向量相似度测度方法

张 宇, 刘雨东, 计 钊

(中国船舶重工集团公司第七六〇研究所, 大连 116013)

摘要: 判断两个向量的相似程度, 一般采用计算相似度的方法。首先对相似度测度法中的几种成熟的计算方法和应用优缺点进行了综合分析和对比。然后, 根据实际的应用需要, 结合两种成熟的距离测度法的方法, 提出了一种改进的距离测度, 即相对误差距离; 又从放大较大相对误差作用的思路出发, 提出一个相似性函数, 即最大相异系数。这两种方法都充分考虑了相对变化量, 具有针对性, 适用于未知的待比较向量与确知的标准向量进行测度的情况。

关键词: 相似度; 距离测度; 相似性函数

中图分类号: TN11

文献标识码: A

文章编号: 1000-3630(2009)-04-0532-05

DOI 编码: 10.3969/j.issn1000-3630.2009.04.021

Vector similarity measurement method

ZHANG Yu, LIU Yu-dong, JI Zhao

(The 760 Research Institute of China Shipbuilding Industry Corporation, Dalian 116013, China)

Abstract: To estimate two vectors' similarity, the methods of computing the similarity are generally adopted. Several mature similarity estimate methods and their merits and drawbacks in applications are analyzed and compared in this paper. According to the demand of practice application and associating with two mature distance measure methods, an improved distance measure method called relative error distance is proposed. And, from the thought of magnifying the effect of the larger relative error, a comparability function called most dissimilarity coefficient is created. These two methods, which have fully considered the relative variation value, are applicable in the case of measuring an unknown comparative vector and a known standard vector.

Key words: similarity; distance measure; comparability function

1 引言

相似度, 顾名思义是指两个对象之间的相似程度。根据研究对象的不同, 出现了多种相似度的细化概念, 如向量相似度、系统相似度、形状相似度等等^[1-7]。相似度测度一般有两种方法: 距离测度法和相似性函数法。

本文首先对几种成熟相似度计算方法和应用优缺点进行了综合分析和对比。然后根据实际应用的特点, 即一个未知的待比较向量和一个确知的标准向量进行测度, 提出了改进的距离测度法和相似性函数。

2 距离测度

两向量的距离测度的方法有很多种, 一般两向量

的距离应满足下面的性质^[8,9]:

设向量 x 和 y 的距离为 $d(x, y)$, 则:

(1) $d(x, y) \geq 0$, 当且仅当 $x=y$ 时, 等号成立;

(2) $d(x, y) = d(y, x)$;

(3) $d(x, y) \leq d(x, z) + d(z, y)$ 。

其中: $x = (x_1, x_2, x_3, \dots, x_n)'$,

$y = (y_1, y_2, y_3, \dots, y_n)'$ 。

2.1 几种成熟的距离测度法

(1) Minkowsky 距离(明氏距离):

$$d(x, y) = \left(\sum_{i=1}^N |x_i - y_i|^m \right)^{1/m} \quad (1)$$

它是距离的通用形式^[10]。

(2) Euclidean 距离(欧氏距离):

$$d(x, y) = \left(\sum_{i=1}^N |x_i - y_i|^2 \right)^{1/2} \quad (2)$$

欧氏距离是明氏距离当 $m=2$ 时的特殊形式, 也是应用较多的一种距离。当 y 是 0 向量时, $d(x)$ 表示通常意义下的长度, 定义为: $d(x) = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}$ 。

(3) Manhattan 距离:

收稿日期: 2008-07-09; 修回日期: 2008-10-08

作者简介: 张宇(1983-), 男, 黑龙江齐齐哈尔人, 硕士研究生, 研究方向为水声工程。

通讯作者: 张宇, E-mail: zhangyuzy@163.com

$$d(x, y) = \left(\sum_{i=1}^N |x_i - y_i| \right) \quad (3)$$

Manhattan 距离是明氏距离当 $m=1$ 时的特殊形式。

(4) Jffreys & Matusita 距离:

$$d(x, y) = \left[\sum_{i=1}^N (\sqrt{x_i} - \sqrt{y_i})^2 \right]^{1/2} \quad (4)$$

