基于统计检验指导的聚类分析方法 [1]

高新波 裴继红 谢维信 *

(西安电子科技大学电子工程学院 西安 710071)
*(深圳大学 深圳 518060)

摘 要 本文提出了一种基于统计检验指导的聚类分析方法，该方法同时处理聚类趋势、聚类分析和聚类有效性三个数据分析中的关键问题，为模式无监督分类的合理性和有效性提供了分析工具．在大样本情况下更能体现该方法的优越性．实验结果证明了它的有效性．

关键词 聚类趋势，聚类分析，聚类有效性，统计检验
中图号 TP391.4

1 引 言

模式聚类是多元统计分析方法之一，也是非监督模式识别的一个重要分支 [1]．其目的是根据数据集的内在结构将一个未标记的模式样本集划分成不同的子集，使相似的样本归为一类，而不相似的样本分在不同类中．

聚类算法的性能与数据集密切相关，没有万能的聚类算法，这也是新的聚类算法层出不穷的原因．现有聚类算法最主要的缺陷是 [2]：不管数据集的结构如何，给定一个有效的聚类数 c，总能将数据集分成 c 类，而不考虑数据集是否具有可分性，也不管分成 c 类是否合理．实际上，要对模式集有效的分析需要同时考虑以下三个问题：(1) 模式集是否有聚类结构，称为聚类趋势问题 [3]；(2) 如果有聚类结构，如何确定这个结构，称为聚类分析问题 [2]；(3) 一旦模式集被聚类，如何确定聚类数 c 的合理性，称为聚类有效性问题 [4]．为此，本文提出一种同时兼顾这三个问题的聚类方法．该方法扩展了 Besag 和 Gleaves 提出的 T-平方空间抽样原理 [5]，结合了基于目标函数的聚类算法，是一种统计检验指导的模式分析方法．由于本方法一旦满足显著性检验就停止，只需做 c 次聚类分析，而且统计检验的复杂度不随样本集的增大而剧烈变化，因此实现简单、运算量较小，尤其适合大数据量样本集的分析处理．

以下，第 2 节引入单峰分布模式的统计检验，第 3 节介绍基于目标函数的聚类算法，第 4 节为本文提出的模式分析自动机，第 5 节给出实验结果和讨论，最后是结论．

2 单峰分布模式的统计检验

2.1 问题表述

设模式集 $X=\{x_1,x_2,\cdots,x_n\}\subset R^p$ 是 p 维实数空间 R^p 中的一个未标记的子集，$x_k=(x_{k1},x_{k2},\cdots,x_{kp})\in R^p$ 称为特征矢量或模式矢量，x_{kj} 为矢量 x_k 的第 j 个特征．那么，按照模式产生的统计模型，概率密度 $p(x)$ 将是可以从集合 X 中导出的唯一特性，当 X 为多类模式的集合时，$p(x)$ 为多峰的；当 X 为单一类模式的集合时，$p(x)$ 为单峰的．而各种聚类算法都可以归结为 $p(x)$ 峰态的间接分解，即

$$p(x)=\sum_{i=1}^{c}f(x|S_i), \tag{1}$$

[1] 1998-06-04 收到，1999-01-23 定稿
国家自然科学基金资助课题

其中 c 为给定的聚类数，$f(x|S_i)$ 是从第 i 类中导出的密度函数.

就聚类分析的目的而言，X 的类组成性质事前并不为人所知，而任何一种聚类算法被应用于 X 时，不论 X 为多类还是单一类模式的集合，都毫无例外要产生密度函数的再分解，这显然是不合适的. 如何避免聚类算法对单一类模式集合的不适当应用，是聚类趋势研究的主要课题，即检验集合 X 是否已经为单峰分布的；而聚类数的有效性则对应于 $p(x)$ 的峰态分解是否完全，即检验每类模式子集 $S_i(i=1,2,\cdots,c)$ 的密度函数是否都是单峰函数. 这样就可以把聚类趋势和聚类有效性两个问题全部归结为模式 (子) 集在特征空间的单峰检验.

本文选用统计检验方法作为判定手段，根据经典的统计检验理论把聚类趋势和聚类有效性两个问题归结为下述零假设 H_0 和备选假设 H_α 的检验：H_0：给定模式 (子) 集在特征空间中呈单峰分布；H_α：给定模式子集在特征空间中仍具有可分性.

