文章编号：1007-791X（2008）05-0435-05

基于多维数据列向量 2D 图表示的多维筛可视化组合分类器

洪文学[1]，张 涛[1]，宋佳霖[1]，王金甲[1]，徐永红[1]

（1. 燕山大学 电气工程学院，河北 秦皇岛 066004）

摘 要：提出一种新型的可视化组合分类器——多维筛分类器。该分类器集成了属性数据的 2D 图表示、图像处理与模式识别技术。其根本思想是将数据矩阵的属性数据映射为 2D 多元图，然后将多元图通过像素图及图像处理技术转换为子分类器，利用组合分类器规则将各子分类器构成多维筛可视化组合分类器。通过引入交互式方法，选择最优多元图用于分类。通过对 Iris 数据集的分类试验表示，散点图多维筛与极坐标多维筛的分类精度可以达到 98.67%和 97.33%。

关键词：模式识别；可视化组合分类器；图表示；交互式

中图分类号：TP391.4 文献标识码：A

0 引言

近年来，可视化在学术研究与应用方面均得到突破式的增长[1-3]。将可视化引入模式识别领域完成用户与系统的交互，从而使人类能力与模式识别算法集成，可使模式识别更易于理解。多元数据图表示方法[4]将抽象的数据以图的形式可视化。尽管在图表示在数据降维[4]，信息融合[5]和模式识别[6]等方面已经进行了深入的研究，但基于多元图形特征的分类器设计仍然是一个需要进一步解决的学术问题。

较理想的可视化分类器至少应满足以下条件：

1）符合数据本身特点。传统分类方法强调同一数据对象中不同属性之间，即数据矩阵的行向量之间的关系，但该关系本身具有一定的不可比性和不确定性。比如，数据测量时量纲的变化会导致其特征的改变。因此，采用列向量（即不同数据对象的同一属性）对数据进行分析从而获得相应的分类依据更为符合数据本身的特点，也符合 Duin 提出的相对描述与概念描述的学术思想[7]；

2）分类器具有良好的解释性。神经网络分类器虽然具有良好的分类性能，但由于解释性较差而受到质疑，数据的可视化为数据分布提供了良好的解释基础。多元图表示的图形特征和原始数据特征间具有确定的映射关系，可作为可解释的分类器基础。而且多元图表示在特征提取、特征选择等方面的成功应用，为基于多元图表示的数据分类提供了良好的理论基础；

3）分界面的生成应符合人类视觉特征与分类习惯。许多分类算法（如线性判别、支持向量机等）的几何解释早就为人所知，但直到最近几年，基于几何的模式识别方法在国内外才开始引起人们的关注。可视化技术的优势是将数据表达为符合人类视觉特性的表达方式，因此，基于可视化技术的分类器分类界面应充分利用这一特性。基于人类视觉的图像处理技术的引入也成了一个必然的解决方法。

本文提出多维筛可视化组合分类器（Visual Combining Classifier, VCC）。该分类器集成了模式识别、多元图表示与图像处理等相关领域知识，并符合理想分类器的设计原则。

1 多维筛分类器基本原理

多维筛分类器的基本原理如图 1 所示。由图可

收稿日期：2008-06-18

作者简介：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融合、可视化模式识别和中医工程学。

以看出，多维筛由 3 部分组成：2D 多元图表示、2D 像素图表示和分类器组合。2D 多元图表示负责将多维数据矩阵的列向量（训练样本的同一属性）映射为 2D 的图表示。2D 像素图表示则将 2D 多元图转换为像素图并利用扩展算法对连通区域扩展，利用混合色表示类别的重叠情况。扩展后的图象可以视为当前数据集的一个分类空间或自分类器。分类器组合则是将这些子分类器集成为一个组合分类器，实现最终的分类。由于像素图为可视化表示，因此该分类器属于可视化组合分类器。

图 1 多维筛分类器框图

Fig. 1 Diagram of the VCC

1.1 多维数据的 2D 图表示

设矩阵X表示多维数据集合

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

其中，x_{ij}为第i个观察对象的第j属性。$[x_{i1}, x_{i2}, \cdots, x_{im}]$为一次观察的$m$个属性。$[x_{1i}, x_{2i}, \cdots, x_{ni}]^{\mathrm{T}}$是所有对象第$i$个属性的集合。对大多数数据集，为了将数据在特定的区域内进行多元图表示，需作归一化处理。由于采用列向量作为分类依据，因此归一化如式 (1)

