高维数据的相似性度量研究

贺 玲[1] 蔡益朝[1] 杨 征[2]

(空军雷达学院预警监视情报系 武汉 430019)[1] (国防科技大学信息系统与管理学院 长沙 410073)[2]

摘 要 数据间的相似性度量是进一步分析数据集整体特性的一个重要基础。针对高维数据的相似性度量问题,提出了一种基于子空间的相似性度量方法。该方法先将高维空间进行基于网格的划分,然后在划分后的子空间内计算数据间的相似性。理论分析表明,在合理选定网格划分参数的前提下,该方法可有效减小"维度灾难"对高维数据相似性度量的影响。

关键词 高维数据,维度灾难,网格划分,子空间,相似度量

中图法分类号 TP311 **文献标识码** A

Researches on Similarity Measurement of High Dimensional Data

HE Ling[1] CAI Yi-chao[1] YANG Zheng[2]

(Department of Early Warning Surveillance Intelligence, Air Force Radar Acdemy, Wuhan 430019, China)[1]

(Information System and Management Department, National University of Defense Technology, Changsha 410073, China)[2]

Abstract The similarity measurement among data is important for further analysis of the data set. Aiming at the similarity measurement of high dimensional data, the paper put forward a new method based on subspace. After dividing high dimensional space into grids, and computing the similarity among data in proper subspaces, the disturbance from the curse of dimensionality can be abated efficiently under the proper dividing parameters.

Keywords High dimensional data, Curse of dimensionality, Grid-based dividing, Subspace, Similarity measurement

1 引言

在很多涉及到高维数据的研究领域,数据间的相似性度量是一个重要的、关键性的问题。而传统的度量方式在应用于高维数据时,由于不能有效消除"维度灾难"的影响,往往无法得到准确的度量结果。

以经典的 L_k 度量为例,文献[1]的研究结果表明,L_k 等传统度量方式在高维空间中的度量结果是不稳定的。设在 d 维空间中,$X_i^d(1\leqslant i\leqslant n)$为服从某种数据分布 $FData^d$ 的 n 个数据点,$\| X_i^d \|$ 为通过某种度量方式得到的 X_i^d 到坐标原点的距离,所有数据点到坐标原点的最近和最远距离分别为 $D_{\min}^d$ 和 $D_{\max}^d$,对于某随机向量 X,其均值和方差分别记为 $E(X)$和 $var(X)$,有以下定理成立[1]。

定理 1

如果 $\lim_{d\to\infty}\mathrm{var}\left(\frac{\| X_i^d \|}{E(\| X_i^d \|)}\right)=0$,那么 $\frac{D_{\max}^d-D_{\min}^d}{D_{\min}^d}\to_p 0$。

该定理表明,在高维空间中,查询点(这里取为原点)的最近邻距离与最远邻距离之间的差异非常小,也就是说,查询点到空间中所有点的距离都几乎相同,于是,"最近邻"的概念没有任何意义。其原因在于,L_k 等传统度量总是在整个数据空间中考虑数据之间的相似特性。而在高维空间中,即使是最为相似的两个数据对象,它们的属性值在某一些数据维上仍然存在较大的差别,但在 L_k 等传统度量方式中,占主导地位的就是这些维,它们对数据对象之间的相似性计算产生了很大的干扰,这些维可称作噪声维。在这样的情况下,如果用 L_k 等传统度量方式来计算高维数据间的相似性,数据对象之间的相似信息通常会被淹没在少数的几个噪声维中,这最终导致了传统度量方式在高维空间中的不稳定性。据此,当数据维数很高时,可以从高维数据空间的子空间出发,在高维数据的某些主要维度上探讨数据间的相似性[2]。

