计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(8):1485~1490, 2006

基于聚类和改进距离的 LLE 方法在数据降维中的应用

王和勇[1] 郑 杰[2] 姚正安[2] 李 磊[1]
[1](中山大学软件研究所 广州 510275)
[2](中山大学数学与计算科学学院 广州 510275)
(zsuwhy@hotmail.com)

Application of Dimension Reduction on Using Improved LLE Based on Clustering

Wang Heyong[1], Zheng Jie[2], Yao Zheng'an[2], and Li Lei[1]
[1](*Software Research Institute, Sun Yat-sen University, Guangzhou* 510275)
[2](*School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou* 510275)

Abstract Locally linear embedding (LLE) is one of the methods intended for dimension reduction. Its extension using clustering and improved LLE for dimension reduction is investigated. Firstly, using clustering can reduce time-consuming. Secondly, the improved LLE is suitable for selecting the number K of the nearest neighbors. When the number K of the nearest neighbors is small, it can obtain good results. While the original LLE algorithm obtains the same results, the number K of nearest neighbors may be much larger. Even if the number K of the nearest neighbors using the improved LLE is selected to be larger, the result is still right. So, the improved LLE is not sensitive to the selection of K. It is shown that the improved LLE based on clustering has less computing than the original LLE algorithm and enlarges the choice of parameter K by experiment.

Key words multimedia database; image retrieval; locally linear embedding

摘 要 局部线性嵌入算法(locally linear embedding, LLE)是解决降维的方法,针对 LLE 计算速度和近邻点个数 K 的选取,研究了该方法的扩展,提出了基于聚类和改进距离的 LLE 方法. 基于聚类 LLE 方法大大缩减了计算 LLE 方法的时间;改进距离的 LLE 方法在近邻点个数取值比较小时的情况下,可得到良好的效果,而原始的 LLE 方法要达到相同的效果,近邻点个数 K 的取值通常要大很多. 同时,改进距离的 LLE 方法可以模糊近邻点个数选取. 实验结果表明,基于聚类和改进距离相结合的 LLE 方法相比原来的 LLE 方法大大提高了降维速度和扩大了参数 K 的选取.

关键词 多媒体数据库;图像检索;局部线性嵌入算法

中图法分类号 TP391.4

维数缩减是模式识别的重要内容. 高维数据含有大量的冗余数据,维数缩减的目的是消除冗余性,以便提高图像的识别速度. 维数缩减有多种方法,传统维数缩减方法主要是线性方法,例如 PCA[1] 方法、KMEANS 方法[2] 和 Fisher 判别方法[3] 等. 文献中介绍很多非线性方法,例如 MDS[4],SOM[5],但 MDS 和 SOM 计算时间比较长,ISOMAP[6] 力图保持流形的全部几何性质,LLE[7] 方法力图保持局部几何性质. 对 LLE 方法,在求近邻点和根据近邻点的权值降维时,与样本点个数有关,为了提高计算速度,必须缩减样本点的个数,所以,本文提出了基于聚类的 LLE 方法. 另外,LLE 方法主要与近邻点个数 K 的选取有关,本文提出改进距离的 LLE 方法,可以模糊近邻点个数 K 的选取.

收稿日期:2005-06-10;修回日期:2005-11-23
基金项目:广东省工业攻关计划项目(2004B10101004)

1 LLE 方法

LLE 是 Roweis 和 Saul[7] 于 2000 年提出的一种非线性降维方法,主要利用局部的线性来逼近全局的非线性,保持局部的几何结构不变,通过相互重叠的局部邻域来提供整体的信息,从而保持整体的几何性质.

