International Journal of Data Warehousing and Mining
Volume 13 • Issue 1 • January-March 2017

Improved Approximation Algorithm for Maximal Information Coefficient

Shuliang Wang, School of Software, Beijing Institute of Technology, Beijing, China

Yiping Zhao, Software Center, Bank of China, Beijing, China

Yue Shu, Tencent Technology (Beijing) Company Limited, Beijing, China

Wenzhong Shi, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

ABSTRACT

A novel statistical maximal information coefficient (MIC) that can detect the nonlinear relationships in large data sets was proposed by Reshef et al. (2011), with emphasis being placed on the equitability, which is a very important concept in data exploration. In this paper, an improved algorithm for approximation of the MIC (IAMIC) is proposed for the development of the equitability. Based on quadratic optimization processes, the IAMIC can search for a more optimal partition on the y-axis rather than use that which was obtained simply through the equipartition of the y-axis, to enable it to come closer to the true value of the MIC. It has been proved that the IAMIC can search for a local optimal value while using a lower number of iterations. It has also been shown that the IAMIC provides higher accuracy and a more acceptable run-time, based on both a mathematical proof and the results of simulations.

KEYWORDS

Accuracy, Big Data, Equitability, MIC, Quadratic Optimization

1. INTRODUCTION

With the large amounts of data that are being generated in various fields, and particularly in the biological and geological information field (Shekar and Xiong, 2007; Ester et al., 2000; Meyer-Schoenberger and Cukier, 2013; Surhone et al., 2010; Wang and Yuan, 2014) information field, which is growing exponentially (Hastie et al., 2009; Howe and Rhee, 2008), there is a critical need to extract the meaning from the large datasets that are obtained (Frankel and Reid, 2008; United Nations Global Pulse, 2012; McKinsey Global Institute, 2011; Rajaraman and Ullman, 2011; Vatsavai et al., 2012). However, the real-world data is always dirty, which prevents researchers to reveal the real meaning behind the data (Hernandez and Stolfo, 1998; Kim et al., 2003; Dasu, 2003; Smets, 1996). Reshef et al. (2011) proposed a new measure to identify relationships between two variable pairs, called the maximal information coefficient (MIC), and it is an interesting approach that can be used to discover relationships between variables in large data sets because of two properties: generality and equitability. Generality of MIC means it can detect more interesting relationships, including functional or not, not limited to certain functional types as Pearson correlation coefficient or Spearman rank correlation coefficienct, etc. Furthermore, Equitability ensures MIC to explore the datasets impartially, with no inclinations to specific relation types.

MIC is a powerful tool for mining the correlation between random variables, which quantify the relationship to the range from 0 to 1. The closer to 1 means closer relationship, oppositely closer to

DOI: 10.4018/IJDWM.2017010104

0 says more likely two independent variables. However, the algorithm used by Reshef et al. (2011) to compute the MIC can only obtain an approximation. In this case, the accuracy of the algorithm will directly affect similarity judgement. The experiment shows that with the improvement of the accuracy of MIC, the rank of pairs according to MIC value changes, and especially some pairs of variables rank much higher. In other words, the accuracy of MIC has a great influence on the measure of dependence for each pair. Because some pairs that are strongly related are more likely be overlooked results from low accuracy of MIC, so as MIC come closer to its true value, we can detect more valuable associations.

Moreover, the standard approximation algorithm to compute the MIC resulted in some deviations from the equitability property. Therefore, the need remains to fully explore its properties to enhance the MIC.

Here, we propose an improved approximation algorithm for MIC, called the IAMIC, which offers a better solution coupled with a reasonable run-time. A mathematical proof of the ability to search for extreme values that requires fewer iterations is also provided. A comparison between the proposed algorithm and the original MIC algorithm provided by Reshef et al. is also given.

Also, the MIC has lower statistical power than distance correlation methods (Szeleky et al., 2007; 2009) in many important relationships, as noted by Simon and Tibshirani (2012), and it is claimed that this power drawback could cause MIC to lose its advantage for general use. In this paper, we propose a probable hypothesis to argue the case of the low power problem of MIC based on the results of our experiments.

The rest of this paper is organized as follows. In section 2, the logic flow of the IAMIC is presented. In section 3, the mathematical proof of the improvement offered by the IAMIC is provided. In section 4, the IAMIC is compared with the algorithms used by Reshef et al. in terms of their effectiveness and accuracy. In Section 5, the work focuses on a discussion of the assumption of associations between the MIC with power and the IAMIC. Section 6 summarizes the work that has been done previously in this field.

2. AN IMPROVED APPROXIMATION ALGORITHM FOR MIC

Reshef et al. (2013) modified the original algorithm to produce an approximation algorithm that was less efficient but had greater accuracy, which exhaustively searched an equipartition of the y-axis for up to 20 rows to find the best subpartition into 2 or 3 rows for all grids with 2 or 3 rows, respectively, instead of a simple equipartition of the y-axis into 2 or 3 rows. As the experimental results have shown, this more exhaustive algorithm has better equitability than the original algorithm, which meant that the deviations from the equitability of the original MIC values were a result of the accuracy of the approximation algorithm, rather than the nature of the MIC.

Reshef et al. (2013) also stated that the use of the approximation algorithm affects the equitability, which motivates us to propose an improved approximation algorithm, which is presented in this paper, to generate a better characteristic matrix, making it possible to obtain higher MIC values to improve the equitability in an acceptable run-time. Also, the experiments with the exhaustive algorithm have inspired us to search for a more optimal y-axis partition than the y-axis equipartition.

In our work, we attempt to find a better partition on the y-axis than the simple equipartition by quadratic optimization as an alternative to the exhaustive searching method. We then provide a simple improved algorithm called IAMIC for MIC. Specifically, the logical path of quadratic optimization is described as follows: given a specific number of rows y_1, first equipartition the y-axis into y_1 rows and optimize the x-axis in the manner of Reshef et al. (2011). Second, select the grid partition of integers (x_1, y_1) with the largest normalized mutual information and get a detailed partition of size x_1. Then, fix the partition of size x_1 on the x-axis and optimize the y-axis into y_1 rows, and then calculate the largest normalized mutual information under the new partition of integers (x_1, y_1) to replace the original value given in the characteristic matrix. After generation of the whole characteristic matrix, we select the maximum score in the characteristic matrix as the MIC.

In other words, the IAMIC uses quadratic optimization on the y-axis just once to search for a more optimal y-axis partition, rather than the approach of the exhaustive method. For each row of the characteristic matrix generated by the algorithm of Reshef et al., we attain a better solution based on the largest value of this row. To present the IAMIC clearly, the detailed steps used to produce the characteristic matrix are listed in Table 1. The parameter *B* in Table 1 is the function of sample size that was introduced by Reshef et al. (2011).

Given y=2, the simple algorithm simulation process is shown in Figure 1 (a) to Figure 1 (e).

Actually, on the y-axis equipartition of Figure 1 (a), the grid partition that is displayed as Figure 1 (b) achieves the largest mutual information, simply by using the original algorithm of Reshef et al. (2011). Also, the IAMIC adds the rest of the steps (Figure 1 (c)-(e)) to search for the optimal grid. From the above statements, the equipartition on the y-axis is a subset of the quadratic optimization on the y-axis with the same partition on the x-axis; thus, the quadratic optimization on the y-axis achieves a better maximal normalized mutual information score than all the scores in one row of the original characteristic matrix under the condition of y-axis equipartition. The new IAMIC algorithm improved the maximum score of the original characteristic matrix for each row, while the MIC has the largest score in the characteristic matrix, and thus the MIC calculated by the use of the IAMIC can possibly achieve greater accuracy. The degree of improvement is equivalent to that of the improved exhaustive algorithm, which is proven later in section 4.

