中国科学 : 信息科学

SCIENTIA SINICA Informationis

论文

基于邻域视角的关联关系挖掘方法

成红红[1], 钱宇华[1,2*], 胡治国[1], 梁吉业[2]

1. 山西大学大数据科学与产业研究院, 山西太原 030006

2. 山西大学计算智能与中文信息处理教育部重点实验室, 山西太原 030006

* 通信作者. E-mail: jinchengqyh@126.com

国家重点研发计划(批准号: 2018YFB1004300)、 国家自然科学基金(批准号: 61672332, 61872226)、 山西省重点研发计划(国际科技合作)项目(201903D421003)、山西省海外归国人员研究项目(批准号: 2017023)、山西省自然科学基金计划资助项目(No.201701D121052)、山西省高等学校科技创新项目(批准号: 201802013)资助项目

摘要 识别海量变量间潜在的复杂关联关系, 判断不同形式关联关系的强弱, 是大数据关联关系挖掘的重要任务之一. 然而, 数据分布的不确定性、关联关系的多样性, 使得基于分布假设的关联关系度量和基于数据驱动的非参数度量方法的适用性、准确性难以保证. 因此, 设计一种对关联关系形式无偏的有效关联度量方法变得至关重要. 本文从大数据背景下潜在关联关系应被公平排序的需求出发, 回顾了目前关联度量的公理化条件, 给出了大数据关联关系度量可能需满足的性质; 讨论了两类基于邻域视角的度量方法存在的不足; 提出了本文基于k-NN粒的关联度量方法, 称为最大邻域系数. 人造数据集和真实数据集实验从不同角度验证了本文所提方法的有效性和优越性. 最后指出了实验中发现的有趣现象和有待解决的理论问题, 以引起对该领域更深入的思考和研究.

关键词 大数据, 复杂关联关系挖掘, 关联度量, 数据驱动, 粒计算, k-NN粒

1 引言

关联关系挖掘旨在探索变量间潜在的同现规律和模式, 可先验性地提供变量间交互信息、协助更高层次的数据建模和统计分析, 其发展和应用广泛渗透到机器学习、生物信息学、社会网络、医疗诊断等多种领域[1~4]. 大数据背景下, 描述事物的变量成千上万, 变量之间隐藏数以万计关联关系的现象普遍存在[5], 逐个检查变量(组)之间潜在的关联关系非常浪费资源物力[6]. 更具挑战性的是多种线性和非线性关系共存, 采用不恰当的关联度量挑选待研究关联关系会产生误判, 例如基于正态分布假设的Pearson相关系数偏向选择线性关系[7], 非参数Spearman相关系数优先识别单调函数关系[8]. 这些现有方法的缺陷经常使研究者在理解底层数据时产生偏差. 因此, 设计独立于关联关系形式的关联度量是大数据复杂关联关系挖掘的重要研究方向之一.

引用格式: 成红红, 钱宇华, 胡治国等. 基于邻域视角的关联关系挖掘方法. 中国科学: 信息科学, 在审文章

Cheng H H, Qian Y H, Hu Z G, et al. Association mining method based on neighborhood (in Chinese). Sci Sin Inform, for review

关联关系挖掘从19世纪Galton研究人类身高遗传问题起一直受到广泛关注[7], 随着科学技术的发展呈现不同的背景需求[6, 9, 10]. 例如生物医学领域, 早期研究者注重研究对象间是否存在关联, 较少关注具体的关联关系形式, 因而所设计的关联度量多对变量独立情况敏感[11]. 然而大数据时代, 从海量变量中准确地识别和遴选关联强度较大的关联变量是复杂关联关系挖掘的挑战之一[4, 5], 这启发研究者重新探索迎接挑战的新思路. 2011年, Reshef等研究者在《Science》上指出, 大数据时代的关联关系度量应启发式地具备普适性(generality)和均衡性(equitability), 并相应地提出了同时满足上述性质的最大信息系数(MIC)[10]. Kinney等研究者指出MIC在大噪声情况下会偏向线性关系, 并指出导致偏差的原因之一是: MIC在寻找最优划分时过于关注网格线的坐落位置而忽略了网格中点的分布[12]. 根据信息论中数据处理不等式过程(DPI), 他们重新定义了均衡性, 提出用Self-均衡性(Self-equitability)代替Reshef等人提出的R^2-均衡性(沿用[12] 中记法), 并证明了基于k-NN统计量估计的互信息(MI_{KSG})能满足此性质[13]. 尽管研究者付出诸多努力试图解决大数据背景下关联关系度量面临的挑战, 依然存在许多问题有待研究:

(1) 现有的关联度量主要针对二元变量间关联关系的遴选任务展开, 对多元变量间相似任务分析较少. 然而在寻找与疾病紧密关联的基因序列任务中, 疾病的发生是由多个基因序列共同表达的结果, 仅分析单个基因序列与疾病的关系会使研究者无法做出正确判断和解释. 在挖掘二元变量关联关系时表现较好的MIC 难以扩展到多元变量情况.

(2) 受参数选择影响较大的关联度量, 不太适合参与多种关系形式共存场景下的关联强度比较, 因为一种参数设置大概率不能适应多个关联关系. 例如MI_{KSG}只有在参数$k = 1$时在不同函数关系上表现出Self-均衡性[13].

(3) 数据分布不确定性、关联关系多样性的特点, 迫切需要研究者设计参数备选空间小且简单有效的关联关系度量. MIC基于网格划分的互信息估计融合形成, 网格的搜索空间对关联度量精度影响较大; Rényi提出的最大相关系数在实践中较难确定变换函数空间[20].

粒计算被认为是现阶段人工智能领域的新兴计算范式. 它通过把复杂问题抽象、划分从而转化为若干较简单的问题. 有助于研究者更好的分析和解决问题[14, 15]. Liang和Qian等认为粒计算理论和方法可能是解决大数据背景下数据挖掘的有效范式, 指出局部数据粒上的模式发现和多粒度融合是解决问题的重要手段之一[16], 这与大数据关联关系度量基于数据驱动的需求相契合. 另外, Hu等研究者已利用δ-邻域粒(δ表示样本的邻域半径)构造邻域熵和邻域互信息衡量连续变量与离散变量之间的相关性, 并在特征选择任务中成功应用[17]. 本文通过固定样本点的邻居个数确定邻域粒大小, 引入k-NN粒代替样本点作为基本运算单位, 定义了单个变量基于k-NN粒的不确定性和两组变量在指定(k_x, k_y)邻域组合下的邻域互信息, 构建了关联关系的邻域特征矩阵. 矩阵中每个元素为归一化后的邻域互信息, 所有元素中的最大值称为最大邻域系数(MNC).

本文主要贡献如下:

(1) 针对大数据关联关系挖掘任务中, 关联度量可识别和遴选多种潜在关联关系并对关联形式无偏的需求, 试给出了大数据关联关系度量需满足的性质;

(2) 剖析了两类传统的基于邻域视角的关联关系度量方法, 并指出在邻域半径选择、局部均匀性假设、参数设置等方面存在的不足;

(3) 引入粒计算理论中的k-NN粒概念和多粒度融合思想, 提出了基于数据驱动的关联关系度量. 实验结果表明, 该度量可无偏地刻画多元变量之间的关联关系, 可缓减度量方法中参数选择对识别精度的影响.

2 大数据关联关系度量可能需满足的性质

早在1959年, Rényi给出依赖关系度量需满足7条公理性条件[20], 并提出满足条件的最大相关系数$MC(X,Y)=\sup_{f,g}\rho(f(X),g(Y))$. 该系数遍历所以可能存在的Borel可测函数$f,g$对原始数据进行变换, 计算变换后变量的$\rho$(Pearson相关系数), 从中选择最大值作为关联强度值. 实际问题中, 先验性地给定合适的变换函数较难, 通常使其难以有效计算[10]. 而且MC基于高斯分布假定, 认为数据经过函数变换呈线性相关. 然而, 随着数据中复杂关联关系的出现, 高斯分布假设不再成立. Schweizer和Wolf指出Rényi的部分公理性条件在非参数度量情况下比较严格, 并提出较松弛的公理条件进行修正[21]. 但是修正后的公理条件中, 关联度量取值为1当且仅当在变量间存在单调函数关系时成立. 这与大数据中单调、非单调函数关系甚至组合函数关系共存的现象不契合. 大数据背景下, 识别和遴选强关联关系的任务, 要求关联度量既能识别多种关联关系又对关联形式无偏向. 因此, 结合此需求和已有关联度量的公理性条件, 给出大数据关联关系度量可能需满足的性质.

以下子节中, 记$\mathbf{X}=(X_1,\cdots,X_p)$, $\mathbf{Y}=(Y_1,\cdots,Y_q)$为随机变量, p和q表示变量维度. $\delta(\mathbf{X},\mathbf{Y})$表示关联关系度量. 当$p=q=1$时, $\delta(\mathbf{X},\mathbf{Y})$衡量二元变量间关联关系强度.

(1) 广泛性: $\delta(\mathbf{X},\mathbf{Y})$ 用于衡量任意变量类型的关联关系, 要求随机向量$\mathbf{X}$, $\mathbf{Y}$不处处为常数.

(2) 对称性: $\delta(\mathbf{X},\mathbf{Y})=\delta(\mathbf{Y},\mathbf{X})$. 若$\mathbf{X}$与$\mathbf{Y}$之间存在某种关联关系, 变量位置改变不影响关联强度大小.

(3) 可比性: $0\leqslant\delta(\mathbf{X},\mathbf{Y})\leqslant 1$. 不同关联关系间, 关联强度可比较大小. δ越接近1表明关联关系越强. $\delta=0$ 表示变量间完全独立; $\delta=1$表示变量间存在强关联关系, 与关系形式无关.

(4) 普适性: 能识别广泛的关联关系形式, 不仅仅是简单的线性关系和单调函数关系.