以此为出发点,本文提出了一种基于子空间的相似性度量方法,试图在高维数据空间的子空间中度量数据间的相似性,以消除维度灾难对高维数据相似性度量的影响。

2 基于子空间的相似性度量方法

设 $x=(x_1,x_2,\cdots,x_d)$,$y=(y_1,y_2,\cdots,y_d)$为 d 维数据空间中的两个数据对象,S 为数据空间的某个子空间,S 的维数 $d(S)<d$,则 x,y 在 S 中的相似性可以通过以下方法来计算:

$$SubDist^{(S)}(x,y)=\sum_{i=1}^{d(S)}Dist^{(i)}(x,y) \quad (1)$$

式中,$Dist^{(i)}(x,y)$表示在 S 的第 i 维上,x,y 之间的相似度。

式(1)所表达的核心思想在于,用数据对象 x,y 在子空间 S 中的相似性来表达它们在原始高维空间中的相似性。因此为了获得合理、准确的相似性度量值,子空间 S 的确立是一个关键的前提。

本文接下将分别介绍如何寻找子空间 S,以及在子空间

到稿日期:2009-07-16 返修日期:2009-09-28 本文受国家自然科学基金(60802080)资助。

贺 玲(1976-),女,博士,讲师,主要研究方向为多媒体信息系统、数据挖掘,E-mail:zjj177@qq.com;蔡益朝(1976-),博士,讲师,主要研究方向为仿真、智能决策技术;杨 征(1978-),博士,讲师,主要研究方向为虚拟现实技术。

S 中怎样计算数据间的相似度。

2.1 一种基于网格的子空间生成方法

利用式(1)计算高维数据对象之间的相似性时，在子空间的选择上首先应将噪声维排除在子空间之外，从而有效消除噪声信息对相似度计算的干扰。基于此，本文选取子空间的方法是，将两个数据对象的属性值相差较大的属性维视为噪声维，从而在那些由数据点的属性值相对靠近的维度构成的子空间中计算它们之间的相似性。

为了更加合理地判断两个数据点之间的属性值是否接近，从而找到相似度计算的子空间，对高维数据空间的各个维度进行区间划分，以形成数据空间的网格结构。选择式(1)中的子空间 S 时，只有当两个数据对象在某一维度上的属性值落入相同或者相邻的区间时，才将该维度加入到 S 中。

在对数据空间的各个维度进行划分时，为了更好地反映数据对象的整体分布特征，采用了等深度的划分方法，即：使得每一区间中落入的数据对象的个数相同。设数据空间的维数为 d，数据集中数据对象的个数为 N，每一个数据对象分别表示为 $x_k(1\leqslant k\leqslant N)$，数据空间的每一维被划分为 m 个连续的区间，则每一区间中落入的数据点的个数为 N/m。如果在数据空间的第 i 维上，对数据集中的各个数据对象按其属性值从小到大的顺序进行排序，排序后数据点 x_k 的序号记为 $Order^{(i)}(x_k)$，用 R_{ij} 表示第 i 维上的第 j 个区间并用 $L_{R_{ij}}$ 和 $U_{R_{ij}}$ 分别表示该区间的下界和上界，则有：

$$\begin{aligned} L_{R_{ij}}&=x_k^{(i)}:Order^{(i)}(x_k)=\left\lfloor\frac{j*N}{m}+1\right\rfloor\\ U_{R_{ij}}&=x_k^{(i)}:Order^{(i)}(x_k)=\left|\frac{(j+1)*N}{m}\right| \end{aligned}\tag{2}$$

式中，$x_k^{(i)}$ 表示数据点 x_k 在数据空间第 i 维上的属性值。

任给 d 维数据空间中的两个数据对象 x 和 y，则在 d 维数据空间的任意一个维度 φ_i 上，x 在 φ_i 上的属性值 $x^{(i)}$ 所在的区间序号 $\eta^{(i)}(x)$ 为：

$$\eta^{(i)}(x)=j:L_{R_{ij}}\leqslant x^{(i)}\leqslant U_{R_{ij}}\tag{3}$$

按同样的方法，可以计算 y 在 φ_i 上的属性值 $y^{(i)}$ 所在的区间序号 $\eta^{(i)}(y)$。对于 x 和 y 而言，它们的属性值落在同一区间的维度构成的集合为：

$$S_1=\{\varphi_i\mid\eta^{(i)}(x)=\eta^{(i)}(y)\}\tag{4}$$

除此之外，设 x 和 y 的属性值落在相邻区间且彼此靠近的维度构成的集合为式(5)：

$$S_2=\left\{\varphi_i\left|\ |(\eta^{(i)}(x)-\eta^{(i)}(y)|=1)\wedge\left(|x_i-y_i|<\frac{|U_{R_{i\eta^{(i)}(x)}}-L_{R_{i\eta^{(i)}(x)}}|+|U_{R_{i\eta^{(i)}(y)}}-L_{R_{i\eta^{(i)}(y)}}|}{2}\right)\right.\right\}\tag{5}$$

