基于半监督距离学习和多模态信息的图像聚类

梁建青 胡清华
(天津大学计算机科学与技术学院 天津 300072)

摘 要 通过融合图像中不同模态的信息并利用少量带标记的图像进行半监督距离学习,来对图像进行聚类。首先,提取彩色图像中RGB颜色空间的直方图信息、纹理信息,并采用SIFT算法提取Bag of Words来重新表达图像,从而基于图像的颜色特征、纹理特征以及语义特征,建立图像的多模态表达机制,将原始图像投射到表达空间;然后,利用少量标记的图像,通过半监督距离学习,获得图像在多模态信息空间的相似性度量;最后,通过半监督聚类方法,实现图像分组,在多个图像数据库中验证提出的方法的有效性。

关键词 半监督,距离学习,多模态,图像聚类
中图法分类号 TP391.4 **文献标识码** A

Image Clustering Based on Semi-supervised Distance Learning and Multi-modal Information

LIANG Jian-qing HU Qing-hua
(School of Computer Science and Technology, Tianjin University, Tianjin 300072, China)

Abstract The project clustered images by fusing the different model information in the images and taking advantage of a small amount of labeled images for semi-supervised distance learning. First, we extracted histogram information of the RGB color space, texture information in the color images, and Bag of Words by using the SIFT algorithm to re-express the image, thus establishing the multi-modal express mechanism of images based on the image's color, texture and semantic features to project the original image onto the space to express. Then, using a small amount of the marked image, we obtained a similarity measure in multi-modal information space of images through the semi-supervised distance learning. Finally, we realized grouping images through the semi-supervised clustering method and verified the validity of the proposed method in the plurality of images in the database as well.

Keywords Semi-supervise, Distance learning, Multi-modal, Image clustering

1 引言

随着大数据时代的到来,以图像、音频和视频等为代表的非结构化数据达到互联网数据量的75%以上,这些数据的产生往往伴随着社交网络、移动计算等新的渠道和技术的不断涌现和应用。此外,这些数据中仅有少部分类别已知,绝大部分数据没有标记。因此,如何充分利用已有信息,在短时间内对海量非结构化数据进行快速分类,从而挖掘出信息潜在的价值,以做出正确合理的决策,成为机器学习、数据挖掘等领域所面临的挑战。

通常,基于内容的图像检索系统采用欧氏度量来计算图像之间的距离。然而,由于在低级特征和高级语义概念之间存在语义鸿沟,欧氏度量无法得到满意的结果。此外,一般的距离度量技术对噪音敏感,当仅有少部分记录数据可用时,无法学习一种可靠的度量。在半监督距离学习方面,Xing等人提出了一种著名的距离度量方法,将任务制定为一个凸优化问题,并将解决方案应用于聚类任务中。此后,大量的距离度量技术逐步产生。近几年,Si等人对协同的图像检索应用提出了一种正则化的度量学习方法。Steven C. H. Hoi, Wei Liu和Shih-Fu Chang通过一个图正规化的学习框架,在距离度量学习中融入无标签数据,提出了一种新的半监督距离度量学习架构用于学习有效、可靠的度量[1]。

针对目前研究过程中出现的问题,本课题主要从以下几方面开展工作:1. 融合图像中颜色、纹理和语义不同模态的信息进行聚类,使得图像聚类算法不仅仅局限于灰色图像。2. 实验中特征提取采用统计信息,可以对大小不一的图像进行聚类。具体而言,颜色和语义特征采用频率表达,纹理则采用灰度共生矩阵得到的特征参数表示。3. 选取的半监督距离学习方法采用最近邻法为测试集样本进行标注,在此基础上进行距离学习,从而有效地避免了过拟合问题的出现。

本文第2节介绍3种特征的提取及表达方法;第3节引入了一种新的半监督距离学习方法;第4节对现有的两种聚类方法融入半监督信息并加以利用;第5节展示实验结果并进行分析;最后总结全文。

2 特征提取及表达

为了对图像不同模态的信息进行相似性度量,首先需要提取图像的3种特征,建立多模态表达机制。考虑到图像的

到稿日期:2013-05-15 返稿日期:2013-06-18 本文受国家优秀青年科学基金(61222210)资助。
梁建青(1990—),女,硕士生,主要研究方向为机器学习,E-mail:1037264519@qq.com;胡清华(1976—),男,博士,教授,主要研究方向为复杂数据机器学习、数据挖掘。