(5) 均衡性: $\delta(f(\mathbf{X}),\mathbf{Y})=\delta(g(\mathbf{X}),\mathbf{Y})$, 如果$\mathbf{Y}=C(f(\mathbf{X}),\eta)$, $\mathbf{Y}=C(g(\mathbf{X}),\eta)$, 其中$f,g$为不同的Borel可测函数, η为噪声项, C表示真实关系和噪声项的组合函数. 若C为加和噪声, 则$C(f(\mathbf{X}),\eta)=f(\mathbf{X})+\eta$. 也即$\delta$只衡量关联强度, 受关系形式影响较小. 该性质也可称为无偏性.

(6) 单调性: 1) 随着$\mathbf{X}$中与$\mathbf{Y}$相关变量的增加, $\mathbf{X}$与$\mathbf{Y}$间关联强度增大, 即$\delta(\mathbf{X},\mathbf{Y})\geqslant\delta(\mathbf{X}\setminus X_i,\mathbf{Y})$, $\mathbf{X}\setminus X_i$表示$\mathbf{X}$中除掉X_i, 其中X_i、X_j相互独立, X_i与Y相关. 2) 随着$\mathbf{X}$中与X_i分量冗余变量的增加, $\mathbf{X}$与$\mathbf{Y}$间关联强度不变, 即$\delta((X_1,X_i,\cdots,X_p),\mathbf{Y})=\delta((X_1,X_i,f(X_i),\cdots,X_p),\mathbf{Y})$, $f(X_i)$表示X_i 的函数. 3) 随着$\mathbf{X}$中与$\mathbf{Y}$统计独立变量的增加, $\mathbf{X}$与$\mathbf{Y}$间关联强度减小, 即$\delta((X_1,\cdots,X_l,X_{l+1},\cdots,X_p),\mathbf{Y})\leqslant\delta((X_1,\cdots,X_l),\mathbf{Y})$, 其中$(X_{l+1},\cdots,X_p)$ 与$\mathbf{Y}$ 统计独立, 且与$(X_1,\cdots,X_l)$中变量相互独立.

(7) 可扩展性: 关联度量随着变量维度的增加容易扩展.

上述性质中, 若δ用于识别和遴选多元变量间潜在关联关系, 则其单调性成立需建立在前5条性质成立的基础之上.

3 基于邻域视角的相关方法

MIC假设存在一种网格划分能将数据的关联关系压缩出来, 通过计算不同网格划分的最大归一化互信息得分判断关联关系强度. 但是MIC在寻找网格的最优划分线时易忽略网格中每个元胞内数据点的分布. Kinney等研究者证明, 基于k-NN统计量的互信息估计MI_{KSG}能克服忽略局部结构引起的偏差. 但是MI_{KSG} 的均衡性表现受参数影响较大. 借鉴两种方法设计策略和估计角度的优点, 并克服它们存在的不足, 本文从样本点的邻域拓扑结构出发构造关联关系度量.

首先剖析两类基于邻域视角的关联度量在大数据背景需求下存在的不足, 然后提出基于k-NN粒的最大信息系数(MNC), 并对其是否满足大数据关联关系度量需满足的性质进行验证.

3.1 基于k-NN统计量的方法

MI_{KSG}是基于k-NN统计量估计的互信息[13]. 该方法首先通过max-范数确定联合分布中k个邻居点的邻域半径ρ_k, 利用此半径确定边际分布的邻居个数n_X和n_Y, 然后在k-NN矩形区域内按照熵定义估计互信息. MI_{KSG}通过邻域半径ρ_k建立起联合分布与边际分布之间的关系, 但采用max-范数会增大边际熵的估计偏差. 因为max-范数一定会增大某个变量的邻域范围, 这在高维情况下尤为明显. 朴素互信息估计方法先利用k-NN统计量分别对边际密度函数和联合密度函数估计(用l_2-范数确定邻居个数), 然后将估计概率插入到熵和互信息定义中完成互信息估计[22]. 插入式估计方式会加大真实值与估计值之间的偏差.

MI_{KSG}和朴素互信息估计都对样本进行了局部均匀性假定: 近似地假定k-NN矩形区域(max-范数)和k-NN球(l_2-范数)中的k个点具有均匀密度. 这种假设在强关联关系情况不成立, 因为k-NN统计量包含k个点所占的体积因子, k个点在这种情况下可能占很小的体积. 有研究者引入PCA估计局部非均匀性, 即用k个邻域点的真实体积修正指定邻域体积(记作MI_{LNC})[6].

基于k-NN统计量的关联度量依赖样本点的具体取值, k个邻居点所在的区域由第k个点决定, 未考虑邻域内散落点的信息, 这些因素会导致该类方法受噪声和参数影响较大.

3.2 基于k-NN图的方法

基于k-NN图的互信息估计方法将样本点当作图顶点, 样本点之间距离看作连边权重. MI_{GNN}是基于广义近邻图(GNN)估计的互信息[23]. 该方法先构造GNN统计量$L_p = \sum\limits_{(x,y)\in E(NN_S(V))} \|x-y\|^p$, 其中$V$表示图的顶点集合, S为指定的邻域集合(k 个正整数集合), $NN_S(V)$ 表示V 中每个顶点到S集合中元素的i近邻构成的有向广义近邻图集合, $E(NN_S(V))$为广义近邻图的连边集. 然后利用GNN 统计量估计Rényi 熵$\widehat{H}_\alpha(X_{1:n}) = \frac{1}{1-\alpha} log \frac{L_p(X_{1:n})}{\gamma n^{1-p/d}}$, 其中$p = d(1-\alpha)$, p表示距离的幂次, d表示数据的维度, γ为依赖于d, p, S的常数. 当$\alpha = 0.99$时, Rényi熵近似于Shanon熵. 互信息估计阶段, 先用copula函数对每个变量进行严格的增函数变换, 变换后变量联合熵的负值为所估计互信息.

邻域相似性(NS)是另一种基于k-NN图构造的关联度量方法[24]. 它假设: 两个变量之间若存在关联关系, 那么样本点在一个变量下的邻居点很大概率是在另一个变量下对应的邻居点. 基于此假设, 首先构建了各变量的邻域图、全连图和交互邻域图, 然后分别确定每个图的连边分布, 通过比较各连边分布的接近程度判断变量间的关联强度.

基于k-NN图的估计方法考虑所有样本的连边分布, 确定连边分布时至少涉及两个参数. 数据分布未知情况下, 该类方法的邻域选择范围较难确定, 引入更多参数会增大估计偏差, 用于多种关联关系识别会导致排序结果不可靠.

3.3 基于k-NN粒的关联度量方法

本节假设: 变量间存在关联关系，则各变量的样本存在相似邻域结构; 通过搜索各变量合适的邻域范围, 则可找到共同邻域结构.

给定样本集$S = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$来自联合分布$(\mathbf{X}, \mathbf{Y})$, 边际变量$\mathbf{X}$, $\mathbf{Y}$的样本分别为$S_{\mathbf{X}} = \{X_1, \cdots, X_n\}$、$S_{\mathbf{Y}} = \{Y_1, \cdots, Y_n\}$. 给定特定邻域组合$(k_x, k_y)$($(k_x, k_y)$为正整数对), 称$N_{\mathbf{X}}^{k_x}(X) = \{X_{j_1}, \cdots, X_{j_{k_x}}\}$为$X$样本的$k_x$-NN粒, 其中下标序列$j_1 < j_2 < \cdots < j_{k_x}$可由$d(X, X_{j_i}) = \|X - X_{j_i}\|_{d_{\mathbf{X}}}$排

序获得, d_X为$\mathbf{X}$空间上的l_p范数(此处$p=2$). $S_{\mathbf{X}}$中所有样本的k_x-NN 粒形成$S_{\mathbf{X}}$的覆盖, 即$\bigcup_{i=1}^{n} N_{\mathbf{X}}^{k_x}(X_i) = S_{\mathbf{X}}$. 同样地样本集$S_{\mathbf{Y}}$也存在覆盖$\bigcup_{i=1}^{n} N_{\mathbf{Y}}^{k_y}(Y_i) = S_{\mathbf{Y}}$. 邻域组合$(k_x, k_y)$形成样本集$S$的覆盖记为$C_{k_x k_y}$. 令$S|_{C_{k_x k_y}}$为样本集$S$在覆盖$C_{k_x k_y}$上的分布, 不同邻域组合形成不同的分布$S|_{C_{k_x k_y}}$.

3.3.1 邻域互信息

利用样本的k-NN粒代替样本作为基本运算单位, 定义特定邻域组合(k_x, k_y)下分布$S|_{C_{k_x k_y}}$的邻域熵和邻域互信息. .

定义1 给定$\mathbf{X}$的样本集$S_{\mathbf{X}}$, $N_{\mathbf{X}}^{k_x}(X_i)$为$X_i$的$k_x$-NN粒, X_i的邻域熵为

$$NH_{k_x}(X_i) = -log\frac{|N_X^{k_x}(X_i)|}{n}, \tag{1}$$

变量$\mathbf{X}$的邻域熵为

$$NH_{k_x}(\mathbf{X}) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{|N_X^{k_x}(X_i)|}{n} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{k_x}{n}. \tag{2}$$

其中$|\cdot|$表示集合的基数.

公式(1)中对数比例可看作样本X_i的局部概率. 由公式(2)知, $\forall X_i$满足$1 \leqslant |N_{\mathbf{X}}^{k_x}(X_i)| \leqslant n-1$, 因此成立$log\frac{n}{n-1} \leqslant NH_{k_x}(\mathbf{X}) \leqslant log(n)$. $NH_{k_x}(\mathbf{X}) = log(n)$当且仅当所有样本都只有一个近邻; $NH_{k_x}(X) = log\frac{n}{n-1}$当且仅当所有样本都将自身之外的点作为邻居. 变量的邻域熵大小随邻居数k_x变化, 这意味在先验信息未知时可以自由指定变量的不确定性.

