DISTANCE BASED PROBABILISTIC CLUSTERING OF DATA

CEM IYIGUN AND ADI BEN-ISRAEL

The problem is to partition a given data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$, into clusters $\{\mathcal{C}_i, \ldots, \mathcal{C}_m\}$, 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 $\mathcal{C}_i$, we associate a **center** $\mathbf{c}_i$, and for any data point $\mathbf{x} \in \mathcal{D}$ we then compute:

• a **distance** $d(\mathbf{x}, \mathbf{c}_i)$, denoted by $d_i(\mathbf{x})$, and
• a **probability** $p_i(\mathbf{x})$ of membership in $\mathcal{C}_i$.

We assume throughout that for all $\mathbf{x}$,

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

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

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

in particular, for $m = 2$,

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

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,bisrael}@rutcor.rutgers.edu`

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