Neighborhood Information-based Method for Multivariate Association Mining

Honghong Cheng, *Member, IEEE,* Yuhua Qian, *Member, IEEE,* Yingjie Guo, Keyin Zheng, and Qingfu Zhang, *Fellow, IEEE,*

Abstract—Most current data is multivariable, exploring and identifying valuable information in these datasets has far-reaching impacts. In particular, discovering meaningful hidden association patterns in multivariate plays an important role. Plenty of measures for multivariate association have been proposed, yet it is still an open research challenge for effectively capturing association patterns among three or more variables, especially the scenario without any prior knowledge about those relationships. To do so, we desire a distribution-free, association type-independent and non-parametrical measure. For practical applications, such a measure should *comparable*, *interpretable*, *scalable*, *intuitive*, *reliability*, and *robust*. However, no exiting measures fulfill all of these desiderata. In this paper, taking advantage of the neighborhood information of a sample, we propose MNA, a maximal neighborhood multivariate association measure that satisfies all the above criteria. Extensive experiments on synthetic and real data show it outperforms state-of-the-art multivariate association measures.

Index Terms—Association mining, multivariate association measure, distribution-free, nonparametric, neighborhood information.

1 Introduction

NOWADAYS bases are large, complex and even unknown distribution [1], [2], [3], [4]. To better understand and utilize the knowledge hiding in them, one needs to explore and evaluate some characteristics of data sets [5], [6], [7], [8]. One of the interesting aspects of this target is to identify meaningful hidden association patterns among three or more variables, such as in Figure 1. Multivariate association analysis is a widely used technology to deal with this issue [9], [10], [11]. However, due to the diverse sources or different records in current datasets, we have no any prior knowledge about such relationships. So non-parametrical analysis technologies with neither assumption on data distribution nor types of association are still needed to explore.

In such contexts, one needs a multivariate association measure satisfying all following desiderata:

D1. Comparable. The association scores of any set of variables should fall into a predetermined range such as [0,1], which means that association measure can be meaningfully compared across differen multivariate sets.

D2. Interpretable. The results of association measure should be interpretable, which helps one easily interpret a given strength from highly association to independent. Such that 0 indicates a set of independent variables and 1 indicates a set of variables with strong relationship.

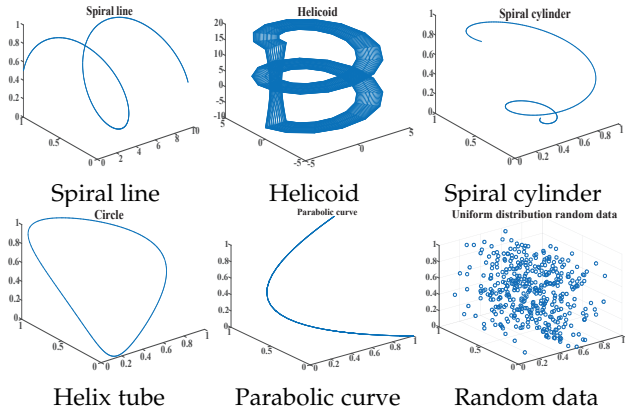

Fig. 1: Some possible potential associations among three variables.

D3. Scalable. To discovery more useful association patterns, not only a particular size subset of variables is concerned with, but also all possible subsets of multivariables. Then an association measure should be easy to compute as the number of data and dimension size increases.

D4. Intuitive. An association measure is intuitive if it involves fewer parameters with clear and explicit functions on the estimation process. Most existing methods need to handle a lot of unintuitive parameters, an appropriate relevant parameters are often provided by the inventors. Different parameters often yield different results. Hence, we target at a method that its parameters are easy to set and intuitive to use.

D5. Reliability. The reliability mentioned here includes two meanings: One is an association measure should be sensitive to any types of relationships, and has no bias to them. One is it should be unaffected to the dimension size,

- *Honghong Cheng is with the School of Information, Shanxi University of Finance and Economics, and also with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, Shanxi Province, China. Yuhua Qian, Yingjie Guo and Keyin Zheng are with the Institute of Big Data Science and Industry, Key Laboratory of Computer Intelligence and China Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi Province, China. Qingfu Zhang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the Shenzhen Research Institute, City University of Hong Kong, Shenzhen 518057, China.*
  E-mail: chhsxdx@163.com, jinchengqyh@126.com (corresponding author), yjguo0625@126.com, zhengkeyin1221@163.com, qingfu.zhang@cityu.edu.hk

Manuscript received Jan. 26, 2022; revised May. 13, 2022.

that is it identifies the fixed degree of association among them. The reliability guarantees a given score evaluates the true hidden association structure.

D6. Robust. Real-world data maybe are sampled from a noisy process. Some poor quality points will far away from their real value. Existing multivariate association measures are easily susceptible to the noisy points. Then an association measure requires to be robust against noise.

Recently, many multivariate association measures have been advanced to quantify the strength of a group variables. However, those measures only fulfil some of desiderata at best. For instance, MAC [12] involves dimension discretization, where grid technology overlooks easily the local information, so it is lack of robustness. CMI [13] couldn't compare across different multivariate sets due to its large value range, so it's poor in comparable. UDS [14] and HICS [15] are susceptible to higher dimensions, especially for independent data, so they behave badly in interpretable. UMC [16] is short of intuitiveness.

In this paper, we aim at proposing a new measure with all above desiderata. Total association framework based on information theory is employed, which has been proved can capture both linear and non-linear associations [17], [18]. For allowing non-parametric assumption on analyzed data, we make full use of the neighborhood information of a sample to estimate the Shannon entropy and total association. Meanwhile, the estimation process ensures robustness owing to the using of neighbors' index rather than their real values. To address the interpretable and reliability, we normalize the estimated total association under different neighborhood parameters, and select the maximal from all possible case as final multivariate association measure. At this point, we propose Maximal Neighborhood multivariate Association measure (MNA) that satisfies the above 6 desiderata.

The main contributions of our paper are summarized as follows:

1) A purely non-parametric neighborhood insight is introduced to estimate the traditional Shannon entropy, joint entropy and total association, and a neighborhood information-based multivariate association measure MNA is proposed, which fulfils all above desiderata, while the existing ones do not.
2) The rationality and validity of k-NN granule of a sample replacing the sample itself are proved theoretically and experimentally.
3) An efficient heuristic algorithm to compute the MNA is provided, which yields high quality.
4) Experimental results on both synthetic and real-world data demonstrate the advantages of MNA amongst existing measures.

The rest of the paper is organized as follows. Section 2 describes the preliminaries work on multivariable association measure. Specifically, it includes the basic framework of total association and several related works based on it. The rationality of neighborhood replacing sample theory is introduced in Section 3. Section 4 gives several concepts of neighborhood information-based association measure, proposes our approach maximal neighborhood multivariate association measure MNA. Section 5 provides an efficient calculating framework for MNA and analyses the time complexity. Numerical experiments are reported in Section 6. Finally, Section 7 summarizes the conclusion and future work of the paper.

