A Distance-Based Approach for Mining Changes in Purchase Behavior in Retail Sale

Pradip Kumar Bala
XIM, Bhubaneswar
Xavier Square, Bhubaneswar
INDIA, PIN - 751013

ABSTRACT

The research paper proposes the relevance of a distance-based approach for change mining and suggests a technique for distance-based change mining. Change mining gives an insight to the retailers on the changing purchase behavior of the shoppers. The research makes an attempt to account for the distance between the attribute-values within a feature of the customer-profile while measuring the similarity between two patterns.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Management, Theory

Keywords

Association Rule, Change Mining, Data Mining, Distance-Based Mining, Pattern Changes.

1.INTRODUCTION, LITERATURE REVIEW AND RESEARCH OBJECTIVE

Retailers improve their marketing strategies by understanding the psychology of how consumers think, feel, reason, and select between different alternative brands or products. To remain competitive, it has become necessary for the retailers to understand the purchase behavior of consumers. The consumer's shopping record (e.g., products purchased, frequency of shopping, monetary value of shopping etc.) and demographic information (e.g., income, educational level of adults in the household, occupations of adults, ages of children, whether the family owns a house etc.) are the primary inputs used for mining purchase behavior of the shoppers. With changes in income, cost of living, number and age of the children, social and cultural forces, purchase behavior of a shopper may change with time. For retailers, it is necessary to learn these changes. In the parlance of data mining, discovering the changes in purchase behavior using data mining techniques is known as change mining. A typical purchase behavior in the context of this research paper is represented by the association rules in the following forms.

- 'Gender = Male' => 'Purchase = Snack Foods'
- 'Gender = Female' and 'Age = [30-40] => 'Purchase = Vegetables and Snack Foods'

The symbol, '=>', is read as 'associates'. LHS of '=>' is known as antecedent or, condition and RHS of '=>' is known as consequent. First example says that if a customer happens to be male, he is very likely to purchase 'snack foods'. The second example says that if a customer happens to be female of age in between 30 to 40, she is very likely to purchase 'vegetables' and 'snack foods'.

Strength of an association rule is measured in terms of *"support, confidence and lift"*. With total no. of transactions (rows) of 'N' in the transaction database of all items which includes X and Y also, support, confidence and lift for the association rule, X => Y, have been defined below.

Support of 'X=>Y' = n (XY)/N, where, n (XY) is number of transactions (rows) with both X and Y present.

Confidence of 'X=>Y' = [n (XY)] / [n (X)], where, n (X) is number of transactions (rows) with X present.

Lift of 'X=>Y' = (Confidence of the rule) / (Support of Y)

= [n (XY) / n (X)] / [n (Y) / N], where, n (Y) is number of transactions (rows) with Y present.

= [n (XY). N] / [n (X). n (Y)]

Various aspects of changes in purchase behavior can be explained as given below.

- A purchase behavior exists in period, P-1, and it does not exist in the next period, P-2.
- A purchase behavior does not exist in period, P-1, and it exists in the next period, P-2.
- A purchase behavior exists in both P-1 and P-2, but with different degrees of intensities given by the strength of the association rule.
- A particular attribute is with different values in P-1 and P-2, although otherwise same purchase behavior exists in both the periods.

In the context of change mining, four types of change patterns in purchase behavior as available in literature are emerging patterns, added patterns, perished patterns and unexpected changes ([1], [2], [3], [4]). These patterns have been explained below.

Emerging patterns are association rules existing in both present and past. Supports of emerging patterns generally vary from period, P-1 to next period, P-2. Emerging patterns are called positive emerging patterns when there is increase in support and they are called negative emerging patterns when there is decrease in support.

Example of positive emerging pattern can be "(a, b) => (c, d)", where, in P-1, association rule, (a, b) => (c, d) holds with support = 30% and in P-2, association rule, (a, b) => (c, d) holds with support = 40%.

Example of negative emerging pattern can be "d => g", where, in P-1, association rule, d => g holds with support = 30% and in P-2, association rule, d => g holds with support = 20%.

Added pattern is a new pattern found in the present, not existing in the past. Generally, all conditional and consequent parts of an association rule (added pattern) in period, P-2, differ significantly from any association rule in previous period, P-1.

Perished pattern is a vanished pattern found in the past and not existing in the present. Generally, all conditional and consequent parts of an association rule (perished pattern) in period, P-1, differ significantly from any rule in next period, P-2.

