Least-Squares Log-Density Gradient Clustering for Riemannian Manifolds

（リーマン多様体に対する最小二乗対数密度勾配推定クラスタリング）

2016 年 3 月

複雑理工学専攻　　芦沢未菜　47146081

指導教員：杉山　将教授

キーワード：Mode-seeking clustering, mean shift, least-squares log-density estimation,Riemannian manifold.

The aim of machine learning is to extract hidden information, e.g., rules and patterns from data. Since more and more digital data have become available, machine learning is in the limelight in recent years. This thesis considers the problem of clustering, which is one of the important tasks of unsupervised machine learning.

The task of clustering is to divide data points into disjoint groups based on their similarities. Among various different types of clustering algorithms, *mode seeking* clustering is a well-studied and practically useful approach. The advantage of mode seeking is that it does not require the prior knowledge of the number of clusters. *Mean shift* is a seminal algorithm for mode seeking clustering. It first estimates the density of given data points with kernel density estimation (KDE) and then moves the data points along the gradient of the estimated density towards the modes. Finally, the data points which converge to the same mode are given the same cluster label. Mean shift has been widely used in various applications, especially in computer vision. To further improve the clustering performance, mean shift has been extended in several directions as follows.

One extension of mean shift is to directly estimate the density gradient. In mode seeking clustering, a good density gradient estimator is needed. Mean shift obtains the density gradient estimator as the derivative of the density estimated by KDE. However, KDE tends to perform poorly in high-dimensional space. Moreover even if we can obtain a good density estimator, it does not necessarily lead to a good density gradient estimator. These facts imply that mean shift does not always work well in practice. To cope with these problems, the *least-squares log-density gradient* (LSLDG), which estimates the log-density gradient directly, was proposed. It was experimentally shown that the clustering based on LSLDG called *LSLDG clustering* (LSLDGC) outperforms the original mean shift.

Another important extension of mean shift is to handle data with some structures. In practice, for example, image data and motion data lie on a structured space such as Lie groups, Stiefel manifolds and Grassmann manifolds. These manifolds are examples of Riemannian manifolds. For such data points on a manifold, the original mean shift does not work properly since the structure of data is not taken into account. To overcome this problem, mean shift was extended to Riemannian manifolds and this extended method showed a better clustering performance in experiments.

In this thesis, we combine the above two extensions of mean shift and propose a novel mode seeking clustering algorithm for Riemannian manifolds that employs direct density gradient estimation. We experimentally demonstrate that the proposed method gives significantly better clustering performance than existing methods for data points on high-dimensional Riemannian manifolds.