大小不一，我们采用基于统计量的方法：对于颜色，采用 RGB 颜色空间的直方图表达；对于纹理，采用灰度共生矩阵的特征参数表达；对于语义，采用 SIFT 算法提取 Bag of Words 来重新表达。

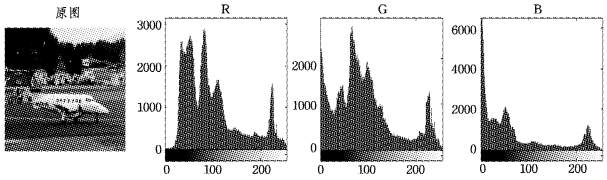
2.1 颜色特征

RGB 模型在颜色的表示方法中较为常见。在 RGB 空间可采用统计直方图表示图像的颜色特征[10]。该统计直方图可用如下函数表示：

$$H(k)=\frac{N_k}{n},k=0,1,\cdots,L-1 \tag{1}$$

式中，k 表示图像像素的颜色取值，L 为颜色特征的维数，n 为像素个数。

airplane 类中一幅图片的 RGB 颜色空间直方图如图 1 所示。



横坐标取值 0～255，表示像素的颜色取值；
纵坐标表示对应频数，实际操作时转换为频率

图 1 RGB 颜色空间直方图

2.2 纹理特征

灰度共生矩阵是一种通过研究灰度的空间相关特性来描述纹理的常用方法。它是通过对图像上保持某距离的两像素分别具有某灰度的状况进行统计而得到的，描述了成对像素的灰度组合分布[10]。

如果用 P_δ 表示灰度共生矩阵，那么矩阵元素可用概率值表示为

$$P_\delta(i,j),i,j=0,1,2,\cdots,L-1 \tag{2}$$

式中，i,j 对应两像素的灰度级，L 为灰度级数，$\delta=(\Delta x,\Delta y)$ 表示两像素在 x 和 y 方向的相对距离。

通常采用以下参数来表示灰度共生矩阵的特征，以此作为对图像纹理的一种描述。

(1)能量

$$Energy=\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}p_\delta(i,j)^2 \tag{3}$$

能量用于度量图像灰度分布的均匀性。对于一幅图像，纹理越粗，能量越大，纹理越细，能量越小。

(2)对比度

$$Contrast=\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}|i-j|^2p_\delta(i,j) \tag{4}$$

对比度反映了图像纹理的清晰程度。图像的纹理越清晰，对比度越大。

(3)相关

$$Correlation=\frac{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i-\mu_i)(j-\mu_j)p_\delta(i,j)}{\sigma_i\sigma_j} \tag{5}$$

其中

$$\begin{aligned}\mu_i&=\sum_{i=0}^{L-1}i\sum_{j=0}^{L-1}p_\delta(i,j)\\ \mu_j&=\sum_{j=0}^{L-1}j\sum_{i=0}^{L-1}p_\delta(i,j)\\ \sigma_i&=\sqrt{\sum_{i=0}^{L-1}(i-\mu_i)^2\sum_{j=0}^{L-1}p_\delta(i,j)}\\ \sigma_j&=\sqrt{\sum_{j=0}^{L-1}(j-\mu_i)^2\sum_{i=0}^{L-1}p_\delta(i,j)}\end{aligned} \tag{6}$$

相关用于度量某一像素与周围像素的相似程度，反映了图像纹理的一致性。

(4)逆差距

$$Homogeneity=\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\frac{p_\delta(i,j)}{1+|i-j|} \tag{7}$$

逆差距反映图像纹理的同质性，用来度量图像纹理局部变化的多少。

上述图片原图、灰度图以及纹理如图 2 所示。

图 2 原图、灰度图以及纹理

上述图片在 0°，45°，90°，135°这 4 个方向得到的特征参数如下：Contrast [118.4461 287.2957 194.6958 268.2423]，Correlation [0.9842 0.9616 0.9740 0.9641]，Energy [8.8520e−004 4.2426e−004 5.2689e−004 4.3025e−004]，Homogeneity [0.4485 0.3015 0.3385 0.3044]。

