数种 Matlab 相似函数和距离函数的建立及其在热带牧草研究中的应用

刘光华[1],刘国道[2],张文军[3]
([1]广东农工商职业技术学院,广州 510507;[2]中国热带农业科学院品资所,儋州 571737;
[3]中山大学生命科学学院,广州 510275)

摘 要:相似函数和距离函数广泛用于聚类分析、相似性分析及差异性分析中。不同的相似函数和距离函数其性质有所不同,可造成分析结果的差异。该研究为 Matlab 计算环境,生成了 13 种相似函数和距离函数:欧氏距离,Manhattan 距离,契比雪夫距离,Jaccard 系数,夹角余弦,联列系数,连关系数,点相关系数,四分相关系数,及变型夹角余弦等。通过实例,说明了一些函数的适用范畴与使用方法。用户可根据需要进行调用,或在此基础上扩充新的功能或函数。同时,给出了热带牧草应用实例。

关键词:相似函数;距离函数;Matlab;建立;热带牧草应用
中图分类号:O212 **文献标识码**:A

Matlab Implementation of Dozens of Similarity and Distance Functions and Their Applications

Liu Guanghua[1], Liu Guodao[2], Zhang Wenjun[3]
([1]*Guangdong AIB Polytech College*, Guangzhou 510507;
[2]*Chinese Academy of Tropical Agricultural Sciences*, Danzhou 571737; [3]*Sun Yat-sen University*, Guangzhou 510275)

Abstract: Similarity functions and distance functions are widely used in cluster analysis, similarity analysis, and difference analysis. These functions have various mathematical forms and will yield different results. In this paper 13 similarity functions and distance functions were developed as Matlab functions. They can be used in various Matlab-based studies. Some applications of these functions were listed.

Key words: similarity functions, distance functions, Matlab, implementation, applications

相似函数和距离函数是聚类分析,相似性分析,以及差异性分析的基础。选择不同的相似函数和距离函数,其结果可能会有显著差异。例如,对若干样品进行聚类分析,选择欧氏距离,Manhattan 距离,或契比雪夫距离,则前两者因数学形式相近,结果可能相似,但契比雪夫距离的聚类结果会有很大的不同。Matlab 是国际上使用最为广泛的科学工程计算软件[1],在该软件中提供有关函数,显然有助于用户直接调用,或进行扩充改进。鉴于此,该研究在 Matlab 环境中建立了 13 种相似函数和距离函数,可供聚类分析等研究使用。同时,给出了应用实例。

1 相似函数和距离函数

相似函数和距离函数种类很多[2,3],笔者选择若干种较为重要的函数。设有 2 个 n 维向量 x 和 y,$x=(x_1, x_2, \cdots, x_n)$,$y=(y_1, y_2, \cdots, y_n)$。相似函数和距离函数的数学形式[2~10]分别如下。

1.1 相似函数

若干种较为重要的相似函数如下。

1.1.1 连续取值函数

夹角余弦:$s=xy^T/(xx^Tyy^T)^{1/2}$

1.1.2 离散多值函数

联列系数:$s=(r^2/(r^2+n..))^{1/2}$

基金项目:中国热带农业科学院科研项目(E2005-1)资助。
第一作者简介:刘光华,男,1965 年出生,广东农垦热带作物研究所,副所长,理学博士,农学博士后,副教授,主要研究领域为生态学、热带作物等。通信地址:510507 广东省广州市天河区粤垦路 198 号广东农工商职业技术学院亚热带作物研究所,Tel:020-85233599,E-mail:liu664@gmail.com。
收稿日期:2008-06-23,修回日期:2007-07-20。

连关系数 1：$s=(r^2/(n..max(p\text{-}1,q\text{-}1)))^{1/2}$

连关系数 2：$s=(r^2/(n..min(p\text{-}1,q\text{-}1)))^{1/2}$

连关系数 3：$s=(r^2/(n..((p\text{-}1)(q\text{-}1))^{1/2}))^{1/2}$

其中，向量 x 有 p 个定性值，$t_1,t_2,\cdots,t_p$，向量 y 有 q 个定性值，$r_1,r_2,\cdots,rq$，n_{kl} 是向量 x 取值 t_k 而向量 y 取值 r_l 的个数，$k=1,2,\ldots,p$；$l=1,2,\ldots,q$，且 $r^2=n(\sum^{p}_{i=1}\sum^{q}_{j=1}n_{ij}^2/(n_in_j)\text{-}1)$，$n=\sum^{p}_{i=1}n_i.$，$n_i=\sum^{q}_{j=1}n_{ij}$，$n_j=\sum^{p}_{i=1}n_{ij}$。

