第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)18—0233—03 文献标识码：A 中图分类号：TP301.6

基于 Mean-Shift 的投影聚类算法 PCMF

黄李国，王士同
(江南大学信息工程学院，无锡 214122)

摘 要：高维数据的聚类都隐含在低维的子空间内。为找出有效的子空间，Agrawal 等人提出了投影聚类概念，通过映射变换转换到子空间里，然后借助其他方法找到聚类。该文基于目前最新的投影聚类算法 EPCH，提出了 PCMF 算法，借助 Mean-Shift 划分子空间聚类。与 EPCH 算法相比，PCMF 在划分子空间中数据时，无须输入参数（EPCH 中是最大聚类个数），能够有效降低划分出的子空间数量，获得与 EPCH 相媲美的实验结果。

关键词：子空间划分；直方图；Mean-Shift；投影聚类

Projective Clustering Algorithm PCMF Based on Mean-Shift

HUANG Li-guo, WANG Shi-tong
(School of Information Engineering, Southern Yangtze University, Wuxi 214122)

【Abstract】 The clusters of a high dimensional dataset are often hidden in the subspaces of the corresponding low dimensional datasets. In order to successfully find the subspaces, Agrawal proposes the conception of projective clustering, converting the data into subspaces with mapping and using another method to find clusters. EPCH is the latest projective clustering algorithm. This paper incorporates Mean-Shift into EPCH to divide a high dimensional dataset into the corresponding subspaces. Experiments demonstrate that the approach is comparable to EPCH in the sense of obtaining the reasonable clusters, however, it doesn't require any parameter and can reduce the number of subspaces.

【Key words】 subspace partition; histograms; Mean-Shift; projective clustering

由于高维数据自身的缺点如稀疏性，使得传统的聚类算法都不适用，如文献[1]中的算法K-Means和K-Medoid，处理高维数据的效果都不理想。采用降维或特征提取来处理，则在不同的维度上都能找到一个分类，同时又会丢失某些分类，即每个维度至少涉及到一个分类。为实现有效聚类，文献[2]中提出了投影聚类方法。投影聚类是把数据集借映射变换投影到低维子空间，再用各种方法划分出子空间内的聚类，它能够有效降低数据集的维度，同时减少数据处理的复杂度。投影聚类算法有EPCH[1]、CLIQUE[2]和PROCLUS[3]和等。本文在EPCH算法的基础上，采用Mean-Shift的子空间划分方法，来找出投影子空间的聚类。与EPCH算法比较，本文提出的PCMF方法更加简单，能够有效降低找到的子空间数量，易于实现，并且聚类结果可与之相媲美。

1 算法 EPCH

算法 EPCH(efficient projective clustering by histograms)是寻找投影聚类及其相关的子空间。在构造直方图前，必须对给定的数据集进行预处理，即处理那些扭曲聚类结果的属性；其次，对于给定数据集，需要采用正交向量集把高维数据投影到低维空间上。

1.1 直方图

数据处理过程中经常用到直方图，特别是在考虑计算效率时。算法 EPCH 把直方图看成是一个近视的函数，则构造一个 d-D 直方图函数：

$$\hat{f}^{d}(x)=\frac{1}{Nh^{d}}Bin(x) \tag{1}$$

其中，N是样本数之和；$Bin(x)$是指与x在同一超立方体中所有样本数之和；h是把每个投影维度按照h划分成许多等间距的超立方体，根据Sturges规则[4]：如果数据服从正态分布，则d-D直方图在理想情况下，$h=(1+\log_2 N)^d$，本文采用较大的数来代替它，不影响实验结果，这是因为Sturges规则要保证数据分布不发生变化，而本文的目的是划分出密度区域。用 $DR_m^{s_i}$ 来标记直方图中划分出的密度区域，其中，S_i 代表某d-D直方图；m是直方图中划分出的第m个超立方体的标号。

1.2 检查密度区域

为了划分直方图中的密度区域，本文采用了不同的极限值，极限值定义如下：假设有 d-D 直方图，对应的样本投影数据为 $X=\{x_1,x_2,\ldots,x_p\}$，则均值为

$$\mu=\frac{1}{p}\sum_{i=1}^{p}x_i$$

标准差为

$$\sigma=\sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(x_i-\mu)^2}$$

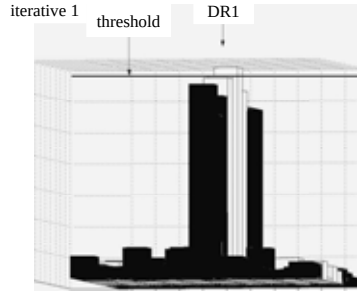
那么极限值[1]为

$$\rho=\mu+\left(\sqrt{\frac{1}{f}-1}\right)\sigma \tag{2}$$

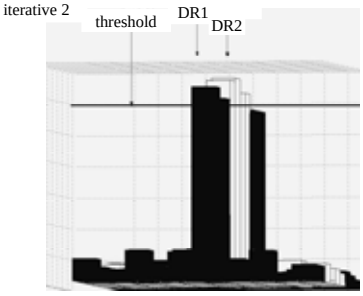
其中，f是某直方图中高度相同的超立方体个数之和的最大值与超立方体个数总和的比（$0\le f\le 1$）。