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

其中，$R_j=\max\limits_{i=1,2,\cdots,n} x_{ij}-\min\limits_{i=1,2,\cdots,n} x_{ij}$。归一化后的数据分布范围是$[0, 1]$，有益于数据的分类。

将抽象数据x_{ij}^{*}映射为一组 2D 多元图的过程如下

$$\{x, y\}=g(x_{i1}^{*}, x_{i2}^{*}, \cdots, x_{im}^{*}) \tag{2}$$

函数$g(\cdot)$根据所使用的多元图类型不同而不同。

图表示虽然可以直观的表示数据的分布，但该分布不一定适合分类，因此需要进行非线性优化。非线性函数$f(x)$的选取原则为：

条件 1：$f(x)\in[x_{\min}, x_{\max}]$，当$x\in[x_{\min}, x_{\max}]$；

条件 2：如果$x_1\ge x_2$，则有$f(x_1)\ge f(x_2)$。其中，$x_1\in[x_{\min}, x_{\max}]$且$x_2\in[x_{\min}, x_{\max}]$。

条件 1 限制了$f(x)$的值域范围，要求与定义域相同，其目的是保证经过非线性变换之后数据分布区间不发生变化，否则可能需要对后续过程重新调整。条件 2 要求$f(x)$为单调函数，以避免数据相对关系的混乱，影响紧支性。任何满足该条件的函数均可用于多元图的非线性优化。

1.2 像素图与区域扩展

多元图的最初目的是表示数据，而不是用于数据分类。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。结合多元图表示与概率分布模型，提出基于色度学的类别分布像素图（简称像素图）的概念，用于深化多元图表示中不同类别数据的概率关系，使多元图表示可以直观的描述类别分布情况，为多元图在可视化分类中的应用提供良好的工具。

在色度学中，色度空间的任何颜色均可由基色按一定比例合成。一般情况下，选择的基色为红色、绿色和蓝色，但不绝对。在像素图中，基色表示相应的类别，而重合点则由该店按类别概率使用混色色表示。当表示的类别超过 3 个时，可以利用补色作为新的基色，并利用新的混合色与饱和度表示。图 2 表示了散点图及和它的像素图。通过比较，可以发现像素图直观表示了重合点的类别分布情况。因此，像素图更适合于分类的情况。更为重要的是，像素图更有利于系统的可视化，从而带来更好的交互式效果。

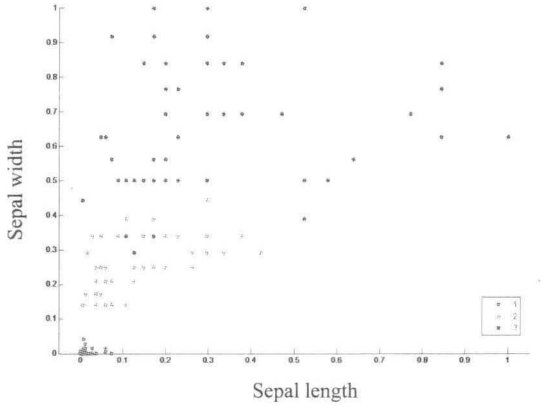
通过像素图，可以更好的对数据进行可视化。但分类的最终目标是要获得分类界面。传统模式识别方法中，一般通过数学方法或几何方法求解分界面。这种方法虽然具有良好的理论基础，但未考虑人类视觉因素。如果采用此方法完成多维筛分类界面的求取，将削弱前期的可视化特色。因此，初步设计了基于图像扩展原理的分类器分类界面计算，其规则如下：

对图像中任一像素$f(x+i,\ y+j)$（$f(x,\ y)$表示在像素图中，坐标为$(x,\ y)$处像素的色度值，$f(x,\ y)$=0表示坐标为$(x,\ y)$处像素的色调为黑色）则区域扩展后该像素点为

