An alternative method for extracting unexpected patterns from huge attributes using conditional contingency table in marketing.

J.M Choi¹, Y. Asami²
¹Department of Urban Engineering, University of Tokyo, Japan.
²Center for Spatial Information Science, University of Tokyo, Japan.

Abstract

Marketing analysis is an indispensable step to establish strategies for the condominium business. In this context, it is crucial to grasp the potential customers' behavioral characteristics as “rule” or “pattern” extracted from historical data. In this paper, the earlier straightforward approach is outlined and the extension based on conditional contingency table is demonstrated as an alternative to extract unexpected rule patterns from a database with relatively few transactions but a huge number of attributes. The paper shows that this approach can better summarize information targeted to unexpected patterns than the earlier method using data mining methodology. For the validity of the proposed method, about 800 condominium purchasers in Tokyo metropolitan area are being analyzed and compared with the result of the earlier method.

1. Introduction

Most of major real estate companies in Japan have been collecting customer data and building massive data warehouses. Few companies however, have been able to realize the actual value stored in it. The question these companies are asking is how to extract this value. One of the possible answers for this could be data mining. Data mining technology has increasingly been fascinating marketers as it can provide valuable hidden business intelligence from historical data. The method can be soundly employed in the context of a need to better understand customers, market segmentation, and to quickly respond to their individual needs and wants. Nevertheless, the actual implications of rule mining in the housing market are very scarce.

A couple of reasons for this can be cited, one being that it is often very hard for a researcher to obtain detailed information about the private housing sector, and even after successful gathering of this information, but owing to business confidentiality there are usually limitations on opening them to the
The other reason is that the focus on so-called data mining is primarily oriented to discover rule patterns from a huge “transaction database”, which is somewhat different from our research database which has relatively few transactions but “massive number of attributes”. This background leads marketers or decision makers in this area to rely heavily on their long period of field experiences or intuitions.

These considerations motivated our study to develop a rather straightforward solution to extract and summarize some unexpected customers’ rules or behavioral patterns from the database with a huge amount of attribute information. In the earlier work [1] we have proposed an intuitive and novel method that extracts unexpected patterns in terms of contingency tables, and demonstrated that the resultant rules are very consistent and quite understandable. Moreover, as an extension to the earlier work, we attempt in this paper to alternatively extend the proposal using conditional contingency tables and make a comparison of the results by both methods.

2. Background and related literature

Most existing data mining techniques, such as association rule mining, are often applied to customers’ behavioral databases. In general however, it is said that conventional data mining techniques produce too many rules and that is a major problem. In addition, most of the mined rules are self-obvious or irrelevant. Especially, for marketers unexpected rules are more important than general rules [3]. One of the reasons is that managerial intuition often comes from a hint that is hidden under strong or general patterns and this sort of unexpected clue motivates marketers to develop a new business domain. On the other hand, unexpected rules are those of weak patterns outside the strong ones. Usually, such patterns are unknown, unexpected, or contradictory to what the user believes. From the opposite viewpoint, it can be said that the main idea of unexpected rule mining is to filter out those uninteresting rules or apparent strong patterns. Certainly the interestingness of a rule is arguable and indeed contingent on individual interest. It is normally clarified that unexpected patterns are more interesting because they contradict our expectations which depend on our system of beliefs [5].

In relation to unexpected rule mining, a good number of methods have been proposed to help the user find exception rules from a large database. Liu et al. [2] provide concrete information about the domain of unexpected rule mining techniques: (1) using some interestingness measures to filter out those uninteresting rules; and (2) using the user’s domain knowledge to help him/her identify unexpected rules.

3. The proposed method

3.1 A contingency table

We focus on the combination of two category variables or contingency tables in order to uncover the relationships among all variables in a dataset (see Figure 1). If we can extract efficiently all unexpected cells over all the possible combinations of contingency tables, then all information about unexpected rules is mined and the data structure is excellently summarized in the context of unexpected rule mining. There is an approach [4] entitled “deviation analysis” in order to mine this sort of unexpected rule pattern, in which the discovery of significant changes or deviations of some pre-defined measures are computed.
from its normative value in a dataset. In most applications, the measured normative value is expressed as the expected value (expectation $E$), or as a forecasted value calculated from applying some mathematical models [4]. In our work, a deviation is given by:

$$I = \log(A_i / E_i)$$

where $A_i$ is actual frequency of observation, and $E_i$ is the expected frequency of marginal product in multinominal distribution in a given cell $C$.

