

原空间中的核 SOM 分类器

潘志松¹, 陈松灿^{1,2}, 张道强¹

(11 南京航空航天大学信息科学与技术学院, 江苏南京 210016;

21 中国科学院自动化研究所模式识别国家重点实验室, 北京 100080)

摘要: 自组织特征映射(SOM)是 Kohonen 提出的一种人工神经网络模型,其整个学习过程是在输入样本空间内进行,并以欧氏距离为度量.这将导致当输入样本分布结构呈高度非线性时,其分类能力下降.核方法通过核函数实现了一个从低维输入空间到高维特征空间的映射,从而使输入空间中复杂的样本结构在特征空间中变得简单. Donald 等人通过核映射将低维输入空间中的非线性问题变换至高维特征空间中,从而使 SOM 聚类形成于映射后的高维特征空间中.但其缺点是失去了对原输入空间聚类中心及结果的直观刻画;本文采用核方法的目的是为原输入空间诱导出一类异于欧氏距离的新的距离度量,并使原 SOM 成为特例.而核的多样性进一步可诱导出原空间中不同的度量,导致各种对应 SOM 分类器的生成.最后,本文侧重通过几种经典的核函数在 Benchmark 上的试验,对该分类器的性能及可靠性进行了验证.

关键词: SOM; 核方法; 分类

中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2004) 02-0222-05

A KernelBased SOM Classification in Input Space

PAN Zhizhong¹, CHEN Songcan^{1,2}, ZHANG Daoqiang¹

(11 Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing, Jiangsu 210016, China;

21 National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Classical SelfOrganizing Maps presented by T. Kohonen is performed in the input sample space based on the Euclidean norm. It fails as the distribution of input patterns becomes highly nonlinear. Kernel means, performing a nonlinear data transformation into some high dimensional feature space, increases the probability of the linear separability of the patterns within the feature space. Donald and others map the data in input space into a high dimension feature space, where SOM algorithm are performed. However, its disadvantage lies in lack of direct descriptions about the clustering center and result. In this paper, a novel kernelbased SOM algorithm is proposed. we replace the Euclidean norm in the SOM training procedure with kernels, which is equivalent to change the metric of distance in input space. Multifomity of kernels leads to different metrics of distance in input space, and correspondingly results in SOM classifications. Finally we discuss the robustness and reliability of KSOM by experimenting on Benchmark based on several classical kernel functions.

Key words: SOM; kernel method; classification

1 引言

Kohonen 提出的自组织特征映射理论 (SelfOrganizing Maps, SOM)^[1] 具有聚类、自组织、自学习以及可视化的功能,已广泛的应用于模式识别、数据挖掘、故障诊断等邻域.该算法在对竞争获胜的神经元及其邻域内的神经元的权值进行调整时,以欧氏距离为度量.对于欧氏距离来讲,等距面是超球体,因而包括 SOM 在内的基于欧氏距离的分类器,其分类面是一个垂直于两类超球体中心连线并通过连线中点的/超平面. SOM 在样本分布类似超球体的数据集上能达到比较好的

分类效果.然而众多实际问题,数据集的分布呈多态形式,具有高度线性不可分,尤其在含有野值点(Outlier)的情况时,传统的 SOM 算法显示出较差的鲁棒性和可靠性.

基于核方法^[2]的学习是近年来发展起来的一种新的机器学习方法,其实质是通过由核诱导的隐映射将低维输入空间中的非线性问题变换至高维(甚至无穷维)特征空间中的较易解决的(可能的)线性问题.并以内积形式刻画.核方法通过核代入最终在特征空间中获得原问题的解决,既避免了计算上的维数灾难,又使问题简化并得到有效解决.因此,核方法已开始广泛地应用于模式分类器设计,如支持向量机

(SVM)^[2], 主分量分析(PCA)^[3]、神经网络设计及聚类分析^[4].