2.3 语义特征

2.3.1 SIFT 算法

SIFT 算法基于尺度空间，是一种对图像的旋转、缩放、光照等变换具有不变性的局部特征描述方法[9]。

(1)尺度空间的建立

高斯函数被证明为具有尺度变换不变性的唯一卷积核。定义如下：

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{8}$$

对于一幅图像，将其与高斯函数卷积，选取不同的 σ 值得到不同尺度空间下的图像，然后将尺度相邻且分辨率相同的图像相减，即可得到 DOG 金字塔：

$$\begin{aligned}D(x,y,\sigma)&=L(x,y,k\sigma)-L(x,y,\sigma)=(G(x,y,k\sigma)-\\&\quad G(x,y,\sigma))*I(x,y)\end{aligned} \tag{9}$$

(2)候选关键点的选取

为了进一步进行极值的选取，将 DOG 金字塔中图像的每一像素与周围 26 个像素点进行比较，如果为极值点，则为候选关键点。

(3)关键点的确定与描述

对于每层的极值点，计算其周围各点梯度大小及方向：

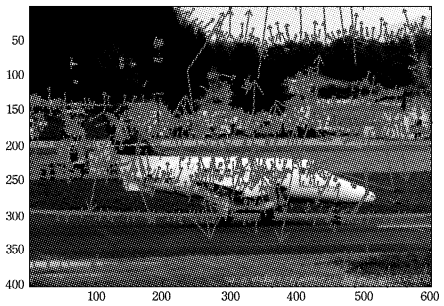
$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \tag{10}$$

$$\theta(x,y)=\tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y))) \tag{11}$$

以关键点为中心采样，用直方图统计其周围像素的梯度方向，该图的峰值作为关键点的领域梯度主方向，即关键点方向。

接下来，将图像以关键点为中心转到其主方向，计算 8 个方向的梯度方向直方图，得到累计值，产生一个种子点。每一关键点使用 16 个种子点描述，得到 128 维的 SIFT 特征向量[8]。

将上述大小为 600×401 的彩色图像进行灰度处理后，提取到 1491 个 SIFT 关键点，该图像的关键点示意图如图 3 所示。



箭头的起点代表位置，长度代表所处的尺度，
方向代表该尺度下关键点邻域的主梯度方向

图 3　关键点示意图

2.3.2　BOW 模型

BOW 模型的主要思想在于将图像作为独立图像块的集合，每一图像块由向量描述。对训练集的向量进行聚类，得到一个由视觉单词组成的词典。参照词典，投票统计图像中的向量，得到特征向量的直方图以表示图像[9]。

(1)BOW 算法

BOW 算法如图 4 所示，基本步骤如下：

Step 1　对图像进行预处理并提取描述算子。将彩色图像灰度处理，采用 SIFT 算法提取特征向量。

Step 2　将描述算子进行聚类，由类中心(视觉单词)组成得到词典。这里采用常见的 K-means 算法。

Step 3　将测试集的描述算子分配到已有的视觉单词类中。考虑到图像大小不一，用归一化后得到的频率向量表示图像。

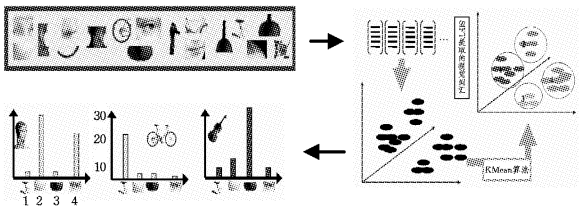


图 4　BOW 算法图示

(2)K-means 算法

K-means 方法采用欧氏距离进行度量，通过迭代使得目标函数取得极值。

K-means 算法的基本步骤如下：

Step 1　任选 k 个数据对象作为初始类中心；

Step 2　计算每个对象到类中心的距离，按照最小距离对数据对象重新划分；

Step 3　重新计算类的中心；

Step 4　计算目标函数，若函数收敛，则算法终止，否则执行 Step 2。

2.4　多模态表达机制

到此为止，每幅图像可以用 1084 维的数值向量表示，其中，对于颜色，R、G、B 3 种颜色分量分别是 256 维，元素范围为 0～1；对于纹理，0°，45°，90°，135°这 4 个方向分别得到的特征参数——能量、对比度、相关以及逆差距共 16 维；对于语义，维数和 k 的大小相同，为 300，元素范围为 0～1。