定义2 给定随机变量$\mathbf{X}$, $\mathbf{Y}$的样本$S_{\mathbf{X}}$, $S_{\mathbf{Y}}$, 以及特定邻域组合(k_x, k_y). $N_{\mathbf{X}}^{k_x}(X_i)$为$X_i$的$k_x$-NN粒, $N_{\mathbf{Y}}^{k_y}(Y_i)$为$Y_i$的$k_y$-NN粒, 联合分布中样本$(X_i, Y_i)$在覆盖$C_{k_x k_y}$下的邻域粒记为$N_{\mathbf{X}*\mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i)$, 其邻域联合熵为:

$$NH_{C_{k_x k_y}}(X_i, Y_i) = -log\frac{|N_{\mathbf{X}*\mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i)|}{n}, \tag{3}$$

$(\mathbf{X}, \mathbf{Y})$的邻域联合熵为

$$NH_{C_{k_x k_y}}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{|NH_{C_{k_x k_y}}(X_i, Y_i)|}{n}. \tag{4}$$

此处定义$N_{\mathbf{X}*\mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i) = N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i)$, 则$(\mathbf{X}, \mathbf{Y})$的邻域联合熵可表示为

$$NH_{C_{k_x k_y}}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{|N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i)|}{n}. \tag{5}$$

特别地, $N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i) = \emptyset$意味着第$i$个样本点在$(k_x, k_y)$邻域组合下没有共同邻居, 此时该样本对变量联合熵的贡献为0, 约定$log\frac{0}{n} = 0$.

定理1 $NH_{C_{k_x k_y}}(\mathbf{X}, \mathbf{Y}) \geqslant NH_{k_x}(\mathbf{X})$, $NH_{C_{k_x k_y}}(\mathbf{X}, \mathbf{Y}) \geqslant NH_{k_y}(\mathbf{Y})$.

证明 $\forall\ (X_i, Y_i) \in S$, $X_i \in S_{\mathbf{X}}$, $y_i \in S_{\mathbf{Y}}$, 有$N_{\mathbf{X}*\ \mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i) \subseteq N_{\mathbf{X}}^{k_x}(X_i)$, $N_{\mathbf{X}*\mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i) \subseteq N_{\mathbf{Y}}^{k_y}(Y_i)$, 那么$|N_{\mathbf{X}*\mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i)| \leqslant |N_{\mathbf{X}}^{k_x}(X_i)|$, $|N_{\mathbf{X}*\mathbf{Y}}^{C_{k_x k_y}}(X_i, Y_i)| \leqslant |N_{\mathbf{Y}}^{k_y}(Y_i)|$ 成立, 因此$NH_{C_{k_x k_y}}(\mathbf{X}, \mathbf{Y}) \geqslant NH_{k_x}(\mathbf{X})$, $NH_{C_{k_x k_y}}(\mathbf{X}, \mathbf{Y}) \geqslant NH_{k_y}(\mathbf{Y})$.

定理1表明两个边际邻域熵共同决定联合邻域熵. 因此, 在先验信息未知情况下, 确定好两个边际邻域熵, 联合邻域熵随之确定.

定义3 给定样本集$S_{\mathbf{X}}$, $S_{\mathbf{Y}}$, 以及特定邻域组合(k_x, k_y)确定的覆盖$C_{k_xk_y}$. X_i, Y_i和(X_i, Y_i)的邻域粒分别为$N_{\mathbf{X}}^{k_x}(X_i)$, $N_{\mathbf{Y}}^{k_y}(Y_i)$和$N_{\mathbf{X}*\mathbf{Y}}^{C_{k_xk_y}}(X_i, Y_i)$, 则联合样本$(X_i, Y_i)$的邻域互信息为:

$$\begin{aligned} NMI_{C_{k_xk_y}}(X_i, Y_i) &= -log\frac{|N_{\mathbf{X}}^{k_x}(X_i)||N_{\mathbf{Y}}^{k_y}(Y_i)|}{n|N_{\mathbf{X}*\mathbf{Y}}^{C_{k_xk_y}}(X_i, Y_i)|} \\ &= -log\frac{|N_{\mathbf{X}}^{k_x}(X_i)||N_{\mathbf{Y}}^{k_y}(Y_i)|}{n|N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i)|} \\ &= log\frac{n|N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i)|}{k_xk_y}, \end{aligned} \tag{6}$$

联合分布$(\mathbf{X}, \mathbf{Y})$的邻域互信息为:

$$\begin{aligned} NMI_{C_{k_xk_y}}(X, Y) &= -\frac{1}{n}\sum_{i=1}^{n} log\frac{|N_{\mathbf{X}}^{k_x}(X_i)||N_{\mathbf{Y}}^{k_y}(Y_i)|}{n|N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i)|} \\ &= \frac{1}{n}\sum_{i=1}^{n} log\frac{n|N_{\mathbf{X}}^{k_x}(\mathbf{X}_i) \cap N_{\mathbf{Y}}^{k_y}(\mathbf{Y}_i)|}{k_xk_y}. \end{aligned} \tag{7}$$

公式(6)可以也称逐点邻域互信息, 联合邻域互信息为逐点邻域互信息之和. 若$N_{\mathbf{X}}^{k_x}(X_i) \cap N_{\mathbf{Y}}^{k_y}(Y_i) = \emptyset$, 则$NMI_{C_{k_xk_y}}(X_i, Y_i) = 0$. $NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) = 0$ 当且仅当所有$NMI_{C_{k_xk_y}}(X_i, Y_i) = 0$, $i = 1, \cdots, n$. 即S 中所有样本在特定邻域组合(k_x,k_y)下都没有共同邻居. 若搜索不同邻域组合, $NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) = 0$ 都成立, 则可认为两个变量间无任何关联关系.

定理2 给定($\mathbf{X}$,$\mathbf{Y}$)在特定邻域组合(k_x,k_y)下的邻域互信息$NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y})$, 如下性质成立:

(1) $NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) = NMI_{C_{k_xk_y}}(\mathbf{Y}, \mathbf{X})$;

(2) $NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) = NH_{k_x}(\mathbf{X}) + NH_{k_y}(\mathbf{Y}) - NH_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y})$;

依照Shannon熵和互信息的定义, 上述性质显而易见.

接着引入有限集合中样本的邻域关联概念.

定义4 给定样本点$(X, Y) \in S$和特定邻域组合(k_x,k_y), $N_{\mathbf{X}}^{k_x}(X)$为X的k_x-NN粒, $N_{\mathbf{Y}}^{k_y}(Y)$为Y的k_y-NN粒. 若$N_{\mathbf{X}}^{k_x}(X) \subseteq N_{\mathbf{Y}}^{k_y}(Y)$, 则称$(X, Y)$是$k_x \times k_y$邻域关联的; 若$N_{\mathbf{Y}}^{k_y}(Y) \subseteq N_{\mathbf{X}}^{k_x}(X)$, 则称$(X, Y)$是$k_y \times k_x$邻域关联的.

由定义4知, 每个样本点在特定邻域组合下存在两种不同的关联方向.

引理1 若样本点(X, Y)是$k_x \times k_y$邻域关联的, 则$NMI_{C_{k_xk_y}}(X, Y) = NH_{k_y}(Y)$; 若$(X, Y)$是$k_y \times k_x$邻域关联的, 则$NMI_{C_{k_xk_y}}(X, Y) = NH_{k_x}(X)$.

定理3 给定样本集S, $S_{\mathbf{X}}$, $S_{\mathbf{Y}}$和特定邻域组合(k_x,k_y), 若S_X中所有样本都满足$k_x \times k_y$邻域关联, 则$NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) = NH_{k_y}(\mathbf{Y})$; 若$S_{\mathbf{Y}}$中所有样本都满足$k_y \times k_x$邻域关联, 则$NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) = NH_{k_x}(\mathbf{X})$.

由定理3可知, $NMI_{C_{k_xk_y}}(\mathbf{X}, \mathbf{Y}) \leqslant min\{NH_{k_x}(\mathbf{X}), NH_{k_y}(\mathbf{Y})\}$. 等号成立的条件是: 所有样本至少满足一种邻域关联. 此时意味着在特定邻域组合(k_x, k_y)下, 一个变量的邻域信息完全由另一个变量决定.

引理1和定量3结合表明, 邻域互信息可衡量变量间的关联程度, 可判断潜在关系的关联方向, 可识别单个样本对整体关联强度的贡献, 同时也可排查偏离整体趋势的样本点. 然而, 海量变量间关联关系复杂多样, 实践中较难通过一组恰当的邻域组合(k_x,k_y)识别出多种关系形式. 不同邻域组合产生不同邻域关联, 融合不同邻域关联设计关联度量是一种可行策略.

3.3.2 最大邻域系数(MNC)

首先定义S的邻域特征矩阵(NM), 然后基于NM定义最大邻域系数(MNC).

定义5 给定样本集合S和邻域组合(k_x,k_y), S邻域特征矩阵(NM)中的元素为:

$$NM(S)_{k_x,k_y} = \frac{NMI(S|_{C_{k_xk_y}})}{log\frac{n}{max(k_x,k_y)}}. \tag{8}$$

$NMI(S|_{C_{k_xk_y}})$表示分布$S|_{C_{k_xk_y}}$上的邻域互信息. 由定理3知, 公式(8)中分母为特定邻域组合下邻域互信息的最大值, 因此邻域矩阵中元素的取值位于$[0,1]$. 归一化处理有助于同一个数据集不同覆盖上关联强度的比较, 也有助于不同数据集之间关联强度的比较.

定义6 给定有限样本集S, 以及邻域搜索范围$NB(n)$, 最大邻域系数(MNC)为:

$$MNC(S) = \max_{1\leqslant k_xk_y\leqslant NB(n)}\{NM(S)_{k_x,k_y}\}. \tag{9}$$

其中$1 \leqslant k_xk_y \leqslant O(n^\alpha)$, $0 < \alpha < 1$.