2.2 数学模型

假设用某种聚类算法把 X 硬划分成 c 个模式子集 $S_i(i=1,2,\cdots,c)$，则有

$$X=\bigcup_{i=1}^{c} S_i,\quad S_i\cap S_j=\phi,\quad i\neq j,\quad 1<i,j<c. \tag{2}$$

对于好的特征集和合理的分类结果而言，每个子集 S_i 在空间中必定致密，不同的子集 S_i 与 S_j 之间必定疏远，也就是说，模式间的距离中蕴涵了研究其空间分布的丰富信息.

Besag 和 Gleaves 为了研究二维空间中的模式分布问题，提出了图 1 所示的 T-平方空间抽样原理 [5]，导出在空间泊松过程条件下服从正态分布 $N(0.5,M/12)$ 的统计量 T_b，即

$$T_b=\frac{1}{M}\sum_{i=1}^{M}\frac{U_i^2}{U_i^2+V_i^2/2}, \tag{3}$$

其中 M 为抽样始点数. 按照图 1，首先在抽样域内随机设置一个与模式相区别的抽样始点 O，并测量它与最近邻模式 P_1 之间的距离 U，然后在与 OP_1 垂直的直线 AB 的另一侧寻找与 P_1 最近邻的模式 P_2，即满足条件 $\cos\angle OP_1P_2<0$，记 P_1 与 P_2 间的距离为 V. 假设独立地在抽样域内设置了 M 个抽样始点，便可以做出统计量 (3) 式. 分析表明，当 $M\geq 10$ 时，统计量 T_b 良好地服从正态分布，即 $T_b\sim N(0.5,M/12)$.

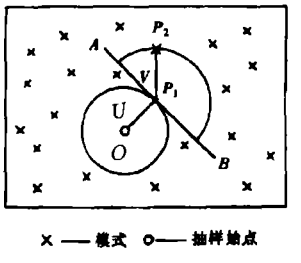

图 1　二维空间中 T-抽样原理
×: 模式,　　o: 抽样始点

曾广周把二维 T-平方抽样原理扩展到 $p(p\geq 3)$ 维情况 [6]，把统计量 T_b 推广为 T_z，在不改变空间泊松过程的条件下 T_z 仍服从正态分布 $N(0.5,M/12)$，并证明统计量 T_z 在半数框架 (即抽样始点在以模式质心为中心，包含一半模式的圆内选取) 的制约条件下，对特

征空间中均匀分布的模式能建立可信的检验基础，对单一高斯分布模式也具有极低的功效检验．因此，完全可以用来进行单峰模式的统计检验．

由于最近邻模式选取的随机性太强，而统计性较差，为了抽取鲁棒性较好的统计量，我们把最近邻模式间的距离替换为 k-近邻模式间的距离，把寻找最近邻模式转化为寻找 k 个近邻模式集，大大降低了模式出现的随机性，从而使所得到的统计量更稳定．因为统计量 T_b 和 T_z 分别表示了关于模式 P_1 和 P_2 搜索的面积或体积之比，这样便可把统计量 T_z 推广为 k-近邻的统计量 T_k ，

$$T_k = \frac{1}{M}\sum_{i=1}^{M}\frac{U_i^p(k)}{U_i^p(k)+V_i^p(k)/2}, \tag{4}$$

其中 $U_i(k)$ 为抽样始点到它的第 k-近邻模式 P_1 的距离，而 $V_i(k)$ 则为模式 P_1 与其第 k-近邻模式 P_2 间的距离，注意在 p 维空间中，半径为 r 的球体体积为

$$V = \frac{r^p\pi^{p/2}}{\Gamma(p/2+1)}. \tag{5}$$

按照大数定理，在空间泊松过程的条件下统计量 T_k 仍然服从 $N(0.5, M/12)$ 为参数的正态分布．当 $k=1$ 时， T_k 就退化为 T_z ，当 $p=2$ 时，又进一步退化为 T_b ．本文将利用统计量 T_k 在半数框架的制约下进行单峰模式的统计检验，进而指导聚类分析的实施．