2 Preliminary Work

Given a finite set samples D=$\{x_1, x_2, \cdots, x_n\}$ be a sequence of independent and identically distributed continuous points, which comes from multivariate random variable $\mathbf{X} = \{X_i, \cdots, X_d\} \in R^d$, and the probability density function and probability distribution function of each marginal random variable X_i are $p(X_i)$ and $P(X_i)$ respectively. Each sample x_i can be represented as $(x_{i,1}, x_{i,2}, \cdots, x_{i,d})$, $1 \leq i \leq n$. S_{X_i} is the sample sets of marginal variable X_i. Here, we assume that sample points in D taking values in the unit cube $[0,1]^d$. Let l_p-norm is a distance metric on R^d, and the distance between two points is defined as $\triangle(x_i, x_j) = (\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p)^{1/p}$ for $1 \leq p < \infty$, and $\triangle(x_i, x_j) = \max_{k=1,2,\cdots,d} |x_{i,k} - x_{j,k}|$ for $p = \infty$. Then for a point x_i, its neighbors can be ranked by $\triangle_{ij} = \triangle(x_i, x_j) : j = 1, 2, \cdots, n, j \neq i$. Noticing that sorting process is same for any choice of l_p-norm for $1 \leq p \leq \infty$, so we drop the p for convenience in the following section.

2.1 Total Association

To discover potential association patterns in multivariate, we need to quantify association strength of a subspace S. We will mainly use $S = \{X_1, \cdots, X_d\}$ represents a subspace in following section. For convenience, the $X_1, \cdots, X_d$ is abbreviated as $X_{1,\cdots,d}$.

A multivariate association measure $M(X_{1,\cdots,d})$ should be able to quantify the difference between their joint probability distribution and the product of their marginal probability distributions. According to the reference [19], the $M(X_{1,\cdots,d})$ is defined as follows:

$$M(X_{1,\cdots,d}) = \text{diff}(p(X_{1,\cdots,d}), \prod_{i=1}^{d} p(X_i)). \tag{1}$$

The larger the difference, the higher $M(X_{1,\cdots,d})$ is. An important property of $M(X_{1,\cdots,d})$ is non-negativity, the value 0 holds iff $p(X_1, \cdots, X_d) = \prod_{i=1}^{d} p(X_i)$.

If the diff is instantiated as the KL-divergence, the equation (1) will turn into the total association measure [19].

$$\begin{aligned} TA(D) &= KL(p(X_{1,\cdots,d}) \| \prod_{i=1}^{d} p(X_i)) \\ &= \sum_{i=1}^{d} H(X_i) - H(X_{1,\cdots,d}). \end{aligned} \tag{2}$$

Theorem 1. *$TA(D) \geq 0$ and 0 if and only if the $X_{1,\cdots,d}$ are statistical independent.*

The $TA(D) > 0$ indicates that there exists at least one variable brings a piece of certainty information to the others

in **X**. That is to say, some types of association structures are hidden in the multivariate **X**.

However, the application of TA is complex in practice, since the probability density functions are always unknown in the entropies. How to make an available estimate of involved probability density or entropy that does not bias the resulting total association value remains an open problem, especially in the case of high dimensional and limited samples. Despite many difficulties, plenty of practical multivariate association measures have been generated with different angles based on this framework.

2.2 Related Work

MAC is a maximal normalized total association measure, which discovers association patterns by identifying the discretizations of all dimensional spaces [12]. Yet, it involves a optimization problem at multiple grid sizes, the parameters in it often cause computational issues. What' more, the discretization process neglects easily the local structure inside grid. CMI is a modified cumulative mutual information, which quantifies subspace strength by aggregating the difference between cumulative entropy of single variable [11] and its conditional cumulative entropy [13]. UDS builds upon on CMI. The difference is that UDS fixes the permutation of dimension and adopts a optimal discretization to compute conditional cumulative entropy terms rather than clusters conditional dimensions. UDS is a normalized cumulative mutual information, which ensures the association strengthes across different subspaces can be compared [14]. UMC is designed for addressing the dimensionality bias issue, which introduces a permutation model in statistical model of independence to make measures suffer from the influence of "correlation-by-chance" value [16]. HICS is an intermediate for outlier mining in high dimensional data, which detects the associated subspaces by computing the cumulative deviation between the marginal probability density of a selected variable and its conditional probability densities [15]. In general, these approaches use the discrepancy between the joint distribution and the product of marginal distributions. In this work, we still follow this general framework, but aim at overcoming those problems.

Granular computing (GrC) is an emerging computing paradigm of information processing, which encourages an approach to data that recognizes and exploits the knowledge present in data at various levels of resolution or scales [20], [21], [22], [23]. That is, it helps us to better analyze and solve problems by abstracting and dividing complex problems into several simpler ones, so it is widely studied in various fields for solving complex problems. "Information granule" is the most critical, fundamental and central concept in GrC, which collects several entities together due to their similarity, functional or physical adjacency, indistinguishability, coherency, or the like [24], [25], [26]. There are many types of granularity that are often encountered in machine mining and data learning, where the equivalence class granulation is a common form. Within an equivalence class, any objects cannot be distinguished from one another based on the equivalence class criterion. In the absence of any prior knowledge from data, collecting the neighbors of samples is the most intuitive method for constructing equivalence classes. There are two types neighborhood granules commonly seen in existing research [27], [28]. One is δ-neighborhood granule [29], [30], where δ represents the radius of an object, objects falling into the circle are regarded as its corresponding neighborhood. Another one is k-NN granule [31], [32], [33], [34], where k is the number of objects in the neighborhood, k neighbors form its corresponding neighborhood. Both they have been applied to estimate entropy, joint entropy and mutual information and succeeded in feature selection, classification and other tasks [35].

Inspired by the data-driven computing advantages of GrC, here we utilize the k-NN granule of each sample instead of the sample itself to construct multivariate joint entropy and total association, and then design a neighborhood information-based multivariate association measure based on those definitions.

3 The Rationality of Neighborhood Replacing Sample

In this section, we focus on analyzing the rationality why the the sample itself can be replaced by its k-NN granule.

For a fixed positive integer $k(k \leq n-1)$, let $N_{i,n,k}(x) = \{x_{N_i(n,1)}, x_{N_i(n,2)}, \cdots, x_{N_i(n,k)}\}$ denotes the k-NN granule of x_i, where $N_i(n,j), j = 1, \cdots, k, j \neq i$ (remove itself) denotes the index of neighbors of x_i among $n-1$ samples, where the equidistant neighbor points are ordered by their index. Then the k-NN granules of all samples constitute a cover of D, that is $\bigcup_{i=1}^{n} N_{i,n,k}(x) = D$.

Let $B_{i,n,k}(x) \subset R^d$ denotes the k-NN ball of x_i, the ball centred at x_i and the radius equals to the distance from x_i to its kth neighbor in D. We assume that the k-NN ball formed by underlying distribution is continuous and smooth enough, the k points falling in the ball can be used to measure its probability. So there is $N_{i,n,k}(x) = B_{i,n,k}(x) \setminus x_i$. That is the k-NN granule can be called a hollow spherical neighborhood of x_i. Let $\omega_{i,n,k}(x)$ denote the probability measure induced by the $p(x)$ on the spherical neighborhood, we have

$$\omega_{i,n,k}(x) = \int_{B_{i,n,k}(x)} dP = \int_{N_{i,n,k}(x)} dP \tag{3}$$

the second equality dues to the fact that the probability of a single point in a continuous distribution is zero.