3　半监督距离学习

半监督距离学习主要考虑如何利用少量有标记数据和大量无标记数据进行距离度量，这对于提高学习性能具有重要意义。

3.1　问题定义

假设有一个数量为 n、维数为 m 的数据样本点集 $C=\{\boldsymbol{x}_i\}_{i=1}^{n}\subseteq R^m$，该集合被划分为以下两个子集：

$S=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \boldsymbol{x}_i,\boldsymbol{x}_j$ 相关$\}$

$D=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\mid \boldsymbol{x}_i,\boldsymbol{x}_j$ 无关$\}$

对于任意两个给定的数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，采用 $\boldsymbol{A}\in\mathbb{R}^{m\times m}$ 作为距离度量，公式如下：

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{x}_j)}$$
$$=\sqrt{\mathbf{tr}(\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}})} \tag{12}$$

通常地，作为一种有效的度量，$\boldsymbol{A}$ 必须是半正定的，即 $A\geqslant 0$。

3.2　一种正则化学习架构

度量学习的一个普遍准则是最小化相似集合中数据点之间的距离，同时最大化相异集合中数据点之间的距离，即“最小最大准则”。一些已有的距离度量工作在最小最大学习架构下进行，例如，将距离度量问题转换为一种凸优化问题：

$$\begin{aligned}&\min_{\boldsymbol{A}\geqslant 0}\ \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2\\&\text{s.t.}\ \ \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}\geqslant 1\end{aligned} \tag{13}$$

该方法试图在使得相异数据点之间的距离和大于 1 的约束下，通过最小化相似数据点之间的距离平方和来求得度量 $\boldsymbol{A}$。然而，当标记数据夹带噪音时，该方法可能因过拟合问题而变得不合理。

为此，我们引入正则化准则来增强距离度量的泛化能力和稳健性，为距离度量学习制定一个广泛的正则化架构：

$$\begin{aligned}&\min_{\boldsymbol{A}} g(\boldsymbol{A})+\gamma_s\upsilon_s(S)+\gamma_d\upsilon_d(D)\\&\text{s.t.}\ \boldsymbol{A}\geqslant 0\end{aligned} \tag{14}$$

式中，$g(\boldsymbol{A})$是一个定义在目标度量 $\boldsymbol{A}$ 上的正则算子，$\upsilon_s(\cdot)$和 $\upsilon_d(\cdot)$是分别定义在相似集和相异集上的损失函数，γ_s 和 γ_d 是为平衡相似和相异约束的两个正则化参数。同时，遵循最小最大准则，我们采用平方距离和表示两个损失函数：

$$\upsilon_s(\cdot)=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2 \tag{15}$$

$$\upsilon_d(\cdot)=-\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2 \tag{16}$$

3.3　拉普拉斯算子正则度量学习

Steven C. H. Hoi，Wei Liu 和 Shih-Fu Chang 通过一个图正规化的学习框架，在距离学习中融合无标签数据信息[1]。具体算法如下：

首先，对训练集和测试集的所有样本采用 KNN 算法，在此基础上计算得到权重矩阵：

$$\boldsymbol{W}_{ij}=\begin{cases}1, & \boldsymbol{x}_i\in N(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_j\in N(\boldsymbol{x}_i)\\0, & \text{otherwise}\end{cases}$$

其中，$N(\boldsymbol{x}_i)$为 $\boldsymbol{x}_i$ 的最邻近样本。

KNN 算法如下：

假设样本空间共有 N 个样本，其中已知 LN 个标记样本分为 C 类。选择欧氏距离对所有未标记样本进行计算，找到 k 个距离最近的样本。这 k 个最近邻样本在 C 类中的个数为 k_i，则判别函数为

$$g_i(x)=k_i,i=1,2,\cdots,C \tag{17}$$

其决策规则如下：

对于某一未标记样本 $\boldsymbol{x}$，若 $j=\arg\max_i g_i(x)$，$i,j=1,2,\cdots,C$，则 $\boldsymbol{x}\in\boldsymbol{w}_j$。

