文章编号：1007-791X（2008）05-0397-04

高维数据的 2D 图单点表示原理

张　涛[1]，洪文学[1]，景　军[1]，彭　勇[1]

（1. 燕山大学 电气工程工程学院，河北 秦皇岛 066004）

摘　要：高维数据的单点 2D 图表示是 2D 图表示的重要分支。由于其可在单幅多元图中显示多个观察，因此适用于模式识别领域中的特征选择与分类空间形成。本文根据多元图中单点映射的变量数目的不同对单点图表示进行了对比，并分析了其各自的适用范围与优缺点。

关键词：单点图表示；高维数据；极坐标映射；散点图；星座图

中图分类号：TP391　　文献标识码：A

0　引言

在多元统计理论中，若每一个观察对象都有$p(p\geq 4)$个变量，用统计语言来说，这种数据就是“高维数据”，用几何语言来说，每个观察对象可以看作是p维空间中的一个样本点，即高维空间点。N个这样的点构成的一个样本，即空间的一个高维数据点集。由于人类视觉只能对低维（1~3 维）的空间进行观察，因此对高维空间点集的直观理解相当困难。这不利于模式识别领域的交互式操作和对数据集中新知识的发现。因此，如何将高维数据映射为低维的空间表示，对于可视化模式识别和知识发现至关重要。3 维空间对于人类视觉虽然可视，但局限于目前的显示技术，其直观性不如 2 维（2D）表示。2 维平面图像是直观形象的数据表示方法，它可以帮助研究者进行思维和判断。20 世纪 70 年来以来，统计学家致力于这方面的研究，发展了许多图表示的方法[1]。从数据与图的关系看，这些图表示方法可以分为两类：2D 多点表示和 2D 单点表示。

2D 单点表示是指将观察对象中的全部或部分变量映射为 2D 图中的一个点。该类表示方法可在同一幅多元图中显示多个观察对象，从而发现观察对象之间的关系，适用于数据的特征选择、聚类与分类。在模式识别中，后期处理主要采用域匹配的分类方法。典型的 2D 单点表示有极坐标映射、散点图、星座图、弹簧力表示等。因此，2D 多点表示与 2D 单点表示在 2D 图表示中具有不同的处理方法与应用领域。本文针对 2D 单点表示的各种情况，分析其在模式识别领域中的作用和各自优缺点。

1　2D 单点表示

将高维数据映射为 2D 空间中的一个点，实际上是$\mathbf{R}^n\rightarrow\mathbf{R}^2$的映射，其中$n\leq p$。在不做任何处理的情况下，目前的多元图表示方法中一幅 2D 图最多仅能表示 3 个变量，对于$p\geq 4$的情况可以利用多幅多元图表示。因此，2D 单点图表示可分为$\mathbf{R}^1\rightarrow\mathbf{R}^2$、$\mathbf{R}^m\rightarrow\mathbf{R}^2$（$m=\{2, 3\}$）与$\mathbf{R}^p\rightarrow\mathbf{R}^2$三种情况，即单点单变量表示、单点多变量表示和单点全变量表示。

从统计观点看，以多维数据矩阵表示所要研究的多维数据样本集合，其中样本为X_1，X_2，…，X_n共n个，每个样本X_i对应的变量为x_{i1}，x_{i2}，…，x_{ip}共p个，整个多维数据集合可以表示为矩阵

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

矩阵中的元素为x_{ij}，其中，i=1, 2, …, n，j=1, 2, …, p。2D 单点图表示实际上基于矩阵$\boldsymbol{X}$的列向量表示，这样的表示方法本质上为 Duin 提出的数据相对描述与概念描述[2]。为了保证表示的标

收稿日期：2008-06-20　基金项目：国家自然科学基金资助项目（60474065；60504035）；河北省自然科学基金资助项目（A1217）

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。

准化，对X作归一化处理，处理方法为极差正规化（归一化）变换，如式（1）所示

$$x_{ij}^{*}=\frac{x_{ij}-\min\limits_{1\le i\le n}x_{ij}}{R_j} \tag{1}$$

其中，$R_j=\max\limits_{i=1,2,\cdots,n}x_{ij}-\min\limits_{i=1,2,\cdots,n}x_{ij}$表示极差。

