Neurocomputing 139 (2014) 408–414

Contents lists available at ScienceDirect

Neurocomputing

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

Localization and regularization of normalized transfer entropy

Heeyoul Choi

Samsung Advanced Institute of Technology, Samsung Electronics 97, Samsung2-ro, Giheung-gu, Yongin-si, Gyeonggi-do 446-712, South Korea

ARTICLE INFO

Article history:
Received 13 February 2013
Received in revised form
29 December 2013
Accepted 22 February 2014
Communicated by S. Choi
Available online 13 April 2014

Keywords:
Information theory
Local transfer entropy
Dirichlet distribution
Regularization

ABSTRACT

To find hidden structures of a data set, it is important to understand the relationship between variables such as genes or neurons. As a measure of such relationship, causality is to find directed relations between the variables, which can reveal more of the structures than undirected relations. As a quantitative measure of such causal relationship, transfer entropy has been proposed and successfully applied to capture the amount of information flow between events and sequences. In order to analyze the flow locally in time, we propose to localize normalized transfer entropy and regularize it to avoid the unstable result. Experiment results with synthetic and real-world data confirm the usefulness of our algorithm.

1. Introduction

In data analysis, it is important to understand the relationship among variables, events and sequences. Once we measure relationships between the variables, we can represent the data set on a metric space [1–4] or build up a model for inference on them [5]. Such relationship can be obtained by symmetric measurements like correlation and distance, or asymmetric ones like transition probability and causality.

Causality is a kind of asymmetric (or directional) relationship between two variables. Although real causality is not possible to find, there are some methods to measure quantitative amounts of causality-like properties. Granger causality test is one of the popular methods to measure such amounts [6], which is based on a linear regression method. To measure the casual relation which cannot be captured by the covariance, transfer entropy (TE) (also known as directed transinformation [7], and directed information [15]) was proposed several decades ago and recently started to draw much attention [7,8]. Transfer entropy has been applied to many research areas like neuroscience and cognitive science [9,10], where it measures how much information flows between variables over the whole time series. Transfer entropy was normalized by the total information of the variable after removing the bias effect, which leads to *normalized transfer entropy (NTE)* [11].

NTE does not measure the flow locally in time, while localized information flow could be more helpful in understanding the relationship. For example, after finding out that neuron A affects neuron B over the whole time, it can be more informative to figure when and how much A affects B. Recently several localized versions of transfer entropy have been proposed [12–14], but normalized transfer entropy has not been localized. So, in this paper, we localize NTE after removing the bias effect as the previous NTE does, which leads to local NTE. The rationale for the need of localization of NTE over TE is that NTE can make the transfer information comparable in different environments or between different pairs of variables without bias effects, while TE cannot. More discussions about normalization can be found in [11].

Another issue in transfer entropy is that when data set is not dense enough, there is a sampling problem as pointed out by [9], since the entropies are calculated based on the probabilities. That is, when the data samples are sparse in the data space, the probability density estimation is too much biased to a few samples. To overcome this sparseness problem, a priori distribution can be used [9]. In this paper, we propose to regularize the local NTE based on Dirichlet distribution as a prior distribution for the probabilities, which leads to *regularized local NTE*.

The rest of this paper is organized as follows. First, we briefly review some previous works including transfer entropy and its localized version in Section 2. Then in Section 3, we define local NTE and regularize it to avoid sparseness problem. Experiment results with synthetic and real-world data sets show how our algorithm works in Section 4, followed by conclusion in Section 5.

2. Previous works

Transfer entropy (TE) was proposed decades ago by Marko [7], and recently started to draw much attention since Schreiber's

E-mail address: heeyoul@gmail.com

http://dx.doi.org/10.1016/j.neucom.2014.02.016
0925-2312/

work [8]. It has been successfully applied to understand how much information flows between variables (or channels) in many research areas [15,11,10]. In this section, we briefly review some previous works including all the steps in transfer entropy and its localized version.

2.1. Symbolization and probability mass function

When the variables in the data set are continuous variables, to make the problem simple, symbolization of the variables can be applied so that the data can be considered as a set of discrete sequences. As in the SAX algorithm [16], we use the distribution of the piecewise aggregate approximates (PAAs) and make a uniformly distributed symbol set based on the histogram of the PAAs. If the data set is given in discrete sequences, symbolization is not necessary.

