

# 自适应子空间选择方法研究

闵 锋,鲁统伟,邹 旭

(武汉工程大学 智能机器人湖北省重点实验室,湖北 武汉 430074)

**摘 要:**由于维数灾难的原因,高维空间的数据聚类是一个具有挑战性的问题。文中提出了一种自适应子空间选择的方法来解决这一难题。该方法采用局部线性嵌入的方法将高维数据映射到低维子空间上,然后采用两步迭代的方法自适应地选择最具有判别力的子空间;固定子空间不变,用K-均值聚类的方法产生类别的标号;固定类别的标号不变,用线性判别分析的方法将样本映射到低维子空间进行子空间选择。通过反复迭代,样本在低维子空间进行有效聚类而避免了维数灾难,同时子空间自适应地调整到全局最优。大量的实验结果表明,该方法聚类效果优于传统的K-均值聚类。

**关键词:**子空间选择;线性判别分析;K-均值聚类

中图分类号:TP391.41

文献标识码:A

文章编号:1673-629X(2013)10-0083-04

doi:10.3969/j.issn.1673-629X.2013.10.021

## Research on Adaptive Subspaces Selection Method

MIN Feng,LU Tong-wei,ZOU Xu

(Hubei Province Key Laboratory of Intelligent Robot,Wuhan Institute of Technology,Wuhan 430074,China)

**Abstract:**Clustering in high dimensional datasets is a challenging problem due to the curse of dimensionality. In this paper,present an adaptive subspaces selection approach to solve this problem. Datasets are projected into lower dimensional subspace through locally linear embedding. Then two iterative steps are implemented to adaptively select the most discriminative subspace;fixing the subspaces,K-means clustering is performed to generate cluster labels;fixing cluster labels,linear discriminant analysis is performed to do subspaces selection. Through iterative steps,clusters are discovered in the lower dimensional subspaces to avoid the curse of dimensionality,while the subspaces are adaptively re-adjusted for global optimality. Extensive experimental results show the benefits of the approach versus traditional K-means clustering.

**Key words:**subspaces selection;linear discriminant analysis;K-means clustering

## 0 引 言

聚类是一种常见的数据分析工具,其目的是把大量数据点的集合分成若干类,使得每个类中的数据之间最大程度的相似,而不同类中的数据最大程度的不同<sup>[1]</sup>。聚类在信息检索<sup>[2]</sup>、图像分割<sup>[3]</sup>和文本挖掘等领域有着大量的应用。通常,聚类的数据的维度是非常高的,达到几百甚至上千维,在如此高维的空间上进行聚类是一个具有挑战性的问题。分析其原因,主要有以下三点:

1)聚类的本质是一个无监督学习问题,很多有监督学习算法不能用;

2)在这样的高维空间,实例间距离会被大量的不相关属性所支配,可能导致相关属性的值很接近的实

例相距很远,聚类的结果不理想;

3)由于维数灾难的原因,维数越高,计算量越大。

对于以上问题,通常采用降维的方法将高维数据映射到低维空间进行子空间选择,然后在子空间上聚类。常用的降维方法有主成分分析(Principal Component Analysis,PCA)<sup>[4]</sup>、多维尺度变化(MultiDimensional Scaling,MDS)<sup>[5]</sup>、局部线性嵌入(Locally Linear Embedding,LLE)<sup>[6]</sup>等方法。由于子空间选择与聚类是分开进行的,所以这种方法的聚类效果并不好。

受到近年来机器学习研究成果<sup>[7-8]</sup>的启发,文中通过整合子空间选择和数据聚类,提出了一种自适应子空间选择方法。该方法通过降维将样本映射到低维子空间,用k-均值聚类(k-means clustering,KM)的方

收稿日期:2012-12-31

修回日期:2013-04-01

网络出版时间:2013-07-24

基金项目:国家自然科学基金资助项目(11001212);国家磷资源开发利用工程技术研究中心开放基金(2012 国磷 k005);武汉工程大学博士启动基金(12106021)

作者简介:闵 锋(1976-),男,湖北红安人,讲师,博士,研究方向为机器学习、计算机视觉。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130724.1007.047.html>

法产生类别的标号,然后用线性判别分析(Linear Discriminant Analysis, LDA)<sup>[9]</sup>的方法进行子空间选择,通过聚类 and 子空间选择的反复迭代,样本在低维子空间进行有效聚类而避免了维数灾难,同时子空间自适应地调整到全局最优。最后算法收敛,得到较好的聚类结果。

