云模型用于特征加权及降维的算法①

秦彩云

(北京石油化工学院 信息工程学院，北京 102617)

摘 要：高维且不独立的样本特征集使分类的质量降低，提出特征权值计算方法，并用于特征加权及特征选择，根据特征的相似性度量函数计算特征的权重，并根据权重排序去除重要性差的特征，用于解决高维样本集的特征降维问题，特征选择结果与主成份分析结果一致。并建立基于保留特征加权的云分类模型，应用于 iris 数据集和复杂矿石图像的分类，效果良好。

关键词：邻接特征选择；云分类器；相似性

Weighting and Degrading Dimension Algorithm Based on Cloud Model

QIN Cai-Yun

(College of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing 102617, China)

Abstract: A cloud classifier based on swarm particle optimization (PSO) is presented, and used in the classification for multi-dimension object. The digital characteristic of cloud model is expected value Ex、entropy and super entropy He, the membership to which every attribute data of classified object belongs to its attribute set center is presents by 1-D cloud model. The digital characteristic of 1-D cloud model is optimized by swarm particle optimization (SPO). The swarm particle optimization cloud classifier (SPOCC) is built from every attribute cloud model, and used in the classification of iris data set, the experiment result is very well.

Keywords: cloud model; classification; swarm particle optimization algorithm; attribute set

1 引言

基于内容的图像理解是一种综合集成技术，它通过分析图像的颜色、形状和纹理等特征，并基于图像特征通过分类聚类技术对图像内容进行区分和理解。图像内容理解已成为国内外学者研究的主要热点问题，并取得了不少的成果，已经有许多原型系统和实际应用[1,2]。其中基于纹理的图像检索是目前的研究热点之一。纹理是图像分析中常用的特征，一般说来可以认为纹理是由许多相互接近的、互相交织的元素。

国内外学者对图像纹理进行了大量的研究，纹理分析取得了很大的进步，并产生了许多纹理的研究方法，如小波变换[1,2]、共生矩阵[3,4]、马尔可夫随机场[5]等。这些方法大体可分为统计分析法、模型法、频域分析法、结构分析法。其中，统计分析法的应用最为广泛，在计算机视觉和模式识别等领域已取得了丰硕研究成果。从灰度图像中计算出灰度共生矩阵，统计纹理特征。当纹理基元很小并构成微纹理时，统计方法比较有效。在 70 年代早期，Haralick[6]等人从灰度共生矩阵提取出的统计量作为纹理特征向量，并最终提出纹理的灰度级间相关性。由于上述纹理特征之间具有相关性，需要基于特征降维或者特征映射的方法进行变换，使映射后的纹理特征更适用于图像理解及识别。

基于云理论建模图像特征。云理论[2,3]，在知识表达上弥补了粗集理论和模糊集理论的弱点，可以建立定性描述的概念和定量表示的数值之间的转换关系，同时反映定性和定量之间影射的随机性和模糊性。云的数字特征完好地把随机性和模糊性结合到一起，构成定性和定量之间的映射关系作为知识表示的基础。

本研究用云模型表达纹理特征，将训练数据集每

① 收稿时间:2010-10-14;收到修改稿时间:2010-11-20

个类别的属性定义为定性概念，用云模型表达定性概念与类别的属性数值的转换关系，从而将不同类别的对象区分开。云划分的边界是模糊且不确定，比起其它的硬分类技术更加符合实际的数据分布和人的思维方式。

同时提出了类别对象集合和类别特征集合的相似度概念，基于类别特征集合的相似度定义特征的权重。基于 KNN 方法对特征集合进行聚类实验，结果表明对图像内容聚类有很好效果。

2 云模型

云模型[4-7]是定性定量间转换的不确定性模型[4]，设 U 是一个论域，U={x}, T 是与 U 相联系的语言值。U 中元素 x 对 T 所表达的定性概念的隶属度 $C_T(x)$(或 x 与 T 的相容度)是一个具有稳定倾向的随机数,隶属度在论域上的分布称为隶属云,简称为云。$C_T(x)$在[0,1]中取值, 云是从论域 U 到区间[0,1]的影射，即 x∈U x→$C_T(x)$。

