

特征空间属性加权模糊核聚类算法

范新南¹, 沈红斌^{1,2}, 陈学忠¹

(1. 河海大学 计算机及信息工程学院, 江苏 常州 213022;

2. 上海交通大学 图像处理与模式识别研究所, 上海 200030)

(fanxn@hhuc.edu.cn)

摘 要: 充分考虑了属性间的不平衡性, 通过 Mercer 核把原始的观察空间映射到高维特征空间, 提出了一种新的特征空间中的加权模糊核聚类算法 WFKCA。众多实例表明, WFKCA 比传统的聚类算法具有更好的性能, 且对于高维数据具有很好的聚类效果。

关键词: 模糊聚类; 核; 模式识别

中图分类号: TP311.131 **文献标识码:** A

New mercer-kernel based Fuzzy clustering algorithm with attribute weights in feature space

FAN Xin-nan¹, SHEN Hong-bin^{1,2}, CHEN Xue-zhong¹

(1. College of Computer and Information Engineering, Hohai University, Changzhou Jiangsu 213022, China;

2. Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Considering the imbalance between the attributes fully, a new weighted fuzzy kernel-clustering algorithm named WFKCA was presented. WFKCA performs clustering in high feature space mapped by mercer kernels. Lots of examples demonstrate that WFKCA is a useful clustering tool.

Key words: Fuzzy clustering; kernel; pattern recognition

0 引言

聚类分析是数据分析、理解与数据可视化的有效工具^[1~10]。给定一个数据集 $\Gamma = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\} \subset R^N$, 聚类分析的目的在于根据某种相似性原则将 Γ 分成 C 个类别。聚类分析是统计理论、数据挖掘和模式识别领域中的一个重要课题^[1~3]。聚类分析理论已经在众多领域得到了广泛应用: 在商业竞争中, 聚类分析能帮助商家在客户数据库中根据客户的购物模式发现不同的组别, 从而更优地决策; 在信息迅速增长的今天, 聚类算法能有效的进行文本分类; 作为数据挖掘的有效工具, 聚类分析能帮助理解数据的不同分布, 观察各个类别的不同性质, 从而对其中感兴趣的类别进行进一步的分析。对于常用的基于分割的聚类算法而言, 其一般的方法是用 C 个类中心向量 $\vec{v}_j (j = 1, 2, \dots, C) \in R^N$ 来代表每个类, 根据相似性原则把样本 \vec{x}_k 分到第 j^{th} 类别中。

假设每个数据样本 \vec{x}_j 有 L 个属性: $\vec{x}_j = \{x_{j1}, x_{j2}, \dots, x_{jL}\}$, 传统的聚类算法, 如 K-means, FCM 等都假设样本的属性对每个类别是平等重要的。但实际上, 不同的属性对于不同类别的贡献可能是不相同的。文献[5]提出了在聚类分析之前进行特征选择的方法选择重要的特征进行聚类。由于文献[5]在聚类分析之前就确定了重要的属性, 因此聚类结果的好坏与产生数据样本的物理背景将密切相关; 另外, 它假设所有的类别都具有相同的重要属性集合, 而实际上, 不同的类别的重要属性集合可能并不是一致的。

为了解决这一问题, 可以在聚类过程中根据不同类的具

体特性动态计算各个属性对于不同类别的重要性。文献[6]提出了观察空间加权距离度量的硬聚类算法; 文献[7]提出了两个观察空间加权距离度量的模糊聚类算法 SCAD1, SCAD2。文献[6,7]所阐述的算法都存在和 FCM 一样的缺点: 它们适用于团状数据集, 但对于非团状的一般高维数据集的聚类则往往得不到理想的结果。一个解决非团状数据集的聚类方法是用核函数把数据从观察空间映射到高维特征空间^[8,9]。但传统的核聚类算法^[8,9]存在两个缺点: 1) 核聚类算法的聚类中心很难理解和表示; 2) 忽略了特征空间中的属性不平衡性。为此, 本文提出了加权模糊核聚类算法 WFKCA, 在聚类分析中动态计算属性间的不平衡性。实验结果表明, 对如高维非团状数据集, WFKCA 比传统的聚类算法具有更好的效果。