那么，本文认为，在 S_1 和 S_2 所包含的各个维度上，x 与 y 的属性值相对较为接近。

2.2 基于子空间的相似性度量

在上述网格划分的基础上，利用 $SubDist^{(S)}()$ 度量 x 与 y 之间的相似性时，子空间 S 取为：

$$S=S_1\cup S_2\tag{6}$$

在此基础上，对于式(1)中 x 和 y 在每一维度 φ_i 上的相似度 $Dist^{(i)}(x,y)$，其计算方法见式(7)：

$$Dist^{(i)}(x,y)=\begin{cases}\dfrac{|x^{(i)}-y^{(i)}|}{|U_{R_{i\eta^{(i)}(x)}}-L_{R_{i\eta^{(i)}(x)}}|}, & \text{if } \varphi_i\in S_1\\ \dfrac{|x^{(i)}-U_{R_{i\eta^{(i)}(x)}}|}{|U_{R_{i\eta^{(i)}(x)}}-L_{R_{i\eta^{(i)}(x)}}|/2}+\dfrac{|y^{(i)}-L_{R_{i\eta^{(i)}(y)}}|}{|U_{R_{i\eta^{(i)}(y)}}-L_{R_{i\eta^{(i)}(y)}}|/2}, & \text{if } (\varphi_i\in S_2)\wedge(x^{(i)}<y^{(i)})\\ \dfrac{|x^{(i)}-L_{R_{i\eta^{(i)}(x)}}|}{|U_{R_{i\eta^{(i)}(x)}}-L_{R_{i\eta^{(i)}(x)}}|/2}+\dfrac{|y^{(i)}-U_{R_{i\eta^{(i)}(y)}}|}{|U_{R_{i\eta^{(i)}(y)}}-L_{R_{i\eta^{(i)}(y)}}|/2}, & \text{if } (\varphi_i\in S_2)\wedge(x^{(i)}\geqslant y^{(i)})\end{cases}\tag{7}$$

由式(7)可见，在计算 x 和 y 在某一维度 φ_i 上的相似性时，都除以了该维度上对应区间的长度。这样处理的原因在于，数据对象在各个维度上通常并不是均匀分布的，越是在数据对象密集分布的区间，数据之间的间隔就越小，区间长度则越短；反之，数据之间的间隔就越大，区间的长度也越长。因此以数据对象在某维上属性值间隔的大小与对应区间的长度的比值来衡量它们在该维上的相似性，更能体现出数据集的统计分布特性，从而使得度量的结果更加合理。

综合式(2)一式(7)，即可得到式(1)式所示的基于子空间的度量方式的一种具体实现方法。与传统的度量方法相比，该相度的计算不是在高维的全空间中进行的，而是在由某些特定维构成的低维的子空间中进行的，因此将具有更大的稳定性。另外，它有效地消除了噪声信息对度量的影响，从而也更加符合人们判定数据对象之间相似性的习惯，具有更强的合理性。

2.3 网格划分参数的确定

对于上述度量方式而言，网格划分参数 m 是影响度量性能的重要因素。合理的 m 值能够使得度量的结果更加准确。通常而言，均匀分布的数据更容易受到高维空间中相似性度量不稳定性的影响。接下来，本文以随机均匀分布的数据为例，分析 m 的取值准则。

在 d 维空间中，设 x 和 y 为从均匀分布的数据集中随机抽取的两个数据对象，定义 0-1 随机变量 B_i 为：

$$B_i=\begin{cases}1, & \text{if } \eta^{(i)}(x)=\eta^{(i)}(y)\\ 0, & \text{if } \eta^{(i)}(x)\neq\eta^{(i)}(y)\end{cases}\tag{8}$$