Lemma 1. *For a continuous sampling distribution, the expected probability measure for any k-NN ball over all sample relations is k/n, that is*

$$E(\omega_{i,n,k}(x)) = k/n. \tag{4}$$

Lemma 2. *Let $P(N_{i,n,k}(x))$ is the probability measure of the k-NN granule, it's a random variable since it depends on the density p being estimated. Let $P(N_{i,n,k}(x)) = \omega_{i,n,k}(x)$, it is a Beta distributions with parameters k and $n-k$.*

Proof. Let $G(\omega)$ is the distribution function of $\omega_{i,n,k}(x)$,

$$\begin{aligned} G(\omega) &= P(\omega_{i,n,k} \leq \omega) \\ &= P(B_{i,n,k}(x)\ contians\ at\ least\ k\ points) \\ &= 1 - \sum_{j=0}^{k-1} P(N_{i,n,j}(x)) \\ &= 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} \omega_{i,n,k}^{j} (1-\omega_{i,n,k})^{n-j-1} \end{aligned}$$

Integrating by parts,

$$\begin{aligned} E(\omega_{i,n,k}^{\alpha}) &= \int_0^1 \omega_{i,n,k}^{\alpha} dG(\omega) \\ &= 1 - \alpha \int_0^1 \omega_{i,n,k}^{\alpha-1} G(\omega) d\omega \\ &= \alpha \sum_{j=0}^{k-1} \binom{n-1}{j} \int_0^1 \omega_{i,n,k}^{j+\alpha-1} (1-\omega_{i,n,k})^{n-j-1} d\omega \end{aligned}$$

According to the relationship of Beta distribution and Bernoulli distribution [36]: $F(y) = \int_0^y \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)} dt = \sum_{i=\alpha}^{n} \binom{n}{i} y^i (1-y)^{n-i}$. We conclude that $\omega_{i,n,k}(x) \sim B(k, n-k)$. □

The integral part in the last equal is the Beta function. With the relationship of Beta function and Gamma function [37], we have $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ with parameters $a = j+\alpha$ and $b = n-j$, where Gamma function $\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$. Thus, we obtain

$$E(\omega_{i,n,k}^{\alpha}) = -\frac{\alpha\Gamma(n)}{\Gamma(n+\alpha)} \sum_{j=1}^{k-1} \frac{\Gamma(j+\alpha)}{\Gamma(j+1)} = \frac{\Gamma(n)\Gamma(k+\alpha)}{\Gamma(n+\alpha)\Gamma(k)}$$

Let $\alpha = 1$, the Lemma 1 was founded.

Let $\alpha = 2$, we obtain the variance of $\omega_{i,n,k}(x)$, $Var(\omega_{i,n,k}) = \frac{k(n-k)}{n^2(n+1)}$.

Theorem 2. *The $\hat{\omega}_{i,n,k} = \frac{k}{n}$ to estimate the probability of $\omega_{i,n,k}$, its convergence rate is $O(1/n)$ under the law of large numbers.*

Proof. Let the expectation and variance of random variables $\omega_{i,n,k}$ exist, for any constant $\varepsilon > 0$, according to the Chebyshev's inequality [38], we have

$$\begin{aligned} P(|\omega_{i,n,k} - \hat{\omega}_{i,n,k}| \geq \varepsilon) &\leq \frac{Var(\omega_{i,n,k})}{\varepsilon^2} \\ &= \frac{k(n-k)}{n^2(n+1)\varepsilon^2} \leq \frac{1}{n\varepsilon^2} \end{aligned} \tag{5}$$

That is

$$\lim_{n\to+\infty} P(|\omega_{i,n,k} - \hat{\omega}_{i,n,k}| \geq \varepsilon) \leq \lim_{n\to+\infty} \frac{1}{n\varepsilon^2} = 0 \tag{6}$$

where k is a specific value in a non-decreasing positive integers sequence $k(n)$ which satisfies $\lim_{n\to+\infty} k(n) = \infty$ and $\lim_{n\to+\infty} k(n)/n = 0$. So the $\hat{\omega}_{i,n,k}$ converges to $\omega_{i,n,k}$ in probability. □

Theorem 2 indicates that for a given k, the probability of large deviation decreases as n increases; On the other side, if the sample size n is fixed, increasing k will increase the probability of large deviation. Therefore we control the range of k in $[1, n^{\alpha}]$ in practice, where $\alpha \in [0, 1)$.

Next, we demonstrate the convergence of $\hat{\omega}_{i,n,k}$ under Mean Squared Error (MSE) [39].

$$\begin{aligned} E(\hat{\omega}_{i,n,k} - \omega_{i,n,k})^2 &= (E(\hat{\omega}_{i,n,k}) - \omega_{i,n,k})^2 + Var(\hat{\omega}_{i,n,k}) \\ &= (E(\hat{\omega}_{i,n,k}) - \omega_{i,n,k})^2. \end{aligned} \tag{7}$$

MSE is the sum of squared bias and variance, the second equality sets up because the variance vanished when the estimator is a constant. Then we have

$$\begin{aligned} E(\hat{\omega}_{i,n,k} - \omega_{i,n,k})^2 &= E(\hat{\omega}_{i,n,k}^2 - 2\hat{\omega}_{i,n,k}\omega_{i,n,k} + \omega_{i,n,k}^2) \\ &= \frac{k^2}{n^2} - 2\frac{k}{n}\frac{k}{n} + \frac{(k+1)k}{(n+1)n} \\ &= \frac{nk - k^2}{n^3 + n^2} = O(1/n) \end{aligned}$$

When $n \to \infty$, the rate of convergence of $\hat{\omega}_{i,n,k}$ under MSE is 1/n.

According to the above analysis, we define the entropy of $\omega_{i,n,k}(x)$ as $H(\omega_{i,n,k}) = -\int \omega_{i,n,k} log(\omega_{i,n,k}) d\omega_{i,n,k}$. Utilizing the $\hat{\omega}_{i,n,k}$ and the empirical $dG_n(\omega_{i,n,k})$ of $G(\omega_{i,n,k})$, we proposed a resubstitution estimation of $\hat{H}(\omega_{i,n,k})$, the specific form is

$$\hat{H}(\omega_{i,n,k}) = -\int log(\hat{\omega}_{i,n,k}) dG_n(\omega_{i,n,k}) = -\frac{1}{n}\sum_{i}^{n} log\frac{k}{n} \tag{8}$$

The above analysis indicates that the sample neighborhood can substitute the sample itself. It also provides a local view of underlying distribution. Based on the rationality analysis, we will introduce the k-NN granule to the total association and provide a non-parametric estimation for total association.

4 Neighborhood Information-based Multivariate Association Measure

In this section, we systematically present a novel multivariate association measure based on neighborhood information.

In the previous section, $N_{i,n,k}(x)$ denotes the k-NN granule of sample x_i from multivariate variables $\mathbf{X}$. To distinguish the granules formed by each marginal variable, we rewrite it as $N_{i,n,k}^{\mathbf{X}}(x)$.

