第 23 卷 第 4 期
2008 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.23 No. 4
Dec. 2008

随机平均欧氏距离的统计性质与分类阈值

张大克，王玉杰
（天津科技大学理学院，天津 300457）

摘 要：平均欧氏距离是系统聚类法中使用最普遍的一个距离指标，它是样品分类的重要依据. 为了能使分类者更科学地确定分类阈值，对随机平均欧氏距离的统计性质进行了研究，确定了它的概率分布，给出了它的数字特征，为分类阈值的确定提供了理论依据和方法.

关键词：系统聚类法；随机平均欧氏距离；概率分布；数字特征；分类阈值
中图分类号：O212.4 **文献标识码**：A **文章编号**：1672-6510（2008）04-0085-04

Statistical Properties of Random Mean Euclidean Distance and Measure of Categorical Threshold

ZHANG Da-ke，WANG Yu-jie
（College of Science，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Mean Euclidean distance is one of the most common distance index used in hierarchical classification method. It is an important basis for sample classification method. To make categorical threshold measure more scientific, statistical properties of random mean Euclidean distance were researched, its probability distribution was determined, its numerical characteristic was given. This research can provide the theoretical basis and the method for determining categorical threshold measure.

Keywords：hierarchical classification method；random Mean Euclidean distance；probability distribution；numerical characteristic；categorical threshold measure

在聚类分析的系统聚类法中，当对 m 个样品（分类的对象）进行分类时，不论采用什么样的样品间距离，采用什么样的聚类方法，其聚类过程都是首先将 m 个样品各自看成一类，然后根据所计算出来的距离以最小为原则进行并类，直至将所有的样品并为一类为止. 而 m 个样品究竟分几类比较科学，分类阈值如何确定，却是系统聚类法至今没能完全解决的问题[1—10]. 1977 年统计学家 Rao C R 曾给出关于类的三种定义[11].

定义 1 若集合 S 中任意两个元素 i 与 j 之间的距离 d_{ij} 满足

$$d_{ij} \leqslant T,\quad i,j \in S$$

T 为给定的阈值，则称 S 对于阈值 T 组成一个类.

定义 2 若集合 S 中任意两个元素间的距离 d_{ij} 满足

$$\frac{1}{k-1}\sum_{j\in S} d_{ij} \leqslant T,\quad i \in S$$

T 为给定的阈值，k 为集合 S 中元素的个数，则称 S 对于阈值 T 组成一个类.

定义 3 若集合 S 中元素间的距离 d_{ij} 对于给定的阈值 T、r（$r>T$），满足

$$\frac{1}{k(k-1)}\sum_{i\in S}\sum_{j\in S} d_{ij} \leqslant T,\quad d_{ij} \leqslant r,\quad i,j \in S$$

其中 k 为集合 S 中元素的个数，则称 S 对于阈值 T 和

收稿日期：2007-12-28；**修回日期**：2008-06-10
基金项目：天津科技大学引进人才基金资助项目（20030402）
作者简介：张大克（1961—），男，吉林人，教授.

r 组成一个类.

由定义 1、定义 2 及定义 3 可见，给定了分类阈值后，样品的分类就确定了. 而恰恰是这一分类阈值系统聚类法至今没能给出一个具体的确定方法. 到目前为止，在实际问题中，分类阈值的确定仍然靠专业经验，其科学性和可靠性无法保证. 而要科学地确定分类阈值，就必须对样品间距离的统计性质进行深入研究，为科学地确定分类阈值建立理论基础和方法. 本文对系统聚类法中使用最普遍的平均欧氏距离的统计性质进行了研究，确定了它的概率分布，给出了它的数字特征，为选用平均欧氏距离，采用最长距离法和最短距离法进行分类时科学地确定分类阈值建立了理论基础和方法.

1　随机平均欧氏距离

设观测数据可连续取值，观测数据经标准差标准化变换后的数据矩阵见表 1[12]. 样品 i 和样品 j 间的平均欧氏距离为

$$d_{ij}=[\frac{1}{n}\sum_{k=1}^{n}(x_{ik}-x_{jk})^2]^{\frac{1}{2}}$$
$$i\neq j;i,j=1,2,\cdots,m \qquad (1)$$

假设 x_{i1}, x_{i2}, …, x_{in} 是随机变量 $\xi_{i1},\xi_{i2},\cdots,\xi_{in}$ 的观测值，$i=1,2,\cdots,m$，则样品 i 和样品 j 间的平均欧氏距离（1）实质上是随机距离

$$D_{ij}=[\frac{1}{n}\sum_{k=1}^{n}(\xi_{ik}-\xi_{jk})^2]^{\frac{1}{2}}$$
$$i\neq j;i,j=1,2,\cdots,m \qquad (2)$$

的观测值，D_{ij} 称为样品 i 和样品 j 间的随机平均欧氏距离，$i,j=1,2,\cdots,m$.

表 1　经标准差标准化变换后的观测数据

Tab. 1　Observational data by standardized transformation

样品	指标					
	x_1	x_2	…	x_s	…	x_n
1	x_{11}	x_{12}	…	x_{1s}	…	x_{1n}
2	x_{21}	x_{22}	…	x_{2s}	…	x_{2n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{i1}	x_{i2}	…	x_{is}	…	x_{in}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	x_{m1}	x_{m2}	…	x_{ms}	…	x_{mn}

2　随机平均欧氏距离的统计性质

假设 n 个分类指标间相互独立，原始观测数据都是来自相互独立的正态总体，且具有相同的方差. 由于表 1 中数据是原始观测数据经标准差标准化变换后的数据，故可假设随机变量 $\xi_{i1},\xi_{i2},\cdots,\xi_{in}$ 相互独立，且都服从标准正态分布[12]，当 $i\neq j$ 时，$\xi_{i1},\xi_{i2},\cdots,\xi_{in}$ 与 $\xi_{j1},\xi_{j2},\cdots,\xi_{jn}$ 也相互独立，$i,j=1,2,\cdots,m$.

若设

$$X_{ij}=\frac{1}{n}\sum_{i=1}^{n}(\xi_{ik}-\xi_{jk})^2$$
$$i\neq j;i,j=1,2,\cdots,m \qquad (3)$$

则随机平均欧氏距离

$$D_{ij}=[\frac{1}{n}\sum_{k=1}^{n}(\xi_{ik}-\xi_{jk})^2]^{\frac{1}{2}}=X_{ij}^{\frac{1}{2}}$$
$$i\neq j;\ \mathrm{i},j=1,2,\cdots,m \qquad (4)$$

定理 1　令

$$W_k=\frac{1}{\sqrt{2}}(\xi_{ik}-\xi_{jk})$$
$$i\neq j;\ i,j=1,2,\cdots,m;\ k=1,2,\cdots,n \qquad (5)$$

则随机变量 $W_1,W_2,\cdots,W_n$ 相互独立，且

$$W_k\sim N(0,1)\quad k=1,2,\cdots,n \qquad (6)$$

定理 2　随机变量 $\sum_{k=1}^{n}W_k^2\sim\chi^2(n)$，其概率密度函数为

$$f_W(w)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}w^{\frac{n}{2}-1}e^{-\frac{w}{2}}, & w\geqslant 0\\ 0, & w<0\end{cases}$$

定理 1 和定理 2 的结论是显然的[13,14].

定理 3　随机变量 X_{ij} 的概率密度函数

$$f_X(x)=\begin{cases}\dfrac{n^{\frac{n}{2}}}{2^n\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{n}{4}x}, & x\geqslant 0\\ 0, & x<0\end{cases} \qquad (7)$$

即

$$X_{ij}\sim\Gamma(\frac{n}{4},\frac{n}{2})\quad i\neq j;\ i,j=1,2,\cdots,m \qquad (8)$$

证明：由式（5）知 $X_{ij}=\frac{2}{n}\sum_{k=1}^{n}W_k^2$，故随机变量 X_{ij} 的分布函数

$$F_X(x)=p\{X_{ij}\leqslant x\}=p\left\{\frac{2}{n}\sum_{k=1}^{n}W_k^2\leqslant x\right\}=$$
$$p\left\{\sum_{k=1}^{n}W_k^2\leqslant\frac{n}{2}x\right\}$$

当 $x<0$ 时，$F_X(x)=p\left\{\sum_{k=1}^{n}W_k^2\leqslant\frac{n}{2}x\right\}=0$

当 $x\geqslant 0$ 时，由定理 2 知 $F_X(x)=p\left\{\sum_{k=1}^{n}W_k^2\leqslant\frac{n}{2}x\right\}=$

$$\int_{-\infty}^{\frac{n}{2}x} f_W(w)\mathrm{d}w = \int_0^{\frac{n}{2}x} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} w^{\frac{n}{2}-1} e^{-\frac{w}{2}}\mathrm{d}w$$

故随机变量 X_{ij} 的概率密度函数

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \begin{cases} \dfrac{n^{\frac{n}{2}}}{2^n\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{n}{4}x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{9}$$

由式（9）和Γ分布的定义知

$$X_{ij} \sim \Gamma(\frac{n}{4}, \frac{n}{2}) \quad i \neq j;\ i,j = 1,2,\cdots,m$$

定理 4　随机变量 X_{ij} 的数学期望和方差都存在, 且

$$E(X_{ij}) = 2,\ V_{ar}(X_{ij}) = \frac{8}{n} \quad i \neq j; i,j = 1,2,\cdots,m \tag{10}$$