LLE 方法是映射数据 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,$\boldsymbol{x}_i\in\mathbb{R}^d$ 到数据集 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$,$\boldsymbol{y}_i\in\mathbb{R}^m$ $(m<d)$. 该方法主要包括 3 步:

第 1 步,对高维空间中的每个样本点 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$,计算它和其他 $n-1$ 个样本点之间的距离,根据距离的大小,选择前 K 个与 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$ 最近的点作为其近邻点,常采用欧氏距离来度量两个点之间的距离,即 $d_{ij}=\left|\boldsymbol{x}_i-\boldsymbol{x}_j\right|$;

第 2 步,对每个 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$,找到它的 K 个近邻点之后,计算该点和它的每个近邻点之间的权重 $w_j^{(i)}$,即最小化:

$$\varepsilon_I(W)=\sum_{i=1}^{n}\left|\boldsymbol{x}_i-\sum_{j=1}^{K}w_j^{(i)}\boldsymbol{x}_j\right|^2,$$

其中,$\sum_{j=1}^{n}w_j^{(i)}=1$,如果 $\boldsymbol{x}_j$ $(j=1,2,\cdots,n)$ 不是 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$ 的近邻,则 $w_j^{(i)}=0$;

第 3 步,根据高维空间中的样点 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$ 和它的近邻 $\boldsymbol{x}_j$ $(j=1,2,\cdots,K)$ 之间的权重 $w_j^{(i)}$ 来计算低维嵌入空间中的值 $\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$. 由于在低维空间中尽量保持高维空间中的局部线性结构,而权重 $w_j^{(i)}$ 代表着局部信息,所以固定权重 $w_j^{(i)}$,使下面的损失函数最小化:

$$\varepsilon_\Gamma(\boldsymbol{Y})=\sum_{i=1}^{n}\left|\boldsymbol{y}_i-\sum_{j=1}^{k}w_j^{(i)}\boldsymbol{y}_j\right|^2=tr(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{Y}),$$

要求 $\sum_{i=1}^{n}\boldsymbol{y}_i=0$ 且 $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}=1$,以使 $\varepsilon_\Gamma(\boldsymbol{Y})$ 对平移、旋转和伸缩变化都具有不变性. 使 $\varepsilon_\Gamma(\boldsymbol{Y})$ 最小化的解为矩阵 $\boldsymbol{M}$ 的最小几个特征值所对应的特征向量构成的矩阵 $\boldsymbol{Y}$. 取 $\boldsymbol{M}$ 最小的 $m+1$ 个特征值对应的特征向量,去掉其中最小的特征值对应的特征向量,剩余的 m 个特征向量组成的矩阵就是低维空间中所得特征向量.

从以上计算过程可以看到,LLE 计算与样本点的个数相关. 当样本点的个数较大时,带来求近邻点及 $\boldsymbol{M}$ 矩阵的计算量增加. 需找到一个新的方法来降低求近邻点及 $\boldsymbol{M}$ 矩阵的计算量.

2 聚类的方法

为了尽可能保持原有样本点的分类信息,使变化后的信息尽量含有原样本点的信息,可行的方法是采用聚类算法. 因为聚类算法是一种多元统计分类方法,这种方法不必事先知道对象的分类结构,而是基于整个数据集内部存在若干‘分组’或‘聚类’为出发点产生的一种数据描述方法,每个子集中的点具有高度的内在相似性. 另外,聚类的中心点含有大量的信息,可用聚类的均值向量即中心点来代表该类,这样,近邻点及矩阵 $\boldsymbol{M}$ 的阶数随着样本点个数减少而减少,可大大降低求近邻点及矩阵的计算量.

聚类分析的算法很多,有系统聚类法、动态聚类法、神经网络聚类法、模糊聚类法、遗传聚类法等. 本文选择基于动态聚类的 K 均值聚类算法[8] 进行实验.

3 改进距离的 LLE 方法

实验发现特别对于分布不均匀的样本集,近邻点个数 K 的选取对实验结果影响较大. 在样本点分布稀疏的区域,K 个近邻点所组成的局部邻域显然要比在样本点分布比较密集的区域大,所以需要对 LLE 进行改进,降低它受样本点分布的影响.

