ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

A hybrid algorithm with cluster analysis in modelling high dimensional data

Burcu Tunga

İstanbul Technical University, Faculty of Science and Letters, Mathematics Engineering Department, Maslak 34469, İstanbul, Turkey

ARTICLE INFO

Article history:
Received 2 August 2015
Accepted 5 September 2017
Available online xxxx

Keywords:
High dimensional problems
Data modelling
Cluster analysis
Approximation
EMPR

ABSTRACT

Multivariate data modelling aims to predict unknown function values through an established mathematical model. It is essential to construct an analytical structure using the given set of high dimensional data points with corresponding function values. The level of multivariance directly affects the modelling process. Increase in the number of independent variables makes the standard numerical methods incapable of obtaining the sought analytical structure. This work aims to overcome the difficulties of high multivariance and to improve the modelling quality by carrying out two main steps: data clustering and data partitioning. Data clustering step deals with dividing the whole problem domain into several clusters by performing k-means clustering algorithm. Data partitioning step performs the Enhanced Multivariance Product Representation method to partition the high dimensional data set of each cluster. The analytical structure is obtained through the partitioned data for each cluster and can be used to predict the unknown function values.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Estimating the tendency of a multivariate system with a large number of independent variables is an important task in many research areas. Multivariate data modelling is used to describe the behaviour of multivariate systems. In this study, it is assumed that a high dimensional data set and the corresponding function values are given and it is expected to predict unknown function values. For this purpose, a mathematical model is established by constructing an analytical structure.

In many cases, there are many independent variables (features) and data points (nodes) to model the considered problem. Increase in multivariance causes difficulties to generate an analytical structure. Therefore, scientists mostly prefer to use divide-and-conquer algorithms. The Enhanced Multivariance Product Representation (EMPR) method is one of these algorithms [14,16,12]. EMPR partitions the high dimensional data into univariate, bivariate and higher variate data sets [13,15]. This process is called data partitioning and overcomes the difficulties coming from high multivariance since it is easier to construct an analytical structure using univariate and bivariate data sets instead of a multivariate data set.

This work aims to offer a new hybrid algorithm with two main steps; data clustering and data partitioning. The first step of the algorithm is to implement a clustering method. Many clusters, having nodes with very similar characteristics, can be generated in this way. K-means clustering algorithm, which is very popular for cluster analysis in data mining, is used to generate the clusters [4,2,8,10,6,9]. The algorithm is also used in solving multidimensional engineering problems [7,18]. K-means partitions m observations into k clusters. The algorithm assigns an observation to the nearest cluster by distance. Euclidean distance is used to determine the nearest cluster to the considered observation. Data clustering in the hybrid algorithm offered in this study is implemented using k-means with Euclidean distance metric.

The second step is to partition the multivariate data set of each cluster using the EMPR method and construct an analytical model for each cluster through the partitioned data sets. EMPR has a finite expansion to represent a multivariate

E-mail address: tungab@itu.edu.tr.

http://dx.doi.org/10.1016/j.dam.2017.09.002
0166-218X/© 2017 Elsevier B.V. All rights reserved.

Please cite this article in press as: B. Tunga, A hybrid algorithm with cluster analysis in modelling high dimensional data, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.09.002.

structure. This expansion has a number of components composed of a constant term, univariate terms, bivariate terms and higher variate terms with some special purpose functions called support functions. The support functions provide flexibility in successfully representing various types of analytical structures. The better the support functions are selected, the better the representation of high dimensional data through EMPR will be. Besides, depending on the selection of the most appropriate support function, less number of components from the EMPR expansion will be utilized [14,15]. This decreases the computational complexity of the method.

The steps of the offered hybrid algorithm are combined in a single method called "Clustering EMPR". This new method has some important advantages when it is compared with the other EMPR based methods. Since a clustering algorithm is used to generate the subdomains, each subdomain consists of data points with similar characteristics. It is better to model high dimensional data of these subdomains instead of the whole domain. In addition, Clustering EMPR has the ability of modelling a scattered multivariate data set while the classical EMPR method is only applicable to model high dimensional data of a full mesh.

The paper is organized as follows. The second section covers the main steps of EMPR. The Clustering EMPR method, which is the proposed algorithm of this study, is given in the third section while the fourth section includes the numerical implementations designed to test the performance of our new method. The concluding remarks of the work are given in the last section.

2. Enhanced multivariance product representation

The EMPR method has the following finite expansion used to represent a given multivariate function, $f(x_1, \ldots, x_N)$, in terms of some less variate functions:

$$f(x_1, \ldots, x_N) = f_0 \prod_{j=1}^{N} s_j(x_j) + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) \prod_{\substack{j=1 \\ j\neq i_1}}^{N} s_j(x_j) + \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{N} f_{i_1 i_2}(x_{i_1}, x_{i_2}) \prod_{\substack{j=1 \\ j\neq i_1,i_2}}^{N} s_j(x_j) + \cdots + f_{1\ldots N}(x_1, \ldots, x_N) \tag{1}$$

where N is the number of independent variables [14]. The expansion also includes the product of certain univariate functions, $s_j(x_j)$, which are called "support functions". The existence of these support functions allows the algorithm to successfully represent different types of analytical structures such as polynomials with additive, multiplicative, hybrid natures, exponential functions, and trigonometric functions. The other terms at the right hand side of the expansion are the orthogonal decomposition components of the multivariate function, $f(x_1, \ldots, x_N)$ which are called "EMPR components". The vanishing conditions on these components are given as follows:

$$\int_{a_{i_\ell}}^{b_{i_\ell}} dx_{i_\ell} W_{i_\ell}\left(x_{i_\ell}\right) s_{i_\ell}\left(x_{i_\ell}\right) f_{i_1,\ldots,i_p}\left(x_{i_1}, \ldots, x_{i_p}\right) = 0,$$
$$1 \le p \le N, \quad 1 \le \ell \le p, \quad 1 \le i_1 < \cdots < i_p \le N \tag{2}$$

where $W_{i_\ell}\left(x_{i_\ell}\right)$ is a component of a product type weight function. The weight function structure is given as follows

$$W(x_1, \ldots, x_N) = \prod_{j=1}^{N} W_j(x_j) \tag{3}$$

and the following normalization criteria on the weight function components are defined for simple derivations [13]:

$$\int_{a_i}^{b_i} dx_i\, W_i(x_i) = 1, \quad \int_{a_i}^{b_i} dx_i\, W_i(x_i) s_i(x_i)^2 = 1, \quad 1 \le i \le N. \tag{4}$$

The main step of the method is to uniquely determine the right hand side components of the EMPR expansion. To obtain the structure of the constant component, f_0, an operator including an N-tuple integration under weight and support functions is defined as follows [13]:

$$\mathcal{I}_0 f(x_1, \ldots, x_N) \equiv \int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N \left(\prod_{j=1}^{N} W_j(x_j) s_j(x_j) \right) F(x_1, \ldots, x_N) \tag{5}$$

Please cite this article in press as: B. Tunga, A hybrid algorithm with cluster analysis in modelling high dimensional data, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.09.002.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات