核聚类算法及其在模式识别中的应用*

吕 佳
(重庆师范大学 数学与计算机科学学院, 重庆 400047)

摘 要:将核学习方法的思想和可能性聚类算法相结合,提出一种基于核的可能性聚类算法,使其能够对非超球体、含有噪音和孤立点的数据进行有效的聚类。将该方法用于模式识别中,仿真实验表明,基于核的可能性聚类算法比模糊 C-均值算法以及可能性聚类算法具有更好的聚类效果,且算法能够很快地收敛。

关键词:模式识别;核学习方法;模糊 C-均值算法;可能性聚类算法;
中图分类号:TP181 **文献标识码**:A **文章编号**:1672-6693(2006)01-0022-03

Clustering Algorithm of Kernel and Its Application in Pattern Recognition

LV Jia
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: Combined kernel learning method with possible C-Means algorithm, a kernel-based possible C-Means algorithm is presented in this paper. Non-hyperspherical samples can effectively be clustered, so data of noise and outliers can be. Applied to pattern recognition, it is indicated that the algorithm has more clustering effect than fuzzy C-Means algorithm and possible C-Means algorithm and it can converge fast.

Key words: pattern recognition; kernel learning method; fuzzy C-Means algorithm; possibilistic C-Means algorithm

聚类分析属于无监督模式识别问题,它无需任何先验知识,只按照某种相似程度的度量,把相似的样本归为一类,不相似的样本归于不同的类。可能性聚类算法[1](Possibilistic C-Means, PCM)是对模糊 C-均值(Fuzzy C-Means, FCM)[2]的一种改进算法。FCM 因算法简单、收敛速度快,具有比较直观的几何意义,且能处理模糊信息而在许多领域特别是模式识别、图像处理、特征提取中得到了广泛的应用。但 FCM 规定了每个样本对各个类的隶属度的和为 1,即假定每个样本对聚类的影响力是相同的,这样就导致当样本中出现噪音和孤立点时,它们被赋予了较大隶属度而被错误地划分到某一类中的现象发生,而 PCM 克服了 FCM 算法的缺点,解决了样本中包含噪音和孤立点时,聚类效果不好的问题。

在使用 FCM 和 PCM 进行聚类分析时,首先假定类分布是超球体或超椭圆体的,当各类样本的边界是线性不可分的或类分布不是超球体或超椭圆体时,聚类常会出现失效或错分的情况。这些方法没有对样本的特征进行优化,而是直接用样本的特征进行聚类,这样这些方法的有效性在很大程度上取决于样本的分布情况。故可引入基于核的学习方法,增加对样本特征的优化。通过利用核函数将在观察空间线性不可分的样本非线性映射到高维的特征空间而变得线性可分,这样样本特征经很好地分辨、提取并放大后,可以实现更为准确的聚类。文献[3]指出只要非线性映射是连续和光滑的,观察空间中样本的拓扑结构将会在高维特征空间中得到保持,并且基于核的聚类算法在类分布不为超球体或超椭圆体时依然有效。

本文利用核学习方法的思想[4],结合可能性聚类算法提出了基于核的可能性聚类算法(Kernel-based Possibilistic C-Means, KPCM),仿真实验表明,该算法能快速有效地进行聚类,并且能克服噪音和孤立点的影响。

* 收稿日期:2005-05-26 修回日期:2005-09-02
资助项目:重庆市教委科学技术研究项目(No. KJ050802);重庆师范大学科研资助项目(No. 05XLY003)
作者简介:吕佳(1978-),女,四川达州人,讲师,硕士研究生,研究方向为数据挖掘、计算机网络。

1 基于核的可能性聚类算法

1.1 可能性聚类算法

FCM 算法要求每个样本对各个类的隶属度有归一化的约束条件,这使得样本的隶属度不但与该类的中心有关而且受其它类中心位置的影响,不能真正表征样本属于该类的程度,且该条件对每个样本的影响力是相同的。当样本中存在噪音时,噪音点被赋予较大的隶属度,从而加大了聚类的误差。

可能性聚类算法放松了样本隶属度的约束,样本隶属度 u_{ij} 只要满足 $\max\limits_{i} u_{ij} > 0$ 即可,而将目标函数[1]改为

$$J_m(U,V) = \sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij})^m d_{ij}^2 + \sum_{i=1}^{C}\eta_i\sum_{j=1}^{N}(1-u_{ij})^m \quad (1)$$