改进距离的 LLE 方法是在第 1 步求 K 个近邻点的距离时,采用

$$d_{il}(\boldsymbol{x}_i,\boldsymbol{x}_l)=\frac{|\boldsymbol{x}_i-\boldsymbol{x}_l|}{\sqrt{M(i)M(l)}}$$

取代计算 LLE 时采用的欧氏距离,其中,$M(i)$,$M(l)$ 分别表示 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$,$\boldsymbol{x}_l$ $(l=1,2,\cdots,n)$ 和其他点之间距离的平均值,采用改进的距离寻找每个样本点 $\boldsymbol{x}_i$ $(i=1,2,\cdots,n)$ 的 K 个近邻点.

$d_{il}(\boldsymbol{x}_i,\boldsymbol{x}_l)$ 的分子是普通欧氏距离,分母是数值,所以容易证明给出新的距离满足距离定义的要求,即

① $d_{il}(\boldsymbol{x}_i,\boldsymbol{x}_l)\geqslant 0$,当且仅当 $\boldsymbol{x}_i=\boldsymbol{x}_l$ 成立,满足距离非负性;

② 满足距离对称性要求 $d_{il}(\boldsymbol{x}_i,\boldsymbol{x}_l)=d_{il}(\boldsymbol{x}_i,\boldsymbol{x}_l)$;

③ 满足三角不等式要求,即 $d_{il}(\boldsymbol{x}_i,\boldsymbol{x}_l)+d_{ik}(\boldsymbol{x}_i,\boldsymbol{x}_k)\geqslant d_{lk}(\boldsymbol{x}_l,\boldsymbol{x}_k)$.

新的距离使处于样本点分布较密集区域的样本点之间的距离增大,而使处于样本点分布较稀疏的区域的样本点之间的距离缩小,这样会使样本点的整体分布趋于均匀化,从而降低由样本点分布对 LLE 的实验结果的影响.

实验图像为半圆柱面(如图 1 所示),半圆柱面的上半部分由 $5\times5=25$ 个点组成,下半部分由 $20\times20=400$ 个点组成:

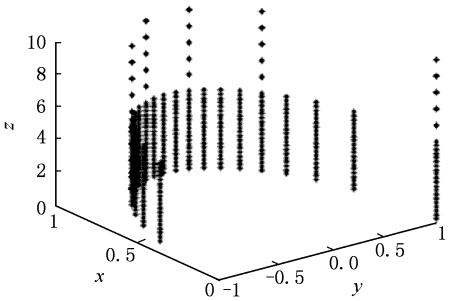

Fig. 1 Half cylindrical image.

图 1 半圆柱面

采用 LLE 方法和改进距离的 LLE 方法,分别将三维的半圆柱面的数据点降到二维平面. 当 K 的取值比较小时,两种方法的效果都不好. 当 $K=4$ 时降维效果如图 2 所示:

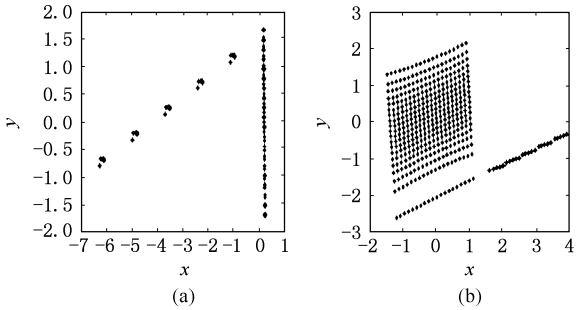

Fig. 2 3-dimensional half cylinder is reduced to 2-dimensional. (a) $K=4$, using LLE and (b) the improved LLE with $K=4$.

图 2 三维半圆柱面的数据点降到二维平面. (a) $K=4$ 时 LLE 方法降维效果;(b) $K=4$ 时改进距离的 LLE 降维效果

当 $K=9$ 时,降维效果如图 3 所示,改进距离的 LLE 方法已得到比较好的结果,而 LLE 方法的效果依然不理想:当 $K=10$ 时,降维效果如图 4 所示,改进距离的 LLE 方法依然得到较好的结果,而 LLE 方法效果还是不太理想.