Make m equal *B/2*, which is the upper limit of the y values. Also, if we make k_x be the number of clumps on the x-axis, then the time complexity of the standard algorithm introduced by Reshef et al. (2011) was $O(mk_x^{\ 2}xy)$. Similarly, let k_y be the number of clumps on the y-axis, while the IAMIC includes two parts, where one is used to find the optimal solution on the x-axis and the other is the second optimal process of the y-axis, and thus we can conclude that the run time is $O(mk_x^{\ 2}xy + mk_y^{\ 2}yx)$, which is a constant multiplied by the time complexity of the algorithm introduced by Reshef et al. (2011).

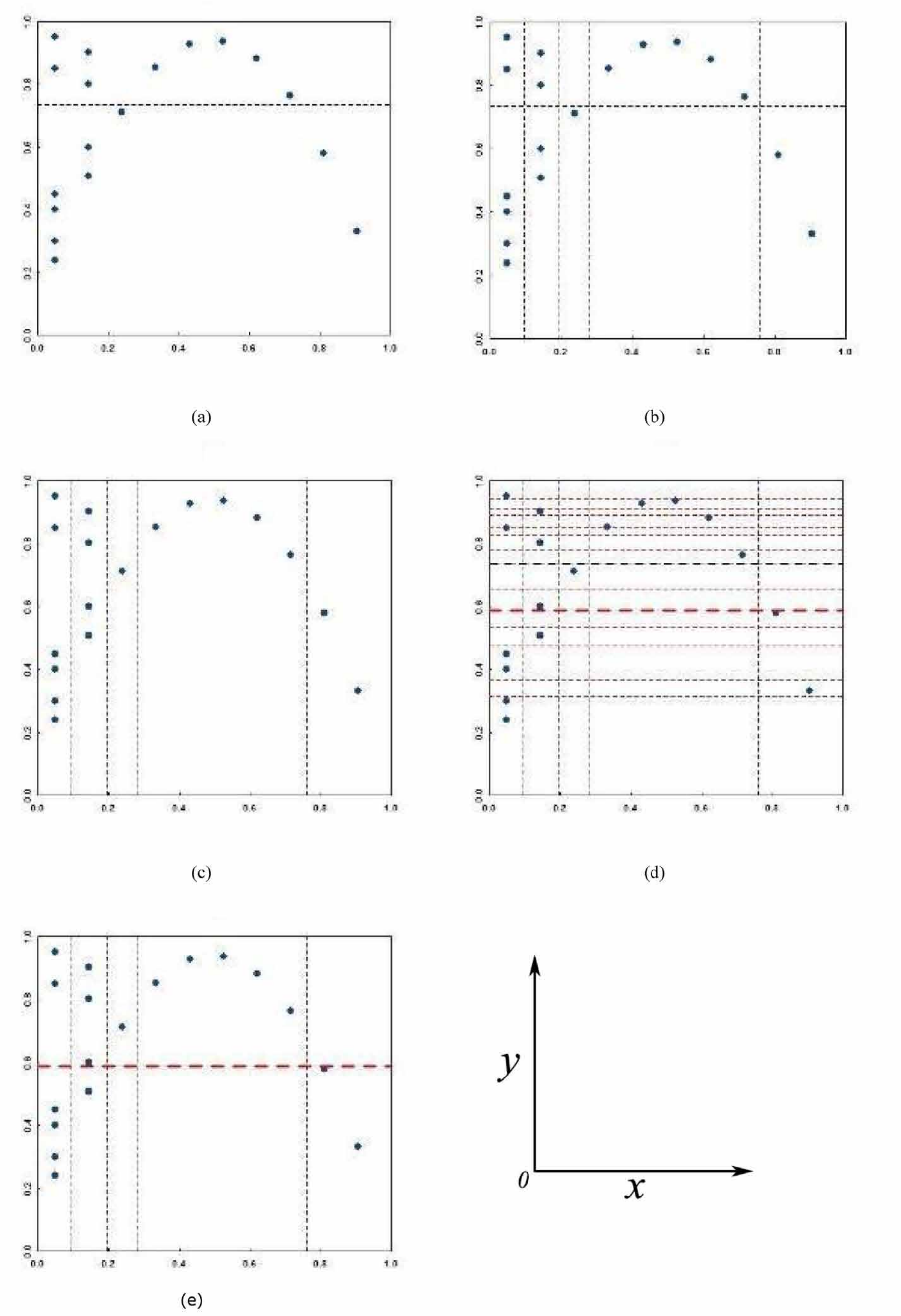
Table 1. The pseudocode of the proposed algorithm

Input		**A set of ordered pairs and some parameters**

Process	a)
	b)
	c)
	d)
	e)
	f)
		for $y\in\{2,3,\ldots B/2\}$, given $y_1 = y$ {
		equipartition y-axis to y1 rows
		for $x\in\{2,3,\ldots B/y_1\}${
		calculate the largest mutual information I(x,y_1)
		}
		let x=x_1, and I(x_1,y_1) is the largest normalized mutual information of all I(x,y_1)
		fix the optimal partition of x1 on x-axis
		for the pair of integers (x_1,y_1), let $\|y_2\|=\|y_1\|$, calculate the largest normalized mutual information I(x_1,y_2) and replace y_1 by y_2.
		}
		Switch the axes, repeat the above steps and obtain the maximal scores

Output	The characteristic matrix M(x,y), each element of which is the largest normalized mutual information achieved by any x-by-y grid	

Figure 1. The simple process of algorithm simulation



(a)

(b)

(c)

(d)

(e)

3. MATHEMATICAL PROOF OF THE IMPROVEMENT OF IAMIC

Previously, we proposed an exhaustive quadratic optimization approach for computation of the MIC (EQOMIC), which is time consuming but offers high accuracy. In this section, we aim to prove that the equitability improvement of the IAMIC is equivalent to that of the EQOMIC.

First, we introduce the EQOMIC algorithm. For each element of the characteristic matrix, we conduct a search of the optimal grid by a second optimization on the y-axis. To be specific, given a y-axis partition of size y_l, we then compute the largest mutual information for each grid partition ((*x*, y_l), *x*∈{*2,3,..., B/* y_l}) using the algorithm provided by Reshef et al. (2011) and fix every kind (*x*∈{*2,3,... B/* y_l}) to repartition the y-axis to correspond to the x-axis partitions. Analogously, the time complexity of this more complex algorithm is $O(mk_x^{\ 2}xy + mk_y^{\ 2}yx^2)$

However, in the IAMIC, we simply fix the x-axis partition only, which is the maximum I(*x*, y_l). Then, for a given *y* row, we search for the largest value in the EQOMIC results by repartitioning the y-axis once only. This means the two algorithms described above for computation of the MIC are equivalent, and thus we can search for a local extremum while using fewer iterations.

Now, we prove that the IAMIC can reduce the EQOMIC time complexity under the premise that the MIC accuracy remains the same.

First, we prove that the IAMIC leads to the same MIC value as EQOMIC. In IAMIC, after dividing the y-axis equally and fixing the y-axis partition $Q = \{0 = r_0, ..., r_l = n\}$, we let $P = \{0 = c_0, ..., c_l = m\}$ be an x-axis partition that maximizes $H(P) - H(P,Q)$. Therefore, for any other partition set $P' = \{0 = c'_0, ..., c'_{l'} = m'\}$,if $H(P') - H(P',Q) \leq H(P) - H(P,Q)$, then we know from the definition of maximal mutual information that $P' \subseteq P$. Then, we fix this x-axis partition and repartition the y-axis, as we did in EQOMIC. Let $S' = \{0 = s'_0, ..., s'_{t'} = n1\}$ and $S = \{0 = s_0, ..., s_t = n2\}$ be the possible y-axis partition sets caused by P' and P, respectively. Then, we have $S' \subseteq S$. Because this y-axis partition size should be equal to $|Q|$, we select Q_1 and Q_1' from S and S' to ensure that $|Q| = |Q_1| = |Q_1'|$. This selection also maximizes $I(P';Q_1') = H(P') - H(P',Q_1') + H(Q_1')$ and $I(P;Q_1) = H(P) - H(P,Q_1) + H(Q_1)$. This means that proving that MIC accuracy is retained is equivalent to proving the following inequality.

$$\begin{aligned} I(P';Q_1') &= H(P') - H(P',Q_1') + H(Q_1') \\ &\leq H(P) - H(P,Q_1) + H(Q_1) = I(P;Q_1) \end{aligned} \tag{1}$$

Because $S' \subseteq S$, we can also pick Q_1' from S. The following inequality is obvious.

$$\begin{aligned} I(P;Q_1') &= H(P) - H(P,Q_1') + H(Q_1') \\ &\leq H(P) - H(P,Q_1) + H(Q_1) = I(P;Q_1) \end{aligned} \tag{2}$$

Let $F(P,P') = I(P;Q_1') - I(P';Q_1')$. Then, we obtain the following function.

$$F(P,P') = [H(P) - H(P,Q_1')] - [H(P') - H(P',Q_1')] \tag{3}$$

Theorem 1. We will have that $F(P,P') \geq 0$.
Proof. Suppose three have three situations.