## 1 k-均值聚类

k-均值聚类是使用最为广泛的无监督聚类方法之一。对于一组数据  $X = \{x_1, x_2, \dots, x_n\}$ , 给定类的个数  $k$ , 将  $\{x_1, x_2, \dots, x_n\}$  中的数据分到  $k$  类中, 使得类类之间的距离最大, 而类之间的距离最小, 这里的距离用的是欧氏距离。从数学表述上说, k-均值聚类是寻找一种对数据的划分, 使得下面的目标函数最小化:

$$E(c) = \arg \min_c \sum_{j=1}^k \sum_{x_i \in c_j} |x_i - m_j|^2 \quad (1)$$

其中,  $k$  是聚类的个数;  $c_j$  是第  $j$  类的集合;  $m_j$  是第  $j$  类的集合的中心, 即均值。k-均值聚类算法简单, 实现容易, 但算法存在初始聚类数  $k$  要事先指定, 初始聚类中心选择存在随机性, 算法容易生成局部最优解, 受孤立点的影响很大等缺点。针对其不足, 有不少研究者对其做出了改进<sup>[10-11]</sup>。

## 2 线性判别分析

线性判别分析是常用的有监督学习方法, 用于数据降维和子空间选择。线性判别分析使投影后的模式样本的类间散布矩阵最大而类内散布矩阵最小, 也就是说, 投影后保证模式样本在新的空间中有最大的类间距离和最小的类内距离, 即模式样本在该空间中具有最佳的可分离性。对于一组  $d$  维的样本  $X = \{x_1, x_2, \dots, x_n\}$ , 它们分属于  $k$  个不同的类别  $C = \{c_1, c_2, \dots, c_k\}$ , 定义总体散布矩阵  $S_T$ , 类内散布矩阵  $S_W$ , 类间散布矩阵  $S_B$  如下:

$$\begin{aligned} S_T &= \sum_{i=1}^n (x_i - m)(x_i - m)^T S_W = \sum_{i=1}^k \sum_{x \in c_i} (x - m_i)(x - m_i)^T \\ S_B &= \sum_{i=1}^k n_i (m_i - m)(m_i - m)^T \end{aligned} \quad (2)$$

其中,  $m_i = \frac{1}{n_i} \sum_{x \in c_i} x$ , 是第  $i$  类的均值,  $n_i$  是第  $i$  类的个数;  $m = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^k n_i m_i$ , 是总体均值。并且  $S_T, S_W, S_B$  之间满足:  $S_T = S_W + S_B$ 。

将  $d$  维空间的原始样本  $X = \{x_1, x_2, \dots, x_n\}$  向  $k-1$  维空间投影  $Y = W^T X$ , 得到新的样本  $Y = \{y_1, y_2, \dots, y_n\}$ , 其中  $W$  是一个  $d \times (k-1)$  的矩阵。这些新的样本

自身又具有它们自己的散布矩阵, 满足:

$$\tilde{S}_W = W^T S_W W \quad \tilde{S}_B = W^T S_B W \quad (3)$$

公式(3)说明了从高维空间向低维空间投影的过程中, 类内散布矩阵和类间散布矩阵的变化与变换矩阵  $W$  相关。线性判别分析的目标是寻找一个合适的  $W$ , 使得投影后的类间距离最大, 类内距离最小, 即最大化下面的目标函数:

$$E(W) = \arg \max_W \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \arg \max_W \frac{|W^T S_B W|}{|W^T S_W W|} \quad (4)$$

容易证明, 最大化目标函数的  $W$  满足:  $S_B W = \lambda S_W W$ 。

## 3 线性判别分析指导的 k-均值聚类

用  $\text{Tr}(M)$  表示矩阵  $M$  的迹, 则 k-均值聚类的目标函数可表示为:

$$E(c) = \arg \min_c \text{Tr}(S_W) = \arg \min_c \text{Tr}(S_T - S_B) \quad (5)$$

可见 k-均值聚类的目的也是最小化类内散布矩阵  $S_W$  或者最大化类间散布矩阵  $S_B$ , 因为  $\text{Tr}(S_T)$  是一常量。由此可见, k-均值聚类和线性判别分析的目标是一致的。由于线性判别分析所挑选的子空间具有最佳的可分离性, 因而被广泛应用于子空间选择。但线性判别分析是有监督学习方法, 需要预先知道数据的类别, 对没分类的数据无能为力。而 k-均值聚类是无监督学习方法, 可以对数据进行自动聚类。因此可以将两者结合起来, 将数据聚类和子空间的划分整合在一起。用 k-均值聚类在子空间产生数据的类别, 再用线性判别分析进行子空间的选择, 如此迭代进行, 直至收敛。该方法称之为线性判别分析指导的 k-均值聚类(LDA-guided K-means, LDA-KM)。