Unexpected change may be with unexpected consequent change, where the conditional parts remain same, but the consequent parts change with time, or, it may be with unexpected conditional change, where, the consequent parts remain same, but the conditional parts change with time.

Measure of rule similarity in two periods under study is used for mining emerging patterns, added patterns and perished patterns, whereas measure of unexpectedness is used for mining the patterns of unexpected change.

Liu and Hsu developed the similarity measure to analyze the degree of similarity between patterns at different periods of time [2]. However, the application of this measure is limited to cover only single item in the consequent (i.e., RHS of an association rule). Chen *et al.* gives a modified similarity measure to address more than one item in the consequent part and rule-similarity is measured as the product of LHS-similarity and RHS-similarity [5].

Hence, Rule-Similarity = LHS-Similarity x RHS-Similarity

To measure rule-similarity of a particular pattern in the form of an association rule in one period with the rules of another period, the concerned rule is compared with all the rules of another period and rule-similarity of the concerned rule is measured with each rule of the other period. Maximum of all such measured rule-similarities is called maximum rule similarity for the rule under consideration. Maximum rule similarity is one for two exactly same rules with same antecedents and consequents, although strength of the rules may be different. Chen *et al.* gives an approach to mine changes in the patterns as discussed in the next paragraphs [5].

A rule will be called an emerging pattern, if its similarity is '1' with one rule mined from the data of another period. In other words, requirement for emerging pattern is that its maximum rule similarity is '1'.

For perished pattern and added pattern, rule matching threshold (RMT) for 'similarity' is used to measure the degree of change. A rule will be called a perished pattern if maximum rule-similarity of this rule with the rules in later period is less than RMT. Similarly, a rule will be called an added pattern if maximum rule-similarity of this rule with the rules in earlier period is less than RMT.

As mentioned earlier, measure of unexpectedness is used for mining the patterns of unexpected change. A threshold value of "unexpectedness" is used to measure the degree of change. Requirement for unexpected consequent pattern is that the value of unexpectedness is '1' and requirement for unexpected conditional pattern is that the unexpectedness is '-1'.

1.1Limitation of the Existing Measure

In the work of [5], while measuring LHS-similarity or, RHS-similarity, attribute similarity comes in the process. An attribute can have different values, e.g., a particular product may be purchased in one unit or in two units or, more. As per the existing work, proximity between two attribute values of an attribute can have values of either '0' or '1', i.e., binary values are possible. As a result, an attribute with two different values in two rules are considered either "similar for same attribute-values for which value of '1' is assigned, or, "equally dissimilar for different attribute-values for which value of '1' is assigned. In fact, for different attribute-values, zero similarity or no similarity is considered for all possible values of the attributes provided they are different**.** However, two attribute-values may be very close to each other in one situation, whereas they may be far apart in a different situation.

Moreover, the paper in [5] does not consider products in LHS of a pattern depicted by association rule. For rules containing products in LHS (antecedent or, conditional part) and RHS (consequent part) both, existing similarity measure can not be simply extended without an appropriate modification.

1.2 Objective of the research

Binary approach of measurement for similarity between two attribute values does not seem to be sufficient in all cases. As per the binary approach, age group of 30-35 years has 'zero' similarity with both age groups of 55-60 years and 35-40 years. However, it is expected that the age group of 30-35 years is more similar with the age group of 35-40 years than with the age group of 55-60 years. This shows the insufficiency of binary approach of measurement.

In view of the above discussion, the objective in the present research paper is to develop a similarity measure for mining changes in purchase behavior in retail sale which can address non-binary distance and hence, non-binary similarity between the attribute-values.

2.DISTANCE-BASED APPROACH FOR CHANGE MINING

A distance-based approach for mining changes in purchase pattern in retail sale has been proposed in this paper. With an example, similarity measure has been computed based on the work of Chen *et al.* [5]. With the same example, similarity measure has also been computed based on the distance-based approach suggested in this paper. The example considers five features (age, income, education, marital status, and number of children) in the demographic profile of the customers. These five features have been taken as antecedent or condition in the patterns in the form of association rule. Three retail items (a, b, and c) have been considered for purchase by the customers and these items have been used as consequent in the patterns. Attributes for each feature of the demographic profiles have been given below.

(i) Age: Three age brackets, A_1, A_2, A_3 and A_4 have been chosen.

