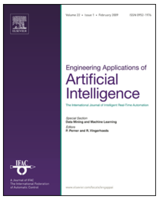
Engineering Applications of Artificial Intelligence 37 (2015) 250–257

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