1. If Q_1' simply divides the point sets belonging to the different x clumps into the different y clumps, i.e. Q_1' does not go through any point set that was partitioned by P, we get the idea that $H(P) - H(P,Q_1') = 0$ from the supporting online material (SOM) of the work

of Reshef et al. (2011). Because $P' \subseteq P$, that means even though Q_1' does not go through any point sets partitioned by P, it may go through some point sets that were partitioned by P'. Thus, it is easy to find that $H(P') - H(P', Q_1') \leq 0$ from the SOM. In this situation, $F(P, P') \geq 0$. This situation is present in Table 2-A.

If Q_1' goes through some point sets that were partitioned by P, we then choose the simplest state to establish the mathematical derivation. Because $P' \subseteq P$, every clump that is divided by P' is equal to some adjacent clumps that were divided by P. Assume that clump D_k' is generated by P' and that D_l and D_{l+1} are generated by P. Let $D_k' = D_l + D_{l+1}$, and Q_1' just crosses D_l, separating it into D_{l1} and D_{l2}. Suppose that D_{l1}, D_{l2}, and D_{l+1} contain u_1, u_2, and v_1 points, respectively; then from the SOM of Reshef et al. (2011), we learned the following equation.

$$\begin{aligned} H(P) - H(P, Q_1') &= \sum_{j=1}^{|P|} \frac{\#_{*,j}}{m} \log \frac{m}{\#_{*,j}} - \sum_{j=1}^{|P|} \sum_{i=1}^{|Q_1'|} \frac{\#_{i,j}}{m} \log \frac{m}{\#_{i,j}} \\ &= \sum_{j=1}^{|P|} \sum_{i=1}^{|Q_1'|} \frac{\#_{i,j}}{m} \log \frac{\#_{i,j}}{\#_{*,j}} \end{aligned} \tag{4}$$

Therefore, the counterpart's residual between P and P' is present below.

$$\begin{aligned} &\sum_{j=l}^{l+1} \sum_{i=1}^{|Q_1'|} \frac{\#_{i,j}}{m} \log \frac{\#_{i,j}}{\#_{*,j}} - \sum_{j=k}^{k} \sum_{i=1}^{|Q_1'|} \frac{\#'_{i,j}}{m} \log \frac{\#'_{i,j}}{\#'_{*,j}} \\ &= \left(\sum_{i=1}^{|Q_1'|} \frac{\#_{i,l}}{m} \log \frac{\#_{i,j}}{\#_{*,j}} + \sum_{i=1}^{|Q_1'|} \frac{\#_{i,l+1}}{m} \log \frac{\#_{i,l+1}}{\#_{*,l+1}} \right) - \sum_{i=1}^{|Q_1'|} \frac{\#'_{i,k}}{m} \log \frac{\#'_{i,k}}{\#'_{*,k}} \\ &= \left(\frac{u_1}{m} \log \frac{u_1}{u_1 + u_2} + \frac{u_2}{m} \log \frac{u_2}{u_1 + u_2} + \frac{v_1}{m} \log \frac{v_1}{v_1} \right) - \left(\frac{u_1 + v_1}{m} \log \frac{u_1 + v_1}{u_1 + u_2 + v_1} + \frac{u_2}{m} \log \frac{u_2}{u_1 + u_2 + v_1} \right) \\ &= \left(\frac{u_1}{m} \log \frac{u_1}{u_1 + u_2} - \frac{u_1 + v_1}{m} \log \frac{u_1 + v_1}{u_1 + u_2 + v_1} \right) + \left(\frac{u_2}{m} \log \frac{u_2}{u_1 + u_2} - \frac{u_2}{m} \log \frac{u_2}{u_1 + u_2 + v_1} \right) \end{aligned} \tag{5}$$

Because
$$\begin{aligned} &\frac{u_2}{m} \log \frac{u_2}{u_1 + u_2} - \frac{u_2}{m} \log \frac{u_2}{u_1 + u_2 + v_1} \\ &= \frac{u_2}{m} \log \frac{u_1 + u_2 + v_1}{u_1 + u_2} = \frac{u_2}{m} \log \left(1 + \frac{v_1}{u_1 + u_2} \right) > 0 \end{aligned} \tag{6}$$

Let $G(x) = \dfrac{u_1 + x}{m} \ln \dfrac{u_1 + x}{u_1 + u_2 + x}$ (7)

where u_1 and u_2 are constants, and x is variable.

Based on $G'(x)$ and $G''(x)$,

$$\begin{aligned} G'(x) &= \frac{1}{m}\ln\frac{u_1+x}{u_1+u_2+x} + \frac{u_1+x}{m}\frac{u_1+u_2+x}{u_1+x}\left(\frac{1}{u_1+u_2+x} - \frac{u_1+x}{(u_1+u_2+x)^2}\right) \\ &= \frac{1}{m}\ln\frac{u_1+x}{u_1+u_2+x} + \frac{1}{m}\frac{u_2}{u_1+u_2+x} \end{aligned} \tag{8}$$

$$\begin{aligned} G''(x) &= \frac{1}{m}\frac{1}{u_1+x}\frac{u_2}{u_1+u_2+x} - \frac{1}{m}\frac{u_2}{(u_1+u_2+x)^2} \\ &= \frac{1}{m}\frac{1}{u_1+x}\left(\frac{u_2}{u_1+u_2+x}\right)^2 > 0 \end{aligned} \tag{9}$$

Because $G''(x) > 0$, we then have $G'(x) < \lim_{x\to\infty} G'(x) = 0$. Now we know that $G(x)$ is a decreasing function. If we bring this back to function (5), we then have $G(x) \le G(0)$.

So, $\frac{u_1}{m}\log\frac{u_1}{u_1+u_2} - \frac{u_1+v_1}{m}\log\frac{u_1+v_1}{u_1+u_2+v_1} \ge 0$

and we get $\sum_{j=l}^{l+1}\sum_{i=1}^{|Q_1'|}\frac{\#_{i,j}}{m}\log\frac{\#_{i,j}}{\#_{*,j}} - \sum_{j=k}^{k}\sum_{i=1}^{|Q_1'|}\frac{\#'_{i,j}}{m}\log\frac{\#'_{i,j}}{\#'_{*,j}} \ge 0$.