1 加权模糊核聚类算法 WFKCA

假设 ϕ 为一非线性映射函数, $\phi: p \rightarrow \phi(p) \in HS$, 其中 $p \in OS$, 是观察空间的一个样本, HS 表示映射以后的高维特征空间。我们提出以下 WFKCA 的目标函数:

$$J_{WFKCA} = \sum_{i=1}^C \sum_{j=1}^N \sum_{k=1}^L \mu_{ij}^m w_{ik}^\beta \|\phi(x_{jk}) - \phi(v_{ik})\|^2 \quad (1)$$

$$\text{s. t. } \mu_{ij} \in [0, 1] \quad \text{And} \quad \sum_{i=1}^C \mu_{ij} = 1 \quad 1 \leq j \leq N \quad (2)$$

$$w_{ik} \in [0, 1] \quad \text{And} \quad \sum_{k=1}^L w_{ik} = 1 \quad 1 \leq i \leq C \quad (3)$$

其中 C 为类别数, $\vec{v}_i = (v_{i1}, v_{i2}, \dots, v_{iL})$ 是第 i^{th} 类中心, μ_{ij} 表示第 j^{th} 样本属于第 i^{th} 类的隶属度; w_{ik} 代表第 k^{th} 个属性对

收稿日期: 2006-02-16; 修订日期: 2006-04-14

作者简介: 范新南(1964-), 男, 江苏常州人, 副教授, 博士, 主要研究方向: 图形图像处理、计算机网络通信; 沈红斌(1979-), 男, 江苏镇江人, 博士, 主要研究方向: 模式识别、生物信息学; 陈学忠(1968-), 男, 江苏常州人, 副教授, 主要研究方向: 数据挖掘、图像处理。

于第 i^h 类的权重, $m > 1, \beta > 1$ 。

根据式(1)可以得到:

$$\begin{aligned} \|\phi(x_{jk}) - \phi(v_{ik})\|^2 &= \phi(x_{jk}) \cdot \phi(x_{jk}) - 2\phi(x_{jk}) \cdot \phi(v_{ik}) + \\ &\quad \phi(v_{ik}) \cdot \phi(v_{ik}) \\ &= K(x_{jk}, x_{jk}) - 2K(x_{jk}, v_{ik}) + K(v_{ik}, v_{ik}) \end{aligned} \quad (4)$$

其中, $K(x, y) = \phi(x) \cdot \phi(y)$ 为用户定义的核映射函数。如果采用常用的高斯核函数:

$$K(\vec{x}, \vec{y}) = \exp\left(-\frac{\|\vec{x} - \vec{y}\|^2}{\sigma^2}\right)$$

那么 $K(x, x) = 1$, 则(1)就演变为:

$$\begin{aligned} J_{WFKCA} &= 2 \sum_{i=1}^C \sum_{j=1}^N \sum_{k=1}^L \mu_{ij}^m w_{ik}^\beta (1 - K(x_{jk}, v_{ik})) \\ &= 2 \sum_{i=1}^C \sum_{j=1}^N \sum_{k=1}^L \mu_{ij}^m w_{ik}^\beta \left(1 - \exp\left(-\frac{\|x_{jk} - v_{ik}\|^2}{\sigma^2}\right)\right) \end{aligned} \quad (5)$$

式(5)对 v_{ik} 求偏导得:

$$\begin{aligned} \frac{\partial J_{WFKCA}}{\partial v_{ik}} &= \frac{-4}{\sigma^2} \sum_{j=1}^N \mu_{ij}^m w_{ik}^\beta \cdot \exp\left(-\frac{\|x_{jk} - v_{ik}\|^2}{\sigma^2}\right) \cdot (x_{jk} - v_{ik}) \\ &= 0 \end{aligned} \quad (6)$$

解式(6)得:

$$v_{ik} = \frac{w_{ik}^\beta \sum_{j=1}^N \mu_{ij}^m \cdot \exp\left(-\frac{\|x_{jk} - v_{ik}\|^2}{\sigma^2}\right) \cdot x_{jk}}{w_{ik}^\beta \sum_{j=1}^N \mu_{ij}^m \cdot \exp\left(-\frac{\|x_{jk} - v_{ik}\|^2}{\sigma^2}\right)} \quad (7)$$

