第 25 卷 第 5 期
2010 年 10 月

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

Vol.25 No. 5
Oct. 2010

随机欧氏距离的统计性质

王玉杰，张大克
（天津科技大学理学院，天津 300457）

摘 要：对随机欧氏距离的统计性质进行了深入的研究,在一定的条件下,确定了它的概率分布,给出了它的数字特征,为科学地分类提供了理论基础.

关键词：系统聚类法；分类；随机欧氏距离；概率分布；数字特征
中图分类号：O212.4 **文献标志码**：A **文章编号**：1672-6510(2010)05-0073-03

Statistical Properties of Random Euclidean Distance

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

Abstract: Statistical properties of random Euclidean distance was studied. The probability distribution was determined and the numerical characteristic was given. It provides the theoretical basis for scientific classification.

Keywords: hierarchical classification method; classification; random Euclidean distance; probability distribution; numerical characteristic

欧氏距离是系统聚类法中使用最普遍的一个距离指标,它主要用来刻画样品间的相近程度,它的大小是分类者进行样品归类的主要依据[1–9]. 如何根据欧氏距离的计算结果进行分类,使分类结果更科学、更可靠,是一个非常值得研究的问题. 本文对随机欧氏距离的统计性质进行了深入的研究,在一定的条件下,确定了它的概率分布,给出了它的数字特征,为科学地进行分类提供了理论基础.

1 随机欧氏距离

假设对 m 个样品进行分类,每个样品观测 n 个分类指标. 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见表 1.

样品 i 和样品 j 间的欧氏距离为

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

假设 x_{i1}, x_{i2}, …, x_{in} 是随机变量 $\xi_{i1},\xi_{i2},\cdots,\xi_{in}$ 的观测值, $i=1,2,\cdots,m$,则式(1)中样品 i 和样品 j 间的欧氏距离是随机距离[10–12]

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

的观测值, D_{ij} 称为样品 i 和样品 j 间的随机欧氏距离($i\neq j$; $i,j=1,2,\cdots,m$).

表 1 观测数据
Tab.1 Observational data

样品	经标准差标准化变换后的指标观测值					
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 随机欧氏距离的概率分布与数字特征

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

收稿日期：2009–11–26；**修回日期**：2009–12–23
基金项目：国家自然科学基金资助项目（60776810）
作者简介：王玉杰（1962—），女，吉林省人，教授，yujiewang@126.com.

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[13]，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（$i,j=1,2,\cdots,m$）.

定理 1　令 $W_{ijk}=\frac{1}{\sqrt{2}}(\xi_{ik}-\xi_{jk})$

$$i\neq j\ ;\ i,j=1,2,\cdots,m\ ;\ k=1,2,\cdots,n \tag{3}$$

则随机变量 $W_{ij1},W_{ij2},\cdots,W_{ijn}$ 相互独立，且

$$W_{ijk}\sim N(0,1)\quad k=1,2,\cdots,n$$

证明　$W_{ij1},W_{ij2},\cdots,W_{ijn}$ 相互独立，且都服从正态分布是显然的. 因为

$$E(W_{ijk})=E[\frac{1}{\sqrt{2}}(\xi_{ik}-\xi_{jk})]=\frac{1}{\sqrt{2}}[E(\xi_{ik})-E(\xi_{jk})]=0$$

$$V_{ar}(W_{ijk})=V_{ar}[\frac{1}{\sqrt{2}}(\xi_{ik}-\xi_{jk})]=\frac{1}{2}[V_{ar}(\xi_{ik})+V_{ar}(\xi_{jk})]=1$$

所以 $W_{ijk}\sim N(0,1),\ k=1,2,\cdots,n$.

引理 1　广义积分 $\int_0^{+\infty}\mathrm{e}^{-\frac{\rho^2}{2}}\rho^{n-1}\mathrm{d}\rho$，当 $n>0$ 时收敛，且 $\int_0^{+\infty}\mathrm{e}^{-\frac{\rho^2}{2}}\rho^{n-1}\mathrm{d}\rho=2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})$.

