A Novel Mixture-Model Cluster Analysis With Genetic EM Algorithm and Information Complexity as the Fitness Function

James E. Wicker[1], **Hamparsum Bozdogan**[2], and Halima Bensmail[3]

[1] Department of Physics and Astronomy,
University of Tennessee,
Knoxville, Tennessee 37996-0532 U.S.A.

[2] Department of Statistics,Operations, and Management Science
University of Tennessee,
Knoxville, Tennessee 37996-0532 U.S.A.

[3] Department of Statistics,Operations, and Management Science
University of Tennessee,
Knoxville, Tennessee 37996-0532 U.S.A.

Abstract. Methods of clustering based on Genetic Algorithms (GAs) offer new and novel approaches to analyzing complex multivariate data to determine the patterns and the number of clusters in the data. The traditional methods based on the iterative EM algorithm can have poor accuracy in parameter estimation, especially when the clusters overlap and show complex covariance structures. To improve this performance,in this paper, for the first time, we introduce a new and novel Genetic EM (GEM) Algorithm that does not show strong dependence on initial conditions and utilizes the global search properties of genetic algorithms (GAs) to accurately estimate model parameters describing the multivariate clusters. Our method is flexible to use Genetic K-Means (GKM) or Genetic Regularized Mahalanobis (GARM) distances to compute the initial cluster parameters, with little difference in the final results. This innovation allows our algorithm to find optimal parameter estimates of complex hyperellisoidal clusters. We develop and score the information complexity (ICOMP) criterion of Bozdogan (1994a,b, 2004) as our fitness function to choose the number of clusters present in the data sets and compare our results with other model selection criteria.

We develop the GEM approach to handle both normal (Gaussian) and non-normal (non-Gaussian) large dimensional heterogeneous data sets using adaptive multivariate mixtures of kernels of Bensmail and Bozdogan (2006a,b) to achieve flexibility in currently practiced mixture-model clustering techniques and to obtain both accurate classification error rates and accurate cluster parameter estimates.

We will highlight and demonstrate our approach on simulated data sets with different configurations and on complex real data sets collected from astronomical observations. We believe that these new methods can be implemented in data mining applications and in complex pattern recognition problems in many cross-disciplinary fields.

References

BOZDOGAN, H. (1994a): Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity. In *Multivariate Statistical Modeling, Vol. 2, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, H. Bozdogan (ed.), Kluwer Academic Publishers, Dordrecht, the Netherlands, 1994, pp. 69-113.

BOZDOGAN, H. (1994b): Choosing the Number of Clusters, Subset Selection of Variables, and Outlier Detection in the Standard Mixture-Model Cluster Analysis. Invited paper in *New Approaches in Classification and Data Analysis*, E. Diday et al. (Eds.), Springer-Verlag, New York, 1994, pp. 169-177.

BOZDOGAN, H. (2004): Intelligent Data Mining with Information Complexity, Statistical Modeling, and Genetic Algorithms: A Three-Way Hybrid. In *Statistical Data Mining Knowledge Discovery*, H. Bozdogan (Ed.), Chapman Hall/CRC, 2004, 15-56.

BENSMAIL, H. and BOZDOGAN, H. (2006a): Adaptive Multivariate Kernel Mixture-Model Cluster Analysis for Mixed Data. Working paper.

BENSMAIL, H. and BOZDOGAN, H. (2006b): Bayesian Cluster Analysis for Gaussian and Non-Gaussian Data With Missing Values. Working paper.

Keywords

MIXTURE-MODEL CLUSTER ANALYSIS, GENETIC EM ALGORITHM, INFORMATION COMPLEXITY