1.1.3 离散二值函数

点相关系数：$s=(ad-bc)/((a+b)(c+d)(a+c)(b+d))^{1/2}$

四分相关系数：$s=sin\ ((a+d-(b+c))/\ (a+b+c+d)\times 3.1415926/2)$

变型夹角余弦 1：$s=(a\times a/((a+b)(a+c)))^{1/2}$

变型夹角余弦 2：$s=(a\times a\times d\times d/((a+b)(a+c)(b+d)(c+d)))^{1/2}$

其中，a 为 n 维中向量 x 和 y 同取 0 的个数，d 为向量 x 和 y 同为非 0 值的个数，b 为向量 x 取 0 且向量 y 取非 0 值的个数，c 为向量 x 取非 0 值且向量 y 取 0 的个数。

1.2 距离函数

数种较为重要的距离函数如下：

1.2.1 连续取值函数

欧氏距离：$d=((x-y)(x-y)^{T})^{1/2}/n$

Manhattan 距离：$d=\sum^{n}_{k=1}|x_k-y_k|/n$

契比雪夫距离：$d=max_k|x_k-y_k|$

1.2.2 离散二值函数

Jacarrd 系数：$d=(b_x+b_y)/(c_x+cy-e)$

其中，b_x 为在向量 x 中出现而不在向量 y 中出现的非 0 值的个数，b_y 为在向量 y 中出现而不在向量 x 中出现的非 0 值个数，c_x 和 c_y 分别为在向量 x 和 y 出现的非 0 值个数，e 为同时在向量 x 和 y 中出现的非 0 值个数。

2 相似函数和距离函数的 Matlab 实现

上述相似函数和距离函数的 Matlab 函数源代码如下：

2.1 欧氏距离

```
function distance=euclideandis (x,y) % x and y: two vectors to be tested.
if (max(size(x))~=max(size(y))) %label1
error('Array sizes do not match.');
end
if ((min(size(x))~=1)|(min(size(y))~=1))
error('Both x and y are vectors');
end %label2
distance=sqrt(sum((x-y).^2))/max(size(x));
```

2.2 Manhattan 距离

```
function distance=manhattandis(x,y) % x and y: two vectors to be tested.
%该段与 label1-label2 之间相同
distance=sum(abs(x-y))/max(size(x));
```

2.3 契比雪夫距离

```
function distance=chebyshovdis(x,y) % x and y: two vectors to be tested.
%该段与 label1-label2 之间相同
distance =max(abs(x-y));
```

2.4 Jaccard 系数

```
function distance=jaccarddis (x,y) %x and y: two vectors to be tested.
%该段与 label1-label2 之间相同
bb=0; cc=0; dd=0; nn1=0; rr1=0;
for kk=1:max(size(x))
if (x(kk)~=0) nn1=nn1+1; end
if (y(kk)~=0) rr1=rr1+1; end
if ((x(kk)==0) & (y(kk)~=0)) bb=bb+1; end
if ((x(kk)~=0) & (y(kk)==0)) cc=cc+1; end
if ((x(kk)~=0) & (y(kk)~=0)) dd=dd+1; end
end
distance=(cc+bb)/(nn1+rr1-dd);
```

2.5 夹角余弦

```
function similarity=angularcosinesim(x,y) %x and y: two vectors to be tested.
%该段与 label1-label2 之间相同
aa=sum(x.×y,2);
bb=sum(x.^2,2);
cc=sum(y.^2,2);
similarity=aa./sqrt(bb. ×cc);
```