Given a neighborhood combination $\{k_{X_1}, k_{X_2}, \cdots, k_{X_d}\}$ (k_{X_i} is a integer less than n), the $N_{i,n,\{k_{X_1},k_{X_2},\cdots,k_{X_d}\}}^{\mathbf{X}}(x)$, $i = 1, \cdots, n$ form a cover of D, record it as $C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$. Let $D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$ is the distribution of D on the cover $C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$.

4.1 Neighborhood Total Association

We first introduce the neighborhood entropy of a marginal variable.

Definition 4.1. *Given a data set $D = \{x_1, \cdots, x_n\}$ sampled from random variables $\mathbf{X} \in R^d$, S_X is the sample of marginal variable X, k_X is a integer, $N_{i,n,k_X}^{X}(S_X)$ is the k_X-NN granule of $S_X(i)$, the neighborhood entropy of $S_X(i)$ is defined as:*

$$NH_{k_X}(S_X(i)) = -log\frac{|N_{i,n,k_X}^{X}(S_X)|}{n}, \tag{9}$$

the neighborhood entropy of X is defined as

$$NH_{k_X}(X) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{|N^X_{i,n,k_X}(S_X)|}{n} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{k_X}{n}. \tag{10}$$

From the formula (10), since $\forall\ S_X(i)$, $N^X_{i,n,k_X}(S_X)$ satisfies $1 \le |N^X_{i,n,k_X}(S_X)| \le n-1$ (other than its own), so we have $log\frac{n}{n-1} \le NH_{k_X}(X) \le log(n)$. $NH_{k_X}(X) = log(n)$ holds if and only if for $\forall\ S_x(i)$, $|N^X_{i,n,k_X}(S_X)| = 1$. That is each point only has one closest neighbor with indistinguishability. For $NH_{k_X}(X) = log\frac{n}{n-1}$, if and only if for $\forall S_x(i)$, $|N^X_{i,n,k_X}(S_X)| = n-1$. That is each point can be represented by the rest. For a given k_X, $NH_{k_X}(X) = log\frac{n}{k_X}$, it means the uncertainty of a variable can be determined freely by reliable neighbors k_X without any prior knowledge.

Next, we introduce the joint neighborhood entropy of multivariate random variables $\mathbf{X}$.

Definition 4.2. *Given a data set $D = \{x_1, \cdots, x_n\}$ sampled from random variables $\boldsymbol{X} \in R^d$, $\{k_{X_1}, k_{X_2}, \cdots, k_{X_d}\}$ is a integer set, $C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$ is a cover on D formed by all granules on d dimensions. The joint neighborhood entropy of sample x_i is defined as:*

$$NH_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(x_i) = -log\frac{|N^{\boldsymbol{X}}_{i,n,\{k_{X_1},k_{X_2},\cdots,k_{X_d}\}}(x)|}{n}, \tag{11}$$

then joint neighborhood entropy of $\boldsymbol{X}$ is defined as:

$$NH_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} NH_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(x_i). \tag{12}$$

Here the $N^{\mathbf{X}}_{i,n,\{k_{X_1},k_{X_2},\cdots,k_{X_d}\}}(x) = N^X_{i,n,k_{X_1}}(S_{X_1}) \cap N^{X_2}_{i,n,k_{X_2}}(S_{X_2}) \cap \cdots \cap N^{X_d}_{i,n,k_{X_d}}(S_{X_d})$. The benefit of this setting is that we don't need to estimate the joint distribution of $\mathbf{X}$, which is determined by the marginal neighborhood entropy of each variable. That is $NH_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\mathbf{X})$ is certain as long as the neighborhood combination is given.

It's easy find that $|NH_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\mathbf{X})| \le \min\{k_{X_1}, \cdots, k_{X_d}\}$, so $NH_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\mathbf{X}) \le \sum_{i=1}^{d} NH_{k_{X_i}}(X_i)$.

The $N^{\mathbf{X}}_{i,n,\{k_{X_1},k_{X_2},\cdots,k_{X_d}\}}(x) = \emptyset$ indicates that the sample x_i on each variable has no common neighbor under the the given neighborhood combination. In this case, the contribution of the sample x_i to multivariate joint neighborhood entropy is 0. We agree that $log\frac{0}{n} = 0$.

According to the definition of total association, next we propose the multivariate total association of $\mathbf{X}$ from the neighborhood sight.

Definition 4.3. *Given the data set $D = \{x_1, \cdots, x_n\}$ sampled from multivariate random variables $\boldsymbol{X} \in R^d$, $\{k_{X_1}, k_{X_2}, \cdots, k_{X_d}\}$ is a group of integer, $C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$ is a cover on D formed by all granules on d dimensions. The neighborhood total association of the sample x_i is defined as:*

$$\begin{aligned} NTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(x_i) = &-\sum_{j=1}^{d} log\frac{|N^{X_j}_{i,n,k_{X_j}}(S_{X_j}))|}{n} \\ &+log\frac{|N^{\boldsymbol{X}}_{i,n,\{k_{X_1},k_{X_2},\cdots,k_{X_d}\}}(x)|}{n}, \end{aligned} \tag{13}$$

and the neighborhood total association of $\boldsymbol{X}$ is defined as:

$$NTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} NTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(x_i). \tag{14}$$

Theorem 3. *$NTA(D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}}) \ge 0$ for any cover $C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$, with equality iff $X_{i,\cdots,d}$ are statistically independent.*

If $X_{i,\cdots,d}$ are statistically independent, there will no common neighbors in their joint space in theory, so $N^{\mathbf{X}}_{i,n,\{k_{X_1},k_{X_2},\cdots,k_{X_d}\}}(x) = \emptyset$ to each sample for any cover $C_{k_{X_1}k_{X_2}\cdots k_{X_d}}$. According to the definition 4.3, the $NTA(D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}}) = 0$.

Theorem 4. *$NTA(D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}}) \le \sum_{i=1}^{d} log\frac{n}{k_{X_i}} - log\frac{n}{\min(k_{X_1},\cdots,k_{X_d})}$.*

The equality in Theorem 4 holds if and only if the joint neighborhood entropy reaches its maximum under given neighborhood combination.

The value of $NTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\mathbf{X})$ relies on the choice of $\{k_{X_i} : k = 1, \cdots, d\}$. For a fair comparison among different covers, one needs to analyze the value range of $NTA(\mathbf{X})$. For convenience, the $NTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(\mathbf{X})$ can be marked as $NTA(D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}})$. To conduct the maximal multivariate association measure on a finite data set D, one can search an optimal cover C^* to maximize the $NTA(D|C^*)$. Thus, for unbiased comparison, we normalize $NTA(D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}})$ under different neighborhood combinations.

Definition 4.4. *Given the data set $D = \{x_1, \cdots, x_n\}$ sampled from $\boldsymbol{X} \in R^d$ and a neighborhood combination $\{k_{X_1}, k_{X_2}, \cdots, k_{X_d}\}$, the normalized neighborhood total association $\boldsymbol{X}$ is defined as:*

$$\begin{aligned} &NNTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(D) = \\ &\frac{NTA(D|C_{k_{X_1}k_{X_2}\cdots k_{X_d}})}{\sum_{i=1}^{d} log\frac{n}{k_{X_i}} - log\frac{n}{\min(k_{X_1},\cdots,k_{X_d})}}. \end{aligned} \tag{15}$$

According to the Theorem 1 and Theorem 4, the $NNTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(D) \in [0, 1]$. However, it's hard to detect a potential relationship just from a specified neighborhood combination in practice. Exploring all possible neighborhood total association and fusing them will be a practical strategy to extract and mining potential association information at different neighborhood combinations. The following definitions show the details.