合适的$NB(n)$设置比较重要: 邻域范围过大会高估独立关系, 过小意味着只能挖掘简单关联关系. 文中从实验上给出有效的邻域范围$n^{0.7} \sim n^{0.8}$. 无特殊指定, 采用$NB(n) = n^{0.8}$进行实验分析.

3.3.3 MNC的相关性质分析

本节验证MNC在大数据关联关系度量需满足性质中的表现:

(1) 广泛性: MNC基于样本点间距离确定邻域, 只需变量空间存在距离度量, 就可获得MNC. 文献[28]中也表明基于邻域构造的互信息能适应多种变量类型.

(2) 对称性: $MNC(\mathbf{X},\mathbf{Y}) = MNC(\mathbf{Y},\mathbf{X})$. 由定理2(1)知, NMI具有对称性. MNC是归一化的NMI, 对称性不变.

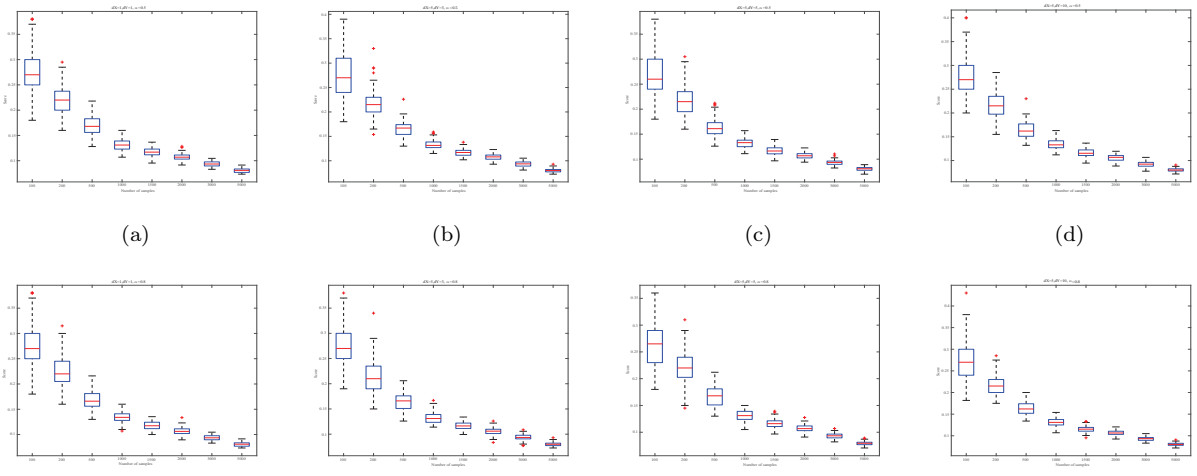
(3) 可比性: $0 \leqslant MNC(\mathbf{X},\mathbf{Y}) \leqslant 1$. 由定义5和定义6易知成立. 由定义3和定理3知, $MNC = 0$表示变量间独立, $MNC = 1$表示变量间存在强关联关系.

(4) 普适性: MNC基于邻域粒设计, 是一种非参数估计方法, 与数据分布无关. 表1中单调性不同的关联关系验证了该性质.

(5) 均衡性: 均衡性概念首次由Reshelf等人提出但是没有理论依据, Kinney等从数据处理不等式过程角度提出Self-均衡性. 两种角度的均衡性引起很多争议[12,25,26], 有研究者认为这两种角度可能同时存在[27]. 图3和图4实验结果表明MNC同时满足两种均衡性. 本文推测两种均衡性定义具有统一的数学形式, 均衡性也可能与互信息估计方式, 后续研究将探明.

(6) 单调性: MNC从数据局部拓扑结构出发构造, 变量间关联强度的变化受关联关系空间结构变化的影响, 通过关联强度的变化识别变量间潜在的关系形式. 多元变量分析中, 关联关系固定, 关联强度随变量维度变化的结果已表明其正确性.

(7) 可扩展性: 只要不同维度变量空间存在距离度量, MNC就可计算.



(a) (b) (c) (d)

(e) (f) (g) (h)

图 1　不同参数在独立数据上的表现

Figure 1　Empirical performance of different parameters on independent data

4　实验分析

为验证MNC的有效性和优越性, 选择两类数据集(模拟数据集、真实数据集), 六种代表性算法作为分析、对比对象. 两类基于邻域视角的方法性能受参数影响较大; Pearson相关系数(ρ)[7]可识别线性关系但不识别非线性关系; Spearman[8]相关系数可识别简单单调关系但不识别复杂周期关系, 文中也用来判断关系的单调程度; MIC[10]可识别多种关联关系但不能刻画变量组间关联关系; dCor[18,19]可衡量多元变量间的关联关系但识别复杂关联关系能力较弱. 通过和这些方法比较, 从不同角度去体现MNC的特点. 实验中数据都用min-max归一化方法处理, 使变量取值位于[0,1]. 涉及参数的比较方法如无特殊说明, 都采用默认设置. 为方便比较, MI_{KSG}、MI_{LNC}和MI_{GNN}先经过$NI=\sqrt{1-e^{-2I}}$变换, 使其取值位于[0,1], 其中I表示互信息.

4.1　参数有效性分析

首先, 检查MNC中参数对不同维度、不同样本量的统计独立数据的影响. 图1每个子图中, x轴表示统计独立数据的样本量, 从100到5000共8种情况; y轴表示MNC得分, 用对应样本100次重复实验的箱线图表示, 标题记录两个独立变量的维度和实验设置的参数. 实验结果显示, 所有子图中MNC值随样本量的变化趋势一致. 即固定变量维度和参数, MNC值随着样本量的增加逐渐减小并趋于0. 这表明独立变量的MNC值在大样本情况下偏差较小. 子图(a)-(d)的变量维度逐渐增加, 邻域参数设置为$\alpha=0.5$. 分析相同样本量不同维度MNC值的变化, 发现100次试验的MNC平均值(箱线图中的红色线)比较接近. 例如样本量为2000时, 4种不同维度下的MNC均值为0.11, 100次试验的波动范围为[0.09,0.12]. 这表明统计独立情况下, 变量维度变化对MNC值的影响较小. 类似地, 子图(e)-(h)的变量维度变化与(a)-(d)中的维度变化相同, 只是邻域参数设为$\alpha=0.8$. 在此设置下, MNC在相同样本量不同维度上的变化与子图(a)-(d)的表现一致, 100次试验的平均值差距也较小. 这表明α的大小对统计独立变量的MNC值影响较小. 结合有效识别海量变量间多种潜在关联关系形式的需求, 后续实验将邻域范围设置为$NB(n)=n^{0.8}$.

表 1　比较方法在无噪声关系上的表现

Table 1　Performance of all methods on noiseless functional relationships

Relationship Type	Figures	Spearman	Pearson	MIC	MI_{KSG}	MI_{LNC}	MI_{GNN}	NS	MNC
Random		0.03	0.03	0.17	0.13	0.13	0.26	0.00	0.21
Linear		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Exponential		1.00	0.87	1.00	1.00	1.00	1.00	0.99	1.00
Cubic		0.78	0.66	1.00	1.00	1.00	0.99	1.00	1.00
Linear Periodic		0.31	0.33	1.00	0.74	0.74	0.93	1.00	1.00
Sin *(Fourier frequency)*		0.14	-0.09	1.00	0.05	0.05	0.93	0.99	1.00
Sin *(Varying frequency)*		-0.11	-0.11	1.00	0.04	0.04	0.98	0.99	1.00
Parabolic		-0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
Sin *(nonFourier frequency)*		0.00	0.00	1.00	0.38	0.38	0.97	0.80	1.00

4.2　模拟数据实验分析

本节在二元变量关联关系、多元变量关联关系的模拟数据上验证MNC可识别复杂关联关系.

4.2.1　二元变量分析

二元变量关联关系数据分析阶段, 在无噪声关联关系上检验MNC的普适性, 在有噪声、单调性不同的函数型关联关系上检验MNC的均衡性.

(1) 普适性: 将MNC应用于多种不同形式的关联关系(表1中第一列为关联关系名, 第二列为对应的关系示意图. 所列关系来自文献[10], 并按$Spearman^2$进行单调性降序排列), 并用不同颜色区分参与比较的关联度量在关系数据集上的分值. 基于邻域视角的两类方法受参数影响较大, 因此表1 中所列结果为多个不同参数设置下的最优值. 观察表1中结果可知, MNC(最后1列)和MIC(第5列)在所有强关联关系上的表现一致, 所给关联强度都为1, 这表明了MNC也能识别不同形式的关联关系. 与其它方法相比, MNC的性能明显较优, 尤其在单调性弱、周期性强的关联关系上的, 其它方法给出的关联强度较低. 另外, 基于k-NN图的两种方法MI_{GNN}、NS在实验上要优于基于k-NN统计量的两种方法MI_{KSG}、MI_{LNC}, 因为它们在单调性弱、周期性强的关联关系上的得分较高. 值得注意的是, MNC在

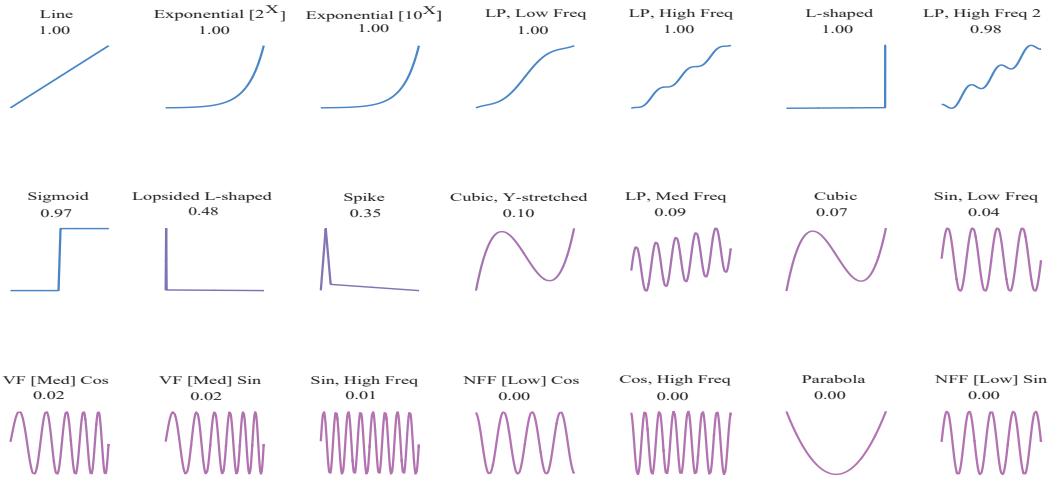


图 2　按单调性排序的函数关系

Figure 2　Functions f used to analyze the equitability of MNC and colored with descending monotonicity

随机变量上的得分偏高, 这主要与数据的样本量小有关(图1可知).

