, 200 and 500. Figure 1 illustrates the F measures of F-ESC with different population sizes, and also compares them against S-ESC with 100 hosts and a dynamic symbiont population limited to a maximum of 500 symbionts. Each algorithm is run 50 times on each data set and one knee solution is selected as the champion solution for each run. The top of each bar defines the median of a distribution composed by 50 solutions selected from 50 runs of the algorithm on each data set. Also first and third quartiles are shown in the form of error bars.

As anticipated the overall performance of F-ESC improves as the population size increases, however, this comes with the price of increased computational expense. Moreover, there does not appear to be any statistical significance to the variation in population size for F-ESC.

[1]In both the cases of S-ESC and F-ESC the mutation rates are deterministically annealed, stopping when a negative outcome is encountered.

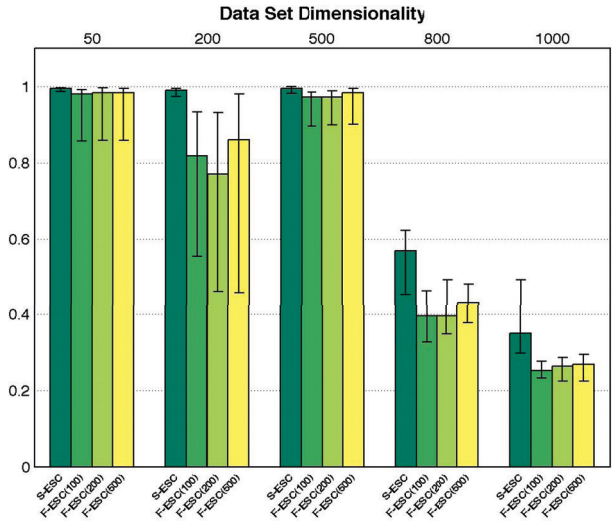

Figure 1: F-measure comparison between S-ESC and F-ESC with different population sizes.

Under the easier data sets (corresponding to the three smaller data dimensions of Figure 1) S-ESC maintains a very tight distribution as well as better median F measure. The 200-d data set is embedded with more challenging cluster configurations than the 500-d data set, resulting in further degradation of the the F-ESC results. S-ESC continues to perform significantly better (as measured under the student T-Test metric) for the remaining 800-d and 1000-d data sets.

In summary, the capacity for symbiotically coevolving solutions under S-ESC in the sub-space clustering task represents a significant advantage in this domain. Future benchmarking might provide further insight as to what particular mechanisms might be responsible for this result.

Acknowledgments

MITACS and NSERC funding is gratefully acknowledged.

4. REFERENCES

[1] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6:90–105, June 2004.

[2] O. Schutze, M. Laumanns, and C. A. Coello Coello. Approximating the knee of an MOP with stochastic search algorithms. In *Parallel Problem Solving from Nature*, volume 5199 of *LNCS*, pages 795–804, 2008.

[3] A. Vahdat, M. Heywood, and A. Zincir-Heywood. Symbiotic evolutionary subspace clustering. In *IEEE Congress on Evolutionary Computation*, pages 2724–2731, june 2012.