3 基于目标函数的聚类算法

聚类分析作为模式识别的一个重要分支，一直是一个研究热点．至今已提出了若干种聚类方法，如谱系法、图论法以及基于目标函数的算法等等．随着计算机的发展和实际问题的需要，基于目标函数的方法已成为聚类分析的主流．这一方面是由于将聚类问题表述为优化问题易于与经典数学的非线性规划领域联系起来，可用现代数学方法求解．另一方面是由于算法的求解过程比较容易用计算机实现．

$J_1(U,v)$ 是最早提出的类内均方误差和目标函数，先后被用来定义硬 c-均值 (HCM) 和 ISODATA 算法．后来， Dunn 引入了模糊集理论，把 $J_1(U,v)$ 推广到 $J_2(U,v)$ ，并提出了模糊 c-均值 (FCM) 算法 [2] ． Bezdek 又把 $J_2(U,v)$ 推广到了一个目标函数的无限簇 [2] ，定义了如下的目标函数：

$$J_m(U,v) = \sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|x_k - v_i\|^2, \quad 1<m<\infty, \tag{6}$$

其中 U 为模糊划分矩阵， u_{ij} 表示 x_j 属于第 i 类 S_i 的隶属度； $v=\{v_1, v_2, \cdots, v_c\}$ 为各类模式的聚类中心； $\|x_j - v_i\|$ 为某种范数，表示 x_j 与 v_i 间的距离； m 为模糊加权指数， m 越大则模式划分的模糊度越大，而 $m \to 1$ 时则退化为硬聚类算法， m 的经验值为区间 $[1.5, 5]$ ．对 (6) 式目标函数的极小化所得到的模式划分即为最终的分类结果．

围绕目标函数的优化问题，目前主要形成了三大研究方向 [7] ：一是基于 Picard 迭代的爬山法或梯度下降法，如 HCM 和 FCM 等算法；二是借助神经网络的并行实现方法，以提高算法的收敛速度，最著名的如 Kohonen 的自组织特征映射网络和 Grossberg 的自适应共振网络；三是结合进化计算的优化方法，以获得全局最优解，如模拟退火算法、遗传算法、进化策略及 Tabu 搜索等方法．

但是，上述算法均要求聚类数 c 事先给定，而且一旦 c 给定，就会以类内均方误差和最小准则，对样本集进行最优划分，获得 c 个子集，算法不管数据集 X 是否有类分性，也不管分成 c 类是否合适．因此，这些算法都是一种条件最佳分类，即在聚类数给定的条件下的最佳分类，而人们所需要的往往是不带任何条件的最佳分类．为此，我们用统计检验与基于目标函数的聚类算法相结合构造一种合理有效的模式分析方法．

4 基于统计检验的聚类自动机

聚类分析是无监督的类分技术，事先没有任何有关数据集的先验知识，包括它的分布和类别数，因此应该是一种机器自动学习方法．通过对数据分布特性的分析，自动给出模式集是否具有类分性，如果可分则自动确定如何分类以及分为几类合适等问题．针对上述要求，我们构造了一个模式分析的自动机，如图 2 所示．

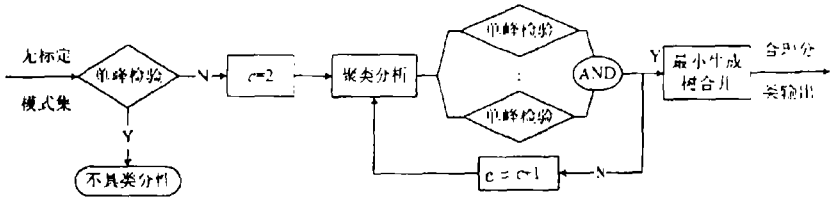

图 2　模式分析自动机

对于一个给定的无标定模式集，图 2 所示的模式分析自动机首先进行单峰分布的统计检验，以确定该模式集是否具有类分性，即聚类趋势检验．其中用到两个算法，一是建立半数框架，二是对统计量 T_k 进行 α-显著性检验．两个算法分别给定如下：

算法 1　建立半数框架

(1) 计算模式集合 X 的均值向量 m，即 $m=(\sum_{i=1}^{n}x_i)/n$；

(2) 计算模式样本 x_i 与均值向量 m 间的距离 $d_i=\|x_i-m\|$，并由小到大排列 $d_1'd_2'\cdots d_n'$，取 $r=d'_{\lfloor n/2\rfloor}$，$\lfloor\ \rfloor$ 为取整符号；

