Factor Analysis in Data Mining

Zu-Hsu Lee
Montclair State University, USA

Richard L. Peterson
Montclair State University, USA

Chen-Fu Chien
National Tsing Hua University, Taiwan

Ruben Xing
Montclair State University, USA

INTRODUCTION

The rapid growth and advances of information technology enable data to be accumulated faster and in much larger quantities (i.e., data warehousing). Faced with vast new information resources, scientists, engineers, and business people need efficient analytical techniques to extract useful information and effectively uncover new, valuable knowledge patterns.

Data preparation is the beginning activity of exploring for potentially useful information. However, there may be redundant dimensions (i.e., variables) in the data, even after the data are well prepared. In this case, the performance of data-mining methods will be affected negatively by this redundancy. Factor Analysis (FA) is known to be a commonly used method, among others, to reduce data dimensions to a small number of substantial characteristics.

FA is a statistical technique used to find an underlying structure in a set of measured variables. FA proceeds with finding new independent variables (factors) that describe the patterns of relationships among original dependent variables. With FA, a data miner can determine whether or not some variables should be grouped as a distinguishing factor, based on how these variables are related. Thus, the number of factors will be smaller than the number of original variables in the data, enhancing the performance of the data-mining task. In addition, the factors may be able to reveal underlying attributes that cannot be observed or interpreted explicitly so that, in effect, a reconstructed version of the data is created and used to make hypothesized conclusions. In general, FA is used with many data-mining methods (e.g., neural network, clustering).

BACKGROUND

The concept of FA was created in 1904 by Charles Spearman, a British psychologist. The term *factor analysis* was first introduced by Thurston in 1931. Exploratory FA and confirmatory FA are two main types of modern FA techniques. The goals of FA are (1) to reduce the number of variables and (2) to classify variables through detection of the structure of the relationships between variables. FA achieves the goals by creating a fewer number of new dimensions (i.e., factors) with potentially useful knowledge. The applications of FA techniques can be found in various disciplines in science, engineering, and social sciences, such as chemistry, sociology, economics, and psychology. To sum up, FA can be considered as a broadly used statistical approach that explores the interrelationships among variables and determines a smaller set of common underlying factors. Furthermore, the information contained in the original variables can be explained by these factors with a minimum loss of information.

MAIN THRUST

In order to represent the important structure of the data efficiently (i.e., in a reduced number of dimensions), there are a number of techniques that can be used for data mining. These generally are referred to as multi-dimensional scaling methods. The most basic one is Principle Component Analysis (PCA). Through transforming the original variables in the data into the same number of new ones, which are mutually orthogonal (uncorrelated), PCA sequentially extracts most of the variance (variability) of

Copyright © 2006, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the data. The hope is that most of the information in the data might be contained in the first few components. FA also extracts a reduced number of new factors from the original data set, although it has different aims from PCA.

FA usually starts with a survey or a number of observed traits. Before FA is applied, the assumptions of correlations in the data (normality, linearity, homogeneity of sample, and homoscedasticity) need to be satisfied. In addition, the factors to extract should all be *orthogonal* to one another. After defining the measured variables to represent the data, FA considers these variables as a linear combination of latent factors that cannot be measured explicitly. The objective of FA is to identify these unobserved factors, reflect what the variables share in common, and provide further information about them. Mathematically, let $\boldsymbol{X}$ represent a column vector that contains p measured variables and has a mean vector $\boldsymbol{\mu}$, $\boldsymbol{F}$ stand for a column vector which contains q latent factors, and $\boldsymbol{L}$ be a $p \times q$ matrix that transforms $\boldsymbol{F}$ to $\boldsymbol{X}$. The elements of $\boldsymbol{L}$ (i.e., factor loadings) give the weights that each factor contributes to each measured variable. In addition, let $\boldsymbol{\varepsilon}$ be a column vector containing p uncorrelated random errors. Note that q is smaller than p. The following equation simply illustrates the general model of FA (Johnson & Wichern, 1998):

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon}.$$

FA and PCA yield similar results in many cases, but, in practice, PCA often is preferred for data reduction, while FA is preferred to detect structure of the data.