核方法的主要技巧一是先将低维(内积)输入空间非线性地变换至高维(内积)特征空间, 然后在该空间中的内积可以用原空间中的变量直接计算得到, 而无需知道该非线性变换的具体形式. 二是将获胜神经元权值采用对偶表示形式^[4,5], 即获胜神经元权值以训练数据的线性组合表示. 当在高维空间中进行学习时, 获胜神经元的权值调整过程在该空间中进行. 其缺点是无法在原输入空间中获得相应获胜神经元和相应邻域的刻画, 并且聚类结果不易在输入空间中获得直观的解释. 其根本原因是在特征空间中的获胜神经元并不一定在原输入空间中存在相应的原像(Pr& image)^[3]; 在文[5]中, Don2 ald 等人即用此法实现了基于核的 SOM 聚类, 其 SOM 聚类的结果难以在原空间中表示, 即可能无法确定其聚类中心. 为了克服^[5]的不足, 我们提出了原空间的核 SOM 算法, 让学习过程仍在原输入空间中进行, 而核代入仅用来为原输入空间诱导出一类不同于欧氏距离的新的距离度量. 而核的灵活性和多样性又可导出不同的距离度量, 构建出基于不同度量的分类器, 并且分类结果可以在输入空间中获得直观刻画. 进一步, 笔者通过几种核函数在 Benchmark 上的试验, 讨论了该分类模型的可靠性和鲁棒性. 同时发现在选择核函数时, 针对不同的数据集, 各核函数分类效果不同. 说明核函数选择是问题依赖的, 即针对不同的问题, 可以选用不同的核函数, 使模型具有推广能力.

2 基于核函数的 SOM(kernel SOM) 算法

自组织特征的映射(selforganizing feature maps, 简称 SOM) 根据样本的内在联系, 能够对样本自动进行聚类. 设网络输入模式为 $X = [x_1, x_2, \dots, x_m]^T$. M 为输入向量的维数. SOM 权值为 $w_j = [w_{j1}, w_{j2}, \dots, w_{jm}]^T, j = 1, 2, \dots, l, l$ 为总的神经元个数.

SOM 算法^[1]中的核心步骤为:

(1) 选择具有最小距离的节点 j^* 作为获胜节点:

$$d_j^* = \min_j \|X - w_j\|, \quad j = 1, 2, \dots, l \quad (1)$$

(2) 调整输出节点及其几何邻域内的节点所连接的权值矢量:

$$w_j(n+1) = w_j(n) + \alpha(n) h_{j,i(x)}(n) (X - w_j(n)) \quad (2)$$

$\alpha(n)$ 为 n 时刻的学习率, 随着训练的进行逐渐减少.

$h_{j,i(x)}(n)$ 为邻域调整函数.

由公式(2)可见, SOM 学习算法在对竞争获胜的神经元 j^* 及其邻域的调整时, 依赖于 X 到各神经元权值 w_j 之间的欧氏距离 $\|X - w_j\|$. 由前面的分析可知, 当样本在输入空间中分布不规则或呈高度非线性分布时, 依赖于欧氏距离的 SOM 分类器分类效果差, 且对噪声和野值点(Outlier)极为敏感. 为此, 核方法提供了解决上述问题的可能性.

如果训练样本集在输入空间线性不可分, 基于核方法, 可通过由核诱导的非线性映射 ϕ 将输入空间中的非线性问题变换至高维(甚至无穷维)特征空间 F 中的较易解决的(可能的)线性问题.

定义非线性映射 $\phi: X \rightarrow F$, 其中 $X \in R^m, R$ 为样本集, F 为特征空间. 可以用形式化目标函数代替欧氏距离:

$$J(w_j) = \frac{1}{2} \|X - \phi(w_j)\|^2 \quad (3)$$

求其极小值. 其中式(3)中的范数可另写为:

$$\begin{aligned} \|X - \phi(w_j)\|^2 &= (X - \phi(w_j))^T (X - \phi(w_j)) \\ &= X^T X - 2X^T \phi(w_j) + \phi(w_j)^T \phi(w_j) \end{aligned} \quad (4)$$

其中的每一项都可看成是特征空间中的内积, 再依据满足 Mercer 条件的核函数的定义:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (5)$$

