基于信息粒度的聚类分析及其应用

陈 洁[1] 张迎春[1] 张燕平[1)2)] 张铃[1)2)]

[1)](安徽大学计算智能与信号处理教育部重点实验室，合肥 230039) [2)](安徽大学人工智能研究所，合肥 230039)

摘 要 在处理复杂问题时，通过改变问题所在的粒度空间，不仅可以有效获取对象的特征，而且可去除干扰和非本质属性，使问题易于分析解决。所谓从粒度计算的观点来讨论聚类分析问题，就是认为聚类是在原问题的粒度下(同一问题的最细粒度空间)进行问题分析。为了简化处理，引入不同的聚类相似性函数，其实质就是得到不同粒度空间的等价类。在实际问题求解中，可以根据问题需要取不同相似性函数，以便将问题变换到所需的粒度空间进行处理。为推广其应用，将该思想应用于车牌二值化，提出了基于信息粒度的聚类变换的二值化算法，实现了从彩色 3 维空间到黑白 1 维空间的粒度变换。实验结果表明，该算法所得结果更加切合实际图像，不仅具有普适性，而且有利于下一步的识别操作，尤其对于各种斜车牌、光照不均车牌更具有一定的优越性。

关键词 信息粒度 粒度空间 聚类 车牌二值化

中图法分类号：TP181 **文献标识码**：A **文章编号**：1006-8961(2007)01-0087-05

Analysis and Application of Clustering Based on Information Granularity

CHEN Jie[1], ZHANG Ying-chun[1], ZHANG Yan-ping[1)2)], ZHANG Ling[1)2)]

[1)](*Key Lab of Intelligent Computing & Signal Processing at Anhui University, Ministry of Education, Hefei* 230039)

[2)](*Institute of Artificial intelligence, Anhui University, Hefei* 230039)

Abstract In dealing with complicated problems, the characters of the object can be obtained effectively when the disturbing and nonessential attribute can be wiped off by changing the granular space where the problem located, which make it easier to analyze and solve the problems. In this paper, the analysis of clustering is discussed according to granularity computing. It is assumed that the clustering problems are analyzed under the same granularity (the finest granular space of the problem). The essential of introducing the different comparability functions of clustering is to get a series of equivalence species of different granular space. In practice, problems can be transformed into required granular space, by selecting different comparability functions according to the problem. The transformation form multicolor three-dimension space to monochrome one-dimension can be realized by proposing The License Plate Binary Algorithm based on Information Granularity. Experiments show that the results of this algorithm are more suitable to actual image, have broad generality, and are in favor of recognition following. It is especially predominant in inclined plates or asymmetrical illumination plates.

Keywords information granularity, granular space, clustering, license plate binary

1 引 言

聚类分析是机器学习领域中的一项重要的研究课题。它既可以作为一个单独的工具，用于发现数据库中数据分布的一些深入信息，也可以作为其他数据挖掘分析算法的一个预处理步骤。聚类在很多领域都有广泛的应用，像模式识别[1]、数据挖掘[2]、

基金项目：安徽省自然科学基金项目(0504200208)；安徽省教育厅自然科学研究项目(2005kj053)；国家重点基础研究发展计划项目(2004CB318108)；国家自然科学基金项目(60475017)；教育部博士点基金项目(20040357002)

收稿日期：2005-03-01；**改回日期**：2006-12-21

第一作者简介：陈洁(1982～)，女。2003 年获安徽师范大学计算机系学士学位，2006 年获安徽大学计算机学院硕士学位，现为安徽大学计算机学院教师。主要研究领域为人工智能在智能交通中的应用、图像识别。E-mail：chenjie200398@163.com

图像分割[3]等等，聚类的目的就是发现样本点之间最本质的“抱团”性质，因为在选定了表示样本的特征之后，样本点就表示为特征空间中的一个点，如果再选定样本点之间的相似性函数，那么聚类的结果就确定了。然而，由于从信息粒度的角度看，会发现聚类操作是在一个统一的粒度下进行计算的[4]，因此依据信息粒度原理，就可以先根据需要改变当前聚类所在的粒度空间，然后通过粒度粗细的变化来改进聚类的结果。

