文章编号：1007-791X（2010）02-0111-04

基于色度学空间的多元图表示

张 涛[1,2]，宋佳霖[2]，刘旭龙[2]，洪文学[2,*]

（1. 燕山大学 信息科学与工程学院，河北 秦皇岛 066004；2. 燕山大学 电气工程学院，河北 秦皇岛 066004）

摘 要：针对传统的多元图表示可以表示信息在空间中的几何分布，但无法表示类别概率的缺点，本文提出基于色度学空间的彩色多元图表示。在保留传统多元图表示优点的基础上，集成了色度学维度，利用色度混合原理表示不同类别数据在空间点所占的比例，有利于通过直观的视觉认识类别分布信息以及引入图像处理方法。比传统的多元图表示更适合可视化模式识别的应用。

关键词：多元图表示；色度学；可视化模式识别；概率分布

中图分类号：TP391 **文献标识码**：A

0 引言

数据可视化是可视化模式识别以及人机交互的重要内容[1]。数据可视化的目的是使大量数据在图形空间中进行直观表示，从而符合统计模式识别的基本要求；同时，数据的可视化将抽象的数据在特定的图形空间中具体化表示，也揭示了数据在该空间下的拓扑关系，从而可以利用结构模式识别方法对数据进行进一步分析。由此可以看到，良好的数据可视化过程，无论对于统计模式识别还是结构模式识别都具有重要意义[2]，因此也成为了可视化模式识别的基础。

目前的数据可视化方法有很多，其中多元图表示是一类最基本也是用的最多的分析方法[3]。多元图表示是多元统计分析的方法之一。而多元统计分析是从经典统计学中发展起来的一个分支，是一种综合分析方法，它能够在多个对象和多个指标互相关联的情况下分析它们的统计规律[4]。利用多元图表示对数据进行可视化不但可以满足在统计意义下进行信息的结构表示的要求，而且有助于利用数学工具完成数据的表示与处理。因此成为了可视化模式识别的重要表示方法之一。

但传统多元图表示的应用场合是进行统计学分析，侧重于对数据空间分布结构的表示，而对于数据的类别信息表示不足。造成了多元图表示在可视化模式识别领域中应用的局限。本文针对传统多元图表示的不足，提出基于色度学空间的彩色多元图表示。该表示方法不但具有传统多元图的统计特性和空间分布特性，还具有类别概率分布特性以及与可视化相适应的数据结构，便于后期应用经典模式识别方法和图像处理方法丰富可视化模式识别的架构。

1 传统多元图表示模型

在多元统计分析中，利用多元图表示的主要目的是观察数据在空间中的分布。根据转换后图形表示方法的不同，可分为单点表示与多点表示[5]。就单点表示而言，其本质是将一个m维多元数据$x=\{x_1, x_2, \cdots, x_m\}$转变为在$j\times k$维空间中一点表示的过程。为了满足可视化的要求，当$j\times k$维空间维数过高时，可利用空间分解方法，将高维空间分解为若干个二维或三维可视空间，进行数据的直观表达。由于多元图表示方法较多，为了描述方便，本文以二维散点图为例进行表示。

二维散点图是数据可视化中最常用的方法之一。它可以显示两个变量之间的关系，实现与物理意义简单。每个数据样本对应一个点或者标记，其位置坐标由两个变量的值决定。通过散点图可以观

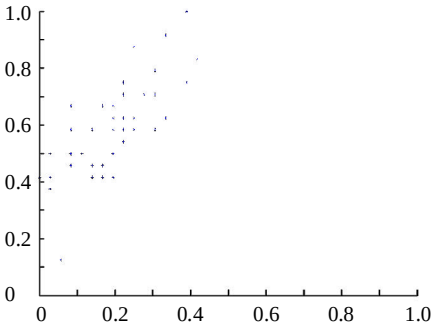
收稿日期：2009-12-11 基金项目：国家自然科学基金资助项目（60904100）

作者简介：张 涛（1979-），男，河北唐山人，博士研究生，讲师，主要研究方向为图像处理、模式识别；*通信作者：洪文学（1953-），男，黑龙江依安人，教授，博士生导师，主要研究方向为信息整合、可视化模式识别和中医工程学，Email：hongwx@ysu.edu.cn。

