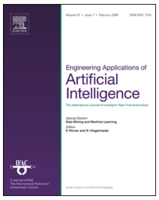
Engineering Applications of Artificial Intelligence 37 (2015) 250–257

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Joint mutual information-based input variable selection for multivariate time series modeling

Min Han*, Weijie Ren, Xiaoxin Liu

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, Liaoning 116023, China

ARTICLE INFO

Article history:
Received 12 May 2014
Received in revised form
10 July 2014
Accepted 19 August 2014
Available online 14 October 2014

Keywords:
Joint mutual information
Copula entropy
Input variable selection
Time series

ABSTRACT

For modeling of multivariate time series, input variable selection is a key problem. This paper presents the estimation of joint mutual information and its application in input variable selection problems. Mutual information is a commonly used measure for variable selection. To improve the performance of input variable selection, we propose a novel high-dimensional mutual information estimator based on copula entropy, which is estimated by the truncated k-nearest neighbor method. Simulations on high dimensional Gaussian distributions substantiate the effectiveness of the proposed mutual information estimator. A relationship between the joint mutual information and the copula entropy is derived, which is used for joint mutual information estimation. Then the proposed estimator is applied to input variable selection for multivariate time series modeling based on the criterion of max dependency and max–min dependency. A stop criterion is proposed to terminate the selection process automatically. Simulation results show that the input variable selection method works well on both synthetic and real life dataset.

1. Introduction

Multivariate time series contains two-dimensional or more variables, which are arranged according to a uniform time interval. There are a wide variety of multivariate time series data in the real world, such as in meteorology (Wu and Chau, 2013), hydrology (Grbić et al., 2013), economics (Keynia, 2012), biomedicine (Han and Liu, 2013) and many other fields. Compared to univariate time series, multivariate time series contains more abundant information of the complex dynamic system. It has been proved that the prediction model with multivariate time series can achieve higher accuracy than those with univariate time series (Du Preez and Witt, 2003). Therefore, modeling of multivariate time series receives more and more attention.

With the development of data acquisition and storage technology, there are a large number of high-dimensional data. As the dimensionality of input variable increases, irrelevant and redundant variables appear which would make it difficult to model multivariate time series. To avoid the curse of dimensionality, dimensionality reduction approaches are necessary (Fu, 2011). Feature extraction and variable selection are two types of commonly used methods. Feature extraction methods reduce dimensionality by mapping or transformation, such as singular value decomposition and principal component analysis (Han and Wang, 2009). However, the new variables obtained by feature extraction methods often lose physical properties of the original variables. In time series analysis, variable selection is more competitive than feature extraction. Variable selection methods (Guyon and Elisseeff, 2003) select the compact subset from the original dataset to improve the performance and interpretability of the prediction model. In this paper, we focus on variable selection methods based on mutual information for multivariate time series modeling.

Mutual information (MI) is one of the most important concepts in the field of information theory. As MI can measure both the linear and nonlinear dependency between variables, it has been applied widely in correlation measurement and variable selection (Wang et al., 2010; Lee and Kim, 2013). The basic idea of variable selection algorithm based on MI is to select the best subset S from the original dataset F by maximizing the joint MI between S and target output Y, namely $I(S;Y)$ (Vergara and Estévez, 2014). The main challenge that limits applications of the above method is to estimate MI between high-dimensional variables. To avoid estimating the joint MI, there are many MI-based variable selection algorithms that use low-dimensional approximation and the heuristic search method, such as mutual information feature selection (MIFS) (Battiti, 1994), mutual information feature selection under uniform information distribution (MIFS-U) (Kwak and Choi, 2002), minimal redundancy maximal relevance (mRMR) (Peng et al., 2005), normalized mutual information feature selection (NMIFS) (Estévez et al., 2009), etc. For most existing variable selection methods based on MI, the major shortcoming is that the candidate variable is selected one by one through

[*] Corresponding author. Tel.: +86 41184708719.
E-mail address: minhan@dlut.edu.cn (M. Han).

http://dx.doi.org/10.1016/j.engappai.2014.08.011
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

evaluating pairwise MI which easily leads to suboptimal results (Chow and Huang, 2005). Moreover, it has been shown that it is infeasible to approximate the high-dimensional MI with algebraic combinations of pairwise MI in any forms (Zheng and Kwoh, 2011). Therefore, we consider a direct estimation of joint MI to measure the dependency between candidate subsets and target output.

The accuracy of MI estimation is always limited by the estimation of the joint probability density function, thus influencing the identification of the dependency between variables. For now, much research about MI estimation has been done (Walters-Williams and Li, 2009). MI can be calculated by entropy, probability density or Kullback–Leibler divergence. They can also be classified as parametric methods and nonparametric methods. Parametric methods include maximum likelihood (ML) estimator, Bayesian estimator, and Edgeworth estimator (EDGE). Nonparametric methods include the histogram based method, kernel density estimator (KDE), the *k*-nearest neighbor (*k*-NN) method, the entropic spanning graph method, etc. Parametric methods assume that the data come from a type of probability distribution and make inferences about the parameters of the distribution. Unlike parametric methods, nonparametric methods make no assumptions about the probability distributions of the data, which is more flexible and convenient for applications (Ethem Alpaydin, 2004).

Maximum likelihood is a parametric technique (Suzuki et al., 2008). It is applicable only if the distribution of data is known. ML is prone to over fitting when the size of the dataset is not large enough compared to the degrees of freedom in the chosen model. This problem can be fixed by the Bayesian method, for the reason that the Bayesian method deals with how to determine the best number of model parameters (Endres and Foldiak, 2005). Therefore, the Bayesian method is very useful when large data sets are hard to obtain. When the underlying distribution of data set is close to normal distribution, EDGE is quite accurate and works well (Van Hulle, 2005). However, when the distribution is far from normal, the approximation error gets large and EDGE becomes unreliable.