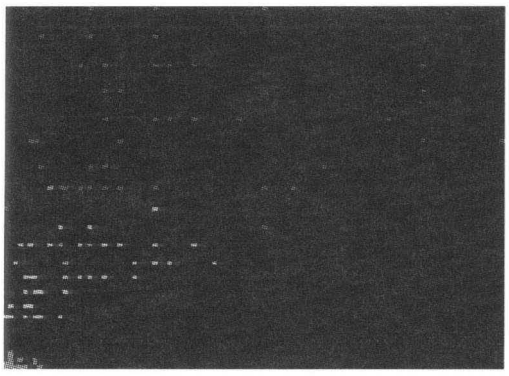
$$g(x+i,\ y+j)=af(x,\ y)+bf(x+i,\ y+j)+$$

$$cf(x+2i,\ y+2j) \tag{3}$$

其中，加权系数a、b、c的大小可根据图 3 的流程进行确定。对该过程进行循环，直至图像中所有的点均非零（即图像中不再存在黑色点，对应于类空间中所有区域均对应类别及相应概率），即获得数据空间的模式分类结构，分类器的设计也由图形空间过渡至类空间。对图 2（b）扩展后的图象如图 4 所示，其可以认为是一个子分类器。对于任意未知类别的数据，仅通过与训练样本相同的数据变换，对照其落入分类图坐标颜色即可判断其所属类别及概率。



(a) 散点图



(b) 像素图

图 2　散点图与像素图比较

Fig. 2　Comparison between a scatter and its pixel graph

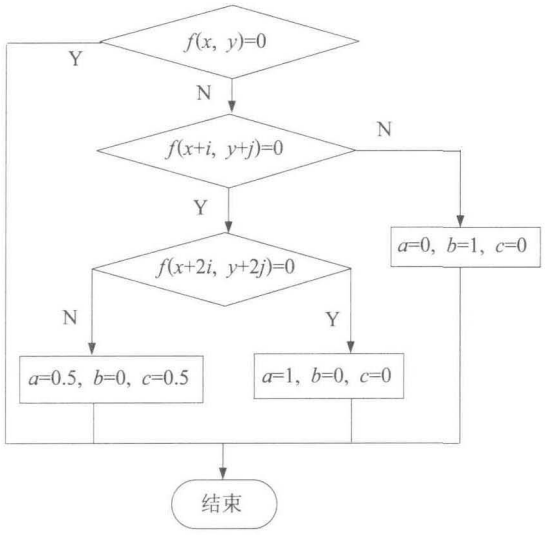


图 3　权重确定流程图

Fig. 3　Flow chart for weighting

图 4　子分类器示例

Fig. 4　An example of sub classifier

1.3　分类器组合

区域扩展后的像素图实际上已经从特征空间过渡到了类空间，所获得的分类模式具有良好的可视化特色，可以直接根据测试样本坐标所对应的颜色值获得其所属类别及概率信息，因此，基于区域扩展技术的多维筛分类器是一种基于多元图表示的可视化分类器，可以作为分类器直接使用。同时，对于特征向量维数高于 1 维的数据，其列向量多元图表示必然由多个多元图共同表示（比如 Iris 数据集），可以认为是由多个分类器共同完成分类任务的组合分类器。多维筛可视化组合分类器的实现过程如图 5 所示。

对于待分类数据，依训练数据参数进行预处理、特征提取与特征组合，对组合后的图形特征与多维筛中对应分类器比较，获得该样本属于c类的概率$p(c|x_i)$，最后通过判决规则对各维分类器的判决结果进行分析，获得最终判决结果。

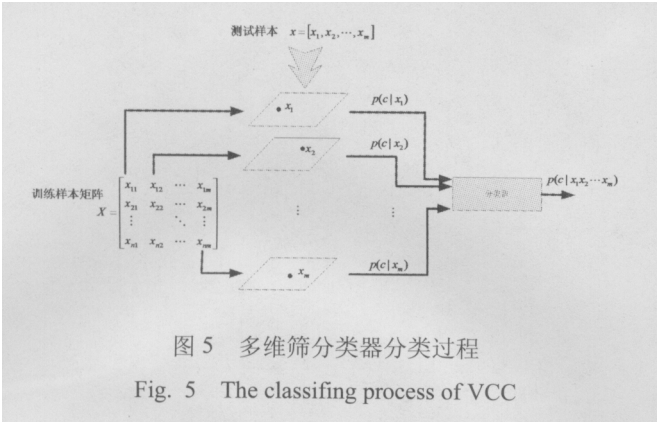


图 5 多维筛分类器分类过程

Fig. 5 The classifing process of VCC

2 基于 Iris 数据集的多维筛实验

本文针对多维筛分类进行了一系列的实验以验证其性能。为了保证通用性，选择的测试数据集为模式识别领域广泛使用的 Iris 数据集，且该数据集各属性值可以直接作为特征使用而无须特征提取过程，这样可以最大限度的保证实验的客观性。