(3) 球 (m,r) 包含的区域即为抽样始点设置域，k-近邻模式 P_1,P_2 允许在球外寻找．

算法 2　α-显著性检验

(1) 在给定模式集 X 中建立半数框架 (m,r)，令 $s=0$;

(2) 在球 (m,r) 中随机设置 M 个抽样始点，并计算标准化的统计量 T_k：

$$T_k=\left[\frac{1}{M}\sum_{i=1}^{M}\frac{U_i^p(k)}{U_i^p(k)+V_i^p(k)/2}-0.5\right]\times\sqrt{12M};$$

(3) 如果 $T_k\geq T(\alpha)$，则 $s=s+1$，其中 α 为选定的显著水平，$T(\alpha)$ 为相应的临界值，可由标准正态分布表中查出，例如 $T(0.05)=1.64485$；

(4) 重复 (2), (3) 步 N 次 (一般取 $N=100$)，计算检验大小 $\bar{s}=s/N$，如果 $\bar{s}>>\alpha$，则认为 X 为多峰分布模式，具有可分性，反之如果 $\bar{s}$ 与 α 相比拟，则认为 X 为单峰模式．

在获知模式集 X 具有类分性后，首先令 $c=2$，用现有的聚类算法获得 X 的 c-划分．这里要求算法要收敛到最优解，否则将影响后续工作．而基于目标函数的聚类算法对初始点敏感，容易陷入局部极值点而得不到最优的分类结果．为此，人们已经提出了多种获得全局最优解的聚类方法，其中有借助数学形态学 [8] 或者势函数 [9] 的初始化方法，有的利用进化计算 [7] 进行全局优化等等．这里可根据实际需要选择合适聚类分析方法．比如可以

选用基于遗传算法的聚类分析算法，首先对聚类中心矢量 $v=\{v_1,v_2,\cdots,v_c\}$ 分别编码链接成基因码链，然后利用 (6) 式所给出的聚类目标函数来构造遗传聚类算法的适应度函数 $f(v)=1/(J_1+C)$，其中 C 为一给定常数．划分矩阵 U 仍按 HCM 算法中的更新方法。这样，就可以利用常规的遗传算法进行优化求解了，具体方法可参见文献 [7]．

得到模式集 X 在给定聚类数 c 条件下的最佳模式划分后，对 c 个子集 $S_1,S_2,\cdots,S_c$ 分别做单峰分布的 α-显著性检验，只要还有一个子集不满足显著性检验，则说明仍存在可分性，令 $c=c+1$ 重新聚类，直到所有 c 个子集均不具有类可分性后，说明 $p(x)$ 的峰态已经分解完全，每个聚类均为单峰分布模式了，则转入聚类后处理。

后处理主要是考虑到聚类的无监督性，为了满足不同的应用场合，为后续决策提供更多的信息，利用最小生成树技术进行聚类合并 [10]，然后给出一个不同水平上类分的谱系图，也就是说给出特征空间中模式矢量的多分辨表示。当然最细微层次上的聚类即为上面得到的分类结果，最粗略的层次上的聚类为整个模式集 X 聚合为一类。

算法 3 建立类分谱系图

(1) 在聚类分析中获得的 c 个聚类中心 $v=\{v_1,v_2,\cdots,v_c\}$ 上构造最小生成树，其边的权值即为节点间的欧氏距离；

(2) 统计最小生成树的权值，逐步在不同水平上删除权值较小的边，合并该边连接的两个类，构造新的最小生成树，直至合并成一个类，最后输出类分谱系图。

按照上述步骤，模式分析自动机即可对任意一种输入模式集，产生一个分类的谱系图，实现简单，过程自动，使聚类分析真正成为机器自动学习算法。

5 实验结果和讨论

大量的实验结果表明本文方法能够合理有效地分析所给的模式样本集，并为后续工作提供不同层次的类分信息，使人们更清楚地认识待分析模式间的亲疏关系、分布形态和内在结构，整个过程自动执行，无须人为干预，为生产自动化提供了条件。由于篇幅所限，这里只选择一个可视化的例子来说明该方法的有效性。