(5) Mahalanobis 距离(马氏距离):

$$d(x, y) = \sqrt{(x-y)' \Sigma^{-1} (x-y)} \quad (5)$$

式中, Σ 表示 x 和 y 的协方差矩阵。

(6) Camberra 距离(Lance 距离、Williams 距离):

$$d(x, y) = \sum_{i=1}^n \frac{|x_i - y_i|}{|x_i + y_i|} \quad (6)$$

2.2 以上几种距离测度法的特点

Minkowsky 距离是距离的通用形式, Euclidean 距离和 Manhattan 距离都是其特殊形式。其中, Manhattan 距离运算量较低, 简单明了, 对于向量中的每个元素的误差都同等对待。而 Euclidean 距离在一定程度上放大了较大元素误差在距离测度中的作用, 被各个领域广泛应用。Jffreys 距离是在 Euclidean 距离的基础上放大了较小元素误差在距离测度中的作用, 对 Euclidean 距离有所修正。Camberra 距离做了自身的标准化, 考虑到了元素误差占本身的比重, 特别适合于高度偏倚的数据。Mahalanobis 距离与上述几种距离不同, 它考虑到了向量中各个元素之间的相关性。

如果假定向量中各个元素相互独立, 即向量的协方差矩阵是对角矩阵, 则 Mahalanobis 距离就退化为 Euclidean 距离, 因此, Mahalanobis 距离又称为广义 Euclidean 距离^[11]。

2.3 相对误差距离

根据向量相互独立的特点, 结合 Euclidean 距离和 Camberra 距离的思路, 提出一种距离测度方法, 即相对误差距离法。

思路: Camberra 距离考虑到了元素差值占自身的比重, 根据这一思路, 给出待比较元素 x_i 与标准元素 y_i 的相对误差 $\frac{x_i - y_i}{y_i}$, 再参考 Euclidean 距离的思路, 求相对误差的长度。计算表达式为:

$$d(x, y) = \left[\sum_{i=1}^n \left(\frac{x_i - y_i}{y_i} \right)^2 \right]^{1/2} \quad (7)$$

其中: 向量 x 表示未知的待比较向量, 向量 y 表示确知的标准向量。

这种方法放大了较大相对误差的作用, 适用于未知的待比较向量与确知的标准向量进行测度的

情况。由于在计算过程中对比较向量和标准向量不同对待, 所以不满足距离性质中的对称性。

3 相似性函数

3.1 几种成熟相似性函数

相似性函数是用函数的方法来表征两向量相似的程度, 根据向量中元素的不同, 相似性函数可分为二元向量的相似性函数和一般向量的相似性函数。

3.1.1 二元向量的相似性函数

也可称作匹配测度。只有两个状态 (0-1) 的变量称为二元变量。由二元变量组成的向量叫做二元向量。二元变量分为对称的和非对称的两种。如果二元变量的两个状态是同等价值的, 具有同样的权重, 那么称其为对称的二元变量, 否则称为不对称的二元变量。设 x 和 y 为二元向量, $x = (x_1, x_2, x_3, \dots, x_n)'$, $y = (y_1, y_2, y_3, \dots, y_n)'$, 则这两个向量匹配数目的定义为:

$$a = \sum_{i=1}^n x_i y_i; \quad b = \sum_{i=1}^n y_i (1 - x_i);$$

$$c = \sum_{i=1}^n x_i (1 - y_i); \quad d = \sum_{i=1}^n (1 - x_i)(1 - y_i)$$

其中: a 为向量 x 和 y 的 (1-1) 匹配的数目, b 为向量 x 和 y 的 (0-1) 匹配的数目, c 为向量 x 和 y 的 (1-0) 匹配的数目, d 为向量 x 和 y 的 (0-0) 匹配的数目。

一般二元向量的匹配测度方法有:

(1) 简单匹配系数(SMC):

$$sim(x, y) = \frac{a+d}{a+b+c+d} \quad (8)$$

简单匹配系数的分子为 (1-1) 和 (0-0) 匹配的数目之和, 分母为所有数目之和。用于分析对称的二元向量。

(2) Jaccard 系数(Tanimoto 系数):

$$sim(x, y) = \frac{a}{a+b+c} \quad (9)$$

Jaccard 系数无论是分子还是分母都完全不考虑 (0-0) 的匹配, 常用来处理非对称的二元向量。

(3) Dice 系数:

$$sim(x, y) = \frac{2a}{2a+b+c} \quad (10)$$

Dice 系数也是分子、分母都不考虑 (0-0) 的匹配, 同时对 (1-1) 匹配进行加权, 体现了对 (1-1) 匹配的重视。

3.1.2 一般向量的相似性函数

一般向量的相似性函数较距离测度应用更为

广泛,如夹角余弦法、相关系数法、广义 Dice 系数法、广义 Jaccard 系数法等。

(1) 夹角余弦法

夹角余弦用来度量两组向量之间夹角的大小,亦称为相和系数,表达式为:

$$\text{sim}(x, y) = \cos(x, y) = \frac{(x, y)}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^n x_i \cdot y_i}{\left(\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i^2\right)^{\frac{1}{2}}} \quad (11)$$

夹角余弦的几何意义是在由 N 个元素组成的 N 维空间中,表征两个向量之间夹角的余弦值。一般在使用前需要对向量中的各元素进行无量纲化处理,使各元素都为正,这时夹角余弦的取值范围为 $[0, 1]$,取值越大表明两向量夹角越小,两者越接近,值为 1 时,两向量完全相同。另外,夹角余弦规范化了向量的长度,这意味着在计算相似度时,不会放大数据对象重要部分的作用^[12]。

(2) 相关系数法

$$\text{sim}(x, y) = r(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{cov}(x, x) \cdot \text{cov}(y, y)}} \quad (12)$$

$$\text{其中: } \text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y});$$

$$\text{cov}(x, x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2;$$

$$\text{cov}(y, y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2;$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i; \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i。$$

所以相关系数可以表示为:

$$\text{sim}(x, y) = r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2\right)^{\frac{1}{2}}} \quad (13)$$

相关系数是多元统计学中用来衡量两组变量之间线性密切程度的无量纲指标,取值范围为 $[-1, 1]$ 。分为正相关、不相关和负相关 3 类情况。通常在计算时,还要经过一定的处理,将负相关与正相关合并到一起,这时它的取值范围为 $[0, 1]$,值越大相关性越强,当值为 1 时,两向量完全相同。从式(13)可以看出相关系数是中心化的夹角余弦,性质与夹角余弦相似。

(3) 广义 Dice 系数法^[13]

$$\text{sim}(x, y) = \frac{2 \sum_{i=1}^n x_i \cdot y_i}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2} \quad (14)$$

广义 Dice 系数是对前面提到的 Dice 系数的应

用推广,它与夹角余弦有些相似之处,两种方法分子相同,区别在于广义 Dice 系数的分母用的是两个向量长度平方的算术平均: $\frac{d^2(x) + d^2(y)}{2}$,而夹角余弦的分母是几何平均: $\sqrt{d^2(x) \cdot d^2(y)}$ 。

(4) 广义 Jaccard 系数法

$$\text{sim}(x, y) = \frac{\sum_{i=1}^n x_i \cdot y_i}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_i \cdot y_i} \quad (15)$$

广义 Jaccard 系数与上面的广义 Dice 系数很相近,是在广义 Dice 系数的分子和分母同时减去两向量的内积,性质与 Dice 系数也很相近。

以上几种一般向量的相似性函数,都是在夹角余弦的基础上的演变,在应用上夹角余弦和相关系数法应用较多。

3.2 最大相异系数

为了判断一个未知的待比较向量和一个确知的标准向量的相似度,提出了最大相异系数法。它的计算公式、原理和特点如下:

设 x 和 y 分别为未知的待比较向量和确知的标准向量。首先,定义“相对误差向量”: $\lambda_i = \left| \frac{x_i - y_i}{y_i} \right|$; 然后,将向量 λ 从大到小排列重组为: $\lambda = \{\lambda_i \geq \lambda_j, i < j\}$,取出向量 λ 的前 k 个值: $\xi = \{\lambda_1, \lambda_2, \lambda_3 \dots \lambda_k\}$; 最后,对向量 ξ 加权平均,得到最大相异系数:

$$z_y = \frac{\sum_{j=1}^k (k-j+1) \xi_j}{\sum_{j=1}^k j} \quad (16)$$

原理:首先,“相对误差向量”的选定是考虑到待比较元素与标准元素的绝对误差的大小不能明确地表明它对相似度测度的影响,例如绝对误差为 2,对于标准元素分别为 100 和 1 两种情况的影响差异很大,如果标准元素是 100,那么误差几乎可以忽略;如果标准元素是 1,那么误差已经成为它的 2 倍,误差非常大,因此元素误差的影响不仅与绝对误差有关,还与标准元素的大小有关,所以采用了相对量;然后,在 n 个元素中选出最大 k 个,由于相对误差小、可以忽略的元素在判决过程中起不到实质的作用,只能为相似度提供基数,所以选取其中影响最大的 k 个;最后,最大相异系数是一个加权平均的过程,采取了不同的加权系数,相对误差越大采用的加权系数就越大,突出了相对误差大的元素对判决的影响。

最大相异系数 z_y 的取值范围为 $[0, \infty)$, z_y 越小则

两向量越相近, 当且仅当 $z_y=0$ 时, 两向量完全相同。

在 3.1 节中所述的一般相似性函数应满足以下性质^[14]:

- (1) $0 < sim(x, y) \leq 1$, 当且仅当 $x=y$ 时, 等号成立;
- (2) $sim(x, y) = sim(y, x)$ 。

这是一般情况下, 计算两个向量的相似度应满足的性质, 计算过程中对两个向量没有偏重, 同等对待, 而本文要解决的是判断一个待比较的未知向量和一个确知的标准向量的相似度, 两个向量不是同等条件的, 文中的最大相异系数法正是根据这一应用特点而提出的, 因此不满足常规的相似性函数的性质。

4 向量相似度方法对比

为了方便各种相似度方法的对比, 现将相似度方法名称、参考变量和优缺点分析列于表 1 中。

5 结语

本文对几种成熟的向量相似度测度方法进行了分析和归纳, 其中包括六种距离测度法、三种针对二元向量的相似性函数和四种针对一般向量的相似性函数, 这些方法从独立到相关, 从长度到角度, 从二元到多元, 全方位多侧面的描述了向量相似度。

根据实际应用的需要, 结合成熟相似度方法的思路, 提出了相对误差距离和最大相异系数, 这两种方法都是从相对误差的概念出发, 放大较大相对误差的作用, 在判断一个未知的待比较向量和一个确知的标准向量的相似度上有其独特的作用和意义。

参 考 文 献

[1] 焦利明, 于伟, 罗均平, 等. 向量相似度在雷达目标识别中的应用[J]. 火控雷达技术, 2006, 35(2): 78-81.
 JIAO Limin, YU Wei, LUO Junping, et al. Application of vector resemblance degree in radar target recognition[J]. Technical Fire Control Radar, 2006, 35(2): 78-81.

[2] 田晓东, 刘忠. 基于形状相似度的水下目标识别算法[J]. 声学技术, 2007, 26(3): 493-497.
 TIAN Xiaodong, LIU Zhong. Mine target recognition based on shape similarity[J]. Technical Acoustics, 2007, 26(3): 493-497.

[3] 王婷. 本体相似度研究[J]. 电脑知识与技术, 2007, 32(14): 1609-1611.
 WANG Ting. A survey on ontology similarity[J]. Computer Knowledge and Technology, 2007, 32(14): 1609-1611.

[4] 许瑞丽, 徐泽水. 区间数相似度研究[J]. 数学的实践与认识, 2007, 37(24): 1-8.
 XU Ruili, XU Zhesui. On similarity degrees of interval numbers[J]. Mathematical Practice and Cognition, 2007, 37(24): 1-8.