根据 w_{ik} 的不同取值, 得到:

$$\begin{cases} v_{ik} = 0 & \text{if } w_{ik} = 0 \\ v_{ik} = \frac{\sum_{j=1}^N \mu_{ij}^m \cdot K(x_{jk}, v_{ik}) \cdot x_{jk}}{\sum_{j=1}^N \mu_{ij}^m \cdot K(x_{jk}, v_{ik})} & \text{if } w_{ik} \neq 0 \end{cases} \quad (8)$$

由 $\sum_{i=1}^C \mu_{ij} = 1$, 根据拉格朗日方程得到以下无限制方程:

$$\begin{aligned} \tilde{J}_{WFKCA} &= 2 \sum_{i=1}^C \sum_{j=1}^N \sum_{k=1}^L \mu_{ij}^m w_{ik}^\beta (1 - K(x_{jk}, v_{ik})) - \\ &\quad \sum_{j=1}^N \lambda_j \left(\sum_{i=1}^C \mu_{ij} - 1 \right) \end{aligned} \quad (9)$$

其中, λ_j 为拉各朗日系数。

式(9)对 μ_{ij} , 并最终解得:

$$\mu_{ij} = \frac{1}{\left(\sum_{r=1}^C \frac{\sum_{k=1}^L w_{ik}^\beta (1 - K(x_{jk}, v_{ik}))}{\sum_{k=1}^L w_{rk}^\beta (1 - K(x_{jk}, v_{rk}))} \right)^{\frac{1}{\beta-1}}} \quad (10)$$

同理, 由 $\sum_{k=1}^L w_{ik} = 1$ 得到以下的无限制函数:

$$\begin{aligned} \tilde{J}_{WFKCA} &= 2 \sum_{i=1}^C \sum_{j=1}^N \sum_{k=1}^L \mu_{ij}^m w_{ik}^\beta (1 - K(x_{jk}, v_{ik})) - \\ &\quad \sum_{i=1}^C \lambda_i \left(\sum_{k=1}^L w_{ik} - 1 \right) \end{aligned} \quad (11)$$

解 w_{ik} 得:

$$\mu_{ij} = \frac{1}{\left(\sum_{i=1}^L \frac{\sum_{j=1}^N \mu_{ij}^m (1 - K(x_{jk}, v_{ik}))}{\sum_{j=1}^N \mu_{ij}^m (1 - K(x_{jk}, v_{ik}))} \right)^{\frac{1}{\beta-1}}} \quad (12)$$

(1), (8), (10), (12) 构成了 WFKCA 聚类算法。

2 实例分析

2.1 IRIS 数据集测试

这个实例中, 我们用标准数据集 IRIS 来测试 FCM, SCAD1, SCAD2 和 WFKCA 算法的性能。IRIS 数据集共有 150 个样本, 分成 3 类, 其中, 每个样本有 4 个属性 $\hat{x}_i = \{x_{i1}, x_{i2}, x_{i3}, x_{i4}\}$ 。设定参数 $m = 2, \beta = 2, \sigma$ 为标准方差。

表 1 给出了各算法在 100 次实验中的平均性能, 除了 SCAD2 外所有算法的起始中心点为随机产生。实验结果发现, SCAD2 对起始中心的选择非常敏感, 表 1 给出了随机选择中心 SCAD2_1 以及把 FCM 迭代 10 次后的中心点作为起始中心点的 SCAD2_2 的性能比较, 可以发现, SCAD2_2 性能比 SCAD2_1 有很大提高。

表 1 各算法在 IRIS 标准数据集的性能比较

算法	平均错分样本数目	聚类精度 (%)	平均迭代步数
FCM	16	89.33	38.58
SCAD1	17.05	88.63	15.20
SCAD2_1	27.85	81.43	19.85
SCAD2_2	11	92.67	15
WFKCA	6	96	19.22