图 3 所示为一组人工产生的 300 个模式的样本集。本文方法对这组数据的分析表明：该模式集具有聚类趋势，因为其单峰 α-显著性检验的大小远大于 α (实验中取 $\alpha=0.05$)．于是，算法从 $c=2$ 开始进行聚类分析，并对所得到的每个聚类分别进行单峰检验，不满足检验则令 $c=c+1$ 重新聚类，直到 $c=6$ 时，所得到的子集才全部通过 α-显著性检验，分类结果如图 4 所示，此时各个聚类皆为单峰分布的子集。

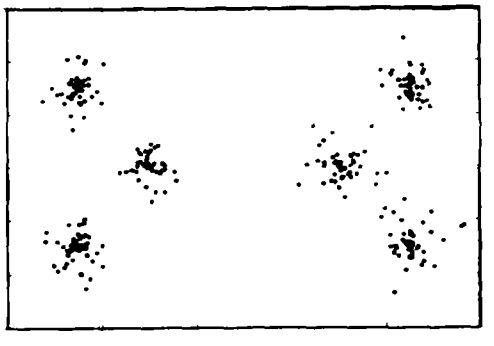

图 3 待分析的模式样本集

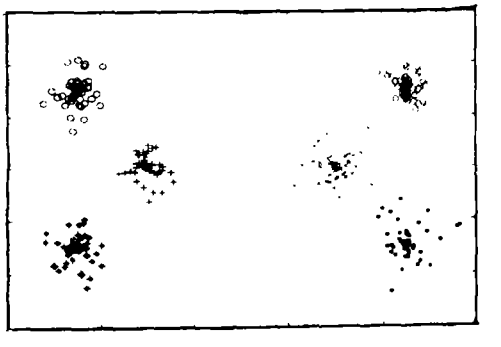

图 4 峰态分解完全后所得到分类结果

表 1 给出了取不同 c 值时，各个聚类进行 T_k 统计检验中得到的第一类错误的最大值，即为最大检验大小 $\bar{s}$，其中 $M=10$，k 取 1 到 5．不同 k 值的 T_k 统计检验结果均表明：直到 c 取 6 时各类才均满足 α-显著性检验，而对应于不同 k 值的检验大小又反映出在 $k=1$

时，统计量 T_z 的检验大小不很稳定，$k=4$ 时则较为理想．因此选择适当的 k 值，将更有利于模式分析的效果．

表 1　在不同聚类数 c 和 k 值条件下各类样本统计检验中最大的第一类错误 (%)

k	聚类数 c					
	1	2	3	4	5	6
1	99	97	47	94	89	9
2	100	65	59	100	99	5
3	100	81	81	99	100	9
4	100	100	100	100	100	0
5	100	96	100	100	100	2

得到样本集的 c-划分后，模式分析自动机进行聚类的合并，以输出样本类分的谱系图．首先以所得到的 6 个聚类中心为节点，构造图 5 所示的最小生成树，其中边权值为节点间的距离，节点上的数值 1-6 表示聚类的序号，利用算法 3，逐次删除权值较小的边并进行聚类合并，得到模式类分谱系图，如图 6 所示．可以看出如果从较粗略的层次上分类，模式集可以分为两大类，而从精细的角度则需要分为 6 类，如此样本集的峰态才能分解完全．因此，根据不同的应用背景可选取不同层次上的分类，方便了用户，而且提供了数据类分的多分辨表示，使用户的推理、决策等操作更具有可靠性．

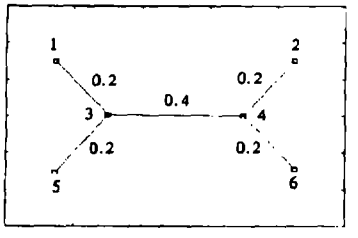

图 5　在聚类中心上构造的最小生成树

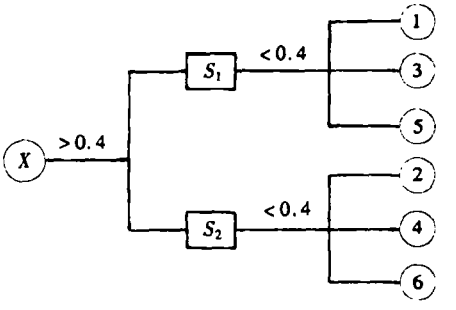

图 6　模式分析自动机输出的类分谱系图

按照上述步骤，对于给定的模式样本集，本文提出的模式分析自动机能够自动、合理而有效地给出特征样本集在不同层次上的类分结果，并且给出不同模式子集之间的亲疏关系，充分显示了该方法的有效性．