在文献[4]中提出了聚类中的粒度理论，聚类处理的对象出于统一粒度。如果改变聚类的相似度函数，则聚类的处理对象将变换到另一粒度空间中。由于粒度的变换可以将复杂的问题简单化，即可将对象简化成若干个保留重要特征和性能的点，以便于分析[5,6]，因此聚类可以进行粒度空间的变换。本文提出一种基于信息粒度的聚类处理方法，该方法首先改变相似度函数，并通过聚类实现不同粒度空间的变换，以解决负责问题，本文将其应用到车牌识别系统的二值化中。实验表明，此种聚类方法相较于当前其他二值化算法，不仅可有效地提高了图像的二值化效果，而且所得结果更接近实际，有利于下一步的识别。

2 粒度信息原理

“人类智能的一个公认特点就是人们能从极不相同的粒度上观察和分析同一问题。人们不仅能在不同的粒度世界上进行问题求解，而且能够很快地从一个粒度跳到另一个粒度的世界，往返自如，毫无困难”[5]。所谓的信息粒度就是指人类在解决处理和存储信息的有限能力上的一种反映，即人类在解决和处理大量复杂信息问题时，由于人类的能力有限，需把大量复杂信息按其各自特征和性能划分成数个较简单的信息块，以方便处理，因此每个如此划分的信息块就被认为是一个粒度[6]。

2.1 信息粒度的形式化描述

文献[2]使用一个三元组(X,f,T)描述一个问题，其中X表示问题的论域；$f(\cdot)$表示论域的属性，可用函数$f: X\to Y$表示；T是论域的结构，其是指论域X中各元素的相互关系。分析或求解问题(X,f,T)，是指对论域X及其有关的结构、属性进行分析、研究。设R是X上的一个等价关系，则对R可以得到对应的商集，记为$[X]$。现在，在$[X]$上定义由T诱导出的拓扑，记为$[T]$，称$[T]$为商拓扑，$([X],[T])$为商拓扑空间。由拓扑学的原理知，从商空间的结构就可以了解原拓扑空间的某些性质。

如果X本身很复杂，则可在X上通过引入一个分类R来得到$[X]$。若R与X的结构T是相容的，则在$[X]$上可诱导出一个商半序$[T]$，于是这就将原来求由w到v的问题转化为在$[X]$中求由$[w]$到$[v]$的问题。由于R是相容的，故$[X,T]\to([X],[T])$是保序的，也就是，当利用适当的分类技术在粗粒度世界讨论问题时，若问题无解，那么在细粒度的原问题上也无解。这样就可缩小求解的范围和加快求解的进度，因为粗粒度世界通常比原世界简单。

从一个较“粗”的角度看问题，实际上就是对X进行简化，即把性质相近的元素看成是等价的，不但可把它们归入一类，并且可将整体作为一个新元素，这样就形成一个粒度较大的论域$[X]$，也就把原问题转化成粗粒度上的问题$([X],[f],[T])$。

2.2 不同粒度世界的关系

不同粒度的划分是为了研究、分析问题的方便，特别是当同一问题需要在不同粒度世界中进行研究时，则要求研究和建立其间的关系。

设R表示由论域X上一切等价关系所组成的集合，可以如下定义等价关系，也就是粒度的“粗”和“细”。

定义 1 设$R_1,R_2\in\mathbf{R}$，如果对于任意元素x，$y\in X$，都有$xR_1y\to xR_2y$，那么就称R_1比R_2细，记为$R_2<R_1$。

可以证明如下的定理[5]。

定理 1 R在如上定义的“$<$”关系下可形成一个完备半序格。

这是一个非常深刻的定理，它揭示了有关粒度的核心性质，因为其他性质都是以此为基础的，有关具体证明请参阅文献[5]。

根据这个定理，可以得到如下的序列：

$$R_n>R_{n-1}>\cdots>R_1>R_0$$

直观地看，如上操作得到的序列和一棵n层的树相对应。设T(即论域的结构)是一棵n层的树，所有叶节点构成集合X(论域)，那么每一层节点都对应着X的一个划分。由于聚类操作得到的聚类谱系图恰好也是一棵n层树，因此必定存在一个等价关系序列与之对应，这也就是粒度和聚类之所以相通的原因。

3　聚类中的粒度原理

众所周知,无论是何种聚类方法,其结果都是将所讨论的空间划分成若干个类,因此在聚类分析中,首先可给出有关论域中各元素之间的相似性的一个度量(如相似度函数),然后给出一个聚类的原则,最后就可根据其相似性所给的聚类原则(最优原则)进行聚类,以便将性质相似的事物聚成同一类,即将原空间划分为若干类,换句话说,就是把当前"细"粒度空间下的不同事物变换成"粗"粒度空间下的同一个事物进行处理,而对于复杂量大的数据情况,聚类也因此表现出它的独特优点。