直到 $K=19$ 时,LLE 方法才开始得到理想的结果(如图 5 所示),而此时改进距离的 LLE 方法一直保持着比较好的效果.

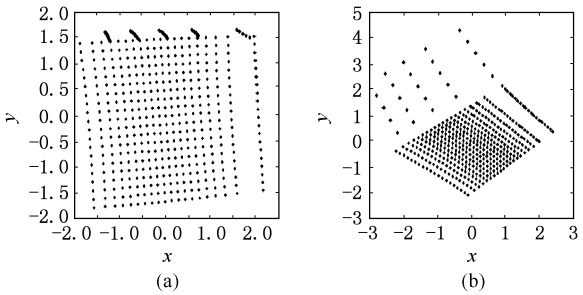

Fig. 3 3-dimensional half cylinder is reduced to 2-dimensional. (a) $K=9$, using LLE and (b) the improved LLE with $K=9$.

图 3 三维半圆柱面的数据点降到二维平面. (a) $K=9$ 时 LLE 方法降维效果;(b) $K=9$ 时改进距离的 LLE 降维效果

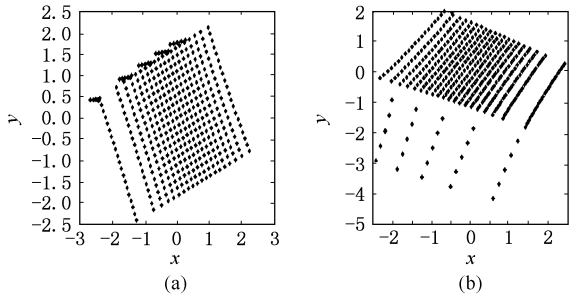

Fig. 4 3-dimensional half cylinder is reduced to 2-dimensional. (a) $K=10$, using LLE and (b) the improved LLE with $K=10$.

图 4 三维半圆柱面的数据点降到二维平面. (a) $K=10$ 时 LLE 方法降维效果;(b) $K=10$ 时改进距离的 LLE 降维效果

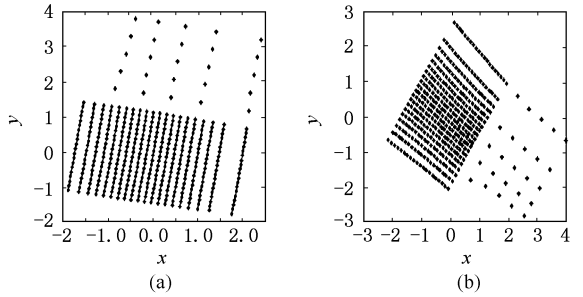

Fig. 5 3-dimensional half cylinder is reduced to 2-dimensional. (a) $K=19$, using LLE and (b) the improved LLE with $K=19$.

图 5 三维半圆柱面的数据点降到二维平面. (a) $K=19$ 时 LLE 方法降维效果;(b) $K=19$ 时改进距离的 LLE 降维效果

由实验不难发现,改进距离的 LLE 方法相对 LLE 方法对多数 K 的取值有较好的结果,从而在一定程度上模糊了 K 的选取.

4 实 验

用纹理图像的特征提取验证算法的正确性,具体方法如下所述:

首先对图像的每一个像素用 $Q\times Q$ 个窗口覆盖(如图 6 所示),每一个窗口包含 $n\times n$ 个像素. 测量是在窗口内进行的,构成了 $\mathbb{R}^{Q^2}$ 维的特征向量. 定义特征向量 $Z=(m_1,m_2,\cdots,m_{Q^2})$,其中 m_j 是第 j 个窗口的度量.

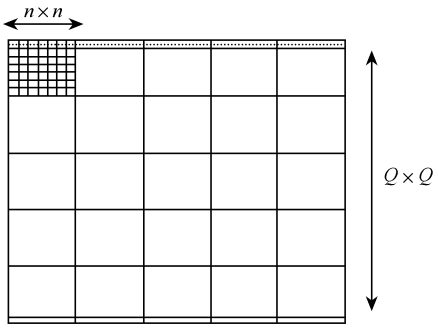

Fig. 6 Multi windows.

