Analysis on E-consumers’ Purchasing Behavior Based on Data-driving Model

Lijuan Huang
Information Management College of Jiangxi University of Finance and Economics, Nanchang, 330013, China
Email: huanglijuan66s@126.com

Abstract—It is the Internet world with vastly purchasing data sea online that makes research model of e-consumers’ purchasing behavior very different from traditional ones. Firstly this paper proposes three kinds of research models of consumers’ purchasing behavior, and then pointed out that data-driving model is the best one to analyze e-consumers’ purchasing behavior on the Internet. Secondly, it adopts the improved SOFM Neural Network as the tool of data-driving model to detailedly analyze e-consumers’ purchasing behavior of Internet marketing. Lastly experiment results demonstrate that the method has more visualization, exactness and robustness. Because consumers’ purchasing behavior analysis based on the SOFM Neural Network is a comparatively novel method, the research fruit in this paper is just for reference.

Index Terms—Internet marketing, purchasing behavior, neural network, data-driving model

I. INTRODUCTION

Research about consumers’ purchasing behavior characteristics dates back to England in eighteenth century. At that time, large number of farmers poured into cities. These new urban residents show faith in the products which were able to demonstrate their social status, and the faith and attitude for these products from the residents brought people’s attention focused on consumer behavior[1]. The research about consumer behavior originated and developed from a western paper named Consumer Analysis published by Guest in Annual Review of Psychology in 1962 [2]. Afterwards, many celebrated scholar did active work on characteristics of consumer behavior. For example, Engel, Kotler and Cliff Allen proposed T-I-K model of consumer behavior in 1993, Solomon, Schiffman and Kanuk raised U-S-E model of consumer behavior in 1999, J. Paul Peter and Jerry C. Olsom presented S-C-T model of consumer behavior in 2000 [3-6]. But these researches were attributed to one of experience-driving research model or theory-driving research model. The author believes that research model of consumer behavior should include data-driving model besides experience-driving research model and theory-driving research model, with the development of modern science and technology, and especially with development of neural network technology, data mining, artificial intelligence, and multi-disciplinary technology. These three kinds of research models are described in Table I.

<table>
<thead>
<tr>
<th>Table I.</th>
<th>RESEARCH MODEL OF CONSUMERS’ PURCHASING BEHAVIOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1: Experience-driving model</td>
<td></td>
</tr>
<tr>
<td>Researcher can communicate with consumers by means of tongue, facial expression and other body language, and then make an analysis of consumers’ purchasing behavior based on the researcher’s own experience. However, in the virtual world of the Internet, there is large sum of data about e-consumers’ purchasing behavior and the researcher lose the chance face to face to communicate with consumers, So analysis of consumers’ purchasing behavior based on experience-driving model loses effect.</td>
<td></td>
</tr>
</tbody>
</table>

Method 2: Theory-driving model

The research steps of theory-driving model are shown in Fig 1. From Fig 1, we can know, in this kind of research mode, researcher first obtains a theory model from purchasing behavior theories; Then makes full use of purchasing data to test and modify the model repeatedly; Finally, based on the last model to deduct and analyze the consumers’ purchasing behavior. This kind of research mode usually can get an unreliable analysis result due to the imperfect and even wrong purchasing behavior theories.

Method 3: Data-driving model

The research steps of data-driving model are shown in Fig 2. From this figure, we can know, in this kind of research mode, researcher first select appropriated intelligent algorithm; Then a model is drawn from purchasing data and is modified repeatedly by these purchasing data; Finally, based on the last model to deduct and analyze the consumers’ purchasing behavior. Obviously, data-driving model is based on real data other than personal experience or pure theories and this kind of model realizes the scientific idea that Let data say for themselves. So, the result of analyzing consumers’ purchasing behavior is more scientific, objective and fair.

Table I shows that it is difficult to adopt experience-driving model to analyze characteristics of online consumer purchase behavior, and adopting theory-driving model or data-driving model may be appropriate. Seen in Table I, Fig. 1 and Fig. 2, it is more objective, scientific and unbiased to adopt data-driving model than to adopt experience-driving model or theory-driving model for analyzing e-consumers’ purchasing behavior.
Therefore, data-driving model is the most suitable for analyzing characteristics of online Consumers’ purchasing behavior, and all input data is from the consumers, it also fully reflected the idea: “Consumer is the God”. Because Self-Organizing Feature Map Neural Network (SOFM NN) belongs to a typical data-driving mode, this paper takes SOFM NN as a tool to analyze e-consumers’ purchasing behavior. The basic principles of SOFM NN are described as follows.

