第 19 卷 第 11 期　　控 制 与 决 策　　2004 年 11 月
Vol 19 No. 11　　Control and Decision　　Nov. 2004

文章编号: 1001-0920(2004)11-1250-05

模糊 CLOPE 算法及其参数优选

李　洁, 高新波, 焦李成
(西安电子科技大学 电子工程学院, 陕西 西安 710071)

摘　要: 提出一种模糊 CLOPE 算法, 并定义了修正划分模糊度, 将其作为新的聚类有效性函数来实现参数的自动优选. 对真实数据测试的实验结果表明, 模糊 CLOPE 算法以及基于修正划分模糊度的参数优选方法是非常有效的.

关键词: 数据挖掘; 聚类分析; 聚类有效性; 类属特征; 参数优选
中图分类号: TP391　　**文献标识码**: A

Fuzzy CLOPE algorithm and its parameter optimal choice

LI Jie, GAO Xin-bo, JIAO Li-cheng
(School of Electronic Engineering, Xidian University, Xi'an 710071, China. Correspondent: LI Jie, E-mail: leejie@mail.xidian.edu.cn)

Abstract: A fuzzy CLOPE algorithm is proposed and a method for the parameters optimal choice is presented by defining a modified partition fuzzy degree as a clustering validity function. The experimental results with real data set show the effectiveness of the proposed fuzzy CLOPE algorithm and parameter optimal choice method based on the modified partition fuzzy degree.

Key words: data mining; cluster analysis; cluster validity; categorical attributes; parameter optimal choice

1 引　　言

在数据挖掘和知识发现领域, 将样本集划分成各种不同的子集(类)是一种基本的操作[1]. 这种方法已在许多领域获得了广泛的应用, 比如分类(无监督)、聚合、划分或解剖[2]等. 聚类分析[3]是其中一种相当流行的样本集近似划分方法.

聚类分析常常需要处理大量的高维数据集[3](具有几十或几百个特征的数千甚至几百万个记录). 这使许多传统的聚类算法不能直接用于数据挖掘; 同时, 数据挖掘中经常会遇到类属特征的数据, 即样本的各维特征为类别、符号或概念. 传统的处理方法是将类属值转化为数值再进行分析. 由于类属域是无序的, 传统的处理方法并不能奏效. 为此, 人们提出一些针对类属型数据的聚类方法. 然而, 现有的聚类分析方法或能处理类属型数据, 但不适合大数据集分析; 或能有效分析大数据集, 却仅限于数值型数据. 只有少数几种算法能兼顾上述两个要求, 如 k-原型算法[4,5]和 CLOPE 算法[6]等.

在实际应用中, 大多数聚类分析算法会遇到另一个问题, 即必须在聚类分析之前, 给定合理的聚类类别数 c 和与 c 相关的其他参数, 而参数的取值正确与否将直接影响到分类结果的合理解释. 分析聚类结果属于聚类有效性研究的课题. 目前, 聚类有效性的研究主要集中在数据集的模糊划分、几何结构以及统计信息这 3 个方面[7]. 这些方法在设计时大多是针对数值型数据, 而没有考虑到类属型数据的情况.

为此, 本文提出了划分模糊度(PFD)的概念, 并

收稿日期: 2004-01-14; **修回日期**: 2004-07-12
基金项目: 国家自然科学基金资助项目(60202004, 60073053).
作者简介: 李洁(1972—), 女, 陕西西安人, 讲师, 博士生, 从事数据挖掘、图像处理等研究; 焦李成(1959—), 男, 陕西白水人, 教授, 博士生导师, 从事非线性系统、神经网络等研究.

兼顾数据集的模糊划分信息和几何结构信息, 将划分熵与划分模糊度相结合, 定义了一种修正的划分模糊度作为聚类有效性评价函数. 该聚类有效性函数能有效地分析类属型数据. 同时, 将模糊集理论引入传统的 CLOPE 算法, 形成模糊 CLOPE 算法, 并利用修正划分模糊度进行参数优选, 实现了真正意义上的类属数据无监督的聚类分析.

