DOI:10.16836/j.cnki.jcuit.2015.04.008

文章编号: 1671-1742(2015)04-0343-06

排序序列对比方法研究

王 娟, 石 磊
(成都信息工程大学信息安全工程学院,四川 成都 610225)

摘要: 排序是很多应用问题的抽象,排序序列的对比是研究这些抽象问题的必要步骤。现有对排序序列的对比多是考察其中重要节点序位变化,缺少对序列整体的评价和对比。针对这点,首先总结归纳现有研究中常见的对比方法,主要分为基于距离和基于相关系数两大类。在此基础上,通过原理分析和实验对比,发现某些方法在应用于排序序列对比时,存在原理缺陷和受排序元素量纲影响等问题。最后推荐排序序列对比使用斯皮尔曼或肯德尔相关系数进行衡量,并给出了二者的适用范围。

关 键 词: 计算机科学与技术; 复杂网络; 排序对比; 排序相似度; 相关系数; 序列距离
中图分类号: TP391 文献标志码: A

0 引言

排序是很多应用问题的抽象。例如,舆情监测中找寻"舆论领袖",其实质是用某种方法对观察范围内的舆情对象进行重要性排序,排在前面的个体就是"舆论领袖"[1];类似的,在交通系统中,对各个站点按某种标准进行重要性排序,排在前面的站点就是需要发现的"交通枢纽"[2];对网络节点按某种标准进行重要性排序,排在前面的节点就是需要重要防护的"网络关键节点"[3-4]。

在研究排序算法的时候,经常遇见需要比较两个排序序列的情况。例如,有一个作为基准(benchmark)的排序序列,可能是专家给出的,也可能是基准算法给出的;研究者设计的排序算法需要跟这个基准排序做比较,以判断新设计的排序算法的效果。或者是两个排序算法需要对比排序效果。目前在分析中大多采用的方法是:

(1) 对排序对象集中,较重要的节点的排序序位做分析,判断其是否获得适当的序号,比较基准序列的序号是高了还是低了? 为什么?

(2) 对某些重要的对象组合,分析其相对位置,即如果基准排序中对象A被排在对象B前面,新设计的排序算法的排序中是否保持了对象A在对象B之前,之前多少位? 相对距离一样吗? 为什么?

这两种常用分析方法着眼点都比较小,分析的是排序序列的局部特征。当需要考察排序的整体效果时,这两种方法就不适用了。目前对排序结果的整体考察还没有比较统一的方法。采用较多的是编码和机器学习等领域里面的基于距离的考察方法和基于相关系数的考察方法。

排序序列对比跟一般编码对比具有自己的的特点: 对比的两个序列中元素是一样的,对比的仅仅是序位的不同。而一般编码序列,两个序列的元素都有可能不一样。这个特征必然导致一部分编码序列对比适用的方法对排序序列对比不适用。

对现有的序列对比方法做了一个比较归纳,通过对原理的讨论和比较实验,探讨现有方法在比较排序序列时的优缺点,最终给出比较适合排序序列整体比较的方法和适用范围。

1 常见排序序列比较方法总结

序列比较常出现于编码、机器学习等领域,现有常用的比较方法有基于距离的方法和基于相关系数又称相似度的方法2大类。将这两类常用的方法其原理和实现方式进行总结归纳。

1.1 基于距离的序列比较方法

1.1.1 汉明距离

在文字处理、信息编码中,汉明距离(Hamming distance)[5]是最常见也是最基础的比较方法。其计算原理为: 2个待比较的编码A与B,其对应位置比特取值不同个数,称为这2个编码的汉明距离。举例如下: 编码A=10101,B=00111,从第一位开始依次有第一位、第四位不同,则汉明距离为2。

1.1.2 欧式距离

收稿日期: 2015-08-04
基金项目: 四川省科技厅应用基础研究计划资助项目(2013JY0064、2014JY0071); 四川省教育厅资助项目(13Z182、13ZB0088); 网络与数据安全四川省重点实验室开放课题资助项目(NDS2015-01)

欧几里得度量(Euclidean metric),也称欧式距离(Euclidean distance)[6-7]。其原理是在 n 维空间中2个点之间的真实距离,或者向量的自然长度(即该点到原点的距离),记为

$$d_E = \| x - y \|_2 = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \quad (1)$$