云的数字特征用期望值 Ex、熵 σ 、超熵 He 这 3 个数值来表征[5]。

期望值 Ex:概念在论域中的中心值，是最能代表这个定性概念的值，即 Ex 隶属于这个定性概念的程度是 100%。

熵 σ :定性概念模糊度的度量，反映在论域中被这个概念所接受的数值范围。σ 越大，概念接受的数值范围越大，概念越模糊。

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Ex)^2}{n-1}}$$

超熵 He:可谓熵 σ 的方差，反映了云滴的离散程度。超熵越大，云滴离散度越大，隶属度的随机性越大，云的“厚度”也越大。

$$He = \sqrt{\frac{\sum_{j=1}^{m}(\sigma_j - \sigma)^2}{m-1}}$$

云模型的 3 个数字特征值把模糊性(概念接受的数值范围越大，概念越模糊)和随机性(隶属度的随机性，反映为云有一定的厚度)完全集成到一起，构成定性（论域中的概念）和定量 μ（论域中属性值的隶属度）相互间的映射，作为知识表示的基础。影射关系由下式表示：

$$\mu = \exp[-\frac{1}{2}\frac{(x_i - Ex)^2}{2\sigma_j^2}] \quad (1)$$

其中，x_i 表示论域范围内的任一数值，Ex 表示论域的中心值，σ_j 表示论域范围这个概念的一组熵中之一。

3 基于特征相似度的特征选择

3.1 特征相似度

定义 1. 两个类别集合 H_i 和 H_j, H_i, $H_j \in R^D$,$d(H_i, H_j)$是两个类别集合的距离，$d(H_{ik}, H_{jk})$是两个类别集合任一维特征 k 的距离。

$$d(H_i, H_j) = \max_{k=1,2,\ldots D}\{d(H_{ik}, H_{jk})\} \quad (4)$$

$$d(H_{ik}, H_{jk}) = \begin{cases} 0, & if(H_{ik} \subseteq H_{jk} 或 H_{ik} \supseteq H_{jk}) \\ \dfrac{Ex_{ik} - Ex_{jk}}{3(\sigma_{ik} + \sigma_{jk})}, & if(H_{ik} \not\subset H_{jk} 或 H_{jk} \not\subset H_{ik}) \end{cases} \quad (5)$$

其中 Ex_{ik} Ex_{jk} 是类别集合 H_i 和 H_j 的第 k 维特征集合中心，σ_{ik} 和 σ_{jk} 是这两个特征集合的熵，$3(\sigma_{ik}+\sigma_{jk})$ 是两个特征集合之和的二分之一。

定义 2. 两个类别集合 H_i 和 H_j，H_i, $H_j \in R^D$, $s(H_{ik},H_{jk})$是两类别集合第 k 维特征集合的相似度。

$$s(H_{ik}, H_{jk}) = \begin{cases} 1 & d(H_{ik}, H_{jk}) = 0 \\ 1 - d(H_{ik}, H_{jk}), & 0 < d(H_{ik}, H_{jk}) < 1 \\ 0 & d(H_{ik}, H_{jk}) \geq 1 \end{cases} \quad (6)$$

$d(H_{ik},H_{jk})=0$, 第 k 维特征对 H_i 和 H_j 的分类中不起作用;$d(H_{ik},H_{jk})>=1$,H_i 和 H_j 是不相交的特征集合,第 k 维特征可将两个类别集合分开;$0<d(H_i,H_j)<1$ 时，H_i 和 H_j 是相交的类别集合，其相似度越大，对分类的作用越小。

定义 3. m(m>2)个 D 维类别集合的距离为

$$d(H_1, \ldots, H_m) = \min_{i=1,\ldots,m, j \neq i}\{d(H_i, H_j)\} \quad (7)$$

m(m>2)个 D 维类别集合的第 k 维特征集合的 距离为:

$$d_k = \min_{i=1,\ldots m, j \neq i}(d(H_{ik}, H_{jk})) \quad (8)$$

含有 D 个特征的若干个类别集合的分类时，具有最小相似度的特征对分类的贡献最大，计算特征的相似度即能计算出特征在分类中的权重。

3.2 图像特征提取

综合纹理与颜色是两个重要的特征，首先提取纹理特征。对图像进行中值滤波，分别从图像中提取四

个颜色分量(R,G,H 和 I，即红、绿、蓝和亮度)，每个颜色分量图像分成 N*N 的小块，并将颜色量化为 8 级，对每个 N*N 的颜色分量量化后的子块计算其共生矩阵。统计提取出相应的纹理特征，即能量、熵、对比度及匀度四个特征参数。