2　修正划分模糊度的定义

令 $X = \{x_1, x_2, \cdots, x_n\}$ 表示一组具有 n 个样本的数据集, 其中 $x_j = [x_{j1}, x_{j2}, \cdots, x_{jp}]^T \in R^p$ 表示第 j 个样本的 p 维特征矢量; $c(1 < c < n)$ 是预先指定的聚类类别数; $c \times n$ 阶矩阵 $U = [u_{ij}]$ 是数据集 X 的模糊划分矩阵, u_{ij} 是第 j 个样本属于第 i 个聚类的隶属度. 为衡量模糊聚类结果的模糊程度, 仿照 Shannon 信息熵的公式, 提出如下模糊划分熵概念:

2.1　模糊划分熵

定义 1　对于给定的聚类数 c 和模糊划分矩阵 U, 数据集 X 的模糊划分熵定义为

$$H(U;c) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_a(u_{ij}). \qquad (1)$$

其中: $a \in (1, \infty)$ 为对数的底数, 约定当 $u_{ij} = 0$ 时 $u_{ij}\log_a(u_{ij}) = 0$.

划分熵具有如下性质:

定理 1　对于 $1 < c < n$ 和任意的模糊划分矩阵 U, 有:

1) $0 \le H(U;c) \le \log_a c$;

2) $H(U;c) = 0$, 当且仅当 U 是硬划分;

3) $H(U;c) = \log_a c$, 当且仅当 $U = [1/c]$.

证明参见文献[8].

从上述性质可知, 划分熵是一个衡量聚类结果模糊程度的标准. 划分结果越分明, $H(U;c)$ 的值就越小; 反之, 划分结果越模糊, $H(U;c)$ 的值就越接近于 $\log_a c$. 记 Ω_c 为所有最优划分矩阵的有限集, 如果存在 (U^*, c^*), 满足

$$H(U^*, c^*) = \min_c\{\min_{\Omega_c} H(U;c)\},$$

则 (U^*, c^*) 对应于最有效的聚类结果, c^* 是最佳的分类数.

除了模糊划分熵外, 本文还从矩阵范数的角度定义了一种划分模糊度, 以度量数据集划分结果的模糊程度.

2.2　划分模糊度

定义 2　对于给定的聚类数 c 和模糊划分矩阵, 数据集 X 的划分模糊度定义为

$$PF(U;c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\left|u_{ij} - (u_{ij})_H\right| \qquad (2)$$

其中

$$(u_{ij})_H = \begin{cases} 1, u_{ij} = \max\limits_{1\le k\le c}\{u_{kj}\}, \\ 0, \text{otherwise}, \end{cases}$$

对应于数据集模糊划分的最贴近的硬划分矩阵.

定理 2　对于 $1 < c < n$ 和任意的模糊划分矩阵 U, 有:

1) $0 \le PF(U;c) \le 2 - 2/c$;

2) $PF(U;c) = 0$, 当且仅当 U 是硬划分;

3) $PF(U;c) = 2 - 2/c$, 当且仅当 $U = [1/c]$.

证明参考文献[8].

从上述性质可知, 划分模糊度可作为分类模糊性的度量. 数据集的划分结果越分明, $PF(U;c)$ 的值就越小; 分类越模糊, $PF(U;c)$ 的值就越接近于 $2 - 2/c$. 为获得最佳的类别数 c, 人们希望得到的模糊划分 $PF(U;c)$ 越小越好. 然而, 划分熵 $H(U;c)$ 以及划分模糊度 $PF(U;c)$ 都随着类别数 c 的增加而呈递增趋势. 这种递增的趋势影响了对 $H(U;c)$ 和 $PF(U;c)$ 的全局或局部极值点的检测, 而这些极值点恰恰对应于一些合理的聚类类别数. 为此, 本文结合 $H(U;c)$ 和 $PF(U;c)$, 提出一个修正的划分模糊度 $MPF(U;c)$.

2.3　修正划分模糊度

定义 3　对于给定的类别数 c 和模糊划分矩阵 U, 数据集 X 的修正划分模糊度定义为

$$MPF(U;c) = \frac{PF(U;c)}{\tilde{H}(U;c)}. \qquad (3)$$

其中: $PF(U;c)$ 如式 (2) 所定义; $\tilde{H}(U;c) = Smooth(H(U;c))$ 是平滑后的模糊划分熵, 可用 3 点线性平滑算子或非线性中值滤波来实现; 约定当 U 是硬划分时, $MPF(U;c) = 0$.