(2) 均衡性: 将MNC应用于21个有噪声、单调性不同的函数型关联关系数据集, 观察所有关系上MNC值随噪声量的变化趋势. 实验数据集生成方式同文献[12]: $Y = f(X) + \eta$, f为图2中所示的单调性递减函数关系, η为服从均匀分布的加和噪声项, 实验中共设置24个不同噪声水平. 图2展示Reshelf等人提出的R^2-均衡性, 用1-$R^2(Y, f(X))$衡量噪声量, R为Pearson相关系数. 图3展示Kinney等人提出的Self-均衡性, 通过$M(X,Y) = M(f(X),Y)$ 成立与否, 判断关联度量M在相同噪声水平时对不同关联关系的识别能力. 两个图中, 前两行展示基于k-NN统计量方法的性能, 中间两行展示基于k-NN图方法的性能, 这两类方法分别考虑在参数k =2、3、5、10四种情况上的性能. 最后一行展示dCor, MIC和本文所提度量MNC的性能. 每个子图中, x轴表示噪声量, y轴表示对应度量的得分值. 观察两种均衡性指标随噪声量的变化趋势, 发现不同函数关系MNC值的变化曲线一致优于其它方法. 即在相同噪声水平下, 不同函数关系上的MNC值接近; 在大噪声情况下, 也未出现偏向某种关联关系的现象. 这从实验上表明MNC满足两种均衡性定义, 对关联关系形式没有偏向. 然而, MIC在大噪声时明显倾向于简单函数关系(蓝色曲线与红色曲线分离); MI_{KSG}在不同函数关系上的得分变化趋势在两种均衡性定义中都较分散, 但在k值较小时的分散程度较小, 说明该方法性能受参数选择影响较大; 改进后的MI_{LNC}在不同关系上的得分随噪声量的变化趋势较MI_{KSG}紧凑, 但也在不同参数设置下表现不同; 基于k-NN图的方法在两种均衡性上无收敛规律, 它给具有大噪声量的关联关系较高的分数(大噪声量对应的蓝色点), 却将具有小噪声量的关联关系判断为0(小噪声量对应的红色点). 这表明该类方法受噪声影响较大, 不适合识别复杂关联关系; dCor在单调性关系上的变化趋势明显分离于非单调关系, 表明dCor对关联关系形式有偏向. 基于k-NN粒设计的MNC度量在两种均衡性上的性能都优于其它两类基于邻域视角的关联度量, 推测与MNC在构造过程中考虑样本k个邻居的相对秩序而非具体取值有关. 另外, MNC在两种均衡性定义上的表现一致, 推测两种均衡性具有一致的数学表征形式.

4.2.2　多元变量分析

上节实验结果已表明, MNC能识别多种复杂关联关系, 且对关系形式具有无偏性. 本节在多元变量数据集上验证MNC的单调性. 基于邻域视角的两类方法受参数影响较大且对二元复杂关联关系的

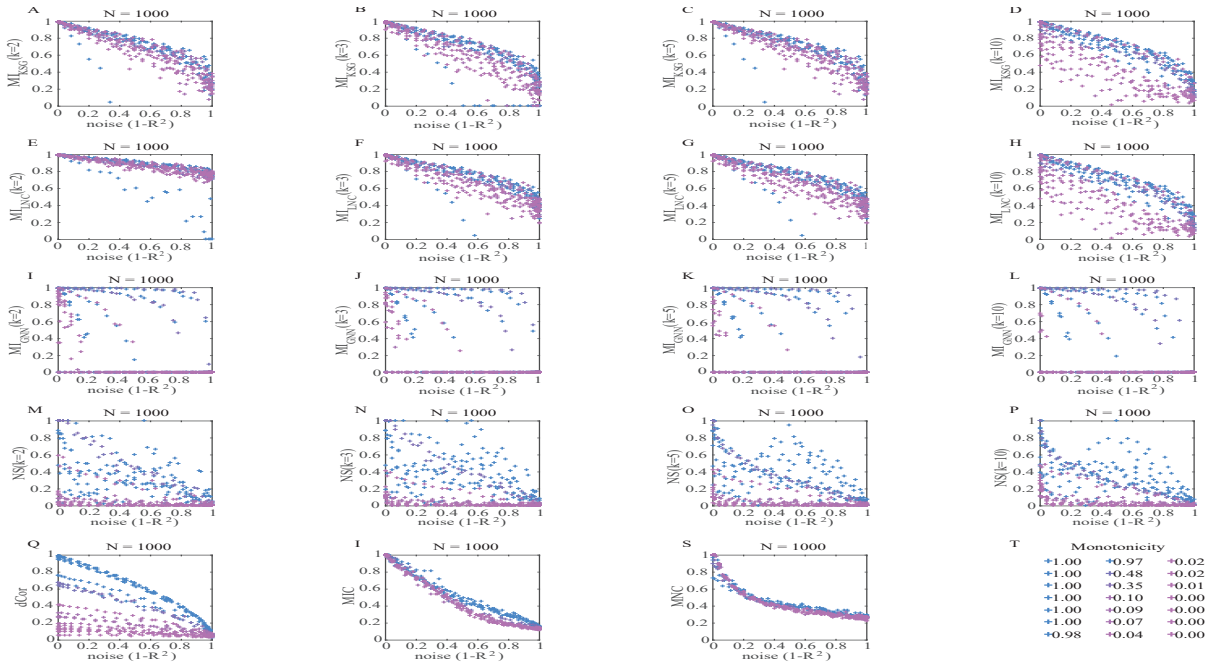


图 3　所有方法的R^2-均衡性表现

Figure 3　Performance of all comparison measures on R^2-equitability

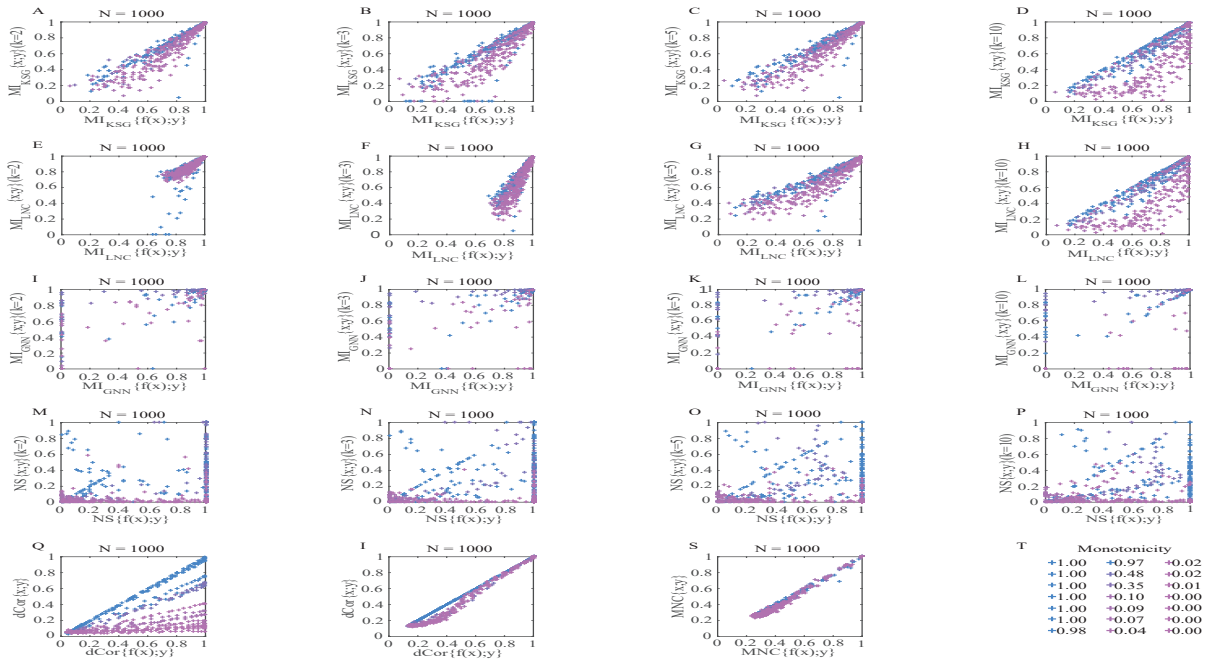
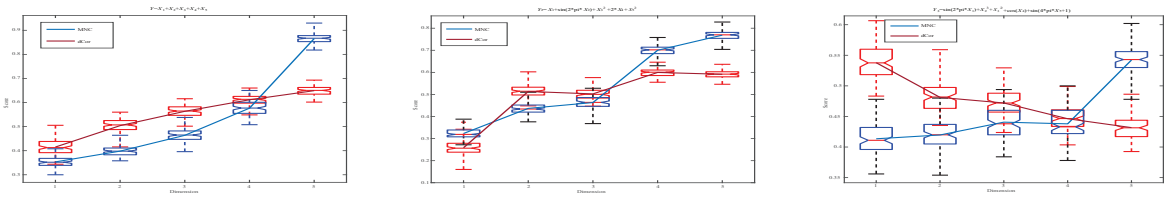