其中:x,y 为 n 维空间点 $x(x_1, x_2, \cdots, x_n)$,$y(y_1, y_2, \cdots, y_n)$。

二维平面上两点 $x(x_1, x_2)$ 与 $y(y_2, y_2)$ 间的欧氏距离为

$$d_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (2)$$

三维空间两点 $x(x_1, x_2, x_3)$ 与 $y(y_1, y_2, y_3)$ 间的欧氏距离为

$$d_3 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \quad (3)$$

1.1.3 曼哈顿距离

想象在曼哈顿要从一个十字路口开车到另外一个十字路口,因为不能穿墙,因此驾驶距离必然不是2点的直线距离即欧式距离。实际沿街区驾驶的这种距离被称为“曼哈顿距离(Manhattan distance)”,也称为城市街区距离(CityBlock distance)[8],记为 d_M。

2个 n 维向量 $a(x_{11}, x_{12}, \cdots, x_{1n})$ 与 $b(x_{21}, x_{22}, \cdots, x_{2n})$ 间的曼哈顿距离计算公式为

$$d_M = \sum_{k=1}^{n} |x_{1k} - x_{2k}| \quad (4)$$

常用的二维平面两点 $a(x_1, y_1)$ 与 $b(x_2, y_2)$ 间的曼哈顿距离为

$$d_2 = |x_1 - x_2| + |y_1 - y_2| \quad (5)$$

1.1.4 契比雪夫距离

契比雪夫距离(Chebyshev distance)又被称为国际象棋距离(Chessboard distance)[9]。在国际象棋中,国王走一步能够移动到相邻的8个方格中的任意一个。那么国王从格子 (x_1, y_1) 走到格子 (x_2, y_2) 最少需要的步数就是契比雪夫距离。2个 n 维向量 $x(x_1, x_2, \cdots, x_n)$ 与 $y(y_1, y_2, \cdots, y_n)$ 间的契比雪夫距离为

$$d_C = \max_i(|x_i - y_i|) \quad (6)$$

1.2 基于相关系数的比较方法

1.2.1 汉明相关系数

很容易看出,汉明距离越大则两个编码的相关系数越低,最大汉明距离就是编码长度,定义汉明相关系数为

$$\mathrm{sim}(x, y) = 1 - \frac{d(x, y)}{len} \quad (7)$$

$d(x, y)$ 是汉明距离,len 是编码总长度。其取值范围为 [0,1],数值越大,相关系数越高,两个集合的关系越大。

1.2.2 余弦夹角相关系数

余弦夹角相关系数[10]与欧式距离有很大关系。2个向量的余弦可以从欧式点积推导出

$$A \cdot B = \| A \| \| B \| \cos\theta \quad (8)$$

其中 $\cos\theta$ 就是向量 A 和 B 的余弦夹角相关系数,可以表示为

$$\cos\theta = \frac{A \times B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (9)$$

1.2.3 杰卡德相关系数

杰卡德相关系数(Jaccard similarity coefficient)[11]是比较2个集合的相关系数一种指标。2个集合 A 和 B 的交集元素在 A,B 的并集中所占的比例,称为2个集合的杰卡德相关系数,用符号 $J(A, B)$ 表示。其相关系数取值范围为 [0,1],数值越大,相关系数越高,2个集合的关系越大。

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (10)$$

但是排序序列是相同元素集合的不同排序的对比,没有不同元素出现,其杰卡德相关系数必然是1。因此从不能使用原始的杰卡德相关系数定义,需要对定义做一点改动。

在排序序列对比中,由于两个序列元素是一样的,因此 $|A \cup B| = |A| = |B|$ 即排序元素的个数,记为len。$|A \cap B|$ 被重新定义为“相同元素在相同的位置”的数目。例如:序列 [1 2 3] 和序列 [2 1 3],元素1和2在两个序列中的位置不同,只有3是一样的因此 $|A \cap B| = 1$,杰卡德系数为:1/3 = 0.33

1.2.4 皮尔森相关系数

皮尔森相关系数(Pearson correlation coefficient)也称皮尔森积矩相关系数(Pearson product-moment correlation coefficient)[12],其计算公式如下,对2个待比较的样本集合 X 与 Y 有

$$\rho_{X,Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} \quad (11)$$

其中 $\bar{X}$ 是 X 的样本均值,$\bar{Y}$ 是 Y 的样本均值。其取值区间为 [-1,1]。相关系数在 -1 与 0 之间,X,Y 呈负相关,即 X 的值增大(减小) Y 值反而减小(增大);如果相关系数为0,则 X 和 Y 两变量无关系;当相关系数相关系数在0与1之间时,X 和 Y 两变量呈正