修正的划分模糊度补偿了由于类别数 c 的增加所引起的划分模糊度的递增趋势, 从而简化了最佳类别数的判定问题.

3　模糊 CLOPE 算法及其参数优选

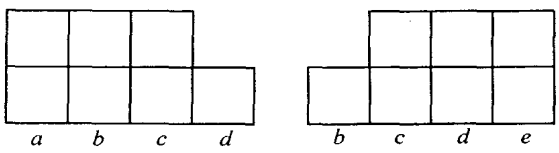
为实现对类属型数据的模糊聚类分析, 本文将 CLOPE 算法扩展到模糊的情况, 并给出一种基于修正划分模糊度的聚类参数优选法.

3.1　CLOPE 算法

Yang 提出一种针对交易数据的聚类算法——CLOPE 算法[6], 并提出一种基于统计直方图的聚类准则函数. 随着每一类中交易数据重合的增多,

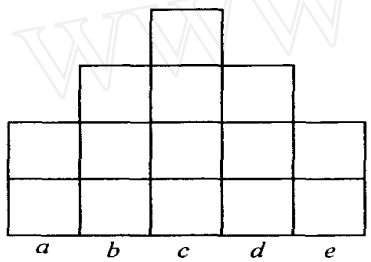
类内数据的统计直方图的高宽比也逐渐增加, 而当各类数据统计直方图的高宽比之和达到最大时, 所对应的分类结果被认为是最优划分.

这里用一个例子来说明该算法: 设有一小型的数据集{*abc*, *abcd*, *bcde*, *cde*}, 将数据集分割为下面两种情况: 1) {{*abc*, *abcd* }, {*bcde*, *cde*}}; 2) {{*abc*, *abcd*, *bcde*, *cde*}}. 对于每一类, 分别统计其直方图及其每类的平均高度H 和宽度W, 如图 1 所示



H=1.75 ,W=4　　H=1.75, W=4

(a) 划分 1)



H=2.8、W=5

(b) 划分 2)

图 1 两种划分的统计直方图

对于上述两种划分, 划分 1) 的高宽比为H/W = 1.75/4 + 1.75/4 = 0.875, 划分 2) 的高宽比为H/W = 2.8/5 = 0.56. 对于CLOPE 算法, 显然划分 1) 比划分 2) 好.

可将 CLOPE 算法描述如下: 令 X = {x_1, x_2, …, x_n} 表示一组具有 n 个样本的数据集, x_j 的所有特征都是类属型的; {X_1, X_2, …, X_c} 表示数据集 X 的 c 个划分, H_i 是 X_i 的不同类属特征的统计直方图. 定义

$$S(X_i) = \sum_{x_j \in X_i} |x_j|, \tag{4}$$

式中 $|x_j|$ 表示 x_j 中特征的维数;

$$W(X_i) = |H_i|, \tag{5}$$

式中 $|H_i|$ 表示统计直方图 H_i 的柄数. 定义判别函数

$$\max\{\text{Profit}(X) = \frac{1}{n}\sum_{i=1}^{c}\frac{S(X_i)}{W(X_i)^2}|X_i|\}, \tag{6}$$

式中 $|X_i|$ 表示集合 X_i 的势, 即 X_i 中所包含元素的个数.

为控制类内的相似度, CLOPE 算法引入了一个排斥因子 r, 则式(6) 可改写为

$$\text{Profit}_r(X) = \frac{1}{n}\sum_{i=1}^{c}\frac{S(X_i)}{W(X_i)^r}|X_i|, \tag{7}$$