图 4　所有方法的Self-均衡性表现

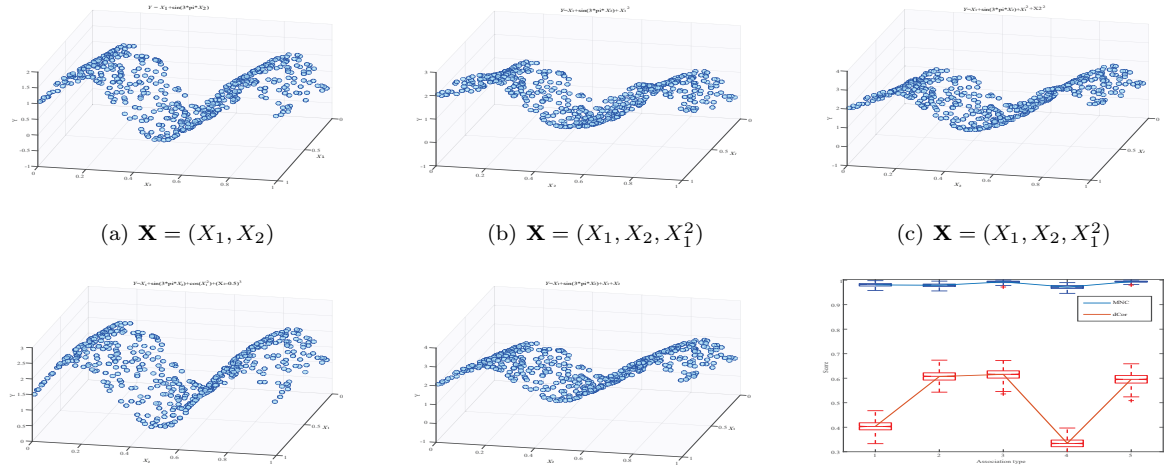
Figure 4　Performance of all comparison measures on Self-equitability



(a) 线性关系　(b) 混合关系　(c) 非线性关系

图 5　随着X中与Y相关变量维度的增加，MNC和dCor在三种不同关系形式上的表现

Figure 5　Empirical performance of MMC and dCor with respect to three different relationship types as the dimension of variables associated with Y in $\mathbf{X}$ increases. (a) Linear relationship; (b) Mixed relationship; (c) Nonlinear relationship



(a) $\mathbf{X}=(X_1,X_2)$　(b) $\mathbf{X}=(X_1,X_2,X_1^2)$　(c) $\mathbf{X}=(X_1,X_2,X_1^2)$

(d) $\mathbf{X}=(X_1,X_2,X_1^2,X_2^2)$　(e) $\mathbf{X}=(X_1,X_2,X_1,X_2)$　(f) Experimental result

图 6　变量X中五种冗余关系类型及MNC和dCor在各情况上的表现

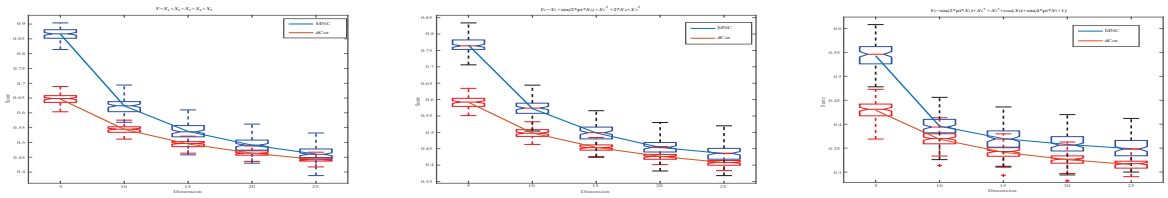
Figure 6　Five redundant relationship types in $\mathbf{X}$ and empirical performance of MMC and dCor in each case

识别能力较弱, MIC、Pearson相关系数和Spearman相关系数不能识别多元变量关联关系, 这些方法不参与本节比较. dCor 存在对关联关系形式有偏向的缺陷, 但可衡量多元变量间关联关系强度, 本节作为基准方法验证MNC的单调性.

首先构造不同形式多元变量关联关系(图5-图7中每个子图的标题表示关系形式), 每种关联关系由500个均匀采样点形成, 其中$\mathbf{X}$变量的子分量间相互独立, 实验结果用100次独立实验的箱线图表示.

(1) 变量Y与变量$\mathbf{X}$之间关联关系形式固定, Y与$\mathbf{X}$中的每个分量都有关联. 图5(a)-(c), 关联关系的复杂程度逐渐增加. 每个子图中, x轴表示增加新分量后变量$\mathbf{X}$的维度, y轴表示关联度量的得分. (a)和(b)子图中, 随着$\mathbf{X}$中与与Y相关分量维度的增加, dCor和MNC的均值逐渐增大(蓝色曲线和红色曲线分别表示两种度量100次实验平均值的变化趋势). (c)中关联关系的非线性程度最大, dCor的均值随着相关维度的增加反而减小. 这表明随着$\mathbf{X}$中与Y相关变量维度的增加, Y与$\mathbf{X}$的关联强度增加.

(2) 变量Y与变量$\mathbf{X}$之间关联关系形式固定, 图6(a)展示了三维变量间存在的关系, 其中Y与X_1具有线性关系与X_2具有非线性周期关系. (b)-(e)$\mathbf{X}$中逐渐增加与X_1和X_2呈线性或非线性关系的冗余维度, 并展示对应散点分布. (f)展示五种关联关系上MNC的性能. 实验结果展示, MNC在五种关系中的



(a) 线性关系　(b) 混合关系　(c) 非线性关系

图 7　随着X中与Y独立变量维度的增加，MNC和dCor在三种不同关系形式上的表现

Figure 7　Empirical performance of MMC and dCor with respect to three relationship types as the dimension of independent variables with Y in $\mathbf{X}$ increases. (a) Linear relationship; (b) Mixed relationship; (c) Nonlinear relationship

得分箱线图较窄且接近1. 这表明MNC受冗余维度和冗余关系形式影响较小. 而dCor在非线性冗余关系(d)和线性冗余关系(e)上的均值相差大约0.3, 说明其受冗余关系形式的影响大, 不适合识别多种关系形式并存的多元变量关联关系挖掘任务.

(3) 变量Y与变量$\mathbf{X}$的关联关系固定, $\mathbf{X}$中逐渐增加与Y和$\mathbf{X}$中分量独立的变量, 观察MNC和dCor值随$\mathbf{X}$维度增加的变化趋势. 与图5中关联关系形式相同, 图7中x轴表示$\mathbf{X}$变量参与计算的维度, y轴为关联度量得分. 实验结果表明, 两种方法在三种关系上的关联强度随噪声维度的增加逐渐减少, 意味着MNC、dCor能识别固定关系中隐藏的噪声.

多元变量间关联关系分析结果显示, MNC的单调性性能优于dCor. 这表明MNC通过识别关联关系的本质结构判断关联强度, 因此可通过MNC值的变化判断变量$\mathbf{X}$中与Y最相关的变量子集.

4.3　真实数据实验分析

本节在真实数据上验证MNC的有效性和优越性. 即MNC不仅具备普适性和均衡性, 还因单调性成立可用于多元变量间潜在关联关系的识别和筛选.

将MNC应用于建筑效能数据集[29]. 该数据集包括768个样本, 8个特征变量: 相对紧实度(X_1), 表面积(X_2), 墙面积(X_3), 屋顶面积(X_4), 总体高度(X_5), 方向(X_6), 玻璃面积(X_7), 玻璃面积分布(X_8)和2个因变量: 热负荷(Y_1), 冷负荷(Y_2). 此处仅挖掘与热负载因变量紧密相关的特征变量子集. 文中设计了MNC-ρ^2统计量用于判断关联关系的非线性程度. 该统计量取值越大, 表明关联关系的非线性程度越强. 二元变量关联关系挖掘阶段, 展示MNC、ρ、Spearman、dCor和MIC的识别和遴选结果, 同时也展示MIC-ρ^2[10]和MNC-ρ^2 对关联关系非线性程度的判断结果. 多元变量关联关系挖掘阶段, 仅展示MNC和dCor的识别和遴选结果.

表2展示按MNC值降序排列后的成对特征变量对及所列统计量的得分. 观察列表前8组特征变量对, MNC和MIC对它们的关联强度判断为1或接近1, 说明两者对强关联特征变量对的相对排序一致; Spearman绝对值、ρ绝对值和dCor对所列变量对关联强度的相对大小判断也一致; 不同之处在于, MNC和MIC在(X_1,X_3)、(X_2,X_3)变量对上的关联强度很高, 而Spearman绝对值、ρ绝对值和dCor在两组变量对上的关联强度较低, 同时发现MIC-ρ^2、MNC-ρ^2在两组变量对上的分值较高. 结合所比较方法的特点和MIC-ρ^2、MNC-ρ^2统计量可判断关联关系非线性程度的功能, 可知MNC对关联关系形式无偏向, 可对潜在关联关系进行公平排序.