即 x 和 y 在数据空间的第 i 维上的属性值如果落入同一区间，则 B_i 取值为 1，否则取值为 0。由于 x 和 y 服从均匀分布，因此，B_i 取值为 1 的概率为 $1/m$，由此得到随机变量 B_i 的均值为 $1/m$，方差为 $(1/m)\cdot(1-1/m)$。设 x 和 y 的属性值落入同一区间的维数为 L，则有：

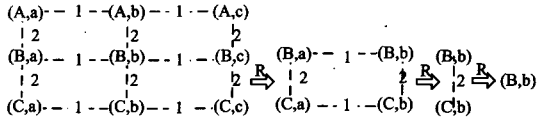
$$L=\sum_{i=1}^{d}B_i\tag{9}$$

由于数据对象在整个数据空间的各个维上都服从均匀分布，因此，所有的 B_i 独立同分布。在这种情况下，当数据空间的维度 d 很大时，根据中心极限定理[3]，随机变量 L 将呈正态分布，其均值为 d/m，标准偏差为 $\sqrt{(d/m)\cdot(1-1/m)}$。

此外，在数据均匀分布的前提下，定理 1 的前提等价于 $\lim\limits_{d\to\infty}\dfrac{\sqrt{(d/m)\cdot(1-1/m)}}{d/m}=0$，即当 $\lim\limits_{d\to\infty}\dfrac{\sqrt{(d/m)\cdot(1-1/m)}}{d/m}$ 取值为 0 时，对应的相似性度量函数是不稳定的。因此，$\lim\limits_{d\to\infty}\dfrac{\sqrt{(d/m)\cdot(1-1/m)}}{d/m}$ 的取值越小时，定理 1 的前提越容易满足。于是为了保证相似性度量的稳定性，必然要求上述极限

(下转第 227 页)

在公开宣告逻辑中重复宣告某一理性可以找到相应的博弈算法求到的博弈解。不断宣告的过程会消除那些理性不成立的世界，这种认知模型不断更新的过程正好和某一博弈算法求解过程一致。这里通过分析例1来展示这种对应性。其中，$R=SR_1 \wedge SR_2$。



重复宣告 R 3次之后，对 R 的任何再次宣告都不会使认知子模型继续收缩。可以清楚地看到这一认知模型的更新过程和图1展示的过程是一样的。

结束语 本文主要讨论策略博弈解。实际上，很多扩展博弈的算法也需要对个体理性做某些假设。文献[9]非形式地分析了一些算法所需要的理性假设。其他学者也在进行这方面的研究，这也是我们进一步研究的方向。

关于理性，如文中分析的那样，一个人是强理性的那么他一定是弱理性，反之不成立。这和文献[9]的说法是一致的。不同的算法需要关于认知和理性做不同强度的假设。文献[8]根据重复删除弱被占优策略算法[5]定义了另外一种理性(R)。正如他们所展示的那样，在以SR为理性前提可求得博弈解的那些博弈中，有些是不能够在以R为理性前提求得博弈解的；反之亦然。同时，对于以WR为理性前提可求得博弈解的那些博弈，以SR或以R为理性前提则均可求得；反之不成立。

文中提出命题1：在一个有限博弈 G 中，参与者关于彼此对理性的认识程度最多需要嵌套($\sum_{i\in N}|S_i|-|N|$)次。文中没有给出严格的形式证明，只是给出非形式出分析说明。但它是不难证明的。文中也提到命题1正是定义3中算法的认知基础。其实在定义3中为 m 设置这个最大值的另一个目的是为了避免这个算法使计算机运行陷于死循环(dead loop)的状态。所以在定义3中，给 m 设定了一个最大值($\sum_{i\in N}|S_i|-|N|$)。

我们可以通过计算机中的"while do"等语句很容易实现IDRDS程序算法在计算机中的执行。如果要进一步考虑系统复杂性的问题，也许可以增加一些算子，如测试算子(test)"?"来检验 S_i^m 或 RD_i^m 的元素个数是否满足某一要求，以便减少或避免无效的重复执行，从而达到减小时间复杂度的目的。在以后的程序设计工作中，我们会朝着减弱复杂度的方向努力。