对于每一个确定的 r, 都能找到一个划分 X^* 和类别数 c, 使得 $\text{Profit}_r(X)$ 最大. 对于上述例子, 要使划分 1) 优于划分 2), 必须满足

$$2 \times 7/4^r + 2 \times 7/4^r > 4 \times 14/5^r,$$

此时应有

$$r > \ln(7/14)/\ln(4/5) = 3.106$$

3.2 模糊 CLOPE 算法

本文在 CLOPE 算法的基础上, 提出一种模糊 CLOPE 算法. 令 X = {x_1, x_2, …, x_n} 表示一组具有 n 个样本的数据集, x_j = {x_1^j, x_2^j, …, $x_{m_j}^j$} 表示第 j 个样本(当 $k \neq l$ 时 $x_k^j \neq x_l^j$, k, l = 1, 2, …, m_j), u_{ij} 表示样本 x_j 属于第 i 类的隶属度, U = [u_{ij}] 是一个 $c \times n$ 阶的模糊划分矩阵, {X_1, X_2, …, X_c} 表示数据集 X 的 c 个模糊划分, H_i 是 X_i 的不同类属特征的统计直方图. 定义

$$S_f(X_i) = \sum_{x_j \in X_i} u_{ij}|x_j|, \tag{8}$$

$$W_f(X_i) = |H_i|, \tag{9}$$

此时, 式(7) 表示的判别函数可写成

$$\max\{\text{Profit}_{fr}(X) = \frac{1}{n}\sum_{i=1}^{c}\frac{S_f(X_i)}{W_f(X_i)^r}|X_i|\}. \tag{10}$$

另外, 定义数据集 X 的特征集

$$C_X = x_1 \cup x_2 \cup \cdots \cup x_n, \tag{11}$$

划分 X_i 的特征集表示为 C_{X_i}, 则 X_i 的统计直方图 H_i 可描述为

$$H_i = \{\alpha_1, \alpha_2, \cdots, \alpha_{W_f(X_i)}\}, \tag{12}$$

式中 α_k (k = 1, 2, …, $W_f(X_i)$) 表示划分 X_i 的第 k 个特征. 将 x_j 对于直方图 H_i 的贴近度作为 x_j 属于第 i 类的隶属度 u_{ij}, 即

$$u_{ij} = \frac{\sum_{\alpha \in C_{X_i} \cap x_j} H_i(\alpha) - \eta|x_j - C_{X_i}|}{|X_i|}, \tag{13}$$

式中 η 为一常数, 称为惩罚因子, 一般取 $\eta \in (1, 2)$.

该算法的步骤如下:

Step1: 设定排斥因子 r 以及惩罚因子 η.

Step2: 读取一个样本 x_j, 将 x_j 分别置入现有的划分 X_i (i = 1, 2, …, k) 以及一个新的划分 X_{k+1} 中, 分别计算 $\text{Profit}_{fr}(X)$. 若 x_j 置入 X_i 划分时, $\text{Profit}_{fr}(X)$ 最大, 则将 x_j 分为第 i 类 ($1 \leq i \leq k+1$); 若 i = k + 1, 则令 k = k + 1.

Step3: 重复 Step2, 直到所有的 x_j 都不再改变类别为止.

这时面临的难题是如何确定 r 的取值, 以确保得到的结果是合理的最优划分. 为解决这一问题, 采用第 2 节介绍的修正划分模糊度作为聚类有效性函数, 以确定 r 的最佳取值.

3.3　基于修正划分模糊度的参数优选

根据以上描述的模糊 CLOPE 算法, 可以看出对于每个确定的 r, 都能找到一个划分 X^* 和类别数 c, 使得 $\mathrm{Profit}_r(X)$ 最大. 显然, 一个最优的 (X^*, c^*) 对应于一个确定的 r, 所以寻找最佳类别数 c^*, 可以转化为确定最优的 r^*.

修正划分模糊度也是 r 的函数. 利用如下准则来确定最佳的 r^*:

$$\mathrm{MPF}(U^*, r^*) = \min_r\{\min_{\Omega_r}\mathrm{MPF}(U; r)\}, \quad (14)$$

进而得到最优的类别数 c^*.

找到最佳的 r^* 后, 便可获得相应的类别数 c^* 以及合理的最优划分 X^*.

4　实验结果

为验证模糊 CLOPE 算法的性能以及基于修正划分模糊度的参数选择效果, 这里利用两组实际类属型数据进行测试实验.

