STATISTICAL TECHNIQUES USED IN DECISION MAKING PROCESS IN VIRTUAL ORGANIZATION

Simona Ghiță, PhD
Emilia Titan, PhD
Vergil Voineagu, PhD
The Bucharest Academy of Economic Studies

Abstract

Actual economical tendencies, such as an obvious evolution towards market-globalization, reducing the production-cycle length, a clearer client-oriented policy imply structure and functional-mode updating of the enterprises according to new conditions. Virtual organizations, as temporary networks of inter-companies cooperation, geographical distributed, commonly participate, with their abilities to use market opportunities. They represent a flexible and efficient solution for optimizing the activity in the new economy.

Data mining is an automat tool based on statistical techniques, which allows transforming data into knowledge and using them in identifying models and non-obvious relations between the information stoked in databases.

In this paper the authors will present statistical fundamentals of data-mining tool for predicting the future behavior and the functioning of virtual organizations.

Keywords: virtual organizations, data mining, statistical techniques.

1. Introduction

Collaborative networks are becoming more important in global and regional business, thanks to their ability to combine organisational competences. But as individual companies seek efficiency gains by focusing on their core competences while outsourcing non-core operations, the degree of inter-firm transactions grows considerably. This makes it imperative to manage network relations well, which in turn calls for the development and deployment of decision support models that assist companies in the management of these relations [6].

A virtual organization is "a temporary network of companies that come together quickly to exploit fast changing opportunities. . . with each partner contributing what it's best at" [1].

As the development of Internet and Intranet, there are virtual organizations in different domains such as commercial networks consisting of several retailers, a health e-business having several hospitals, a web-based education system comprising several schools and so on. Two main features of such temporary collaborations are:

– Each partner stores detailed data locally. Only summary or high level information is allowed to be shared for security reasons.

– Different partners are likely to have significant contextual and implementation differences. For examples, different retailers of a commercial virtual organization adopt different business policies, sell different products and have different price standards.

To successfully discover knowledge, data mining in virtual organizations should consider both of the above two features. The first feature requires that data mining in virtual organizations has to be done in a distributed environment.The second feature implies that the distributed data in virtual organizations can not be regarded as identically and independently distributed because there exist context heterogeneity across different parters.

Data Mining (DM) or Knowledge Discovery in Databases (KDD) [5] is an interdisciplinary field with major impact in the scientific and commercial environments. Data Mining is the iterative and interactive process of discovering valid, novel, useful, and understandable patterns or models in massive databases. Data Mining means searching for valuable information in large volumes of data, using exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. The major data mining tasks [2] are prediction and description. Prediction methods use some variables to predict unknown or future values of other variables: these include classification, regression, and deviation detection. The description methods find human-interpretable patterns that describe the data: these include clustering, association rules discovery and sequential pattern discovery. KDD consists of an iterative sequence of the following steps: data selection, data cleaning, data transformation, pattern generation, validation and visualisation.

2. Models of Distributed Data in Virtual Organizations

In a virtual organization, distributed data usually do not strictly have a hierarchical structure because there are no variables describing context of partners, although there exists context heterogeneity. In practice, it is difficult to determine the exact sources of the context heterogeneities. Although a distinction is made between homogeneous and heterogeneous data, it is limited on whether database schemata being similar or not, and assuming that the distributed data sets are identically and independently distributed.

Distributed data mining (DDM) is an active research sub-area of data mining, usually considering as algorithms for regression or classification. It accepts that data may be inherently distributed among different loosely coupled sites that are connected by a network. Through transmitting high level or summary information, DDM techniques can discover new knowledge from dispersed data. Such high level or summary information not only has reduced storage and bandwidth requirements, but also maintains the privacy of individual records.

Hierarchical models (multilevel models, HM) have been well studied both in theory and in practice [3, 4]. They were initially developed for analyzing hierarchically structured data.

Individuals can be in various types of groups. There are variables describing individuals, as well as variables describing groups. Through performing regression analysis at more than one level (e.g. both individual level and group level) simultaneously, HM involves a statistical integration of the different models specified at the level of interest. This kind of statistical integration reflects the fact that homogeneity and heterogeneity coexist across different groups. All traditional HM algorithms are iterative based (e.g. Iterative Generalized Least Squared, IGLS) and for centralized data. In the field of DDM, HM were used to discover multi-level association rules and exceptions from distributed hierarchical data. HM's application was extended to express uncertainty at several levels of aggregation. However, all the work is for centralized data.

Obviously, DDM techniques are suitable for the first feature of virtual organizations. However, the implicit assumption of identically and independently distributed data contradicts the second feature. For hierarchically structured data, HM techniques can well deal with context heterogeneity of higher levels (e.g. group level). Although distributed data in virtual organizations are not hierarchically structured in a strict sense, the idea of using HM to express uncertainty at different levels is suitable for the second feature of virtual organizations.

In statistical analysis, a popular way to model unobservable or immeasurable context heterogeneity is to assume that the heterogeneity across different sites is random. In other words, context heterogeneity derives from essentially random differences among partners in the virtual organization whose sources cannot be identified or are unobservable. Given the context, data of a certain partner is described to be identically and independently distributed (IID). So distributed data across various partners having randomly distributed context heterogeneity are conditional IID [12]. This leads to a two-level hierarchical model which describes context heterogeneity with mixture models and employs latent variables in a hierarchical structure.