Therefore,

$$\begin{aligned} F(P,P') &= [H(P) - H(P,Q_1')] - [H(P') - H(P',Q_1')] \\ &= \left(\sum_{j=l}^{l+1}\sum_{i=1}^{|Q_1'|}\frac{\#_{i,j}}{m}\log\frac{\#_{i,j}}{\#_{*,j}} - \sum_{j=k}^{k}\sum_{i=1}^{|Q_1'|}\frac{\#'_{i,j}}{m}\log\frac{\#'_{i,j}}{\#'_{*,j}}\right) + \\ &\quad \left(\sum_{\substack{j=1\\ j\ne l,l+1}}^{|P|}\sum_{i=1}^{|Q_1'|}\frac{\#_{i,j}}{m}\log\frac{\#_{i,j}}{\#_{*,j}} - \sum_{\substack{j=1\\ j\ne k}}^{|P'|}\sum_{i=1}^{|Q_1'|}\frac{\#'_{i,j}}{m}\log\frac{\#'_{i,j}}{\#'_{*,j}}\right) \end{aligned} \tag{10}$$

Recall that $\sum_{\substack{j=1\\ j\ne l,l+1}}^{|P|}\sum_{i=1}^{|Q_1'|}\frac{\#_{i,j}}{m}\log\frac{\#_{i,j}}{\#_{*,j}} - \sum_{\substack{j=1\\ j\ne k}}^{|P'|}\sum_{i=1}^{|Q_1'|}\frac{\#'_{i,j}}{m}\log\frac{\#'_{i,j}}{\#'_{*,j}} \ge 0$

Therefore, $F(P,P') \ge 0$. This situation is present in Table 2-B.

3. If Q_1' crosses more than one point set, as Reshef et al. indicated in their SOM [1], we achieve the optimal solution when we draw the x-axis partition on the edges of clumps (runs of consecutive points that fall in the same row of the y-axis partition); this means that the key factor for maximization of mutual information is the partitioning of more clumps to gather points. Because the region obtained from P is a kind of subdivision of that obtained from P', in the same y

partition Q_1', P will make the points gather together more effectively, resulting in larger mutual information. Thus, $F(P,P')\geq 0$. This situation is present in Table 2-C.

In conclusion, $F(P,P')\geq 0$.

Therefore, $I(P';Q_1')\leq I(P;Q_1')\leq I(P;Q_1)$. This means that the IAMIC leads to the same MIC value as the EQOMIC.

Now, we prove that the IAMIC reduces the time complexity. As we described above, in EQOMIC, every time we compute the largest possible mutual information under a specific partition size, that represents an element in the feature matrix. Thus, in EQOMIC, we must perform n^2 computations. However, when we use IAMIC, the largest possible mutual information that we compute at one time represents the maximum value of one row in the feature matrix. Therefore, we need only perform *n* computations in total.

4. EXPERIMENTS AND COMPARASIONS

It has been proved that the IAMIC provides higher accuracy with an acceptable time performance, as described above. In this section, we present the performance of the IAMIC in terms of the experimental results for two aspects: the characteristic matrix and the equitability. The experiments show that the IAMIC realizes a better equitability.

4.1. Imrovements of the Characteristic Matrix

For each function in Table 3, we first generated one dataset of 2300 points spaced evenly along the curve described by the function, each of which is displayed in Figure 2 (a0)-(e0); then, we created another dataset with the same sample size by adding uniform vertical noise, as shown in Figure 3 (f0)-(j0), in which R^2 (the coefficient of determination) equals 0.64. We then compute the characteristic

Table 2. EQOMIC and IAMIC

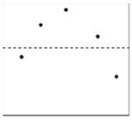
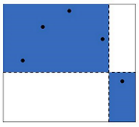
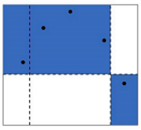
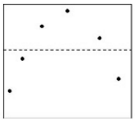
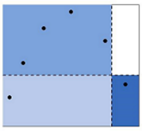
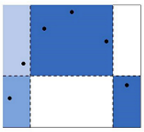
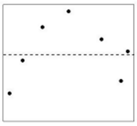
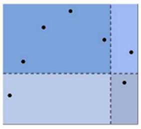
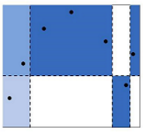
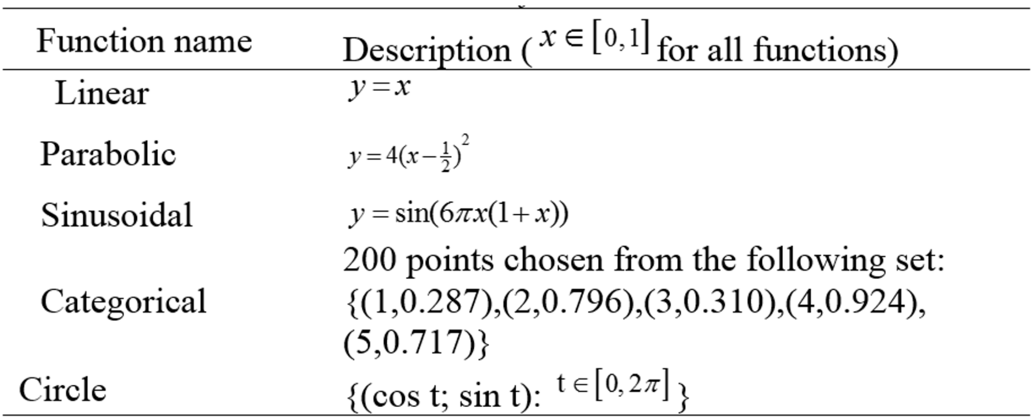
	Equipartition y-axis Q	EQOMIC $I(P';Q_1')$	VS.	IAMIC $I(P;Q_1')$
A			≤	
B			≤	
C			≤	

Table 3. The function list



Function name	Description ($x \in [0,1]$ for all functions)
Linear	$y = x$
Parabolic	$y = 4(x - \frac{1}{2})^2$
Sinusoidal	$y = \sin(6\pi x(1+x))$
Categorical	200 points chosen from the following set: {(1,0.287),(2,0.796),(3,0.310),(4,0.924),(5,0.717)}
Circle	{(cos t; sin t): $t \in [0, 2\pi]$ }

matrix of the two data sets using both the original algorithm and the new algorithm. In this work, we set the value of the exponent α as 0.7 in the function $B(n) = n^{\alpha}$, and we only show the results for the range where $1 < x \leq 15$ and $1 < y \leq 15$.

Figure 2 (a1)-(e1) and Figure 3 (f1)-(j1) correspond to the visualizations of the characteristic matrices of the relationships shown in Figure 2 (a0)-(e0) and Figure 3 (f0)-(j0), respectively, where the characteristic matrices are generated by the previous algorithm of Reshef et al. (2011).

Figure 2 (a2)-(e2) and Figure 3 (f2)-(j2) correspond to the visualizations of the characteristic matrices of the relationships shown in Figure 2 (a0)-(e0) and Figure 3 (f0)-(j0), respectively, where the characteristic matrices are created by the IAMIC algorithm that is provided in this work.

Figure 2 (a3)-(e3) and Figure 3 (f3)-(j3) correspond to the visualizations of the residual matrices of the relationships shown in Figure 2(a0)-(e0) and Figure 3 (f0)-(j0), respectively, where the residual matrices are produced by 10 times $(x1\text{-}x2)(x \in \{a,b,\ldots,e\})$.

4.2. Improvement of the Equitability

Here, we select the 16 different functional relationships that are described in Table 3. For each relationship, we produce a noiseless data series with a sample size n=1000 and 249 additional data series with the same sample size by adding incremental uniform vertical noise to analyze the equitability. We also iterate the above steps 100 times. Figure 4 (a) shows the results of the standard algorithm for approximation of the MIC, while Figure 4 (b) shows the results of using the IAMIC to calculate the MIC under the same conditions, which comes closer to the true value of the MIC and develops the equitability of the reported MIC values.

