基于 PCA 预处理的图像特征属性约简

孙颖楷 王光学
(北京师范大学珠海分校信息技术与软件工程学院, 珠海 519085)

摘 要 讨论了主分量分析在图像特征属性约简中的应用。运用主成分分析 PCA(principal component analysis)对特征向量进行降维处理,并引入粗糙集理论,对其在特征参数属性优化中的运用进行了探索,利用约简算法剔除识别决策表中不必要的属性,揭示出 CBIR(content based image retrieval)系统中特征条件判断属性内在的冗余性。UCI 数据集处理结果表明 PCA 预处理可排除无关特征量的影响,有效进行特征提取,降低图像识别处理的复杂性。

关键词 PCA 图像 粗糙集 约简
中图法分类号: TP391 **文献标识码**: A **文章编号**: 1006-8961(2007)10-1897-04

Image Feature Attributes Reduction Based on PCA Pre-processing

SUN Ying-kai, WANG Guang-xue
(*School of Information Technology and Software Engineering, Beijing Normal University Zhuhai Campus, Zhuhai* 519085)

Abstract The paper discusses the application of Principle Component Analysis (PCA) in image's feature attributes reduction. After PCA pre-processing, Rough Set theory was introduced, and its application in characterized parameters' attribute optimization was also explored. The unnecessary attributes were eliminated with an attribute reduction algorithm. The inner redundancy of CBIR was revealed. The result of attribute reduction using UCI dataset proved the algorithm can exclude the influence of unused attributes and decrease the complexity of CBIR effectively.

Keywords PCA, image, rough sets, reduction

1 引 言

由于图像本身具有内容丰富,与语言无关等突出特点,图像的数字化应用目前已非常普及,而对图像进行识别及检索,获取有用信息,在军事、医学、农业等许多领域都有广泛和重要的应用。

目前基于内容的图像检索 CBIR(content based image retrieval)已成为研究的热点之一[1],其主要思想是根据图像中物体或区域的颜色、形状、纹理和空间位置关系等底层特征来进行分析和检索。颜色、纹理、形状等是图像的重要信息和特征,描述的是图像或图像区域所对应的景物表面性质,如在医学领域,正常器官表面纹理与病变器官表面纹理具有很大的差异,医生根据纹理特征即可实现病理诊断。在实际应用中,一般通过对图像的颜色、纹理、形状等反映基本物理特征的查找,即通过比较由图像中自动抽取的特征间的相似性,来进行图像检索。

在颜色信息提取方面,Swain 等人提出了直方图交集算法[2],袁等人提出了基于聚类分析的主色提取方法[3]。在纹理特征提取方面,Haralick 等人提出了基于二阶灰度统计特征的共生矩阵方法[4]。在形状特征提取方面,有模板匹配法、形状因子提取法等等。也有综合颜色、纹理、形状、亮度等特征进行识别的算法[5]。

近年来,只采用底层特征的快速图像信息检索技术已得到快速发展,虽然许多学者对采用底层特征图像的自动分类技术进行了研究,但是应用底层特征来有效表达底层语义概念仍很困难。

一个有效的图像检索系统不能仅仅考虑单一的特征,如颜色、纹理或形状来完成,但建立合成的特征向量将导致高维的特征空间,直接对该高维向量

收稿日期:2007-07-16;改回日期:2007-07-25

第一作者简介:孙颖楷(1973 -),男,副教授。2001 年于重庆大学获控制理论与控制工程博士学位。主要研究方向为粗糙集、模式识别等。已在国内外发表论文 20 余篇。E-mail: sunykai@hotmail.com

进行处理比较困难。如何在不影响识别效果的前提下,对高维特征向量进行有效筛选,剔除冗余的、不相容的判断属性,获取最优特征判断核属性,从而降低处理难度,提高识别效率,已是 CBIR 研究方向之一[6,7]。

主成分分析(principal component analysis, PCA)算法[8]是一种常用的特征压缩算法,在许多领域都有广泛运用。本文首先通过 PCA 对特征向量进行降维,然后引入粗糙集(rough set, RS)理论来进行后续属性约简,基于 UCI 数据库的图像识别属性约简结果表明了该方法的有效性。