将式(5)代入(4), 则有

$$\begin{aligned} J(w_j) &= \frac{1}{2} \|X - \phi(w_j)\|^2 \\ &= K(X, X) + K(w_j, w_j) - 2K(X, w_j) \end{aligned} \quad (6)$$

求函数 $J(w_j)$ 极小值, 可以利用梯度下降法^[11]. 推导得到 w_j 的新的调整公式:

$$\begin{aligned} w_j(n+1) &= w_j(n) - \alpha(n) \nabla J(w_j) \\ &= w_j(n) - \alpha(n) \left(\frac{\partial K(w_j, w_j)}{\partial w_j} - 2 \frac{\partial K(X, w_j)}{\partial w_j} \right) \end{aligned} \quad (7)$$

这样, 我们就按照梯度下降法推导出基于核的 SOM 算法(Kernel SOM, 简称 KSOM). 与其他基于核函数的算法(SVM、KPCA 等)一样, 我们仅在推导过程中概念性的使用非线性变换从低维空间变换到高维空间, 由于核变换后只是改变了内积运算, 不需要知道非线性变化的具体形式, 并没有使算法复杂性随着维数的增加而增加. 而与这些基于核函数的算法不同的是: KSOM 的整个学习过程依然在原空间中进行, 保持了整个权矩阵在原空间中的直观刻画.

根据核映射的灵活性, 不同的核函数又可以诱导出不同的距离度量. 我们采用以下 4 种经典的满足 Mercer 条件的核函数:

(1) 多项式 $K(x, y) = (x^T y)^d, \quad d \in 2$ (8)

(2) 径向基 $K(x, y) = e^{-\|x - y\|^2 / 2R^2}$ (9)

(3) 柯西 $K(x, y) = \frac{1}{1 + \|x - y\|^2 / R^2}$ (10)

(4) 对数形式 $K(x, y) = \log(1 + \|x - y\|^2 / R^2)$ (11)

分别将式(8), (9), (10), (11)代入式(7), 有

$$w_j(n+1) = w_j(n) - \alpha(n) (2d(w_j^T)^{d-1} w_j - (X^T w_j)^{d-1} X) \quad (12)$$

$$w_j(n+1) = w_j(n) - \alpha(n) \left[\frac{2}{R^2} e^{-\|X - w_j\|^2 / R^2} (X - w_j) \right] \quad (13)$$

$$w_j(n+1) = w_j(n) + \alpha(n) \left[\frac{1}{R^2 \#(1 + \|X - w_j\|^2 / R^2)^2} (X - w_j) \right] \quad (14)$$

$$w_j(n+1) = w_j(n) + \alpha(n) \left[\frac{1}{R^2 \#(1 + \|X - w_j\|^2 / R^2)} (X - w_j) \right] \quad (15)$$

以上式(12)、(13)、(14)、(15)就是基于以上四种核函数的 KSOM 的权值调整公式. 和 SOM 不同之处在于:

由公式(4), 在新的距离度量下, 重新定义获胜神经元 j^* :

$$d_j^* = \min_j \|X - \phi(w_j)\|^2$$

$$= \min(K(X, X) + K(w_j, w_j) - 2K(X, w_j)), j = 1, 2, \dots, l \quad (16)$$

调整获胜神经元及其邻域内的节点所连接的权值矢量. 以径向基核函数为例, 将 SOM 算法中式(2)的权值调整部分修正如下:

$$w_j(n+1) = w_j(n) + G(n) \# h_{j,i(x)}(n) \# 2/R^2 \# e^{-\|x - w_j\|^2/R^2} (X - w_j(n)) \quad (17)$$

令 $G(n) = G(n) \# 2/R^2$, 则

$$w_j(n+1) = w_j(n) + G(n) \# h_{j,i(x)}(n) \# e^{-\|x - w_j\|^2/R^2} (X - w_j(n)) \quad (18)$$