从表 1 可以看出, WFKCA 最终的聚类结果在四个算法中是最好的。同时, 表 1 也证明了 WFKCA, SCAD1 及 SCAD2 的算法收敛迭代次数要比 FCM 少, 这和 WFKCA, SCAD1 和 SCAD2 的权重系数在迭代过程中所产生的引导作用密切相关。

2.2 高维数据聚类分析

本实例采用文献[11]中鉴别宋代官汝窑器与民汝窑器的数据作为测试数据集, 该数据集中每个样本有 10 个属性。表 2 给出了各算法在 100 次试验中在该数据集上的平均聚类结果。从结果可以看出来, WFKCA 具有最好的性能。这表明 WFKCA 算法在处理高维数据集中同样存在着优势。

表 2 高维数据各算法性能比较

算法	平均错分样本数	聚类精度 (%)
FCM	3	88.89
SCAD_1	2.5	90.74
SCAD_2	7	74.1
KCA ^[8]	3.14	88.37
WFKCA	1	96.3

3 结语

传统的聚类算法, 如 FCM 等对于“非团状”的数据集很难得到理想的聚类结果, 主要是由于下面的原因:

(1) 算法在观察空间中进行聚类分析, 对“团状”数据集有效, 但对一些更一般的数据集, 特别是高维数据集将不能得到理想的结果;

(2) 传统的聚类算法, 如 FCM 等, 忽略了属性间的不平衡性, 而这一点在实际生活中常常存在。

本文基于核函数, 将数据集从观察空间映射到高维特征空间, 提出了特征空间中新的属性加权模糊核聚类算法 WFKCA。实验表明, 对于一般高维数据集, WFKCA 能更好地反映属性间的不平衡性, 并得到很好的聚类结果。

(下转第 1915 页)

3) 将特性 LTL 公式翻译为一个自动机。

4) 计算自动机与全局有限状态自动机的同步积, 形成一个自动机。

5) 检查最后这个自动机的语言是否为空。如果为非空, 则所描述的进程满足指定的特性模式。

6) SPIN 中, LTL 公式用来描述坏的特性模式, 如果自动机的语言为非空, 则知道描述的系统不能满足特性模式要求, 此时可以发现反例, 这对调试非常有用。

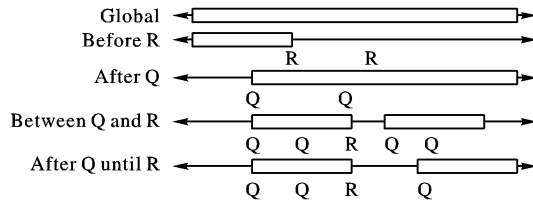


图2 特性作用范围

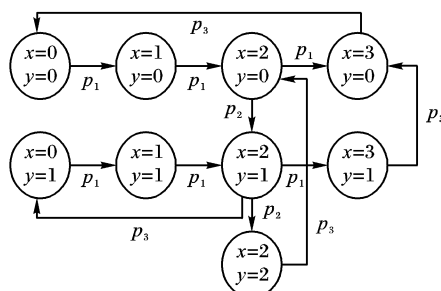


图3 特性模式的分类

这里, 给出一个如图3所示的状态图, 其对应的 SPIN 代码如下所示:

```
int x = 0, y = 0;
proctype p1() {
  do
    :: atomic{ (y < 2) && (x < 3) -> x = x + 1 }
  od
}
proctype p2() {
  do
    :: atomic{ (x + y) == 3 -> x = 0 }
    :: atomic{ (x + y) == 4 -> y = 0 }
  od
}
proctype p3() {
  do
    :: atomic{ (y < 2) && (x < 3) -> x = x + 1 }
  od
}
init {
  run p1();
```

```
run p2();
```

```
run p3();
```

在 SPIN 4.2.5 下, 分别验证其下列特性:

1) 安全性示例: 验证系统是否满足 LTL 公式 $\Box \Diamond (x + y) = 1$, 通过 SPIN 验证和仿真功能, 发现 (x, y) 的冲突序列为: $(0, 0), (1, 0), (2, 0), (2, 1), [(2, 2), (0, 2), (1, 2)]^*$ 。