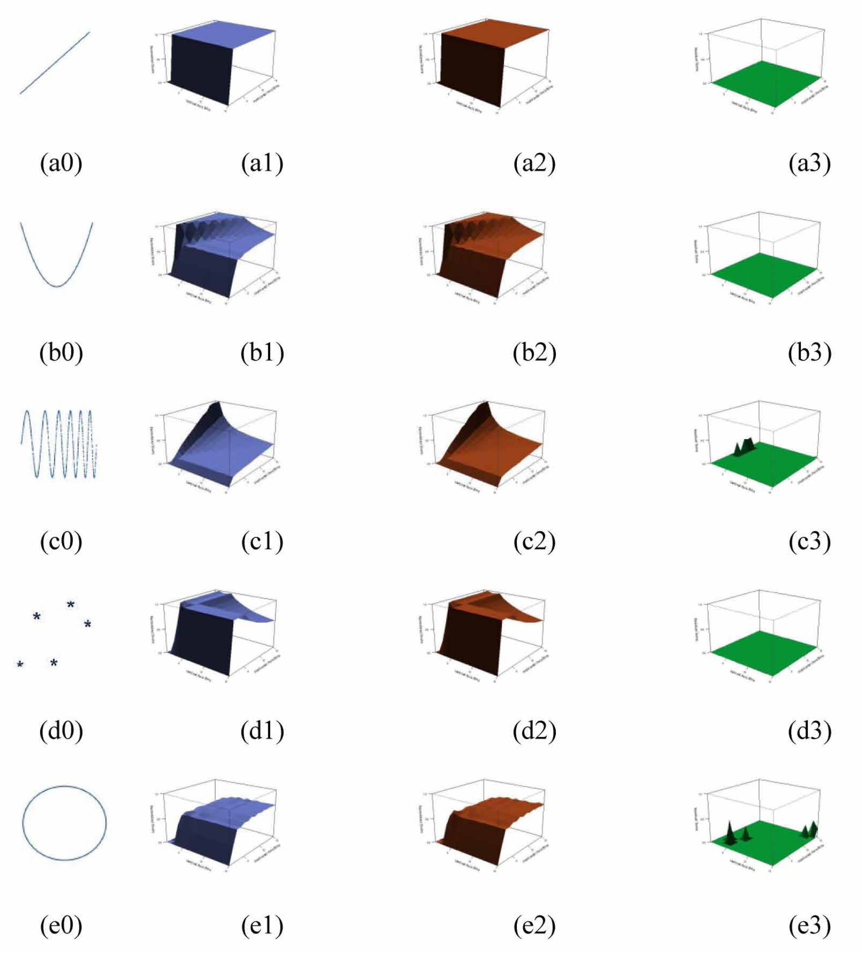
The legend of Figure 4 (a)-(b) is shown in Figure 4 (c), which lists all the colors that correspond to each type of functional relationship.

In this paper, we also add Gaussian noise rather than uniform vertical noise to perform the same equitability analysis that was performed in Figure 4, except that we define the function of the exponential as $x \in [0, 10]$ instead of $x \in [0, 1]$ in Table 4. Figure 5(a)-(c) correspond to Figure 4(a)-(c), respectively.

5. MIC AND POWER

While the MIC has the advantage of equitability, Simon and Tibshirani (2012) indicated that the MIC has lower statistical power than distance correlation, which is another measure of the dependence

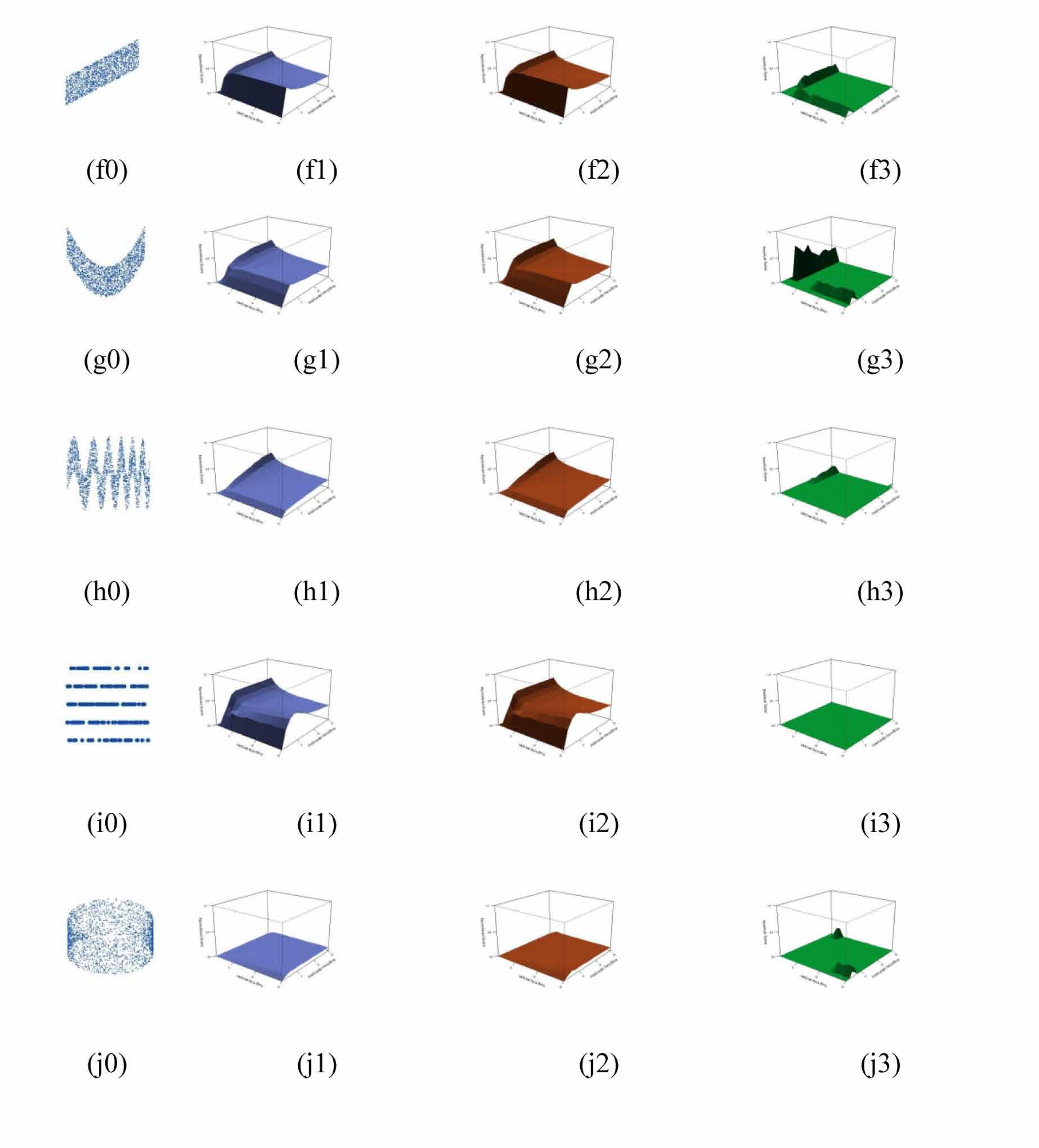
Figure 2. A comparison of the characteristic matrices of noiseless relationships



given by Székely and Rizzo (2009). Sometimes, the MIC is less powerful than Pearson correlation. Because of the power drawbacks, it has been said that the MIC is not an appropriate measure for data exploration. Here, we run the same simulations to compare the power of the values calculated with IAMIC in our paper to that of the values calculated with MIC in Reshef et al (2011), which is computed by the standard approximation algorithm, Pearson correlation and distance correlation (dcor).

It is shown that the improved algorithm has greater accuracy, and thus the IMIC comes closer to the nature of the MIC values. Also, as shown in Figure 6, the IMIC is always more powerful than the MIC, and IMIC has higher power than dcor for the Circle function in particular, so we propose an assumption that the approximation algorithm affects the power of the MIC rather than the intrinsic behavior of the MIC. Because the current algorithm cannot reach the true value of the MIC, the conclusion that the MIC has serious power deficiencies is arbitrary. In other words, if a globally optimal algorithm is found for the MIC, there may be little difference between the MIC and other methods in terms of power properties. Consequently, the power issues of the MIC may have to be reconsidered.

