

Misty Mountain - A parallel clustering method. Application to fast unsupervised flow cytometry gating

István P. Sugár*, and Stuart C. Sealfon

Department of Neurology and Center for Translational Systems Biology, Mount Sinai School of Medicine, New York, NY, USA

Model based clustering requires serial clustering for all cluster numbers within a user defined interval. The final cluster number is then selected by various criteria. These supervised serial clustering methods are time consuming and frequently different criteria result in different optimal cluster numbers. We developed a new, unsupervised density contour clustering algorithm, called Misty Mountain, that is based on percolation theory and that efficiently analyzes large data sets. The approach can be envisioned as a progressive top-down removal of clouds covering a data histogram relief map to identify clusters by the appearance of statistically distinct peaks and ridges. This is a parallel clustering method that finds every cluster after analyzing only once the cross sections of the histogram.

The multi-dimensional data is first processed to generate a histogram containing an optimal number of bins by using Knuth's data-based optimization criterion. Then cross sections of the histogram are created. The algorithm finds the largest cross section of each statistically significant histogram peak. The data points belonging to these largest cross sections define the clusters of the data set. The algorithm is unbiased for cluster shape, robust to noise and fast (The clustering of 10^6 data points in 2D data space takes place within about 15 seconds on a standard laptop PC). It is unsupervised (it does not need estimation for cluster number) and the computation time linearly increases with the number of data points. Its performance with various datasets supports its reliability and utility for automating the analysis of FCM data.

This work from the Program for Research in Immune Modeling and Experimentation (PRIME) was supported by contract NIH/NIAID HHSN266200500021C.