参考文献

[1] Brandenburger. Epistemic Game Theory: Complete Information 2007[M]. The New Palgrave Dictionary of Economics(2nd edition) edited by Steven Durlauf and Lawrence Blume, Palgrave Macmillan, 2008

[2] Brandenburger A, Friedenberg A. Finite-Order Rationality [OL]. 2009. http://pages.stern.nyu.edu/~abranden/papers.html

[3] Baltag A, Moss L S, Solecki S. The logic of public announcement, common knowledge and private suspicious [R]. SEN-R9922, CWI, Amsterdam, 1999

[4] Lorini E, Schwarzentruber F, Herzig A. Epistemic game in model logic: joint action, knowledge and preferences all together[C]// Logic Rationality and Interaction. LORI-II. Springer Press, 2009

[5] Bonanno G. Epistemic foundations of game theory[R]. LLC, University of Amsterdam, 2007

[6] Bonanno G. A syntactic approach to rationality in game with ordinal payoffs, volume 3[M]. Amstrdam: Amsterdam University Press, 2008

[7] van Ditmarsch H, van der Hoek W, Kooi B. Dynamic Epistemic Logic[M]. Berlin: Springer, 2007

[8] Cui J Y, Guo M Y, Tang X J. Characterizations of iterated adimissiblity based on PEGL[C]// Logic Rationality and Interaction, Chongqing. LORI-II. Springer Press, 2009

[9] van Benthem J. Rational dynamics and epistemic logic in games [J]. Game and Economic Behavior, 2006

[10] van Benthem J. Exploring Logical Dynamics [M]. Stanford, Mass: ESLI Publications, 1996

[11] Apt K R. The many faces of rationalizability [J]. The Journal of Theorrtical Economics, 2007, 7

[12] Apt K R, Zvesper J A. Common belief and public announcements in strategic games with arbitrary strategy sets[J]. Journal of Logic and Computation, 2007. http://arxiv.org/PS_cache/arxiv/pdf/0710/0710.3536v2.pdf

[13] Osborne M, Rubinstein A. A Course in Game Theory[M]. Cambridge (Mass.): MIT Press, 1994

[14] Battigalli P, Bonanno G. Recent results on belief, knowledge and the epistemic foundations of game theory[J]. Research in Economics, 1999, 53: 149-225

(上接第156页)

值大于0，即应当满足：

$$\lim_{d\to\infty}\frac{\sqrt{(d/m)\cdot(1-1/m)}}{d/m}=\lim_{d\to\infty}\sqrt{\frac{m-1}{d}}>0 \qquad (10)$$

而且，上述极限值不能太小。为了达到这一目的，由式(10)可见，m 的取值不能过低。

但 m 的取值也不能过高，因为过高的 m 值会导致落入每一个区间的数据对象的个数很少，这样会在过滤掉噪声信息的同时，也将许多包含非噪声信息的维度排除在最终选择的子空间之外，从而得到不准确的度量结果。

综合两方面的因素，m 值可取作维数 d 的线性倍数，即有 $m=\lceil c\cdot d\rceil$，其中 c 为某常数。

结束语 为有效消除维度灾难对高维数据相似性度量的影响，本文提出了一种基于子空间的相似性度量方法。通过将高维数据进行基于网格的划分，进而在划分后的合理子空间中进行数据的相似性度量，可较好地避免噪声属性对高维数据相似性度量的影响。

网格划分参数是影响本文度量方法的一个重要方面，本文从理论分析的角度讨论了网格划分参数的取值准则。进一步的工作主要在于将提出的度量方法与具体的应用实例结合起来，如将其应用于高维数据的聚类处理，从而通过实践进一步完善网格划分参数的确定原则。

参考文献

[1] Hinneburg A, Aggarwal C C, Keim D A. What is the nearest neighbor in high dimensional spaces[C]// 26th VLDB Conference. 2000: 506-515

[2] Kriegel H-P, Kröger P, Zimek A. Clustering high-dimensional data: a Survey on subspace clustering, pattern-based clustering, and correlation clustering[J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3(1): 1-58

[3] 盛骤，谢式千，潘承毅. 概率论与数理统计[M]. 北京：高等教育出版社，1989