2 PCA 算法

PCA 是设法将原来众多具有一定相关性的指标,重新组合成一组新的互相无关的综合指标来代替原来的指标,使得新指标既包含原始数据的主要信息,又能更集中地显示出研究对象的特征。

设 n 维输入向量 $\boldsymbol{x}$ 的相关系数矩阵用 $\boldsymbol{R}$ 表示,即 $\boldsymbol{R}=E[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$。其特征值即 $\lambda_1>\lambda_2>\cdots>\lambda_n$ 按降序排列,对应的特征向量分别为 $\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n$。

PCA 算法的目的是寻求正交矩阵 $\boldsymbol{W}$,使得 $\boldsymbol{W}$ 对 $\boldsymbol{x}$ 变换后的矩阵为对角矩阵

$$\boldsymbol{W}=[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n]^{\mathrm{T}},\text{且 }\boldsymbol{\omega}_i\boldsymbol{\omega}_j^{\mathrm{T}}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \tag{1}$$

则 $y_j=\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{x}$, $j=1,2,\cdots,n$,其中,y_j 为向量 $\boldsymbol{x}$ 在单位向量 $\boldsymbol{\omega}_j$ 表示的主方向上的投影,即主元。

通过选取 m 个较大特征值对应的特征向量,从而将高维向量 $\boldsymbol{x}\in\mathbf{R}^n$ 转换为低维向量 $\boldsymbol{y}\in\mathbf{R}^m$, $m<n$。

PCA 通过寻找变量最大的投影轴,判断有多少个独立变量,并将相关量组合成新量,在保留了向量 $\boldsymbol{x}$ 的绝大多数特征信息前提下,通过使用低维的向量 $\boldsymbol{y}$ 来替代原先的维数较多的向量 $\boldsymbol{x}$,实现降维目的,这可大大减少计算的复杂性,同时保证尽可能少地丢失信息,相比于反射映射或交叉相关方法,对样本要求不高。

3 粗糙集理论

粗糙集理论[9,10]由 Pawlak 提出,主要思想是在保持分类能力不变的前提下,通过对知识的约简,导出概念的分类规则。其中知识被认为是一种对抽象或现实对象进行分类的能力。分类过程中相差不大的个体被归于一类,它们的关系就是不可分辨关系,也称等价关系。

不可分辨关系是粗糙集理论的基石,揭示了论域中知识的颗粒结构,也是定义其他概念的基础。

给定一有限的非空集合论域 $\boldsymbol{U}$,定义 $\boldsymbol{R}$ 代表论域 $\boldsymbol{U}$ 中的一种等价关系,则称 $\boldsymbol{U}$ 上的分类族,即知识库 $\boldsymbol{K}=(\boldsymbol{U},\boldsymbol{R})$ 为一个近似空间。

对于子集 $\boldsymbol{X},\boldsymbol{Y}\subset\boldsymbol{U}$,若根据关系 $\boldsymbol{R}$,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的属性不可分辨时,用 $[\boldsymbol{X}]_{\boldsymbol{R}}$ 来表示,它代表子集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 同属 $\boldsymbol{R}$ 中的一个范畴。

若 $\boldsymbol{P}\subset\boldsymbol{R}$,且 $\boldsymbol{P}\neq\varnothing$,则 $\boldsymbol{P}$ 中全部等价关系的交集也是一种等价关系,称为 $\boldsymbol{P}$ 上的不可分辨关系,记为 $ind(\boldsymbol{P})$:

$$[\boldsymbol{X}]_{ind(\boldsymbol{P})}=\cap[\boldsymbol{X}]_{\boldsymbol{R}},\boldsymbol{P}\subset\boldsymbol{R} \tag{2}$$

为了确定对某一概念的支持程度,可分为以下 3 种情况:肯定支持此概念、肯定不支持此概念、可能支持此概念,并分别用 3 个近似集合来表示:正域、负域和边界域。

$$\text{正域:}Pos_{\boldsymbol{B}}(\boldsymbol{X})=\boldsymbol{B}\_(\boldsymbol{X}) \tag{3}$$

$$\text{负域:}Neg_{\boldsymbol{B}}(\boldsymbol{X})=\boldsymbol{U}-\boldsymbol{B}^{-}(\boldsymbol{X}) \tag{4}$$

$$\text{边界域:}Bn_{\boldsymbol{B}}(\boldsymbol{X})=\boldsymbol{B}^{-}(\boldsymbol{X})-\boldsymbol{B}\_(\boldsymbol{X}) \tag{5}$$