1）能量　$E=\sum_{i=1}^{g}\sum_{j=1}^{g}m^2(i,j)$　(9)

2) 对比度

$$I=\sum_{i=1}^{g}\sum_{j=1}^{g}\left[(i-j)^2\times m(i,j)\right] \quad (10)$$

3) 熵

$$S=-\sum_{i=1}^{g}\sum_{j=1}^{g}m(i,j)\times\log_{10}m(i,j) \quad (11)$$

4) 匀度　$H=\sum_{i=1}^{g}\sum_{j=1}^{g}m(i,j)/\left[1+(i-j)^2\right]$　(12)

分别计算每个子块的四个颜色分量的能量、熵、对比度及匀度四个统计量，每个子块的纹理特征向量为 16 维：

$$F_i=[FR,FG,FH,FI]=\{f_1,f_2,...,f_{16}\} \quad (13)$$

图像的纹理特征并不相互独立，有一定的关联，这些特征对区分图像内容的贡献率不同，采用特征选择方法，筛除掉不重要的特征，对后续图像内容理解的计算量和准确度都存在的作用。

3.3 基于特征相似度的特征筛选

分类问题的训练样本集是整个类别空间样本中的一部分，从整个空间中随机抽取出的。进行类别建模的首要问题是类别边界的不确定性，云理论最适合不确定问题的建模。类别建模的另一个问题是定义高维样本各特征在分类中的不同作用，即确定特征权重。

假定样本集存在 m 个类别，每个类别 D 个特征)，计算出每个特征的相似度（m 个类别中最大的两特征相似度)，排列各特征相似度，可以决定各特征对分类的重要度，并计算特征的权重。

基于特征相似度的云分类器算法：

1) 类别 i 的样本集 $X_i=\{x_1,\ldots,x_p\}$，类别 i 的第 j 个特征集合用集合 X_{ij} 表示，其中 i=1,2,…,m; j=1,2,…,D;

集合 X_{ij} 的中心 $Ex_{ij}=mean(X_{ij})$;

集合 X_{ij} 左边界 $min_{ij}=min(X_{ij})$;

集合 X_{ij} 右边界 $max_{ij}=max(X_{ij})$;

熵 $\sigma_{ij}=\frac{Ex_{ij}-min_{ij}}{6}$ 或 $\sigma_{ij}=\frac{max_{ij}-Ex_{ij}}{6}$　(14)

2) 计算特征 j 的相似度

按照升序排列特征 j 内的各类别 i 的中心值 Ex_{ij}，用 id(i,j)表示排序后的特征 j 内的类别标识。计算特征 j 内类别的距离 d_j

$$d_j=\min_{i=1,\ldots m-1}\left(\frac{Ex_{id(i+1,j),j}-Ex_{id(i,j),j}}{3(\sigma_{id(i+1,j),j}+\sigma_{id(i,j),j})}\right) \quad (15)$$

特征 j 的相似度用 s(j)表示

if $d_j\geq 1$, then s(j)=0;

else　s(j)=1－d_j。

3) 计算各特征的权重 weight(j)

$$weight(j)=\frac{1-s(j)}{\sum_{j=1}^{D}(1-s(j))} \quad (16)$$

4) 选择特征

从 D 维特征中选择出权值最大的前 m 维特征，使前 m 维的权值之和为总权值的 85%以上。

5) 云分类模型为：

$$A_i=\sum_{j=1}^{m}weight(j)\exp\frac{(x-Ex_{ij})^2}{2\sigma_{ij}^{\ 2}} \quad (17)$$

分类样本在第 i 个云模型中获得最大值，则样本属于第 i 个类别。

$$\mu_i=\exp\left(-\frac{1}{2}\sum_{j=1}^{D}\frac{(x_{jk}-Ex_{ij})^2}{2\sigma_{ij}^{\ 2}}\right) \quad (18)$$