1.1 单点单变量 2D 图表示

单点单变量 2D 图表示即多元图中的每个点仅对应观察对象的 1 个变量，该类图表示适用于观察单个变量的分类能力，代表性方法为极坐标映射。

极坐标映射（Polar Mapping）的主要思想是通过对单一变量加权，映射为极坐标下的幅度值与角度值，从而实现单变量向 2D 空间的映射。对于归一化数据$x_{ij}^{*}\in[0,\ 1]$向极坐标映射，可得极坐标下的坐标值为

$$\begin{cases}r=f(x_{ij}^{*})\\ \theta=g(x_{ij}^{*})k\pi\end{cases} \tag{2}$$

其中，$f(\cdot)$与$g(\cdot)$为优化函数，要求为单调增函数且值域为$[0,\ 1]$，k为旋转因子，决定了数据在空间的角度范围。对 Iris 数据的极坐标表示如图 1 所示，其参数选取为$f(x)=x$，$g(x)=x$，$k=1$。

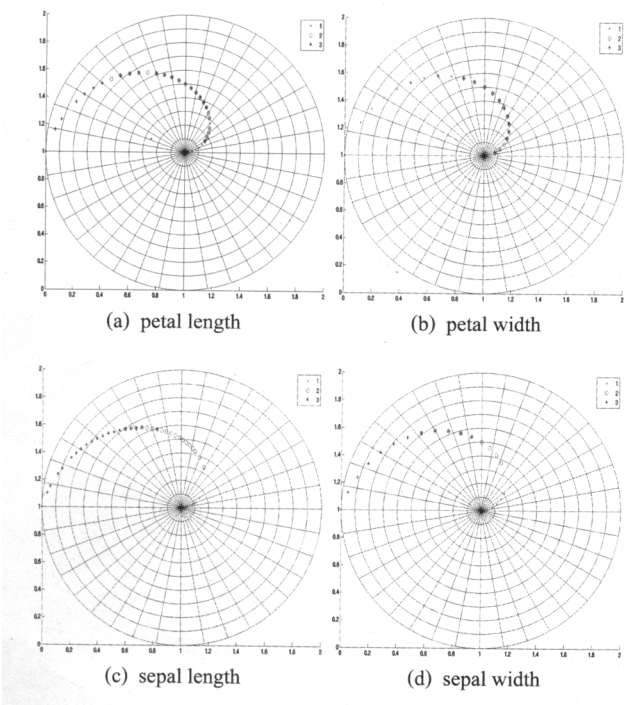

图 1 Iris 数据的极坐标表示

Fig. 1 Polar mapping for Iris data set

由图 1 可见，对于 petal length 与 petal width 特征，若不做任何处理的使用，不同类别间数据混叠严重，不利于分类过程。而 sepal length 特征重叠部分较少，可以作为分类的参考。因此，$\mathbf{R}^1\rightarrow\mathbf{R}^2$适用于对已有特征或变量进行特征选择。但由于每幅多元图只能表示 1 个变量，因此对于p个变量的一组观察，需要使用p个多元图来分析。

1.2 单点多变量 2D 图表示

单点单变量多元图表示的是不同观察对象相同变量间的空间分布，但很多的特征提取算法所提取的特征向量各维之间并非完全相互独立，因此需要在同一多元图中观察多个特征，即通过特征组合完成升维操作，这就需要进行单点多变量 2D 图表示。如需要描述两个特征之间的相互关系，可采用单点双变量图表示。

平面直角散点图（Scatter）是单点双变量 2D 图表示中最常用的方法之一。它可以显示两个变量之间的关系。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。2 维数据的平面散点图实际上就是以 2 维数据变量为坐标在平面直角坐标系中描点表示。对于完成列归一化的数据矩阵$[x_{ij}^{*}]$，其对应的散点图坐标为

$$\begin{cases}x=x_{ij}^{*}\\ y=x_{ik}^{*}\end{cases} \tag{3}$$

对于高于 2 维的多维数据，常用散点图矩阵表示。散点图矩阵可以看作一个大的图形方阵，其每一个非主对角线元素的位置上是对应行的变量与对应列的变量的散点图，而主对角线元素可视为各变量的单点单变量表示。这样，借助散点图矩阵能够清楚地看到所研究的多个变量两两之间的关系。Iris 数据的散点图矩阵表示如图 2 所示。

