# 基于核方法的聚类算法及其应用

纪秋颖 林 健

(五邑大学 管理学院, 江门 529020) (北京航空航天大学 经济管理学院, 北京 100083)

**摘 要**: 在分析核方法的核心概念基础上,提出了一种基于核方法的聚类算法. 通常,传统聚类算法只在数据特征差异较大时才有效,当数据特征差异较小时,很难取得较好的聚类效果. 引入核函数,将原始数据由数据空间映射到特征空间,在特征空间中进行聚类. 核函数的非线性映射使得原始数据的特征更完整地显现出来,从而能够更客观准确地聚类. 与传统聚类方法相比,该方法聚类结果更客观有效. 以 16 组实际数据为例,将该方法应用于数据分类研究中,聚类结果表明了该方法的可行性和有效性,从而为数据分类提供了一种新的可行方法.

**关 键 词**: 聚类算法; 核方法; 特征空间; 核函数; 分类

**中图分类号**: TH 166

**文献标识码**: A **文 章 编 号**: 1001-5965(2006)06-0747-04

## Clustering algorithm based on kernel methods and its application

Ji Qiuying

(School of Management, Wuyi University, Jiangmen 529020, China)

Lin Jian

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

**Abstract**: Based on the analysis of the core concepts of the kernel methods, a clustering algorithm based on kernel methods was put forward. In general, traditional clustering algorithms are suitable to implement clustering only if the feature differences of data are large. If the feature differences are small and even cross in the original space, it is difficult for traditional algorithms to cluster correctly. By using kernel functions, the data in the original space was mapped into a high-dimensional feature space, in which more features of the data were exposed so that clustering could be performed efficiently. Compared with the traditional clustering methods, this clustering method had superiorities in dealing with the nonlinear data, which made its clustering result more objective and valid. This method was applied to the classification of 16 groups of data, and results show the feasibility and effectiveness of the kernel clustering algorithm.

**Key words**: clustering algorithm; kernel methods; feature space; kernel function; classification

核方法(kernel method)[1-3]是一系列非线性数据处理方法的总称,其共同特征就是这些处理方法都应用了核函数. 核方法的基本操作过程是首先采用非线性映射将原始数据由数据空间映射到特征空间,进而在特征空间中进行对应的线性操作. 核方法通过运用非线性映射,大大增强了非线性处理能力,已经成为当前机器学习研究领域的热点课题. 目前国内与核方法相关的研究工作主要集中在支持向量机(SVM,Support Vector Machines)和支持向量回归(SVR,Support Vector Regression)2 个方面,应用领域包括模式识别[4]、预测[5]、综合评价[6]等,但将核方法应用于聚类研究的文献不多. 本文开发了一种基于核方法的聚类算法,并将其应用于实际问题,效果良好.

**收稿日期**: 2006-03-22

**基金项目**: 国家自然科学基金资助项目(70471074);广东省科技攻关资助项目(2004B36001051)

**作者简介**: 纪秋颖(1962-), 女, 吉林吉林人, 副教授, 现为北航在职博士生,jiqiuying@126.com.

## 1 核方法的核心概念

### 1.1 最优分类面

对于2类线性可分问题,存在着无穷多个分割线(多维情况下对应超平面),可以准确无误地将2类样本分开,其中必然存在一条不仅误差最小、而且离2类样本的间隙最大的分割线(平面),这条分割线(平面)称为最优分类线(平面).

对于一个给定的线性可分样本集$(\boldsymbol{x}_i, y_i)$,$i=1,2,\cdots,n$,$\boldsymbol{x}_i \in \mathbf{R}^d$,$y_i \in \{-1,1\}$是类别标号,$d$维空间中线性判别的一般形式为$g(\boldsymbol{X})=\boldsymbol{W}\cdot\boldsymbol{X}+b$,分类面方程为

$$\boldsymbol{W}\cdot\boldsymbol{X}+b=0 \tag{1}$$

满足$|g(\boldsymbol{X})|=1$的样本点,离最优分类线(平面)距离最近,对于线性可分问题,这些样本点完全决定了最优分类线(平面),被称为支持向量(SV, Support Vectors)[4]. 因此,最优分类面就由如下的优化问题决定[1]:

$$\left.\begin{aligned}&\min\ \phi(\boldsymbol{W})=\frac{1}{2}\|\boldsymbol{W}\|^2=\frac{1}{2}(\boldsymbol{W}\cdot\boldsymbol{W})\\&\text{s.t.}\ \ y_i[(\boldsymbol{W}\cdot\boldsymbol{x}_i)+b]-1\geqslant 0\\&\qquad i=1,2,\cdots,n\end{aligned}\right\} \tag{2}$$