其中,N 为样本个数;C 为聚类中心数 $2 \leqslant C \leqslant N$;$u_{ij} = u_i(x_j)$ 表示第 j 个样本属于第 i 类的隶属度;m 为权重指数,表征模糊化程度;d_{ij} 为第 j 个样本到第 i 个聚类中心的距离。(1)式中,等号右边的第1项即为 FCM 的目标函数,要求各个样本点到每个类的聚类中心的距离之和尽可能的小,第2项为惩罚函数,要求 u_{ij} 尽可能大,从而避免无效解。η_i 为一合适的正数,推荐取值为

$$\eta_i = K\frac{\sum_{j=1}^{N}u_{ij}^m d_{ij}^2}{\sum_{j=1}^{N}u_{ij}^m} \quad (2)$$

式中 K 为正整数,通常取 $K=1$。

隶属度 u_{ij} 公式为

$$u_{ij} = \frac{1}{1+\left(\frac{d_{ij}^2}{\eta_i}\right)^{\frac{1}{m-1}}} \quad (3)$$

从上式可以看出当 $\eta_i = d_{ij}^2$ 时,$u_{ij} = 0.5$,故 η_i 决定了聚类时类的范围,η_i 越大,在迭代时属于该类的点越多。另外还要说明的是,在 PCM 中权重指数 m 的增加表示数据集中的所有点完全属于指定的一个类的可能性增加了,m 过大会导致聚类产生一致的类中心,m 一般取为1.5。

PCM 本质上是一个穷举型搜索算法[5],算法需要进行适当的初始化才能收敛到全局最小点。一般推荐用 FCM 算法的结果作为 PCM 算法的初始划分,其算法流程如下:

1)确定聚类中心数 C,权重指数 m,最大迭代次数 l 及迭代终止误差 ε,并用 FCM 的聚类结果初始化其隶属度 u_{ij};

2)根据(2)式计算出 η_i;

3)根据(3)式计算出聚类中心 v_i^l,并根据隶属度计算公式[2]得到隶属度 u_{ij}^{l+1};

4)判断 $\|u_{ij}^l - u_{ij}^{l+1}\| < \varepsilon$,满足条件则迭代终止;否则转至3),继续迭代。

1.2 基于 Mercer 核的可能性聚类算法

设 $x_k \in \mathbf{R}^n, k = 1,2,\cdots,l$ 是原空间的样本点,利用一非线性映射 Φ 将原空间的样本映射到一个高维的核空间 H 中,得到 $\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_l)$,则原空间的点积在高维核空间可以用 Mercer 核来表示[6]为

$$K(x_i,x_j) = (\Phi(x_i),\Phi(x_j))。 \quad (4)$$

核函数具有两个特征:对称性和满足 Cauchy-Schwarz 不等式。常用的核函数为高斯核函数,它对应的特征空间是无穷维的,有限的样本在该特征空间肯定是线性可分的。高斯核函数为 $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right)$,其中 δ 为高斯核函数的宽度。

PCM 算法的目标函数在高维核空间 H 中改为

$$J_H = \sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij})^m \|\Phi(x_j) - \Phi(v_j)\|^2 + \sum_{i=1}^{C}\eta_i\sum_{j=1}^{N}(1-u_{ij})^m \quad (5)$$

式中,

$$Q_{ij} = \|\Phi(x_j) - \Phi(v_i)\|^2 = K(x_j,x_j) + K(v_i,v_i) - 2K(x_j,v_i) \quad (6)$$

$$\eta_i = K\frac{\sum_{j=1}^{N}u_{ij}^m Q_{ij}}{\sum_{j=1}^{N}u_{ij}^m} \quad (7)$$

其中,Q_{ij} 为高维核空间中第 j 个样本到第 i 个聚类中心的距离。当取高斯核函数时,$K(x,x) = 1$,故上式可简化为

$$Q_{ij} = 2(1 - K(x_j,v_i))。 \quad (8)$$

为使目标函数值达到最小,得到聚类中心 v_i 和隶属度 u_{ij} 的更新公式为

$$v_i = \frac{\sum_{j=1}^{N}u_{ij}^m K(x_j,v_i)x_j}{\sum_{j=1}^{N}u_{ij}^m K(x_j,v_i)} \quad (9)$$

$$u_{ij} = \frac{1}{1+\left(\frac{Q_{ij}}{\eta_i}\right)^{\frac{1}{m-1}}} \quad (10)$$

KPCM 的算法流程如下:

1)确定聚类中心数 C,权重指数 m,最大迭代次数 l 及迭代终止误差 ε,并用 FCM 的聚类结果初始化其隶属度 u_{ij};

2)根据(6)式计算高维核空间中样本到类中心的距离 Q_{ij};

3)根据(7)式 计算 η_i;

4)根据(9)、(10)式计算聚类中心 v_i^l 和隶属度 u_{ij}^{l+1};

5)判断 $\|u_{ij}^{l}-u_{ij}^{l+1}\|<\varepsilon$,满足条件则迭代终止;否则转至 3),继续迭代。

2　仿真实验[7]

