On Mean Shift Clustering for Directional Data on a Hypersphere

Miin-Shen Yang[1,*], Shou-Jen Chang-Chien[1], and Hsun-Chih Kuo[2]

[1] Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan
[2] Department of Statistics, National Chengchi University, Wenshan District, Taipei 11605, Taiwan
msyang@math.cycu.edu.tw

Abstract. The mean shift clustering algorithm is a useful tool for clustering numeric data. Recently, Chang-Chien et al. [1] proposed a mean shift clustering algorithm for circular data that are directional data on a plane. In this paper, we extend the mean shift clustering for directional data on a hypersphere. The three types of mean shift procedures are considered. With the proposed mean shift clustering for the data on a hypersphere it is not necessary to give the number of clusters since it can automatically find a final cluster number with good clustering centers. Several numerical examples are used to demonstrate its effectiveness and superiority of the proposed method.

Keywords: Clustering, Mean shift, Directional data on a hypersphere.

1 Introduction

In 1918, von Mises [13] first introduced a distribution of circular data. Following Watson and Williams [14] investigation of inference problems for the von Mises distribution with the construction of statistical methods on circular data, there has been increasing research on directional data and it has been applied in biology, geology, medicine, meteorology, oceanography [3,6,10,11,12].

Clustering is a useful tool for finding clusters of a data set, grouped with most similarity in the same cluster and most dissimilarity between different clusters [8,16]. Clustering methods can generally be divided into two categories, the probability model-based approach and the nonparametric approach. The nonparametric approach includes partitional and kernel-based methods, while the kernel-based approach includes supervised and unsupervised learning methods. The support vector machine is a well-known supervised kernel-based method [2] and mean shift clustering is a popular unsupervised kernel-based method [5,15]. However, the mean shift clustering is generally used for clustering numeric data.

Recently, Chang-Chien et al. [1] proposed a mean shift clustering algorithm for circular data. In this paper, we extend the mean shift clustering for directional data on a hypersphere. The three types of mean shift procedures are considered. The proposed mean shift clustering for the data on a hypersphere can automatically find a final

* Corresponding author.

L. Rutkowski et al. (Eds.): ICAISC 2014, Part II, LNAI 8468, pp. 809–818, 2014.

cluster number with good clustering centers. Several numerical examples are used to demonstrate the effectiveness and superiority of the proposed method.

2 Mean Shift Clustering for the Data on a Hypersphere

In this section, we extend Chang-Chien et al. [1] to the directional data on a hypersphere. Since Chang-Chien et al. [1] used the distance measure for angles to modify the mean shift based clustering algorithm to two dimensional directional data (i.e. circular data), we need to extend the distance measure used in Chang-Chien et al. [1] to high-dimensional directional data. For any two circular (angle) observations θ_1 and θ_2 ($\theta_2 > \theta_1$), Chang-Chien et al. [1] considered the distance between θ_1 and θ_2 with $1-\cos(\theta_2-\theta_1)$. For extending two dimensions to high dimensions, we consider high dimensional directional data as points on the unit hypersphere. Thus, the distance measure $1-x_1^T x_2$ can be used for high dimensional directional data x_1 and x_2. Now, we use this distance measure to define a kernel on high dimension directional data. Let $X=\{x_1, x_2, .., x_n\}$ be a data set on the unit hypersphere and let $H: X \to [0,1]$ be a kernel function with $H(x)=h(1-x^T x_j)$. The kernel density estimate using the kernel H is given by

$$\hat{f}_H(x)=\sum_{j=1}^{n} h(x)w(x_j)$$

where $w(x_j)$ is a weight function. By maximizing $\hat{f}_H(x)$ with the constraint $x^T x=1$, we obtain a general formula for the mode a as follows:

$$a=x=\frac{\sum_{j=1}^{n} h'(1-x^T x_j)x_j w(x_j)}{\| \sum_{j=1}^{n} h'(1-x^T x_j)x_j w(x_j) \|} \tag{1}$$

where $w(x_j)$ is a weight function and h' is the derivative of h. Throughout this paper, we have all data points with equal weights. Now, we choose a suitable kernel for equation (1). In general, the performance of the kernel density estimate depends on the bandwidth selection. However, Wu and Yang [15] gave another method to obtain good kernel density estimation where it is different from the bandwidth selection. This technique is to normalize the distance measure and then estimate the stabilization parameters. In the high dimensional directional data case, we define the kernel function $K^p: X \to [0,1]$ as follows:

$$K^p(x)=\begin{cases} \left(1-\dfrac{1-x^T x_j}{\beta}\right)^p = \left(\dfrac{\beta-(1-x^T x_j)}{\beta}\right)^p, & \text{if } 1-x^T x_j \le \beta \\ 0, & \text{if } 1-x^T x_j > \beta \end{cases} \tag{2}$$

where the parameters β and p are called the normalization and stabilization, respectively. We use the sample standard deviation of the directional data set to