Let $A=\{a_1,a_2,\ldots,a_K\}$ be a set of the symbols of the two discrete variables along the time $t=1,\ldots,N:\boldsymbol{x}=[x_1,x_2,\ldots,x_N]$ and $\boldsymbol{y}=[y_1,y_2,\ldots,y_N]$, and $x_t\in A$ and $y_t\in A$. $p(x)$ be the probability mass function of $\boldsymbol{x}$. Basically, the probability functions are obtained by frequency of the symbols, that is, $p(x)=n_x/N$, where n_x is the number of times that symbol x happens in the sequence.

2.2. Transfer entropy

There are several entropy measures related to TE [7]. Shannon entropy of $\boldsymbol{x}$ and conditional entropy (or entropy rate) of $\boldsymbol{x}$ given $\boldsymbol{x}^p$ are, respectively, defined by

$$H(\boldsymbol{x})=-\sum_{x\in A}p(x)\log\ p(x), \tag{1}$$

$$H(\boldsymbol{x}|\boldsymbol{x}^p)=-\sum_{x,x^p\in A}p(x,x^p)\log\ p(x|x^p), \tag{2}$$

where $\boldsymbol{x}^p$ is the past values of $\boldsymbol{x}$. Conditional entropy means the total information of current $\boldsymbol{x}$ knowing the previous values. Free information of $\boldsymbol{x}$ is given by $\boldsymbol{x}^p$ and $\boldsymbol{y}^p$, $H(\boldsymbol{x}|\boldsymbol{x}^p,\boldsymbol{y}^p)$ is defined in a similar way and it means the information of current $\boldsymbol{x}$ is independent of $\boldsymbol{y}$:

$$H(\boldsymbol{x}|\boldsymbol{x}^p,\boldsymbol{y}^p)=-\sum_{x,x^p,y^p\in A}p(x,x^p,y^p)\log\ p(x|x^p,y^p). \tag{3}$$

Transfer entropy from $\boldsymbol{y}$ to $\boldsymbol{x}$ is given by

$$TE(\boldsymbol{y},\boldsymbol{x})=H(\boldsymbol{x}|\boldsymbol{x}^p)-H(\boldsymbol{x}|\boldsymbol{x}^p,\boldsymbol{y}^p), \tag{4}$$

which means the information of current $\boldsymbol{x}$ is coming from $\boldsymbol{y}$ but not from the previous values of $\boldsymbol{x}$ itself.

2.3. Normalized transfer entropy

Transfer entropy might include another kind of information from $\boldsymbol{y}$ to $\boldsymbol{x}$ that is not what we want to count [17]. For example, the future of $\boldsymbol{y}$ can incidentally give some amount of information to $\boldsymbol{x}$, especially when the data samples are sparse. To remove such inevitable information from $\boldsymbol{y}$ to $\boldsymbol{x}$ in the transfer entropy, we need to keep the distribution of $\boldsymbol{y}$ as independent of $\boldsymbol{x}$ as possible over the whole sequence. If $\boldsymbol{y}$ is independent of $\boldsymbol{x}$, $H(\boldsymbol{x}|\boldsymbol{x}^p,\boldsymbol{y}^p)=H(\boldsymbol{x}|\boldsymbol{x}^p)$ which makes $TE(\boldsymbol{y},\boldsymbol{x})=0$. That is, even after shuffling (or reordering randomly) $\boldsymbol{y}$, if $TE(\boldsymbol{y}^s,\boldsymbol{x})>0$, where $\boldsymbol{y}^s$ is the shuffled sequence of $\boldsymbol{y}$, that is not the information we want to measure. So we remove that from the real $TE(\boldsymbol{y},\boldsymbol{x})$ as in [17]. Instead of shuffling, we can shift $\boldsymbol{y}$ assuming that if a signal is shifted enough, there is no mutual information between two signals.

Finally, normalized transfer entropy (NTE) from $\boldsymbol{y}$ to $\boldsymbol{x}$ [11] is given by

$$NTE(\boldsymbol{y},\boldsymbol{x})=\frac{TE(\boldsymbol{y},\boldsymbol{x})-TE(\boldsymbol{y}^s,\boldsymbol{x})}{H(\boldsymbol{x}|\boldsymbol{x}^p)}, \tag{5}$$

where $TE(\boldsymbol{y}^s,\boldsymbol{x})$ is a bias term with a shuffled (or shifted) sequence of $\boldsymbol{y}$. That is, the bias term indicates the amount of information flowing even when the sequence is shuffled. NTE represents the ratio of information in the current $\boldsymbol{x}$ transferring from $\boldsymbol{y}$ to the total information of $\boldsymbol{x}$ after removing the bias effect. The rationale for the need of the bias term and normalization was discussed in [11]. As briefly described, NTE makes comparison between transfer entropies possible.