相关关系,即 X 的值增大(减小),Y 值也随之增大(减小)。相关系数的绝对值的大小代表相关性的强弱,相关系数越接近于 1 或 -1,相关度越强,相关系数越接近于 0,相关度越弱。一般强弱相关性按表 1 判断。

表 1　相关系数判断相关性强弱

下界	上界	相关性
0.8	1.0	极强相关
0.6	0.8	强相关
0.4	0.6	中等程度相关
0.2	0.4	弱相关
0.0	0.2	极弱相关或无相关

皮尔森相关系数要求: (1) 两个变量之间是线性关系,都是连续数据。(2) 两个变量的总体是正态分布,或接近正态的单峰分布。(3) 两个变量的观测值是成对的,每对观测值之间相互独立。

1.2.5　斯皮尔曼相关系数

由上可知皮尔森相关系数(Spearman correlation)[13] 适用范围较窄,有较大的使用限制。斯皮尔曼相关系数的定义与皮尔森相关系数比较类似,但是限制条件要宽松很多。只要两个变量的观测值是成对的等级评定资料,或者是由连续变量观测资料转化得到的等级资料,不论两个变量的总体分布形态、样本容量的大小如何,都可以用斯皮尔曼相关系数来进行研究。由于它是以元素被评价的等级而不是元素的数值,所以又被称为“斯皮尔曼等级相关系数(Spearman's rank correlation coefficient)”。

计算方法如下: 假设 2 个随机变量(2 个集合)分别为 X、Y(也可以看做 2 个集合),元素个数 N,对 X、Y 进行排序(同时为升序或降序),得到 2 个元素排行集合 x、y,对 x,y 进行以下计算,跟皮尔森系数公式类似,为

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i}^{N}(y_i - \bar{y})^2}} \tag{12}$$

1.2.6　肯德尔相关系数

肯德尔相关系数(Kendall correlation coefficient)[14] 也是通过元素的位次而不是绝对数值来考察序列的相关性的,这点与斯皮尔曼系数类似,但是计算方法不同。

设 (x_1, y_1),(x_2, y_2),…,(x_n, y_n) 是随机变量 X 和 Y 的一组联合观测集合。(x_i) 和 (y_i) 的值是唯一的。其中满足以下条件的被称为“和谐对(Concordant pairs)”: 有 $x_i > x_j$ 则 $y_i > y_j$ 或 $x_i < x_j$ 则 $y_i < y_j$。如果不满足则被称为“不和谐对(Discordant pairs)”。对要比较的排序序列,分别按顺序取出元素对,并计数“和谐对”和“不和谐对”的个数。肯德尔相关系数由以下公式计算出,其中 n 是元素数目

$$\tau = \frac{\text{和谐对数目} - \text{不和谐对数目}}{\frac{1}{2}n(n-1)} \tag{13}$$

举例: 有两个排序: $O1 = [a,c,b,d]$,$O2 = [a,c,d,b]$

计算:

$P1 = \{[a,c], [a,b], [a,d], [c,b], [c,d], [b,d]\}$,$P2 = \{[a,c], [a,b], [a,d], [c,b], [c,d], [d,b]\}$.

和谐对: $[a,c]$,$[a,b]$,$[a,d]$,$[c,b]$,$[c,d]$ 共 5 对,不和谐 $[d,b]$ 1 对,元素 a,b,c,d 4 个,因此有

$$\tau = \frac{5-1}{\frac{1}{2}\times 4\times(4-1)} = \frac{4}{2\times 3} = 0.67$$

1.3　方法原理分析比较

1.3.1　基于距离的方法分析

首先,对比距离和相关系数两种度量,前者只能给出距离的绝对值,在没有相对参照的情况下,只能简单感知对比序列的差异,没有整体度量。从整体度量考虑,实用性不如相关系数。例如,有 n 个元素进行排序,序列 x 为基准序列,序列 $x1$ 与 x 距离 $m1$。$m1$ 到底是说两者区别不大,还是区别很大? 整体上难以判断。只能用两组对比互相比较,即另有序列 $x2$ 与基准排序 x 距离 $m2$,对比 $m1$ 与 $m2$ 的值可以分析两个排序跟基准排序的差异大小。

距离计算中的汉明距离比较特别,因为汉明距离的最大值就是元素个数,因此很容易将汉明距离转化为汉明相关系数(公式 7)。而其他距离计算方法在多维度下的最大距离很难计算,因此没有类似相关系数定义。