II. BASIC PRINCIPLES OF THE SOFM NEURAL NETWORK

In 1981, Finnish scholar Teuvo Kohonen firstly raised the concept of SOFM NN[7], which can simulate the function of the brain that reflects to different kinds of input signals (e.g. light signal, sound signal) and automatically sort these input signals into different zones of the brain layer[8]. Through inputting large sum of purchasing data of consumers into SOFM NN, these e-consumers can be objectively, scientifically, and automatically clustered and divided into different groups based on the similarity of consumers’ purchasing data, and this means minimizing difference between the consumers in the same group and maximizing the difference between different groups. Analyzing and aiming directly at the different feature of these different consumer groups, it would be helpful to make some aimed marketing strategies for promotion, service, price etc, avoid the risk of taking the uniform strategies for all the consumers and with high cost for not important consumers or taking the unscientific ranked service to lost the potential VIP consumers.

A. Topology Structure of the SOFM NN

The typical SOFM NN (seen in Fig. 3) forms topology structure of input signals based on one-dimension or two-dimension cellular array [8], so the SOFM NN has the ability to extract the feature of the input signals’ model[9]. The SOFM NN commonly only includes a one-dimensional or two-dimensional arrays, but could also be extended to handle the multi-dimensional cellular array [10-12]. In order to have better stability and operating efficiency of SOFM NN, we add a feedback loop on the traditional SOFM NN to obtain improved SOFM NN (seen in Fig. 4).

The improved SOFM NN is composed of the following four parts.

- Cellular array for recognizing: This is mainly used for receiving the input signals and forming the “discrimination function” to recognize the input signals.
- Mechanism for comparing and choosing: This is used for comparing these “discrimination functions” and making a decision to choose a processing unit with stronger functional output signals.
- Local inter-connection and inter-action: This is used for stimulating both the chosen signals processing unit and its nearby signals processing unit.
- Self-adapting process: This is used for modifying the parameters of stimulated processing unit so that it can increase the output value of the given “discrimination function”.

B. The SOFM NN’s Algorithm

The SOFM NN’s algorithm are described as follows.

1) Initialization: choose “nearby neuron” set $S_j(0)$ with output neurons $j$, and the connection weight value $w_{ij}(0)$ for both the input neuron $i$ and the output neuron $j$ is computed as equation(1).
2) Calculating the Euclidean distance: euclidean distance means the distance between the input sample and every output neuron \( j \). Calculating the Euclidean distance \( d_j(t) \) is shown in equation (2).

\[
d_j(t) = \ln(||X - w_j||) = \ln\left(\sqrt{\sum_{i=1}^{n}[x_i(t) - w_{ij}(t)]^2}\right)
\]  

(2)

3) Defining a neighborhood function: neighborhood function \( S_{ij}(t) \) is expressed in equation (3), where \( S_{ij}(t) \) gets decreased as the time goes on.

\[
S_{ij}(t) = S_{ij}(0) \exp\left(-\frac{d_{ij}(t)}{2\sigma^2}\right)
\]  

(3)

4) Working out the minimum distance: the minimum distance \( \min(d_j) \) among these corresponding neurons is calculated as equation (4).

\[
\min(d_j) = \arg\min \sum_{i=1}^{n}[x_i(t) - w_{ij}(t)]^2
\]  

(4)

5) Setting learning rate: learning rate \( \eta \) may be computed according to equation (6), where \( \eta \) gets decreased to zero as time \( t \) goes on.

\[
\eta(t) = \eta(0) \exp\left(-\frac{t}{\tau}\right)
\]  

(5)

6) Modifying the weight value: When the weights variation \( \Delta w_{ij}() \) reduces to zero, topology structure of the SOFM NN is most stable, and \( \Delta w_{ij}() \) is computed as equation (6).

\[
\Delta w_{ij}(t) = \begin{cases} \eta(t)[x_i(t) - w_{ij}(t)], X \in S(k) \\ 0, X \notin S(k) \end{cases}
\]  

(6)

7) Offering new learning samples to repeat the learning process mentioned above, then \( t \rightarrow t+1 \), till \( \eta(t) \) decreases to 0 or enough small, and process of network learning is terminted.

III. AN EXAMPLE OF ANALYZING E-CONSUMERS’ PURCHASING BEHAVIOR

Because selling book is one typical choice to do E-bussiness, this paper takes consumers of book bussiness website for example to analyze e-consumers’ purchasing behavior[13].

A. Main Clustering Variables

Most data of customers come from online dealing records of a famous book website (dingdang.com) in China[1]. These data could be divided into two groups: customers’ attributes data, and transaction data. Customers’ basic attributes data mainly include: customer’s name, gender, age, income, educational status, occupation, city, marriage status, enrolment time, home address, hobby etc. Transaction data mainly include: shopping time, frequency of shopping, consumption of shopping, product name, price, way of paying (e.g. cash on delivery, cash on postage and credit Card), latest shopping time etc.