2.4. Local transfer entropy

Given two sequences of symbols, TE or NTE gives us just one value which indicates how much information is flowing between the two sequences over the whole time. To understand the information locally in space or time, local transfer entropy can be used [12,13]. Local TE is almost the same as Eq. (4), except not taking the expectation. The local entropies at time $t\in[1,N]$ are defined as follows.

Local entropy, also known as surprise or information, of $\boldsymbol{x}$ at a specific time t is defined by

$$H_t(\boldsymbol{x})=-\log\ p(x_t), \tag{6}$$

where $\sum_t H_t(\boldsymbol{x})=H(\boldsymbol{x})$ in Eq. (1). That is, $H_t(\boldsymbol{x})$ indicates the contribution to the entropy $H(\boldsymbol{x})$ at a specific time t. The conditional entropy is localized in the same way as follows:

$$H_t(\boldsymbol{x}|\boldsymbol{x}^p)=-\log\ p(x_t|x_{t-1}), \tag{7}$$

where the range of t is $[2,N]$. Here, we consider just the order 1 Markov process to reduce the computation complexity so that $\boldsymbol{x}^p$ is replaced by x_{t-1}. Also, free information of $\boldsymbol{x}$ is localized by

$$H_t(\boldsymbol{x}|\boldsymbol{x}^p,\boldsymbol{y}^p)=-\log\ p(x_t|x_{t-1},y_{t-1}). \tag{8}$$

Finally, local transfer entropy from $\boldsymbol{y}$ to $\boldsymbol{x}$ [12] is defined by

$$TE_t(\boldsymbol{y},\boldsymbol{x})=H_t(\boldsymbol{x}|\boldsymbol{x}^p)-H_t(\boldsymbol{x}|\boldsymbol{x}^p,\boldsymbol{y}^p), \tag{9}$$

and $\sum_t TE_t(\boldsymbol{y},\boldsymbol{x})=TE(\boldsymbol{y},\boldsymbol{x})$, where $TE_t(\boldsymbol{y},\boldsymbol{x})$ indicates the contribution to the transfer entropy $TE(\boldsymbol{y},\boldsymbol{x})$ at a specific time t.

3. Proposed algorithm

3.1. Localization of normalized transfer entropy

Although NTE is more appropriate than TE to compare many pairs, there is no localized version of NTE. So, in light of localizing TE, we localize NTE, which leads to *local NTE*, as follows:

$$NTE_t(\boldsymbol{y},\boldsymbol{x})=\frac{TE_t(\boldsymbol{y},\boldsymbol{x})-\frac{1}{N-1}\sum_t TE_t(\boldsymbol{y}^s,\boldsymbol{x})}{\sum_t H_t(\boldsymbol{x}|\boldsymbol{x}^p)}, \tag{10}$$

where

$$\sum_t TE_t(\boldsymbol{y}^s,\boldsymbol{x})=TE(\boldsymbol{y}^s,\boldsymbol{x}),$$

$$\sum_t H_t(\boldsymbol{x}|\boldsymbol{x}^p)=H(\boldsymbol{x}|\boldsymbol{x}^p).$$

To make the shuffled effect consistent over the whole time series, we take the average of $TE_t(\boldsymbol{y}^s,\boldsymbol{x})$ over the whole time which is $TE(\boldsymbol{y}^s,\boldsymbol{x})/(N-1)$, and we normalize the equation by $\sum_t H_t(\boldsymbol{x}|\boldsymbol{x}^p)$ to make sure that $\sum_t NTE_t(\boldsymbol{y},\boldsymbol{x})=NTE(\boldsymbol{y},\boldsymbol{x})$ as follows:

$$\begin{aligned}\sum_t NTE_t(\boldsymbol{y},\boldsymbol{x})&=\frac{\sum_t TE_t(\boldsymbol{y},\boldsymbol{x})-\sum_t\frac{1}{N-1}TE(\boldsymbol{y}^s,\boldsymbol{x})}{H(\boldsymbol{x}|\boldsymbol{x}^p)}\\&=\frac{TE(\boldsymbol{y},\boldsymbol{x})-TE(\boldsymbol{y}^s,\boldsymbol{x})}{H(\boldsymbol{x}|\boldsymbol{x}^p)}\\&=NTE_{(\boldsymbol{y},\boldsymbol{x})}.\end{aligned}$$