(ii) Income: Five income brackets, I_1, I_2, I_3, I_4 and I_5 have been chosen.

(iii) Education: Four different educational levels, E_1, E_2, E_3 and E_4 have been taken.

(iv) Marital status: Four different attributes, M_1, M_2, M_3 and M_4 have been taken.

(v) Number of children: Four different attributes, C_1, C_2, C_3 and C_4 have been taken.

Let us consider two following cases of pattern changes in two periods, T_1 (former period) and T_2 (later period). It is being observed how a pattern in T_1 changes to a different pattern in T_2 in two different cases. Similarity is measured in both cases and analyzed.

Case-1:

In **T_1**, the pattern is, "Age = A_1, Income = I_2, Marital Status = M_2, Number of Children = C_2} => {a, b}" and it is written as **'{A_1, I_2, M_2, C_2} => {a, b}'**.

In **T_2**, the pattern is, 'Age = A_1, Income = I_3, Education = E_3, Marital Status = M_2, Number of Children = C_3} => {a, b, c}' and it is written as **'{A_1, I_3, E_3, M_2, C_3} => {a, b, c}'**.

Case-2:

In **T_1**, the pattern is, "Age = A_1, Income = I_2, Education = E_4, Marital Status = M_2, Number of Children = C_2} => {a, b}" and it is written as **'{A_1, I_2, E_4, M_2, C_2} => {a, b}'**.

In **T_2**, the pattern is, 'Age = A_1, Income = I_3, Education = E_3, Marital Status = M_2, Number of Children = C_3} => {a, b, c}' and it is written as **'{A_1, I_3, E_3, M_2, C_3} => {a, b, c}'**.

Applying the method suggested in the existing work, 'similarity' in between the patterns of two periods has been computed separately for two cases as below.

$$\text{LHS-Similarity} = \frac{\text{(Number of common features in two rules)}}{\text{(Total number of features included in both the rules)}} \times \frac{\text{(Number of features with same attribute-values)}}{\text{(Number of common features in two rules)}}$$

Although the expression given above for LHS-similarity can be simplified, it has been kept in the present form for better explanation of the measure. Similarly,

$$\text{RHS-Similarity} = \frac{\text{(Number of common items in two rules)}}{\text{(Total number of items included in both the rules)}} \times \frac{\text{(Number of items with same attribute-values)}}{\text{(Number of common items in two rules)}}$$

If the patterns given by the association rules deal with only binary status of items (i.e, presence or absence of items in a rule) as discussed in the example of this paper, then for computing RHS-similarity, only the first part of the product will suffice. This is because in second part of the product, numerator and denominator will hold same values.

Hence,

$$\text{RHS-Similarity} = \frac{\text{(Number of common items in two rules)}}{\text{(Total number of items included in both the rules)}}$$

Finally, rule similarity = LHS-Similarity x RHS-Similarity.

For case-1:

LHS-Similarity = (4/5) x (1+0+1+0)/4 = 2/5 = 0.4

RHS-Similarity = 2/3 = 0.67

Rule Similarity = 0.4 x 0.67 = 0.268

For case-2:

LHS-Similarity = (5/5) x (1+0+0+1+0)/5 = 2/5 = 0.4

RHS-Similarity = 2/3 = 0.67

Rule Similarity = 0.4 x 0.67 = 0.268

Hence, it is observed that rule-similarity takes same value in both the cases. However, if we look into the patterns of case-1 and case-2, it is logical to think that two rules in case-2 are more similar to one another than two rules in case-1 which has been explained here. In case-1, the feature of 'education' does not appear in the pattern of **T_1** as a part of the antecedent, whereas 'education' appears in the pattern of **T_2** as a part of the antecedent. However, in case-2, the feature of 'education' appears in the patterns of **T_1** and **T_2** both as a part of their antecedents, although with different attribute values of E_4 and E_3 respectively in two periods. Hence, in case-1, two rules are more dissimilar compared to two rules in case-2. It is to be noted that other features and product items have remained same in two cases. In both cases, the products in **T_1** are {a, b} and the products in **T_2** are {a, b, c}. Moreover, in both the cases, in the demographic profiles of the patterns in both the periods, features of age, income, marital status and number of children are present, although with different attribute-values for some of the features.

