# The Rough Set-Based Algorithm for Two Steps

Shu-Hsien Liao[1], Yin-Ju Chen[2], and Shiu-Hwei Ho[3]

[1] Department of Management Sciences, Tamkang University,
No.151 Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan R.O.C
[2] Graduate Institute of Management Sciences, Tamkang University,
No.151 Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan R.O.C
[3] Department of Business Administration, Technology and Science Institute of Northern Taiwan,
No. 2, Xueyuan Rd., Peitou, 112 Taipei, Taiwan, R.O.C
michael@mail.tku.edu.tw, s5515124@ms18.hinet.net,
succ04.dba@msa.hinet.net

**Abstract.** The previous research in mining association rules pays no attention to finding rules from imprecise data, and the traditional data mining cannot solve the multi-policy-making problem. Furthermore, in this research, we incorporate association rules with rough sets and promote a new point of view in applications. The new approach can be applied for finding association rules, which has the ability to handle uncertainty combined with rough set theory. In the research, first, we provide new algorithms modified from Apriori algorithm and then give an illustrative example. Finally, give some suggestion based on knowledge management as a reference for future research.

**Keywords:** Rough sets, Association rules, Data mining, Knowledge management.

## 1 Introduction

In data mining, it is often unclear which algorithm is best suited for the problem. Here, we require some decision support for data mining. To ensure that appropriate data is recorded when the collection process begins, it is useful to first build a decision model and use it as a basis for defining the attributes that will describe the data [2]. First, the quick growth of the Internet has led the global economy and social reforming. The traditional business model has been changed; the network shopping has become a part of our daily lives. Therefore, the fast growth of Internet stores changes not only the way customers buy goods but also the way customers receive goods. This means that the customers will pay more and more attention to the logistics services of the Internet stores, and then shopping on the Internet has become an important marketing channel. Second, most customers will rely on the recommendation system for a web-based computer retail store. It is a systematic attempt to strengthen a firms' competitive ability or give useful information to the end user. For the trend of diverseness, we propose a mining concept for knowledge refinement that is based on the discovery of unexpected patterns and uncertain information from the multi-database and the result of accurate prediction can be used for recommending products. The remainder of this paper is organized as follows. Section 2 reviews relevant literature and the problem statement. Section 3 new algorithms modified from Apriori algorithm. Section 4 an illustrative example. Closing remarks and future work are presented in Section 5.

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part II, LNCS 7063, pp. 63–70, 2011.

## 2 Literature Review and Problem Statement

Rough set theory, proposed by Pawlak in the 1980s, is a theory for the study of intelligent systems characterized by inexact, uncertain or vague information. Now rough set theory has found successful applications in such fields of artificial intelligence as machine learning, knowledge discovery, decision analysis, process control, pattern recognition, etc. It has become one of flash points in the research area of information science [3, 10]. An information system is a 4-tuple S={U, A, V, f}, where U is a finite set of objects, called the universe, A is a finite set of attributes, V is a domain of attribute a, and f:U×A→V is called an information function such that f(x,a) [5]. RST is an approach to aid decision making in the presence of uncertainty. It classifies imprecise, uncertain or incomplete information expressed in terms of data acquired from experience. In RST, a set of all similar objects is called an elementary set, which makes a fundamental atom of knowledge. Any union of elementary sets is called a crisp set and other sets are referred to as rough set. As a result of this definition, each rough set has boundary-line elements. For example, some elements cannot be definitively classified as members of the set or its complement. In other words, when the available knowledge is employed, boundary-line cases cannot be properly classified. Therefore, rough sets can be considered as uncertain or imprecise [7]. The association rule has become one of the most important techniques in data mining. The Apriori algorithm is a great discovery for data mining and the association rule. By reducing insignificant candidate item sets, it successfully increases processing speed and reduces the usage of space [1]. Traditional data mining cannot solve the multi-policy-making problem, such as, the consumer buys two things at the same time. In the real world, if the customers transaction data is either A or B, it does not differ to transform categorical or nominal data into Boolean. If we examine from Table 1, Table 2 does not look different. However, if the customers transaction dates are A and B, and if we examine from Table 3, Table 4 will look different. It will be the key factors for mining consumer purchasing patterns in transaction databases.