4.1　大豆疾病数据集测试实验

实验中使用的数据是大豆疾病的实际数据[9]. 大豆疾病数据共有 47 个记录, 每个记录由 35 个特征描述. 每个记录都被标记为 4 种疾病中的一种: Diaporthe stem canker, Charcoal rot, Rhizoctonia root rot 和 Phytophthora rot. Phytophthora rot 有 17 个记录, 其余每种疾病都有 10 个记录.

首先将每个特征作为交易数据中的一个交易项, 这样 35 个特征便可转换成 84 种交易项, 每个记录用 35 个交易项表示. 然后令 r 取遍 $r_{\min}$～$r_{\max}$ ($r_{\min}$ 是使类别数 $c = 2$ 的实数, 当类别数 c 不再随 r 而增加或 $c = 2\ln n$ 时, $r = r_{\max}$), 采用模糊 CLOPE 算法 ($\eta = 1$) 得到的划分熵和划分模糊度如图 2 所示. 可以看出, 无论是划分熵还是划分模糊度, 都随 r 的增加而呈递增趋势, 所以不可能获得最优类别数.

采用本文的修正划分模糊度如图 3 所示. 为便于直观显示 r 与 c 的关系, 图中同时画出了 c 随 r 的变化曲线. 可以看出, 在 $r = 1.6$～1.8 时, 修正划分模糊度达到最小, 这时对应的类别数为 4, 正好与真实情况相符.

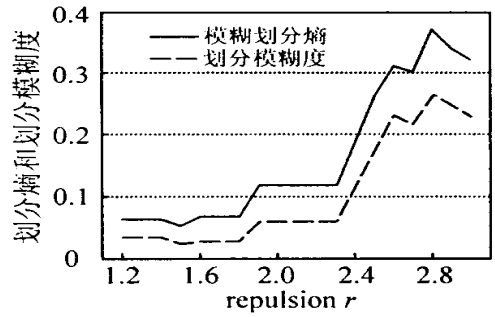


图 2　PF(U; c) 及 H(U; c) 随 r 变化曲线

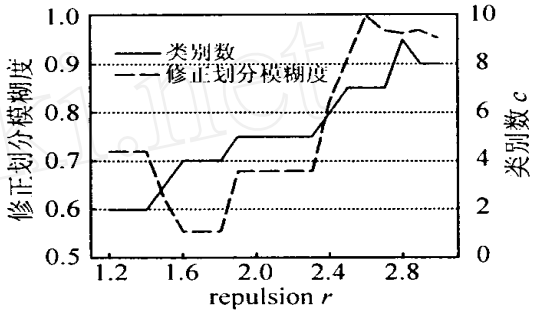


图 3　MPF(U; c) 及 c 随 r 变化曲线

表 1　r = 1.6～1.8 时的分类结果

	Cluster1	Cluster2	Cluster3	Cluster4
D			10	
C	10			
R				10
P		17		

所得分类结果如表1所示, 其中 D, C, R, P 分别表示每一种大豆疾病. 此时所有的样本都被正确分类, 说明选择的参数是合理的, 所提出的模糊 CLOPE 算法是有效的.

4.2　蘑菇数据集测试实验

在数据挖掘中, 经常需要处理大数据集. 本实验采用蘑菇数据集[6], 该数据集包含 8 124 个记录, 每个记录由 22 个特征描述, 分别表示蘑菇伞的形状、表皮、颜色等信息. 将 22 个特征转换为 116 个交易项, 这样每个记录便可用 22 个交易项来表示. 数据集中有 2 480 个记录缺少第 11 个特征 Stalk-root, 认为这 2 480 个记录每个记录只包含 21 个交易项即可. 数据集分为两类: 4 208 个可食用的和 3 916 个有毒的.

令 $r_{\min} = 0.1$, $r_{\max} = 1$ ($r_{\min}$ 和 $r_{\max}$ 的取值原则与 4.1 节实验相同), 采用模糊 CLOPE 算法得到数据集的隶属度矩阵, 并得到相应的修正划分模糊度曲线, 如图 4 所示. 从图中可以看到, 当 $r = 0.1$ 时, 修正划分模糊度达到极小值, 这时对应的类别数 $c = 2$, 与实际情况相符.