The histogram based method (Hacine-Gharbi et al., 2012) and kernel density estimator are the two principal differentiable estimators of MI. There are mainly two types of histogram based estimators, namely equidistant and equiprobable. The equiprobable histogram based estimator is more accurate than the equidistant one. KDEs are more accurate than histogram based methods, but they are more time-consuming. For example, the Parzen window method (Kwak and Choi, 2002) has a quadratic complexity with respect to the number of dimensionality. Compared with histogram based methods and kernel density estimators, *k*-NN is a better choice as fine partitions capture the fine structure of chaotic data and it is not significantly corrupted with noise. But the estimation accuracy depends on the value of k and there seems no systematic strategy to choose the value of k appropriately (Kraskov et al., 2004). Entropic spanning graph is a “non plug-in” method as it estimates entropy directly from the sample set. The entropy estimator based on entropic graph has a linear complexity with variable dimensionality and has $O(N \log N)$ complexity for constructing an entropic spanning graph over N training samples (Balagani and Phoha, 2010). So it is not bounded by the curse of dimensionality. However, it cannot estimate Shannon entropy directly. Different parameters α must be used so that the Shannon entropy can be extrapolated with the α-entropy.

Above all, every MI estimator has its advantages and scope of applications. In this paper, we propose a new MI estimator based on copula entropy to avoid the estimation of both the marginal and joint probability density functions. And truncating *k*-NN is used to estimate the copula entropy on the basis of a group of pseudo-observations calculated from the given samples. Then, the proposed MI estimator is applied to input variable selection based on MD and MmD criterion. The rest of the paper is organized as follows. In Section 2, the background of MI will be introduced and several kinds of *k*-nearest neighbor estimators will be discussed and compared in detail. In Section 3, we will give a detailed presentation for the proposed MI estimator. And the experimental results are analyzed in Section 4. Finally, the conclusions are given in Section 5.

2. Background on mutual information

In this section, we briefly review the definition of MI and its estimation based on *k*-nearest neighbor method.

2.1. Definition of mutual information

The MI is a commonly used concept in the field of information theory. To understand the meaning of MI, entropy is an essential prior knowledge. Shannon's entropy (Shannon, 2001), first introduced in 1948, is a measure of uncertainty of random variables. If X is a continuous random variable with probability density function $p(x)$, the entropy of X is defined as

$$H(X) = -\int p(x)\log\, p(x)dx \tag{1}$$

The joint entropy is used to examine the amount of information among multiple variables. The joint entropy of two continuous random variables X and Y is as follows:

$$H(X,Y) = -\iint p(x,y)\log\, p(x,y)dx\, dy \tag{2}$$

where $p(x,y)$ is the joint probability density function of X and Y.

For two continuous random variables X and Y, the MI is defined as

$$I(X;Y) = \iint p(x,y)\log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy \tag{3}$$

where $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y respectively. The MI describes the shared information of X and Y, and can be used to measure the dependency between two random variables without any prior knowledge. Generally, the stronger correlation between two random variables is, the larger MI they will have. A relationship between the MI and the entropy can be drawn from the above definitions

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

Extend the MI to more than two continuous random variables $\{X_1, X_2, \ldots, X_m\}$, and then we can obtain high-dimensional MI with m variables,

$$I(X_1, X_2, \ldots, X_m) = \iint\cdots\int p(x_1, x_2, \ldots, x_m)\log \frac{p(x_1, x_2, \ldots, x_m)}{p(x_1)p(x_2)\ldots p(x_m)}\, dx_1\, dx_2 \cdots dx_m \tag{5}$$

where $p(x_1, x_2, \ldots, x_m)$ is joint probability density function and $p(x_1), p(x_2), \ldots, p(x_m)$ are marginal probability density functions.

The joint MI measures the dependency between multiple variables $\{X_1, X_2, \ldots, X_m\}$ and Y. The joint MI is defined as follows:

$$I(X_1, X_2, \ldots, X_m; Y) = \iint\cdots\int p(x_1, x_2, \ldots, x_m, y)\log \frac{p(x_1, x_2, \ldots, x_m, y)}{p(x_1, x_2, \ldots, x_m)p(y)}\, dx_1\, dx_2 \cdots dx_m\, dy \tag{6}$$

Unlike the MI between two random variables, the joint MI not only concerns the dependency between $\{X_1, X_2, \ldots, X_m\}$ and Y, but also involves the internal correlation of $\{X_1, X_2, \ldots, X_m\}$. Therefore, the joint MI is highly suited to solve the input variable selection problems.

2.2. K-nearest neighbor estimation of mutual information

The *k*-NN method has been widely used in the field of pattern recognition. As for the estimation of MI, there are three kinds of *k*-NN methods at present according to the different ways they are used. They are k-NN for the estimation of probability density function, k-NN for the estimation of Shannon entropy, and k-NN for the estimation of MI directly. The specific calculation processes of three methods are shown below.

Recent research (Blumentritt and Schmid, 2012) showed that MI depends on the copula density as follows:

$$I(\boldsymbol{X}) = \int_{[0,1]^m} c(\boldsymbol{u})\log\ [c(\boldsymbol{u})]d\boldsymbol{u} \tag{7}$$

where $\boldsymbol{X} = [X_1, X_2, ..., X_m]$ is a group of random variables, and $c(\boldsymbol{u})$ is the corresponding copula density. Therefore, the estimation of MI can be converted to the problem of copula density estimation. A group of pseudo-observations $\boldsymbol{U}$ are first generated by estimating the marginal distribution functions of $\boldsymbol{X}$. Let d_k denote the distance from the sample point to its kth nearest neighbor, and then the copula density is estimated by k-NN as

$$\hat{c}(\boldsymbol{u}) = \frac{k/N}{(2d_k)^m} \tag{8}$$

where N is the number of samples. So the MI can be estimated as

$$I_{trnc}(\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N} \log(\hat{c}(\boldsymbol{u}_i)) \tag{9}$$