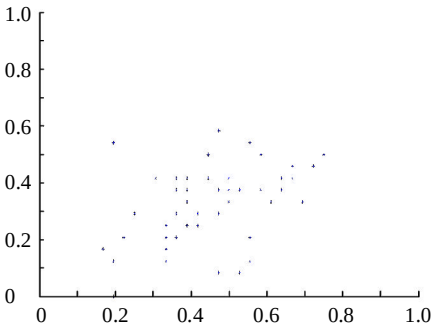
察和理解聚类、离群点、趋势以及相关等数据结构信息[4]，是典型的具有统计与结构表示功能的多元图表示方法之一。二维数据的平面散点图实际上就是以二维数据变量为坐标在平面直角坐标系中描点表示。对于数据矩阵$[x_{ij}]$，其对应的散点图坐标为

$$\begin{cases} x=x_{ij} \\ y=x_{ik} \end{cases} 。 \tag{1}$$

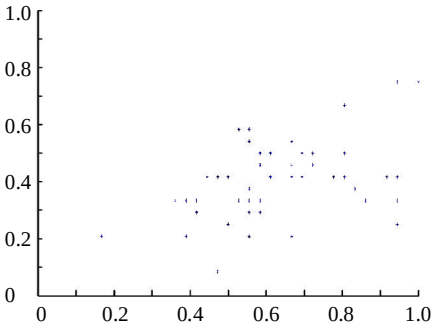
以 Iris 数据的 petal length 与 petal width 特征为例，若使用传统多元图表示，其 setosa 类、versicolor 类与 virginica 的散点图分别如图 1 所示。



(a) setosa 类分布



(b) versicolor 类分布



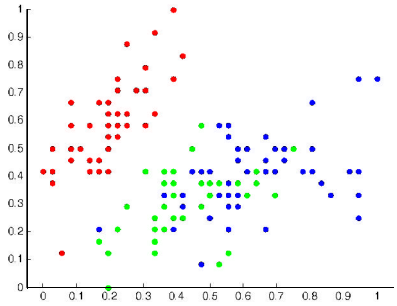
(c) virginica 类分布

图 1 Iris 数据不同类别的分布散点图

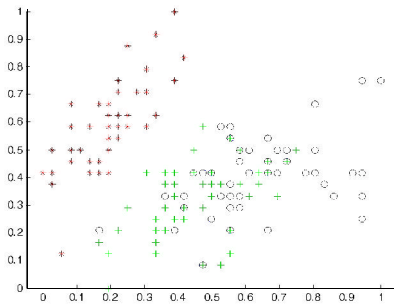
Fig. 1 Scatter plot of different category in Iris

从图 1 的各子图中，可以观察每个类别数据各自的分布状况。但在模式识别领域，更关心的是不同类别数据的相对分布而不是单独类别的绝对分布。而将图 1 各图直接叠加所获得的图虽然可以表示样本点的相对分布，但无法表示类别。为了引入类别信息，需要在不影响可视化的基础上，在传统多元图表示的维数上加入类别维以区分不同类别的分布状况。

目前常用的类别区分办法是将不同类别的数据表示为不同颜色或不同符号，如图 2 所示。图 2 (a) 利用不同颜色对 3 个类别进行了表示。从图中可以直观的观察不同类别数据在空间中的分布，但由于其定义每一个点只能对应为某种类别的颜色，因此对于重叠点数据却无法表示。而重叠点数据恰恰是模式识别中分类器设计的重要参考点。图 2(b) 利用了颜色与符号结合的方式进行了数据表示，由此可以看到，部分 versicolor 类（绿色十字表示）与 virginica 类（蓝色空心圆表示）的样本点发生了重叠，改进了图 (a) 的表示。但无法表示重叠点的类别概率分布，且该表示方法由于同时加入了颜色了符号两种信息，使得后期基于图的处理困难，不适合模式识别领域的应用。