将分类对象送入分类器，得到的最大 μi,值，即表明分类对象属于类别 i。

4 实验及讨论

本文对复杂的矿石图像内容进行分类，为了验证方法的有效性，参照 iris 数据集进行了分类实验对比。

4.1 iris 和矿石图像特征筛选

基于特征相似度的计算方法，充分考虑分类过程各特征的不同作用，计算了 iris 数据四个属性的相似度和权重，表 1 所示。

表 1　iris 数据集的特征相似度和权重特征

特征	Sepal length	Sepal width	Petal length	Petal width
相似度	0.821	0.883	0.580	0.506
权重	0.148	0.097	0. 346	0.408

将 iris 数据由主成分分析方法进行特征映射（表 2），结果与特征相似度方法得出的特征评价结果一致。

表 2 iris 数据集的 PCA 映射的特征值和权重特征

特征	Sepal length	Sepal width	Petal length	Petal width
相似度	0.147	0.021	0.921	0.911
权重	0.053	0.005	0.217	0.728

从采矿现场采集的矿石图像图 1 所示，

图 1 矿石原图

将矿石图像分为 3×3 的子块，每个子块均中提取出 16 维的纹理特征向量，分别为 R、G、H、I 四个颜色分量的能量、熵、对比度及匀度。应用 PCA 方法对图像特征向量进行映射，取贡献率之和超过 90%的前 5 维特征，分别为第 12，3，9，10 和 6 维特征，即 H 分量的匀度、红色分量的对比度、H 分量的能量、H 分量的熵和绿色分量的熵这 5 维特征为主要特征。

通过计算每个特征的相似度，得到的特征权重值进行排序，与 PCA 特征映射的结果一致，筛选出与 PCA 相同的 5 维特征。

表 3 矿石图像的特征值和特征权重重

特征	PCA特征值	特征相似度	特征权重	PCA特征排序	特征权重
λ 1	0.009	0.009	0.003	λ 12	
λ 2	0.023	0.011	0.004	λ 3	
λ 3	2.888	0.532	0.203	λ 9	
λ 4	0.006	0.007	0.002	λ 10	
λ 5	0.193	0.058	0.022	λ 6	
λ 6	1.003	0.312	0.119	λ 7	
λ 7	0.883	0.124	0.049	λ 5	
λ 8	0.010	0.010	0.004	λ 15	
λ 9	1.541	0.442	0.168	λ 16	
λ 10	1.241	0.330	0.127	λ 14	
λ 11	0.005	0.006	0.002	λ 13	
λ 12	7.795	0.973	0.370	λ 2	
λ 13	0.024	0.017	0.007	λ 8	
λ 14	0.098	0.035	0.013	λ 1	
λ 15	0.175	0.061	0.024	λ 4	
λ 16	0.107	0.048	0.019	λ 14	

4.2 iris 和矿石图像分类

根据公式 2 和 3 建立 iris 数据集及矿石图像的云分类器，将表 2 和表 3 中权重代入分类器模型（方程 16），即得到特征相似度云分类器。利用 iris 数据集对特征相似度云分类器进行分类测试，结果如表 4 所示。A1,A2 和 A3 分别表示类别 setosa, versicolor 和 versicolor 的普通云分类器(四维)，模型的分类结果有误差，将少量样本错误地划分到其它类别中，尤其是类别 versicolor 和 virginica 的有错分现象。普通云模型不能很好处理类别边缘数据的分类。

表 4 普通云模型和分解云模型的分类结果

错误率 / 类别	云模型分类结果(%)			基于特征相似度的云分类器的分类结果(%)		
	A_1	A_2	A_3	A_1	A_2	A_3
setosa	100	0	0	100	0	0
versicolor	0	92	8	0	94	6
virginica	0	6	94	0	6	94

矿石图像云分类器和特征相似度云分类器的分类结果如图 2 和图 3 所示，根据公式 15 建立 iris 和矿石图像的基于特征相似度的云分类器。云分类器中不对矿石图像特征进行筛选，保留全部特征。Iris 数据集的分类结果见表 4，特征筛选后 versicolor 类别的分类正确率提高 2%。

图 2 中图像子块特征包括全部 16 维纹理特征及 RGB 三个颜色特征，共 19 维。图 3 中图像子块特征包括特征筛选后保留的 5 维纹理特征及 RGB 三个颜色特征，共 8 维。云分类器的分类结果与特征相似度云分类器的分类效果相比性能明显下降。