LDA-KM 的目标是用无监督学习的方法找到具有最佳的可分离性的子空间和数据的类别, 其目标函数如下:

$$E(W, C) = \arg \max_{W, C} \frac{|W^T S_B W|}{|W^T S_W W|} \quad (6)$$

这个目标函数与线性判别分析的多出一个变量  $C$ ,  $C$  表示数据的类别。因此目标函数有两个变量  $C, W$ , 采用依次优化  $C, W$  的方法求解目标函数。

(1) 固定  $W$  不变, 优化  $C$ 。这种情况下, 目标函数为:

$$\begin{aligned} E(C) &= \arg \max_C \frac{|W^T S_B W|}{|W^T S_W W|} = \arg \max_C \frac{\text{Tr}(W^T (S_T - S_W) W)}{\text{Tr}(W^T S_W W)} = \arg \max_C \left( \frac{\text{Tr}(W^T S_T W)}{\text{Tr}(W^T S_W W)} - 1 \right) \end{aligned} \quad (7)$$

因为  $\text{Tr}(\mathbf{W}^T \mathbf{S}_T \mathbf{W})$  与  $\mathbf{C}$  无关,所以目标函数变为:

$$\begin{aligned} \arg \min_{\mathbf{C}} \text{Tr}(\mathbf{W}^T \mathbf{S}_W \mathbf{W}) &= \arg \min_{\mathbf{C}} \text{Tr}(\sum_{i=1}^k \sum_{x \in c_i} \mathbf{W}^T (x - \\ m_i)(x - m_i)^T \mathbf{W}) &= \arg \min_{\mathbf{C}} (\sum_{i=1}^k \sum_{x \in c_i} \|\mathbf{W}^T x - \\ \mathbf{W}^T m_i\|^2) \end{aligned} \quad (8)$$

由此可见,优化  $\mathbf{C}$  的值就是用  $k$ -均值聚类在子空间  $\mathbf{Y} = \mathbf{W}^T \mathbf{X}$  聚类的结果。

(2) 固定  $\mathbf{C}$  不变,优化  $\mathbf{W}$ 。这种情况下,目标函数与线性判别分析的目标函数是一样的,所以优化  $\mathbf{W}$  的值就是用线性判别分析求解的结果。

结合(1)和(2),LDA-KM 的算法描述如下:

- ①对数据  $\mathbf{X}$  降维,得到初始化的  $\mathbf{W}$ ;
- ②执行(1),得到优化后的  $\mathbf{C}$ ;
- ③执行(2),得到优化后的  $\mathbf{W}$ ;
- ④循环执行②、③直到  $\mathbf{C}, \mathbf{W}$  的值不再发生变化为止。

## 4 局部线性嵌入

局部线性嵌入 (LLE) 是一种非线性降维方法,相对于传统的 PCA 降维方法,在处理非线性高维数据中效果较好<sup>[12]</sup>,在文中用于图像的降维。其算法的主要思想:对于一组具有嵌套流形的数据集,在嵌套空间与内在低维空间局部邻域间的点的关系应该不变。即在嵌套空间每个采样点可以用它的近邻点线性表示,在低维空间中保持每个邻域中的权值不变,重构原数据点,使重构误差最小。设高维空间中有  $N$  个数据属于同一类,记为:  $\vec{X}_i = \{x_1, x_2, \dots, x_n\}$ 。假设有足够的数据点,并且认为空间中的每一个数据点可以用它的  $k$  个近邻线性表示。 $W_{ij}$  表示第  $i$  个数据点用其第  $j$  个近邻线性表示时的权值,可以用如下的代价函数表示:

$$\psi(\mathbf{W}) = \sum_{i=1}^N \|\vec{X}_i - \sum_{j=1}^K W_{ij} \vec{X}_j\|^2 \quad (9)$$