Figure 3. A comparison of the characteristic matrices of noisy relationships



(f0) (f1) (f2) (f3)

(g0) (g1) (g2) (g3)

(h0) (h1) (h2) (h3)

(i0) (i1) (i2) (i3)

(j0) (j1) (j2) (j3)

6. RELATED WORK

A variety of data mining algorithms are provided to expose the relation ship between variables (Wang et al., 2011; Wu et al., 2014). Working on the basis of mutual information, which was first introduced by Linfoot (1957), Reshef et al. (2011) suggested a new statistical method, the MIC, which could identify arbitrary relationships between pairwise variables. They proposed a new concept where, if there is an association between two variables, a grid can be drawn on the scatter plot of the pairwise variables, while the MIC searches for the optimal grid resolution.

Several main properties of the MIC were also introduced, including equitability. Equitability means that the MIC will give similar scores to different functional relationships with similar noise

Figure 4. Comparison of the equitability

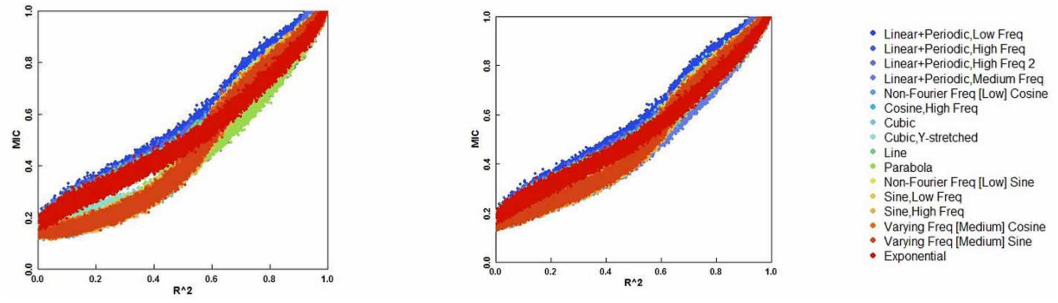
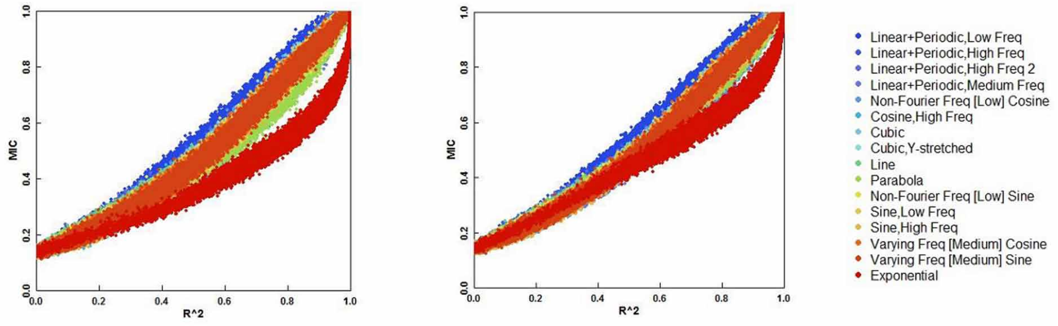


Figure 5. Comparison of the equitability



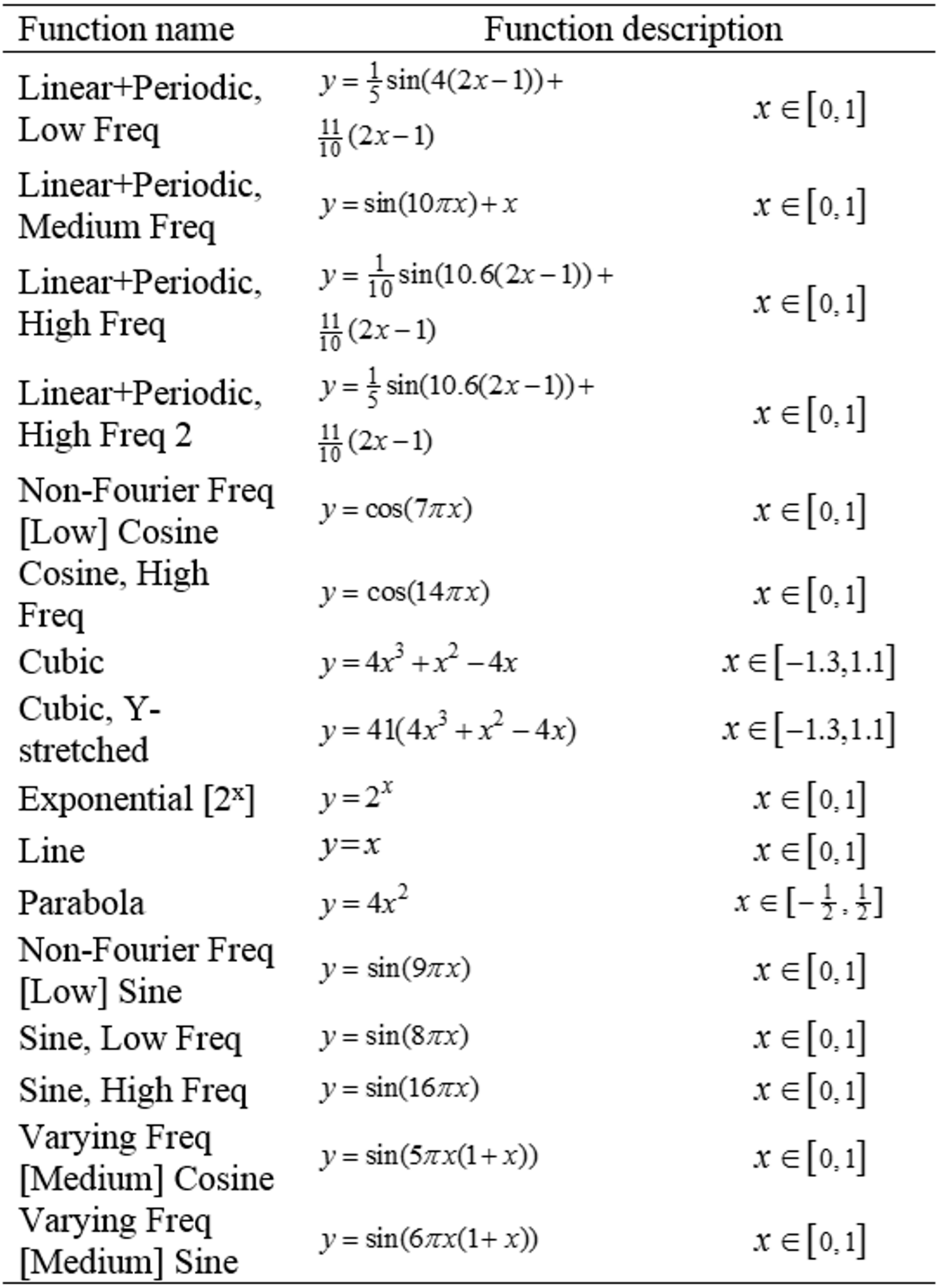
levels or similar R^2 (coefficient of determination) values. By comparison with other methods, including mutual information estimation, distance correlation, the Spearman correlation coefficient, principal curve-based methods, and maximal correlation, the MIC showed its equitability property through simulations. The MIC has now been applied in various fields, including clinical data, genomics and virology applications, as shown in the literature (Lin et al., 2012; Das et al., 2012; Anderson et al., 2012). Also, Karpinets et al. (2012) and Zhang et al. (2013) suggested that MIC would help to improve their applications.

The proposal of the MIC has had a considerable effect on many fields and has received high acclaim. Speed (2011) stated that the MIC, which can be used to determine nonlinear correlations in data sets equitably, had reached the crest of the domain called mutual information that had been developing for more than 50 years. Nature Biotechnology (2012) also showed great interest in MIC and asked eight experts to discuss its usefulness.

However, the MIC also stimulated questions. Simon and Tibshirani (2012) noted that the MIC would cause too many false positives in data analysis because of its low power. Gorfine et al. (2012) argued that the equitability of the MIC was less practical than a new test, HHG (Heller et al., 2012), for data exploration by simple power comparisons. Kinney and Atwal offered a mathematical proof to support mutual information rather than the MIC.