表3展示不同方法对所有特征变量与Y_1关联程度的判断, 并按MNC值降序排列. 对比结果发现, MNC和MIC将(X_1,Y_1)、(X_2,Y_1)置于列表顶端, 说明两者都认为X_1、X_2与Y_1最关联, 这与文献[29]基

表 2　不同方法衡量8个特征变量两两之间的关联强度

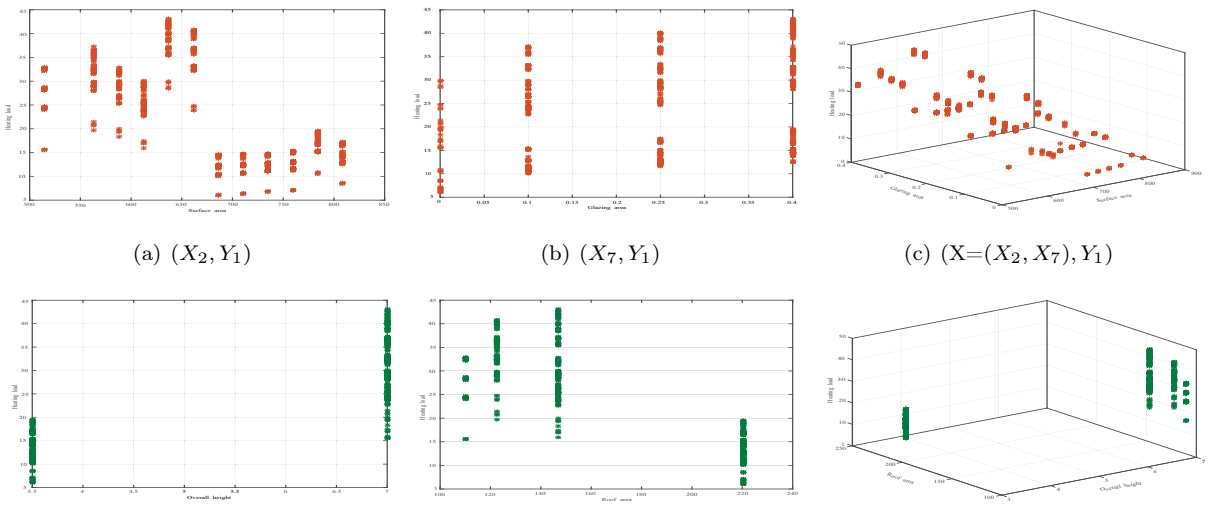
Table 2 Different measures to compute the associations strength of pairwise variables

Xvar	Yvar	MNC	Spearman	dCor	MIC	MNC-ρ^2	MIC-ρ^2	rho
X_1	X_2	1	-1	1	1	0.02	0.02	-0.99
X_1	X_3	1	-0.26	0.45	1	0.96	0.95	-0.20
X_1	X_4	1	-0.87	0.88	1	0.25	0.25	-0.87
X_1	X_5	1	0.87	0.86	1	0.31	0.31	0.83
X_2	X_3	1	0.26	0.45	0.99	0.96	0.95	0.20
X_2	X_4	1	0.87	0.89	1	0.22	0.22	0.88
X_2	X_5	1	-0.87	0.89	1	0.26	0.26	-0.86
X_4	X_5	1	-0.94	0.99	1	0.05	0.05	-0.97
X_5	X_6	0.79	0	0	0	0.79	0	0
X_3	X_5	0.78	0.22	0.31	0.37	0.71	0.30	0.28
X_3	X_4	0.72	-0.19	0.34	0.39	0.63	0.30	-0.29
X_4	X_6	0.66	0	0	0	0.66	0	0
X_1	X_6	0.58	0	0	0	0.58	0	0
X_2	X_6	0.58	0	0	0	0.58	0	0
X_3	X_6	0.56	0	0	0	0.56	0	0
X_5	X_8	0.5	0	0	0	0.5	0	0
X_3	X_8	0.42	0	0	0	0.42	0	0
X_6	X_8	0.40	0	0	0	0.40	0	0
X_7	X_8	0.38	0.19	0.21	0.34	0.33	0.29	0.21
X_3	X_7	0.36	0	0	0	0.36	0	0
X_5	X_7	0.34	0	0	0	0.34	0	0
X_4	X_8	0.25	0	0	0	0.25	0	0
X_4	X_7	0.25	0	0	0	0.25	0	0
X_1	X_8	0.25	0	0	0	0.25	0	0
X_2	X_8	0.25	0	0	0	0.25	0	0
X_6	X_7	0.25	0	0	0	0.25	0	0
X_1	X_7	0.23	0	0	0	0.23	0	0
X_2	X_7	0.23	0	0	0	0.23	0	0

表 3　8个特征变量与热载变量的关联关系

Table 3　Associations of 8 variables against heating load

Xvar	Yvar	MNC	MIC	Spearman	dCor	MIC-ρ^2	MNC-ρ^2
X_1	Y_1	0.81	1	0.62	0.76	0.61	0.43
X_2	Y_1	0.81	1	-0.62	0.78	0.57	0.38
X_3	Y_1	0.72	0.67	0.47	0.43	0.46	0.51
X_4	Y_1	0.66	1	-0.80	0.91	0.26	-0.09
X_7	Y_1	0.65	0.68	0.32	0.25	0.60	0.57
X_5	Y_1	0.51	1	0.86	0.92	0.21	-0.28
X_8	Y_1	0.45	0.26	0.07	0.09	0.25	0.44
X_6	Y_1	0.39	0.14	0	0.01	0.14	0.39



(a) (X_2, Y_1)　(b) (X_7, Y_1)　(c) $(\mathrm{X}=(X_2, X_7), Y_1)$

(d) (X_5, Y_1)　(e) (X_4, Y_1)　(f) $(\mathbf{X} = (X_5, X_4), Y_1)$

图 8　ENB数据集中代表性关系展示: (a)-(c) 由MNC发现; (d)-(f) 由dCor发现

Figure 8　Demonstration some representative associations of ENB: (a)-(c) by MNC, (d)-(f) by dCor

于互信息判断X_1、X_2与Y_1最相关的结论一致; MIC、dCor和Spearman给(X_4,Y_1)、(X_5,Y_1)较高的关联强度值, 将X_4、X_5与Y_1的关联关系优先于其它变量, 然而文献[29]表明X_4、X_5与Y_1呈线性相关, 这反映了dCor、Spearman指标优先识别线性关系的特点. MNC将(X_3,Y_1)置于列表第三位, 结合MIC-ρ^2和MNC-ρ^2给出较高值和dCor、Spearman识别关联关系有偏向的特点, 我们推断X_3与Y_1之间存在非线性关联关系, 这留给领域专家解释.

为分析MNC识别和筛选多元变量间关联关系的能力, 文中比较了所有可能特征变量组合(256-8=248种)与Y_1的关联强度, 并展示每种组合情况MNC、dCor值位于前5的变量组合(表4). 表3中, MNC将X_2判断为Y_1的最关联变量之一; 表1中, X_2和X_7的MNC值位于列表底端(可认为两者统计独立); 而当X_2 与X_7 组合之后, MNC值增加: $MNC(X_2, Y_1) = 0.81$, $MNC(X_7, Y_1) = 0.65$, $MNC(X_2, X_7) = 0.81$, $MNC((X_2, X_7), Y_1) = 0.94$. 统计独立特征变量联合后与因变量的关联强度增大, 这表明MNC能挖掘与应变量真正关联的自变量. 表3中dCor优先选择X_5、X_4为Y_1的最关联变量, 在表4中也将(X_5,X_4)组合位于列表之首. 然而, 表2中显示X_5与X_4存在明显的线性关系($\rho = -0.97$), 这说明dCor偏向线性关联关系. 图8 分别展示了MNC、dCor排列首位的关联关系, MNC 遴选出的关联关系(c)明显比dCor遴选出的关联关系(f)复杂, 具体原因留给专家解释.

观察表4所列的变量组合, 发现每种组合中X_7都可被MNC选中, 说明X_7与其它变量组合后增强了与Y_1的关联程度. 这与专业领域文献[29]的结论: X_7与Y_1弱相关, 但对提高Y_1预测精度最重要, 相吻合. 另外发现, 随着特征变量个数的增加, MNC先保持不变而后减少. 以每种组合的首行变量组合为例, 逐渐增加的特征变量依次为X_5、X_4、X_3、X_1、X_8和X_6, 其中X_8是MNC变化的拐点. 结合表2、表3可知, X_5、X_4、X_1与X_2高度线性相关(ρ值较大), X_3与X_2存在非线性关联关系(MNC, MIC, MIC-ρ^2和MNC-ρ^2在其上得分较高). X_5、X_4、X_3和X_1可看作X_2的冗余变量, 因而随着这些变量的增加, MNC值不变; X_8和X_6与所有特征变量的关联强度较小(参与比较的方法在其上的分值都较小)、与Y_1的关联强度列于表3最底端. X_8和X_6可看作Y_1的统计独立变量, 随着这些变量的增加, MNC值减小. 这均表明MNC单调性成立.

综上, 模拟数据和真实数据实验结果均表明, MNC通过探索关联关系的本质拓扑结构识别潜在复杂关系, 受关系形式影响较小, 可用于多元变量间复杂关联关系挖掘任务.

5 总结

复杂关联关系的探索与识别是大数据背景下数据挖掘的重要前沿课题, 具有重要的学术意义和广泛的应用价值. 本文针对海量变量间多种复杂关联关系可被公平识别的需求, 尝试给出了大数据背景下关联关系度量需满足的性质, 并设计了一种能够满足该性质的最大邻域系数MNC. 该系数克服了两类基于邻域视角的关联度量性能受参数影响较大的缺点, 弥补了MIC只能识别二元复杂关联关系、dCor偏向于简单关联关系的不足.

均衡性实验发现, MNC同时呈现出R^2-均衡性和Self-均衡性的特点, 推测上述两种均衡性定义具有统一的数学表征, 厘清两者之间关系有助于对复杂关联关系本质结构的探索和判别. 此外, 邻域粒的统计性质、邻域粒的有效维度等理论解释对复杂关联关系挖掘领域起着基础性的影响作用, 这将是未来复杂关联关系挖掘领域的主要研究方向之一.