5 结论

本文提出了基于特征相似度进行特征筛选的方法，并与云理论结合提出了特征相似度云分类器，对

(下转第 168 页)

果图像受光照不均匀影响较大时，Niblack 算法会过分夸大图像细节部分，Otsu 算法会因为受到强光照影响产生二值化误差。文中算法能够消除光照不均对 QR 码图像二值化阈值选取的影响，对于不均匀光照的图像二值化时由于增加了均匀校正的处理过程而带来了一些时间上的开销，经过试验验证能够满足 QR 识别实时性需求，且二值化效果较好。

5 结语

本文提出的改进的 QR 图像自适应二值化算法，具有兼顾均匀光照与不均匀光照下 QR 码图像二值化阈值选取的特点，能够适应灰度变化范围比较大、以及光照不均匀的 QR 码图像，实用性更加广泛，对后续的 QR 码识别提供了可靠保证。

参考文献

1 中国物品编码中心.GB/T 18284-2000 中华人民共和国国家标准快速响应矩阵码.北京:中国标准出版社,2000.

2 Sahoo PK, SoltaniS, Wong AKC, Chen YC. Survey of thresholding techniques. Computer Graphics, Vision and Image Processing, 1988(41):233–260.

3 Otsu N. A threshold selection method from gray-level histograms. IEEE Trans. Oil Systems, Man and Cybernetics, 1979,9(1):62–66.

4 Niblack W. An Introduction to Digital Image Processing Prentice Hall, Englewood Cliffs, NJ, 1986(25):115–116.

5 Jackway PT. Improved morphological Top-Hat. IEEE Electronics Letters, 2000, 36(14):1194–1195.

6 Hsia SC, Chen MH, Chen TM. A cost-effective line-based light-balancing technique using adaptive processing. Proc. of IEEE Trans. on Image Processing, 2006,15(9):2719–2729.

7 刘悦,刘明业,刘明军.快速响应矩阵码自动识别算法的设计.计算机系统应用,2006,15(6):51–54.

8 孙忠贵.数字图像光照不均匀校正及 Matlab 实现.微计算机信息,2008,24(4-3):313–314.

9 Zhu KH, Qi FH, Jaing RJ. Automatic character detection and segmentation in natural scene images. Journal of Zhejiang University Science A, 2007,8(1):63–71.

10 尤玉虎,周孝宽.数字图像最佳插值算法研究.中国空间科学技术,2005,(3):14–18.

11 沈庭芝,方子文.数字图像处理及模式识别.北京:理工大学出版社,1998.56–58.

(上接第 199 页)

iris 数据集和矿石图像进行特征筛选的实验，特征选择结果与主成分分析的结果一致，分类结果表面特征筛选并加权后分类器的性能明显提高。

参考文献

1 卜东波,白硕,李国杰.聚类/分类中的粒度原理.计算机学报,2002,25(8):810–816.

2 宋远骏,李德毅,杨孝宗.电子产品可靠性云模型评价方法.电子学报,2000,28(12):74–76.

3 扬朝辉,李德毅.二维云模型及其在预测中的应用.计算机学报,1998,21(11):962–968.

4 邸凯昌,李德毅,李德仁.理论及其在集合数据发掘和知识发现中的应用.中国图象图形学报,1999,4(A)(11):930–935.

5 李德毅,孟海军,史雪梅.隶属云和隶属云发生器.计算机研究与发展,1995,2(6):16–21.

6 Li DY. Soft inference mechanism based on cloud models. Proc. of the Joint Int'l Conf and Symposium on Logic Programming. Martin, Germany,1996.38–63.l.

7 邸凯昌.集合数据发掘与知识发现.武汉:武汉大学出版社,2001.

8 陆建江,钱祖平,宋自林.正态云关联规则在预测中的应用.计算机研究与发展,2000,37(11):1317–1320.

9 范建华.基于云理论的数据开采技术及其在指挥自动化系统中的应用.南京:解放军理工大学,1999..

10 Kennedy J, Eberhart RC. Particle Swarm Optimization. Proc. IEEE International Conference on Neural Networks, 1995:1942–1948.