In a recent follow-up study, Reshef et al. (2013) proposed that the use of the approximation algorithm affects the equitability through the comparison to a lower efficiency algorithm that provides a more intensive search for optimal grids, and suggested that the essence of the MIC does not lead to the deviations from equitability of the recently reported MIC values. In particular, Reshef et al. had been expecting the approximation algorithms to have better accuracy-time tradeoffs.

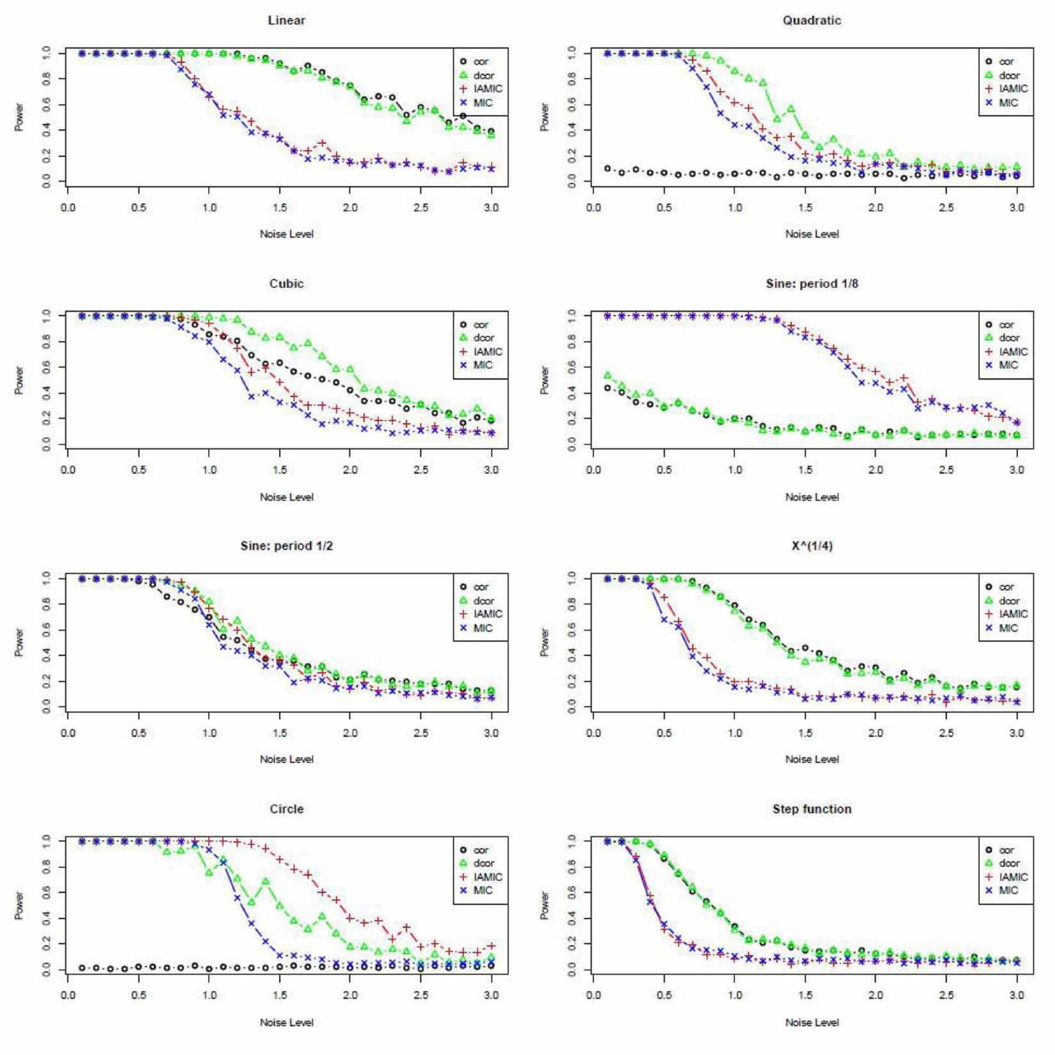
Table 4. Descriptions of the functions



Function name	Function description	
Linear+Periodic, Low Freq	$y=\frac{1}{5}\sin(4(2x-1))+\frac{11}{10}(2x-1)$	$x\in[0,1]$
Linear+Periodic, Medium Freq	$y=\sin(10\pi x)+x$	$x\in[0,1]$
Linear+Periodic, High Freq	$y=\frac{1}{10}\sin(10.6(2x-1))+\frac{11}{10}(2x-1)$	$x\in[0,1]$
Linear+Periodic, High Freq 2	$y=\frac{1}{5}\sin(10.6(2x-1))+\frac{11}{10}(2x-1)$	$x\in[0,1]$
Non-Fourier Freq [Low] Cosine	$y=\cos(7\pi x)$	$x\in[0,1]$
Cosine, High Freq	$y=\cos(14\pi x)$	$x\in[0,1]$
Cubic	$y=4x^3+x^2-4x$	$x\in[-1.3,1.1]$
Cubic, Y-stretched	$y=41(4x^3+x^2-4x)$	$x\in[-1.3,1.1]$
Exponential [2^x]	$y=2^x$	$x\in[0,1]$
Line	$y=x$	$x\in[0,1]$
Parabola	$y=4x^2$	$x\in[-\frac{1}{2},\frac{1}{2}]$
Non-Fourier Freq [Low] Sine	$y=\sin(9\pi x)$	$x\in[0,1]$
Sine, Low Freq	$y=\sin(8\pi x)$	$x\in[0,1]$
Sine, High Freq	$y=\sin(16\pi x)$	$x\in[0,1]$
Varying Freq [Medium] Cosine	$y=\sin(5\pi x(1+x))$	$x\in[0,1]$
Varying Freq [Medium] Sine	$y=\sin(6\pi x(1+x))$	$x\in[0,1]$

Using MIC with high accuracy for data minging can avoid losing some important pairs of variables that are closely related. For example, we apply MIC with two different algorithms to global indicators from the WHO datasets, which is provided in Reshef et al. (2011), and select 11 relationships (A-K) of D-value are more than 0.1.D-value means the difference of MIC valure calculated by two different

Figure 6. Comparison of power performance



accuracy algorithms. From Table 5, the relationship A ranks at 2763 when using standard algorithm, but increased to 71 after using an improved algorithm. The rank of other relationships also changed to a large extent.

Table 5 indicates that when MIC is used as the measure of dependence for general recommend system or data minging system, the approximation with low accuracy will result in the deficiency of some important variables in the process of feature engineering, and then reduce the accuracy of recommend system or data mining system.

In the research for MIC, an unavoidable problem is to find the global optimum MIC, reshef provide a violent search algorithm to get a better approximation. However, because of the extensive computation, the mesh partition is limited. On the other side, MIC is limited between two vectors. In fact, those relationships could generate between multi-vectors. That is to say, those vectors, which seem have no relationship, may do have strong relationship between multi-vectors. And those researchers found that the power value of MIC, which represents the effect of the algorithm

Table 5. The comparasion of rank scores

Relationship	D-value	$\text{Rank}_{\text{IMIC}}$	Rank_{MIC}
A	0.46306	71	2963
B	0.18028	89	459
C	0.14498	90	321
D	0.12688	47	155
E	0.12479	75	210
F	0.12204	54	165
G	0.11485	56	147
H	0.1135	44	126
I	0.10588	95	207
J	0.10308	18	86
K	0.10159	81	183

on statistics, is low. And if the low power value could not be explained reasonably, the reliability of MIC would be greatly reduced.

Hence, the research presented in this paper intends to provide that an improvement over the original approximation algorithm that was developed by Reshef et al (2011).