那么，如何利用极限值划分密度区域？文献[1]指出：使用通过迭代计算出的新极限值划分密度区域，直到极限值不能划分出任何密度区域为止(见图 1)。

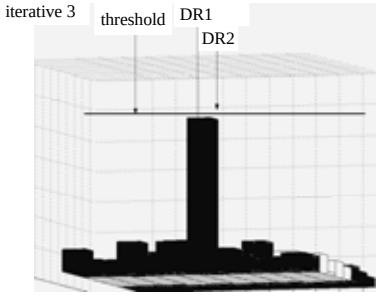
作者简介：黄李国(1979－)，男，硕士研究生，主研方向：模式识别，聚类分析；王士同，教授、博士生导师
收稿日期：2006-11-20 **E-mail：**hlg5141979@yahoo.com.cn



(a)



(b)



(c)

图 1　迭代极限值划分过程

2　基于 Mean-Shift 的投影聚类算法 PCMF

在阐述Mean-Shift划分子空间前，首先构造Mean-Shift过程需要的核密度估计函数(即影射)。假设在d-维空间R^d中给定N个数据点 $X_i(i=1,...,N)$，则X处的核密度估计函数为

$$\hat{f}(X)=\frac{1}{N}\sum_{i=1}^{N}K_H(X-X_i) \tag{3}$$

式(3)中，N是样本个数；$K_H(X)=\|\boldsymbol{H}\|^{-1/2}K(\boldsymbol{H}^{-1/2}X)$，其中，$K(X)$是多维核密度估计函数；$\boldsymbol{H}$是个 $d\times d$ 的对称正定窗体宽度矩阵，且完全参数化会增加密度估计的复杂度[4]。实际上，$\boldsymbol{H}$要么是对角阵 $\boldsymbol{H}=\mathrm{diag}[h_1^2,...,h_d^2]$，要么是$\boldsymbol{H}$=$h\boldsymbol{I}$，后者只需要一个参数$h$>0，根据式(3)，首先要找一个有效的特征空间的欧式矩阵，因此只需要输入h。那么，核密度估计的一般表达式为

$$\hat{f}(X)=\frac{1}{Nh^d}\sum_{i=1}^{N}K\left(\left\|\frac{X-X_i}{h}\right\|^2\right) \tag{4}$$

核密度估计的评价标准由密度和估计值的平方差的均值决定，实际上这个标准是个估算。在近似估计下，随着样本数量 $N\to\infty$，那么窗体宽度 $h\to0$的速率是低于N^{-1}。另外，h 可以用最近邻居规则来定义：假设 $X_{i,k}$ 是点 X_i 第 k 个最近邻居点，则 $h_i=\|X_i-X_{i,k}\|_2$。采用L_2规则是因为它适用于实验中的数据结构，要注意的是，k必须选取得足够大，保证大部分核在给定的h_i下，核密度是不断增大的[6]。由于本文中的数据都是高维的，结合上面 2 个取h的规则，采用同一h，并使h足够大以找到更多的邻居点。

实验中用到的核：

(1)Epanechnikov kernel

$$K_E(X)=\begin{cases}\frac{1}{2}c_d^{-1}(d+2)(1-\|X\|^2) & \|X\|\le1\\ 0 & otherwise\end{cases} \tag{5}$$

其中，c_d^{-1}是单个 d-维空间密度区域的数量。

(2)normal kernel

$$K_N(X)=(2\pi)^{-d/2}\exp\left(-\tfrac{1}{2}\|X\|^2\right) \tag{6}$$

用核密度函数 $f(X)$去分析特征子空间，目的是找到子空间中的密度区域中心点，而中心点却在 $\nabla f(X)=0$ 处。那么，如何得到子空间中的这些中心点？

Mean-Shift 迭代过程如下：

首先使用式(4)求给定核密度估计函数的梯度：

$$\nabla\hat{f}(X)=\frac{2}{Nh^{d+2}}\sum_{i=1}^{N}(X-X_i)K'\left(\left\|\frac{X-X_i}{h}\right\|^2\right) \tag{7}$$

令 $G(X)=-K'(X)$(