Center–Based Clustering and Regression Clustering

Bin Zhang
Hewlett-Packard Research Laboratories, USA

INTRODUCTION

Center-based clustering algorithms are generalized to more complex model-based, especially regression-model-based, clustering algorithms. This article briefly reviews three center-based clustering algorithms—K-Means, EM, and K-Harmonic Means—and their generalizations to regression clustering algorithms. More details can be found in the referenced publications.

BACKGROUND

Center-based clustering is a family of techniques with applications in data mining, statistical data analysis (Kaufman et al., 1990), data compression (vector quantization) (Gersho & Gray, 1992), and many others. K-means (KM) (MacQueen, 1967; Selim & Ismail, 1984), and the Expectation Maximization (EM) (Dempster et al., 1977; McLachlan & Krishnan, 1997; Rendner & Walker, 1984) with linear mixing of Gaussian density functions are two of the most popular clustering algorithms.

K-Means is the simplest among the three. It starts with initializing a set of centers $M = \{m_k \mid k = 1,...,K\}$ and iteratively refines the location of these centers to find the clusters in a dataset. Here are the steps:

K-Means Algorithm

- **Step 1:** Initialize all centers (randomly or based on any heuristic).
- **Step 2:** Associate each data point with the nearest center. This step partitions the data set into K disjoint subsets (Voronoi Partition).
- **Step 3:** Calculate the best center locations (i.e., the centroids of the partitions) to maximize a performance function (2), which is the total squared distance from each data point to the nearest center.
- **Step 4:** Repeat Steps 2 and 3 until there are no more changes on the membership of the data points (proven to converge).

With guarantee of convergence to only a local optimum, the quality of the converged results, measured by the performance function of the algorithm, could be far from its global optimum. Several researchers explored alternative initializations to achieve the convergence to a better local optimum (Bradley & Fayyad, 1998; Meila & Heckerman, 1998; Pena et al., 1999).

K-Harmonic Means (KHM) (Zhang, 2001; Zhang et al., 2000) is a recent addition to the family of center-based clustering algorithms. KHM takes a very different approach from improving the initializations. It tries to address directly the source of the problem—a single cluster is capable of trapping far more centers than its fair share. This is the main reason for the existence of a very large number of local optima under K-Means and EM when $K>10$. With the introduction of a dynamic weighting function of data, KHM is much less sensitive to initialization, demonstrated through a large number of experiments in Zhang (2003). The dynamic weighting function reduces the ability of a single data cluster, trapping many centers.

Replacing the point-centers by more complex data model centers, especially regression models, in the second part of this article, a family of model-based clustering algorithms is created. Regression clustering has been studied under a number of different names: Clusterwise Linear Regression by Spath (1979, 1981, 1983, 1985), DeSarbo and Cron (1988), Hennig (1999, 2000) and others; Trajectory clustering using mixtures of regression models by Gaffney and Smith (1999); Fitting Regression Model to Finite Mixtures by Williams (2000); Clustering Using Regression by Gawrysiak, et. al. (2000); Clustered Partial Linear Regression by Torgo, et. al. (2000). Regression clustering is a better name for the family, because it is not limited to linear or piecewise regressions.

Spath (1979, 1981, 1982) used linear regression and partition of the dataset, similar to K-means, in his algorithm that locally minimizes the total mean square error over all K-regressions. He also developed an incremental version of his algorithm. He visualized his piecewise linear regression concept in his book (Spath,

1985) exactly as he named his algorithm. DeSarbo (1988) used a maximum likelihood method for performing clusterwise linear regression. Hennig (1999) studied clustered linear regression, as he named it, using the same linear mixing of Gaussian density functions.

MAIN THRUST

For K-Means, EM, and K-Harmonic means, both their performance functions and their iterative algorithms are treated uniformly in this section for comparison. This uniform treatment is carried over to the three regression clustering algorithms, RC-KM, RC-EM and RC-KHM, in the second part.

Performance Functions of the Center-Based Clustering

Among many clustering algorithms, center-based clustering algorithms stand out in two important aspects—a clearly defined objective function that the algorithm minimizes, compared with agglomerative clustering algorithms that do not have a predefined objective; and a low runtime cost, compared with many other types of clustering algorithms. The time complexity per iteration for all three algorithms is linear in the size of the dataset *N*, the number of clusters *K*, and the dimensionality of data *D*. The number of iterations it takes to converge is very insensitive to *N*.

Let $X = \{x_i \mid i = 1,...,N\}$ be a dataset with *K* clusters, iid sampled from a hidden distribution, and $M = \{m_k \mid k = 1,...,K\}$ be a set of *K* centers. K-Means, EM, and K-Harmonic Means find the clusters—the (local) optimal locations of the centers—by minimizing a function of the following form over the *K* centers,

$$Perf(X,M) = \sum_{x \in X} d(x,M) \qquad (1)$$

where *d(x,M)* measures the distance from a data point to the set of centers. Each algorithm uses a different distance function:

(a) K-Means: $d(x,M) = \underset{1 \le k \le K}{MIN}(\| x - m_k \|)$, which makes (1) the same as the more popular form

$$Perf_{KM}(X,M) = \sum_{k=1}^{K} \sum_{x \in S_k} \| x - m_k \|^2, \qquad (2)$$

where $S_k \subset X$ is the subset of *x* that are closer to m_k than to all other centers (the Voronoi partition).

(b) EM: $d(x,M) = -\log\left(\sum_{k=1}^{K} p_k * \frac{1}{(\sqrt{\pi})^D} EXP(-\| x_i - m_l \|^2) \right)$, where $\{p_k\}_1^K$ is a set of mixing probabilities.

A linear mixture of *K* identical spherical (Gaussian density) functions, which is still a probability density function, is used here.

(c) K-Harmonic Means: $d(x,M) = \underset{1 \le k \le K}{HA}(\| x - m_k \|^p)$, the harmonic average of the *K* distances,

$$Perf_{KHM}(X,M) = \sum_{x \in X} \frac{K}{\sum_{m \in M} \frac{1}{\| x_i - m_l \|^2}}, \text{ where } p > 2. \qquad (3)$$

K-Means and K-Harmonic Means performance functions also can be written similarly to the EM, except that only a positive function takes the place where this probability function is (Zhang, 2001).

Center-Based Clustering Algorithms

K-Means' algorithms are shown in the Introduction. We list EM and K-Harmonic Means' algorithms here to show their similarity.

EM (with Linear Mixing of Spherical Gausian Densities) Algorithm

- **Step 1:** Initialize the centers and the mixing probabilities $\{p_k\}_1^K$.
- **Step 2:** Calculate the expected membership probabilities (see item <B>).
- **Step 3:** Maximize the likelihood to the current membership by finding the best centers.
- **Step 4:** Repeat Steps 2 and 3 until a chosen convergence criterion is satisfied.

K-Harmonic Means Algorithm

- **Step 1:** Initialize the centers.
- **Step 2:** Calculate the membership probabilities and the dynamic weighting (see item <C>).