Main clustering variables of the SOFM neural network are seen in Table II, where main variables labeled by (*) indicates to be clustering variables.

<table>
<thead>
<tr>
<th>Total amount of purchase</th>
<th>Monthly income</th>
<th>Frequency of shopping</th>
<th>Latest time of shopping</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1(*)</td>
<td>x2(*)</td>
<td>x3(*)</td>
<td>x4(*)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Educational status</th>
<th>District</th>
</tr>
</thead>
<tbody>
<tr>
<td>x5</td>
<td>x6</td>
<td>x7</td>
<td>X8</td>
</tr>
</tbody>
</table>

B. Sample Data of Consumers’ Behavior

There are 5000 sample records but limited by the length of this paper, we will only list part of the samples as demonstrated in Table III, where capitalized variables in Table III means to be standardized in the domain [0, 1].

<table>
<thead>
<tr>
<th>Cust-ID</th>
<th>Total amount of purchase (X1)</th>
<th>Monthly income (X2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>0.9260</td>
<td>0.9454</td>
</tr>
<tr>
<td>1002</td>
<td>0.7549</td>
<td>0.6950</td>
</tr>
<tr>
<td>1003</td>
<td>0.8118</td>
<td>0.8975</td>
</tr>
<tr>
<td>1004</td>
<td>0.7982</td>
<td>0.6825</td>
</tr>
<tr>
<td>1005</td>
<td>0.6532</td>
<td>0.5816</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cust-ID</th>
<th>Frequency of shopping (X3)</th>
<th>Latest shopping time (X4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>0.9720</td>
<td>0.9335</td>
</tr>
<tr>
<td>1002</td>
<td>0.7273</td>
<td>0.6918</td>
</tr>
<tr>
<td>1003</td>
<td>0.7586</td>
<td>0.7324</td>
</tr>
<tr>
<td>1004</td>
<td>0.8180</td>
<td>0.7817</td>
</tr>
<tr>
<td>1005</td>
<td>0.6609</td>
<td>0.5141</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Through system function premnmx() or user-defined functions, sample data can be normalized in the domain [0, 1]. In this paper, we adopt the Min-Max standardize method shown in equation (7).

\[
X(i) = \frac{x(i) - \min{x(i)}}{\max{x(i)} - \min{x(i)}}
\]  

(7)
C. Design of the SOFM Neural Network

1) Topology Structure

There are three kinds of topology structures: rectangular topology structure, hexagonal topology structure and random topology structure, which can take the corresponding three functions (namely gridtop(), hextop() and randtop() ) to describe the different topology structure of these neuron areas [14]. Here we take the 6*4 random topology structure (shown in Fig. 5).

2) Main Programming Codes

We firstly use function newsom() to create a SOFM neural network; then we use function train() and function sim() to train and simulate the new created network in order. Different training steps have different effects over efficiency of self-recognizing. Here, we set the training steps as 1000, 3000, 5000 and 10000 and observe the efficiencies of clustering respectively. The main programming codes are shown as follows:

```matlab
net=newsom(minmax(X),[6,4],'randtop');
a=[1000 3000 5000 10000];
yc=rands(1,10);  
for i=1:4
    net.trainParam.epochs=a(i);
    net=train(net,X);
    figure;
    w1=net.IW{1,1};  
    plotsom(w1,net.layers{1}.distances);
    y=sim(net,X);
    yc=vec2ind(y)
end
```

D. Analysis on the Result of Training and Computing

1) Network’s Weight Value Structure

There are great differences of SOFM neural network’s performance when we take different training steps. In the paper, we only take four kind of different training steps namely 1000, 3000, 5000 and 10000, and the corresponding Network’s weight value structure are shown in Fig. 6, Fig. 7, Fig. 8, and Fig. 9 respectively as follows.

From the above 4 figures, we can easily find that Network’s weight value figure comes to a comparatively stable status when the training steps is 5000 and 10000.
2) Network’s Clustering Result

Through training and simulating according to the four different kinds of training steps, we can also acquire a clustering result as shown in Table IV, where only 20 sample records are listed for demonstration.