其中  $\sum_{j=1}^K W_{ij} = 1$ , 求解  $W_{ij}$  就是最小化  $\psi(\mathbf{W})$ , 实际是求解一个最小二乘问题。求出  $W_{ij}$  后,保持权值不变,在低维空间中对原数据点进行重构。设低维空间的数据点为  $\vec{Y}_i = \{y_1, y_2, \dots, y_n\}$ , 其代价函数为:

$$\varphi(\mathbf{Y}) = \sum_{i=1}^N \|\vec{Y}_i - \sum_{j=1}^K W_{ij} \vec{Y}_j\|^2 \quad (10)$$

LLE 的算法描述如下:

- (1) 在高维空间,找到每一个数据点的  $k$  个近邻点;
- (2) 计算每一个数据点用它的  $k$  个近邻线性表示的权值,使得  $\psi(\mathbf{W})$  最小;
- (3) 保持权值不变,在低维空间中对原数据点进

行重构,使得  $\varphi(\mathbf{Y})$  最小。

## 5 实验

为了验证 LDA-KM 的有效性,以 500 张分辨率为  $40 \times 50$  的鼻子的灰度图像为例,进行聚类。为了得到初始子空间,将 LLE 应用于图像的降维,其中 LLE 中的近邻数为 8。希望鼻子形状相似的图像聚类在一起,采用了图像的梯度特征来消除光照的影响,用其梯度图像进行降维。依据 LLE 降维的结果,将鼻子图像映射到二维平面,如图 1 所示。以此空间为初始子空间,分别应用  $k$ -均值和 LDA-KM 进行聚类。聚类结果如图 2 所示,用  $k$ -均值聚类得到结果(a),利用(a)得到的类别用 LDA 进行子空间划分,然后新的子空间用  $k$ -均值聚类得到结果(b),如此迭代下去,得到中间结果(c),直到最后收敛得到结果(d)。从图 2 可以看出,分别用三种不同符号“.”,“+”,“\*”表示不同的三类,在 LDA-KM 的迭代算法下,三类逐渐分离并最终收敛。很明显,LDA-KM 的聚类结果比  $k$ -均值的好。为了进一步比较在不同聚类数上, $k$ -均值和 LDA-KM 聚类的正确率,在公共测试数据集 UCI<sup>[13]</sup> 上进行了实验。UCI 数据集标注了数据的属性和类别,用户可以用自己的数据挖掘方法去将 UCI 数据集分类,将结果与数据说明的结果对比,说明自己算法的正确性。比较了聚类数从 2 到 20 之间, $k$ -均值和 LDA-KM 聚类的正确率,结果如图 3 所示。从图中可以看出,随着聚类数的增多,两种算法的正确率都在下降,但 LDA-KM 的正确率始终高于  $k$ -均值的,这是因为 LDA-KM 算法是在  $k$ -均值结果的基础上采用反复迭代的方式,子空间自适应地调整到全局最优。