与所采用的相似度函数对应,就形成了由细到粗的一族等价关系。当采用的相似度较高,则得到的是较细的粒度空间分布,反之采用较低的相似度,则得到粗粒度空间的分布。在实际应用中可以根据处理的实际情况通过适度的聚类来简化计算。本文所提出的彩色车牌二值化就是根据该原理,将原来多类别的 RGB 空间像素点,通过聚类变换粒度空间来得到一粗粒度空间(给出各像素点的决策属性,使其分为 2 个等价类),即是本文的二值化图像。

4　基于粒度信息的聚类应用

通过聚类变换粒度空间来解决问题比仅仅基于原问题空间的方法更加简单,其工作效率也相对更高。下面以它在车牌识别技术中二值化的应用为例来说明。

大家知道,车辆牌照自动识别技术是智能交通系统中的重要研究课题。通常完整的车牌识别系统分为车牌定位、字符分割和字符识别 3 个部分,其中车牌分割是车牌识别的一个重要环节,而车牌图像正确的二值化则是字符分割的基础,目前车牌识别率不高的一个最主要原因就是图像二值化的结果不能正确显示车牌的特征。现今车牌识别系统虽可采用多种方法进行二值化,但其主要方法均是通过考察二值化图像从黑色到白色或者从白色到黑色的变化来进行分割[7],或者在整个二值化图像中搜索怀疑连通区域,再进行筛选[8]。这类二值化方法,由于在光照不均、摄像机畸变、曝光不足、动态范围太窄或者车辆牌照被污染等车辆牌照图像的质量不佳时,处理结果中会产生大量噪音,从而影响了分割的效果,并使系统的整体识别率不高。究其原因还是由于图像二值化方法的不理想,从而使整个自动识别系统难以应用到实际动态识别中。

4.1　彩色车牌二值化简介

所谓图像的二值化即是通过求出一合适的平面来对该空间进行切割,使其一分为二,首先得到一个两色的平面图像。现今车牌识别系统虽可采用多种方法进行分割,但其基本上仍是对二值化后的车牌图像进行处理。这些传统的二值化方法均是通过灰度分割方法实现的,其灰度公式对所有图像均相同,换句话说,就是采用一组平行的平面来切分所有的 RBG 空间(由灰度阈值决定截距)。在车牌自动识别的实际应用中,由于种种外界和人为因素影响而造成各种差异,致使用固定的斜率来切分空间所得的结果往往不能得到理想的二值化图像。

图像二值化就是将像素点从原来"细"粒度空间,变换到"粗"粒度空间。本文提出了经过二次聚类最终实现二值化的算法,即首先通过聚类实现粒度空间变换,而后为了改进聚类效果,在新的粗粒度空间中再次聚类,如此所得结果即为二值化图像。

4.2　基于信息粒度的聚类算法

由相似度函数进行聚类实现粒度变换的算法如下:

(1) 粒度变换

设一个彩色图(求图中各点的 R、G、B 值),则可得原粒度空间(3 维"细"粒度空间)中的像素点集 $\boldsymbol{K}=\{\boldsymbol{p}_i=(r_i,g_i,b_i)\}$。

在 $\boldsymbol{K}$ 中取两点 a,b(即是对初始的两类中心赋初始值,此处所选点应保证所选 a,b 点能大致分别代表字符和背景点,根据经验先取车牌最上面 5 行像素点与最下面 5 行像素点 RGB 的平均值作为点 b 的初始值,而将最中间 5 行的像素值作为点 a 的初始值);然后计算出以 a,b 点连线为法线的平面方程,作为聚类的相似度函数;最后将 $\boldsymbol{K}$ 中的点按"就近原则"进行聚类得 $\boldsymbol{A},\boldsymbol{B}$ 两类,再求类 $\boldsymbol{A},\boldsymbol{B}$ 新的重心 a,b,并重新计算相似度函数,以重新聚类。循环上面的步骤,直到聚类结果满足条件(根据经验值,此处要求类 $\boldsymbol{A}$ 中像素点数目应大于总体像素点数的 30%)。至此原 RGB 空间的多类像素点就分为 2 个等价类,并可分别得到各自的聚类决策属性,其粒度空间由细到粗。

