

基于可变 k 近邻 LLE 数据降维的图像检索方法

李 勇, 陈贺新, 赵 刚, 孙中华, 陈绵书

(吉林大学 通信工程学院, 长春 130022)

摘要: 在基于内容的图像检索中, 其计算复杂度随着描述图像内容的特征向量的维数的增加而急剧增加, 而应用局部嵌入算法(LLE)进行数据降维时, 需要确定近邻点 k 的个数。根据图像特征在原空间的分布情况, 提出了一种可变 k 近邻 LLE 的数据降维方法, 使得降维后的特征向量有效地保持了其在高维空间中的拓扑结构。实验结果表明, 提出的可变 k 近邻 LLE 数据降维方法在基于内容的图像检索中有较高的检索准确率。

关键词: 信息处理技术; 局部嵌入算法; 可变 k 近邻; 数据降维

中图分类号: TN911.73 **文献标识码:** A **文章编号:** 1671-5497(2008)04-0946-04

Image retrieval based on variable k -nearest neighbor locally linear embedding data dimension reduction algorithm

LI Yong, CHEN He-xin, ZHAO Gang, SUN Zhong-hua, CHEN Mian-shu

(College of Communication Engineering, Jilin University, Changchun 130022, China)

Abstract: In content based image retrieval, computation complexity will increase seriously as the demension of the image feature increases. When reducing the vector dimension with locally linear embedding algorithm, the number k of nearest neighbors should be determined. In this paper we propose a variable k neighbors LLE (VK-LLE) method based on the distribution of the image feature. The VK-LLE method can reduce the vectors dimension with keeping their original topology structure into a lower dimension space. Experiment shows that the proposed VK-LLE method can achieve higher precision rate in content based image retrieval problem.

Key words: information processing; locally linear embedding; variable k nearest neighbor; data dimension reduction

基于内容的图像检索指的是通过比较描述图像内容的特征而不是文本来进行相似图像匹配。一般情况下, 图像特征用高维向量进行表示, 一幅图像可以看作高维空间中的一个点, 相似图像检索则可以转化成寻找高维特征空间中与指定点距离最近的一组点的问题。对于大型的图像数据

库, 由于图像数据量巨大, 需要进行数据降维来降低存储空间和检索匹配复杂度。维数缩减可以降低数据的冗余性。传统的数据降维方法(如 PCA)是一种线性降维方法, 针对线性分布的高维数据进行降维, 具有局部最优问题^[1]。针对非线性降维问题, Roweis 和 Saul^[2]于 2000 年提出了

收稿日期: 2007-03-26.

基金项目: 国家自然科学基金项目(60672100, 60702036)

作者简介: 李勇(1974-), 男, 讲师, 博士. 研究方向: 图像内容检索. E-mail: liyong99@jlu.edu.cn

通信作者: 孙中华(1978-), 男, 讲师, 博士. 研究方向: 图像内容检索, 视频内容分析, 全景视频拼接.

E-mail: sunzh@bjut.edu.cn

一种局部线性嵌入 (Locally Linear Embedding, LLE) 方法,它解决了高度非线性分布数据的降维问题,同时不存在局部最优的问题。LLE 利用局部线性重建反映高维数据空间中的非线性结构,使降维后的数据保持原始空间的拓扑结构^[3-4]。LLE 通过最小化所有特征向量局部 k 近邻重建误差而达到全局最优^[5]。在求近邻点和根据近邻点的权值降维时,降维结果与样本点个数有关^[6]。若 k 值选取太小,则不足以反映特征与其近邻的拓扑结构;若 k 值选择太大,则会造成降维后低维空间的重叠,这些都会影响检索的准确度。针对这一问题,作者提出了一种可变 k 近邻的 LLE 降维算法——VK-LLE (Variable K-nearest neighbors LLE),根据图像特征向量在高维空间的分布情况,对不同特征向量选择不同数量的近邻个数。

1 LLE 算法描述

给定 N 个输入特征向量集合 $\{x_1, x_2, \dots, x_N\}$, $x_i \in R^d$, 通过 LLE 算法,得到降维后的输出向量 $y_i, i \in [1, N]$, $y_i \in R^m, m \ll d$ 。LLE 算法分为三步:

(1) 寻找每个特征向量的 k 个最近邻点。对高维空间中的样本点 x_i , 计算该点与其他 $N-1$ 个样本点之间的距离,将距离排序,选择前 k 个与 x_i 最近的点作为其邻近点。

(2) 用每个特征向量的近邻点对该特征向量进行重建,求取使重建误差最小的近邻局部重建权值矩阵。

重建误差函数定义如下

$$\epsilon(\mathbf{W}) = \sum_{i=1}^N \left\| x_i - \sum_{j=1}^k \omega_j^{(i)} x_{(i)j} \right\|^2 \quad (1)$$

式中: $x_{(i)j} (j=1, 2, \dots, k)$ 为 x_i 的第 j 个近邻点;

$\omega_j^{(i)}$ 为 x_i 与 $x_{(i)j}$ 之间的权值,且要满足 $\sum_{j=1}^k \omega_j^{(i)} = 1$ 。使误差函数值达到最小的 \mathbf{W} 即为所求的近邻局部重建权值矩阵。

(3) 由该特征向量 x_i 的局部重建权值矩阵 \mathbf{W} 和其近邻点 $x_{(i)j}$ 计算出该特征向量在低维嵌入空间的输出向量 y_i 和 $y_{(i)j}$ 。

为了使输出向量在低维空间尽量保持高维空间的拓扑结构,构造代价函数如式(2)所示,使得在构造过程中,代价函数值达到最小。

$$\epsilon(\mathbf{Y}) = \sum_{i=1}^N \left\| y_i - \sum_{j=1}^k \omega_j^{(i)} y_{(i)j} \right\|^2 \quad (2)$$

式中: y_i 为 x_i 的输出向量; $y_{(i)j} (j=1, 2, \dots, k)$ 为 y_i 的 k 个近邻点,且要满足两个条件

$$\sum_{i=1}^N y_i = 0, \frac{1}{N} \sum_{i=1}^N y_i y_i^T = \mathbf{I} \quad (3)$$

式中: \mathbf{I} 是单位矩阵。在计算最小化代价函数 $\epsilon(\mathbf{Y})$ 的最优解 y_i 时, $\epsilon(\mathbf{Y})$ 可写成如下形式

$$\epsilon(\mathbf{Y}) = \sum_{i=1}^N \sum_{j=1}^N \mathbf{M}_{i,j} y_i^T y_j = \text{tr}(\mathbf{YMY}^T) \quad (4)$$

式中: \mathbf{M} 是一个 $N \times N$ 的对称矩阵,其表达式为

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W}) \quad (5)$$

使代价函数 $\epsilon(\mathbf{Y})$ 最小的解为矩阵 \mathbf{M} 的最小 m 个特征值所对应的特征向量构成的矩阵 \mathbf{Y} , 则 \mathbf{Y} 的列向量即为 m 维空间的降维向量表示。

综上所述, LLE 算法在计算过程中与样本点近邻个数选择相关。图 1 中给出了应用颜色特征进行图像检索的结果^[7], 选择不同的 k 值得到的降维后图像检索的查准率和查全率也不同。

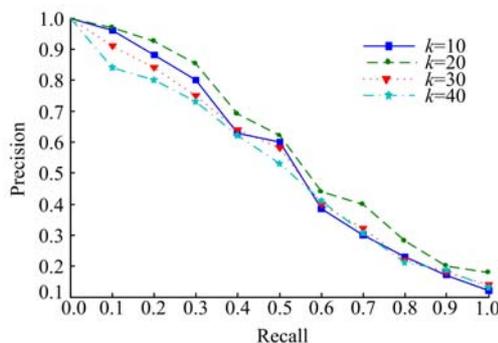


图 1 不同 k 值对应的图像检索查准率-查全率

Fig. 1 Precision-recall for image retrieval result with different k value

2 可变 k 近邻的 LLE 算法

2.1 k 值的确定

根据样本点的分布情况来确定 k 值。由于样本分布未知,假设相对于某个样本,用于对该样本进行局部重建的样本到该样本的距离服从高斯分布,那么在区间 $[\mu - 3\sigma, \mu + 3\sigma]$ 上包含了 99.73% 的这类样本,由这些样本对该样本进行局部重建具有一定的合理性,因此确定 k 为此范围内包含的样本数。该方法的具体实现如下:给定特征向量集合 $\mathbf{X} = \{x_1, x_2, \dots, x_N\}, x_i \in R^d$ 对于某一图像对应的特征向量 x_i , 首先计算它与其他特征向量之间的距离 $(d_{i1}, d_{i2}, \dots, d_{i(N-1)})$, 然后计算均