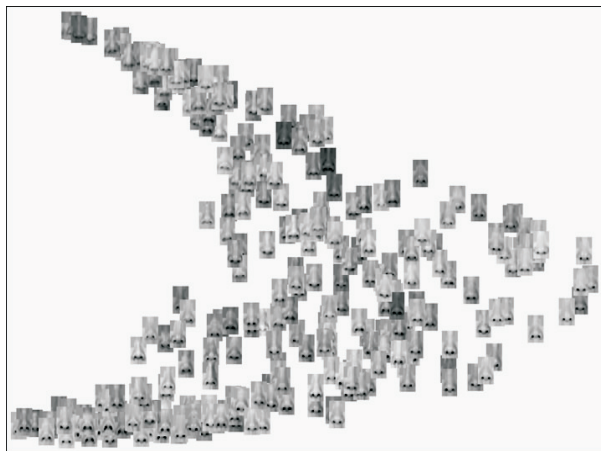


图 1 使用梯度特征,鼻子图像映射到 LLE 二维子空间的结果

## 6 结束语

文中通过整合 LDA 和  $k$ -均值,提出了一种自适

适的降维方法确定最初的低维子空间? 在以后的工作中,将对这些问题进行研究。

#### 参考文献:

- [1] 贺玲,吴玲达,蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究,2007(1):10-13.
- [2] 于洪涛,段军义,杜照丰. 一种基于聚类技术的个性化信息检索方法[J]. 计算机工程与应用,2008,44(8):187-188.
- [3] 李旭超,刘海宽,王飞,等. 图像分割中的模糊聚类方法[J]. 中国图象图形学报,2012,17(4):447-458.
- [4] Jolliffe T. Principal component analysis[M]. New York: Springer-Verlag,1986.
- [5] Cox T, Cox M. Multidimensional scaling[M]. London: Chapman-Hall,1994.
- [6] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science,2000,290:2323-2326.
- [7] De la Torre F, Kanade T. Discriminative cluster analysis[C]//Proc of International Conference on Machine Learning. New York: ACM,2006:241-248.
- [8] Ding C, Li T. Adaptive dimension reduction using discriminant analysis and k-means clustering[C]//Proc of International Conference on Machine Learning. New York: ACM,2007:521-528.
- [9] Fukunaga K. Introduction to statistical pattern recognition[M]. Boston: Academic Press,1990.
- [10] 周爱武,于亚飞. K-Means 聚类算法的研究[J]. 计算机技术与发展,2011,21(2):62-65.
- [11] 黄韬,刘胜辉,谭艳娜. 基于 k-means 聚类算法的研究[J]. 计算机技术与发展,2011,21(7):54-57.
- [12] 黄移军. 基于局部线性嵌入法的流形学习[J]. 数学理论与应用,2009,29(4):38-42.
- [13] Frank A, Asuncion A. UCI machine learning repository[D]. Irvine, CA: University of California,2010.

图 2 k-均值和 LDA-KM 聚类结果

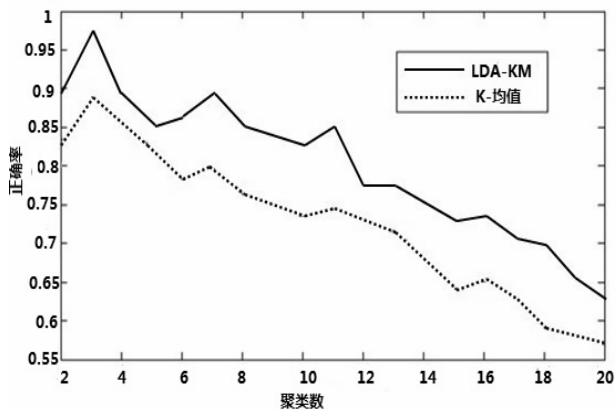


图 3 不同聚类数上, k-均值和 LDA-KM 聚类的正确率比较结果

应于空间选择方法。该方法采用反复迭代的方式,用无监督学习的方法找到具有最佳的不可分离性的子空间和数据的类别,实验结果表明该方法的聚类效果优于传统的 k-均值聚类。然而,还有一些问题需要进一步的研究。例如,如何确定最佳的聚类数?如何选择合

(上接第 82 页)

- [3] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis[M]. Cambridge: Cambridge University Press,2004.
- [4] 庞剑锋,卜东波,白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001(9):23-26.
- [5] 汪洪桥,孙富春,蔡艳宁,等. 多核学习方法[J]. 自动化学报,2010,36(8):1037-1050.
- [6] Lanckriet G R, Cristianini N, Bartlett P. Learning the kernel matrix with semidefinite programming[J]. Journal of Machine Learning Research,2004(5):27-72.
- [7] Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm[C]//Proceedings of the 21th International Conference on Machine Learning. New York: ACM Press,2004.
- [8] Sonnenburg S, Ratsch G, Schafer C. A general and efficient multiple kernel learning algorithm[C]//Proc of Neural Information Processing Systems. Cambridge: MIT Press,2006.
- [9] Gai K, Chen G, Zhang C. Learning kernels with radiuses of minimum enclosing balls[C]//Proc of 24th Annual Conference on Neural Information Processing System. Cambridge: MIT Press,2010.
- [10] 刘勇,廖士中. 基于支持向量机泛化误差界的多核学习方法[J]. 武汉大学学报(理学版),2012,58(2):149-156.
- [11] Cortes C, Mohri M, Rostamizadeh A. Generalization bounds for learning kernels[C]//Proceedings of the 27th International Conference of Machine Learning. New York: ACM Press,2010.

# 自适应子空间选择方法研究

作者：[闵锋](#)，[鲁统伟](#)，[邹旭](#)，[MIN Feng](#)，[LU Tong-wei](#)，[ZOU Xu](#)  
作者单位：[武汉工程大学 智能机器人湖北省重点实验室, 湖北 武汉, 430074](#)  
刊名：[计算机技术与发展](#)

英文刊名：

ISTIC

[Computer Technology and Development](#)

年，卷(期)：

[2013\(10\)](#)

本文链接：[http://d.g.wanfangdata.com.cn/Periodical\\_wjfz201310021.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjfz201310021.aspx)