其次,除汉明距离以外的其他 3 个距离: 曼哈顿距离、欧氏距离和契比雪夫距离都存在明显的缺点: 不能处理不同量纲的值,即将各个分量的量纲(scale),也就是“单位”当作相同的看待了,也没有考虑各个分量的分布(期望,方差等)的不同。当元素量纲不同时,计算出的距离会失真,这点在实验中有具体体现。

1.3.2　基于相关系数方法的分析

首先,分析余弦相关系数,该相关系数考察的是两个向量在方向上的差异,而不是绝对距离。如果保持 A 点位置不变,B 点朝原方向远离坐标轴原点,那么这

个时候余弦距离 $\cos\theta$ 是保持不变的(因为夹角没有发生变化),而 A、B 两点的距离显然在发生改变。反应这种变化的是欧式距离。因此可以推论,如果排序序列构成的两个向量在方向上变化不大,但是绝对距离有较大变化,余弦相关系数是体现不出来的。这明显不是排序序列对比要的结果。实验数据也支持这点。

其次,杰卡德相关系数主要用于考察含不同元素集合的相关系数,文中将其重新进行排序中含义的定义后,其表现有待考察。其计算复杂度较小算是一个优点。

最后,皮尔森、斯皮尔曼和肯德尔相关系数,三者关系紧密,但又各有特点。

(1) 皮尔森和斯皮尔曼相关系数的计算公式一样(公式11与12),二者的区别主要在于皮尔森的计算对象是元素的具体值,要求对比的两个变量是连续且呈线性相关的。该要求很高,排序序列的元素很难满足,因此适用范围非常受限。而斯皮尔曼的计算对象是元素的秩次跟其具体值无关。因此又称秩相关系数,是利用两变量的秩次大小作线性相关分析,对原始变量的分布不作要求,适用范围要广些。因此,对于不连续的,元素量纲不同的排序,皮尔森相关系数是不适合的。可以采用斯皮尔曼或者肯德尔等级相关系数。

(2) 斯皮尔曼与肯德尔等级相关系数的计算对象都是元素的秩次跟其具体值无关,很多情况下二者适用范围都类似。但是斯皮尔曼相关系数的计算复杂度比肯德尔相关系数要低,因此通常都使用斯皮尔曼相关系数进行计算。

(3) 但是,对于数学精度要求较高的场合,肯德尔系数更有优势。在均方差,偏差等方面,肯德尔的表现比斯皮尔曼好,即如果数据集中噪音和异常数据较多,用肯德尔系数能取得较好的数学精度,偏差较小[15]。

(4) 对大规模数据来说,因为数据量比较大,一般推荐用计算复杂度较小的斯皮尔曼系数;但是,数据量越大一般来说噪音异常数据越多,影响系数精度,如果数据难于清洗而对计算精度要求较高则推荐用肯德尔系统;反之,用斯皮尔曼系数。

2 实验与分析

实验工具采用的是Matlab,各方法的Matlab命令和使用方法见表2。总体来说,距离在Matlab中用pdist命令,带不同类型参数;相关系数是用corr命令带不同类型参数。其中余弦夹角相关系数又称为余弦距离,杰卡德相关系数基于杰卡德距离计算,也被Matlab归为距离类。

需要注意的几处:

(1) 最后3个相关系数的输入是两个序列,序列是列向量;

(2) 其余方法的输入是一个矩阵,待对比排序序列是矩阵的两行,是行向量;

(3) 中间有两处需要进行1-pdist操作是因为Matlab本身在计算该值时候是用取的1减去本身值的结果,要原来结果就必须进行反向操作。

表2 所用Matlab命令及使用方式

方法	对比序列输入格式举例	命令
欧式距离	t12 = [1 2 3; 2 1 3]	pdist(t12,'euclidean')
曼哈顿距离	同上	pdist(t12,'cityblock')
契比雪夫距离	同上	pdist(t12,'chebychev')
余弦夹角相关系数	同上	1 - pdist(t12,'cosine')
杰卡德相关系数	同上	1 - pdist(t12,'jaccard')
皮尔森相关系数	t1 = [1;2;3], t2 = [2;1;3]	corr(t1,t2)
斯皮尔曼相关系数	同上	corr(t1, t2, 'type','spearman');
肯德尔相关系数	同上	corr(t1, t2, 'type', 'Kendall');