引入拉格朗日乘子$\alpha_i\geqslant 0$,$i=1,2,\cdots,n$,建立拉格朗日函数:

$$L(\boldsymbol{W},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{W}\|^2-\sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{W}\cdot\boldsymbol{x}_i+b)-1] \tag{3}$$

对$\boldsymbol{W}$,$b$取极小,同时对$\alpha_i$取极大,有

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{n}\alpha_i y_i=0 \tag{4}$$

$$\frac{\partial L}{\partial \boldsymbol{W}}=0\Rightarrow\boldsymbol{W}=\sum_{i=1}^{n}\alpha_i\boldsymbol{x}_i y_i \tag{5}$$

由式(3)~式(5)得式(2)的对偶优化问题:

$$\left.\begin{aligned}&\min\ Q(\boldsymbol{\alpha})=\frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{n}\alpha_i\\&\text{s.t.}\ \ \alpha_i\geqslant 0\\&\qquad\sum_{i=1}^{n}\alpha_i y_j=0\end{aligned}\right\} \tag{6}$$

式中$\boldsymbol{x}_i$,$\boldsymbol{x}_j$为样本输入变量.

根据KKT条件,式(6)的解满足:

$$\alpha_i[y_i(\boldsymbol{W}\cdot\boldsymbol{x}_i+b)-1]=0 \tag{7}$$

式中$i=1,2,\cdots,n$. 仅有部分$\alpha_i$满足$\alpha_i>0$,它们对应前面所提及的支持向量. 在$\alpha_i$基础上,通过式(5)即可得到$\boldsymbol{W}$,$b$通过式(8)获得:

$$b=-\frac{1}{2}\boldsymbol{W}\cdot(\boldsymbol{x}_r+\boldsymbol{x}_s) \tag{8}$$

式中$\boldsymbol{x}_r$,$\boldsymbol{x}_s$分别为2类样本中任意一个支持向量.

在线性可分情况下,最优分类线(平面)为

$$f(\boldsymbol{X})=\operatorname{sgn}\Big(\sum_{i=1}^{n}\alpha_i y_i(\boldsymbol{x}_i,\boldsymbol{X})+b\Big) \tag{9}$$

### 1.2 支持向量机

对于线性不可分问题,无法获得满足式(6)的最优分类线(平面). 此时可以通过将样本$\boldsymbol{X}$映射到高维特征空间中,使其线性可分.

引入核函数

$$\boldsymbol{K}(\cdot,\cdot)=(\phi(\cdot),\phi(\cdot)) \tag{10}$$

即可实现特征空间与数据空间之间的数据转换.

这样,式(6)变为

$$\left.\begin{aligned}&\min\ Q(\boldsymbol{\alpha})=\frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{n}\alpha_i\\&\text{s.t.}\ \ \alpha_i\geqslant 0\\&\qquad\sum_{i=1}^{n}\alpha_i y_j=0\end{aligned}\right\} \tag{11}$$

最优分类面为

$$f(\boldsymbol{X})=\operatorname{sgn}\Big(\sum_{i=1}^{n}\alpha_i y_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{X})+b\Big) \tag{12}$$

式中$b$由式(13)给出[3]:

$$b=\frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i[\boldsymbol{K}(\boldsymbol{x}_r,\boldsymbol{x}_i)+\boldsymbol{K}(\boldsymbol{x}_s,\boldsymbol{x}_i)] \tag{13}$$

式(12)被定义为支持向量机.

### 1.3 核函数

不难发现,核函数是实现从线性可分的最优分类面到非线性可分的支持向量机的关键. 不同的核函数形式决定不同的支持向量机表示形式. 依据统计学习理论,满足Mercer条件的函数就可视为一种核函数.

由于核函数的引入,在使用核方法时,并不需要了解具体的非线性映射形式,而只要了解映射后向量的内积核函数的形式. 这正是核方法的优势,可以设计出非常复杂的非线性映射形式,而实际上用到的核函数则比较简单. 目前,常用的核函数主要有:线性核函数$(\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}_k)=\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x})$、$d$阶多项式核函数$(\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}_k)=(\boldsymbol{x}_k^{\mathrm{T}}+1)^d)$、高斯径向基核函数$(\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}_k)=\exp\{-\|\boldsymbol{x}-\boldsymbol{x}_k\|^2/\sigma^2\})$.