Bonev et al. applied the Kozachenko–Leonenko's entropy estimator based on k-NN for MI estimation in Bonev et al. (2013). Let $d(i)$ denote the distances between the samples $\boldsymbol{x}_i$ and their respective k-NN, the Shannon entropy can be estimated by

$$\hat{H}(\boldsymbol{X}) = -\psi(k) + \psi(N) + \log \frac{V_m}{2^m} + \frac{m}{N}\sum_{i=1}^{N} \log 2d(i) \tag{10}$$

where k is the number of neighbors, N is the number of samples, m is the number of variable dimensionality, V_m is the volume of the unit ball and $\psi(\bullet)$ is the digamma function which can be calculated by the following recursions:

$$\psi(1) = -0.5772516$$
$$\psi(k+1) = \psi(k) + 1/k \tag{11}$$

Thus the MI can be estimated by entropy with (4) or

$$I_{leknn}(X;Y) = H(X) - H(X|Y) \tag{12}$$

This method enables us to estimate the high-dimensional MI for the computational time increases linearly with the number of dimensionality. But the errors made in the individual estimated entropy would presumably not cancel in the MI.

Two straightforward MI estimation algorithms based on k-NN were proposed in Kraskov et al. (2004) by very similar arguments with the Kozachenko–Leonenko's Shannon entropy estimator. Given variables X, Y and $\boldsymbol{Z} = [X, Y]$, $d(i)$ represents the distance from $\boldsymbol{z}_i$ to its kth neighbor, and $d_x(i)$ and $d_y(i)$ represent the distances in the subspaces of X and Y between the same points. Then the number of points whose distance from x_i is strictly less than $d(i)$ is counted as $n_x(i)$. With the same way, we can get $n_y(i)$. Therefore, MI can be estimated as follows:

$$I_{knn}(X;Y) = \psi(k) - \langle \psi(n_x+1) + \psi(n_y+1) \rangle + \psi(N) \tag{13}$$

where $\langle ... \rangle$ represents the average both over all $i \in [1, \ldots, N]$ and all realizations of the random samples.

$$\langle ... \rangle = \frac{1}{N}\sum_{i=1}^{N} E[\ldots(i)] \tag{14}$$

Another algorithm used the following equation:

$$I_{knn}(X,Y) = \psi(k) - 1/k - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \tag{15}$$

To estimate the joint MI between $\{X_1, X_2, \ldots, X_m\}$ and Y, the high-dimensional variables $\{X_1, X_2, \ldots, X_m\}$ should be treated as a whole and n_x would be defined as the number of points in the m-dimensional space.

3. Mutual information estimation based on copula entropy

In time series analysis, the input variable often has high dimensionality. Therefore, estimation of high-dimensional MI is of great value. In this section, estimation of high-dimensional MI and joint MI are presented. We first derive the relationship between MI and copula entropy. Then, estimator of high-dimensional MI and joint MI based on copula entropy are proposed. The truncating k-nearest neighbor method is adopted for copula entropy estimation. Finally, the basic steps are given for the estimation of high-dimensional MI and joint MI.

3.1. Joint mutual information estimation based on copula entropy

According to the Sklar theorem (Sklar, 1959), the joint probability density function of two random variables X and Y is

$$p(x,y) = c(u_x, u_y)p(x)p(y) \tag{16}$$

where $u_x = \int_{-\infty}^{x} p(x)dx$, $u_y = \int_{-\infty}^{y} p(y)dy$, and $c(u_x, u_y)$ is the copula density function. We can infer a new formula for MI based on (16)

$$\begin{aligned} I(X;Y) &= \iint p(x,y)\log \frac{p(x,y)}{p(x)p(y)}dx\,dy \\ &= \iint c(u_x,u_y)p(x)p(y)\log \frac{c(u_x,u_y)p(x)p(y)}{p(x)p(y)}dx\,dy \\ &= \iint c(u_x,u_y)p(x)p(y)\log\ c(u_x,u_y)dx\,dy \\ &= \iint c(u_x,u_y)\log\ c(u_x,u_y)du_x\,du_y \end{aligned} \tag{17}$$

According to (17), we can use the copula density function to estimate MI instead of estimating the joint probability density function (Zeng and Durrani, 2011). Extending it to the condition of high-dimensional MI with m variables, we can get

$$\begin{aligned} I(X_1, X_2, \ldots, X_m) = &\iint\cdots\int c(u_1, u_2, \ldots, u_m)p(x_1)p(x_2)\ldots \\ &\times p(x_m)\log\ c(u_1, u_2, \ldots, u_m)dx_1\,dx_2\cdots dx_m \\ &= \iint\cdots\int c(u_1, u_2, \ldots, u_m)\log\ c(u_1, u_2, \ldots, u_m)du_1\,du_2\cdots du_m \end{aligned} \tag{18}$$

Copula entropy (Davy and Doucet, 2003) is defined as

$$\begin{aligned} H_C(U_1, U_2, \ldots, U_m) = -&\iint\cdots\int c(u_1, u_2, \ldots, u_m)\log\ c(u_1, u_2, \ldots, u_m) \\ &\times du_1\,du_2\ldots du_m \end{aligned} \tag{19}$$

According to (18) and (19), we can derive a relationship between high-dimensional MI and copula entropy

$$I(X_1, X_2, \ldots, X_m) = -H_C(U_1, U_2, \ldots, U_m) \tag{20}$$

Assuming that $S_{m-1} = \{X_1, X_2, \ldots, X_{m-1}\}$ and $S_m = \{S_{m-1}, X_m\}$, the joint MI of S_m and Y can be calculated based on the relationship between MI and entropy

$$I(S_m;Y) = H(Y) + H(S_m) - H(S_m, Y) \tag{21}$$

Based on the definition of high-dimensional MI shown in (5), we can easily obtain

$$H(S_m) = \sum_{i=1}^{m} H(X_i) - I(S_m) \tag{22}$$

$$H(S_m, Y) = H(Y) + \sum_{i=1}^{m} H(X_i) - I(S_m, Y) \tag{23}$$

From (21) to (23), the joint MI can be calculated by

$$I(S_m;Y) = I(S_m, Y) - I(S_m) \tag{24}$$

Therefore, the relationship between joint MI and copula entropy is as follows:

$$I(S_m;Y) = H_C(U_1, U_2, \ldots, U_m) - H_C(U_1, U_2, \ldots, U_m, U_y) \tag{25}$$

From the above analysis, the problem of joint MI estimation is converted to the problem of copula entropy estimation, which can avoid the estimation of joint probability density function.