Note that eqn (1) provides a degree of deviation, but no information about the basis or measure for extracting unexpected cell(s). Therefore, one of the main issues in our methods is establishing a basis for extracting unexpected patterns which is outlined in section 3.2. If any deviation is detected, which exceeds a user-specified “minimum deviation threshold”, it suggests that an unexpected event has occurred. Note that any cell highly correlated with two variables in a contingency table is regarded as significant cell(s) and vice versa under the user specified deviation threshold. That is, there are two types of significant cells; a cell surpasses theoretical expectation and a cell fall below theoretical expectation.

On the other hand, all the possible combinations of variables could lead to a very large number of contingency tables that affect calculation time. Therefore, as an alternative to this problem in our research, we separate all the underlying variables into two sets of variables with respect to research framework. That is “target variable” and “non-target variable” as shown in Figure 1(a). In some circumstances like our research, it is more beneficial to designate a group of variables in terms of their characteristics. For instance, since in many real situations in marketing our concern is focused on the specific target variables, we pay attention to describe the characteristics of target variables with the relating attribute variables (non-target variables). This approach could be thought, for example, as deploying variables in designing the relationship between the response variable and the explanatory variable(s) in a regression analysis. This framework also gives an if–then structure, where “condition” is the non-target variable, “then” is the target variable and supplementary information is obtained from eqn (1). Thus, a set of target variables can be any set of attribute variables that the researcher pays attention to, and a typical example of target variables could be the response variables used for such as market segmentation.
As a result, the meaning of “the combinations of contingency table” refers to the set of combinations which consist of pairs of target variables and non-target variables. An example of a contingency table from the combinations of contingency tables is illustrated in Figure 1(b).

3.2 The earlier method for extracting significant fragmented rules

Since significant deviations from the norms are unexpected, they should be interesting to the user. Likewise, we have proposed an unexpected rule mining technique using contingency tables. Our earlier proposal [1] comprises two major parts in extracting unexpected rule patterns: extracting unexpected rules (or significant cells) and integrating significant cells using association rule mining method.

3.2.1 Extracting significant cells

The detailed process of extracting significant cells is as follows.

Step 1: Prepare an \( r \times c \) contingency table
The number of contingency tables to be made is the number of combinations between target variables and non-target variables, not all the possible combinations of underlying variables. A pair of category variables in a contingency table consists of a target variable (in row) and a non-target variable (in column) as exemplified in Figure 1(a).

Step 2: Regard distribution of a contingency table as a binomial distribution and then approximate it to a normal distribution
Under the assumption that two variables in a contingency table are independent, it is possible to regard the distribution of frequencies for each cell in the table as binomial distribution. This is because the task of counting observations in a contingency table can be cast into the form of a repeated sequence of binary trials comparing each observation in a cell with the case being counted. If the number of count is large enough, then the cumulative probability of a binomial distribution can be approximated into cumulative probability of a normal distribution in terms of normal function.

Step 3: Remove cells with scarce frequencies in the table
If the contingency table is sparse, i.e. with many zero entries (frequencies), this is problematic. Moreover, from the practical aspect, given any cell with zero or close to zero frequencies, which frequently occur in a real situation, all these cells could fall into significant cells on the basis of the proposed measure. For this, we propose the least frequency in a cell by calculating the theoretical number when the frequency on binomial distribution is zero under a user-defined probability and total count. Then we can exclude each cell that is under the least frequency in a table.