然后，对样本引入相应的线性映射 $U:R^m\to R^r$。定义两输入样本之间的距离如下：

$$\begin{aligned}d(\boldsymbol{x}_i,\boldsymbol{x}_j)&=\|\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)\|^2\\&=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)\\&=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{x}_j)\end{aligned}\tag{18}$$

其中，$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ 是将被学习得到的度量。

接下来，通过权重矩阵 $\boldsymbol{W}$ 引入无标签数据，得到如下算子：

$$\begin{aligned}g(\boldsymbol{A})&=\frac{1}{2}\sum_{i,j=1}^{n}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_j\|^2W_{ij}\\&=\sum_{k=1}^{r}u_k^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{X}^{\mathrm{T}}u_k\\&=\sum_{k=1}^{r}u_k^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}u_k=\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U})\\&=\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})\end{aligned}\tag{19}$$

式中，$\boldsymbol{D}$ 为对角矩阵，其对角元素为 $\boldsymbol{W}$ 每行元素之和，即 $D_{ii}=\sum_j W_{ij}$。$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$ 是拉普拉斯矩阵。

通过引入该正则算子，得到了“拉普拉斯算子正则度量学习”，其目标函数如下：

$$\min_{\boldsymbol{A}\geqslant 0}tr(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A})+\gamma_s\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in S}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2-\gamma_d\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in D}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{A}}^2\tag{20}$$

由于缺乏先验信息，实验设置 $\gamma_s=\gamma_d=1$，KNN 算法中参数 k 的取值范围设为 1～20，以观测其对聚类精度的影响。

4 半监督聚类

已有的半监督聚类方法通常归为两类：基于约束的和基于度量的。前者将约束作为目标函数的一部分用于聚类算法并作为指导，以获得合理的划分；后者采用一种经过训练的距离度量方法，根据标记数据构造一种距离度量，在此基础上进行聚类。

实验在拉普拉斯度量矩阵的基础上，采用类平均法和重心法作为距离度量方法，最后将未标记样本划分至与其距离最短的类中。

4.1 类平均法

类平均法把未标记样本与已知类之间的距离定义为未标记样本与已知类中所有样本间的平均距离[6]。设已知类 $\boldsymbol{G}_L$ 中有 n 个样本 $\boldsymbol{x}_{L1},\boldsymbol{x}_{L2},\cdots,\boldsymbol{x}_{Ln}$，则未标记样本 $\boldsymbol{x}$ 与 $\boldsymbol{G}_L$ 之间的距离为

$$D(\boldsymbol{x},\boldsymbol{G}_L)=\frac{1}{n}\sum_{i=1}^{n}d(\boldsymbol{x},\boldsymbol{x}_{Li})=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}-\boldsymbol{x}_{Li})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_{Li})\tag{21}$$

该方法充分利用全部样本的信息，通常被认为是一种较好的系统聚类法，如图 5 所示。

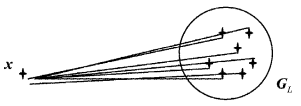


图 5 类平均法

4.2 重心法

重心法将未标记样本与已知类之间的距离定义为样本与类重心（均值）之间的距离[6]。设已知类 $\boldsymbol{G}_L$ 中有 n 个样本 $\boldsymbol{x}_{L1},\boldsymbol{x}_{L2},\cdots,\boldsymbol{x}_{Ln}$，其重心为 $\bar{\boldsymbol{x}}_L=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{Li}$，则未标记样本 $\boldsymbol{x}$ 与 $\boldsymbol{G}_L$ 之间的距离为

$$D(x,\boldsymbol{G}_L)=d(\boldsymbol{x},\bar{\boldsymbol{x}}_L)=(\boldsymbol{x}-\bar{\boldsymbol{x}}_L)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\bar{\boldsymbol{x}}_L)\tag{22}$$

该方法通常对异常值不敏感，如图 6 所示。

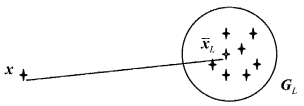


图 6 重心法

重心能够较好地代表类，但没有充分利用各样本的信息。

5 实验结果及分析

本课题采用 Computational Vision Group 所提供的图片进行测试，共分为‘airplane’、‘background’、‘car’、‘face’、‘leave’以及‘motorbike’ 6 种类别，每种类别的图片数量相同。为了选择合适的参数和算法，我们固定测试集图片数量为 240，等比例增加训练集图片数量，同时逐步增加 KNN 算法中参数 k 的取值，进行多次实验，观察参数对聚类精度的影响。