其中,$\boldsymbol{B}\_(\boldsymbol{X})$ 及 $\boldsymbol{B}^{-}(\boldsymbol{X})$ 分别为 $\boldsymbol{X}$ 的 $\boldsymbol{B}$ 下近似集和 $\boldsymbol{B}$ 上近似集。

$$\boldsymbol{B}\_(\boldsymbol{X})=\cup\{\boldsymbol{Y}_i\in\boldsymbol{U}\mid ind(\boldsymbol{B}):\boldsymbol{Y}_i\subseteq\boldsymbol{X}\} \tag{6}$$

$$\boldsymbol{B}^{-}(\boldsymbol{X})=\cup\{\boldsymbol{Y}_i\in\boldsymbol{U}\mid ind(\boldsymbol{B}):\boldsymbol{Y}_i\cap\boldsymbol{X}\neq\varnothing\} \tag{7}$$

设 $\boldsymbol{P}$、$\boldsymbol{Q}$ 是 $\boldsymbol{U}$ 上的两个等价关系族,且 $\boldsymbol{Q}\subseteq\boldsymbol{P}$,如果满足:$ind(\boldsymbol{Q})=ind(\boldsymbol{P})$;$\boldsymbol{Q}$ 是独立的;则称 $\boldsymbol{Q}$ 是 $\boldsymbol{P}$ 的一个约简。

$\boldsymbol{P}$ 中所有必要关系组成的集合,称为关系族 $\boldsymbol{P}$ 的核,记作 $core(\boldsymbol{P})$。

令 $\boldsymbol{DS}=(\boldsymbol{U},\boldsymbol{A},\boldsymbol{V},f)$ 为决策系统,$\boldsymbol{C}$ 为条件属性集,$\boldsymbol{D}$ 为决策属性集,$\boldsymbol{A}=\boldsymbol{C}\cup\boldsymbol{D}$,$\boldsymbol{C}\cap\boldsymbol{D}=\varnothing$;$\boldsymbol{V}$ 为属性值的集合;$f:\boldsymbol{U}\times\boldsymbol{A}\rightarrow\boldsymbol{V}$ 指定 $\boldsymbol{U}$ 中每一对象 $\boldsymbol{x}$ 的属性值。

通过对决策表中条件属性进行简化,使得化简后的决策表具有与化简前的决策表相同功能,但条件属性数目更少。可见决策表约简在工程应用中非常实用,同样的决策可基于更少量的条件,而这就是本文引入粗糙集理论进行判断属性约简的目的。

化简后的决策表是一个不完全的决策表,它仅包含那些在决策时所必需的条件属性值,但具有原始知识系统的所有知识。

一个属性子集可以有不止一种约简,因此一个知识表达系统的决策表的约简不是唯一的,也即问题的最小解不是唯一的。这样可以按照某些要求对问题的解进行优化,从中选取适合要求的解的表达形式。下面给出一种基于可辨识矩阵的属性约简方法。

令 $a(x)$ 是记录 x 在属性 $a, a \in C$ 上的值,v_{ij} 表示可辨识矩阵 V 中第 i 行第 j 列的元素,则可辨识矩阵 V 可表示为

$$(v_{ij}) = \begin{cases} a \in C & a(x_i) \neq a(x_j), D(x_i) \neq D(x_j) \\ 0 & D(x_i) = D(x_j) \\ -1 & a(x_i) = a(x_j), D(x_i) \neq D(x_j) \end{cases} \quad (8)$$

其中,$i, j = 1, 2, \cdots, n$。

令 C_{core} 为核属性集合,$p(a_k)$ 为可辨识矩阵 V 中属性 a_k 的属性频率函数,则可以得到基于可辨识矩阵的属性约简算法如下:

(1) 令 $R = C_{core}$;

(2) $Q = \{v_{ij}: v_{ij} \cap R \neq \varnothing, i \neq j\}$, $V = V - Q$, $B = A - R$;

(3) 对所有的 $a_k \in B$,计算在 V 中的 $p(a_k)$,并且令 $p(a_q) = \max_k \{p(a_k)\}$;

(4) $R \Leftarrow R \cup \{a_q\}$;

(5) 重复上述过程,直至 $V = \varnothing$。

由此可见,粗糙集理论实际上与人类的认知特性很相似。而知识的不精确性主要是由于其粒度太大而引起的,解决的方法就是形成知识库的偏序结构,并通过运算,求出最小约简,从而形式化条件属性和决策属性之间的最小依赖关系,不必再针对所有的条件属性集。

4　PCA 预处理

选取 UCI 中的 Image 数据集作为约简对象[11],包含 19 个图像分类特征属性,为了提高运算速度,降低识别所需维数,首先运用 PCA 对数据集进行预处理,由于属性 REGION-PIXEL-COUNT 的样本值一直保持为固定值,该属性可确定为冗余属性,处理时可先行剔除,因此使用其余的 18 个图像特征属性进行处理,计算获得如表 1 所示特征值表。

从表中可以得知,前 14 个主成分分量的累计贡献率已达 100%,也就是说,原所需的 19 维判断分量,现在只需 14 个主成分就可以完全代表先前的系统,而不会带来信息的损失。

根据处理的不同要求,也可选取不同数量的分量进行后续处理。本文选取前 11 个分量,而此时其累计贡献率也已高达 99.238%,再进一步计算在主成分上的载荷值,由于该表较大,因此仅给出其中前 4 个主元的载荷值,如表 2 所示。

表 1　总方差解释情况表

Tab.1　Total variance explained

成分序号	特征值	初始特征值	
		占总方差百分比	积累百分比
1	7.645	42.472	42.472
2	3.183	17.685	60.157
3	1.870	10.389	70.546
4	1.177	6.536	77.082
5	1.024	5.689	82.771
6	0.762	4.235	87.005
7	0.617	3.429	90.434
8	0.545	3.028	93.462
9	0.485	2.692	96.154
10	0.325	1.804	97.959
11	0.230	1.280	99.238
12	0.097	0.541	99.779
13	0.039	0.219	99.999
14	0.000	0.001	100.000
15	3.41E-015	1.90E-014	100.000
16	2.38E-015	1.32E-014	100.000
17	1.21E-015	6.71E-015	100.000
18	5.91E-016	3.28E-015	100.000

表 2　载荷矩阵

Tab.2　Component matrix

属性名称	成分			
	1	2	3	4
REGION-CENTROID-COL	0.011	-0.149	0.106	-0.605
REGION-CENTROID-ROW	-0.539	0.069	0.489	0.054
SHORTLINE-DENSITY-5	-0.037	-0.003	0.430	0.636
SHORTLINE-DENSITY-2	0.148	0.471	0.007	0.522
VEDGE-MEAN	0.144	0.785	0.000	0.032
VEDGE-SD	0.103	0.881	0.018	-0.202
HEDGE-MEAN	0.142	0.844	0.067	-0.017
HEDGE-SD	0.107	0.884	0.034	-0.213
INTENSITY-MEAN	0.969	-0.096	0.174	-0.029
RAWRED-MEAN	0.961	-0.106	0.171	-0.032
RAWBLUE-MEAN	0.984	-0.085	0.119	-0.014
RAWGREEN-MEAN	0.947	-0.098	0.239	-0.044
EXRED-MEAN	-0.858	-0.015	-0.165	-0.003
EXBLUE-MEAN	0.905	-0.004	-0.238	0.083
EXGREEN-MEAN	-0.641	0.022	0.580	-0.138
VALUE-MEAN	0.977	-0.088	0.159	-0.018
SATURATION-MEAN	-0.599	0.025	-0.470	0.084
HUE-MEAN	-0.510	-0.057	0.795	-0.079

表中，第 1 主成分与 INTENSITY-MEAN、RAWRED-MEAN、RAWBLUE-MEAN、RAWGREEN-MEAN、EXBLUE-MEAN、VALUE-MEAN 有较大的正相关；第 2 主成分与 VEDGE-SD、HEDGE-MEAN、HEDGE-SD 有较大的正相关，依此类推。通过主成分计算，可使用新获取的 11 个特征量来代替原先的 19 个特征量进行系统分析，使得识别的处理数据量大为降低。