在两点的情况下,"就近原则"就是将图 1 用点 a、b 连线的垂直平分线分成两部分,一部分属于类

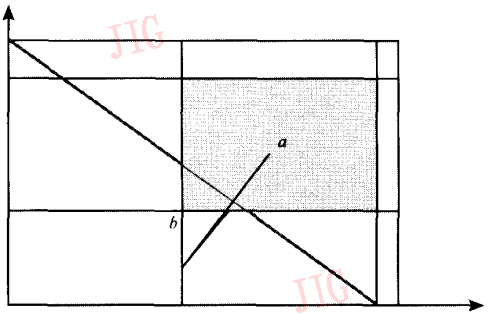

图 1 平面切分

Fig. 1 Plane syncopating

A,另一部分属于 B。原灰度公式为

$1W = 0.3R + 0.59G + 0.11B$[10](单位:lm)

其转换为空间平面方程即为 $F = 0.3x + 0.59y + 0.11z$,由此可以看出,该切平面的斜率是固定的,并由灰度阈值来决定截距,这样就使平面有所改进。

而本算法的任意切分平面方程为

$F(x) = \langle (x - (a-b)/2), (a+b)/2 \rangle = 0$

$F(x) > 0$,x 属于类 A,反之属于类 B。

求类 A 的重心:$a = \left(\sum_i p_i \right) \Big/ n_A$,其中 $\sum$ 表示对类 A 中的 p_i 取和,i 取值从 1 到 n_A,p_i 表示类 A 中的第 i 个像素点,n_A 为类 A 中的像素点数目。

(2) 在新的粒度空间再次聚类

通过上一步的粒度空间变换,就使当前的粒度空间转换为新的粗(1 维)粒度空间。经过此次粒度空间变换后,所得二值化效果依然比较粗糙。如果在当前空间,依据最后得到的字符类中心点再次聚类,则可使字符与背景的边界更加清晰,这也是该算法最终结果图像更切合实际的原因。

设最后得到的两像素点 a,b 分别为字和背景的中心,且车牌中字的像素点是属于类 A,则取 a 为聚类中心点进行聚类。为使聚类结果更优,可动态变化相似度。取 r(开始时,可取 $r = (a-b)/2$)为半径进行聚类,其所得等价类数(即球面 S 中像素点个数)为 n,然后调整 r 的大小 ,使 $n/N \approx 0.3$(N 是车牌中所有的点数)。循环计算直到满足条件。

(3) 二值化

取 K 中的点 x,若落在球面 S 内,则取值为 1,反之,则取值为 0。

通过上述处理,最后得到的即是一幅二值化的图像。对于该图像,可继续进行下一步的分割、识别工作,以最终实现车牌的自动识别(这方面的工作另文讨论)。

4.3 算法分析

本文提出的二值化算法与目前常用的基于灰度的二值化比较,有以下几个特点:

(1)目前流行的二值化算法均是采用同一平面切分 3 维空间的方法,使其分为 2 个等价类,一个记为 0,一个记为 1,以使图像转化为二值图像,与这些算法相比,本文提出的方法则是以要识别的车牌为依据,由于其是以车牌中"字符"与背景的区别为依据来确定切分两类的切分平面,故其精度更高,其二值化过程是,首先进行的是粒度空间的变换,然后通过空间变换直接转换到粗(1 维)粒度空间,最后再次通过聚类实现二值化;

(2)本文所提出的方法比"灰度平面切分的二值化算法"具有更强的泛化能力,这是因为该算法中增加了迭代的过程(相当于 k-聚类法),这是一个逐步求优的过程,故其精度更高;

(3)本算法充分利用了车牌中"字符"与背景的面积比例,同时通过动态地调整相似函数中的阈值,使划分后"字符"的面积保持在 30% 左右,这样就使划分后得到的"字符"的粗细与车牌中字符的粗细基本相同,故其识别精度也就更高;