参考文献

1 Young A I, Benonisdottir S, Przeworski M, et al. Deconstructing the sources of genotype-phenotype associations in

表 4　不同特征变量组合情况下, 分别由MNC和dCor度量排在前5的关系

Table 4　Top 5 associations ranked by MNC and dCor on different combined variables

X	Y	MNC	X	Y	dCor
(X_2,X_7)	Y_1	0.94	(X_4,X_5)	Y_1	0.92
(X_1,X_7)	Y_1	0.94	(X_1,X_5)	Y_1	0.91
(X_4,X_7)	Y_1	0.86	(X_2,X_5)	Y_1	0.91
(X_5,X_7)	Y_1	0.84	(X_3,X_5)	Y_1	0.91
(X_3,X_4)	Y_1	0.84	(X_2,X_4)	Y_1	0.89
(X_2,X_5,X_7)	Y_1	0.94	(X_3,X_4,X_5)	Y_1	0.92
(X_1,X_5,X_7)	Y_1	0.94	(X_1,X_4,X_5)	Y_1	0.91
(X_2,X_4,X_7)	Y_1	0.94	(X_2,X_4,X_5)	Y_1	0.91
(X_1,X_4,X_7)	Y_1	0.94	(X_4,X_5,X_7)	Y_1	0.91
(X_3,X_4,X_7)	Y_1	0.94	(X_4,X_5,X_8)	Y_1	0.90
(X_2,X_4,X_5,X_7)	Y_1	0.94	(X_3,X_4,X_5,X_7)	Y_1	0.91
(X_3,X_4,X_5,X_7)	Y_1	0.94	(X_2,X_4,X_5,X_7)	Y_1	0.91
(X_1,X_4,X_5,X_7)	Y_1	0.94	(X_1,X_4,X_5,X_7)	Y_1	0.91
(X_2,X_3,X_5,X_7)	Y_1	0.94	(X_2,X_3,X_4,X_5)	Y_1	0.91
(X_1,X_3,X_5,X_7)	Y_1	0.94	(X_1,X_3,X_4,X_5)	Y_1	0.91
(X_2,X_3,X_4,X_5,X_7)	Y_1	0.94	(X_2,X_3,X_4,X_5,X_7)	Y_1	0.91
(X_1,X_3,X_4,X_5,X_7)	Y_1	0.94	(X_1,X_3,X_4,X_5,X_7)	Y_1	0.91
(X_1,X_2,X_4,X_5,X_7)	Y_1	0.94	(X_3,X_4,X_5,X_7,X_8)	Y_1	0.90
(X_1,X_2,X_3,X_5,X_7)	Y_1	0.94	(X_1,X_2,X_4,X_5,X_7)	Y_1	0.90
(X_1,X_2,X_3,X_4,X_7)	Y_1	0.94	(X_2,X_3,X_4,X_5,X_8)	Y_1	0.90
$(X_1,X_2,X_3,X_4,X_5,X_7)$	Y_1	0.94	$(X_1,X_2,X_3,X_4,X_5,X_7)$	Y_1	0.90
$(X_2,X_3,X_4,X_5,X_7,X_8)$	Y_1	0.88	$(X_2,X_3,X_4,X_5,X_7,X_8)$	Y_1	0.90
$(X_1,X_3,X_4,X_5,X_7,X_8)$	Y_1	0.88	$(X_1,X_3,X_4,X_5,X_7,X_8)$	Y_1	0.90
$(X_1,X_2,X_4,X_5,X_7,X_8)$	Y_1	0.88	$(X_1,X_2,X_4,X_5,X_7,X_8)$	Y_1	0.89
$(X_1,X_2,X_3,X_5,X_7,X_8)$	Y_1	0.88	$(X_1,X_2,X_3,X_5,X_7,X_8)$	Y_1	0.89
$(X_1,X_2,X_3,X_4,X_5,X_7,X_8)$	Y_1	0.88	$(X_1,X_2,X_3,X_4,X_5,X_7,X_8)$	Y_1	0.89
$(X_1,X_2,X_3,X_4,X_5,X_6,X_7)$	Y_1	0.84	$(X_1,X_2,X_3,X_4,X_5,X_6,X_7)$	Y_1	0.88
$(X_1,X_3,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.74	$(X_2,X_3,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.88
$(X_1,X_2,X_3,X_4,X_6,X_7,X_8)$	Y_1	0.74	$(X_1,X_3,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.88
$(X_2,X_3,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.73	$(X_1,X_2,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.88
$(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.74	$(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)$	Y_1	0.88

humans. Science, 2019, 365: 1396−1400
2 Fan J Q, Liu H. Statistical analysis of big data on pharmacogenomics. Advanced Drug Delivery Reviews, 2013, 65: 987−1000
3 Fang Z Y, Fan X W, Chen G. A study on specialist or special disease clinics based on big data. Frontiers of Medicine, 2014, 8: 376−381
4 Fan J Q, Han F, Liu H. Challenges of Big Data analysis. National Science Review, 2014, 1: 293−314
5 Speed T. A correlation for the 21st century. Science, 2011, 334: 1502−1503
6 Gao S, Ver Steeg G, Galstyan A. Efficient estimation of mutual information for strongly dependent variables. Artificial Intelligence and Statistics. 2015: 277−286
7 Galton F. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London, 1888, 45: 135−145
8 Spearman C. The proof and measurement of association between two things. The American Journal of Psychology, 1904, 15: 72−101
9 Spearman C. The Proof and Measurement of Association between Two Things. American Journal of Psychology, 1987.
10 Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. science, 2011, 334: 1518−1524
11 Picornell A C, Echavarria Diaz-Guardamino I, Alvarez Castillo E L, et al. 186PBreast cancer PAM50 subtypes: Correlation between RNA-Seq and multiplexed gene expression platforms. Annals of Oncology, 2017, 28
12 Kinney J B, Atwal G S. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences, 2014, 111: 3354−3359
13 Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical review E, 2004, 69: 066138
14 Bargiela, A. & Pedrycz, W. Granular computing. In Handbook on Commputational Intelligence: Fuzzy Logic, Systems, Artificial Neural Networks, and Learning Systems, 2016, 43−66
15 Qian Y H, Cheng H H, et al . Grouping granular structures in human granulation intelligence. Information Sciences 2017, 382: 150−169
16 Liang J Y, Qian Y H, Li D Y,et al.Theory and method of big data mining.Scientia Sinica: Information, 2015, 45: 1355−1369
17 Hu Q , Zhang L , Zhang D , et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information. Expert Systems with Applications, 2011, 38: 10737−10750
18 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. The Annals of Statistics, 2007, 35: 2769−2794
19 Székely G J, Rizzo M L. Brownian distance covariance. The Annals of Applied Statistics, 2009, 3: 1236−1265
20 Rényi, A. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungarica, 1959, 10: 441−451
21 Schweizer, B, Wolff, E F. On nonparametric measures of dependence for random variables. The Annals of Statistics, 1981, 9: 879−885
22 Singh H, Misra N, Hnizdo V, et al. Nearest neighbor estimates of entropy. American Journal of Mathematical and Management Sciences, 2003, 23: 301−321
23 Pál D, Póczos B, Szepesvári C. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In: Advances in Neural Information Processing Systems. 2010, 1849−1857
24 Xu Y, Qiu P, Roysam B. Unsupervised discovery of subspace trends. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37: 2131−2145
25 Reshef Y A, Reshef D N, Finucane H K, et al. Measuring dependence powerfully and equitably. The Journal of Machine Learning Research, 2016, 17: 7406−7468
26 Reshef D N, Reshef Y A, Mitzenmacher M, et al. Cleaning up the record on the maximal information coefficient and equitability. Proceedings of the National Academy of Sciences, 2014, 111: E3362
27 Murrell B, Murrell D, Murrell H. R2-equitability is satisfiable. Proceedings of the National Academy of Sciences, 2014, 111: E2160
28 Hu Q H, Yu R D. Applied rough calculation(The 2ed). Science Press, 2012
29 Tsanas A, Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings, 2012, 49: 560−567

Association mining method based on neighborhood

Honghong CHENG[1], Yuhua QIAN[1,2*], Zhiguo HU[1] & Jiye LIANG[2]

1. *Institute of Big Data Science and Industry, Shanxi University, Taiyuan* 030006*, China*;
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006*, China*
* Corresponding author. E-mail: jinchengqyh@126.com

Abstract One of an important tasks in big data association mining is to identify potentially complex associations among massive variables and to determine the strength of different forms of associations. However, the uncertainty of data distribution、the diversity of associations make the measures based on distribution assumptions and data-driven non-parametric measurement methods are difficult to ensure their applicability and accuracy. Therefore, it is urgent to design an effective association measure that is unbiased to the relationship types. In this article, starting from the fair ordering requirement of potential relationships in big data, we review the current axiomatic conditions of association metrics, provide some possible properties that associations measures in big data should to satisfy; discuss some shortages of two kinds association methods based on neighborhood; and propose a new associations measure based on k-NN granule, called maximum neighborhood coefficient. The experiments on artificial datasets and real datasets verify the effectiveness and superiority of the proposed method from different perspectives. Finally, some interesting phenomena in the experiment and theoretical issues to be solved are pointed out, we hope they will arouse deeper thinking and research in this field.

Keywords big data, complex associations mining, association measure, data-driven, granular computing, k-NN granule

Honghong Cheng was born in 1986. She received a B.S. degree at school Mathematical Sciences from Shanxi University, China, in 2012. She is a PhD candidate at Institute of Big Data Science and Industry, Shanxi University. Her research interests includes associations mining in big data and machine learning.

Yuhua Qian was born in 1976. He received the M.S. degree and the Ph.D. degree in Computers with Applications at Shanxi University in 2005 and 2011, respectively. He is a professor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He is actively pursuing research in artificial intelligence, granular computing, machine learning, and deep learning. He is winner of Excellent Youth Fund of China National Natural Science Foundation in 2013.

Zhiguo Hu was born in 1977. He received Ph.D. degree in computer science from Tongji University, China in 2012. He is now working in the school of computer and information technology. His research interests include network measurement, data mining and machine leaning.

Jiye Liang was born in 1962. He received the M.S. and Ph.D. degrees from Xi.an Jiaotong University, Xi.an, China, in 1990 and 2001, respectively. He is a professor at the School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education at Shanxi University. His current research interests include computational intelligence, granular computing, data mining, and knowledge discovery.