Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 114 (2017) 216–223

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA

Knowledge-based Comparable Predicted Values in Regression Analysis

Mika Sato-Ilic

Faculty of Engineering, Information and Systems, University of Tsukuba, Tennodai 1-1-1, Tsukuba, Ibaraki 305-8573, Japan

Abstract

Regression analysis has a major role in predicting values of a dependent variable by using values of independent variables. The estimate of predicted values is obtained as projected values in a linear subspace spanned by vectors of independent variables. However, if the data set has been observed simultaneously from multiple different data sources, then we must create different linear subspaces to estimate the different predicted values corresponding to the different data sources. Then, we cannot compare the different predicted values, since the linear subspaces are different. In order to solve this problem, we propose a method to obtain comparable predicted values obtained from different data sources by utilizing a fuzzy clustering result and an orthogonal projector which projects two different vectors corresponded with the two different dependent variables to the same intersection of the two different linear subspaces. From this, since the different predicted values from different data sources can be obtained in the common space, we can compare the different predicted values.

Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.

Keywords: Regression analysis; Projection; Subspace

1. Introduction

The recent growth of data means the speed of analysis has become increasingly important, as has the problem of how to predict the future value when data is observed in real time. For this prediction, especially, if the data observed from the multiple data sources, then we often need to compare the predicted values over the multiple data sources. For example, if the data is observed as time series data with respect to independent variables and a dependent variable, and such data sets are observed over several subjects, then in many occasions, we need to

1877-0509
Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.
10.1016/j.procs.2017.09.063

compare the predicted values over the subjects. However, the data of times with respect to variables are obtained differently at each subject, the estimates of predicted values are obtained in the different subspaces, so we cannot compare the predicted values over the subjects. In another example, if we observe a data set of student's examinations with respect to independent and dependent variables over the several semesters, then we often need to compare the predicted values of a dependent variable over the several semesters. In this case, we have to have the predicted values in the same linear subspace in order to obtain the comparable predicted values over the semesters; however, in the ordinary linear regression analysis, we cannot obtain the comparable predicted values over the semesters.

The main reason for this problem is that the ordinary regression model has the estimate of a dependent variable in a linear subspace spanned by vectors of independent variables, and the vectors of independent variables are mostly different to each other over the different data sources. Therefore, in this case, we have difficulty in obtaining the intersection of the linear subspaces whose each linear subspace has a projected vector of each dependent variable for each subject.

In order to solve this problem, in this paper, we utilize the common clusters over the subjects, obtained as the result of fuzzy clustering for 3-way data, as the common scale over the subjects, and use this common scale as the basic axis to measure the degree of contribution of independent variables of all the subjects to the basic axis. And by using the value of the degree of the contribution, we select the efficient independent variables which span a linear subspace for each subject. Since the independent variables of each subject are selected from all independent variables of all subjects, we can obtain the intersection of the linear subspaces consisted of each subspace corresponding to each subject. Therefore, we can estimate the predicted values in the same linear subspace so as to compare the estimates over the subjects. In addition, the obtained clusters are able to be considered as knowledge to explain the latent structure of data. So, selecting the independent variables according to the similarity to the basic axis of the clusters means selecting the independent variables based on the common knowledge through a data structure over the different data sources.

This paper consists of the following: Section 2 describes a fuzzy clustering for 3-way data [6] whose result is used in the proposed method. Section 3 describes the ordinary linear regression model. Section 4 proposes a method to obtain the comparable predicted values in the regression analysis. Section 5 presents numerical examples of the proposed method, and Section 6 contains the conclusions.

2. Fuzzy Clustering for 3-Way Data

Suppose Z_t be a given data matrix consisted of n objects with respect to p variables at a subject t called a 3-way data and shown as follows:

$$Z_t = (z_{ir}^{(t)}), \quad i = 1,\cdots,n, \ r = 1,\cdots,p, \ t = 1,\cdots,T.$$

In order to obtain the same clusters over the T subjects, the following $nT \times p$ super matrix $\widetilde{Z}$ is created.

$$\widetilde{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_T \end{pmatrix} = (\widetilde{z}_{jr}), \quad j = 1,\cdots,nT, \ r = 1,\cdots,p. \tag{1}$$

The purpose of this fuzzy clustering is to classify the nT objects into K clusters. The state of the fuzzy clustering is represented by a partition matrix:

$$\widetilde{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_T \end{pmatrix} = (\widetilde{u}_{jk}), \quad U_t = (u_{ik}^{(t)}), \quad j = 1,\cdots,nT, \ i = 1,\cdots,n, \ k = 1,\cdots,K, \ t = 1,\cdots,T, \tag{2}$$

where $\widetilde{u}_{jk}$ is a degree of belongingness of an object j which is shown as $\widetilde{\mathbf{z}}_j = (\widetilde{z}_{j1},\cdots,\widetilde{z}_{jp})$ to a fuzzy cluster k and $u_{ik}^{(t)}$ is a degree of belongingness of an object i to the same fuzzy cluster k at a subject t. From (2), it can be seen