证明　令 $u=\frac{\rho^2}{2}$，则

$$\int_0^{+\infty}\mathrm{e}^{-\frac{\rho^2}{2}}\rho^{n-1}\mathrm{d}\rho=2^{\frac{n}{2}-1}\int_0^{+\infty}\mathrm{e}^{-u}u^{\frac{n}{2}-1}\mathrm{d}u=2^{\frac{n}{2}-1}\Gamma(\frac{n}{2}).$$

若设

$$X_{ij}=\sum_{k=1}^{n}(\xi_{ik}-\xi_{jk})^2\quad i\neq j\ ;\ i,j=1,2,\cdots,m \tag{4}$$

则由式(2)知样品 i 和样品 j 间的随机欧氏距离

$$D_{ij}=[\sum_{k=1}^{n}(\xi_{ik}-\xi_{jk})^2]^{\frac{1}{2}}=X_{ij}^{\frac{1}{2}}\quad i\neq j\ ;\ i,j=1,2,\cdots,m \tag{5}$$

定理 2　随机变量 X_{ij} 的概率密度函数为

$$f_X(x)=\begin{cases}\dfrac{1}{2^n\Gamma(\frac{n}{2})}\mathrm{e}^{-\frac{x}{4}}x^{\frac{n}{2}-1}, & x\geqslant 0\\ 0\ , & x<0\end{cases} \tag{6}$$

即 $X_{ij}\sim\Gamma(\frac{1}{4},\frac{n}{2})$.

证明　由式(3)和式(4)知 $X_{ij}=2\sum_{k=1}^{n}W_{ijk}^2$，故随机变量 X_{ij} 的分布函数

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

当 $x<0$ 时，$F_X(x)=0$；

当 $x\geqslant 0$ 时，

$$F_X(x)=\iint_{\Omega}\cdots\int(\frac{1}{\sqrt{2\pi}})^n\mathrm{e}^{-\frac{1}{2}\sum_{k=1}^{n}w_{ijk}^2}\mathrm{d}w_{ij1}\mathrm{d}w_{ij2}\cdots\mathrm{d}w_{ijn}\ ,$$

其中

$$\Omega=\left\{(w_{ij1},w_{ij2},\cdots,w_{ijn})\,|\,\sum_{k=1}^{n}w_{ijk}^2\leqslant\frac{1}{2}x,\ (w_{ij1},w_{ij2},\cdots,w_{ijn})\in R^n\right\}.$$

球坐标变换得

$$F_X(x)=\int_{-\pi}^{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cdots\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_0^{\sqrt{\frac{x}{2}}}\frac{1}{(\sqrt{2\pi})^n}\mathrm{e}^{-\frac{\rho^2}{2}}\cdot\rho^{n-1}J(\theta_1,\theta_2,\cdots,\theta_{n-1})\mathrm{d}\rho\mathrm{d}\theta_1\mathrm{d}\theta_2\cdots\mathrm{d}\theta_{n-1} \tag{7}$$

其中 $J(\theta_1,\theta_2,\cdots,\theta_{n-1})$ 为坐标变换的 Jacobi 行列式.

令　$C_{n-1}=\frac{1}{(\sqrt{2\pi})^n}\int_{-\pi}^{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cdots\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}J(\theta_1,\theta_2,\cdots,\theta_{n-1})\cdot\mathrm{d}\theta_1\mathrm{d}\theta_2\cdots\mathrm{d}\theta_{n-1}$ (8)

则式(7)转化为

$$F_X(x)=C_{n-1}\int_0^{\sqrt{\frac{x}{2}}}\mathrm{e}^{-\frac{\rho^2}{2}}\rho^{n-1}\mathrm{d}\rho \tag{9}$$