Using the existing methodology for measuring LHS-similarity, similarity between two rules is taken as proportional to the "number of features with same attribute-values" which finds place in the numerator of the expression for LHS-similarity. In this process, it is inherently assumed that when a particular feature takes same attribute-values in two periods, they are considered similar and a value of '1' is assigned (considering 'zero' distance), whereas if the feature takes different attribute-values, they are considered totally dissimilar and a value of '0' is assigned (considering 'a very large distance' between two). It has been mooted in the present paper that two attribute-values of a feature should not be considered at equal distance for any two sets of different attribute-values. Two attribute-values may be close to one another while comparing with another set of two different attribute-values. Hence, the distance between the attribute-values must be taken into account while measuring rule-similarity.

Concept of attribute similarity matrix has been introduced in this work for capturing distance. In fact, distance is considered to be inversely proportional to the attribute similarity. Although the distance between attribute-values is considered to be inversely proportional to the similarity of attribute-values, similarity matrix for an attribute can be used to understand the distance between two attribute-values.

For the feature, age, with four attribute-values, the following similarity matrix has been considered below in table 1.

Table 1: Attribute Similarity Matrix for Age

	A1	A2	A3	A4
A1	1	0.8	0.2	0
A2		1	0.8	0.2
A3			1	0.8
A4				1

As shown above, a triangular matrix can be used for describing attribute similarity of the features where the values are populated in the range of [0, 1]. The similarity between A1 and A2 is 0.8 whereas between A1 and A3 is 0.2. This implies that the distance between A1 and A2 are closer compared to the distance between A1 and A3. Similarity for the same attribute-values, we assign a value of '1' which is the maximum similarity. Attribute similarity matrices for other features are given below in table 2, table 3, table 4 and table 5.

Table 2: Attribute Similarity Matrix for Income

	I1	I2	I3	I4	I5
I1	1	0.8	0.2	0	0
I2		1	0.8	0.2	0
I3			1	0.8	0.2
I4				1	0.8
I5					1

Table 3: Attribute Similarity Matrix for Education

	E1	E2	E3	E4
E1	1	0.8	0.2	0
E2		1	0.8	0.2
E3			1	0.8
E4				1

Table 4: Attribute Similarity Matrix for Marital Status

	M1	M2	M3	M4
M1	1	0.8	0.2	0
M2		1	0.8	0.2
M3			1	0.8
M4				1

Table 5: Attribute Similarity Matrix for Number of Children

	C1	C2	C3
C1	1	0.8	0.2
C2		1	0.8
C3			1

Hence, measurement or LHS-similarity has been modified to account for the attribute-similarity of the features and the expression is as given below.

$$\text{LHS-Similarity} = \frac{\text{(Number of common features in two rules)}}{\text{(Total number of features included in both the rules)}} \times \frac{\text{(Sum of attribute-similarity-values of the features)}}{\text{(Number of common features in two rules)}}$$

In the new expression, 'sum of attribute-similarity-values of the features' replaces 'number of features with same attribute-values' in the old expression for LHS-similarity. Hence, using the proposed methodology, rule-similarities are computed in two cases as given below.

For case-1:

LHS-Similarity = (4/5) x (1+0.8+1+0.8)/4 = 0.72

RHS-Similarity = 2/3 = 0.67

Rule Similarity = 0.72 x 0.67 = 0.48

For case-2:

LHS-Similarity = (5/5) x (1+0.8+0.8+1+0.8)/5 = 0.88

RHS-Similarity = 2/3 = 0.67

Rule Similarity = 0.88 x 0.67 = 0.59

As anticipated, the rule-similarity in case-2 is more than that in case-1.

3.CONCLUSIONS

A distance-based change mining has been proposed in this research paper. The suggested approach can be used for mining change patterns in purchase behavior of the shoppers in retailing. As the suggested measurement of similarity is based on the distance of attribute-values, it is expected to mine more realistic change patterns.

4.REFERENCES

[1] Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences, KDD-99

[2] Liu, B., & Hsu, W. (1996). Post-analysis of learned rules. Proceedings of the Thirteenth National Conference on KDD (PAKDD), 220-232.

[3] Padmanabhan, B., & Tuzhilin (1999). Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, 27, 303-318.

[4] Song, H.S., Kim, J.K., & Kim, S.H. (2001). Mining the change of customer behavior in an internet shopping mall. Expert System with Applications, 21(3), 157-168.

[5] Chen, M.C., Chiu, A.L., & Chang, H.H. (2005). Mining changes in customer behavior in retail marketing. Expert System with Applications, 28, 773-781.