算法其余部分不变. 当 $R \rightarrow \infty$ 时, KSOM 退化为 SOM, 即 SOM 可作为 KSOM 的特例. 由式(18)和式(2)比较可见, 在 KSOM 算法中, 虽然我们使用了核函数代替了样本空间中的内积运算, 但从本质上说, 其聚类的过程仍然在原样本空间中进行, 不同的是, 其度量样本和权向量之间的尺度不再使用欧氏距离, 而是使用核代入方法, 这是一种新的核方法的应用. 文献[5]在使用核方法时, 获胜神经元的在高维特征空间中形成, 故无法在原输入空间中获得相应神经元和相应邻域的刻画, 其根本原因是在特征空间中的获胜神经元并不一定在原输入空间中存在相应的原像; 此外, 聚类结果不易在输入空间中获得直观的解释. KSOM 克服了文[5]的不足, 其权矩阵的调整过程仍在原输入空间中进行, 整个学习过程在输入空间中获得了直观的刻画. 下面针对几个 Benchmark 数据集对不同的距离度量的 KSOM 进行了测试, 并就不同核函数下分类器的鲁棒性和可靠性进行了分析.

3 实验结果及分析

3.1.1 Iris 数据集

实验采用 Iris 数据集^[8], Iris 数据集包括 3 个类, 每个类由 50 个 4 维模式组成; 并且第 2、3 两类是线性不可分的. 我们使用 75 组数据训练, 75 组数据用于测试.

在实验中, 我们分别对四种核函数进行了 10 次独立的无监督学习的分类实验. 其实验结果如表 1. 表中列出了各个核函数每次实验的错分模式数及平均错分数和平均错分率. 同时, 为了保证实验的可行性、有效性和可比性对模式的一些参数作了如下的规定:

表 1 SOM, 及四种核函数的 KSOM 的 Iris 分类实验结果

模型	次数										平均错分数	平均错分率
	1	2	3	4	5	6	7	8	9	10		
SOM	4	5	7	10	2	8	6	2	11	3	5.8	7.733%
多项式核	5	5	12	3	5	1	3	3	4	2	4.4	5.866%
径向基核	2	6	5	3	9	1	7	1	11	1	4.6	6.133%
柯西核	7	6	4	4	1	6	0	7	3	4	4.2	5.200%
对数核	3	6	7	3	2	4	1	4	3	2	3.5	4.666%

(1) 二维神经元矩阵为 9@9; (2) 排序阶段: 学习的最大次数为 1000; 学习率的初始值为 0.19; 学习率的最终值为 0.0105; 邻域的初始值为 8; 邻域的最终值为 1; (3) 收敛阶段: 学习的最大次数为 2000; 学习率的初始值为 0.0105; 学习率的最终值为 0.0101; 邻域的初始值为 0; 邻域的最终值为 0; (4) $d = 3, R = 2$.

我们发现对数核函数对 Iris 效果较好, 错分率下降了 31.067%, 但由于参数选择的多样性, 其他的核函数形式对不同的问题通过参数的调整, 可能有更好的效果.

由于 KSOM 的聚类的过程仍是在原样本空间中进行的, 所以分类的结果在输入空间中可获得直观的刻画. 通过对输出结果可视化^[7]后发现, KSOM 输出的二维神经元的映射图和 U 矩阵^[7,9]和传统的 SOM 模型有较大的不同. 见图 1-4.

U 矩阵表示了 SOM 映射图上相邻结点之间的距离, 因此可以从 U 矩阵看出映射图的聚类结构. 由图 1 可以看出, 在类别映射图上被明显分为两类, 两类间的分界面在 U 矩阵中对应的邻结点距离比较大, 即颜色比较深. 而二三类的分类边界很模糊, 直接导致两类间的高错分率.

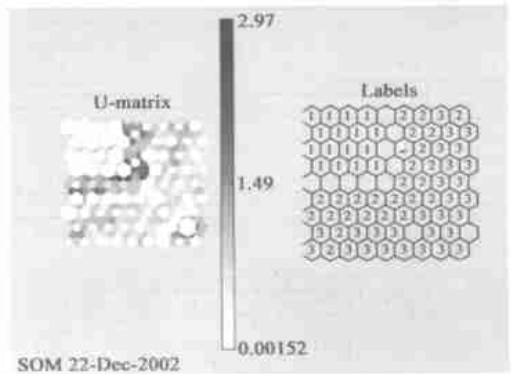


图 1 SOM 的 U 矩阵和带标记的 SOM 映射图

图 2(a) 为该 SOM 的 9@9 的神经元的网格; 图 2(b) 为三维的 U 矩阵图; 图 2(c) 为 SOM 的原型图; 图 2(d) 为原型和数据图. 从图 2(c) 的 SOM 原型图上也可看出, 二三类的神经元聚类的效果较差.