4.2 Maximal Neighborhood Multivariate Association Measure (MNA)

Now we present the maximal neighborhood multivariate association measure based on normalized neighborhood total association.

Definition 4.5. *Given the data set $D = \{x_1, \cdots, x_n\}$ sampled from $\boldsymbol{X} \in R^p$ and the neighborhood range $NB(n)$, the maximal neighborhood multivariate association coefficient is defined as:*

$$MNA(D) = \max_{\substack{C=\{k_{X_1},\cdots,k_{X_d}\} \\ 1\le k_{X_i}k_{X_j}\le n^{\alpha}, 0<\alpha<1}} NNTA_{C_{k_{X_1}k_{X_2}\cdots k_{X_d}}}(D). \tag{16}$$

The setting of $NB(n)$ is important: $NB(n)$ too high will lead to non-zero score even for statistical independent data, while $NB(n)$ too low means only detect simple patterns. To avoid the invalid parameter design, we need to limit the maximum neighborhood combination $NB(n)$ of covers C. Here we heuristically provide the neighborhood range that satisfy the $k_{X_i}k_{X_j} < n^{\alpha}, i \neq j, i, j = 1, \cdots, d, \alpha \in (0, 1)$. In this paper, $n^{0.7} \sim n^{0.8}$ is effective experimentally. Unless specified otherwise, we use $NB(n) = n^{0.8}$, which works well in practice.

Some properties of MNA are as follows:

D1. Comparable. $MNA(D) \in [0, 1]$. Based on the definition 4.4, the $NNTA(D) \in [0, 1]$ for any sets and any neighborhood combinations, so the MNA(D) not only can be compared across different neighborhood combinations of the same data but also different dimensional multivariate sets.

D2. Interpretable. $MAC(D) = 0$ means the $X_{1,\cdots,d}$ are statistically independent. Due to the limited sample size, the $MNA(D)$ will deviate from theoretical 0 in practical application. The higher the MNA score, the stronger association among $X_{1,\cdots,d}$. $MAC(D) = 1$ indicates that there exists at least a variable X_i such that each $X_j \in \{X_{1,\cdots,d} \setminus X_i\}$ is a function of X_i.

5 Calculating MNA

5.1 Brute-force Approach

To use MNA in practice, a brute-force search strategy is easiest to be thought of. It traverses all possible neighborhood combinations over every dimension and selects an optimal one that maximize the MNA. Specially, for every neighborhood combination $\{k_{X_i}\}_{i=1}^{d-1}$, it fixes the number of neighbors of samples on $d-1$ dimension firstly, and then try to find the optimal neighborhood size of the remaining dimension. For every neighborhood combination, the above operation is repeated per dimension and the maximal value is reported over all dimensions.

However, using the $k_{X_i}k_{X_j} \leq n^{\alpha}, i \neq j$, the search space on all dimensions is $k_{X_1} \times \cdots \times k_{X_d} \leq n^{\alpha d/2}$. Then the size of search space for d-dimensional is $O(N^d)$. It's obviously not going to work for large d in practice.

5.2 Our Approach

In this section, a simple and efficient greedy method for estimating MNA is provided. To compute the MNA(D), the computational intractability is embodied in finding concurrently the optimal neighborhood size of all dimensions that maximize the normalized neighborhood total association (see Eq.(14) and (15)). Thus an efficient search method is required. The intuition is to serialize the above traversal search. That is, the dimension and its corresponding optimal number of neighbors are gradually sought to maximize the normalized neighborhood total association with all selected dimensions and their neighborhood size.

In practice, we first ascertain two dimensions $X_1^{'}$ and $X_2^{'}$ such that $MNA(X_1^{'}, X_2^{'})$ is the maximum among all pairs of variables. Then, at each subsequent step $l \in [2, d-1]$, let $C_k = \{X_1^{'}, X_2^{'}, \cdots, X_k^{'}\}$ is the subset of variables have been picked, and their optimal neighborhood sizes have been determined. The $R_{d-k} = \{X_{k+1}, \cdots, X_d\}$ is the subset of remaining variables. The purpose of the $k+1$ step is: select the variable $X_{k+1}^{'} \in R_{d-k}$ that maximize the association strength with C_k, find the optimal neighborhood size for $X_{k+1}^{'}$ which satisfies $k_{X_{k+1}^{'}} < \frac{N^{\alpha}}{\max(\{k_{X_i^{'}}\}_{i=1}^{k})}$, and then calculate the $MNA(X_{k+1}^{'}, C_k)$ score. Repeat the above steps, we obtain the approximate score of $MNA(D)$ and its corresponding optimal neighborhood combination. Reviewing the above process, it is equivalent to splitting the overall maximization goal into stepwise maximization subgoals. Details of the proposed algorithm is shown in Algorithm 1.

Algorithm 1 Maximal Neighborhood Multivariate Association Measure (MNA)

Input: A finite data set $D = \{x_1, \cdots, x_n\}$, a integer set satisfies $k_{X_i}k_{X_j} \leq n^{\alpha}$, neighborhood parameter $\alpha \in [0, 1)$, current variables set C_1, remaining variables set R_d, optimal neighborhood combination set ON_1, the distance measure d_{X_i} on the $X_i, i = 1, \cdots, d$ space.
Output: $MNA(D)$, the variables sequence C_d, neighborhood sequence ON_d.
Compute the similarity matrix $S_m(X_i), i = 1, \cdots, d$ of each marginal sample set S_{X_i}.
repeat
  for $k = 2$ **to** d **do**
    if $k = 2$ **then**
      Calculate the MNA values of all pairs variables in R_d, select the variable pairs corresponding to the maximum MNC value and add into the set $C_2 = \{X_1^{'}, X_2^{'}\}$, record the neighborhood combination of selected variables and add into the set $ON_2 = \{(k_{X_1^{'}})^*, (k_{X_2^{'}})^*\}$, update the $R_{d-2} = R_d - C_2$.
    if $k > 2$ **then**
      Select the variable $X_k^{'}$ from R_{d-k-1}, search the neighborhood combination while satisfying $k_{X_k^{'}} < \frac{N^{\alpha}}{\max(\{k_{X_i^{'}}\}_{i=1}^{k-1})}$ on similar matrix $S_m(X_k^{'})$, and record neighbors size $(k_{X_k^{'}})^*$ that make $MNA(C_{k-1}, X_k)$ reach the largest. Update the ON_k and R_{d-k}.
  until $R_d = \emptyset$

Note that since all dimensions in C_k have already been fixed gradually, we need not to determine them again, so it can be computed much more efficiently. In addition, with gradual addition of variables, the neighborhood range of remaining variables gradually decreases, it also contributes to boost computing efficiency.

By the way, according to our computing rule, it will occur a phenomena that different neighborhood combinations yield the same MNA score. Here, we agree to take the neighborhood combination with smaller neighborhood sizes, this will offer larger neighborhood search range for remaining variables.