由图 2 可以看到，由于散点图考虑了特征间的相互联系，与极坐标映射相比其可分性大大增强。同时，由于散点图表示的是特征间的相互关系，因此散点图矩阵中的散点图个数为p^2，对于高维数据，p^2数值过大，很可能产生“组合爆炸（Combinatorial Explosion）”问题。但由于散点图矩阵中为以主对角线为对称轴的对称位置散点图表达数据与空间分布相同，仅进行了空间上的旋转，对分

类贡献相同，因此实际有效散点图为散点图矩阵的上三角或下三角的散点图，个数为C_p^2+p。即便如此，使用散点图进行数据表示时，散点图数目仍然较多。如何自动筛选以保留最具有分析价值的散点图是一个值得深入研究的问题。

在各种散点图中，三角形散点图以正三角形的 3 条高分别表示 3 个变量的坐标轴，从而实现单点 3 变量表示。具体方法参见文献 [3]。

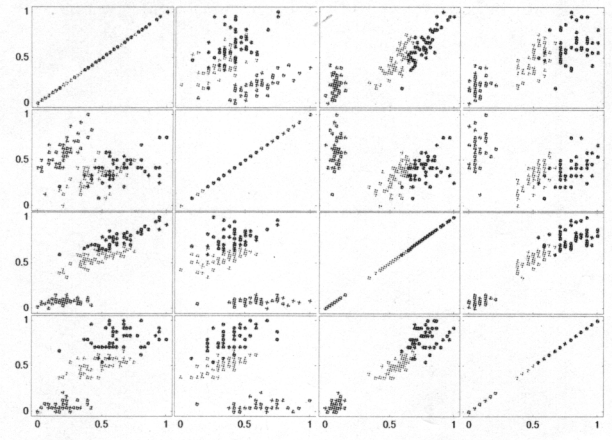

图 2　Iris 数据的散点图矩阵表示

Fig. 2　Scatter matrix of Iris data set

1.3　单点全变量 2D 图表示

为了避免“组合爆炸”问题，可以利用单点全变量 2D 图表示。该表示方法实质上是对各变量信息进行融合或降维，然后在多元图上进行表示。与前两种表示方法相比，单点全变量表示丢失了原始高维数据的大部分信息，是一种最简单、功能最弱的多维数据处理方法。从映射的角度，可以做如下解释：设$\vec{P}_g^*=(x_{g1},\ x_{g2},\ \cdots,\ x_{gn})$与$X_g$意义相同，$\mu$为间隔尺度，$1\le\mu\le n$，$m=[n/\mu]$为将要降到的维数。将$\vec{P}_g^*$映射到一个 2 维复平面$C$上的点$\vec{Q}_g^*$：

$$\vec{Q}_g^*=\psi\left(\vec{P}_g^*\right)=\sum_{k=0}^{m-1}(\lambda_k x_{gk+1})e^{i\frac{2\pi}{n}k} \tag{4}$$

其中，$\lambda_k\in[-1,\ 1]$是一个自调整权值。由于降维过程中存在一定的信息损失，对于分类来说是维数大幅度减少后会使类间的可行性降低，信息补偿就是通过调整参数实现类间可分性达到甚至超过原始高维数据的情况。采用调整参数λ_k来实现损失信息的补偿。λ_k参数的调整过程，首先，根据数据的统计信息，得到不同变量的显著性指标，将得到的变量显著性曲线给对应的λ_k赋值，得到最优的λ_k曲线，也可根据先验知识，手工调整部分λ_k的值，使得不同类别的样本可分性最好，即实现了对多维数据降维的信息补偿。星座图、弹簧力表示等都是单点全变量表示的典型方法。