5.1 参数 k 的取值对聚类精度的影响

为了确定参数 k 的取值对聚类精度的影响，在大小为 120、216 和 240 的训练集上进行多次实验，实验结果如图 7－图 9 所示。

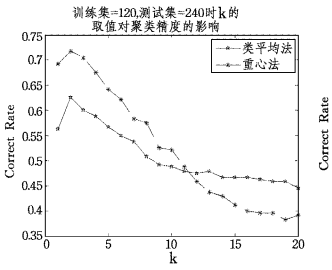


图 7 训练集/测试集＝0.5 时，k 的取值对聚类精度的影响

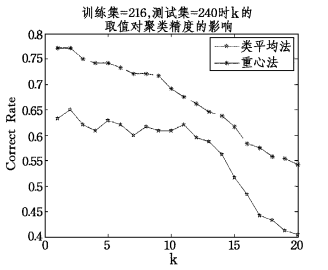


图 8 训练集/测试集＝0.9 时，k 的取值对聚类精度的影响

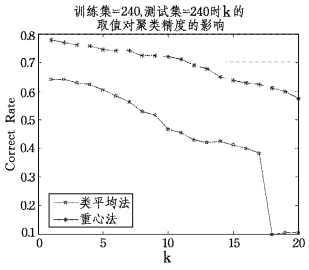


图 9 训练集/测试集＝1.0 时，k 的取值对聚类精度的影响

实验结果分析如下：

（1）训练集和聚类算法不变时，参数 k 的不同取值对精度的影响很大。例如，在图 8 中，随着 k 从 1 开始递增到 20，重心法的精度从 0.6917 变化到 0.3917，$k=2$ 时达到最大值 0.7167，$k=19$ 时取得最小值 0.3833。

（2）训练集/测试集的大小不同时，随着 k 的增长，两种聚类算法的结果表现出较大的差异。具体而言，随着 k 的增大，比值为 0.5 时，两种聚类算法的精度都呈现出下降趋势，重心法更为明显；比值为 0.9 时，类平均法局部波动并精度逐步减

小，重心法的精度也逐渐减小；比值为 1.0 时，类平均法在 $k=17$ 处其精度骤减，重心法的精度也逐渐减小。

5.2 训练集/测试集对聚类精度的影响

一般地，固定测试集，增加训练集，将会提高聚类算法的精度。为了验证猜想，实验设置参数 $k=2$，增大训练集，使训练集/测试集从 0.1 增加到 1.0，进行多次实验，插值后的实验结果如图 10 所示。

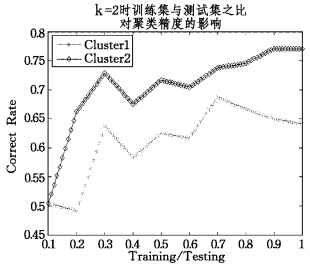


图 10 $k=2$ 时，训练-测试集之比对聚类精度的影响

实验结果证实了我们的猜想。总体上，随着训练-测试集之比的增长，聚类算法的精度呈现上升趋势。事实上，当训练集的规模增长时，就能为距离学习提供更多的类别信息，并且聚类时也利用了更充分的先验信息，因此得到更好的效果。然而，并非所有训练图像的添加都会有助于聚类结果的改善，当两幅不同类别的图像差异不显著时，可能带来误导信息，这就导致了局部波动现象的出现。

5.3 聚类精度的整体趋势

基于上述分析，我们希望对聚类精度的整体趋势有所了解，为此，增加训练集图片数量，使训练-测试集之比以 0.1 为间隔从 0.1 增加到 1.0，同时设置 KNN 算法中参数 k 的取值范围为 1～20，进行多次实验，对每一种算法得到的坐标点进行曲面拟合，实验结果如图 11、图 12 所示。

