Information Sciences 280 (2014) 98–110

Contents lists available at ScienceDirect

Information Sciences

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

On establishing nonlinear combinations of variables from small to big data for use in later processing

Jerry M. Mendel *, Mohammad M. Korjani

Signal and Image Processing Institute, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2564, United States

ARTICLE INFO

Article history:
Received 29 December 2013
Received in revised form 29 March 2014
Accepted 25 April 2014
Available online 9 May 2014

Keywords:
Big data
Causal combination
Fast processing
Nonlinear combination
Parallel and distributed processing
Preprocessing

ABSTRACT

This paper presents a very efficient method for establishing nonlinear combinations of variables from small to big data for use in later processing (e.g., regression, classification, etc.). Variables are first partitioned into subsets each of which has a linguistic term (called a *causal condition*) associated with it. Our *Causal Combination Method* uses fuzzy sets to model the terms and focuses on interconnections (*causal combinations*) of either a causal condition or its complement, where the connecting word is AND which is modeled using the minimum operation. Our *Fast Causal Combination Method* is based on a novel theoretical result, leads to an exponential speedup in computation and lends itself to parallel and distributed processing; hence, it may be used on data from small to big.

1. Introduction

Suppose one is given data of any size, from small to big, for a group of v input variables that one believes caused[1] an output, and that one does not know which (nonlinear) combinations of the input variables caused the output. *This paper presents a very efficient method for establishing the initial (nonlinear) combinations of variables that can then be used in later modeling and processing.* For example, in nonlinear regression (e.g., [21,22]) one needs to choose the nonlinear interactions among the variables as well as the number of terms in the regression model,[2] in pattern classification (e.g., [7,2]) that is based on mathematical features (e.g., [23]) one needs to choose the nonlinear nature of those features as well as the number of such features, and in some neural networks (e.g., [11]) one needs to know which combinations of the inputs and how many such combinations should be fanned out to one or more of the network's various layers. Our *Causal Combination Method* (CCM) that is described in Section 3 provides the initial combinations of the variables as well as their number, and can also be used in later processing to readjust the combinations of the variables as well as their number that are used in a model. Our *Fast Causal Combination Method* (FCCM) that is also described in Section 3 is a very efficient way of implementing CCM for data of any size.

Establishing which combinations of variables to use in a model can be interpreted as a form of data preprocessing. According to [24]: "Data preprocessing is a data mining technique that involves transforming raw data into an understandable

* Corresponding author. Tel.: +1 213 740 4445; fax: +1 213 740 4456.
E-mail address: mendel@sipi.usc.edu (J.M. Mendel).

[1] How to choose the variables is crucial to the success of any model. This paper assumes that the user has already established the variables that (may) affect the outcome.

[2] According to [5, p. 20], "... in practice, due to complex and often informal nature of a priori knowledge, ... specification of approximating functions may be difficult or impossible."

http://dx.doi.org/10.1016/j.ins.2014.04.042
0020-0255/

format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing." According to [13] data preprocessing includes cleaning, normalization, transformation, feature extraction and selection. Our preprocessing is about transformation of raw data into patterns.

CCM focuses on interconnections of either a causal condition (defined in Section 2) or its complement where the connecting word is AND which is modeled using the minimum operation. Note that, because you might be wrong about postulating a cause you protect yourself against this by considering both the cause and its complement (an idea that was first suggested[3] by Ragin in [19 p. 131]). The interconnection of either a causal condition or its complement for all of the v variables is called a *causal combination*. As will be seen in Section 2, there can be a (very) large number of candidate causal combinations. CCM prunes the (very) large number of candidate causal combinations (using data about the v input variables) to a much smaller subset of *surviving causal combinations*, and FCCM does this in a very efficient way. These surviving causal combinations are the nonlinear combinations of the input variables that can then be used in later modeling or processing.

In summary, *this paper presents a very efficient method for establishing the initial (nonlinear) combinations of variables that can then be used in later modeling and processing by using a novel form of preprocessing that transforms raw data into patterns through the use of fuzzy sets. We will show that our method lends itself to massive distributed and parallel processing which makes it suitable for data of all sizes from small to big.*

The rest of this paper is organized as follows: Section 2 describes the terminology and approach that is used in the rest of the paper; Section 3 provides the main results for CCM and FCCM; Section 4 quantifies the computational speedup for FCCM; Section 5 provides some additional ways to speed up FCCM; and Section 6 draws conclusions.