5 RS 属性约简

这时可根据所获取的 11 个特征属性作为系统的输入参数来构建系统，此时也已经达到降低识别属性数量的目的。由于 PCA 进行降维处理时，并未考虑决策属性 D，但如果最终的处理目的是进行决策判断，我们会继续考虑这些特征量是否都是决策判断中所必需的，也即根据这 11 个条件属性来进行决策，是否依然还会存在冗余的特征量？如果存在，如何将其区分开来并剔除出？这就可考虑利用粗糙集方法来处理。

由于属性值是连续值，而在粗糙集理论中，粗糙集方法对数据的约简是建立在离散数据表的基础上的，因此首先需要将连续属性值进行离散化处理，转化为决策表后再进行后续分析。目前粗糙集理论中的数据离散化处理算法较多，如等距离离散算法、等频率离散算法、Naïve 算法、布尔逻辑算法等，不同的离散化处理可导致最终不同的约简结果。本文选取等频率离散算法来进行离散化处理，将条件属性离散化为 4 个值。

经计算后，该 UCI 图像数据集的判断属性可取 11 个主成分分量中的 6 个，用原数据集进行检验，均能准确识别，可见原先取 11 个主分量还有可约简的空间。

6 结 论

由此可见，在未经属性压缩、约简之前，原来需要用 19 个特征属性来对图像进行识别，在运用 PCA 进行预处理后，再利用粗糙集方法进行属性约简，现在可以只需其中 6 个征兆，即完成了高维判断向量的降维操作。处理结果揭示了原先特征属性内含的冗余性，同时也降低了后续处理的复杂性。

参考文献(References)

1 Wei Na, Geng Guo-hua, Zhou Ming-quan. An overview of performance evaluation in content-based image retrieval[J]. Journal of Image and Graphics, 2004, 9(11): 1271 ~ 1276. [韦娜,耿国华,周明全. 基于内容的图像检索系统性能评价[J]. 中国图象图形学报, 2004, 9(11): 1271 ~ 1276.]

2 Swain M J, Ballard D H. Color indexing [J]. International Journal of Computer Vision, 1991, 17(1): 11 ~ 32.

3 Yuan Xin, Zhu Miao-liang. Content based image retrieval (CBIR) system based on dominant color matching[J]. Journal of Computer-Aided Design & Computer Graphics, 2000, 12(12): 917 ~ 921. [袁昕,朱淼良. 基于主色匹配的图像检索系统[J]. 计算机辅助设计与图形学学报, 2000, 12(12): 917 ~ 921.]

4 Haralick R M, Shanmugam K, Dinstein I. Texture features for image classification [J]. IEEE Transactions on Systems, Man and Cybernetics, 1973, 3(6): 610 ~ 621.

5 Ma W Y, Manjunath B S. Netra: a toolbox for navigating large image databases[J]. Multimedia Systems, 1999, 17(3): 184 ~ 198.

6 Yang Guan-liang, Li Zhong-jie, Xu Xiao-jie. Algorithm of image retrieval based on color-space[J]. Journal of Engineering Graphics, 2005, 26(3): 50 ~ 53. [杨关良,李忠杰,徐小杰. 基于颜色-空间的图像检索算法[J]. 工程图学学报, 2005, 26(3): 50 ~ 53.]

7 Wang Shao-bin, Hao Hong-wei. A new method of CBIR by utilizing feature synthesis and their complement[J]. Control & Automation, 2006, 22(22): 177 ~ 179. [王少彬,郝红卫. 利用综合特征的图像检索及特征互补性研究[J]. 微计算机信息, 2006, 22(22): 177 ~ 179.]

8 Oja E. Subspace Methods of Pattern Recognition [M]. England: Research Studies Press, 1983.

9 Pawlak Z. Rough Sets: The Oretical Aspects of Reasoning About Data [M]. Netherlands: Kluwer Academic Publishers, 1991.

10 Pawlak Z. Rough sets and intelligent data analysis[J]. Information Sciences, 2002, 147: 1 ~ 12.

11 UCI Repository of Machine Learning Database and Domain Theories [DB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html