本文选择散点图与极坐标映射作为多维筛的图表示方法。作为典型的图表示方法，散点图可以在一幅图中反映两个变量的相互关系，而极坐标映射仅表示单个变量。测试方法为留一法，即 LOOCV (leave one out cross validation)，该方法被认为是精度测评的基本方法，且比交叉验证更可靠 [8]。

Iris 数据集的散点图表示如图 6 所示。由于矩阵关于对角线对称，因此仅使用下三角 6 个矩阵描述数据分布即可。图 7 和图 8 即针对这 6 个矩阵优化后的像素图表示和分类界面表示。图 9 给出了不同非线性参数对散点图多维筛分类精度的影响。从图中可以看出非线性参数对精度的影响，精度最高可以达到 94.72%, 接近于利用 1 近邻和 3 近邻分类器 95.33%的精度。

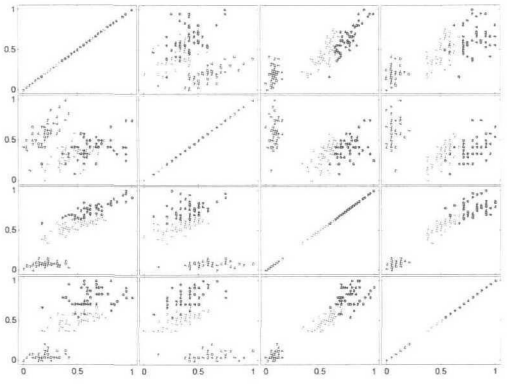


图 6 Iris 数据集的散点图矩阵表示

Fig. 6 Scatter matrix of Iris date

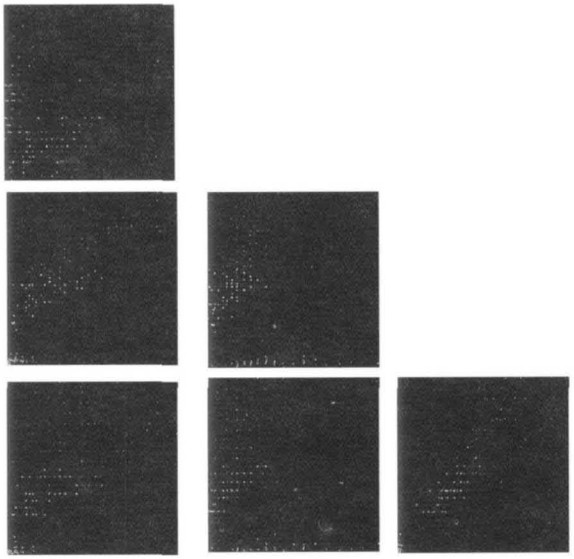


图 7 像素图表示

Fig. 7 Pixel representation

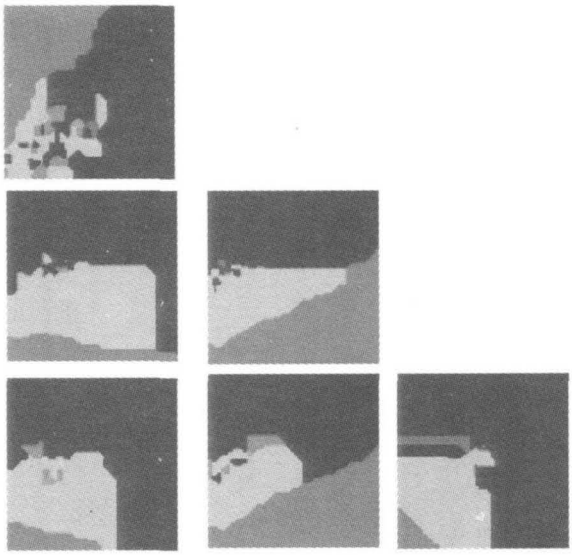


图 8 子分类器空间

Fig. 8 Sub classifier space

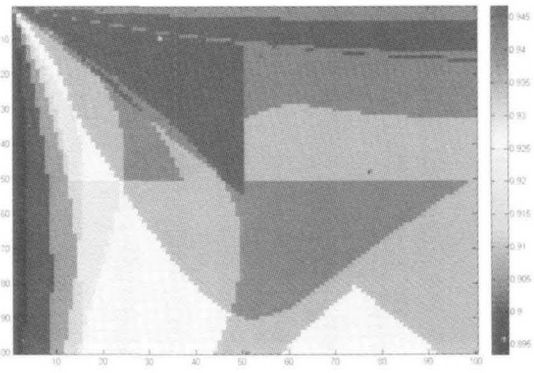


图 9 散点图多维筛非线性参数与精度关系

Fig. 9 Relationship between nonlinear parameters and precision for scatter VCC

以上实验为自动分类过程，没有用户参与。作为可视化分类器，其重要特点就是与用户的交互。通过交互引入人类知识，从而降低复杂度并提高分类性能。