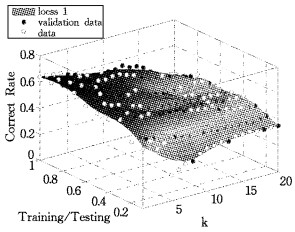


图 11 类平均法曲面图

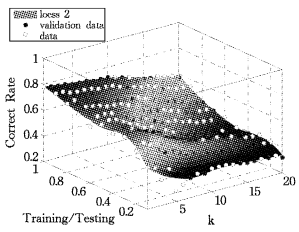


图 12 重心法曲面图

在此，我们先对涉及到的统计参数进行解释：

(1)SSE(和方差)

该参数计算原始数据和对应拟合数据点的误差平方和，计算公式如下：

$$SSE=\sum_{i=1}^{n}w_i(y_i-\hat{y}_i)^2 \tag{23}$$

其值越接近 0，则拟合效果越好，预测越准确。

(2)RMSE(均方根)

该参数也称为回归模型的拟合标准差，计算公式如下：

$$RMSE=\sqrt{SSE/n}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}w_i(y_i-\hat{y}_i)^2} \tag{24}$$

(3)R-square(确定系数)

$$R\text{-}square=\frac{SSR}{SST}=\frac{SST-SSE}{SST}=1-\frac{SSE}{SST} \tag{25}$$

其中，$SSR=\sum_{i=1}^{n}w_i(\hat{y}_i-\bar{y}_i)^2$ 为预测数据与原始数据均值之差的平方和，$SST=\sum_{i=1}^{n}w_i(y_i-\bar{y}_i)^2$ 为原始数据与其均值之差的平方和。确定系数采用数据的变化表示拟合的效果，其值越接近 1，模型的拟合效果越好。

实验中，类平均法的曲面拟合结果为 $SSE=0.1585$，$RMSE=0.0282$，$R\text{-}square=0.9291$；重心法的曲面拟合结果为 $SSE=0.0629$，$RMSE=0.0177$，$R\text{-}square=0.9802$。由此判断，以上两种算法曲面拟合的 SSE 较小，R-square 接近于 1，效果还是比较好的。

结束语 本课题通过融合图像中颜色、纹理以及语义特征 3 种不同模态的信息，利用少量带标记的图像，采用改进的拉普拉斯算子规范度量学习算法进行半监督距离学习，在此基础上融入半监督信息进行聚类，从而实现在短时间内对大量图像较为准确的标记。

需要注意的是，考虑到先验信息的缺乏以及计算量的增长，本课题并没有对 γ_s，γ_d 以及 K-means 算法中的参数 k 进行过多的探讨，它们对实验结果的影响较小。

今后，我们将关注多模态信息距离度量学习的算法改进、训练图像的选取、先验信息的获取以及相关参数的设置，进而提高算法的性能。

参 考 文 献

[1] Hoi S C H, Liu W, Chang S F. Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008. Anchorage, Alaska, IEEE Computer Society, 2008: 1-7

[2] Weinberger K Q, Blitzer J, Saul L K. Distance Metric Learning for Large Margin Nearest Neighbor Classification [J]. Journal of Machine Learning Research, 2009, 10: 207-244

[3] Bilenko M, Basu S J, Mooney R. Integrating Constraints and Metric Learning in Semi-Supervised Clustering[C]// ICML'04 Proceedings of the 21st International Conference on Machine Learning, 2004. Banff, Canada: ACM New York, 2004: 81-88

[4] McFee B, Lanckriet G. Learning Multi-modal Similarity [J]. Journal of Machine Learning Research, 2011, 12: 491-523

[5] 熊建斌，李振坤，刘怡俊. 半监督聚类算法研究现状[J]. 现代计算机，2009，12：61-64

[6] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005：233-234

[7] 李有锋. 基于颜色和纹理特征的图像检索相关算法研究[D]. 成都：电子科技大学，2009

[8] 杨璐宇. 基于图像 SIFT 特征的图像检索方法[J]. 科技资讯，2009，34：81-82

[9] 雷兰一菲. 基于局部图像特征的目标识别和分类方法研究[D]. 北京：北京化工大学，2011

[10] 王耀南，李树涛，毛建旭. 计算机图像处理与识别技术[M]. 北京：高等教育出版社，2001：149-155

[11] 张琳波，王春恒，肖柏华，等. 基于 Bag-of-phrases 的图像表示方法[J]. 自动化学报，2012(1)：46-54