5.3 Complexity Analysis

The time complexity of MNA(D) includes three parts: (1) the cost of computing similarity matrix for all dimensions $O(dN^2)$; (2) the cost of sorting similarity matrix for all dimensions $O(dN^2 logN)$; (3) the cost of find $X_1^{'}$ and $X_2^{'}$

$O(d^2N^{2\alpha})$ and the cost of find subsequent dimensions $O(dN^{2\alpha})$. In fact, the time complexity of subsequent dimensions is less than $O(N^{\alpha})$ for $d \geq 3$ due to the gradually restricted neighborhood range. So the overall time complexity is $O(d^2N^{2\alpha})$. We adopt $\alpha = 0.8$ in implementation, the complexity MNA(D) is $O(d^2N^{1.6})$.

6 Experiments

In this section, we empirically assess the performance of MNA (here the $p = 2$ for l_p-norm. In fact, different ps yield the same result on a single dimension.) First, we study the validity of parameter $NB(n)$ on synthetic datasets. Second, we verify some properties of MNA experimentally. Third, we apply MNA to detect novel associations on real-world data.

We also compare our approach to three state-of-the-art multivariate association measures, namely HICS, MAC and UDS. CMI was not included here due to it is not normalized. The parameters involved in those comparison methods are set according to their respective papers. The results presented are the average values of 100 experiments.

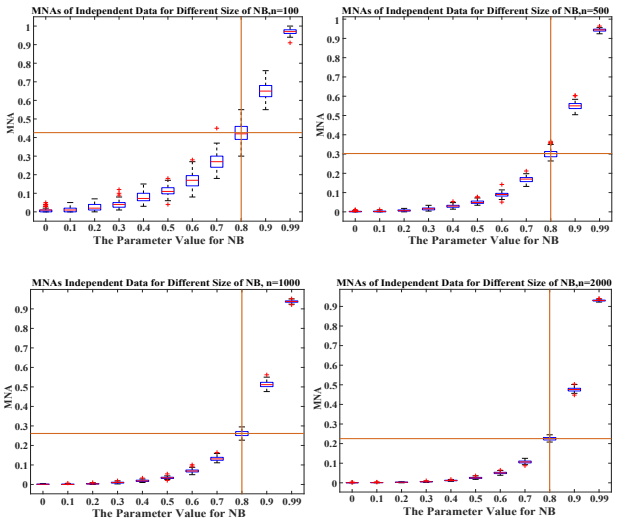

Fig. 2: The performance of different parameter values in $NB(n)$ on bivariate independent data as increasing n.

6.1 Performance on Synthetic Data

Parameter validity. Figure 2 reports different neighborhood parameter $NB(n) = O(n^{\alpha})$ results in MNA scores move towards zero for bivariate statistically independent data as n grows. For every sample size, 100 random data sets are sampled from a uniform distribution for each neighborhood parameter $\alpha \in [0 : 0.1 : 0.9, 0.99]$. The distribution of 11 boxplots in each subfigure demonstrates the incremental changes of MNA with gradually larger parameters, which indicates setting too large neighborhood parameters will increase the strength between independent variables. Further, we see that the MAC changes greatly at the brown cross line in each sample size, where the cross point represents the mean MNA at $NB(n) = n^{0.8}$. In addition, we can see the MNC score at $NB(n) = n^{0.8}$ moves toward 0 with increasing sample size n. This suggests $\alpha = 0.8$ is a statistically reasonable parameter value. Without special assignment , we adopt it in our experiment.

D3. Scalable. The parameter validity results partly verify the scalable of MNA, that is easy to compute as the increasing number of data size. The scalable with respect to dimension can be verified by reliability results.

D4. Intuitive. Only one neighborhood parameter α in MNA(D) should be considered, which controls the range of neighborhood information to be referenced. Users can easily understand the impact of the parameter and control the estimation process.

D5. Reliability. In this section, we identify and analyze the impact of relationship types and dimension deviation on MNA experimentally. To study reliability, six different types relationships with ten different dimensions are used in experimental. The specific forms of the dataset are from [14], [16]: (1) Independent relationship: each X_i variables are mutual independent; (2) Identical relationship: $X_i = X_1$; (3) Power law relationship: $X_i = X_1^i$; (4) Sin relationship: $X_i = \sin(X_{i-1})$; (5) Quadratic relationship: $X_i = X_{i-1}^2 + 2X_{i-1}$; (6) Log relationship: $X_i = log(|X_{i-1}| + 1)$. For each relationship, $i = 2, \cdots, 20$. In each dataset, samples of the variable X_1 comes from a uniform distribution [0,1], samples of other variables are generated depend on the corresponding relationship. Each relationship is formed by 500 sample points. The results in each type dimension are the average of 100 trials.

We report the results in Figure 3. By looking at the variation curves of compared measures and MNA with increasing data dimensions, we see MNA and MAC appear a desired flat line behavior at 1 for (2)-(6) strong different associations, which reveals the MNA can detect different style associations. For (1) independent data, the curve of MNA nearly presents a flat line at 0, (the deviation on 2 dimension may be due to the limited samples), which manifests the MNA is not influenced by the dimensions for detecting independent variables. However, none of other compared measures behave correctly across all six cases. Hence MNA has excellent reliability.

On the side, the above experiments partly explain the **D3. Scalable** of MNA, that is it can be used to measure multivariate variable associations among different dimensions.

D6. Robust. Noisy data may lead to wrong estimates, such as an random data is declared as a strongly association. MNA is a rank statistic, because we use the neighborhood subscript of a point rather than its sample value, so it is sensitive to noise data in theory. Here we test the power of MNA experimentally. The relationships are still adopted in above section, but we report the results for four complex relationships. Each data set is 500 points and fixed 20 dimensions, the 10 noise level $\epsilon \sim \text{Gaussian}(0, \delta)$, we control noise by varying δ.

Figure 4 displays the results. The higher score the better. Though no one measure is consistently superior to the compared measures, MNA outperforms others on three relationships except Log relationship. MAC degenerates very quickly may due to neglecting the local structures of relationship especially on noisier data.

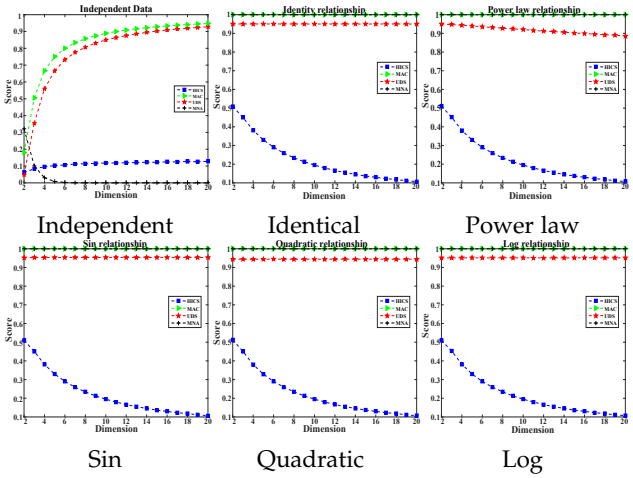

Fig. 3: The performance of MNA and the comparison methods on six relationships.

