The problem is to partition a given data set $D = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$, into clusters $\{C_i, \ldots, C_m\}$, where the number $m$ of clusters is either given, or to be determined by an optimality criterion.

Clusters consist of similar points, and are themselves dissimilar, where similarity is in a sense of a distance $d(\cdot, \cdot)$ on $\mathbb{R}^n$.

With a cluster $C_i$, we associate a center $c_i$, and for any data point $x \in D$ we then compute:
- a distance $d(x, c_i)$, denoted by $d_i(x)$, and
- a probability $p_i(x)$ of membership in $C_i$.

We assume throughout that for all $x$,

$$p_i(x) d_i(x) = \text{constant, depending on } x, \quad \text{for } i = 1, \ldots, m, \quad (1)$$

making membership in nearby clusters more probable. Since probabilities add to 1, assumption (1) implies

$$p_i(x) = \frac{\prod_{j \neq i} d_j(x)}{\sum_{k=1}^{m} \prod_{j \neq k} d_j(x)}, \quad i = 1, \ldots, m, \quad (2)$$

in particular, for $m = 2$,

$$p_1(x) = \frac{d_2(x)}{d_1(x) + d_2(x)}, \quad p_2(x) = \frac{d_1(x)}{d_1(x) + d_2(x)}. \quad (3)$$

We present a new clustering algorithm that iterates on centers, distances and probabilities, compare it with existing methods, and illustrate its advantages.

Keywords: Distance based clustering, probabilistic clustering.

E-mail address: {iyigun,benisrael}@rutcor.rutgers.edu

RUTCOR–Rutgers Center for Operations Research, Rutgers University, 640 Bartholomew Rd., Piscataway, NJ 08854-8003, USA