3.2. Copula entropy estimation

To estimate the copula entropy, we should first have a group of observations of $\boldsymbol{U}$. Give N random samples drawn from m variables $\boldsymbol{X}=[X_1,X_2,...,X_m]$, and the pseudo-observations $\hat{\boldsymbol{U}}_i=[\hat{U}_{i1},\hat{U}_{i2},...,\hat{U}_{im}]$, $i=1,2,\ldots,N$ of $\boldsymbol{X}$ can be generated as follows (Blumentritt and Schmid, 2012):

$$\hat{U}_{ij}=\frac{1}{N+1}r_{ij} \tag{26}$$

where r_{ij} is the rank of X_{ij} in $X_{1j},X_{2j},...,X_{Nj}$ $(j=1,2,\ldots,m)$. As a result, the pseudo-observations are strictly restricted to $[0,1]^m$. Based on these pseudo-observations we can estimate the copula entropy with the k-nearest neighbor estimator (Kraskov et al., 2004).

$$\hat{H}_{cknn}(\boldsymbol{U})=-\psi(k)+\psi(N)+\log\, c_m+\frac{m}{N}\sum_{i=1}^{N}\log\, 2d(i) \tag{27}$$

where $d(i)$ is the distance from the ith sample $\boldsymbol{u}_i$ to its k-nearest neighbor. And the maximum norm is used herein, as it is noted to be useful in high dimensionality (Blumentritt and Schmid, 2012). c_m equals to 1 when the maximum norm is used. $\psi(x)$ is the digamma function which can be defined in terms of the $\Gamma(\bullet)$ function

$$\psi(x)=\Gamma(x)^{-1}\, d\Gamma(x)/dx \tag{28}$$

It also satisfies the recursion as follows:

$$\psi(1)=-\gamma,\gamma\simeq 0.5772516 \tag{29}$$

$$\psi(x+1)=\psi(x)+1/x \tag{30}$$

where γ is the Euler–Mascheroni constant. $\Gamma(\bullet)$ is the Gamma function defined as

$$\Gamma(z)=\int_0^{\infty} t^{z-1}e^{-t}\, dt \tag{31}$$

for those z with positive real part. And it is a generalization of the factorial function $n!=\Gamma(n+1)$, $n\in\mathbb{N}$ to the complex domain.

As we know, the k-nearest neighbor estimator is not restricted to $[0,1]^m$. So it is possible that mass is placed in the exterior of the unit cube when the center of the grid locates near the boundaries. To solve this problem, we propose to use the truncating k-nearest neighbor so that all grids which are not a subset of the unit cube are truncated. We denote the modified entropy estimator as $\hat{H}_{ctknn}$ which is shown as follows:

$$\hat{H}_{ctknn}(\boldsymbol{U})=-\psi(k)+\psi(N)+\log\, c_m+\frac{1}{N}\sum_{i=1}^{N}\log\, \varepsilon_i \tag{32}$$

$$\begin{aligned}\varepsilon_i=\prod_{j=1}^{m}[&d_j(i)+d_j(i)I_{(u_{ij}-d_{k,ij})\geq 0}(u_{ij})I_{(u_{ij}+d_{k,ij})\leq 1}(u_{ij})+u_{ij}I_{(u_{ij}-d_{k,ij})<0}(u_{ij})\\ &+(1-u_{ij})I_{(u_{ij}+d_{k,ij})>1}(u_{ij})]\end{aligned} \tag{33}$$

where $d_j(i)$ is the distance from the ith sample $\boldsymbol{u}_i$ to its k-nearest neighbor in the jth dimension and

$$I_{s(x)}(x)=\begin{cases}1 & \text{if } s(x) \text{ is true}\\ 0 & \text{if } s(x) \text{ is false}\end{cases} \tag{34}$$

Above all, the algorithm of high-dimensional MI and joint MI estimation based on copula entropy can be described as the following steps:

1) Generate the pseudo-observations $\hat{\boldsymbol{U}}_i$ according to (26) based on the N given random samples $\boldsymbol{X}=[X_1,X_2,...,X_m]$;
2) Estimate the copula entropy $\hat{H}_{ctknn}$ of $\hat{\boldsymbol{U}}_i$ based on (32) using the truncating k-nearest neighbor method;
3) Substitute $\hat{H}_{ctknn}$ into (20), we can get the high-dimensional MI as

$$I_{ctknn}(\boldsymbol{X})=-\hat{H}_{ctknn}(\boldsymbol{U}) \tag{35}$$

And substitute $\hat{H}_{ctknn}$ into (25), we can get the estimation of the joint MI.

4. Experimental results and analysis

In this section, the experiments are divided into two parts. First, we perform experiments on high dimensional Gaussian distributions to illustrate the effect of the proposed high-dimensional MI estimator. Then the proposed MI estimator is applied to input variable selection based on the criterion of max dependency (MD) and max–min dependency (MmD) (Bonev et al., 2008). To determine the optimal size of the variable subset, a stop criterion is introduced for both MD and MmD.