(4)本算法的计算量比灰度平面切分算法稍高。

4.4 实验结果

为更加清楚证明变化粒度空间的优越性,该算法与当前普遍使用的最大类间方差的方法[11]进行了比较。结果表明,根据所得粒度空间粗细程度的不同,二值化图像亦更加贴近原图,作者同时给出了在一个 2 维粒度空间上的二值化图像结果(由于 2 维介于 3 维和 1 维之间,因此更能说明粒度空间的渐变效果)。所谓 2 维粒度空间即每个像素点具有 2 维属性(h,b),其中 h 是 HIS 颜色空间的 H 分量,b 是RGB 颜色空间的 B 分量。在由这些属性构成的 2 维空间中进行聚类分析,即可得到二值化结果(其算法思想同 3 维空间)。由于此 2 维分量需要另外转换才能获得,其后期处理与 3 维空间无异,因此该算法可直接对 3 维属性进行聚类。实验是利用以上聚类算法对图像进行处理,聚类结果属于 0 的用黑色像素点显示,属于 1 的用白色像素点显示,则可得如图 2 所示的实验结果。由图 2 可见,特别是在有噪音的干扰下,随着粒度变粗,其二值化效果也就更好。

图 2　车牌二值化实验结果

Fig. 2　The result of license plate binary algorithm

例如图 2(b)、图 2(c)均是由于摄像角度造成的图像倾斜,由于采用最大类间方差算法进行处理后的二值化图像易产生噪音,使下一步的分割增加了难度。而基于信息粒度思想算法,所得结果则基本上与原图像一致,而且每个数字的粗细程度也基本保持原样(除了图 2(b)图中数字"7"二值化之后,可能会被误认为"1"),这就为下一步的分割和识别都打下很好的基础。

5　结　论

本文从粒度计算观点来讨论聚类问题,以便将聚类的相似性信息、聚类要求和粒度空间的粗细联系在一起。当面对复杂、难于准确把握的问题时,人们就是采用概略的、由细到粗、由粗到细的多粒度分析法来避免计算复杂度高的困难。人们对聚类结果的要求就是要改变当前事件所处的粒度空间的粒度粗细程度,反过来就可以通过粒度空间变换来改进聚类结果。本文还将粒度变换的思想应用到车牌识别系统中,实验结果表明,不论是蓝车牌还是黄车牌,基于信息粒度的聚类算法都可以很好地处理,而且结果更符合实际车牌。

参考文献(References)

1　Chuang K H, Chiu M J, Lin C C, *et al*. Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means [J]. IEEE Transactions on Medical Imaging, 1999, **18**(12):1117 ~ 1128.

2　Horota K, Pedrycz W. Fuzzy computing for data mining [J]. Proceeding of the IEEE, 1999, **87**(9):1575 ~ 1600.

3　Nozha Boujemaa. On competitive unsupervised clustering[A]. In: 15th International Conference on Pattern Recognition (ICPR'00)[C], Barcelona, Spain, 2000: 1631 ~ 1634.

4　Bu Dong-bo, Bai Shuo, Li Guo-jie. Principle of granularity in clustering and classification [J]. Chinese Journal of Computers, 2002, **25**(8):810 ~ 816. [卜东波,白硕,李国杰. 聚类/分类中的粒度原理[J]. 计算机学报, 2002, **25**(8): 810 ~ 816.]

5　Zhang Bo, Zhang Ling. Theory and Application of Problem Solving [M]. Beijing: Tsinghua University Press, 1990(in Chinese)[张钹,张铃. 问题求解的理论及应用[M]. 北京:清华大学出版社, 1990.]

6　Zhang Yan-ping, Zhang Ling, Wu Tao. The representation of different granular worlds: a quotient space[J]. Chinese Journal of Computers, 2004, **27**(3): 328 ~ 333. [张燕平,张铃,吴涛. 不同粒度世界的描述法——商空间法[J]. 计算机学报, 2004, **27**(3):328 ~ 333.]

7　Choudhury A Rahman, Wael Badawy. A real time vehicle's license plate recognition system[A]. In: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance(AVSS'2003)[C], Miami, FL, USA, 2003: 163 ~ 166.

8　Naito T, Tsukada T, Yamada K, *et al*. License plate recognition method for inclined plates outdoors[A]. In: Proceedings of IEEE International Conference on Information Intelligence and Systems (ICIIS'99)[C], Washington, DC, USA: 304 ~ 312.

9　Ruan Qiu-qi. Digital Image Processing (Second Edition) [M]. Beijing: Publishing House of Electronics Industry, 2003:48 ~ 51. [阮秋琦. 数字图像处理学[M]. 北京:电子工业出版社, 2003: 48 ~ 51.]

10　Duan Zhen. Location of Vehicle Plate and Recognition System Research Based on Structural Machine Learning[D]. Hefei: Anhui University, 2004. [段震. 基于构造性学习的车牌定位与识别系统研究[D]. 合肥:安徽大学, 2004.]