2) 活性示例: 验证系统是否满足 $\Box (2x > y)$, 通过 SPIN 验证和仿真功能, 发现冲突序列为: $(0, 0)$ 。

3) 响应性示例: 验证系统是否满足 $\Box ((x + y = 1) \Rightarrow \Diamond (x + y = 3))$, 通过 SPIN 验证功能, 特性可满足。

6 结语

高层次的定义和抽象是自动编写形式化规格说明的重要方法, 线性时态逻辑特性模式特别适用于有限状态验证工具 SPIN 中的规格说明。本文从特性模式的分类和作用范围描述 LTL 公式, 对自动编写系统规格说明特性有很大帮助, 如文献[7]给出了 SPIN 中常用的特性模式是其应用的一个方面。

参考文献:

- [1] CLARKE EM, SCHLINGLOFF BH. Model Checking[A]. Handbook of Automated Reasoning[C]. Band II, S. Elsevier, 2001. 1637 - 1790.
- [2] BÉRARD B, BIDOIT M, FINKEL A. Systems and Software Verification: Model-Checking Techniques and Tools[M]. Berlin: Springer, 1999.
- [3] KRIPKE SA. Semantical considerations on modal logic[J]. Acta Philosophica Fennica, 1963, 16: 83 - 94.
- [4] PNUELI A. The Temporal Semantics of Concurrent Programs[A]. Proceedings of the International Symposium on Semantics of Concurrent Computation, Lecture Notes In Computer Science[C]. Springer-Verlag, 1979, Vol 70.
- [5] BEN-ARI M, PNUELI A, MANNA Z. The Temporal Logic of Branching Time[J]. Acta Informatica, 1983, 20(3): 207 - 226.
- [6] EMERSON EA, HALPERN JY. "Sometimes" and "Not Never" revisited: on branching versus linear time temporal logic[J]. Journal of the ACM, 1986, 33(1): 151 - 178.
- [7] HOLZMANN GJ. Spin Model Checker: The Primer and Reference Manual[M]. New York: Addison Wesley, 2003. 608.
- [8] HOLZMANN GJ. Design and validation of computer protocols[M]. London: PRENTICE-HALL, 1991. 22 - 78.
- [9] HOLZMANN GJ. The Model Checker SPIN[J]. IEEE transactions on software engineering, 1997, 23(5).

(上接第 1889 页)

参考文献:

- [1] HOPPER F. Fuzzy cluster analysis[M]. Chichester: John Wiley, 1999.
- [2] HAN JW, KAMBER M. Data mining: Concept and Techniques[M]. San Mateo: Morgan Kaufmann, 2001.
- [3] BEZDEK JC. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press, 1981.
- [4] 沈红斌, 王士同. 离群模糊核聚类算法[J]. 软件学报, 2004, 15(7): 1021 - 1029.
- [5] YEUNG DS, WANG XZ. Improving Performance of Similarity-Based Clustering by Feature Weight Learning[J]. IEEE Transactions On PAMI, 2002, 24(4): 556 - 561.
- [6] CHANA EY, CHINGA WK. An optimization algorithm for clustering using weighted dissimilarity measures[J]. Pattern Recognition, 2004, 37(5): 943 - 952.
- [7] FRIGUI H, NASRAOUI O. Unsupervised learning of prototypes and attribute weights[J]. Pattern Recognition, 2004, 37: 567 - 581.
- [8] 张莉, 周伟达, 焦李成. 核聚类算法[J]. 计算机学报, 2002, 25(6): 587 - 590.
- [9] GIROLAMI M. Mercer Kernel - Based Clustering in Feature Space[J]. IEEE Transactions on Neural Networks, 2002, 13(3): 780 - 784.
- [10] 沈红斌, 杨杰, 王士同. 基于信息理论的合作模糊聚类算法研究[J]. 计算机学报, 2005, 8: 1287 - 1294.
- [11] 王焜, 陆文聪. 宋代汝窑古瓷的微量元素——支持向量机算法研究[J]. 计算机与应用化学, 2004, 2: 191 - 194.