因为 $\lim\limits_{x\to+\infty}F(x)=1$，所以由引理 1 知

$$C_{n-1}\int_0^{+\infty}\mathrm{e}^{-\frac{\rho^2}{2}}\rho^{n-1}\mathrm{d}\rho=C_{n-1}2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})=1$$

$$C_{n-1}=\frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}$$

故由式(9)得

$$F_X(x)=\frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}\int_0^{\sqrt{\frac{x}{2}}}\mathrm{e}^{-\frac{\rho^2}{2}}\rho^{n-1}\mathrm{d}\rho \tag{10}$$

对式(10)求导得随机变量 X_{ij} 的概率密度函数

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

因此 $X_{ij}\sim\Gamma(\frac{1}{4},\frac{n}{2})$.

定理 3　随机变量 X_{ij} 的数学期望和方差都存在，且 $E(X_{ij})=2n$，$V_{ar}(X_{ij})=8n$.

因为 $X_{ij}\sim\Gamma(\frac{1}{4},\frac{n}{2})$，所以定理 3 的结果是显然成立的[14-15].

定理 4　样品 i 和样品 j 间随机欧氏距离 D_{ij}（$i\neq j\ ;\ i,j=1,2,\cdots,m$）的概率密度函数为

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

证明　随机欧氏距离 D_{ij} 的分布函数

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

当 $z<0$ 时，$F_D(z)=0$；

当 $z\geqslant 0$ 时，

$$F_D(z)=p\left\{0\leqslant X_{ij}\leqslant z^2\right\}=\frac{1}{2^n\Gamma(\frac{n}{2})}\int_0^{z^2}\mathrm{e}^{-\frac{x}{4}}x^{\frac{n}{2}-1}\mathrm{d}x \tag{12}$$

对式(12)求导得随机欧氏距离 D_{ij} 的概率密度函数

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

定理 5　样品 i 和样品 j 间随机欧氏距离 D_{ij}（$i\neq j$；$i,j=1,2,\cdots,m$）的数学期望和方差都存在，并且

$$E(D_{ij})=2\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$$

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

证明

$$E(D_{ij})=E(X_{ij}^{\frac{1}{2}})=\int_{-\infty}^{+\infty}\sqrt{x}f_X(X)\mathrm{d}x=$$

$$\frac{1}{2^n\Gamma(\frac{n}{2})}\int_0^{+\infty}\mathrm{e}^{-\frac{x}{4}}x^{\frac{n-1}{2}}\mathrm{d}x$$

令 $u=\dfrac{x}{4}$，则

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

由定理 3 知 $E(D_{ij}^2)=E(X_{ij})=2n$，故

$$V_{ar}(D_{ij})=E(D_{ij}^2)-[E(D_{ij})]^2=$$

$$2n-[2\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}]^2=2\left\{n-2[\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}]^2\right\}$$

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] 裴鑫德. 多元统计分析及其应用[M]. 北京：北京农业大学出版社，1991：89–172.

[4] 安希忠. 实用多元统计方法[M]. 长春：吉林科学技术出版社，1992：114–160.

[5] 叶皓，沈顺，张祥民. 液相指纹图谱结合欧氏距离对野菊花质量控制的研究[J]. 世界科技研究与发展，2006，28(2)：72–74.

[6] 叶君武，陈淑琴，周丽琴，等. 用欧氏距离系数预报舟山海域赤潮发生指数等级[J]. 海洋环境科学，2010，29(1)：108–111.

[7] 曾现来，刘晓红，张增强. 应用欧氏距离聚类法综合评价环境质量[J]. 中国给水排水，2003，19(12)：99–100.

[8] 耿协鹏，胡鹏. 基于最短欧氏距离的空间点集聚类的栅格算法[J]. 测绘科学，2008，33(3)：35–37.

[9] 吴成东，贾子熙，张云洲，等. 基于欧氏距离的分布式网格定位估计方法[J]. 东北大学学报：自然科学版，2009，30(3)：325–328.

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

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

[12] 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.

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

[14] 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.

[15] 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.