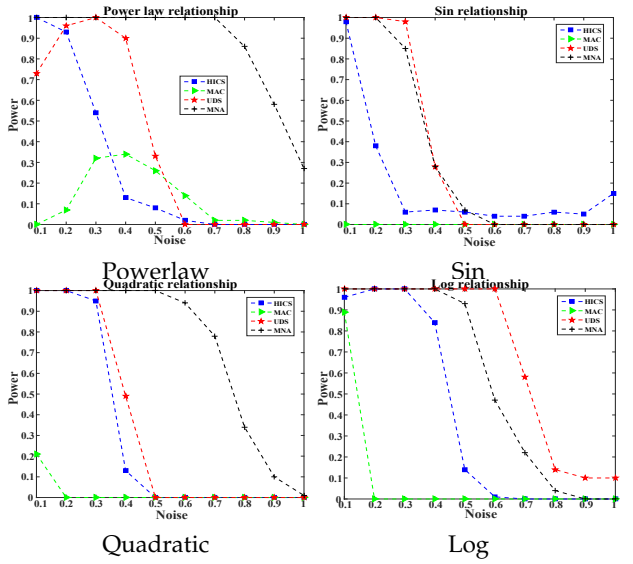

Fig. 4: Power of MNA with regards to different noise.

6.2 Performance on Real Data

Discovering novel associations. For unlabeled data exploratory analysis, one hopes an association measure identify and detect interesting and valuable relationships among multivariable. To evaluate the efficacy of MNA in exploratory analysis, we apply MNA to WHO data set [1], which includes 357 global indicators for 202 countries from 1960 to 2005. There many missing values exist in the data sets, we use linear interpolation method to fill the missing values [40].

Below we present some interesting associations discovered by MNA but its competitors can't discovered. First, we sort the MNA scores and compared methods scores on all pairwise variables, and then compute their scores among three variables based on the top 50 detected bivariate associations. Associations with high MNA score on both two variables and three variables are displayed. Figure 5 shows the detected linear relationships, those associations are intuitively understandable. Figure 6 presents interesting nonlinear associations discovered using MNA, it is obvious a exponential and linear superposition tendency, but other methods fail to. We verified all findings with a domain expert, some haven already known as associations, and others that are novel.

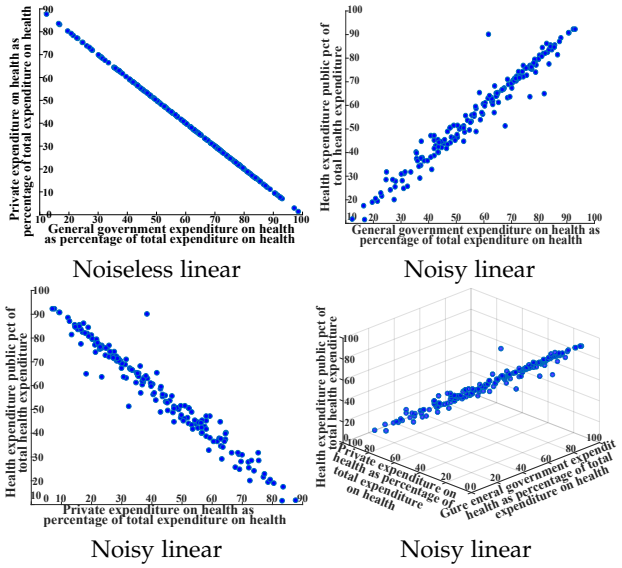

Fig. 5: Linear associations discovered by MNA.

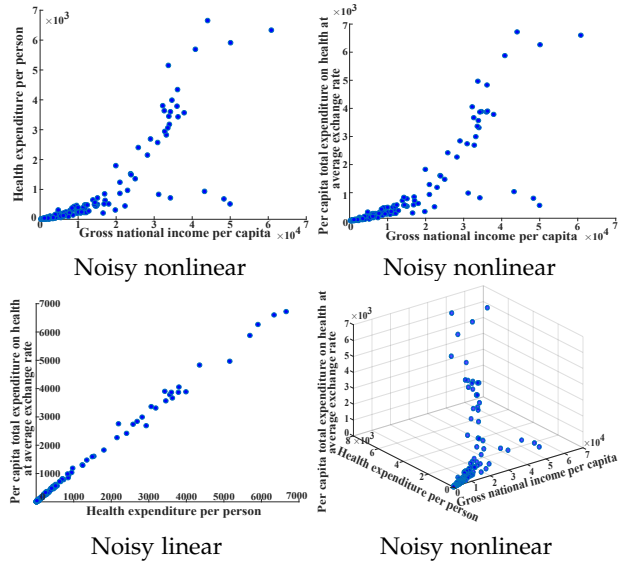

Fig. 6: Nonlinear associations discovered by MNA.

7 Conclusion and Future Work

Exploring and detecting potential association patterns in multivariate data sets is a kind of important tasks in data mining. Data-driven approaches that making no assumption of data distributions nor types of associations are still urgent. In this paper, we introduced the neighborhood information into total association and proposed a MNA measure for multivariate association mining, which fulfils comparable, interpretable, scalable, intuitive, reliability, and robust requirements statistically.

Experimental results convincingly demonstrated that MNA outperforms existing measures. MNA discovers association patterns by exploring the neighborhood structures

of all dimensions, its successful applications provides a novel perspective to complex multivariate association mining task. For further research, we will gain more insight into the statistical nature of neighborhood.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (No.2021ZD0112400), Key Program of the National Natural Science Foundation of China (No. 62136005), Young Scientists Fund of the Natural Science Foundation of Shanxi (No. 20210302124549), Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (2021L286) and Research Project Supported by Shanxi Scholarship Council of China (No. 2020-095).

REFERENCES

[1] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[2] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[3] R. C. Carlos, C. E. Kahn, and S. Halabi, "Data science: big data, machine learning, and artificial intelligence," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 497–498, 2018.

[4] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," *Artificial intelligence*, vol. 174, no. 9-10, pp. 597–618, 2010.

[5] E. Martínez-Gómez, M. T. Richards, and D. S. P. Richards, "Distance correlation methods for discovering associations in large astrophysical databases," *The Astrophysical Journal*, vol. 781, no. 1, p. 39, 2014.

[6] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, no. 2. ACM, 2001, pp. 37–46.

[7] D. E. Zhuang, G. C. Li, and A. K. Wong, "Discovery of temporal associations in multivariate time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2969–2982, 2014.

[8] S. Agrawal, M. Steinbach, D. Boley, S. Chatterjee, G. Atluri, A. T. Dang, S. Liess, and V. Kumar, "Mining novel multivariate relationships in time series data using correlation networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 9, pp. 1798–1811, 2019.

[9] A. C. Rencher, *Methods of multivariate analysis*. John Wiley & Sons, 2003, vol. 492.

[10] T. Metsalu, "Statistical analysis of multivariate data in bioinformatics," 2016.

[11] F. Wang, B. C. Vemuri, M. Rao, and Y. Chen, "A new robust information theoretic measure and its application to image alignment," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2003, pp. 388–400.

[12] H. V. Nguyen, E. Müller, J. Vreeken, P. Efros, and K. Böhm, "Multivariate maximal correlation analysis," in *International Conference on Machine Learning*, 2014, pp. 775–783.

[13] H. V. Nguyen, E. Müller, J. Vreeken, F. Keller, and K. Böhm, "Cmi: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 198–206.