7. CONCLUSION

In this paper, the IAMIC iterative optimization algorithm has been proposed, which produces results that are close to the real MIC results by searching just n times, compared to the n2 computations required for the previous method. This algorithm, which is based on the axis equipartition method of Reshef et al. (2011), has presented a new concept for optimization of the partition on the y-axis to approximate the MIC with better accuracy. Mathematical analysis supports the fact that IAMIC can find the local maximum value to enable it to develop the important equitability property of the MIC. We also verified the performance improvement of IAMIC experimentally.

Finally, we recreated the experiments of Simon and Tibshirani (2012) using our algorithm. Our experimental results indicated that the power drawback is not an intrinsic defect of the MIC, and may draw researchers' attention back to this potential power issue.

The next step of our research will concentrate on calculation of the global maximum value of the MIC with acceptable time complexity. We believe that this process may reveal the essence of the MIC at a deeper level.

ACKNOWLEDGMENT

This work was supported in part by a grant from National Natural Science Fund of China (61472039) and National Key Research and Development Program (2016YFC0803000 and 2016YFB0502603).

REFERENCES

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2012). Cmine, minerva & minepy: a c engine for the mine suite and its r and python wrappers. arXiv:1208.4271

Das, J., Mohammed, J., & Yu, H. (2012). Genome-scale analysis of interaction dynamics reveals organization of biological networks. *Bioinformatics (Oxford, England)*, *28*(14), 1873–1878. doi:10.1093/bioinformatics/bts283 PMID:22576179

Dasu, T. (2003). *Exploratory Data Mining and Data Cleaning*. New York: John Wiley & Sons. doi:10.1002/0471448354

Ester, M., Frommelt, A., Kriegel, H.-P., & Sander, J. (2000). Spatial data mining: Databases primitives, algorithms and efficient DBMS support. *Data Mining and Knowledge Discovery*, *4*(2/3), 193–216. doi:10.1023/A:1009843930701

(2012). Finding correlations in big data. *Nature Biotechnology*, *30*(4), 334–335. doi:10.1038/nbt.2182 PMID:22491290

Frankel, F., & Reid, R. (2008). Distilling meaning from data. *Nature,* 455.

Gorfine, M., Heller, R., & Heller, Y. (2012). Comment on "Detecting novel associations in large datasets". Retrieved from http://emotion.technion.ac.il/~gorfinm/files/science6.pdf

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Verlag. doi:10.1007/978-0-387-84858-7

Heller, R., Heller, Y., & Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. arXiv:1201.3522

Hernàndez, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, *2*(1), 1–31. doi:10.1023/A:1009761603038

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., … Rhee, S.Y. (2008, September). The future of biocuration. *Nature,* 455, 47-50.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, *7*(1), 81–99. doi:10.1023/A:1021564703268

Kinney, J.B., & Atwal, G.S. (2013). Equitability, mutual information, and the Maximal information coefficient. arXiv:1301.7745

Lin, C., Canhao, H., Miller, T., Dligach, D., Plenge, R. M., Karlson, E. W., & Savova, G. (2012). Maximal information coefficient for feature selection for clinical document classification. *Proceedings of the ICMLWorkshop on Machine Learning for Clinical Data.*

Linfoot, E. H. (1957). An informational measure of correlation. *Information and Control*, *1*(1), 85–89. doi:10.1016/S0019-9958(57)90116-X

McKinsey Global Institute. (2011). Big Data: the Next Frontier for Innovation, Competition, and Productivity.

Meyer-Schoenberger, V., & Cukier, K. (2013). *Big data: a Revolution That will Transform How We Live, Work and Think*. London: John Murray.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge University Press. doi:10.1017/CBO9781139058452

Anderson, T.K., Laegreid, W.W., Cerutti, F., Osorio, F.A., Nelson, E.A., Christopher-Hennings, J., & Goldberg, T.L. (2012). Ranking viruses: Measures of positional importance within networks define core viruses for rational polyvalent vaccine development. *Bioinformatics (Oxford, England)*, *28*(12), 1624–1632. doi:10.1093/bioinformatics/bts181 PMID:22495748

ReshefD.ReshefY.MitzenmacherM.SabetiP. (2013). Equitability Analysis of the Maximal Information Coefficient, with Comparisons. arXiv:1301.6314

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P., & Sabeti, P. C. et al. (2011). Detecting novel associations in large data sets. *Science*, *334*(6062), 1518–1524. doi:10.1126/science.1205438 PMID:22174245

Shekar, S., & Xiong, H. (Eds.). (2007). *Encyclopedia of GIS*. New York: Springer.

Simon, N., & Tibshirani, R. (2012). Comment on “Detecting novel associations in large data sets”. Retrieved from http://www-stat.stanford.edu/_tibs/reshef/comment.pdf

Smets, P. (1996). Imperfect information: imprecision and uncertainty. In *Uncertainty Management in Information Systems* (pp. 225–254). London: Kluwer Academic Publishers.

Speed, T. (2011). A correlation for the 21st century. *Science*, *334*(6062), 1502–1503. doi:10.1126/science.1215894 PMID:22174235

Surhone, L. M., Tennoe, M. T., & Henssonow, S. F. (2010). *Big Data: BigTable, Cloud Computing, Database Theory*. Betascript Publishing.

Szekely, G., Rizzo, M., & Bakirov, N. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. doi:10.1214/009053607000000505

Szekely, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, *3*(4), 1236–1265. doi:10.1214/09-AOAS312 PMID:20574547

Tatiana, V. (2012). Karpinets, Byung H. Park and Edward C. Uberbacher. Analyzing large biological datasets with association networks. *Nucleic Acids Research*, *40*(17).

United Nations Global Pulse. (2012). Big Data for Development: Challenges & Opportunities.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012, November 6-9). Spatiotemporal data mining in the era of big spatial data: algorithms and applications. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, Redondo Beach, CA, USA. doi:10.1145/2447481.2447482

Wang, S. L., Gan, W. Y., Li, D. Y., & Li, D. R. (2011). Data Field for Hierarchical Clustering. *International Journal of Data Warehousing and Mining*, *7*(4), 43–63. doi:10.4018/jdwm.2011100103

Wang, S. L., & Yuan, H. N. (2014). Spatial data mining: A perspective of big data. *International Journal of Data Warehousing and Mining*, *10*(4), 50–70. doi:10.4018/ijdwm.2014100103

Wu, X., Zhu, X., Wu, G., & Ding, W. (2014). Data Mining with Big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. doi:10.1109/TKDE.2013.109

Zhang, X., Liu, K., Liu, Z.-P., Duval, B., Richer, J.-M., Zhao, X.-M., & Chen, L. et al. (2013). Jin-Kao Hao3, Luonan Chen. NARROMI: A noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics (Oxford, England)*, *29*(1), 106–113. doi:10.1093/bioinformatics/bts619 PMID:23080116

Shuliang Wang, PhD, is the associate head and professor in software engineering with the School of Software, the Beijing Institute of Technology in China. For his innovatory study of spatial data mining, he was awarded one of the best national thesis, the Fifth Annual InfoSciR-Journals Excellence in Research Awards, IGI Global and so on. His research interests include spatial data miningand software engineering.

Yiping Zhao, received her B.Sc. and M.Se degree from the Beijing Institute of Technology, Beijing, China. She is currently a regular employee at the Software Center, Bank of China. Her research interests include software engineering and big data analytics.

Yue Shu, received his B.Sc. degree from the Beijing Institute of Technology, Beijing, China, in 2012 and obtained his M.Se degree from the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, in 2015. He is currently a Master candidate at Tencent Technology (Beijing) Company Limited. His research interests include computer graphics, texture synthesis and big data analytics.

Wenzhong Shi, received the Ph.D. degree from the University of Osnabrück, Vechta, Germany, in 1994.He is a Head of Department and Chair in GIS and remote sensing with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University,Hung Hom, Kowloon, Hong Kong. His current research interests include GIS and remote sensing, uncertainty and spatial data quality control, and image processing for high-resolution satellite images.