值 μ 和方差 σ

$$\mu_i = \frac{1}{N-1} \sum_{j=1}^{N-1} d_{ij} \quad (6)$$

$$\sigma_i^2 = \frac{1}{N-1} \sum_{j=1}^{N-1} (d_{ij} - \mu_i)^2 \quad (7)$$

定义算子 $K(x)$ 为到 x_i 的距离小于 x 的特征向量的个数, 则对于 x_i , 其近邻个数 k_i 为

$$k_i = K(|\mu_i - 3\sigma_i|) \quad (8)$$

2.2 可变 k 近邻 LLE(VK-LLE)数据降维

VK-LLE 和 LLE 算法的步骤类似, 在构造局部重建权值矩阵的误差代价函数时, 采用了同模糊 C 均值算法相类似的权值计算方法^[8], 避免了规范化参数 r 的选取。局部重建权值矩阵 W 为

$$\omega_{ij}^{(v)} = \frac{1}{\sum_{q=1}^{k_i} \left[\frac{d_{(i)j}}{d_{(i)q}} \right]^2} \quad (9)$$

式中: $d_{(i)j}$ 为 x_i 的近邻点 $x_{(i)j}$ 到 x_i 的距离; $d_{(i)j}^2 = (x_i - x_{(i)j})^T (x_i - x_{(i)j})$ 。

3 实验与结果

本文实验所用的数据库为 Corel 图像库, 库中共有 1000 幅图像, 首先对图像库中的图像进行人工分类, 将相似的图像归于一类。然后计算库中图像的颜色直方图^[7]作为原始特征向量, 之后对原始图像分别进行基于降维前和经过 PCA、LLE、VK-LLE 降维后的特征进行相似图像检索。

图 2 为降维前后检索的结果(由 256 维颜色



(c) LLE降维检索

图 2 降维前后检索结果

Fig. 2 Image retrieval result before and after dimension reduction

特征向量降到 32 维)。图 3 为降维前后图像检索的查准率和查全率曲线。图 4 和为示例图像降维

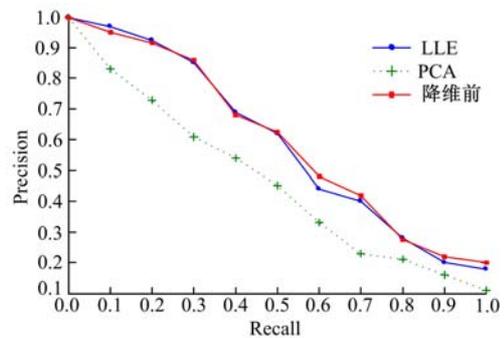


图 3 降维前后查准率-查全率曲线

Fig. 3 Precision-recall curve for image retrieval before and after dimension reduction



(a) 降维前检索



(b) PCA降维检索



(a) LLE降维检索



(b) VK-LLE降维检索

图 4 LLE 和 VK-LLE 降维检索结果比较

Fig. 4 Result comparison between LLE and VK-LLE dimension reduction

后的检索结果。

从图 3 中的检索结果曲线可以看出,LLE 降维后检索的准确率不但没有降低,反而有所提高,降维后的特征向量仍然保持高维空间的拓扑结构,大大降低了由于降维引起的检索精度损失。