6 结　束　语

为了对数据集进行综合分析，同时处理非监督模式分析中聚类趋势、聚类分析和聚类有效性等三个关键问题，本文提出基于统计检验指导的聚类分析方法．该方法自动检验样本集的聚类趋势，如果不存在聚类结构，则停止对样本集的聚类，以避免算法应用不适当所造成的评价聚类结果有效性的困难；反之则利用现有聚类算法在不同聚类数下分别进行聚类分析，直到所得到的每个聚类的模式分布满足单峰统计检验，最后利用最小生成树技术进行聚类的合并，得到模式集的类分谱系图．

为了辨识模式在特征空间中的单峰分布，曾光周 [6] 把 Besag 和 Gleaves 提出的统计量 T_b 推广到 p 维，我们又把它推广到 k-近邻距离的统计量 T_k，在半数框架的制约下，T_k-统计量对均匀分布的模式能够建立可信的检验基础，而对单一高斯模式有极低的功效检验，从而可以用来检验单峰分布的模式集．利用此性质，可以指导模式集分布的峰态分解，直到把模式集分解成单峰分布的模式子集．

最小生成树技术则方便了模式子集的合并，产生模式集的类分谱系图，从而为用户提供更为综合的信息。总之，本文提出的模式分析自动机能够自动、合理地分析模式，为促进聚类分析真正成为智能的机器学习方法提供了新思路。

参 考 文 献

[1] Anderberg M R. Cluster Analysis for Applications. New York: Academic Press, 1973, chapter 1.

[2] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. NewYork: Plenum Press, 1981.

[3] Panayirci E, Dubes R C. A test for multidimensional clustering tendency. Pattern Recognition, 1983, 16(4): 433–444.

[4] Dubes R C, Jain A K. Validity studies in clustering methodologies. Pattern Recognition, 1979, 12(11): 235–247.

[5] Besag J E, Gleaves J T. On the detection of spatial pattern in plant communities. Bulletin of the International Statistical Institute, 1973, 45(1): 153–158.

[6] 曾广周. 随机分布模式的统计检验. 山东工业大学学报, 1986, 16(4): 12–18.

[7] 高新波. 基于进化计算和神经网络的模糊聚类新算法研究: [硕士论文]. 西安: 西安电子科技大学, 1996.

[8] Gao Xinbo, *et al.* An initialization method for multi-type prototype fuzzy clustering. Proceeding of International Conference on Signal Processing, ICSP'98, Beijing: 1998, 1205–1208.

[9] 裴继红. 基于模糊信息处理的图象分割方法研究: [博士论文]. 西安: 西安电子科技大学, 1998.

[10] Hoffman R, Jain A K. A test of randomness based on the minimal spanning tree. Pattern Recognition Lett., 1983, 4(1): 175–180.

A NOVEL CLUSTER ANALYSIS METHOD SUPERVISED BY STATISTICAL TESTS

Gao Xinbo Pei Jihong Xie Weixin*

(*School of E.E., Xidian University, Xi'an* 710071)

*(*Shenzhen University, Shenzhen* 518060)

Abstract A novel cluster analysis method supervised by statistical tests is proposed in this paper, which processes three key problems in data analysis, cluster tendency, cluster analysis and cluster validity, simultaneously. So, it provides an analysis tool for the validity and reasonableness of pattern unsupervised classification, especially in the case of large number of samples. The experimental results demonstrate its effectiveness.

Key words Cluster tendency, Cluster analysis, Cluster validity, Statistical tests

高新波： 男， 1972 年生，博士生，主要研究方向有模式识别、图象处理、模糊信息处理和人工智能.

裴继红： 男， 1966 年生，讲师，主要研究方向有信号处理、模式识别、图象处理和模糊信息处理.

谢维信： 男， 1941 年生，教授，博士生导师，主要研究方向有信号处理、模式识别和智能信息处理.