4.1. High dimensional Gaussian distributions

In all cases we shall deal with Gaussian distributions with zero mean and unit variance. The analytical expression of MI for such m-dimensional Gaussian distributions with covariance matrix σ is given as follows (Kraskov et al., 2004):

$$I(X_1,X_2,\ldots,X_m)=-\frac{1}{2}\log\left[\det(\sigma)\right] \tag{36}$$

We generated all the data points from the one-parameter family of m-dimensional Gaussian distributions, which means that all the correlation coefficients equal to ρ, namely $\sigma_{ij}=\rho$ $(i\neq j)$. In this simulation, the sample size N is set to 500, 1000 and 2500, the correlation coefficient ρ is set to 0, 0.25, 0.5 and 0.75, and the dimensionality of input variable is arranged from 2 to 5. In order to validate the effectiveness of the high-dimensional MI estimator, the proposed I_{ctknn} is compared to I_{trnc} (Blumentritt and Schmid, 2012), I_{leknn} (Bonev et al., 2013) and I_{knn} (Kraskov et al., 2004), which are calculated based on (9), (12) and (13) respectively. In addition, we also take I_{cknn} which is calculated based on (27) into consideration for comparison to illustrate the effect of the truncated strategy.

Figs. 1–4 show the averaged absolute errors within 20 runs when N equals to 500 for different dimensions. From the results showed above, the MI estimator I_{ctknn} based on copula entropy with truncated strategy performs much better than the one I_{cknn}

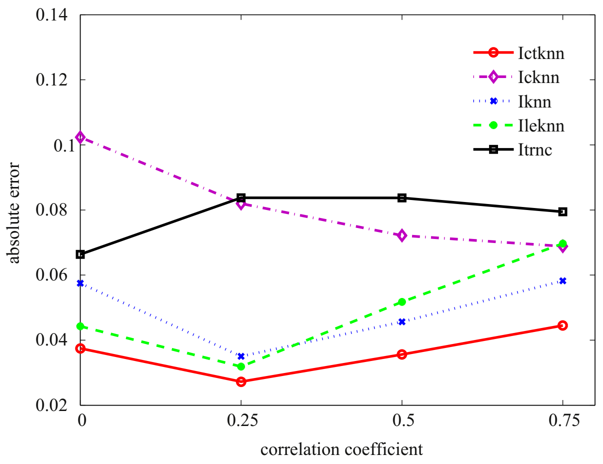


Fig. 1. Absolute errors for five different MI estimators with m=2 and N=500.

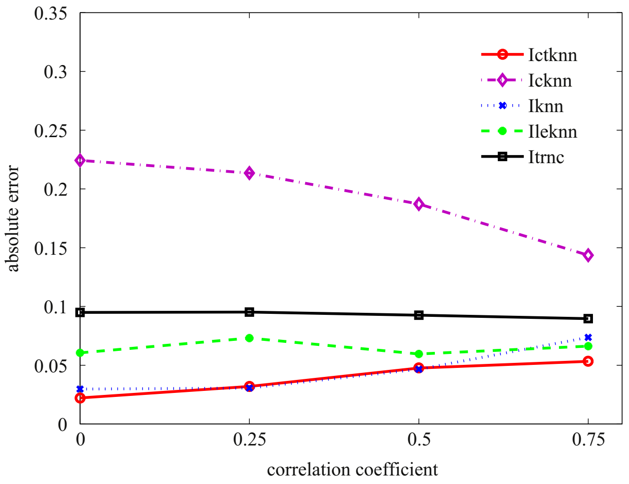


Fig. 2. Absolute errors for five different MI estimators with $m=3$ and $N=500$.

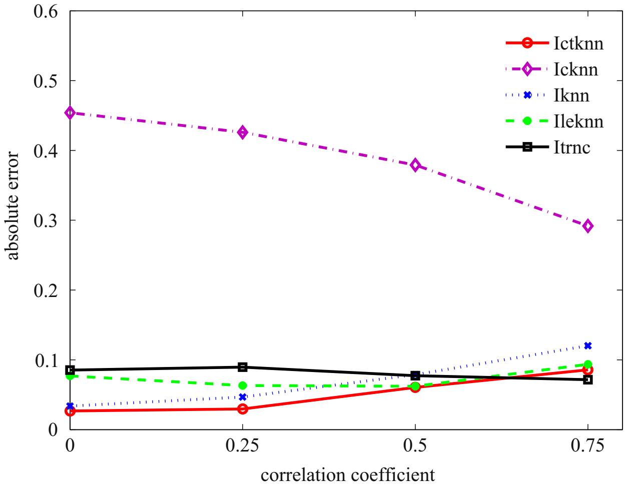


Fig. 3. Absolute errors for five different MI estimators with $m=4$ and $N=500$.

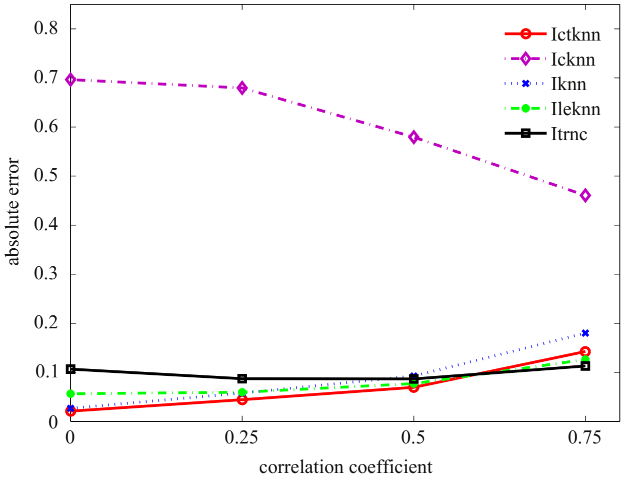


Fig. 4. Absolute errors for five different MI estimators with $m=5$ and $N=500$.

