# Electronic Commerce Based on Self-Organizing Data Mining Customer Churn Prediction Model

**Ai-hua Ren, Wei-wei Zhao**

Department of Insurance, Hebei Finance University, Baoding, China

**Abstract**

In order to solve the high dimensional and nonlinear problems of churn prediction of E-business customers, this paper proposes a novel model for churn prediction of E-business customer based on Self-Organized Data Mining ( SODM ) . In this model, Objective System Analysis ( OSA ) and improved Group Method of Data Handling ( GMDH ) , two important SODM algorithms, are integrated for the churn prediction of E-business customer . At first, the critical attributes of E-business customer chum are chosen with OSA and then the training samples are sent to improved GMDH for studying and training anthe status of customer chum of testing sample is identified. The approach has been applied to the empirical analysis on the prediction of E-customer chum, which proves that compared with some common approaches, this integrated model based on SODM is an efficient and practical tool for the prediction of business chum and provides E-business enterprises with a new forecasting tool in customer relationship management.

**Keywords**: Customer churn prediction; Self－organize data mining ( SODM ); Objective system analysis ( OSA ); Group method of data handling ( GMDH ); E－Business

## 1. Introduction

As a heuristic automatic modeling technique, self-organizing data mining (SODM) is a class of multivariate analysis of complex systems modeling and identification methods. SODM according to guidelines and terminating outside the law to find the optimal complexity of the model, automate the modeling process, an effective solution "over learning" issue, with good generalization performance and higher prediction accuracy. In recent years, SODM in engineering, scientific, economic and other fields have been widely used, but the existing literature rarely applied to customer churn prediction, therefore, this paper attempts to self-organize several dug excavation in the objective system analysis algorithm (OSA) and data packet processing network (Group Method of Data Handing, GMDH) to introduce e-commerce customer churn prediction, a more effective solution to the e-commerce customer churn prediction exist in high-dimensional, nonlinear problems, thereby improving the accuracy of customer churn prediction.

 

## 2. Based on Customer Churn Prediction Model SODM

Churn refers to the suspension of corporate customers continue to buy the original enterprise goods or business services, turn to accept the competitor's goods or services. An important task is the enterprise to identify which customers are likely to be lost, and then take appropriate measures to minimize the loss, to achieve maximum benefits. Self-organizing data mining technology can assist in this task. This article will self-organizing data mining algorithms and improved GMDH network OSA combine to build e-commerce customer churn prediction model, the basic idea is: the OSA GMDH algorithm as the improvement of the network prefix system through the OSA algorithm reduces customer churn prediction system the number of attributes, thereby reducing system complexity GMDH network, but also reduces GMDH network training time. GMDH network using the information as a rear recognition system, fault-tolerant and anti-jamming capability. Based on self-organizing data mining e-commerce customer churn prediction model specific steps shown in Figure 1.

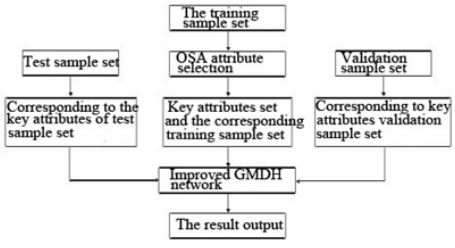


Fig.1. OSA algorithms and GMDH network integration model diagram

Since customer data is usually noisy mass data, to accelerate operational efficiency, customer data must first be analyzed to identify closely related with customer churn key attributes, in this based on customer churn prediction model. However, in the analysis of complex systems and modeling, they tend to use the data sample contains many variables and these variables generally there is a very complex relationship, it is difficult to get directly through qualitative analysis convincing conclusion, especially when researchers lack of complex systems necessary prior information. From the modeling point of view, the first problem to be solved is what the variable is the dependent variable, which is the independent variable, and which are the object and being independent study of economic variables.

Self-organizing data mining algorithms in an objective system analysis method (OSA) is one kind that can be objectively solve the basic problems of the algorithm. OSA is called "discovery rule " algorithm, able to establish equations model the way there from a number of interrelated factors ( variables ) to find the most essential characteristic variables and found a causal relationship between them. OSA was not available in the professional scholars of human survival and reliable information on the relationship between environmental pollution cases, be used to Ukraine and Asia -speed sea pollution prediction proved its validity on the modeling of complex systems. But so far, no in customer churn prediction literature involved.

OSA is a basic principle of evolution mechanisms and evaluation criteria, and ultimately determine a system of equations to describe the system being studied. Equations contained in the structure and the number of equations is based on the consistency of the model obtained. Its mechanism is the evolution equations complexity continues to increase, while the basic evaluation criteria used outside the guidelines as minimum deviation criteria.

OSA algorithm based on the basic principles, this paper build e-commerce

customer data key attribute selection algorithms:

The first step: the customer in mind customer data attribute number is *m +1*, data length to n. The state of the target variable churn denoted *x0*. For the quantification of qualitative data. With Boolean logic values 0 and 1 to indicate churn state, where 0 indicates that the client does not drain, 1 churn. To eliminate the influence of dimension, each of the variable data be standardized. Standardized formula is:

$x_{ij}' = \frac{x_{ij} - \overline{x_i}}{\sigma_i}$, *i=1,2,...,m+1; j=1,2,...,n+1.*

Step two: the customer data sets *W* into two sets of samples of the same size *A* and *B*, such that $W = A \cup B$. If the customer data length is not even, it can be discarded or any one customer data duplication join A, B two data sets. Since customer data is massive data, this will not affect the modeling conclusions.

*A* and *B* data record length is *N*.

Note *P ={ 1,2,..., m }, Q = {1,2,..., N };*

So *h = 1*.