2.6 联列系数

```
function similarity=linkagesim(x,y)
%该段与 label1-label2 之间相同
pn=1;qn=1; %label3
pp(1)=y(1); ww(1)=x(1);
for kk=1:max(size(x))
jj=0;
for ii=1:pn if (y(kk)~=pp(ii)) jj=jj+1; end; end
if (jj==pn) pn=pn+1; pp(pn)=y(kk); end
jj=0;
for ii=1:qn if (x(kk)~=ww(ii)) jj=jj+1; end; end
if (jj==qn) qn=qn+1; ww(qn)=x(kk); end
end
```

```
for kk=1:pn
for jj=1:qn
temp(kk,jj)=0;
for ii=1: max(size(x)) if ((y(ii)~=pp(kk)) & (x(ii)~=ww(jj))) temp(kk,jj)=temp(kk,jj)+1; end; end
end
end
summ=0;
for kk=1:pn
pp(kk)=0;
for jj=1:qn pp(kk)=pp(kk)+temp(kk,jj); end
summ=summ+pp(kk);
end
for kk=1:qn
ww(kk)=0;
for jj=1:pn ww(kk)=ww(kk)+temp(jj,kk); end;
end
xsquare=0;
for kk=1:pn
for jj=1:qn xsquare=xsquare+temp (kk,jj)*temp(kk,jj)/(pp(kk)*ww(jj)); end;
end
xsquare=summ×(xsquare-1); %label4
similarity=sqrt(xsquare/(xsquare+summ));
```

2.7 连关系数 1

```
function similarity=colinkage1sim(x,y)
%该段与 label1-label2 之间相同
%该段与 label3-label4 之间相同
similarity=sqrt(xsquare/(summ×max(pn-1,qn-1)));
```

2.8 连关系数 2

```
function similarity=colinkage2sim(x,y)
%该段与 label1-label2 之间相同
%该段与 label3-label4 之间相同
similarity=sqrt(xsquare/(summ×min(pn-1,qn-1)));
```

2.9 连关系数 3

```
function similarity=colinkage3sim(x,y)
%该段与 label1-label2 之间相同
%该段与 label3-label4 之间相同
similarity =sqrt(xsquare/(summ×sqrt((pn-1) ×(qn-1))));
```

2.10 点相关系数

```
function similarity=pointcorresim(x,y)
%该段与 label1-label2 之间相同
aa=0; bb=0; cc=0; %label5
dd=0; %label6
for kk=1:max(size(x))
if ((x(kk)==0) & (y(kk)==0)) aa=aa+1; end
if ((x(kk)==0) & (y(kk)~=0)) bb=bb+1; end
if ((x(kk)~=0) & (y(kk)==0)) cc=cc+1; end
if ((x(kk)~=0) & (y(kk)~=0)) dd=dd+1; end %label7
end %label8
similarity= (aa×dd-bb×cc)/sqrt ((aa+bb)×(cc+dd)×(aa+cc)×(bb+dd));
```

2.11 四分相关系数

```
function similarity=quadraticcorresim(x,y)
%该段与 label1-label2 之间相同
%该段与 label5-label8 之间相同
similarity=sin ((aa+dd- (bb+cc))/ (aa+bb+cc+dd)×3.1415926/2);
```

2.12 变型夹角余弦 1

```
function similarity=angularcosine1sim(x,y)
%该段与 label1-label2 之间相同
%该段去掉 label6 和 label7，其余与 label5-label8 之间相同
similarity=sqrt(aa×aa/((aa+bb)×(aa+cc)));
```

2.13 变型夹角余弦 2

```
function similarity=angularcosine2sim(x,y)
%该段与 label1-label2 之间相同
%该段与 label5-label8 之间相同
similarity=sqrt (aa×aa×dd×dd/ ((aa+bb)×(aa+cc)×(bb+dd)×(cc+dd)));
```

3 案例应用

3.1 连续取值函数

已测得 5 个柱花草株系的干物质产量，旱季产量，增产性，比例，病级，长势，及存活率，属于连续取值类型，可计算各株系之间的 Manhattan 距离，Matlab 源代码为：

```
m=5;
for i=1:m
for j=i:m
distancematrix(i,j)=manhattandis(data(i,:),data(j,:));
distancematrix(j,i)=distancematrix(i,j);
end
end
distancematrix
```