为了验证本算法的有效性,将基于核的可能性聚类算法用于电力变压器的油中溶解气体分析。油中溶解气体分析样本和变压器发生故障模式类型有一种非线性对应关系,通过对样本的分析可以诊断出变压器的故障模式类型,但是由于测量和故障的复杂性使得样本中含有噪音,降低了故障诊断的精度。表 1 为 15 组变压器油中气体分析的样本数据。

表 1　15 组变压器油中气体分析的样本数据($\times 10^{-6}$)

序号	H_2	CH_4	C_2H_4	C_2H_6	C_2H_2
x_1	14.7	3.8	2.7	10.5	0.2
x_2	980	73	12	58	0
x_3	181	262	28	41	0
x_4	173	334	813	172	37.7
x_5	127	107	154	11	224
x_6	200	48	117	14	131
x_7	6.7	10	71	11	3.9
x_8	220	340	480	42	14
x_9	170	320	520	53	3.2
x_{10}	27	90	63	42	0.2
x_{11}	565	93	47	34	0
x_{12}	32.4	5.5	12.6	1.4	13.2
x_{13}	56	286	928	96	7
x_{14}	160	130	96	33	0
x_{15}	650	53	20	34	0

实际应用中,即使是同一种故障类型的变压器油中溶解气体数据,其原始数据相互间的差异也可能较大,因而宜对原始数据先进行规格化预处理,提取故障特征信息,这对提高聚类效果是非常重要的。规格化的方法较多,本文选用比例规格化方法得

$$x_{i1}=\frac{x_{i1}}{\sum_{j=1}^{5}x_{ij}},x_{ij}=\frac{x_{ij}}{\sum_{j=1}^{5}x_{ij}} \qquad (11)$$

$i=1,2,\cdots,n$; $j=2,3,\cdots,5$

经上述处理后,就得到了样本的规格化矩阵

$$X=(x_{ij})_{n\times 5}。$$

对上述样本分别应用 FCM、PCM、KPCM 算法进行聚类分析。应用 KPCM 算法,初始值 $m=1.5$,$\delta=2$,程序迭代 $l=3$ 次达到收敛,其收敛速度优于 FCM、PCM。得到的 5 个故障类型的聚类中心如表 2 所示,聚类结果如表 3 所示。

表 2　最优聚类中心

聚类中心	v_{i1}	v_{i2}	v_{i3}	v_{i4}	v_{i5}
V_1	45.76	22.48	60.46	15.84	1.21
V_2	11.61	26.37	8.43	63.34	1.85
V_3	82.84	51.13	30.45	18.18	0.24
V_4	36.51	81.01	3.74	35.81	42.44
V_5	28.82	57.87	15.46	26.55	0.17

表 3　聚类结果表

类别	样本序号
第 1 类	x_4,x_7,x_8,x_9,x_{13}
第 2 类	x_2,x_{11},x_{15}
第 3 类	x_5,x_6,x_{12}
第 4 类	x_1
第 5 类	x_3,x_{10},x_{14}

从聚类结果和变压器实际故障对比可以看出,应用 KPCM 算法 15 个样本中除第 7 个样本外,其余 14 个分类完全正确,聚类效果较好,正确判别率为 93.33%,而 FCM 算法的诊断精度为 80%,PCM 算法的诊断精度为 86.7%。

3　结论

基于核的可能性聚类算法 KPCM 算法利用核函数对原输入空间样本点的特征进行处理,突出了样本点之间的特征差异,能够较好地处理噪音和孤立点。收敛速度快,能提供更准确的聚类效果。仿真实验表明,其聚类效果明显优于 FCM 和 PCM 算法。

参考文献:

[1] KRISHNAPURAM R, KELLER J M. A Possibilistic Approach to Clustering[J]. IEEE Trans on Fuzzy Systems, 1993,1(2):98-110.

[2] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York:Plenum Press,1981.

[3] GIROLAMI M. Mercer Kernel-Based Clustering in Feature Space[J]. IEEE Trans on Neural Networks,2002,13(3):780-784.

[4] 张莉,周伟达,焦李成. 核聚类算法[J]. 计算机学报,2002,25(6):587-590.

[5] YANG M S,WU K L,YU J. A Novel Fuzzy Clustering Algorithm[A]. In Proc of the 2003 IEEE Int'l Symp. On Computational Intelligence in Robotics and Automation[C],Kobe:IEEE,2003,647-652.

[6] 沈红斌,王士同,吴小俊. 离群模糊核聚类算法[J]. 软件学报,2004,15(7):1021-1029.

[7] 吕佳. 可能性 C-Means 聚类算法的仿真实验[J]. 重庆师范大学学报(自然科学版),2005,22(3):129-132.

(责任编辑　游中胜)