Improved Niching and Encoding Strategies for Clustering Noisy Data Sets

Olfa Nasraoui and Elizabeth Leon

Department of Electrical and Computer Engineering, The University of Memphis
Memphis, TN 38152
{onasraou, eleon}@memphis.edu

Abstract. Clustering is crucial to many applications in pattern recognition, data mining, and machine learning. Evolutionary techniques have been used with success in clustering, but most suffer from several shortcomings. We formulate requirements for efficient encoding, resistance to noise, and ability to discover the number of clusters automatically.

1 Introduction

The Need for a Robust Fitness Measure: Most existing evolutionary clustering techniques, such as [1], and [2], rely on a fitness that is based on a Sum of Squared Errors that is sensitive to noise because it increases indefinitely with distance. A *robust* fitness function can resist noise, for example by weighting the points' contributions by a robust weight function that decreases the influence of outliers.
The Need for a Scalable Chromosome Encoding: The chromosome in most existing evolutionary clustering techniques either encodes a possible partition of the entire data set, or encodes all the cluster prototypes. The former encoding leads to an explosion of the search space size as the data set gets larger. The latter assumes a known number of clusters and leads to a search space size that explodes exponentially with the number of clusters. A scalable encoding that is independent of the number of clusters and the size of the data, encodes a single cluster prototype in each chromosome.
The Need for Niching and Automatic Niche Size Estimation: An optimal single cluster encoding strategy will cause the fitness to have a different mode for each cluster. Therefore, niching methods are required. As in nature, niches in our context correspond to different subspaces of the environment (clusters) that can support different types of life (data samples).

2 The Unsupervised Niche Clustering and Comparison to Existing Evolutionary Clustering Techniques

The Unsupervised Niche Clustering (UNC) [3] is a recent approach to evolutionary clustering. UNC uses a chromosome representation encoding a single cluster prototype, and optimizes a density based fitness function that reaches a maximum at every good cluster center, hence requiring a niching strategy. A hybrid scale updating strategy is used to estimate the niche sizes reliably, and thus improve the niching. Because UNC

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 1324–1325, 2004.

uses robust weights in its cluster fitness definition, it is less sensitive to the presence of noise. Furthermore, the combination of the single-cluster chromosome encoding with niching offers a simple and efficient approach to automatically determine the optimal number of clusters.

Table 1 compares some evolutionary clustering techniques, including UNC.

Table 1. Comparison of UNC with Other Evolutionary Clustering Algorithms for data of size N, population of size N_P, and C clusters

Approach →	*UNC* [3]	*GGA* [1]	*Lee* [2]	*G-C-LMedS* [4]	*k-d-Median* [5]
Search Method	GA	GA	ES	GA	GA
Robustness to noise	yes	no	no	yes	yes
Automatic Scale Estimation	yes	no	no	no	no
Complexity per Generation	$O(NN_P)$	$O(CNN_P)$	$O(CNN_P)$	$O(CNN_P)$	$O(N_P CN \log(N))$
Hybrid	yes	no	no	no	yes
Does not require No. of Clusters	yes	no	yes	no	no
Handles ellipsoidal clusters	yes	no	no	no	no
Density/Partition	Density	Partition	Partition	Partition	Partition

3 Conclusion

Most existing clustering techniques necessitate the derivation of the optimal prototypes by differentiation to guarantee convergence to a *local* optimum, which can be impossible for most subjective and non-metric dissimilarity measures. For this reason, Evolutionary clustering methods are preferrable. Unfortunately most evolutionary clustering techniques are sensitive to noise, and assume a known numer of clusters. We summarized requirements to ensure efficient encoding, resistance to noise, and ability to discover the number of clusters automatically.

Acknowledgment. This work is supported by a National Science Foundation CAREER Award IIS-0133948 to O. Nasraoui.

References

1. L. O. Hall, I. O. Ozyurt, and J. C. Bezdek, "Clustering with a genetically optimized approach," *IEEE Trans. Evolutionary Computations*, vol. 3, no. 2, pp. 103–112, July 1999.
2. C.-Y. Lee and E. K. Antonsson, "Dynamic partitional clustering unsing evolution strategies," in *3rd Asia Pacific Conf. on simulated evolution and learning*, Nagoya, Japan, 2000.
3. O. Nasraoui and R. Krishnapuram, "A novel approach to unsupervised robust clustering using genetic niching," in *Ninth IEEE International Conference on Fuzzy Systems*, San Antonio, TX, May 2000, pp. 170–175.
4. O. Nasraoui and R. Krishnapuram, "Clustering using a genetic fuzzy least median of squares algorithm," in *North American Fuzzy Information Processing Society Conference*, Syracuse NY, Sep. 1997.
5. V. Estivill-Castro and J. Yang, "Fast and robust general purpose clustering algorithms," in *Pacific Rim International Conference on Artificial Intelligence*, 2000, pp. 208–218.