The third step:

1) The i-th variable, with the least squares method on the sample Collection W parameter estimation, too

$$x_i = a_0 + b_{0x}, i \in P$$

On the data sets A and B respectively, the least squares method is used to estimate parameters:

$$x_i^A = a_0^A + b^A x_0, x_i^B = a_0^B + b^B x_0, i \in P$$

2) Calculate the minimum deviation criterion value:

$$\eta_{1i} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_i^A(k) - x_i^B(K)}{x_i(k)}\right), i \in P \quad (1)$$

Formula (1), $x_i(k)$, $x_i^A(k)$ and $x_i^B(K)$ denote the set of the data *W*, the data sets *A* and *B*, the data sets obtained by fitting parameter estimates.

if $\eta_1 = \min(\eta_{1i})$,

The fourth step:

*h=h+1*

1) That represents the customer attributes in m variables, take any type of attribute variables of *h xi, xj,..., xr (i, j,..., r ∈ P)*, the least square method in the sample Works W on parameter estimation, get h element equations:

$$\begin{cases} x_i = a_{01} + a_{11}x_j + \ldots + a_{(k-1)1}x_r + b_1 x_0 \\ x_j = a_{02} + a_{12}x_i + \ldots + a_{(k-1)2}x_r + b_2 x_0 \\ \vdots \\ x_r = a_{0k} + a_{1k}x_i + \ldots + a_{(k-1)k}x_r + b_k x_0 \end{cases}$$

*xi, xj,..., xr (i, j,..., r ∈ P)*, using least squares parameter estimation on the set of A and B respectively, get :

$$\begin{cases} x^A{}_i = a_{01}{}^A + a_{11}{}^A x_j + \ldots + a_{(k-1)1}{}^A x_r + b_1{}^A x_0 \\ x_j{}^A = a_{02}{}^A + a_{12}{}^A x_i + \ldots + a_{(k-1)2}{}^A x_r + b_2{}^A x_0 \\ \vdots \\ x_r{}^A = a_{0k}{}^A + a_{1k}{}^A x_i + \ldots + a_{(k-1)k}{}^A x_r + b_k{}^A x_0 \end{cases}$$

$$\begin{cases} x^B{}_i = a_{01}{}^B + a_{11}{}^B x_j + \ldots + a_{(k-1)1}{}^B x_r + b_1{}^B x_0 \\ x_j{}^B = a_{02}{}^B + a_{12}{}^B x_i + \ldots + a_{(k-1)2}{}^B x_r + b_2{}^B x_0 \\ \vdots \\ x_r{}^B = a_{0k}{}^B + a_{1k}{}^B x_i + \ldots + a_{(k-1)k}{}^B x_r + b_k{}^B x_0 \end{cases}$$

2) For each equations calculating the minimum deviation criterion value:

$$\eta_{i,j,\ldots,r} = \frac{1}{h}(\eta_{ki} + \eta_{kj} + \ldots + \eta_{kr})$$

Among them, $\eta_{ki}, \eta_{kj}, \ldots, \eta_{kr}$ all got by (1).

$\eta_h = \min(\eta_{i,j,\ldots,r})$

The fifth step:

Compare with $\eta_h$ and $\eta_{h-1}$, if $\eta_h < \eta_{h-1}$, return to step 4. Otherwise stop algorithm, record system minimum deviation criterion value best $\eta = \eta_{h-1}$.

Minimum deviation criterion value of $\eta_{h-1}$ when the corresponding variables to the characteristics of the system of equations. These variables corresponding customer attributes, namely key attributes for the customer.

## 3. Conclusion

We build self- organizing data mining based on e-commerce customer churn prediction model, and an online shopping mall using the actual sample data empirical results show that :

(1) OSA algorithm using the churn attribute variables are selected without loss of information obtained under the impact of electronic commerce on the premise lost four key attributes of customer repeat purchase frequency, the number of purchases during the day and night, the number of purchases, the number of purchases late at night, making the GMDH network input significantly reduce the number of data, simplifying the network structure of the network system to improve the speed and efficiency of learning prediction.

(2) The GMDH network as a post-information processing system to improve the model of fault tolerance and noise immunity. GMDH network by using improved after reduction of the sample data for training and testing samples for identification, the results show that the proposed e-commerce customer churn prediction model is not only effective but also efficient. With OSA GMDH, OSA BP neural network and a simple improvement GMDH, GMDH model is compared, OSA improved GMDH network integration model of non- discrimination churn prediction lose customers

Accuracy, loss of customers and overall prediction accuracy of prediction accuracy has dramatically increased. The model can accurately explore e-commerce businesses churn the real situation, to carry out a comprehensive and effective customer relationship management to provide better decision support, the model in China has a broad application prospects.

Of course, this model also has some disadvantages, such as the loss of customers misjudge larger number of non- loss of customers, allowing companies to provide certain non- losing customers when special services pay bigger. How to further improve the model of e-commerce customer churn prediction ability, will likely be one of the important follow-up study.

## 4. References


[1] Reichheld Frederick F. Thomas Teal. The Loyalty Effect: The Hidden Force behind Growth and Lasting Value [M]. Boston. MA: Business Scholl Press, 1996.

[2] Reichheld F. Learning from Customer Defections [J].Harvard Business Review, 1999, ( 2 ) : 56-61.

[3] Baesens, Bart, Verstraeten et al. Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers [J]. European Journal of Operational Research ( 2004 ),156 ( 2 ).

[4] Sum Kim, Kyung-Shik Shin, Kyungdo Park. An application of support vector machines for customer churn analysis: credit card case [C]. Advances in Natural Computation. First International Conference, ICNC 2005.

[5] A G Ivakhnenko. Heuristic Self-organization in Problems of Engineering Cybernetics [J]. Avtomatika,1970, ( 2 ).