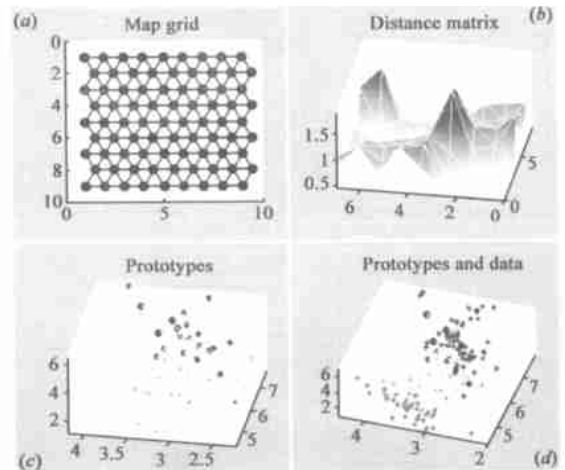


图 2 传统 SOM 的 U 矩阵和输出原型

而基于核函数的 KSOM 对可靠性有较高的特性. 我们以径向核 SOM 为例, 从图 3 中可以看出, 在 KSOM 的 U 矩阵中, 一二类以及二三类之间的边界都较明显, 从图 4(b) 的 U 矩阵三维图中, 可以更直观地看出. 从图 4(c) 的 KSOM 原型(Proto types)^[18]图中, 第二类和第三类的神经元自动聚类后明显被分为两类.

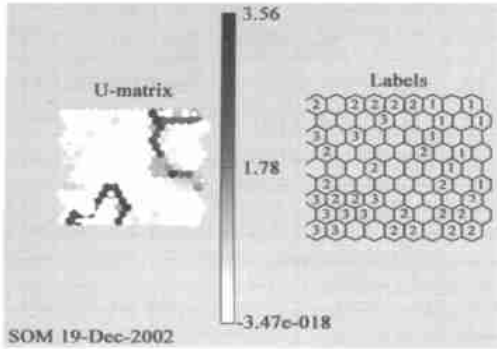


图 3 基于径向核 SOM 的 U 矩阵和带标记的 SOM 映射图

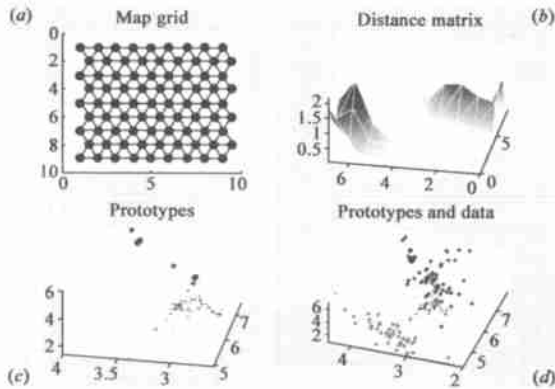


图 4 基于径向核 SOM 的 U 矩阵和输出原型

衡量分类器的好坏由两个因素, 一个是错分率, 另一个就是可靠性. 在 KSOM 算法中, 定义可靠性度量如下:

$$D = \frac{1}{M} \sum_{i=1}^M \frac{d_{i^*} - d_{i^*c}}{d_{i^*}} \quad (19)$$

其中 M 为输入样本的个数, 而 d_{i^*} 定义见公式(16), d_{i^*c} 为 X 到次匹配单元的距离:

$$d_{i^*c} = \min(K(X, X) + K(w_j, w_j) - 2K(X, w_j)), \\ j = 1, 2, \dots, j^* - 1, j^* + 1, \dots, l \quad (20)$$