**Table 1.** Original data (Either A or B)

| U \ Q | Gender | Age | Purchasing |
|---|---|---|---|
| 001 | Male | 20 | Coca-Cola |
| 002 | Female | 23 | Pepsi |
| 003 | Male | 30 | Coca-Cola |
| 004 | Male | 22 | Pepsi |

**Table 2.** Original data (Boolean type)

| U \ Q | Gender | | Age | | | Purchasing | |
|---|---|---|---|---|---|---|---|
| | Male | Female | Under 20 | 20~25 | 26~30 | Coca-Cola | Pepsi |
| 001 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 002 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 003 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| 004 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |

**Table 3.** Original data (Both A and B)

| Q / U | Gender | Age | Purchasing |
|---|---|---|---|
| 001 | Male | 20 | Coca-Cola & Pepsi |
| 002 | Female | 23 | Pepsi |
| 003 | Male | 30 | Coca-Cola |
| 004 | Male | 22 | Coca-Cola & Pepsi |

**Table 4.** Original data (Boolean type)

| Q | Gender | | Age | | | Purchasing | |
|---|---|---|---|---|---|---|---|
| U | Male | Female | Under 20 | 20–25 | 26–30 | Coca-Cola | Pepsi |
| 001 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 002 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 003 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| 004 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |

## 3 Rough Set-Based Algorithm for Two-Step

In the classic rough set theory (RST), the attributes considered in an information system, including decision attributes, are taken on nominal values [4]. Mining generalized association rules between items in the presence of taxonomies has been recognized as an important model in data mining [6]. For the trend of diverseness market, the traditional association rule algorithm concept does not meet the needs of consumers anymore. The RST has been successfully applied in selecting attributes for improving the effectiveness in deriving decision trees/rules for decisions and classification problems [2]. The new algorithm modified from Apriori algorithm that is based on the RST has the ability to handle the uncertainty in the classing process and finding association rule. The new algorithm modified from rough set and Apriori is reproduced below.

**Table 5.** Sample data set

| Q | Attributes | | | Decision |
|---|---|---|---|---|
| U | Age | Gender | Shopping frequency | Brand loyalty |
| $t_1$ | 20–29 | Male | Once a month | High |
| $t_2$ | 30–39 | Female | Under fortnight | Median |
| $t_3$ | 20–29 | Male | Once a month | High |
| $t_4$ | 20–29 | Female | Once a month | Median |
| $t_5$ | 10–19 | Male | Other | Low |
| $t_6$ | 30–39 | Female | Under fortnight | Median |
| $t_7$ | 10–19 | Male | Other | Low |

**Step 1:** The basic concepts of the RST, an information system, can be seen as a system. We input data to reduce attributes as shown in Table 5. There are ten objects DT = {$t_1$, $t_2$,…, $t_{10}$}, three attributes A = {Age, Gender, Shopping frequency}, and a

decision attribute D = {Brand loyalty}. Assume the decision attribute has only three possible values: {High, Low, Median}, and the other input data is as shown in Table 5.

**Step 2:** Then we try to find $B = [t_i]_{Ind(A)}$ in Table 5. In the example, we classified customers into three major profitable groups with decision attribute. Therefore, three classes exist in the data set shown in Table 6.

**Table 6.** Attributes set

| Q / U | Attributes | | | Decision |
|---|---|---|---|---|
| | Age | Gender | Shopping frequency | Brand loyalty |
| $\{t_1, t_3\}$ | 20–29 | Male | Once a month | High |
| $\{t_2, t_6\}$ | 30–39 | Female | Under fortnight | Median |
| $\{t_4\}$ | 20–29 | Female | Once a month | Median |
| $\{t_5, t_7\}$ | 10–19 | Male | Other | Low |

**Table 7.** Ind(A)= Ind(A-$a_j$)( $a_j$=Age)

| Q / U | Attributes | | Decision |
|---|---|---|---|
| | Gender | Shopping frequency | Brand loyalty |
| $\{t_1, t_3\}$ | Male | Once a month | High |
| $\{t_2, t_6\}$ | Female | Under fortnightly | Median |
| $\{t_4\}$ | Female | Once a month | Median |
| $\{t_5, t_7\}$ | Male | other | Low |

**Table 8.** Ind(A) ≠ Ind(A-aj)($a_j$ = Shopping frequency)

| Q / U | Attributes | | Decision |
|---|---|---|---|
| | Age | Gender | Brand loyalty |
| $\{t_1, t_3\}$ | 20–29 | Male | High |
| $\{t_2, t_6\}$ | 30–39 | Female | Median |
| $\{t_4\}$ | 20–29 | Female | Median |
| $\{t_5, t_7\}$ | 10–19 | Male | Low |

**Table 9.** Ind(A) ≠ Ind(A-$a_j$) ($a_j$ = Gender)

| Q / U | Attributes | | Decision |
|---|---|---|---|
| | Age | Shopping frequency | Brand loyalty |
| $\{t_1, t_3, t_4\}$ | 20–29 | Once a month | High |
| $\{t_2, t_6\}$ | 30–39 | Under fortnightly | Median |
| $\{t_5, t_7\}$ | 10–19 | other | Low |

**Step 3:** According to Table 6, discriminate the efficient attribute set, which is used in mining association rules and building purchase profile. If Ind(A) = Ind(A-$a_j$) are shown in Table 7, Ind(A) ≠ Ind(A-$a_j$) are shown in Table 8 and Table 9. Finally, we find Ind(B) = {Gender, Shopping frequency} is the indiscernibility relation (see Table 10). Then, we use the reduce attribute set to find the customer who has high brand loyalty.

