MI tech
Multimedia Imaging Technology

university of groningen

THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE

A Variant of Adaptive Mean Shift-Based Clustering

Fajie Li and Reinhard Klette

Outline and Objective

We are interested in clustering sets of highly overlapping clusters. For example, given is an observed set of stars (considered to be a set of points); how to find (recover) clusters which are the contributing galaxies of the observed union of those clusters? Below we propose a modification of an adaptive mean shift-based clustering algorithm (called Algorithm 1) proposed in 2003 by B. Georgescu, I. Shimshoni and P. Meer.

Our Algorithm

Algorithm 2 *Locally-Adaptive Mean-Shift Clustering*

Input: *Three positive integers k, N (number of iterations) and T (threshold of the number of merged points to apply one of the traditional clustering algorithms, such as kmeans or clusterdata, as (e.g.) implemented in MATLAB), n old clusters C_i, where $i = 1, 2, \ldots, n$.*

Output: *m new clusters G_i, where $i = 1, 2, \ldots, m$.*

1: $C = \cup_{i=1}^{n} C_i$ and $S = \emptyset$
2: **for** each $\mathbf{x} \in C$ **do**
3: Let k, C, $\mathbf{x}$ and N be the input for Algorithm 1; compute an approximate local maximum of the density, denoted by $\mathbf{x}'$; and let $S = S \cup \{\mathbf{x}'\}$.
4: **end for**
5: Sort S according to lexicographic order.
6: Merge duplicated points in S into a single point. Let the resulting set be S'.
7: **if** $|S'| > T$ **then**
8: $C = S'$ and goto Step 2
9: **end if**
10: Sort S' according to the cardinalities of associated sets of points in S'.
11: Let the last m points in S' be the initial centers, apply *kmeans* to cluster S'; the resulting (new) clusters are denoted by G'_i, where $i = 1, 2, \ldots, m$.
12: **for** each $i \in \{1, 2, \ldots, m\}$ **do**
13: Output $G_i = (\cup_{\mathbf{x}' \in G'_i} S'_{\mathbf{x}'}) \cup \{\mathbf{x}'\}$
14: **end for**

Results

We use a common test data set of simulated astronomic data; see [A.Helmi and P.T. de Zeeuw. Mapping the substructure in the Galactic halo with the next generation of astrometric satellites. Astron. Soc., 319:657-665, 2000]. Algorithm 2 ensures a mean recovery rate (see [Li & Klette, PSIVT 2009, Tokyo]) of 35.45% (using *kmeans*) or of 35.73% (using *clusterdata*). The best possible upper bound, estimated in Subsection 4.2 in [Li & Klette, MI-tech TR, 2008] for this data set, is between 39.68% and 44.71%. Thus, the obtained mean recovery rate is close to this estimated upper bound.