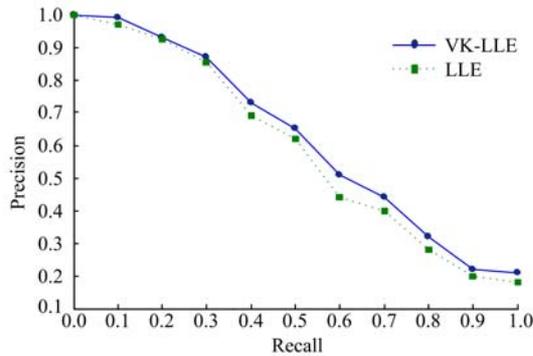


图 5 LLE 和 VK-LLE 查准率-查全率曲线

Fig. 5 Precision-Recall curve for LLE and VK-LLE

图 5 为示例图像经过 LLE 降维和 VK-LLE 降维之后的查准率-查全率曲线,图 6 为 PCA、LLE、VK-LLE 算法降到不同维数时,图像检索的准确率变化情况。从图 6 中可以看出,当 LLE 和 VK-LLE 算法将图像特征降低到 8 维时,检索结

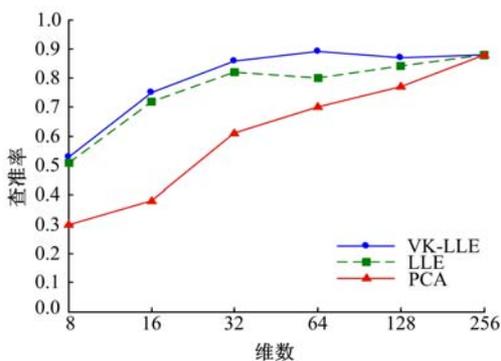


图 6 不同降维维数的查准率曲线

Fig. 6 Precision curve for different number of dimension reduction

果仍然保持较高的查准率,而 PCA 算法的查准率却在迅速下降。这是因为 PCA 算法在将特征向量映射到低维空间后破坏了原始高维空间中特征向量间的拓扑结构,导致相似性度量失常所致。

4 结束语

在对 LLE 非线性降维方法研究和实验的基础上,对其进行了改进,提出了基于可变 k 近邻的近似 LLE 降维算法——VK-LLE 算法。从降维

的检索结果、查准率-查全率曲线以及分类的错误率三个方面的比较结果来看,VK-LLE 算法表现了比较好的性能。

从实验来看,PCA 恰恰抛弃了原始空间数据分离性和局部的拓扑结构。LLE 和 VK-LLE 基于与聚类、检索相一致的距离函数,考虑近邻问题,构建局部权值矩阵,以此局部权值矩阵进一步构建降维误差代价函数,转化为最小化此代价函数的问题,使降维后特征向量同其近邻的距离权重的误差最小,也就使得在低维空间仍然能够保持原始空间近邻间的拓扑结构。提出的 VK-LLE 算法在 LLE 的基础上进行了改进,考虑到每个特征向量,其近邻分布是不同的,因此近邻 k 值也应该是不同的,并给出了 k 值及近邻的确定方法,使算法更加合理,从而得到了更好的检索效果。

参考文献:

- [1] Richard O D, Peer E H, David G S. Pattern Recognition[M]. 2nd ed. London: Wiley Interscience, 2001.
- [2] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290:2323-2326.
- [3] Wu F C, Hu Z Y. The LLE and a linear mapping[J]. Pattern Recognition, 2006, 39:1799-1804.
- [4] Hadid A, Kouropteva O, Pietikainen M. Unsupervised learning using locally linear embedding: experiments with face pose analysis[C]//Proceedings of the 16th International Conference on Pattern Recognition, 2002: 111-114.
- [5] Hong C, Yeung D Y. Robust locally linear embedding [J]. Pattern Recognition, 2006, 39:1053-1065.
- [6] 王和勇,郑杰,姚正安,等. 基于聚类和改进距离的 LLE 方法在数据降维中的应用[J]. 计算机研究与发展, 2006, 43(8):1485-1490.
Wang He-yong, Zheng Jie, Yao Zheng-an, et al. Application of dimension reduction on using improved LLE based on clustering[J]. Journal of Computer Research and Development, 2006, 43(8):1485-1490.
- [7] Lee S M, Xin J H, Westland S. Evaluation of image similarity by histogram intersection[J]. Color Research and Application, 2005, 30(4):265-274.
- [8] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.