<table>
<thead>
<tr>
<th>Training Steps</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>8</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>16</td>
<td>19</td>
<td>13</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>19</td>
<td>13</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>5000 10000</td>
<td>11</td>
<td>18</td>
<td>7</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Training Steps</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>15</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>7</td>
<td>19</td>
<td>13</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>19</td>
<td>16</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>5000 10000</td>
<td>13</td>
<td>19</td>
<td>7</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>18</td>
<td>11</td>
<td>13</td>
<td>19</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

To observe Table IV, we can find some rules as follows:
- When the training steps are 1000, all the samples are divided into 1 group.
- When the training steps are 3000, all the samples are divided into 2 groups.
- When the training steps are 5000, all the samples are divided into 3 groups.
- When the training steps are 10000, all the samples are divided into 3 groups.

From Fig. 10, we can also find that there is the unique minimum from a single neuron’s error surface, so the structure of the above improved SOFM NN is comparatively stable. This means Customers clustering stability are robust.

3) Customers’ Recognition and the Corresponding Marketing Strategies

According to the above 4 network’s weight value structure Figures (namely Fig.6-9) and one network’s clustering result table (namely Table IV), We can also reach a further conclusion: when the training steps are more than 5000 (including 5000), the samples are steadily clustered and divided into 3 groups. To observe these 3 groups and make an analysis of customers’ purchasing behavior, we find each group has its own special features as illustrated in Table V, where 3 distinguished marketing strategies are strongly suggested aiming at these 3 groups’ special features. Obviously, recognizing customers’ features and taking the distinguishing marketing strategies can help to reach a win-win result between customers and business website, increase the loyalty of customers (esp. VIPs), and maximize the profit of e-marketing.

<table>
<thead>
<tr>
<th>Cluster NO 1</th>
<th>Customers (5.71%)</th>
<th>Consumption (0.13%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features of customers’ purchasing behavior: occasional customers, most of the occasional customers are teenagers who come from different districts of the nation, and the total amount of purchase is low with low income and low shopping frequency. Most of them have a low-level educational status.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing strategy: These customers deserve the normal service, such as racking up points for discount, getting the book information through e-mail but reading e-books not free on the Internet.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster NO 2</th>
<th>Customers (74.71%)</th>
<th>Consumption (17.86%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features of customers’ purchasing behavior: main customers, most of the main customers are youths who come from different districts of the nation, and the total amount of purchase is higher with middle-level income and higher shopping frequency. They have a middle-level educational status.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2011 ACADEMY PUBLISHER
Marketing strategy: these customers deserve the middle-class service, such as ordering individualized information of book through e-mail, racking up points for higher discount, reading some e-books free on the Internet when the amount of purchase accumulate to a certain point, enjoying free e-cards or e-flowers on their birthday and so on.

Features of consumers’ purchasing behavior: most is VIPs, who are young women who often come from highly developed districts or remote places, and the total amount of purchase is the highest with high income or low income and the highest shopping frequency. Most of the VIPs have a middle-level or high-level educational status.

Marketing strategy: these customers deserve the top service, such as enjoying VIP service to have free private cyberspace and fastest green passage, downloading or reading some e-books free on the Internet, conferring the latest book catalogue in both paper’s form and e-mail’s form, free biggest cards and best flowers on their birthday, the highest discount and so on.

Table V strongly proves the Pareto 80/20 Principle: 20% of all customers are the VIPs (Cluster NO 3), and their contribution is 80%. In this table, we can also find some interesting phenomena. For example, VIPs would not definitely be customers with high income, and most of VIPs are young women rather than men, VIP customers are not only from developed regions, but also from less developed regions.

IV. CONCLUSION

Famous economist Christopher pointed out: in today’s unpredictable business competition, the market is no longer on the sellers’ side but on the buyers’ side [14]. “Customer is the god”. So exactly to analyze consumers’ purchasing behavior on the Internet and accordingly to make some scientific Internet marketing strategy for sale promotion are key factor to success for assuring the profit of E-business website. As for how to analyze e-consumers’ purchasing behavior, this paper proposes and compares three kinds of research models, and pointed out the data-driving model is best one to analyze e-consumers’ purchasing behavior. SOFM NN belongs to a typical data-driving model, so this paper improves the traditional SOFM NN and takes the improved one as a tool to analyze e-consumers’ purchasing behavior. Because e-consumers’ purchasing behavior analysis based on the SOFM Neural Network is a comparatively novel method, the result of research in this paper is just for reference.

ACKNOWLEDGMENT

The author thanks the anonymous reviewers for their valuable remarks and comments. This work is supported by 2010 National Social Science Fund of China (Grant No. 10BGL028), National Natural Science Fund of China (Grant No. 70861002), China Postdoctoral Science Fund (Grant No. 200902535), 2010 Science and Technology Project of education department of Jiangxi Province (Grant No. GJJ10430), and 2010 Social Science Planning Project of Jiangxi Province (Grant No. 10GL35).

REFERENCES