为了关注对比方法本身,选用非常简单的排序序列做例子,待比较的序列为以下3个

实验1:$t1$ = [1;2;3],$t2$ = [2;1;3],$t3$ = [1;3;2]

$t1$ 与 $t2$:两者的区别是1号节点和2号节点的排序不同,但1,2号节点在3号节点之前都是一样的,相对位置不变;3的序位没有变。

$t1$ 与 $t3$:两者的区别是2号节点和3号节点的排序不同,它们都在1号节点之后,相对位置不变,1的序位没有变。因此,$t1 \sim t2$ 与 $t1 \sim t3$ 改变的规律是一样的,其评价应该相同。

$t2$ 与 $t3$:$t2 \sim t3$ 的改变是1、3相对位置不变,但是整体从2的后面移动到了2的前面,即1,3与2之间的序位方向出现不同,导致3个元素的序位全部发生变化。该组的相关系数应该是最低、而距离最远的。但是1,3相对位置不变这点需要体现。

考察各方法具体给出值的区别:基于距离的实验结果如表3,基于相关系数的实验结果如表4。

表3 基于距离的对比方法结果

	汉明距离	欧式距离	曼哈顿距离	契比雪夫距离
$t1$ 与 $t2$	2	1.4142	2	1
$t1$ 与 $t3$	2	1.4142	2	1
$t2$ 与 $t3$	3	2.4495	4	2

表 4　基于相关系数的对比方法结果/%

	汉明相关系数	余弦夹角相关系数	杰卡德相关系数	皮尔森相关系数	斯皮尔曼相关系数	肯德尔相关系数
*t*1 与 *t*2	33.33	92.86	33.33	50	50	33.33
*t*1 与 *t*3	33.33	92.86	33.33	50	50	33.33
*t*2 与 *t*3	0	78.75	0	-50	-50	-33.33

对 *t*1 ~ *t*2,与 *t*1 ~ *t*3 的对比所有方法都给出了同样的评价,说明这些方法都反应一定的排序相似性,但是在最后 t2 ~ t3 的评价上区别较大,说明这些方法之间存在不同特性,使用于不同情况。

汉明距离与汉明相关系数: 二者联系紧密,一同分析。*t*2 ~ *t*3 的汉明距离为 3,即元素的排序序位完全不同,达到汉明巨鹿的最大值; 相对的相关系数降为 0。完全忽视了"1,3 相对位置不变"这点,只关心单个元素的排位,不能反应多个元素的相对位置。因此不适合用来对比排序序列。首先被排除出候选对比方案。与此类似,杰卡德相关系数,也被排除。

而余弦夹角考察的是两个向量的方向差异,对同方向的绝对数值不敏感。给出的相关系数过高,从实验来看前两个排序给出了 92.86% 的相关系数,对只有 3 个元素的排序来说这个值太高,不能很好反应不同排序的区别,也不建议使用。

剩下的 3 个距离——曼哈顿距离、欧氏距离和契比雪夫距离都受元素量纲影响,而皮尔森相关系数也受影响,只有斯皮尔曼和肯德尔相关系数不受影响,将 *t*1 ~ *t*3 元素的量纲改变如下

实验 2: *t*1 = [1; 10; 200]; *t*2 = [10; 1; 200]; *t*3 = [1; 200; 10]。

可以看到除了元素量纲有变化,元素 2 从单位 1 变为单位 10,元素 3 从单位 1 变为单位 100。而排序序位的改变和前面实验是一样的,对比的结果应该也一样才是需要的方法,实验结果如表 5。

表 5　调整量纲后各方法对比数据

	欧式距离	曼哈顿距离	契比雪夫距离	皮尔森相关系数/%	斯皮尔曼相关系数	肯德尔相关系数/%
*t*1 与 *t*2	200.4046	218	200	99.68	0.5	33.33
*t*1 与 *t*3	268.7006	380	190	-42.91	0.5	33.33
*t*2 与 *t*3	199.4543	218	199	-50	-0.5	-33.33

对比表 5 和表 3、表 4 的数据,可以很明显的看到,3 个距离受量纲影响明显。距离的绝对值和变化的百分比都与实验 1 差距甚远。在实验 1 中,*t*2 与 *t*1 距离1.4142,*t*3 与 *t*1 距离 2.4495,*t*2 与 *t*3 与基准相差距离为 1.0353; 而在实验 2 中,*t*2 与 *t*3 与基准相差距离为 68.296,相差巨大。其他距离请类似,受量纲影响巨大。相关系数中,皮尔森相关系数受影响较大,完全不符合排序变动的事实。不受影响的就只有斯皮尔曼和肯德尔相关系数。因此这两个方法是推荐方法。具体的区别在 1.3.2 中有详细叙述。