应用公式(19)计算了各个核 SOM 分类器的可靠性. 传统的 SOM 为 $613443 @ 10^{-5}$, 径向基核为 $112484 @ 10^{-4}$, 柯西核为 $111198 @ 10^{-4}$, 对数核为 $11929 @ 10^{-4}$. 我们发现三种核函数都能明显的提高对 IRIS 数据集的可靠性. 而由图 1 和图 3, 比较两者带标记映射图可以更直观地看出, 图 1 中 SOM 映射图的分布密集, 而图 2 的 KSOM 映射图稀疏, 将导致误识率下降以及可靠性提高. 在图 4(c) 的 KSOM 原型图上, 三个类别的神经元明显地被从上至下地聚类为三层, 这些都从直观上验证了 KSOM 分类器的可靠性.

3.1.2 Wine Recognition 数据集

Wine Recognition 数据集^[8]是三种意大利葡萄酒的化学分

析结果. 该数据集含有 178 个样本, 每个样本具有 13 维特性. KSOM 初始设置同试验一. 我们采用 90 个样本训练, 88 个样本测试, 试验结果如表 2.

表 2 四种核函数的 KSOM 的 Wine recognition 分类实验结果

模型 \ 次数	次数										平均错分率	平均错分率
	1	2	3	4	5	6	7	8	9	10		
SOM	5	9	4	9	9	8	10	12	10	9	8.5	9.659%
径向基核	5	6	6	5	6	5	7	9	8	6	6.3	7.159%
柯西核	6	6	5	8	3	6	5	7	5	8	5.9	6.704%
对数核	36	41	27	37	35	32	42	41	34	40	33.03	37.54%

从表 2 可以看出, 基于两种核函数 KSOM 对 Wine recognition 数据的分类效果比 SOM 好, 错分率分别下降了 2.5%、2.955%, 达到了较好的效果. 而对数核和多项式核(结果未列出)的 KSOM 对于 Wine recognition 无法正确识别, 误分率达到 37% 和 50%, 当二维神经元矩阵为 $5 @ 5$, $R=6$ 时, 对数核的 KSOM 错分率为 17.6%, 这也说明参数的调整也是影响分类结果的重要因素. 从上面的分析可以看出, KSOM 算法是问题依赖的, 并不是对所有问题都适用, 但由于核方法的灵活性, 针对不同的问题, 可以使用不同的核函数, 或通过经验调整参数, 从而使 KSOM 模型具有更好的推广能力.

3.1.3 对野值点的处理

为了比较 KSOM 对野值点的处理能力, 我们在如下人工数据集上做了模拟实验. 数据集 D 是以 $(0, 0)$ 和 $(3, 0)$ 为中心呈高斯分布两类二维随机数, 并且包括三个野值点 (Outlier) $(200, 200)$, $(200, 190)$, $(200, -100)$ ^[10].

SOM 算法对数据集 D 的聚类结果如图 5 所示, 而 KSOM 算法对数据集 D 的聚类结果见图 6 所示, KSOM 中使用的核函数均为径向基核, 参数 $R=2$. 从两图的比较可看出, SOM 受到野值点的干扰非常严重, 把本属于两类的样本看成是一类, 野值点则作为另一类. 而基于核的聚类算法则完全避免了野值点的影响, 从 U 矩阵中可看出, 两类间隔颜色较深, KSOM 将两类样本正确分开. 错分率为 0. 说明 KSOM 算法的抗野值点性能相比传统的 SOM 算法有明显的提高, 具有鲁棒性^[10].