(a) 颜色类别散点图



(b) 颜色与符号类别散点图

图 2 类别散点图

Fig. 2 Scatter in category

2 基于色度学空间的多元图表示

可视化模式识别对数据表示的基本要求除了

要能够对数据进行直观表示外，还要求可在可视化中体现不同类别的概率分布，以及便于后期处理。

2.1　色度学在多元图表示的应用

由第 1 章分析可知，传统的多元图表示可以很好的描述各类数据在图形空间中的分布。但对于分类数据的某些特征，可能出现几个同类别或不同类别数据完全一致的情况，对这类重合点的描述需要使用概率模型而不能单纯的通过其空间位置进行。传统的多元图表示方法无法从空间分布和类别概率两个角度同时对数据进行直观表示。因此，本文结合多元图表示与概率分布模型，提出基于色度学空间的多元图表示（简称色度多元图或色度图表示）概念，用于深化多元图表示中不同类别数据的概率关系，使多元图表示在原有对数据进行结构和单独类别统计表示的基础上，可以直观的描述不同类别间的概率分布情况，为多元图在可视化分类中的应用提供良好工具。

色度学认为：几种不同波长的光以一定比例进行混合，会得到一种全新的主观感受的颜色，该颜色的色度取决于参与混合的各颜色的比例[6]。图 3 为颜色空间的圆表示，其中几乎概括了所有人类可视的颜色。由图 3 可知，人类可感知的颜色可以看作是一个连续分布的圆，即尽管可见光的波长分布为一段线性分布，但对于人类主观感觉而言，可视颜色却是一个连续的循环分布。因此，可以利用有限种颜色作为基色，由基色按照不同比例进行混合，获得整个颜色空间中的所有颜色。著名的三基色定理即以此为依据。

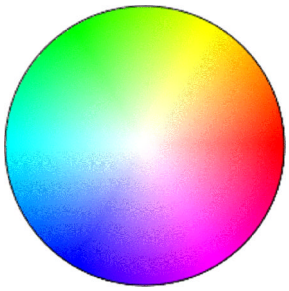


图 3　颜色空间

Fig 3　Color space

因此，在色度散点图中，可以根据类别数目选择适当的基色进行类别的标识。在非重合点，色度直接使用基色表示。在重合点处，若为同类别重合，则基色与自身混合，仍为原始基色，不影响信息表示；若为不同类重合点，则根据该点出类别的概率分布对不同类别的基色进行混合，得到的混合色用于当前点的着色。

2.2　色度多元图的生成方法

对于数据$x=\{x_1,\ x_2,\ \cdots,\ x_m|c_k\}$，其中，$x_i$表示对应特征，$c_k$表示该数据对应的类别。其在直角坐标系下的多元图映射的强度为$f(x_i,\ x_j,\ c_k)$。与传统多元图相比，在数据结构的表示上，色度多元图增加了c_k维表示类别，c_k维将在可视空间内以色度进行表示，因此不会影响原有的空间结构与表示过程。

为了在二维直角坐标下对$f(x_i,\ x_j,\ c_k)$进行表示，则需要对c_k维进行色度学的合成。设在该坐标$(x_i,\ x_j)$下共有l个类别，需要l个基色进行表示。设选取的基色在颜色空间坐标为$\vec{h}_k(r_k,\ \theta_k)$，幅值$r_k$表示饱和度，用于表示类别的混合程度。对于基色，由于表示单一类别，因此饱和度最大，可令$r_k=1$。相角θ_k表示色调，不同的色调对应不同相角。

由于在模式识别中关心的是类别的概率分布而不是绝对数目，需要对空间中相同坐标点的不同类别数据做归一化处理，即

$$f(x_i,\ x_j,\ c_k)=\frac{f(x_i,\ x_j,\ c_k)}{\sum_{k=1}^{l} f(x_i,\ x_j,\ c_k)},\tag{2}$$