Step 4: Extract significant cells only at a given significance level
It is required to determine whether a given cell is suitable to be selected as a significant cell or not. To this end, we compare all cells over a contingency table using statistical significance level on normal distribution \( N(0, 1) \). In this way, significant cells can be selectively extracted by picking up the cells in which frequencies are out of the theoretical frequencies calculated in Step 2 under a given statistical confidence level, excluding the cells which are the case of scarce
3.2.2 Integrating significant cells using association rule method

By arranging the extracted significant cells according to the target variables, some of the latent information can be uncovered outside one's domain knowledge. This is because the proposed method also gives an if-then rule structure as mentioned in section 3.1. However, it should be noted that the extracted significant cells are only a collection of pairs in contingency tables or fragmented information that is regarded as significant from the statistical framework. Therefore, it is necessary to synthesize the significant cells systematically for consistency and compact interpretation of relations among variables. We pay attention to the formation of significant cells sorted by target variables. Because if we regard a set of significant cells sorted by the levels in a given target variable, then it can be inferred that in a cluster of frequent significant cells there is close relation among the pairs of the cells. In fact, this concept is none other than association rule mining, which searches for frequent pairs among items in a given data set.

3.3 Conditional contingency table

In the earlier work as outlined in section 3.2, the formation of a contingency table is composed of a pair of two-dimensional category variables, where one is designated as a target variable and the other as a non-target variable. In this paper, we attempt to consider the formation of contingency table on the basis of three-dimensional conditional contingency table. Conditional probability has recently been gaining greater interest in the data mining community in the context of Bayesian discovery methods. Moreover, with the conditional contingency table we expect to obtain more abundant unexpected rule patterns at the expense of a slight limitation of levels in conditional variables or the number of records that is needed to fill employment of our algorithm.

The two-dimensional contingency tables in section 3.2 can easily be extended to three-dimensional contingency tables (see Figure 1(c)). A conditional contingency table,

\[ CT(I_i, J_j | K_k = l), \quad (i \in I, j \in J, k \in K) \]  

(2)

is the contingency table for the subset of records in the dataset that satisfy the condition to the right of the \(|\) symbol. In eqn (2), \( I, J, K \) is the superset of each attribute variable \( i, j, k \) and \( l \) is the level of a condition variable \( k \) \((l \in L_k)\). As we want to extract all significant cells over all possible combinations of contingency tables using the same method illustrated in section 3.2.1, the number of contingency tables which should be executed exhaustively is given as eqn (3).

\[ \sum_{l=1}^{L_L} \sum_{j=1}^{L_J} \sum_{k=1}^{L_K} \sum_{l=1}^{L_L} CT(I_i, J_j | K_k = l) \]  

(3)

Note that in the case of the conditional contingency table, since the records in an unconditional contingency table are divided into each subset of records by the condition, the number of records decreases in proportion to the ratio of condition to total records. This implies that the possibility of zero entries
in a table will increase. In addition, it is necessary to have a large enough number of records to apply the algorithm shown as in section 3.2.1. As a result, a measure should be taken into consideration to cope with the problem. Alternatively there could be two options; 1) To increase the number of records 2) To limit the number of levels in a conditional variable lest frequencies should fall into the sparse table.

With respect to summarizing the extracted significant cells, in this paper we have adopted a different approach unlike the association rule technique used in the earlier work outlined in section 3.2.2. We have paid attention to the relationship between a pair of variables and a given condition variable, where both variables have at least one of the significant cells at the same time. That is, we link a pair of variables to each level of conditional variable in a given significant cell in the context of response-explanatory framework. This approach provides promising information about the relationship between a pair of variables and condition variable at that time.

4. Application of the proposed method and discussions

4.1 Data and framework of conditional contingency table

We tested our algorithm on the condominium customer database with 798 household records in five wards of the Tokyo metropolitan area in Japan. The selected variables are categorized into three main sets of variables as shown in Table 1. In this paper, the design of contingency tables demonstrated in section 3.3 is set to $CT(b,c \mid a)$ where $b, c, a$ are variables indicating “behavior, condominium, and attribute” respectively shown as Table 1. The background of this framework is that we would like to know the relationship between the behavioral characteristics by “behavior variables” and the characteristics of actually purchased properties by “condominium variables” in terms of the conditional layer or “attribute variables” in order to uncover the potential customers’ behavioral characteristics.