In any experiment, any one scenario may be delineated by a large number of factors. Identifying important factors and putting them into more general categories generates an environment or structure that is more advantageous to data analysis, reducing the large number of variables to smaller, more manageable, interpretable factors (Kachigan, 1986). Technically, FA allows the determination of the interdependency and pattern delineation of data. It "untangles the linear relationships into their separate patterns as each pattern will appear as a factor delineating a distinct cluster of interrelated data" (Rummel, 2002, Section 2.1). In other words, FA attempts to take a group of interdependent variables and create separate descriptive categories and, after this transformation, thereby decrease the number of variables that are used in an experiment (Rummel, 2002). The analysis procedures can be performed through a geometrical presentation by plotting data points in a multi-dimensional coordinate axis (exploratory FA) or through mathematical techniques to test the specified model and suspected relationship among variables (confirmatory FA).

In order to illustrate how FA proceeds step by step, here is an example from a case study on the key variables (or characteristics) for induction machines, conducted by Maté and Calderón (2000). The sample of a group of motors was selected from a catalog published by Siemens (1988). It consists of 134 cases with no missing values and 13 variables that are power (*P*), speed (*W*), efficiency (*E*), power factor (*PF*), current (*I*), locked-rotor current (*ILK*), torque (*M*), locked-rotor torque (*MLK*), breakdown torque (*MBD*), inertia (*J*), weight (*WG*), slip (*S*), and slope of *M-s* curve (*M_S*). FA can be implemented in the following procedures using this sample data.

Step 1: Ensuring the Adequacy of the Date

The correlation matrix containing correlations between the variables is first examined to identify the variables that are statistically significant. In the case study, this matrix from the sample data showed that the correlations between the variables are satisfactory, and thus, all variables are kept for the next step. Meanwhile, preliminary tests, such as the Bartlett test, the Kaiser-Meyer-Olkin (KMO) test, and the Measures of Sampling Adequacy (MSA) test, are used to evaluate the overall significance of the correlation. Table 1 shows that the values of MSA (rounded to two decimal places) are higher than 0.5 for all variables but variable *W*. However, the MSA value of variable *W* is close to 0.5 (MSA should be higher than 0.5, according to Hair, et al. (1998)).

Step 2: Finding the Number of Factors

There are many approaches available for this purpose (e.g., common factor analysis, parallel analysis). The case study first employed the plot of eigenvalues vs. the factor number (the number of factors may be 1 to 13) and found that choosing three factors accounts for 91.3% of the total variance. Then, it suggested that the solution be checked

Table 1. Measures of the adequacy of FA to the sample: MSA

E	*I*	*ILK*	*J*	*M*	*M_S*	*MBD*	*MLK*	*P*	*PF*	*S*
0.75	0.73	0.86	0.78	0.76	0.82	0.79	0.74	0.74	0.76	0.85
W	*WG*									
0.39	0.87									

3 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/factor-analysis-data-mining/10648

Related Content

Material Acquisitions Using Discovery Informatics Approach
Chien-Hsing Wuand Tzai-Zang Lee (2005). *Encyclopedia of Data Warehousing and Mining (pp. 705-709).*
www.irma-international.org/chapter/material-acquisitions-using-discovery-informatics/10688

Privacy-Preserving Data Mining and the Need for Confluence of Research and Practice
Lixin Fu, Hamid Nematiand Fereidoon Sadri (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 2402-2420).*
www.irma-international.org/chapter/privacy-preserving-data-mining-need/7770

Expanding Data Mining Power with System Dynamics
Edilberto Casado (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 2688-2696).*
www.irma-international.org/chapter/expanding-data-mining-power-system/7792

Web Usage Mining Data Preparation
Bamshad Mobasher (2005). *Encyclopedia of Data Warehousing and Mining (pp. 1226-1230).*
www.irma-international.org/chapter/web-usage-mining-data-preparation/10785

Facilitating and Improving the Use of Web Services with Data Mining
Richi Nayak (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 2022-2035).*
www.irma-international.org/chapter/facilitating-improving-use-web-services/7746