基于 *K*-Means动态聚类的投影寻踪分类模型

姚 奕[1,2], 倪 勤[2]

(1. 南京师范大学数学科学学院, 江苏 南京 210046)
(2. 南京航空航天大学经济与管理学院, 江苏 南京 210016)

[摘要] 投影寻踪分类模型作为一种多因素影响问题的综合评价方法, 已经被研究者广泛应用在各个领域并取得了良好的效果. 然而模型本身还存在密度窗宽不确定以及模型无分类规则等尚需解决的问题. 针对这些问题, 提出一个基于 *K*-Means动态分类的投影寻踪分类模型, 定义了一个新的投影指标. 实证分析说明了该模型的可靠性和可操作性.

[关键词] 投影寻踪分类, 动态聚类, 投影指标, 遗传算法
[中图分类号] O212; TP18 [文献标识码] A [文章编号] 1001-4616(2009)04-0016-05

A Projection Pursuit Classification Model Based on *K*-Means Dynamic Cluster

Yao Yi[1,2], Ni Qin[2]

(1. School of Mathematical Sciences, Nanjing Normal University, Nanjing 210046, China)
(2. College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract As a comprehensive evaluation model, projection pursuit classification model has been applied in research widely and gained successful results. However, there are some drawbacks, such as the uncertain cutoff radius, lacking classification rules. In order to solve these problems, a projection pursuit classification model based on *K*-Means dynamic cluster is proposed and a new projection index is constructed in this paper. Finally, the empirical study shows this model is credible and feasible.

Key words projection pursuit classification, dynamic cluster, projection index, genetic algorithm

20世纪 60年代末 70年代初出现的投影寻踪 (Projection Pursuit, 简称 PP)模型[1,2], 是通过数值优化计算将高维数据投影到低维空间, 从而找到反映数据结构特征的最优投影的一种多元数据处理方法. 此模型的主要特点可概括为以下几点[3]:

(1) PP模型是少数几种能克服“维数祸根”的多元数据分析方法之一;

(2) PP模型能忽略与数据结构或特征不相关变量的影响;

(3) PP模型是把一维统计技术应用到多维数据最有力的方法, 对数据本身无特别的要求.

其中的投影寻踪聚类模型是一种数据聚类处理技术, 用于多因素影响问题的综合评价, 目前此模型的主要应用分为投影寻踪分类 (Projection Pursuit Classification)模型和投影寻踪等级评价 (Projection Pursuit Grade Evaluation)模型. 在运用投影寻踪等级评价模型时, 被评价问题事先必须具备等级评价的标准. 除了资源、工程和环境等问题外, 许多问题并不具备这样的评价标准. 投影寻踪分类模型不受评价标准的限制, 因而应用范围更广. 虽然投影寻踪分类模型在许多方面被成功应用[4-7], 但模型本身还存在着有待解决的问题, 主要体现在:

(1) Friedman和 Tukey在文献[2]中用标准差和局部密度来构造投影指标, 其中密度窗宽是需要确定的参数, 不同的密度窗宽得到不同的投影方向, 因此, 这个参数的取值在模型中非常关键, 它的取值合理与

收稿日期: 2009-06-16
基金项目: 国家自然科学基金 (60875001)资助项目.
通讯联系人: 姚 奕, 博士生, 讲师, 研究方向: 管理科学与工程. E-mail: yaoyi@njnu.edu.cn