定理 5　随机平均欧氏距离 D_{ij}（$i \neq j$；$i,j = 1,2,\cdots,m$）的概率密度函数

$$f_D(y) = \begin{cases} \dfrac{n^{\frac{n}{2}}}{2^{n-1}\Gamma(\frac{n}{2})} y^{n-1} e^{-\frac{n}{4}y^2}, & y \geqslant 0 \\ 0, & y < 0 \end{cases} \tag{11}$$

证明：由式（4）知 $D_{ij} = X_{ij}^{\frac{1}{2}}$（$i \neq j$；$i,j = 1,2,\cdots,m$）, 所以随机平均欧氏距离 D_{ij} 的分布函数

$$F_D(y) = p\{D_{ij} \leqslant y\} = p\left\{X_{ij}^{\frac{1}{2}} \leqslant y\right\}$$

当 $y < 0$ 时, $F_D(y) = p\left\{X_{ij}^{\frac{1}{2}} \leqslant y\right\} = 0;$

当 $y \geqslant 0$ 时,

$$F_D(y) = p\left\{X_{ij}^{\frac{1}{2}} \leqslant y\right\} = p\{X_{ij} \leqslant y^2\} = \int_0^{y^2} \frac{n^{\frac{n}{2}}}{2^n\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{n}{4}x}\mathrm{d}x$$

故随机平均欧氏距离 D_{ij}（$i \neq j$；$i,j = 1,2,\cdots, m$）的概率密度函数

$$f_D(y) = \frac{\mathrm{d}F_D(y)}{\mathrm{d}y} = \begin{cases} \dfrac{n^{\frac{n}{2}}}{2^{n-1}\Gamma(\frac{n}{2})} y^{n-1} e^{-\frac{n}{4}y^2}, & y \geqslant 0 \\ 0, & y < 0 \end{cases}$$

定理 6　随机平均欧氏距离 D_{ij} 的数学期望和方差都存在, 且

$$E(D_{ij}) = \frac{2}{\sqrt{n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})},\quad V_{ar}(D_{ij}) = 2\left\{1 - \frac{2}{n}[\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}]^2\right\}$$
$$i \neq j;\ i,j = 1,2,\cdots,m$$

证明：根据定理 3

$$E(D_{ij}) = E(X_{ij}^{\frac{1}{2}}) = \int_{-\infty}^{+\infty} f_X(x) x^{\frac{1}{2}}\mathrm{d}x = \int_0^{+\infty} \frac{n^{\frac{n}{2}}}{2^n\Gamma(\frac{n}{2})} x^{\frac{n+1}{2}-1} e^{-\frac{n}{4}x}\mathrm{d}x$$

令 $u = \dfrac{n}{4}x$, 则

$$E(D_{ij}) = \frac{2}{\sqrt{n}\Gamma(\frac{n}{2})} \int_0^{+\infty} e^{-u} u^{\frac{n+1}{2}-1}\mathrm{d}u = \frac{2}{\sqrt{n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$$

根据定理 4

$$V_{ar}(D_{ij}) = E(D_{ij}^2) - [E(D_{ij})]^2 = E(X_{ij}) - [E(D_{ij})]^2 = 2 - [\frac{2}{\sqrt{n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}]^2 = 2\left\{1 - \frac{2}{n}[\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}]^2\right\}$$

3　分类阈值的确定

定理 5 和定理 6 给出了随机平均欧氏距离 D_{ij}（$i \neq j$；$i,j = 1,2,\cdots,m$）的概率密度函数、数学期望和方差, 并且其分布、数学期望及方差完全由分类指标的个数 n 来决定, n 的取值不同, 其分布、数学期望及方差也不同. $n = 4,10,30,60$ 时随机平均欧氏距离 D_{ij} 的概率密度曲线图见图 1.

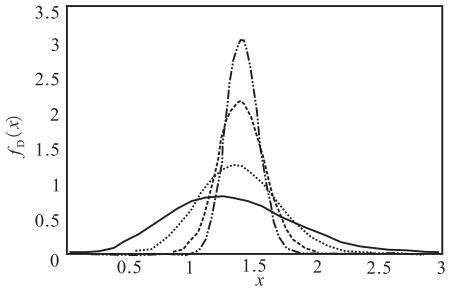

图 1　n =4, 10, 30, 60 时 D_{ij} 的概率密度曲线图

Fig. 1　Probability density curve of D_{ij} when $n = 4,10,30,60$

图 1 中的曲线按其峰值从低至高依次是 $n = 4$、$n = 10$、$n = 30$ 及 $n = 60$ 时随机平均欧氏距离 D_{ij} 的概率密度曲线. 由图 1 可见当 n 取不同值时, 随机平均欧氏距离 D_{ij} 的概率密度曲线的峰值沿横轴方向移动很小, 只是随着 n 的增大, 密度曲线的峰值在不断地增高, 方差在不断地减小.