2. Terminology and approach

A *data pair* is $(\mathbf{x}(t), y(t))$, where $\mathbf{x}(t) = col(x_1(t),\ldots,x_v(t))$, $x_i(t)$ is the ith input variable and $y(t)$ is the output for that $\mathbf{x}(t)$. Each data pair is treated as a "case," index t denotes a data case and there does not have to be a unique natural ordering of the cases over t (in a multi-variable approximation application there is no natural ordering of the data cases, but in a time-series forecasting application the data cases would have a natural temporal ordering). We assume that N data pairs are available, and refer to the collection of these data pairs as S_{Cases}, where:

$$S_{Cases} = \{(\mathbf{x}(t), y(t))\}_{t=1}^{N} \tag{1}$$

For Big Data [10, Table I], N ranges from *huge* ($O(N) = 10^{10}$) to *monster* ($O(N) = 10^{12}$) to *very large* ($O(N) = 10^{>12}$).

We begin by partitioning each *input variable* into subsets each of which may be thought of as having a linguistic term[4] associated with it, e.g., the variable *Pressure* can be partitioned into *Low Pressure*, *Moderate Pressure* and *High Pressure*. Because it is very difficult to know where to draw a crisp line between each of the subsets, so as to separate one from the other, they are modeled herein as fuzzy sets, and, there can be from 1 to n_v subsets (*terms*) for each input variable. The terms for each input variable that are actually used in CCM are called *causal conditions*.

If one chooses to use only one term for each variable (e.g., *Profitable Company*, *Educated Country*, *Permeable Oil Field*, etc.), then the words "variable," "term" and "causal condition" can be interchanged, i.e., they are synonymous. If, on the other hand, one chooses to use more than one term for each variable (e.g., *Low Pressure*, *Moderate Pressure* and *High Pressure*), i.e., to *granulate* [1] each variable, as is very commonly done in engineering and computer science applications, then one must distinguish between the words "variable," "term" and "causal condition." We elaborate further on this next.

If, e.g., there are V variables, each described by n_v terms ($v = 1,\ldots,V$) then (as in fsQCA [16,17,19,20]) each of the terms will be treated as a separate causal condition[5] (this is illustrated below in Example 1), and, as a result, there will be $k = n_1 + n_2 + \cdots + n_V$ causal conditions.

We let ξ_v ($v = 1,\ldots,V$) denote a variable and $T_l(\xi_v)$ ($l = 1,\ldots,n_v$) denote the terms for each variable. For simplicity, in this paper the same numbers of terms are used for each variable, i.e. $n_v = n_c$ for $\forall v$, so that the total number of causal conditions is $k = n_cV$.

The terms are organized according to the (non-unique) ordering of the V input variables, as $\{T_1(\xi_1),\ldots, T_{n_c}(\xi_1),\ldots,T_1(\xi_V),\ldots,T_{n_c}(\xi_V)\}$. This set of n_cV terms is then mapped into an ordered set of *possible causal conditions*, $S_{C'}$, as follows:

$$\begin{aligned}\{T_1(\xi_1),\ldots,T_{n_c}(\xi_1),\ldots,T_1(\xi_V),\ldots,T_{n_c}(\xi_V)\} &\rightarrow \left\{C'_1(\xi_1),\ldots,C'_{n_c}(\xi_1),\ldots,C'_{n_c(V-1)+1}(\xi_V),\ldots,C'_{n_cV}(\xi_V)\right\}\\ &\equiv \left\{C'_1,\ldots,C'_{n_c},\ldots,C'_{n_c(V-1)+1},\ldots,C'_{n_cV}\right\}\end{aligned} \tag{2}$$

[3] Traditional interconnections usually do not consider both a cause and its complement; in fact, one almost never sees the complement of a cause in an interconnection of causes (e.g., in the antecedents of either a crisp or fuzzy rule).

[4] The actual names that are given to the subsets are not important for this paper, e.g., they may be given linguistically meaningful names (as in our example of *Pressure*) or symbolic names (e.g., *A*, *B*, *C*; *T1*, *T2*, *T3*; etc.).

[5] One may raise an objection to doing this because of perceived correlations between terms for the same variable (e.g., perhaps *Low Pressure* and *Moderate Pressure* are highly correlated, or *Moderate Pressure* and *High Pressure* are highly correlated). Such perceptions depend on how overlapped the fuzzy sets are for the terms and does not have to be accounted for during CCM because the mathematics for CCM will take care of the overlap automatically.