[14] H.-V. Nguyen, P. Mandros, and J. Vreeken, "Universal dependency analysis," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 792–800.

[15] F. Keller, E. Muller, and K. Bohm, "Hics: high contrast subspaces for density-based outlier ranking," in *2012 IEEE 28th international conference on data engineering*. IEEE, 2012, pp. 1037–1048.

[16] Y. Wang, S. Romano, V. Nguyen, J. Bailey, X. Ma, and S.-T. Xia, "Unbiased multivariate correlation analysis," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[17] X. Zeng, Y. Xia, and H. Tong, "Jackknife approach to the estimation of mutual information," *Proceedings of the National Academy of Sciences*, vol. 115, no. 40, pp. 9956–9961, 2018.

[18] P.-R. Loh, G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler *et al.*, "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis," *Nature genetics*, vol. 47, no. 12, pp. 1385–1392, 2015.

[19] T. S. H., "Multiple mutual informations and multiple interactions in frequency data," *Information and Control*, vol. 46, pp. 26–45, 1980.

[20] A. Bargiela and W. Pedrycz, "Granular computing," in *HANDBOOK ON COMPUTATIONAL INTELLIGENCE: Volume 1: Fuzzy Logic, Systems, Artificial Neural Networks, and Learning Systems*. World Scientific, 2016, pp. 43–66.

[21] W. Pedrycz, *Granular computing: analysis and design of intelligent systems*. CRC press, 2018.

[22] S. Xia, Z. Zhang, W. Li, G. Wang, E. Giem, and Z. Chen, "Gbnrs: a novel rough set algorithm for fast adaptive attribute reduction in classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1231–1242, 2020.

[23] K. Liu, X. Yang, H. Fujita, D. Liu, X. Yang, and Y. Qian, "An efficient selector for multi-granularity attribute reduction," *Information Sciences*, vol. 505, pp. 457–472, 2019.

[24] X. S. Rao, J. J. Song, X. B. Yang, K. Y. Liu, and P. Wang, "Neighborhood classifier for label noise," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2020.

[25] Y. Qian, H. Zhang, Y. Sang, and J. Liang, "Multigranulation decision-theoretic rough sets," *International journal of approximate reasoning*, vol. 55, no. 1, pp. 225–237, 2014.

[26] S. Zhao, Z. Dai, X. Wang, P. Ni, H. Luo, H. Chen, and C. Li, "An accelerator for rule induction in fuzzy rough theory," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 12, pp. 3635–3649, 2021.

[27] K. Liu, X. Yang, H. Yu, H. Fujita, and D. Liu, "Supervised information granulation strategy for attribute reduction," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 5150, 2020.

[28] Q. Wang, Y. Qian, X. Liang, Q. Guo, and J. Liang, "Local neighborhood rough set," *Knowledge-Based Systems*, vol. 153, pp. 53–64, 2018.

[29] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, and W. Pedrycz, "Measuring relevance between discrete and continuous features based on neighborhood mutual information," *Expert Systems with Applications*, vol. 38, no. 9, pp. 10 737–10 750, 2011.

[30] X. Yang, H. Chen, T. Li, J. Wan, and B. Sang, "Neighborhood rough sets with distance metric learning for feature selection," *Knowledge-Based Systems*, vol. 224, p. 107076, 2021.

[31] G. Song, J. Rochas, L. El Beze, F. Huet, and F. Magoules, "K nearest neighbour joins for big data on mapreduce: a theoretical and experimental analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2376–2392, 2016.

[32] O. El Barbary, A. Salama, and E. S. Atlam, "Granular information retrieval using neighborhood systems," *Mathematical Methods in the Applied Sciences*, vol. 41, no. 15, pp. 5737–5753, 2018.

[33] Y. Qian, H. Cheng, J. Wang, J. Liang, W. Pedrycz, and C. Dang, "Grouping granular structures in human granulation intelligence," *Information Sciences*, vol. 382, pp. 150–169, 2017.

[34] H. Cheng, Y. Qian, H. U. Zhiguo, and J. Liang, "Association mining method based on neighborhood perspective," *Scientia Sinica Informationis*, vol. 50, no. 6, p. 824, 2020.

[35] S. Mukherjee, H. Asnani, and S. Kannan, "Ccmi: Classifier based conditional mutual information estimation," in *Uncertainty in artificial intelligence*. PMLR, 2020, pp. 1083–1093.

[36] D. V. Lindley and L. Phillips, "Inference for a bernoulli process (a bayesian view)," *The American Statistician*, vol. 30, no. 3, pp. 112–119, 1976.

[37] R. A. Askey and R. Roy, "Gamma function." 2010.

[38] J. G. Saw, M. C. Yang, and T. C. Mo, "Chebyshev inequality with estimated mean and variance," *The American Statistician*, vol. 38, no. 2, pp. 130–132, 1984.

[39] K. Das, J. Jiang, J. Rao *et al.*, "Mean squared error of empirical predictor," *Annals of Statistics*, vol. 32, no. 2, pp. 818–840, 2004.

[40] M. N. Noor, A. S. Yahaya, N. A. Ramli, and A. M. M. Al Bakri, "Filling missing data using interpolation methods: Study on the effect of fitting distribution," *Key Engineering Materials*, vol. 594-595, pp. 889–895, 2014.

Honghong Cheng received the BS degree from the School of Mathematical Sciences, Shanxi University, Taiyuan, China, in 2012 and the PhD degree from the Institute of Big Data Science and Industry, Shanxi University, Taiyuan, China, in 2020. She is currently a teacher at the School of Information, Shanxi University of Finance and Economics. She was a visiting scholar with the City University of Hong Kong, Hong Kong, China, in 2019. Her research interests include associations mining, multi-modal learning, data mining.

Yuhua Qian received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively. He is currently a Director at the Institute of Big Data Science and Industry, Shanxi University, where he is also a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education. His research interests include artificial intelligence, data mining, machine learning, granular computing and machine vision. He has authored over 100 articles on these topics in international journals.

Yingjie Guo received a BS degree in 2009 from the School of Computer and Information Technology, Shanxi University, Taiyuan, China, and a PhD degree in 2019 from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. From 2014-2016, She was a visiting scholar at Cornell University, Ithaca, USA. She is currently a teacher at the Institute of Big Data Science and Industry, Shanxi University. Her research interests include bioinformatics, association analysis and machine learning.

Keyin Zheng received a B.S. degree in information and computing science and Master's degree in pattern recognition and intelligent system at school of Mathematical Sciences from Shanxi University, China, in 2012 and 2015, respectively. She is a PhD candidate at Institute of Big Data Science and Industry, Shanxi University. Her research interest includes concept learning, associations mining and machine learning.

Qingfu Zhang (M01-SM06-F17) received the BSc degree in mathematics from Shanxi University, China in 1984, the MSc degree in applied mathematics and the PhD degree in information engineering from Xidian University, China, in 1991 and 1994, respectively. He is a Chair Professor of Computational Intelligence at the Department of Computer Science, City University of Hong Kong. His main research interests include evolutionary computation, optimization, neural networks, data analysis, and their applications. Dr. Zhang is an Associate Editor of the IEEE Transactions on Evolutionary Computation and the IEEE Transactions on Cybernetics.