否直接影响到分类结果的合理性. 以往大多是通过试算或经验值来确定, 例如文献 [2]中提出采用投影样本方差的 10%; 文献[6]采用了一个统计经验值, 这些值尚缺乏理论依据, 影响到模型的推广应用. 事实上, 不管采用哪种投影指标, 其实质都是要度量偏离正态分布的程度, 而仿射不变的指标作为投影指标可以度量正态性[8], 因此投影指标应该满足仿射不变性, 即满足任意实数 α, β 且 α ≠ 0, 对于投影 z有 $Q(\alpha z + \beta) = Q(z)$. 而Huber在文献[3]中指出用标准差和局部密度构造的投影指标不是仿射不变的. 文献[9]于 2007年提出了投影寻踪动态聚类模型, 首次将动态聚类方法引入投影寻踪模型, 并构造了一个不包含密度窗宽的投影指标, 为投影寻踪分类模型的研究提供了一个极好的思路, 开辟了一个新的途径. 该文中构造了投影指标 $Q(a) = ss(a) - dd(a)$, 其中 $ss(a) = \sum_{z_i, z_j \in \Omega} d(z_i, z_j)$ 表示所有点之间的距离和, $dd(a) = \sum_{h=1}^{p} \sum_{z_i, z_j \in \Theta_h} d(z_i, z_j)$ 表示所有属于同一个类的点的距离和, 由 $dd(a)$ 与 $ss(a)$ 的意义可知, $dd(a)$ 是 $ss(a)$ 的一部分, 因此 $Q(a)$ 反映了不同类的点的距离和, 该指标值越大则说明不同类之间越分散. 但是, 由于 $ss(a)$ 也随投影方向 a变化而变化, 因而同一类内点的距离之间的关系并不能明确反映.

(2) 由投影寻踪分类模型可以得到基于投影特征值大小的样本排序, 却没有严格的分类标准[8], 分类时通常需要研究者根据特征值的散布情况或凭借经验做出分类判断, 得到的分类结果往往带有主观性.

针对上述问题, 本文提出一个基于 K-Means动态聚类的投影寻踪分类模型, 定义了一个新的投影指标, 该投影指标满足仿射不变性, 也符合同类点尽量集中、不同类点尽量分散的分类要求. 实际应用说明了该模型的可靠性和可操作性.

1　基于 K-Means动态聚类的投影寻踪分类模型的步骤

步骤 **1**　对样本评价指标进行归一化处理. 设各评价指标值的样本集为 $\{x^*_{ij} \mid i = 1, 2, \cdots, n; j = 1, 2, \cdots, p\}$, 其中 x^*_{ij} 为第 i个样本的第 j个指标值, n, p 分别为样本容量和指标数量. 归一化处理可以消除指标量纲并统一各评价指标值的变化范围.

对越大越优的指标, 归一化公式为

$$x_{ij} = \frac{x^*_{ij} - \min_i x^*_{ij}}{\max_i x^*_{ij} - \min_i x^*_{ij}}, \tag{1}$$

对越小越优的指标, 归一化公式为

$$x_{ij} = \frac{\max_i x^*_{ij} - x^*_{ij}}{\max_i x^*_{ij} - \min_i x^*_{ij}}, \tag{2}$$

其中 $\max_i x^*_{ij}$, $\min_i x^*_{ij}$ 分别为第 j个评价指标值的最大值和最小值, $\{x_{ij} \mid i = 1, 2, \cdots, n, j = 1, 2, \cdots, p\}$ 为评价指标值归一化后的序列.

步骤 **2**　线性投影. 投影寻踪方法就是把 p维数据 $\{x_{ij} \mid j = 1, 2, \cdots, p\}$ 线性投影为以 $a = \{a_1, a_2, \cdots, a_p\}$ 为投影方向的一维投影值 z_i, $i = 1, 2, \cdots, n$,

$$z_i = \sum_{j=1}^{p} a_j x_{ij}, \ i = 1, 2, \cdots, n. \tag{3}$$

步骤 **3**　K-Means动态聚类. 把 n个样本点 z_i, $i = 1, 2, \cdots, n$分成 K类. 聚类的方法如下:

以每一类的均值为聚类中心, $\bar{z}_i$是第 i类的聚类中心, n_i表示类 P_i中的点的数目, $\sum_{i=1}^{K} n_i = n$

(1) 随机选取 K个点作为 K个聚类中心, 记为 $L^0 = \{\bar{z}_1^0, \bar{z}_2^0, \cdots, \bar{z}_K^0\}$, 根据 L^0, 把所有的点分为 K类, 记为 $P^0 = \{P_1^0, P_2^0, \cdots, P_K^0\}$, 其中 $P_i^0 = \{z \mid d(z, \bar{z}_i^0) \leqslant d(z, \bar{z}_j^0), j = 1, 2, \cdots, K, j \neq i\}$, 分类完成后将每一类中的点记为 z_{ij}^0, $i = 1, 2, \cdots, K$, $j = 1, 2, \cdots, n_i$.