由于样品 i 和样品 j 间的平均欧氏距离 d_{ij} 是随机平均欧氏距离 D_{ij} 的观测值，因此，当两样品确实同属一类时，其平均欧氏距离的值不应过大，大到使一个小概率事件发生的程度.

定义 4　若对给定的 $\alpha, 0<\alpha\leqslant 0.20$，有 $P\{d_{ij}>d_{\alpha}(n)\}=\alpha$ 成立，则称 $d_{\alpha}(n)$ 为在 α 水平下选用平均欧氏距离，采用最长距离法和最短距离法进行分类时阈值的上确界.

由定义 4 知，对于给定的 α（$0<\alpha\leqslant 0.20$），当选用平均欧氏距离，采用最长距离法和最短距离法进行分类时取分类阈值大于其上确界 $d_{\alpha}(n)$ 是不科学、不可靠的. 当然 α 在区间（0，0.20]上的取值不同，分类阈值的上确界也不同，至于 α 取多少比较合适，这要视具体问题的可靠性要求来决定. $d_{\alpha}(n)$ 的值可通过下式来计算.

$$\int_0^{d_{\alpha}(n)} \frac{n^{\frac{n}{2}}}{2^{n-1}\Gamma(\frac{n}{2})} y^{n-1} e^{-\frac{n}{4}y^2}\,\mathrm{d}y = 1-\alpha \qquad (12)$$

为了便于分类者应用上述方法确定分类阈值的上确界，根据式（12）计算出 n 取不同值时一些常用的分类阈值上确界见表 2.

表 2　分类阈值上确界 $d_{\alpha}(n)$

Tab. 2　Supremum $d_{\alpha}(n)$ of categorical threshold measure

α	n												
	4	8	10	12	14	16	18	20	22	24	26	28	30
0.200	1.730	1.661	1.640	1.623	1.610	1.599	1.590	1.582	1.575	1.569	1.563	1.559	1.550
0.150	1.836	1.734	1.705	1.683	1.665	1.650	1.638	1.627	1.619	1.611	1.604	1.597	1.591
0.100	1.972	1.828	1.788	1.758	1.735	1.715	1.699	1.686	1.674	1.663	1.653	1.646	1.638
0.050	2.178	1.969	1.913	1.872	1.839	1.813	1.791	1.772	1.756	1.742	1.729	1.718	1.708
0.025	2.360	2.094	2.024	1.972	1.932	1.899	1.872	1.848	1.829	1.811	1.796	1.782	1.770
0.010	2.576	2.241	2.154	2.090	2.040	2.000	1.967	1.938	1.914	1.892	1.874	1.857	1.842

参　考　文　献：

[1] 张尧庭，方开泰. 多元统计分析引论[M]. 北京：科学出版社，2003：314—361.

[2] Allen D M. 数量分类学[M]. 张克坚，译. 上海：上海科技出版社，1995：168—218.

[3] 于春海，樊治平. 特征指标信息不完全的系统聚类方法[J]. 系统工程，2006，24（2）：101—105.

[4] 李　伟，王黎勇，杨瑞贞. 运用系统聚类法综合评价农村社区卫生服务中心功能[J]. 中国医院统计，2006，13（3）：201—203.

[5] 刘　江，赵卫国，李小龙，等. 多元统计分析在产品设计要素分析中的应用[J]. 机电产品开发与创新，2007，20（5）：73—75.

[6] 郭　奎，张亚楠，毛利强. 用聚类法确定专题要素的分类方法研究[J]. 测绘信息与工程，2006，31（2）：22—23.

[7] 蔡　健，兰　伟，罗瑞丽，等. 皖北小麦主栽品种遗传多样性的系统聚类法分析[J]. 中国农学通报，2006，22（11）：143—146.

[8] 于洪敏，孙　琰，蔡延曦，等. 装备物资混装配载效益化特征聚类分析[J]. 军械工程学院学报，2005，17（2）：49—52.

[9] Larsen R J，Marx M L. Phonetic Classification [M]. New York：Academic press，1995：78—185.

[10] Fitch W M. Principles of Numerical Taxonomy [M]. New York：Academic press，1993：105—197.

[11] Rao C R. Cluster Analysis Applied to a Study of Race Mixture in Human Populations，Classification and Clustering [M]. New York：Academic press，1977：178—285.

[12] 徐克学. 数量分类学[M]. 北京：科学出版社，1994：54—62.

[13] 茆诗松，周纪芗. 概率论与数理统计[M]. 北京：中国统计出版社，2003：285—308.

[14] 茆诗松，王静龙，濮晓龙. 高等数理统计[M]. 北京：高等教育出版社，1998：167—253.