Table 1. Selected variables

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Behavior</th>
<th>Condominium</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1. Type of address change</td>
<td>B1. Motivation of purchasing</td>
<td>C1. Price</td>
</tr>
<tr>
<td>A2. Experience of residence</td>
<td>B2. Important factor of purchase</td>
<td>C2. Time to walk from nearest</td>
</tr>
<tr>
<td>A3. Number of tenant</td>
<td>B3. Pattern of gathering</td>
<td>railway station</td>
</tr>
<tr>
<td>A5. Means of transportation</td>
<td>B5. Media environment to access</td>
<td>C4. Unit type</td>
</tr>
<tr>
<td>of HH</td>
<td>B6. Customer satisfaction</td>
<td>C5. Ownership space</td>
</tr>
<tr>
<td>A7. Sex of HH</td>
<td>B8. Resigned factor of purchase</td>
<td>C7. Total number of floors</td>
</tr>
<tr>
<td>A8. Type of employment for HH</td>
<td></td>
<td>C8. Total number of houses</td>
</tr>
<tr>
<td>A9. Occupation of HH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A10. Category of business for HH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A11. Annual income of HH</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HH: Head of a household
4.2 Results and discussion

To begin with we compared the ratio of extracted significant cells to the total possible number of cells. In the case of the unconditional contingency table the ratio is 13.15%, whereas in the case of the conditional contingency table it is 4.4% (2,138 significant cells / 47,730 total number of cells). In addition, we compared the distribution of frequency of significant cells by each level of “attribute variable”. The result shows that the most frequent significant conditional attribute and level is Gender of household (A7): “male”, Experience of residence purchase (A2): “for the first time”, followed by Annual income of household (A11): “over 15 million Yen”. This result implies that Gender of household (A7): “male” is the most dominant factor which affects the extracting of unexpectedness in this dataset.

We attempted to compare both results of unexpected patterns by the unconditional contingency table method used in the earlier work and the conditional contingency table method cited in this paper. Due to limitations of space, we omit detailed results of comparison. Instead, a sample of the comparison is illustrated in Table 2, where we can notice that the conditional contingency table method provides more detailed information as well as the key unexpected findings. Comparing the consistency and abundance of unexpected rules, we believe that the alternative using the conditional contingency table is superior to the earlier method using the association rule technique for summarizing significant cells. Because the resultant rules using the association rule technique in the earlier work depend on the thresholds “support” and “confidence”, rule patterns are rather unstable. Optionally the conditional contingency table method is more applicable, since it has one more variable “condition” as a “layer” for analysis.

Table 2. Example of the patterns extracted by both methods

<table>
<thead>
<tr>
<th>Using unconditional contingency table in the earlier work</th>
<th>Using conditional contingency table in this paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) If access time to the nearest railway station to be over 10 minutes, then “accessibility” as a resigned factor increases, which implies that potential accessibility in mind would be around 10 minutes on walk or within about 350 meters from nearest railway station.</td>
<td>In addition to the left, age of household: “40s”, annual income of household: “over 15 million Yen”, family type of household: “couple” is especially high.</td>
</tr>
<tr>
<td>2) Households with average annual salary of 15–20 million Yen are targeted to high-rise building especially over 16 floors, and in this case “building facilities” is prone to be a decisive factor of purchase. In particular, the possibility that their category of business for household to be “financial” is very high.</td>
<td>In addition to the left, annual income of household: “6–8 million Yen”, and occupation of household: “businessman” is especially high.</td>
</tr>
</tbody>
</table>

5. Conclusion

We have presented an alternative, a rather straightforward method of extracting unexpected rules from a dataset with a huge number of attributes. Some of characteristics in our proposed method are as follows.
1) As mentioned above the motivation for our approach was to summarize solely unexpected rules or information among a huge number of attribute variables, not a huge number of transaction records. 2) Our proposal provides a relatively sound statistical ground in extracting significant cell(s) in the contingency table. 3) The extracted rules are robust in the perspective of consistency and compact interpretation.

Application of the alternatively proposed method here also produced quite encouraging results. From the author’s experience with a series of contingency table-based approach, the results from the earlier study and the extension in this paper are very promising. As such, many ideas presented here can in fact be modified to suit various applications of similar needs and situations. For future work, different thresholds in extracting significant cells can be experimented and some alternative effective methods to summarize the extracted significant cells can be compared. For the reliability and validity of the proposed methods, datasets in different domains can be evaluated and compared to their domain knowledge.

References