Suppose a virtual organization consists of *K* partners. The data set stored at the *k*th partner consists of data $f(y_{ik}, x_{ik})$, $i = 1...N_k$ where the y_s are numerical response for regression problems, N_k is the sample size of the data, and $k = 1...K$. Assuming that the context of different partners is distributed normally with mean μ and variance δ^2, data y_{ik} of the *k*th partner has normal distribution with mean μ_k and variance σ^2, then the two-level hierarchical model of distributed data in the virtual organization is:

$$\begin{cases} \text{between partener level}: \mu_k \sim N(\mu, \delta^2) \\ \text{within a partener level}: \ y_{ik}/\mu_k \sim N(\mu_k, \sigma^2) \end{cases} \tag{1}$$

If d_k is the random sampling error of context across different partners, e_{ik} is the random sampling error of the *k*th partner, $f(x_{ik})$ is the real global regression model, and the different level residuals d_k and e_{ik} are independent, then (1) can be rewritten as:

$$\begin{aligned} &y_{ik} = \mu + e_{ik} + d_k = f(x_{ik}) + e_{ik} + d_k, \\ &e_{ik} \sim N(0, \sigma^2), \ d_k \sim N(0, \delta^2) \end{aligned} \tag{2}$$

When $\delta^2 = 0$ or $d_k = 0$, (1) and (2) can be respectively rewritten as:

$$y_{ik} \sim N(\mu, \sigma^2) \tag{3}$$

and

$$y_{ik} = \mu + e_{ik} = f(x_{ik}) + e_{ik}, \ e_{ik} \sim N(0, \sigma^2) \tag{4}$$

Equation (3) and (4) are the standard models that traditional DDM techniques to model distributed data. It means that all the partners in the virtual organization are homogeneous, and the differences among them are just random sampling errors. So a distributed scenario with homogeneous probability distribution is only a special case of the generic situation.

3. Distributed Linear Hierarchical Modeling (DLHM)

Once distributed data in a virtual organization are formulated as (2), the main task of data mining is how to fit it in the distributed environment.

When the underlying global regression model $f(x_{ik})$ is linear, if β is the fixed effects matrix (matrix of regression parameters), (2) can be rewritten as:

$$\begin{aligned} &y_{ik} = x_{ik}\beta + e_{ik} + d_k \\ &e_{ik} \sim N(0, \sigma^2), \ d_k \sim N(0, \delta^2) \end{aligned} \tag{5}$$

The most popular linear HM algorithm to solve (5) is based on *maximum likelihood via iterative generalized least squares* (IGLS)[1]. Assuming $X = \{x_{ik}\}, Y = \{y_{ik}\}$, V is the variance matrix, β^* is the random effects matrix and Z^* is the design matrix for random effects parameters, then:

$$\beta = \left(X^T V^{-1} X\right)^{-1}\left(X^T V^{-1} Y\right) \tag{6}$$

$$\beta^* = \begin{bmatrix} \delta^2 \\ \sigma^2 \end{bmatrix} = \left(Z^{*T} V^{*-1} Z^*\right)^{-1}\left(Z^{*T} V^{*-1} Y^*\right) \tag{7}$$

where $V^{*-1} = V^{-1} \otimes V^{-1}$ and $Y^* = \left[(Y - X\beta)(Y - X\beta)^T\right]$.

Starting with an initial estimate of β from ordinary least squares, IGLS iterates between (6) and (7) to convergence. However this algorithm is only suitable for centralized data. So we need to extend it to fit our distributed environment.

Since V is block diagonal, (6) and (7) can be respectively rewritten as:

$$\begin{cases} \beta = \left(X^T V^{-1} X\right)^{-1}\left(X^T V^{-1} Y\right) \\ X^T V^{-1} X = \sum_{k=1}^{K} X_k^T V_k^{-1} X_k \\ X^T V^{-1} Y = \sum_{k=1}^{K} X_k^T V_k^{-1} Y_k \end{cases} \tag{8}$$

$$\begin{cases} \beta^* = \left(Z^{*T} V^{*-1} Z^*\right)^{-1}\left(Z^{*T} V^{*-1} Y^*\right) \\ Z^{*T} V^{*-1} Z^* = \sum_{k=1}^{K} Z_k^{*T} V_k^{*-1} Z_k^* \\ Z^{*T} V^{*-1} Y^* = \sum_{k=1}^{K} Z_k^{*T} V_k^{*-1} Y_k^* \end{cases} \tag{9}$$

Where $X_k = \{f(x_{ik})/k\}$ and similarly for the other terms. Thus each partner contributes its component to the total matrix. In this case, X, Y do not need to be transformed.

Concluding, modelling the distributed data in a virtual organization with a two-level hierarchical model can be realised using the existing DDM approaches (DLHM) and statistical multi-level models to mine data.

References

1. Byrne, J.A.: The virtual corporation. Business Week (1993)
2. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R.Advances in knowledge discovery and data mining. Menlo Park, Calif: AAAI Press; 1996.
3. Goldstein, H.: Multilevel Statistical Models. Second edition edn. ARNOLD (1995)
4. Kreft, I., Leeuw, J.D.: Introducing Multilevel Modeling. Sage Publications (1998)
5. Piatestsky-Shapiro G. Knowledge discovery in databases. AAI/MIT Press; 1991.
6. R. Gulati, N. Nohria, and A. Zaheer, "Strategic networks," Strategic Management Journal, vol. 21, pp. 203–215, 2000
7. Yan Xing, Michael G. Madden, Jim Duggan, Gerard J. Lyons, Context-based Distributed Regression in Virtual Organizations, Technical Report No. : NUIG-IT-160503

[1] Goldstein, H.: Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73 (1986) 43–56