(2) 由 P^0出发, 计算新的聚类中心 L^1, $L^1 = \{\bar{z}_1^1, \bar{z}_2^1, \cdots, \bar{z}_K^1\}$, 其中 $\bar{z}_i^1 = \frac{1}{n_i} \sum_{z_{ij}^0 \in P_i^0} z_{ij}^0$, 根据 L^1, 记为 $P^1 = \{P_1^1, P_2^1, \cdots, P_K^1\}$, 其中 $P_i^1 = \{z \mid d(z, \bar{z}_i^1) \leqslant d(z, \bar{z}_j^1), j = 1, 2, \cdots, K, j \neq i\}$, 分类完成后将每一类中的点记

为 z_{ij}^{1}, $i = 1, 2, \cdots, K$, $j = 1, 2, \cdots, n_i$; 循环.

(3) 定义: $u_m(K) = \sum_{i=1}^{K}\sum_{z_{ij}^{m}\in P_i^m} d(z_{ij}^{m}, \bar{z}_i^{m})$, 当 $\frac{|u_{m+1}(K) - u_m(K)|}{u_{m+1}(K)} \leqslant \varepsilon$, ε是一个充分小的允许误差, 则算法终止.

步骤 **4**　构造投影指标函数 $Q(z)$①.

投影指标应该满足仿射不变性. 除此以外, 确定投影指标时, 要求投影值 z_i, $i = 1, 2, \cdots, n$ 应具有如下散布特征: (1) 局部投影点尽可能密集, 最好能凝聚成若干个点团; (2) 整体上投影点团之间尽可能散开.

根据上述对投影指标的要求, 本文提出的投影指标为

$$Q(z) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{z}_i - \bar{z})^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z})^2} = \frac{\sum_{i=1}^{K} n_i(\bar{z}_i - \bar{z})^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z})^2}, \tag{4}$$

其中 $i = 1, 2, \cdots, K$, $j = 1, 2, \cdots, n_i$, $\bar{z}_i = \frac{1}{n_i}\sum_{z_{ij}\in P_i} z_{ij}$, $\bar{z} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} z_{ij}$, $\sum_{i=1}^{K} n_i = n$.

投影指标 $Q(z)$ 满足下列性质:

(1) $Q(z)$ 满足仿射不变性, 即任意实数 α, β, 且 $\alpha \neq 0$, 对于投影 z有 $Q(\alpha z + \beta) = Q(z)$.

这个性质显然成立.

(2) 投影指标 $Q(z)$ 满足不同类尽量分散, 同一类的点尽量集中的要求.

在投影指标 $Q(z)$ 中, 记 $SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z})^2$, 它反映了所有点与均值 $\bar{z}$的总距离, 这个量可以进行如下分解:

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{z}_i - \bar{z})^2 =$$
$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2 + \sum_{i=1}^{K} n_i(\bar{z}_i - \bar{z})^2 = SSE + SSA.$$

$SSE = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2$ 度量了各类内的点 z_{ij} 与该类聚类中心 $\bar{z}_i$ 之间的距离, 反映了同一个类内点之间的离散程度, 这个值越小, 说明同一个类内的点越集中; $SSA = \sum_{i=1}^{K} n_i(\bar{z}_i - \bar{z})^2$ 度量 K 个聚类中心与均值 $\bar{z}$之间的距离, 反映了不同类之间的离散程度, 显然这个值越大, 说明不同类越分散.

根据上面的记号, 投影指标可以记为 $Q(z) = \frac{SSA}{SSE + SSA} = \frac{SSA}{SST}$, 由此看出, $Q(z)$越大, 分类效果越好.