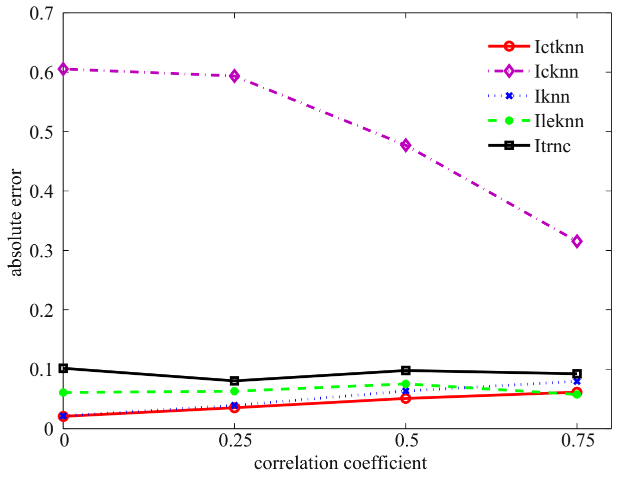


Fig. 5. Absolute errors for five different MI estimators with $m=5$ and $N=1000$.

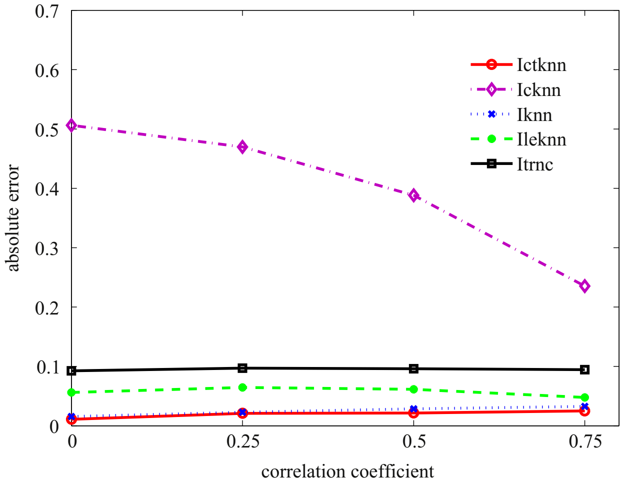


Fig. 6. Absolute errors for five different MI estimators with $m=5$ and $N=2500$.

without truncated strategy, especially in high-dimensional situation. And in general, the proposed MI estimator achieves the smallest errors for different dimensions together with the estimator I_{knn}. The simulation results for $m=5$ are presented in Fig. 5 and Fig. 6, when N equals to 1000 and 2500 respectively. All the estimators except for I_{cknn} have a stable performance with the correlation coefficients as the number of samples grows large. Therefore, the proposed estimator can be successfully applied to the estimation of high-dimensional MI, as well as joint MI.

4.2. Applications for input variable selection

The experiments are carried out on two datasets: the synthetic dataset of Friedman (1991) and the real life dataset of meteorological series of Dalian in China.

4.2.1. Overview of mutual information-based criterion

The MI-based variable selection criterions we used in this section are max dependency (MD) and max–min dependency (MmD) (Bonev et al., 2008), as the joint MI can be estimated by I_{ctknn}. Given an original set F with M variables, the goal of variable selection is to find the subset S_m that can satisfy the MD or the MmD criterion:

$$MD : \max_{S_m \subseteq F} I(S_m;Y) \tag{37}$$

$$MmD : \max_{S_m \subseteq F} [I(S_m;Y) - I(F - S_m;Y)] \tag{38}$$

Since there are $2^M - 1$ candidate subsets of F, exhaustive search is not applicable in practical applications. Hence, the forward greedy search strategy is used instead of exhaustive search. The forward greedy search strategy selects the best variable one by one, of which the computation complexity is $O(M^2)$. A rank of the variables is given with the above criterion, which also takes a large amount of time cost when M is large. Indeed, we do not need the rank of all the variables. To make the variable selection procedure more efficient, we propose a simple stop criterion for both MD and MmD.

$$I(S_{m-1}, X_m; Y) \le I(S_{m-1}; Y) \tag{39}$$

When (39) is satisfied, the variable selection procedure is terminated automatically.

Based on forward greedy search and joint MI-based criterion, calculation process of input variable selection algorithm is as follows:

(1) Initialization. Set the original set $F = \{F_1, F_2, \ldots, F_M\} \in \mathbb{R}^M$ and the selected subset $S_m = \{X_1, X_2, \ldots, X_m\} \in \mathbb{R}^m$, where $S_m \subseteq F$ and $m \in [1, M]$ is the number of selected variables. The target output is Y.

(2) Selection of the first variable. Calculate the MI between each candidate variable and Y. Then select the first variable with the largest MI and add it to the selected subset S.

(3) Selection of the mth variable. Based on forward greedy search strategy, select the mth variable according to the following criterion:

$$MD: \hat{X}_m = \underset{X_m \in F - S_{m-1}}{\arg\max}\ I(S_{m-1}, X_m; Y) \tag{40}$$

$$MmD: \hat{X}_m = \underset{X_m \in F - S_{m-1}}{\arg\max}\ [I(S_{m-1}, X_m; Y) - I(F - S_{m-1} - X_m; Y)] \tag{41}$$

The selected subset is updated as $S_m = S_{m-1} \cup \{\hat{X}_m\}$ and the new joint MI between S_m and Y is calculated, $I(S_m; Y) = I(S_{m-1}, \hat{X}_m; Y)$.

(4) Repeat step (3), until the stop criterion (39) is satisfied. Output the subset S_{m-1} as the optimal subset.

4.2.2. Friedman dataset

The Friedman model is shown in (42),

$$Y = 10\ \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon \tag{42}$$

where $X_1, X_2, \ldots, X_5$ are relevant input variables ranging in $[0, 1]$ and ε is the independent Gaussian white noise with zero mean and unit variance. Another five irrelevant variables $X_6, X_7, \ldots, X_{10}$ are generated randomly from $[0, 1]$, and two redundant variables X_{11} and X_{12} are generated as

$$X_{11} = 0.5X_1 + 0.5\varepsilon$$
$$X_{12} = 0.5X_2 + 0.5\varepsilon \tag{43}$$

This dataset can be used to examine whether the variable selection algorithm can select relevant variables or not in the condition of both irrelevant and redundant variables. The performance of MD and MmD are compared to two popular MI-based criterion, mRMR (Peng et al., 2005) and NMIFS (Estévez et al., 2009). The simulation results are presented in Table 1, from which we can see that MD and MmD outperform mRMR and NMIFS. Both of mRMR and NMIFS select four relevant variables and one irrelevant variable, while MD and MmD can successfully avoid the influence of irrelevant and redundant variables. Moreover, the stop criterion is shown to work well, as it terminates the selection process immediately after five relevant variables being selected. Therefore, joint MI-based input variable selection methods are more efficient than those low-dimensional approximation ones.