## 2 基于核方法的聚类

基于核方法的聚类实际上是一类分类方法的应用[4,7]. 一类分类方法就是用一个具有最小体

积的超球体将现有样本包含起来. 对于样本集:

$$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\} \tag{14}$$

设将样本集中全体样本完全包围的最小球体的半径为 $R$,球心为 $\boldsymbol{a}$,则满足优化方程:

$$\left.\begin{aligned}&\min\ L(R)=R^2\\&\text{s.t.}\quad R^2-(\boldsymbol{x}_i-\boldsymbol{a})(\boldsymbol{x}_i-\boldsymbol{a})^{\mathrm{T}}\geqslant 0\end{aligned}\right\} \tag{15}$$

由式(15)可定义 Lagrange 函数:

$$L(R,a,\boldsymbol{\alpha})=R^2-\sum_{i=1}^{N}\alpha_i\cdot[R^2-(\boldsymbol{x}_i^2-2\boldsymbol{a}\boldsymbol{x}_i-\boldsymbol{a}^2)] \tag{16}$$

式中 $\alpha_i\geqslant 0$ 为 Lagrange 系数. 将式(16)对 $R$ 和 $\boldsymbol{a}$ 求偏微分,并令它们等于 0,即

$$\frac{\partial L}{\partial R}=2R-2R\sum_{i=1}^{N}\alpha_i=0 \tag{17}$$

$$\frac{\partial L}{\partial \boldsymbol{a}}=\sum_{i=1}^{N}\alpha_i(2\boldsymbol{x}_i-2\boldsymbol{a})=0 \tag{18}$$

由式(17)和式(18)有

$$\sum_{i=1}^{N}\alpha_i=1 \tag{19}$$

$$\boldsymbol{a}=\sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i \tag{20}$$

将式(19)和式(20)代入式(16),有

$$\max\ L=\sum_{i=1}^{N}\alpha_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \tag{21}$$

约束条件为式(19)及

$$\alpha_i\geqslant 0 \tag{22}$$

根据 KKT 条件,只有很少部分 $\alpha_i>0$,其余大部分 $\alpha_i=0$,与这些 $\alpha_i$ 对应的样本点决定了边界的构成.

由式(19)、式(21)和式(22)获得 $\alpha_i$ 后,利用式(20)即可求出球心 $\boldsymbol{a}$,任选一支持对象由式(23)可求出 $R$:

$$R^2-(\boldsymbol{x}_i^2-2\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}^2)=0 \tag{23}$$

引入核函数$(\boldsymbol{x}\cdot\boldsymbol{y})\rightarrow\boldsymbol{K}(\boldsymbol{x},\boldsymbol{y})$,在特征空间中的优化问题为

$$\max\ L=\sum_{i=1}^{N}\alpha_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{24}$$

约束条件不变,即式(19)和式(22).

引入松弛变量,样本集满足:

$$\left.\begin{aligned}&\min\ L(R)=R^2+C\sum_{i=1}^{N}\xi_i\\&\text{s.t.}\quad(\boldsymbol{x}_i-\boldsymbol{a})^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{a})\leqslant R^2+\xi_i\\&\qquad\xi_i\geqslant 0\\&\qquad i=1,2,\cdots,N\end{aligned}\right\} \tag{25}$$

定义 Lagrange 函数:

$$L(R,a,\boldsymbol{\alpha},\xi)=R^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\alpha_i[R^2+\xi_i-(\boldsymbol{x}_i^2-2\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}^2)]-\sum_{i=1}^{N}\gamma_i\xi_i \tag{26}$$

式中 Lagrange 系数 $\alpha_i\geqslant 0,\gamma_i\geqslant 0,i=1,2,\cdots,N$.

对式(26)分别关于 $R$ 和 $\boldsymbol{a}$ 求偏微分,并令它们为 0,得

$$\left.\begin{aligned}&\max\ L=\sum_{i=1}^{N}\alpha_i(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s.t.}\quad 0\leqslant\alpha_i\leqslant C\\&\qquad\sum_{i=1}^{N}\alpha_i=1\end{aligned}\right\} \tag{27}$$

引入核函数,得

$$\left.\begin{aligned}&\max\ L=\sum_{i=1}^{N}\alpha_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s.t.}\quad 0\leqslant\alpha_i\leqslant C\\&\qquad\sum_{i=1}^{N}\alpha_i=1\end{aligned}\right\} \tag{28}$$

通过调整高斯径向基核函数中参数 $\sigma$ 值的大小,来比较所取得的聚类效果,当 $\sigma$ 取一定数值时,正常域分成很多个互不相通的小区域,这就是基于核方法的聚类原理. 其中,聚类的边界可任选一支持对象 $\boldsymbol{z}$ 到球心 $\boldsymbol{a}$ 的距离来表示:

$$f(\boldsymbol{x})=f(\boldsymbol{z})=\boldsymbol{K}(\boldsymbol{z},\boldsymbol{z})-2\sum_{i=1}^{N}\alpha_i\boldsymbol{K}(\boldsymbol{z},\boldsymbol{x}_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

## 3　实证研究

以 16 组实际数据来验证本文提出的基于核方法的聚类算法的可行性和有效性.