图 6 多窗口

度量使用的是每个窗口的灰度值的标准偏差:

$$m_j=\sqrt{\frac{\sum_{r=1}^{n^2}i_r^2}{n^2}-\left(\frac{\sum_{r=1}^{n^2}i_r}{n^2}\right)^2},$$

$$j=1,2,\cdots,Q^2,1\leqslant r\leqslant n^2,$$

其中,i_r 代表像素的灰度值,并且 $\sum_{r=1}^{n^2}i_r^2$ 表示第 j 个窗口所有像素的平方和,$\sum_{r=1}^{n^2}i_r$ 表示第 j 个窗口所有像素的和.

本文实验图像的宽度为 128 个像素,高度为 128 个像素. 取 $Q=5$, $n=7$,窗口个数为 25,窗口大小为 49,因此对每个像素都有 $Z=(m_1,m_2,\cdots,m_{25})$维的向量. 对图像的每个像素分别用上述所讲的多窗口来覆盖,所以图像共有128×128($m_1,m_2,\cdots,m_{25}$)个向量. 把 128×128 个像素作为样本点,每个样本点($m_1,m_2,\cdots,m_{25}$)是 25 维向量.

按照 K 均值聚类的方法对 128×128($m_1,m_2,\cdots,m_{25}$)个向量聚类,以聚类的中心点作为新的样本点,这样近邻点的个数及 M 矩阵的阶数只与样本点聚类的个数有关,大大缩减了求近邻点和求 M 矩阵特征向量的计算量;同时,再利用改进距离的 LLE 算法进行降维,可以降低提取出的特征由于分布不均匀所带来的影响.

非常明显,聚类的个数越多检索的效果就越好,但是也带来计算量大的麻烦,图 7 是聚类的个数和检索精度的分析,通过实验得到合适的聚类个数是 70.

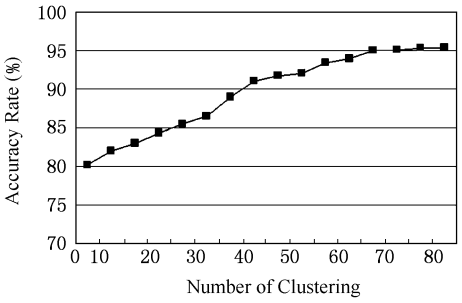

Fig. 7 The comparison between the number of clustering and the retrieval accuracy rate.

图 7 聚类的个数与查准率的比较

对 LLE 方法分别与聚类和改进距离的方法组合,形成 3 种降维方法,即改进距离的 LLE 方法、基于聚类的 LLE 方法、基于聚类和改进距离相结合的 LLE 方法,对各种方法选取 $K=13$. 各种方法降维时间如表 1 所示:

Table 1 The Comparison Among Using Four Methods to Reduce Dimension

表 1 四种方法降维所用时间的比较

Method	Time (ms)
LLE	3300
Improved LLE	3270
LLE based on clustering	610
Improved LLE based on clustering	600

各种方法效果趋势如图 8 所示:

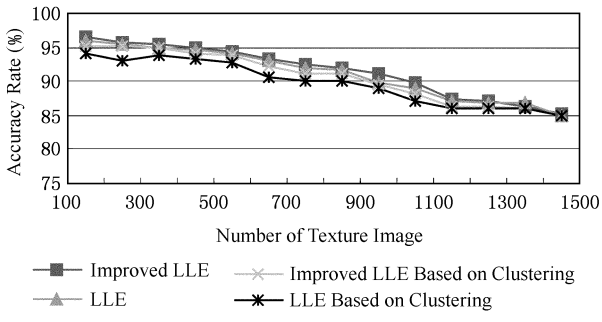

Fig. 8 The comparison of retrieval accuracy rate among four methods while test images are increased.