步骤 **5**　优化投影指标函数. 当各指标值的样本集给定时, 投影指标函数 $Q(z)$ 只随投影方向 a变化. 考虑下列优化问题

$$\max Q(z) = \frac{SSA}{SST},$$
$$s.t. \sum_{j=1}^{p} a_j^2 = 1 \tag{5}$$

这是一个 p 维变量 a 的非线性优化问题, 本文利用实数编码的加速遗传算法[10]来求解问题(5).

2　应用实例

为了便于比较, 本文利用文献[9]的实验数据, 取南京站 10个洪水样本作为研究实例, 分类指标包括: 洪峰水位、洪水位超过 9m的天数、大通洪峰流量、5 ~ 9月洪量以及流量与历时综合指标, 这些指标的样本数据见表 1, 按照本文的模型步骤, 计算出最优投影方向 $a^* = (0.128\,8, 0.848\,7, 0.267\,5, 0.146\,2,$

① 投影指标中投影方向是所求的变量, 为了下面性质表述的便利, 这里的表示方法与 Huber的表示方法一致.

0.413 7), 将最优投影方向代入 (3) 式, 得到各样本的投影值, 如表 1. 其中投影值越大, 表明洪水强度越大.

表 1　南京站洪水样本的投影值及分类结果

Table 1　Projection values and classifications of flood samples in Nanjing Station

年份	洪峰水位 /m	洪水位超过9m的天数 /d	大通洪峰流量 / (m^3 /s)	5~9月洪量 / ($10^8 m^3$ /s)	流量与历时综合指标	投影值	分类结果
1954	10.22	87	92 600	8 891	7 800	1.804 9	第一类
1969	9.2	8	67 700	5 447	1 710	0.077 53	第三类
1973	9.19	7	70 000	6 623	3 280	0.239 3	第三类
1980	9.2	10	64 000	6 340	2 730	0.168 5	第三类
1983	9.99	27	72 600	6 641	3 560	0.583 8	第二类
1991	9.7	17	63 800	5 576	1 930	0.212 2	第三类
1992	9.06	13	67 700	5 295	1 575	0.099 9	第三类
1995	9.66	23	75 500	6 162	2 390	0.434 4	第二类
1996	9.89	34	75 100	6 206	2 702	0.595 5	第二类
1998	10.14	81	82 100	7 773	5 283	1.422 1	第一类

根据表 1中投影值的大小, 洪水强度按从大到小排列为: 1954年、1998年、1996年、1983年、1995年、1973年、1991年、1980年、1992年、1969年, 与文献 [9]不同的是, 本文结果中 1991年的洪水强度大于1980年的洪水强度, 与《中国水灾年表》的实际统计数据符合.

根据投影值的大小排序作出相应的雷达图, 本文投影结果的雷达图如图 1所示, 由文献 [9]的投影值得到的雷达图如图 2所示. 比较图 1和图 2可以发现, 本文的模型在分类效果上使得同类数据更加集中, 不同类数据更加分散.

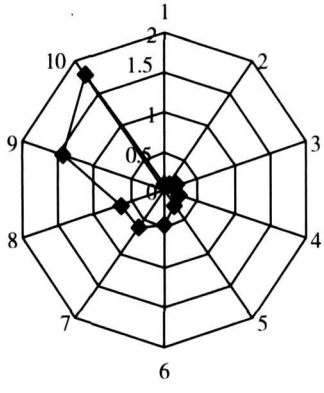

图 1　本文模型投影值的雷达图

Fig.1　Radar images of projection values of model proposed by this paper

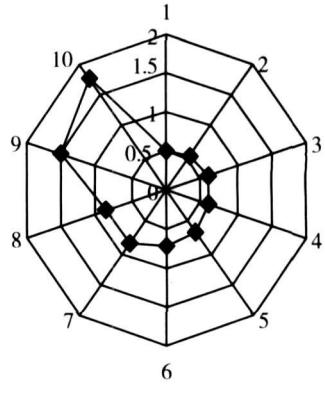

图 2　文献[10]投影值的雷达图

Fig.2　Radar images of projection values of literature [10]

本文模型计算得到的分类结果列在表 1中. 其中 1954年和 1998年属于第一类, 判定为特大洪水; 1983年、1995年和 1996年属于第二类, 判定为大洪水; 1969年、1973年、1980年、1991年和 1992年属于第三类, 判定为中等洪水, 样本分类的结果与文献 [9]的结果是一致的, 与实际情况也是吻合的.