Table 1
Variable selection results for Friedman dataset.

Method	Selected variables
mRMR	X_4, X_2, X_1, X_5, X_6
NMIFS	X_4, X_2, X_1, X_5, X_6
MD	X_4, X_2, X_1, X_5, X_3
MmD	X_4, X_2, X_1, X_5, X_3

4.2.3. Dalian meteorological time series

The dataset of Dalian meteorological series can be downloaded from the database of the China meteorological data sharing service system. It contains six series in all, including wind speed X_1, percentage of sunshine X_2, atmospheric pressure X_3, temperature X_4, relative humidity X_5 and rainfall X_6. All the data are monthly average value ranging from January 1951 to June 2010, resulting in 715 months totally. The original dataset contains a lot of noise, which is difficult to build the prediction model. For this reason, the singular spectrum analysis method (Wu and Chau, 2013) is adopted for de-noising original data to get the main evolution characteristics of time series. Then, we set the embedding dimension as 12 and the delay time as 1 based on phase space reconstruction theory (Takens, 1981), as these series have an apparent period of 12. We get a group of 72 input variables as follows:

$$\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5, \boldsymbol{X}_6\} \tag{44}$$

$$\boldsymbol{X}_i = \{X_i(t-12), X_i(t-11), \ldots, X_i(t-1)\}, \quad i = 1, 2, \ldots, 6 \tag{45}$$

In this section, we do one-step prediction for both rainfall and temperature. The output variables are

$$\boldsymbol{Y} = \{X_4(t), X_6(t)\} \tag{46}$$

To evaluate the prediction performance, we use root mean square error (RMSE) and normalized mean square error (NMSE) for results comparison, which are defined as follows:

$$E_{RMSE} = \left(\frac{1}{N-1}\sum_{i=1}^{N}[\hat{y}_i - y_i]^2\right)^{1/2} \tag{47}$$

$$E_{NMSE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_m)^2}. \tag{48}$$

where N is the number of samples, y_i is the real output value, $\hat{y}_i$ is the predicted output value, and $\hat{y}_m$ is its mean. The more close to zero they are, the better are the prediction results.

The extreme learning machine (ELM) (Chen et al., 2014) is employed as the prediction model with 100 hidden nodes for all experiments. The dataset is divided into two parts: 75% samples for training and 25% samples for testing. Table 2 shows the averaged RMSE and NMSE of 20 runs with ELM for rainfall prediction. And the prediction performance of MD and MmD are compared with that of all the 72 variables and 24 variables of $\boldsymbol{X}_4$ and $\boldsymbol{X}_6$. It is shown that MD and MmD select the same four input variables for rainfall one-step prediction, and their prediction errors are much smaller than that of 72 variables and 24 variables. Table 3 presents the averaged RMSE and NMSE of 20 runs with ELM for temperature one-step prediction. MD and MmD select the same three input variables. The prediction errors show that the models based on MD, MmD and 24 variables perform much better than the one based on 72 variables. Though the model of 24 variables outperforms that of MD and MmD, the model performance with just three inputs is acceptable in applications.

Above all, the variable selection criteria MD and MmD based on the proposed MI estimator work well for dimensionality reduction and the variables they select perform well for prediction models.

Table 2
Comparison on the one-step prediction performance for rainfall.

Method	Input variables	E_{RMSE}	E_{NMSE}
Six series	$\boldsymbol{X}$	29.2165 ± 0.3976	5.4426 ± 0.1209
Two series	$\boldsymbol{X}_4$, $\boldsymbol{X}_6$	19.6266 ± 0.1925	3.6594 ± 0.0774
MD	$X_4(t-6)$, $X_6(t-1)$, $X_6(t-2)$, $X_4(t-12)$	7.9728 ± 0.0314	0.9137 ± 0.0076
MmD	$X_4(t-6)$, $X_6(t-1)$, $X_6(t-2)$, $X_4(t-12)$	7.9728 ± 0.0314	0.9137 ± 0.0076

Table 3
Comparison on the one-step prediction performance for temperature.

Method	Input variables	E_{RMSE}	E_{NMSE}
Six series	$\boldsymbol{X}$	1.3339 ± 0.0190	0.2510 ± 0.0076
Two series	$\boldsymbol{X}_4$, $\boldsymbol{X}_6$	0.3596 ± 0.0013	$0.0275 \pm 2.0241\mathrm{e}-4$
MD	$X_4(t-1)$, $X_4(t-3)$, $X_4(t-12)$	0.5541 ± 0.0032	$0.0346 \pm 3.9147\mathrm{e}-4$
MmD	$X_4(t-1)$, $X_4(t-3)$, $X_4(t-12)$	0.5541 ± 0.0032	$0.0346 \pm 3.9147\mathrm{e}-4$

5. Conclusions

A novel high-dimensional MI estimator is proposed in this paper and applied for variable selection. On the basis of the relationship between MI and copula entropy, the MI estimation is transformed into the problem of copula entropy estimation. To estimate the copula entropy, a group of pseudo-observations are generated by estimating the marginal probability functions of the given samples at first. Then the truncated k-NN method is used for entropy estimation. Simulations on Gaussian distributions substantiated the effectiveness of the proposed estimator and the truncated strategy. Additionally, a relationship between the joint MI and copula entropy is derived in this paper. Thus the joint MI can be estimated with the proposed estimator, and variable selection can be realized as a result. We use the MD and MmD criteria and forward greedy search strategy to select variables for datasets of Friedman and Dalian meteorological series. A new stop criterion is introduced to make the searching process terminate and to alleviate time cost. Simulation results show that the proposed MI estimator works well for variable selection and the stop criterion can make the variable selection process efficiently.

Although the motivation in this paper is to solve input variable selection for multivariate time series modeling, the proposed high-dimensional MI and joint MI estimators can be used to various practical issues. As the proposed MI estimators are based on nonparametric methods, there is no limit in assumption of data distribution. Whether the proposed method can achieve good results in other complex problems, such as gene expression and EEG classification, has to be studied further. Besides, for input variable selection algorithm, subset generation is a key problem. Exhaustive search with high computational complexity is not applicable in high-dimensional applications. Forward greedy search strategy used in this paper is more efficient, but it usually leads to suboptimal solutions. Therefore, the focus of future research is to improve efficiency and nature of the solution of input variable selection algorithm.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61374154) and the National Basic Research Program of China (973 Program) (2013CB430403).

References

Balagani, K.S., Phoha, V.V., 2010. On the feature selection criterion based on an approximation of multidimensional mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 32 (7), 1342–1343.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. 5 (4), 537–550.

Blumentritt, T., Schmid, F., 2012. Mutual information as a measure of multivariate association: analytical properties and statistical estimation. J. Stat. Comput. Simul. 82 (9), 1257–1274.

Bonev, B., Escolano, F., Cazorla, M., 2008. Feature selection, mutual information, and the classification of high-dimensional patterns. Pattern Anal. Appl. 11 (3–4), 309–319.

Bonev, B., Escolano, F., Giorgi, D., Biasotti, S., 2013. Information-theoretic selection of high-dimensional spectral features for structural recognition. Comput. Vis. Image Underst. 117 (3), 214–228.

Chen, H., Peng, J., Zhou, Y., Li, L., Pan, Z., 2014. Extreme learning machine for ranking: generalization analysis and applications. Neural Netw. 53, 119–126.

Chow, T.W., Huang, D., 2005. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. IEEE Trans. Neural Netw. 16 (1), 213–224.

Davy, M., Doucet, A., 2003. Copulas: a new insight into positive time–frequency distributions. IEEE Signal Process. Lett. 10 (7), 215–218.

Du Preez, J., Witt, S.F., 2003. Univariate versus multivariate time series forecasting: an application to international tourism demand. Int. J. Forecast. 19 (3), 435–451.

Endres, D., Foldiak, P., 2005. Bayesian bin distribution inference and mutual information. IEEE Trans. Inf. Theory 51 (11), 3766–3779.

Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M., 2009. Normalized mutual information feature selection. IEEE Trans. Neural Netw. 20 (2), 189–201.

Ethem, Alpaydin, 2004. Introduction to Machine Learning. MIT Press, Cambridge, MA.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.

Fu, T.C., 2011. A review on time series data mining. Eng. Appl. Artif. Intell. 24 (1), 164–181.

Grbić, R., Kurtagić, D., Slišković, D., 2013. Stream water temperature prediction based on Gaussian process regression. Expert Syst. Appl. 40 (18), 7407–7414.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hacine-Gharbi, A., Ravier, P., Harba, R., Mohamadi, T., 2012. Low bias histogram-based estimation of mutual information for feature selection. Pattern Recognit. Lett. 33 (10), 1302–1308.

Han, M., Liu, X., 2013. Feature selection techniques with class separability for multivariate time series. Neurocomputing 110, 29–34.

Han, M., Wang, Y., 2009. Analysis and modeling of multivariate chaotic time series based on neural network. Expert Syst. Appl. 36 (2), 1280–1290.

Keynia, F., 2012. A new feature selection algorithm and composite neural network for electricity price forecasting. Eng. Appl. Artif. Intell. 25 (8), 1687–1697.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69 (6), 066138.

Kwak, N., Choi, C.H., 2002. Input feature selection for classification problems. IEEE Trans. Neural Netw. 13 (1), 143–159.

Kwak, N., Choi, C.H., 2002. Input feature selection by mutual information based on Parzen window. IEEE Trans. Pattern Anal. Mach. Intell. 24 (12), 1667–1671.

Lee, J., Kim, D.W., 2013. Feature selection for multi-label classification using multivariate mutual information. Pattern Recognit. Lett. 34 (3), 349–357.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27 (8), 1226–1238.

Shannon, C.E., 2001. A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. 5 (1), 3–55.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231.

Suzuki, T., Sugiyama, M., Sese, J., Kanamori, T., 2008. Approximating mutual information by maximum likelihood density ratio estimation. In: JMLR Workshop and Conference Proceedings, vol. 4, pp. 5–20.

Takens, F., 1981. Detecting strange attractors in turbulence, Dynamical Systems and Turbulence. Springer, Berlin Heidelberg, pp. 366–381.

Van Hulle, M.M., 2005. Edgeworth approximation of multivariate differential entropy. Neural Comput. 17 (9), 1903–1910.

Vergara, J.R., Estévez, P.A., 2014. A review of feature selection methods based on mutual information. Neural Comput. Appl. 24 (1), 175–186.

Walters-Williams, J., Li, Y., 2009. Estimation of mutual information: a survey. Rough Sets Knowl. Technol., 389–396.

Wang, X., Han, M., Wang, J., 2010. Applying input variables selection technique on input weighted support vector machine modeling for BOF endpoint prediction. Eng. Appl. Artif. Intell. 23 (6), 1012–1018.

Wu, C.L., Chau, K.W., 2013. Prediction of rainfall time series using modular soft computing methods. Eng. Appl. Artif. Intell. 26 (3), 997–1007.

Zeng, X., Durrani, T.S., 2011. Estimation of mutual information using copula density function. Electron. Lett. 47 (8), 493–494.

Zheng, Y., Kwoh, C.K., 2011. A feature subset selection method based on high-dimensional mutual information. Entropy 13 (4), 860–901.