**Table 10.** Final data set

| Q / U | Attributes | | Decision |
|---|---|---|---|
| | Gender | Shopping frequency | Brand loyalty |
| $\{t_1, t_3\}$ | Male | Once a month | High |
| $\{t_2, t_6\}$ | Female | Under fortnightly | Median |
| $\{t_4\}$ | Female | Once a month | Median |
| $\{t_5, t_7\}$ | Male | other | Low |

**Step 4:** Using the Apriori algorithm and mining association rules by minimum support and minimum confidence.

$$\text{Support}((\text{Male}) \cap (\text{Once a month})) = \frac{(\text{Male}) \cap (\text{Once a month}) \text{Total of trades in database}}{\text{Total of trades in database}} = \frac{1}{5} = 20\%$$

$$\text{Confidence}((\text{Male}) \cap (\text{Once a month}) \rightarrow (\text{Brand loyalty - high})) = \frac{(\text{Male}) \cap \text{Once a month} \cap (\text{Brand loyalty - high}) \text{Total of trades in database}}{(\text{Male}) \cap (\text{Once a month}) \text{Total of trades in database}} = \frac{1}{1} = 100\%$$

**Algorithm-First stage**

```
Input:
Information System (IS);
Output:
{ Attribute Sets};
Method:
 1. Begin
 2.   IS = (U, A);
 3.    x_1, x_2, ..., x_n ∈ U ; /* where x_1, x_2, ..., x_n are the objects of
       set U */
 4.    a_1, a_2, ..., a_m ∈ A ; /* where a_1, a_2, ..., a_m are the elements of
       set A */
 5.  For each a_m do;
 6.   compute f(t, a); /* compute the information function
          in IS */
 7.   compute Ind(A − a_j); /* compute the relative reduct of
          the elements for element j */
 8.  Endfor;
 9.   Output { Attribute Sets };
10. End;
```

**Algorithm-Second stage**

```
Input:
Decision Table (DT);
Output:
{Classification Rules};
```

```
Method:
 1. Begin
 2. DT = (U, Q);
 3.     t1, t2, ⋯ ti ∈ U ; /* where t1, t2, ⋯ ti are the objects of
        set U */
 4.     Q = (A, D);
 5.      a1, a2, ⋯, aj ∈ A ; /* where a1, a2, ⋯, aj are the
          elements of set A */
 6.      d1, d2, ⋯, dl ∈ D ; /* where d1, d2, ⋯, dl are the decision
          elements of set D */
 7.  For each dl do;
 8.    compute f(t, a); /* compute the information function
        in DT */
 9.    compute Sup(aj); /* compute the support */
10.    compute Conf(aj → Dk); /* compute the confidence */
11.  Endfor;
12.    Output {Classification Rules};
13.End;
```

## 4 Illustrative Example

The two steps approach is based on a though market understanding and has to be managed in such a way as to effectively meet differing customer needs. Link up database marketing concept in the first step with the help of segmentation. Then, Apriori algorithm provides us with this possibility to categorize customers with most similarities in a group. In the present competitive environment within the service industries, retailing customers a maximizing profit from an existing customer base has become at least as important as attracting new customers. The input variable for each customer includes "education," "age," "shopping frequency," "product using state," "e-paper," et al. According to those attributes, we find some rules for decision criteria. We arrange the rank of decision rules (Table 11 and Table 12).

In first step, the rough set is applied to reduce attributes, and based on those attributes, we find some rules for the decision criteria. In the second step, the association rule is help to find an association diagram from customer's transaction data (Table 13 and Table 14). The two-step approach is easier than traditional method in finding association rule.

**Table 11.** Decision rules (output: price sensitive)

| No | Attribute set | rules | Generate rule |
|---|---|---|---|
| 1 | Education, product using state | 227 | (education) & (product using state) => (price sensitive) |
| 2 | Age, product using state | 191 | (age) & (product using state) => (price sensitive) |

**Table 12.** Decision rules (output: brand loyalty)

| No | Attribute set | rules | Generate rule |
|---|---|---|---|
| 1 | Education, product using state, shopping frequency | 227 | (education) & (product using state) & (shopping frequency) => (brand loyalty) |
| 2 | Education, product using state | 203 | (education) & (product using state) => (brand loyalty) |
| 3 | Education, e-paper | 185 | (education) & (e-paper) => (brand loyalty) |