然后对归一化后的数据进行色度学合成，混合色的色调θ和色饱和度r分别为

$$\theta=\sum_{k=1}^{l} f(x_i,\ x_j,\ c_k)\theta_k,\tag{3}$$

$$r=\max\left(\sum_{k=1}^{l} f(x_i,\ x_j,\ c_k)r_k,\ l\right)。\tag{4}$$

通过色度学计算，将类别信息转化为色度信息对当前像素点进行着色。

2.3　生成范例与分析

根据 2.2 节所用方法，对于图 2 所示数据利用色度图表示则如图 4 所示。由于 Iris 数据为三类，因此选用常用的三基色表示，分别用红色、绿色和蓝色表示 setosa 类、versicolor 类和 virginica 类。

由图 4 可以看到，图中部分样本点表现为青色，即绿色与蓝色的混合，由此表示了 versicolor

类（绿色表示）与 virginica 类（蓝色表示）的混合比例。由图 5 可以清晰看到，重叠点的颜色均为青色，但部分青色更偏绿色，说明其类别分布中 versicolor 类的样本更多；而另一些更偏蓝色，说明在该点处 virginica 类样本更多。由此完成了对不同类别概率进行表示的过程。另外，在该表示方法中，当当前点无样本时，依据色度学表示为黑色，该表示方法与图像处理习惯一致，便于利用图像处理方法进行后期计算，也使得图像处理方法可以有机的集成至可视化模式识别的框架内。

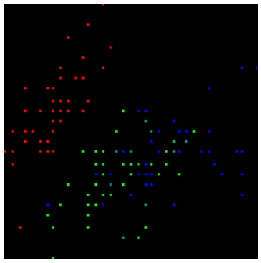


图 4 色度图表示

Fig. 4 Representation in color space graph

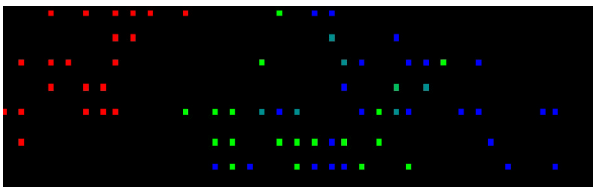


图 5 图 4 的部分放大

Fig. 5 Zoom out a part of fig 4

3 结论

利用色度学空间的多元图表示，在不损失原有多元图表示信息的基础上，增加了类别分布概率信息。将不同类别样本的概率分布进行了直观的表示，更符合人类的分类习惯，同时也便于后期图像处理技术的引入。但当表示类别较多时，如何在保证人类视觉习惯的同时选择合适的基色，是下一步将要继续研究的课题。

参考文献

[1] Ben Fry. 可视化数据[M]. 张羽, 译. 北京: 电子工业出版社, 2009.

[2] Elzbieta PeRkalska, Robert P W Duin, Pavel Paclik. Prototype selection for dissimilarity-based classifiers[J].Pattern Recognition, 2006,39 (2): 189- 208.

[3] 张涛, 洪文学, 景军, 等. 模式识别中的表示问题 [J]. 燕山大学学报, 2008,32 (5): 382-388.

[4] 约翰逊, 威克恩. 实用多元统计分析 [M]. 6 版. 陆璇, 叶俊, 译. 北京: 清华大学出版社, 2008.

[5] 张涛, 洪文学, 景军, 等. 高维数据的 2D 图单点表示原理 [J]. 燕山大学学报, 2008,32 (5): 397-400.

[6] 胡威捷, 汤顺青, 朱正芳. 现代颜色技术原理及应用 [M]. 北京: 北京理工大学出版社, 2007.

Graphical representation based on chromatic space

ZHANG Tao[1,2], SONG Jia-lin[2], LIU Xu-long[2], HONG Wen-xue[2]

(1. College of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China; 2. College of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China)

Abstract: The graphical representation is a excel tool for representing the geometrical distribution of data, but the category of data could not be illustrated by it. In this paper, a new type of representation named chromatic graphical representation is proposed. The novel method integrates the chromatic space into graphical representation and represents the data based on chromatic theory. As the result of that, chromatic graphical representation inherits the merits of traditional and represent the category information by chromatic of the current sample. The illustration of this paper shows the dominance in intuitive color mixing and introducing image processing in visual pattern recognition.

Key words: graphical representation; chromatics; visual pattern recognition; probability distribution