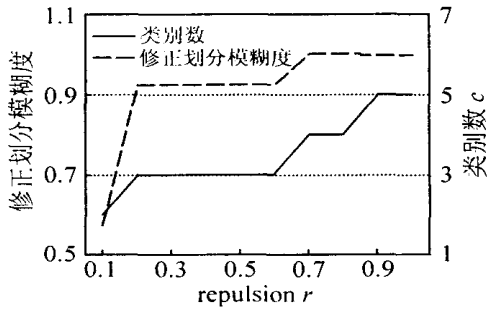


图 4 MPF(U; c) 及 c 随 r 变化曲线

5 结 论

本文提出一种模糊 CLOPE 算法, 实现了对类属型数据的聚类分析; 同时提出一种修正的划分模糊度, 用它构造了一个聚类有效性函数, 以实现算法中参数的自动选取. 通过对实际数据的分析表明, 该算法能有效地分析类属型数据, 基于修正划分模糊度的参数选取是合理而有效的.

参考文献(References):

[1] Klosgen W, Zytkow J M. Knowledge discovery in databases terminology [A]. *Advances in Knowledge Discovery and Data Mining* [C]. AAAI Press/The MIT Press, 1996. 573-592.

[2] Cormack R M. A review of classification [J]. *J Roy Statist Soc Serie A*, 1971, 134: 321-367.

[3] Anderberg M R. *Cluster Analysis for Applications* [M]. New York: Academic Press, 1973.

[4] Zhexue Huang, Michael K Ng. A fuzzy *k*-modes algorithm for clustering categorical data [J]. *IEEE Trans on Fuzzy Systems*, 1999, 7(4): 446-452.

[5] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining [A]. *Proc of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery* [C]. ACM Press, 1997. 1-8.

[6] Yiling Yang, Xudong Guan. CLOPE: A fast and effective clustering algorithm for transactional data [A]. *The Eighth ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining* [C]. Edmonton, 2002.

[7] Gao Xinbo, Xie Weixin. Advances in theory and applications of fuzzy clustering [J]. *Chinese Science Bulletin*, 2000, 45(11): 961-970.

[8] **高新波. 模糊聚类算法的优化及应用研究** [D]. **西安: 西安电子科技大学**, 1999.

[9] Michalski R S, Stepp R E. Automated construction of classifications: Conceptual clustering versus numerical taxonomy [J]. *IEEE PAMI*, 1983, 5: 396-410.

(**上接第 1249 页**)

[7] Vincezo L, Bruno S, Luigi V. Position and orientation estimation based on kalman filtering of stereo images [A]. *Proc of the 2001 IEEE Int Conf on Control Applications* [C]. Banff: IEEE Press, 2001. 702-707.

[8] Vince M. Dynamics and system performance of visual servoing [A]. *IEEE Conf on Robotics and Automation* [C]. San Francisco: IEEE Press, 2000. 644-648.

[9] Liu H, Ramon R C, Paulo L S A. Stable adaptive visual servoing for moving targets [A]. *Proc of the American Control Conf* [C]. Arlington: Machinery, 2000. 2008-2012.

[10] **林靖, 徐强华, 陈辉堂, 等. 基于图像差的平面大范围视觉伺服控制** [J]. **控制与决策**, 2000, 15(5): 581-584.
(Lin J, Xu Q H, Chen H T, et al. Image-error-based planar global visual servoing [J]. *Control and Decision*, 2000, 15(5): 581-584.)

[11] **夏利民, 谷士文, 樊晓平, 等. 基于活动轮廓的机器人视觉伺服** [J]. **机器人**, 2000, 22(5): 359-364.
(Xia L M, Gu S W, Fan X P, et al. Robotic visual servoing based on active control [J]. *Robot*, 2000, 22(5): 359-364.)

[12] Prokop R J, Reeves A P. A survey of moment-based object representation and recognition [J]. *CVGIP: Graphical Model and Image Processing*, 1992, 54(5): 438-460.

[13] Martinez J, Thomas F. A reformation of gray-level image geometric moment computation for real-time application [A]. *Proc of IEEE Int Conf on Robotics and Automation* [C]. Minncapolis: IEEE Service Center, 1996. 2315-2320.