**Table 13.** Association rule (min Sup=10, min Con=80)

| Sup | Conf | Lift | Consequent | Antecedent | | |
|---|---|---|---|---|---|---|
| 15.5 | 92.5 | 1.2 | Brand loyalty (Medium) | Price Sensitivity (High) | University or college degree | e-paper (No) |
| 12.3 | 88.1 | 1.2 | Brand loyalty (Medium) | Price Sensitivity (Medium) | University or college degree | e-paper (No) |
| 11.4 | 87.2 | 1.1 | Brand loyalty (Medium) | aged 20–29 | Family use | e-paper (No) |

**Table 14.** Association rule (min Sup = 5, min Con = 50)

| Sup | Conf | Lift | Consequent | Antecedent | | | |
|---|---|---|---|---|---|---|---|
| 6.5 | 59.1 | 2.6 | Price Sensitivity (Low) | Brand loyalty (Low) | University or college degree | e-paper (No) | - |
| 6.2 | 52.4 | 1.5 | Price Sensitivity (High) | Personal use | University or college degree good | Brand loyalty (Medium) | e-paper (No) |
| 7.0 | 50.0 | 1.6 | Price Sensitivity (Medium) | Family use | Brand loyalty (Medium) | e-paper (No) | aged 50–59 |

## 5 Conclusion

Rough Set Association Rule (RSAR) algorithm that is based on the RST has the ability to handle the uncertainty in the classing process and finding the association rule. Using the suggested methodology can reduce the dimension of transactions and help the decision maker to find useful association rules more effective than tradition association rules. Because using the traditional association rule to calculate support and confidence, as we changed min support and min confidence once, it will calculate again. In the example, we have to calculate $C_1^3 \times C_1^2 \times C_1^4 \times C_1^3 = 72$. In the new algorithm given in the example, we only need to calculate $C_1^2 \times C_1^4 \times C_1^3 = 24$. The method in the paper, we can use them to calculate itemsets' support and confidence from imprecise and multiple criteria data.

In this paper, we focus on the concept of rule and the management of customer relationship in the market downturn. Using the suggested methodology, decision maker can accurate at segmentation and assist the development of a new product. The

result of accurate prediction also can be used for recommending products to the customers and suggesting useful links. The strategic challenges faced by organizations are often framed in information- and knowledge-based terms such as uncertainty or complexity [9]. Consequently, how to find appropriate solutions to their particular knowledge problems is a most important problem for organizations to implement KM and make decisions. Real-world data tends to be imprecise due to human errors, instrument errors, recording errors, and so on. The RSAR algorithm has the ability to handle uncertainty and missing data in the classing process and discovering meaningful patterns and rules. This study proposes a new algorithm that can transform knowledge into information. Meanwhile, the knowledge conversion process to resolve the complexity that exists and provide more accurate information to make decisions of policy makers. Mining activities help us to know data patterns.

**Acknowledgements.** This research was funded by the National Science Council, Taiwan, Republic of China, under contract NSC 100-2410-H-032 -018-MY3.

## References

1. Chen, C.M., Liao, S.H.: Association Rule Algorithms for Logical Equality Relationships. In: IEEE 8th International Conference on Computer and Information Technology, Sydney, Australia, July 8-11 (2008)
2. Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., Kobler, A.: Data mining and visualization for decision support and modeling of public health-care resources. Journal of Biomedical Informatics 40, 438–447 (2007)
3. Lee, J.W.T., Yeung, D.S., Tsang, E.C.C.: Soft Computing - A Fusion of Foundations, Methodologies and Applications. Soft Computer 49(1), 27–33
4. Lee, J.W.T., Yeung, D.S., Tsang, E.C.C.: Rough sets and ordinal reducts. Soft Computing - A Fusion of Foundations, Methodologies and Applications 10, 27–33 (2006)
5. Li, R., Wang, Z.-o.: Mining classification rules using rough sets and neural networks. European Journal of Operational Research 157(2), 439–448 (2004)
6. Lian, W., Cheung, D., Yiu, S.M.: An efficient algorithm for finding dense regions for mining quantitative association rules. Computers and Mathematics with Applications 50(3-4), 471–490 (2005)
7. Parmar, D., Wu, T., Blackhurst, J.: MMR: An algorithm for clustering categorical data using. Rough Set Theory, Data & Knowledge Engineering 63(3), 879–893 (2007)
8. Uta, J., Martin, C., Susan, B.: Demand chain management-integrating marketing and supply chain management. Industrial Marketing Management 36(3), 377–392 (2007)
9. Zack, M.H.: The role of decision support systems in an indeterminate world. Decision Support Systems 43(4), 1664–1674 (2007)
10. Zhang, W.X., Qiu, G.F., Wu, W.Z.: A general approach to attribute reduction in rough set theory. Science in China Series F: Information Sciences 50(2), 188–197 (2007)