通过可视化，可以发挥人类快速选取子分类器界面的能力，不必过于担心计算复杂度，因此可以引入组合特征作为新特征。图 10 为本研究组选出的最优像素表示组合。左图为 $(\text{petal width})^{1.1}$-$(\text{sepal length})^{3}$ 与 $(\text{sepal width})^{3}$ 的散点图得到的像素图，右图为 $(\text{sepal length})^{3}$ 与 $(\text{sepal width})^{3}$ 的组合，其中，petal width, sepal length, sepal width 均为 Iris 特征名称。由这两个子分类器构成组合分类器。显然，通过引入交互式方法，分类复杂度降低。同时，分类精度达到了 98.67%，高于目前的主流分类器。

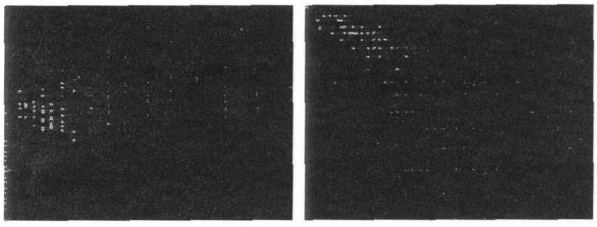


图 10　最优子分类器像素图

Fig. 10　Pixel graphs for best sub classifiers

多维筛另一个重要的优点为灵活性大。由于图表示方法很多，理论上每种图表示方法都可以生成一种多维筛分类器，不同类型的图表示多维筛对应的分类精度也会不同。本文对极坐标映射下的多维筛进行了测试。由于极坐标映射每幅图中仅表示一个变量，因此其多维筛的组合结构比散点图简单。当 k=1 时，其精度如图 11 所示。其最佳精度为 97.33%，高于 SVM 的 96.66%和 3 近邻的 95.33%，但低于交互式散点图多维筛分类器的 98.67%。以上实验证明，多维筛分类器的性能已经接近甚至超过主流分类器水平。

3　结束语

本文提出一种基于多元图表示的新型可视化组合分类器，该分类器是模式识别、多元图表示、图像处理及相关技术的有机结合，这些技术保证了多维筛分类器可视化特性。所做的分类器实验不仅证明了其可行性，更重要的是，由于引入了交互方式，使得分类性能得到提高。关于自适应非线性优化、像素图改进等还需进一步的研究。