图 8 实验图像个数增多时的各种方法查准率比较

从图 8 和表 1 可以看出,基于聚类和改进距离相结合的 LLE 方法大大提高了 LLE 计算的速度,而且随着图像个数的增多几乎不影响原来的检索精度.

5 结 论

本文首先分析了 LLE 方法的不足,着重讲述了 LLE 针对样本点计算带来的计算量的分析,介绍了基于聚类和改进距离相结合的 LLE 方法,大大缩减了计算近邻点和 M 矩阵阶数的计算量,通过实验验证了算法的优越性.

参 考 文 献

[1] I T Jolliffe. Principal Component Analysis [M]. Berlin: Springer, 1986

[2] B Scholkopf, A Smola, K R Muller. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10(5): 1299-1319

[3] S Mika, G Ratsch, J Weston, *et al*. Fisher discriminant analysis with kernels [J]. Proceedings of IEEE Neural Networks for Signal Processing Workshop, 1999, 8(9): 41-48

[4] I Borg, P Groenen. Modern Multidimensional Scaling: Theory and Applications [M]. New York: Springer-Verlag, 1997

[5] T Kohonen. The self-organizing map [J]. Proceedings of the IEEE, 1990, 78(9): 1464-1480

[6] J B Tenenbaum, Vin de Silva, C John. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290: 2319-2323

[7] S T Roweis, L K Saul. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290: 2323-2326

[8] Bian Zhaoqi, Zhang Xuegong. Pattern Recognition [M]. Beijing: Tsinghua University Press, 2001 (in Chinese)
(边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2001)

Wang Heyong, born in 1973. Received his M A's degree in the Software Research Institute, Sun Yat-sen University, Guangzhou, China, in 2002. Since 2003, he has been a Ph D candidate in the Software Research Institute, Sun Yat-sen University. His current research interests include pattern recognition and data minning.

王和勇,1973 年生,博士研究生,主要研究方向为模式识别、数据挖掘.

Zheng Jie, born in 1979. Received his M A's degree in mathematics from the School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou in 2005. His current research interets include data mining and knowledge discovery.

郑杰,1979 年生,硕士,主要研究方向为数据挖掘、知识发现.

Yao Zhengan, born in 1960. Received his M A's degree and Ph D degree in Jilin University, China, in 1988 and 1994 respectively. He is a professor and the director in the School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou, where he also serves as a supervisor of doctor graduate. His current research interests include data minning, pattern recognition, and net safety.

姚正安,1960 年生,博士,教授,博士生导师,主要研究方向为数据分析、图像处理、计算机网络安全 .

Li Lei, born in 1951. Received his Ph D degree in computer science from Claude Bernard Lyon University, France, in 1988, and he finished his postdoctoral research in computer science from Jilin University, China, in 1990. He is a professor and the director in the Software Research Institute of Sun Yat-sen University, Guangzhou, where he also serves as a supervisor of doctor graduate. His main research interests focus on databases and logic programming, software engineering, and data minning.

李磊,1951 年生,博士,教授,博士生导师,主要研究方向为数据库与知识库、软件工程、数据挖掘.

Research Background

Dimension reduction is an important operation for pattern recognition. Because high-dimensional data have a lot of redundancies, the purpose of this operation is to eliminate the redundancies and lessen the amount of data to be processed.

Locally linear embedding (LLE) is one of the methods intended for dimensionality reduction, which relates to the number K of nearest-neighbors points to be initially chosen. So, clustering and improved LLE for dimension reduction is proposed. Firstly, using clustering can reduce time-consuming. Secondly, a method of improved LLE is given, which uses a new approach for computing weight of K nearest neighbor points in LLE. Thus, even when the number K is little, the improved LLE can get good results of

dimension reduction, while the traditional LLE needs a larger number of K to get the same results. When the number K of the nearest neighbors gets large, test has proved that the Improved LLE can still get correct results. This research is supported by the Key Industrial Technologies Research and Development Program of Guangdong Province (2004B10101004).