表 1 向量相似度方法对比
 Table 1 Comparison between vector similarity methods

方法名称	参考变量	优缺点分析
Minkowsky 距离	一般向量	距离的通用形式。
Euclidean 距离	一般向量	应用广泛的距离形式, 一定程度上放大了较大元素误差在距离测度中的作用。
Manhattan 距离	一般向量	简单明了, 对向量中的每个元素同等对待。
Jffreys & Matusita 距离	一般向量	在 Euclidean 距离的基础上, 放大了较小的元素误差的作用, 对 Euclidean 距离有所修正。
Mahalanobis 距离	一般向量	又称广义 Euclidean 距离, 考虑到了向量中各个元素之间的相关性。
Camberra 距离	一般向量	做了自身的标准化, 考虑到了元素误差占本身的比重, 特别适合于高度偏倚的数据。
简单匹配系数	二元向量	用于分析对称的二元向量。
Jaccard 系数	二元向量	不考虑(0-0)的匹配, 常用来处理非对称的二元向量。
Dice 系数	二元向量	不考虑(0-0)的匹配, 同时对(1-1)匹配进行加权, 体现了对(1-1)匹配的重视, 常用来处理非对称的二元向量。
夹角余弦	一般向量	表征两个向量之间夹角的余弦值, 规范化了向量的长度。但在计算相似度时, 不会放大数据对象重要部分的作用。
相关系数	一般向量	相关系数是多元统计学中用来衡量两组变量之间线性密切程度的无量纲指标, 是中心化的夹角余弦, 性质与夹角余弦相似。
广义 Dice 系数	一般向量	与夹角余弦有些相似, 改两向量长度平方的几何平均为算术平均。
广义 Jaccard 系数	一般向量	广义 Dice 系数的基础上, 分子分母同时减去两向量的内积。
相对误差距离	一般向量	结合 Euclidean 距离和 Camberra 距离的优点, 放大了较大相对误差的作用, 适用于待比较向量与标准向量进行测度的情况。
最大相异系数	一般向量	针对待比较向量和标准向量做相似性判断的特点, 使用相对误差、加权、平均等手段, 放大了较大相对误差的作用, 同时减少了较小相对误差的基数作用。

- [5] 戚晓明, 陆桂华, 吴志勇等. 水文相似度及其应用[J]. 水利学报, 2007, **38**(3): 355-360.
QI Xiaoming, LU Guihua, WU Zhiyong et al. Hydrology similarity degree and its application[J]. Hydrotechny Transaction, 2007, **38**(3): 355-360.
- [6] Inderjeet Mani, Eric Bloedorn. Summarizing similarities and differences among related documents[J]. Information Retrieval, 1999, **1**: 35-67.
- [7] 郑丽萍, 李光耀, 梁永全, 等. 本体中的概念相似度的计算[J]. 计算机工程与应用, 2006, **30**: 25-27.
ZHENG Liping, LI Guangyao, LIANG Yongquan, et al. The computation of conceptual similarity in ontology[J]. Computer Engineering and Application, 2006, **30**: 25-27.
- [8] 孙即祥. 现代模式识别[M]. 长沙: 国防科技大学出版社, 2002.
SUN Jixiang. Modern patten recognition[M]. Changsha: National Defense Science and Technology University Press, 2002.
- [9] 数据挖掘导论 [DB/OL]. <http://book.csdn.net/bookfiles/327>
Introduction of data mining[DB/OL]. <http://book.csdn.net/bookfiles/327>
- [10] 沈清, 汤霖. 模式识别导论[M]. 长沙: 国防科技大学出版社, 1991.
SHEN Qing, TANG Lin. Introduction of patten recognition[M]. Changsha: National defense Science and Technology University Press, 1991.
- [11] 赵艳玲. 舰船噪声异常判断[D]. 大连: 中船重工第 760 研究所, 2008: 23-24.
ZHAO Yanling. Ship noise abnormal judgment[D]. Dalian: The 760 Research Institute of CSIC, 2008: 23-24.
- [12] TIAN Runtao, XIE Peishan. Study on the standardization of similarity evaluation method of chromatographic fingerprints(Part I)[J]. Traditional Chinese Drug Research & Clinical Pharmacology, 2006, **17**(1): 40-42.
- [13] 陶跃华. 基于向量的相似度方案[J]. 云南师范大学学报, 2001, **21**(5): 17-19.
TAO Yuehua. Vector-based computing scheme of similarity[J]. Journal of Yunnan Normal University, 2001, **21**(5): 17-19.
- [14] 程勇, 黄河, 邱莉榕, 等. 一个基于相似度计算的动态多维概念映射算法[J]. 小型微型计算机系统, 2006, **27**(6): 975-979.
CHENG Yong, HUANG Hei, QIU Lirong, et al. Similarity-based dynamic multi-dimension concept mapping algorithm[J]. Mini-micro systems, 2006, **27**(6): 975-979.