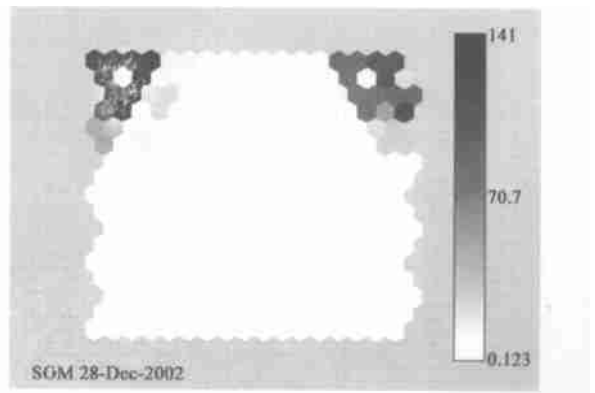


图 5 SOM 的聚类结果

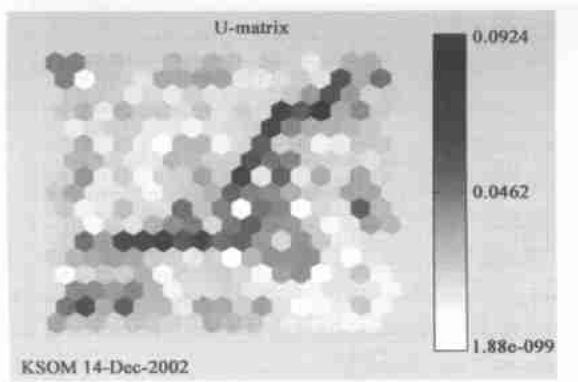


图6 KSOM的聚类结果

4 结论

本文使用核技巧设计了一种新的基于核的KSOM分类器。KSOM算法的实质是核代入为原输入空间诱导出了一类不同于欧氏距离的新的距离度量,和一般的基于核的算法不同的是,其聚类过程是在原输入空间中形成,而核的灵活性可导出不同的距离度量,从而在原空间中实现不同度量的聚类,这是一种全新的应用核的观点。进一步,文章讨论了该度量的高可靠性和鲁棒性。而诱导新距离度量的新观点可应用于其它类似问题,如K均值的聚类算法的设计,矢量量化以及对于高度线性不可分的异常检测问题等。最后上述算法在Benchmark实验中获得了令人满意的结果。另外,从文中可以看出,对于不同的数据集,而可以采用不同的核函数,以求达到较好的效果。

参考文献:

- [1] Simon Haykin. Neural Networks: A Comprehensive Foundation[M]. 北京: 清华大学出版社, 2001. 443- 483.
- [2] Scholkopf B, Burges C J C, Smola A J. Advances in Kernel Methods: Support Vector Learning[M]. Cambridge, MA, The MIT Press, 1999.

- [3] Scholkopf B, Smola A J, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998, 10(5): 1299 - 1319.
- [4] 陈松灿, 张道强. 输入空间中的核聚类算法[R]. 南京: 南京航空航天大学四院, 2002.
- [5] Donald MacDonald, Colin Fyfe. The Kernel Self Organising Map[A/OL]. <http://cis.paisley.ac.uk/fyfe2ci0/kernel/kmeans.ps>, 2000.
- [6] Andreas Namberger, Marcin Detyniecki. Visualizing changes in data collections using growing selforganizing maps[A]. International Joint Conference on Neural Networks (IJCNN 2002)[C]. Honolulu, Hawaii, USA, 2002. 1912-1917.
- [7] SOMTOOLBOX[CP/OL]. <http://www.cis.hut.fi/projects/somtoolbox/>.
- [8] UCI repository of machine learning databases[DB/OL]. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
- [9] Kraaijveld M A, Mao J, Jain A K. A nonlinear projection method based on kohonen's topology preserving maps[J]. IEEE Transactions on Neural Networks, 1995, 6(3): 548- 559.
- [10] Kadir L. Robust error measure for supervised neural network learning with outliers[J]. IEEE Trans. Neural Network, 1996, 7(1): 246- 250.
- [11] Peter Andras. Kernel kohonen networks[J]. International Journal of Neural Systems, 2002, 12(2): 117- 135.

作者简介:

潘志松 男, 1973年生于江苏南京, 南京航空航天大学信息科学与技术学院博士研究生, 主要研究方向信息系统安全、入侵检测系统、神经网络。Emails: ldsystem@sina.com; hotpzs@hotmail.com; s.chen@nuaa.edu.cn

陈松灿 男, 1962年生于浙江余姚, 南京航空航天大学信息科学与技术学院教授, 博士研究生导师, 主要研究方向神经网络、模式识别。

张道强 男, 1978年生于山东聊城, 博士研究生, 主要研究方向神经网络、模式识别。