3　结束语

归纳总结常见的衡量序列之间关系的方法包括距离和相关系数两大类。原理分析和实验对比指出,距离方法和部分相关系数受原理或者元素量纲的影响,不适合应用在排序序列对比上面。最后推荐使用的是斯皮尔曼相关系数或者是肯德尔相关系数进行排序序列比较,两者都提供整体性评价,且都不受数据量纲影响。前者计算复杂度低,但是精算精度也低,适用于对精度要求不高的大规模数据集排序研究。例如: 社交网络的用户重要性排序,其用户数量巨大但用户排序的序位差距一两位影响不大; 后者,计算复杂度高,但是数学精度高、容忍度高,对有噪音异常或对精度有较高要求且数据规模有限的数据集适用。例如: 交通网络中的节点重要性排序研究,对比社交网络动辄上亿的用户量,交通网络节点数量有限,但是其节点排序的精度要求较高,序位的一位差距就可能造成决策的误差,因此建议使用精度较高的肯德尔相关系数,而不是斯皮尔曼相关系数。

参考文献:

[1] 丁雪峰,刘嘉勇,吴越,等. 基于 SNA 的网络舆论意见领袖识别研究 [J]. 高技术通讯,2011,(2): 167 - 172.

[2] 党亚茹,孟彩红. 基于复杂网络的航空货运枢纽城市研究 [J]. 交通运输工程与信息学报,2012,(2): 12 - 18.

[3] 刘建国,任卓明,郭强,等. 复杂网络中节点重要

性排序的研究进展[J]. 物理学报,2013,(17).
[4] 任晓龙,吕琳媛. 网络重要节点排序方法综述[J]. 科学通报,2014,(13):1175-1197.
[5] Derek J S,Robinson. An Introduction to Abstract Algebra [M]. Berlin&New York: Walter de Gruyter,2003:255-257.
[6] James E,Gentle. Matrix Algebra: Theory, Computations, and Applications in Statistics [M]. New York: Springer-Verlag,2007:299.
[7] Michel Marie Deza,Elena Deza. Encyclopedia of Distances [M]. New York: Springer,2009:104.
[8] Eugene F. Krause. Taxicab Geometry: An Adventure in Non-Euclidean Geometry [M]. New York: Dover Publications,Inc,1975:2-5.
[9] Cantrell C D. Modern mathematical methods for physicists and engineers [J]. Modern Mathematical Methods for Physicists & Engineers Cambridge Uk Cambridge University Press,2000,10(8):83.
[10] 王应明. 基于相关性的组合预测方法研究[J]. 预测,2002,(2):58-62.
[11] Jaccard,Paul. Étude comparative de la distribution florale dans une portion des Alpes et des Jura [J]. Bulletin de la Société Vaudoise des Sciences Naturelles ,1901,37: 547-579.
[12] FISHER R A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population [J]. Biometrika,1915,10 (4):507-521.
[13] Myers Jerome L,Well Arnold D. Research Design and Statistical Analysis [M](2nd ed.). Mahwah,NJ: Lawrence Erlbaum,2003:508.
[14] Kendall M G. A New Measure of Rank Correlation [J]. Biometrika, 1938, 30(3):81-93.
[15] Xu Weichao, Hou Yunhe, Hung Y S,et al. A Comparative Analysis of Spearmans Rho And Kendalls Tau in Normal and Contaminated Normal Models [J]. Signal Processing,2013,93(1):261-276.

The Research on Rank Series Comparison Methods

WANG Juan, SHI Lei
(College of Information Security Engineering, Chengdu University of Information Technology, Chengdu 610225,China)

Abstract: Ranking is abstraction of a lot of application problems. Rank comparison is necessary step of these researches. Most of existing rank series comparison methods are evaluating the sequence change of important nodes, lack of evaluation and comparison of the whole sequence. In order to address the problem, the existing comparison methods are summarized firstly, and they are mainly divided into two types: based on the distance and based on similarity. And then through the theory analysis and perimental comparison, the inherent vice of some methods and they are also affected by element dimension are found. Finally the spearman or Kendall correlation coefficient are recommended to do rank series comparison and the applicable scope of both are presented.

Key words: rank comparison; rank similarity; correlation coefficient; series distance