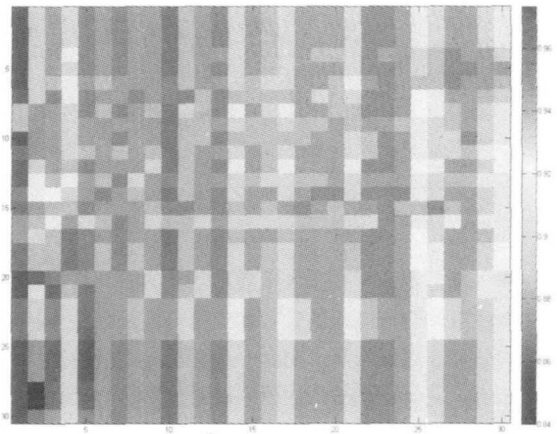


图 11　极坐标多维筛非线性参数与精度关系

Fig. 11　Relationship between nonlinear parameters and precision for polar VCC

参考文献

[1] Janez Demšar, Gregor Leban, Blaž Zupan. FreeViz—An intelligent multivariate visualization approach to explorative analysis of biomedical data [J]. Journal of Biomedical Informatics, 2007,40 (6): 661-671.

[2] Peter C C Wang. Graphical representation of multivariate data [M]. Orlando: Academic Press, 1978.

[3] Daniel A Keim, George G Robertson, Jim J Thomas, et al.. Guest editorial: special section on visual analytics [J]. IEEE Transactions on Visualization and Computer Graphics, 2006,12 (6): 1361-1362.

[4] Liu Wenyuan, Meng Hui, Hong Wenxue, et al.. A new method for dimensionality reduction based on multivariate feature fusion [C] //IEEE ICIT. Shenzhen, China, 2007: 108-111.

[5] Xu Yonghong, Hong Wenxue, Li Xin, et al.. Visual pattern recognition method based on optimized parallel coordinates [C] // IEEE ICIT. Shenzhen, China, 2007: 127-132

[6] Xu Yonghong, Hong Wenxue, Li Xin, et al.. Parallel dual visualization of multidimensional multivariate data [C] // IEEE ICIT. Shenzhen, China, 2007: 263-268.

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

[8] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection [C] //Proc of the 14th Int Joint Conference on Artificial Intelligence. Montreal, Canada, 1995: 1137-1143.

（下转第 460 页）

算法在各种空间下，其分类误差都是随着样本集数的增加而减少的，而 SVM 的性能在各种空间下表现都比较稳定。

参考文献

[1] 王爱民，赵忠旭，沈兰荪. 中医舌象自动分析中舌色苔色分类方法的研究 [J]. 北京生物医学工程，2000,19 (3): 136-142.

[2] 张衡翔，李斌，姚鹏，等. 中医舌象自动分类方法研究 [J]. 北京生物医学工程，2006,25 (1): 47-50.

[3] 李晓宇，张新峰，沈兰荪. 基于支撑向量机的中医舌色苔色识别算法研究 [J]. 北京生物医学工程，2006,25 (1): 43-46.

[4] 李博聪，黄庆梅，陈松鹤，等. 基于 CIELAB 空间的中医舌色分析方法[J].世界科学技术——中医药现代化,2007,9(3):28-32.

Evaluation of tongue classification based on different color-space

HONG Wen-xue[1], YANG Qing[1], MENG Hui[1]

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

Abstract: Classification error is the best evaluated performance. Using 360 images and the performance of tongue classification is researched by cross validation method compared 9-3D color-spaces and 18-2D chrominance planes widely used in methods of tongue classification. The system of tongue-color classification performance is built to choose proper color-space to detect character and choose better classify arithmetic. The result shows that the better classify performance of tongue-color is in aesthesia color-space, apperceive color-space and 2D chrominance planes include luminance.

Key words: tongue color; color-space; classification; cross validation; classification error

(上接第 439 页)

A novel visual combining classifier based on 2D graphical representation of the column in multivariate matrix

HONG Wen-xue[1], ZHANG Tao[1], SONG Jia-lin[1], WANG Jin-jia[1], XU Yong-hong[1]

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

Abstract: A novel visual combining classifier (VCC), which integrates two-dimensional graphical representation of the attribute data, image processing and pattern recognition techniques together, is proposed. The basic process of the VCC is as follows: map attribute data of a data matrix to the two-dimensional graphs, transform these graphs to sub classifiers by pixel graphs, and combin the sub classifiers by decision rules. By interactive approaches, the optimum graphs for classification could be chosen and then pattern recognition could be realized automatically. The two experiments of the scatter and pole graphical representations based on Iris database have been made and classification precisions are 98.67% and 97.33% by LOOCV respectively.

Key words: pattern recognition; visual combining classifier; graphical representation; interactive