对原始数据进行规范化处理:$x'_{ij}=\dfrac{x_{ij}}{x_{\max}}$,$i=1,2,\cdots,16$;$j=1,2$. 式中 $x_{\max}=\max(x_{ij})$,结果如表 1 所示.

**表 1　16 组基本数据**

| $i$ | $x_{i1}$ | $x_{i2}$ | $i$ | $x_{i1}$ | $x_{i2}$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 9 | 0.44 | 0.66 |
| 2 | 0.82 | 0.75 | 10 | 0.86 | 0.42 |
| 3 | 0.45 | 0.42 | 11 | 0.94 | 0.95 |
| 4 | 0.37 | 0.27 | 12 | 0.83 | 0.79 |
| 5 | 0.23 | 0.33 | 13 | 0.73 | 0.47 |
| 6 | 0.32 | 0.59 | 14 | 0.46 | 0.82 |
| 7 | 0.72 | 0.86 | 15 | 0.76 | 0.65 |
| 8 | 0.72 | 0.34 | 16 | 0.21 | 0.71 |

采用高斯径向基核函数,调节参数 $\sigma$ 直至 $\sigma=0.15$,可以得到聚类的结果为如下 5 类:

$$c_1 = \{1,11\}$$
$$c_2 = \{6,9,14,16\}$$
$$c_3 = \{2,7,12,15\}$$
$$c_4 = \{3,4,5\}$$
$$c_5 = \{8,10,13\}$$

聚类的效果如图 1 所示. 由图可见,基于核方法的聚类算法可以获得较为理想的聚类边界.

根据各组数据的评价得分等进行综合分析,该聚类结果是符合实际情况的. 因此,基于核方法的聚类算法是可行和有效的.

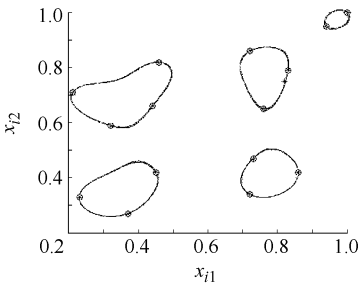



图 1 16 组数据聚类效果示意图

## 4 结 束 语

实证结果表明,采用基于核方法的聚类算法对实际问题分类研究,可以得到比较满意的结果. 该方法可以在一定程度上消除以往数据分类过程中的随机性和分类人员主观上的不确定性及其认识上的模糊性,保证了分类结果的客观性.

## 参考文献 (**References**)

[1] Muller K R, Mika S, Ratsch G, et al. An introduction to kernel-based learning algorithms [J]. IEEE Trans on Neural Networks, 2001,12(2): 181-201

[2] Mika S, Ratsch G, Weston J, et al. Fisher discriminant analysis with kernels[C]//Neural Networks Signal Process Proc IEEE. Piscataway, NJ: IEEE, 1999:41-48

[3] Klinke S, Cook D. Binning of kernel-based projection pursuit indices in XGobi[J]. Computational Statistics & Data Analysis, 1997,25(3): 363-369

[4] 肖健华. 基于支持对象的野点检测方法[J]. 计算机工程, 2003,29(11):43-45
Xiao Jianhua. Approach of outlier detection based on support objects [J]. Computer Engineering,2003,29(11):43-45(in Chinese)

[5] 魏宏业,王建华,何葳. 销售量预测的支持向量机建模及参数选择研究[J]. 系统仿真学报,2005,17(1):33-36
Wei Hongye, Wang Jianhua, He Wei. Study on support vector machines model for sales volume prediction and parameters selection[J]. Journal of System Simulation, 2005,17(1): 33-36(in Chinese)

[6] 肖健华,吴今培,杨叔子. 基于 SVM 的综合评价方法研究[J]. 计算机工程,2002,28(8):28-30
Xiao Jianhua, Wu Jinpei, Yang Shuzi. Approach of evaluateon system based on support vector machine[J]. Computer Engineering, 2002, 28(8): 28-30(in Chinese)

[7] 李焕荣,林健. 基于一类分类方法的多类分类及其应用[J]. 华南理工大学学报(自然科学版),2004,32(8):82-88
Li Huanrong, Lin Jian. Multiclass classification based on the one-class classification and its application [J]. Journal of South China University of Technology (Natural Science), 2004, 32(8):82-88(in Chinese)

(责任编辑:文丽芳)