Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 850 – 859

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

A variable selection method considering cluster loading for labeled high dimension low sample size data

Jiaxin Chen[a], Mika Sato-Ilic[b,*]

[a] *Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan*
[b] *Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan*

Abstract

As the information society rapidly develops, there is an increased importance placed of dealing with high dimension low sample size (HDLSS) data, whose number of variables is much larger than the number of objects. Moreover, the selection of effective variables for HDLSS data is becoming more crucial. In this paper, a variable selection method considering cluster loading for labeled HDLSS data is proposed. Related to cluster loading, the conventional model considering principal component analysis has been proposed. However, the model can not be used for HDLSS data. Therefore, we propose a cluster loading model using a clustering result. By using the obtained cluster loading, we can select variables which belong to clusters unrelated with the given discrimination information represented by the labels of objects. Several numerical examples show a better performance of the proposed method.

Peer-review under responsibility of KES International

Keywords: HDLSS data; variable selection; cluster loading; clustering

1. Introduction

With the rapid development of the information society, data analysis of HDLSS data[1,2,3,4] is getting more and more valued, especially in the specific application areas such as microarray gene data and image data. In HDLSS data, the number of variables is much larger than the number of objects, so how to select the effective variables plays an important role in analyzing HDLSS data. In this paper, a variable selection method considering cluster loading for labeled HDLSS data is proposed.

The cluster loading is defined as the relationship between given labels of objects and obtained clusters of variables. Related to the cluster loading, a method for constrained principal component analysis (CPCA)[5] has been proposed, which is based on principal component analysis with external information on both objects and variables. In this model, the relationship between objects and variables is estimated by using least squares method. As a mathematically similar model to the CPCA, a model for interpreting principle components by using discrimination information[6] has

* Corresponding author. Tel.: +81-29-853-5006 ; fax: +81-29-853-5006.
E-mail address: mika@risk.tsukuba.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.248

been proposed. In this model, the discrimination information is represented as the labels of objects and relationship between given labels and obtained principal components is estimated. The cluster loading is defined by using the framework of this model mathematically. However, the previously proposed model for the interpretation of principal components can not be used for HDLSS data. Because the result from principal component analysis is associated with the eigenvalues which are calculated through variance-covariance matrix. In HDLSS data, these obtained eigenvalues of samples are not consistent with the eigenvalues of population. As a result, this model is not applicable to HDLSS data.

By using the obtained cluster loading, we can select variables which belong to clusters unrelated with the given labels. This means that we can select variables which do not relate to the discrimination of objects. The selected variables are the targets to be deleted variables.

The construction of this paper is as follows: In Section 2, a method for constrained principal component analysis (CPCA) is referred. In Section 3, a method for interpreting principal components using discrimination information is described. In Section 4, a cluster loading model for HDLSS data is proposed, and section 5 proposes a variable selection method using the cluster loading. Section 6 shows numerical examples of the proposed method. In Section 7, conclusions are stated.

2. Constrained principal component analysis

Let $\mathbf{X}$ be a data matrix consisted of n objects and p variables, where $n > p$. Suppose $\mathbf{G}$ a $n \times Q$ object information matrix and $\mathbf{H}$ a $p \times K$ variable information matrix. A model of constrained principal component analysis[5] has been defined as follows:

$$\mathbf{X} = \mathbf{GMH^T} + \mathbf{BH^T} + \mathbf{GC} + \mathbf{E}, \tag{1}$$

where $\mathbf{M}$, $\mathbf{B}$ and $\mathbf{C}$ are the matrices of coefficients to be estimated, and $\mathbf{E}$ is a matrix of errors. In equation (1), $\mathbf{X}$ is defined by three terms, where the first term can be explained by both $\mathbf{G}$ and $\mathbf{H}$, the second term is only defined by $\mathbf{H}$, the third term is defined by $\mathbf{G}$. That is, $\mathbf{M}$ shows the relationship between summarized objects and summarized variables, $\mathbf{B}$ shows relationship of objects through summarized variables and $\mathbf{C}$ shows relationship of variables through summarized objects.

Estimates of $\mathbf{M}$, $\mathbf{B}$ and $\mathbf{C}$ are obtained by minimizing $SS(\mathbf{E}) = tr(\mathbf{E^T E})$, where $SS(\mathbf{E})$ shows the sum of square of errors. From equation (1), we consider the first term of the model and define a model as follows:

$$\mathbf{X} = \mathbf{GMH^T} + \mathbf{E_1}. \tag{2}$$

From equation (2),

$$\begin{aligned} SS(\mathbf{E_1}) &= tr(\mathbf{E_1^T E_1}) \\ &= tr((\mathbf{X} - \mathbf{GMH^T})^{\mathbf{T}}(\mathbf{X} - \mathbf{GMH^T})) \\ &= tr(\mathbf{X^T X}) - tr(\mathbf{X^T GMH^T}) - tr(\mathbf{HM^T G^T X}) + tr(\mathbf{HM^T G^T GMH^T}). \end{aligned} \tag{3}$$

From equation (3) and

$$\begin{aligned} \frac{\partial tr(\mathbf{E_1^T E_1})}{\partial \mathbf{M}} &= \frac{\partial tr(\mathbf{X^T X})}{\partial \mathbf{M}} - \frac{\partial tr(\mathbf{X^T GMH^T})}{\partial \mathbf{M}} - \frac{\partial tr(\mathbf{HM^T G^T X})}{\partial \mathbf{M}} + \frac{\partial tr(\mathbf{HM^T G^T GMH^T})}{\partial \mathbf{M}} \\ &= -2(\mathbf{G^T XH}) + 2(\mathbf{G^T GMH^T H}) \\ &= 0, \end{aligned}$$

we obtain $\hat{\mathbf{M}}$ which is the estimate of $\mathbf{M}$ as follows:

$$\hat{\mathbf{M}} = (\mathbf{G^T G})^{-}\mathbf{G^T XH}(\mathbf{H^T H})^{-}, \tag{4}$$

where $(\mathbf{G^T G})^{-}$ and $(\mathbf{H^T H})^{-}$ are g-inverses of $\mathbf{G^T G}$ and $\mathbf{H^T H}$, respectively.