结果见表 1，该表中，株系 1-5 分别代表 2001-1，2001-2，2001-3，2001-4，2001-5。由该表可进行 5 个柱花草之间的差异性分析。显然，2001-3 与 2001-4 最接近，2001-2 与 2001-3 差异最大。

表 1 柱花草株系之间的 Manhattan 距离

	1	2	3	4	5
1	0	4.9976	5.34	3.8887	2.5331
2	4.9976	0	9.6239	8.1726	4.3724
3	5.34	9.6239	0	1.6956	5.3194
4	3.8887	8.1726	1.6956	0	4.1613
5	2.5331	4.3724	5.3194	4.1613	0

表 2 热带人工草地放牧地的植被生物量分布

	10cm	20cm	30cm	40cm	50cm
6 月	62.3	61.3	48	31.3	19.7
8 月	25	35.5	39.5	41.3	43.8
10 月	31	74	104	115.8	105.3
12 月	17	38.3	51	55.5	57.3

3.2 离散多值函数

已测得热带人工草地放牧地 6,8,10,12 月在 10,20,30,40,50cm 处的植被生物量(表 2),对植被生物量按如下方法分级:(0,30)→1,(31,60)→2,(61,90)→3,(91,∞)→4,共有 4 级,数据属于离散多值类型。选用联列系数,分析各月份植被生物量分布的相似性。

各月份之间植被生物量的联列系数见表 3。

从系数值看,在植被生物量分布的相似性上,8 月与 12 月最相似,6 月与其它月份相似性较小。

3.3 离散二值函数

只考虑植被生物量的大与小,即在表 2 中,对植被生物量按如下方法分级:(0,40)→0,(41,∞)→1,共有 2 级,则原始数据成为离散二值类型。选用点相关系数,则得如下各月份植被生物量分布的相似性。此时,8 月与 12 月最相似,6 月与 8 月相似性较小(表 4)。

表 3 各月份之间植被生物量的联列系数

	6 月	8 月	10 月	12 月
6 月	0.3492	0.3206	0.3206	0.3206
8 月	0.3206	0.7071	0.4237	0.7071
10 月	0.3206	0.4237	0.3676	0.4237
12 月	0.3206	0.7071	0.4237	0.7071

表 4 各月份之间植被生物量的点相关系数

	6 月	8 月	10 月	12 月
6 月	1	-1	-0.4082	-0.6667
8 月	-1	1	0.4082	0.6667
10 月	-0.4082	0.4082	1	0.6124
12 月	-0.6667	0.6667	0.6124	1

4 结论

Matlab 是国际上应用最广泛的科学工程计算软件。研究给出了 13 种相似函数和距离函数的 Matlab 函数，可供用户直接调用，或进行扩充改进。这些 Matlab 函数可用于聚类分析,相似性分析,以及差异性分析中。应用实例表明,它们是客观可用的。

参考文献

[1] 魏巍.MATLAB 应用数学工具箱技术手册.北京:国防工业出版社,2004.

[2] 张尧庭,方开泰.多元统计分析引论.北京:科学出版社,1982.

[3] 张文军.生态学研究方法.广州:中山大学出版社,2007.

[4] 张文军. 一种随机化检验算法及其 Matlab 实现. 生物数学学报,2007,22(4):745-752.

[5] 张文军,齐艳红,张治国.生境均质性分析及随机化检验的算法与网络软件.计算机应用与软件,2004,21(11):66-69.

[6] 张文军,齐艳红,张治国.生态样带边界分析的改进算法与网络计算.生物数学学报,2005,20(4):477-486.

[7] 齐艳红,张文军.CorreDetector:一种用于信息资料相关性分析的网络共享软件.情报学报,2003,22(Suppl.):266-268.

[8] 齐艳红. 生境异质性与有害生物在生境中扩散的模型与算法研究:[学位论文].中山大学:中山大学计算机科学系,2003:46-47.

[9] Zhang WJ. Computer inference of network of ecological interactions from sampling data. Environmental Monitoring and Assessment, 2007, 124:253-261.

[10] Krebs CJ. Ecological Methodology. Happer Collins Publishers, Inc. New York, 1989.