上述实例分析表明本文取得了很好的分类效果, 聚类结果更加清晰, 本文的模型算法有效且切实可行.

3　结语

本文提出了基于 *K*-Means动态聚类的投影寻踪分类模型, 该模型中提出的新的投影指标不仅具有仿射不变性, 而且同时满足同类投影点尽量集中和不同类投影点尽量分散的分类要求. 该指标克服了投影寻踪分类模型密度窗宽不确定的问题. 同时, *K*-Means动态聚类的特点也决定了分类结果由实际数据客观确定, 无需人为判断, 因此模型便于操作和推广.

实例分析的结果说明该模型的应用效果令人满意, 期望能为分类评价问题提供一个新的思路.

[参考文献]

[1] Kruscal J B. Linear transformation of multivariate data to reveal clustering[M]// Multidimensional scaling: Theory and application in the behavioral sciences(I) theory. New York and London: Seminar Press, 1972.
[2] Friedman J H, Turkey J W. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Trans on Computer, 1974, 23(9): 881-890.
[3] Huber P J. Projection pursuit(invited paper)[J]. The Annals of Statistics, 1985, 13(2): 435-475.
[4] 金菊良,张欣莉,丁晶.评估洪水灾情的投影寻踪模型[J].系统工程理论与实践, 2002, 22(2): 140-144.
[5] 金菊良.投影寻踪模型在水资源工程方案优选中的应用[J].系统工程理论方法应用, 2004, 13(1): 81-84.
[6] 张欣莉,任仕泉,罗利.企业竞争力评价的投影寻踪模型[J].数理统计与管理, 2005, 25(4): 53-55.
[7] Liu Bao, Shen Zhenkang, Sun Zhougkang. A pattern recognition method using projection pursuit[C]// IEEE Aerospace and Electronics Conf, 1990: 300-302.
[8] 付强,赵小勇.投影寻踪模型原理及其应用[M].北京:科学出版社, 2006.
[9] 倪长健,崔朋.投影寻踪动态聚类模型[J].系统工程学报, 2007, 22(6): 634-638.
[10] 杨晓华, 陆桂华,郦建强.混合加速遗传算法在流域模型参数优化中的应用[J].水科学进展, 2002, 13(3): 340-344.

[责任编辑: 丁　蓉]

(上接第 15页)

$\circ R$是$(\mathscr{C})$收敛于r的. 进而, 如果$p \geqslant m$, 则$U \circ R(p, f) = U(p, f(p)) \in B_m$; 即$U \circ R(p, f) \geqslant m$. 这可以直接推出$T \circ U \circ R$是$T$的一个子网, 从而定理成立.

推论 **2**　设$\mathscr{C}$是$(X, \&)$中的一个收敛类. 由$\mathscr{C}$确定的X的s-量子拓扑是X上的拓扑.

[参考文献]

[1] Susan B Niefield, Kimmo I Rosenthal. Constructing locales from quantales[J]. Math Proc Camo Phil Soc, 1988, 104(2): 215-233.
[2] Borceux F, Bossche G Van Den. Quantales and their sheaves[J]. Order, 1986, 3(1): 61-87.
[3] Borceux F, Bossche G Van Den. An essay on non-commutative topology[J]. Topology and its Applications, 1989, 31(2): 203-223.
[4] Johnstone P T. Stone Spaces[M]. Cambridge: Cambridge University Press, 1982.
[5] John L Kelly. General Topology[M]. New York: Springer-Verlag, 1955.

[责任编辑: 丁　蓉]