星座图（Constellation Graph）就是将n个观察对象显示在一个半圆内，每个对象用一颗星表示，同类的对象组成一个星座，很像天文学上表示星座的图象，故名星座图。星座图法是一种非常直观的方法，对多个变量的观察对象在不同权重下进行汇总时，既能体现统计数据的统计结果，同时还能反映数据的均衡性，使用起来极其方便。根据样本点的位置可以直观地对各样本点之间的相关性进行分析。利用星座图还可以方便地对样本点进行分类，在星座图上比较靠近的样本点比较相似，可以分为一类，相距较远的点则说明相应样本点的差异性较大。样本数据X_i的星座图坐标为

$$\left(\sum_{j=1}^{n}\omega_j\cos\xi_{\omega j},\ \sum_{j=1}^{n}\omega_j\sin\xi_{\omega j}\right) \tag{5}$$

其中，$\xi_{ij}=\frac{x_{ij}-x_{\min,\,j}}{R_j}\pi$，$\{\omega_j\}$为路径权值，$\omega_j\ge0$，$\sum_{j=1}^{n}\omega_j=1$。

图 3 为 Iris 数据的星座图表示，不同类别的数据分布于不同的空间。

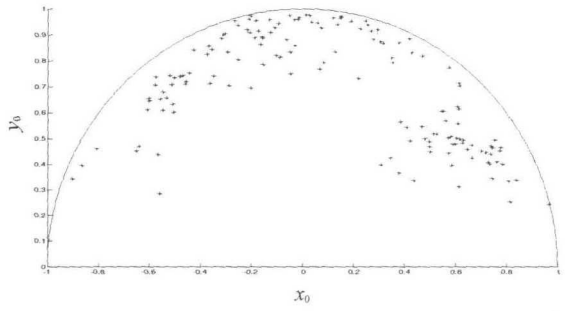

图 3　Iris 数据的星座图表示

Fig. 3　Constellation graph for Iris data set

由式（5）可知，星座图中星的位置实际上可以看作是各变量的加权和，该过程为简单的信息融合过程，可以通过交互方式完成。Mayumi Oyama-Higa 的研究小组开发了交互式星座图表示软件并将其应用于老年痴呆症的分析中 [4]。

2　结论

本文针对典型的高维数据 2D 单点表示问题进

行了分析。由于单点表示具有单图表示多个观察对象的特性，因此其适用于模式识别领域中的特征选择与分类空间形成问题。通过前文的对比可以发现：在 2D 单点表示中，单点单变量是对各变量的独立表示，从中可以有效地分析出各变量的作用，方便进行特征选择；单点多变量表示是对各变量的组合表示，从中可以分析各变量间的相互关系，其结果更有利于分类，但可能产生组合爆炸问题；单点全变量表示则相当于对各变量的融合表示，其很好地解决了组合爆炸问题，但由于其实质上是对各变量的融合表示，因此其最终效果与图的选择和加权关系密切。在实际使用中可以根据不同的应用场合进行选择。

参考文献

[1] 李伟明. 多元描述统计方法 [M]. 上海: 华东师范大学出版社, 2001.

[2] Duin R P W, Pekalska E, Paclík P, et al.. The dissimilarity representation: a basis for domain based pattern recognition?[C]// Pattern representation and the future of pattern recognition. ICPR 2004 Workshop Proceedings. Cambridge, United Kingdom, 2004: 43-56.

[3] 洪文学, 李昕, 徐永红, 等. 基于多元统计图表示原理的信息融合和模式识别技术 [M]. 北京: 国防工业出版社, 2008.

[4] Mayumi Oyama-Higa, Miao Teijun, Yuko Mizuno-Matsumoto. Analysis of dementia in aged subjects through chaos analysis of fingertip pulse waves [C] //IEEE International Conference on Systems, Man and Cybernetics. Taipei, 2006: 2863-2867.

2D graphical representation reflecting high-dimensional observation with sole point

ZHANG Tao[1], HONG Wen-xue[1], JING Jun[1], PENG Yong[1]

(1. College of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China)

Abstract: 2D graphical representation which reflects high-dimensional observation with sole point is one important branch of 2D graphical representations. The representation is competent for feature selection and class space building as it can display a set of observations in one graph. In this paper, the typical representations are analyzed according to the relationship between the sole point and number of attributes. The advantages, disadvantages of each representation and its applicability are also discussed.

Key words: graph representation with sole point; high-dimensional data; polar mapping; scatter; constellation graph