An Incremental Approach for Rule Induction under Coarsening and Refining of Attribute Values in E-Business Systems

Dun Liu[1] **Tianrui Li**[2] **Junbo Zhang**[3]

[1]Department of Economics and Management, Southwest Jiaotong University
Chengdu, 610031, China
[2,3]Department of Information Science and Technology, Southwest Jiaotong University
Chengdu, 610031, China

Abstract

The variation of attribute values is an important case in the dynamic E-Business systems. With the rapid increase and update of data sets in business database everyday, a new incremental model, approach as well as its algorithm is presented for rule induction under coarsening and refining of attribute values. An example with online-shopping illuminate our method and experiments validate the feasibility of the incremental approach.

Keywords: Rough sets theory, incremental learning, coarsening, refining, E-business.

1. Introduction

In the machine learning community, there exists a number of approaches to learning classification rules for E-business applications. The inductive learning is a popular approach; its object is used to find the classification rules. As a component of hybrid solutions in machine learning and data mining, rough set theory has been found to be particularly useful for rule induction and feature selection. Successful applications, such as multimedia, web and text mining, signal and image processing, software engineering, robotics, and engineering, have proved that the rule induction approaches from the view of rough sets are helpful in obtaining interesting knowledge (rules) from the databases. The researches of rule induction based on rough set models assume that the procedure of classification will eventually converge to a stable state, but the volume of data is growing rapidly in real-life applications. As a famous E-business company, the eBay's massive oracle database has over 212 million registered users in 2006, holding two Petabytes of user Data. This large scare database is running on Teradata with over 20 billion transactions per day [1]. For management and market decision in such a business environment, an efficient rule induction method with the real-time processing ability is extraordinarily valuable. As an efficient data analysis' technique, the incremental approach has been paid much attention to by

the machine learning and data mining community. Since a dataset is the collection of data items (objects or records) with its features (attributes) and feature values (attribute value), recent studies mainly focus on variation of the database. Those changes are finite with a relatively small portion of the original training examples. Based on this assumption, the incremental approaches for data updating are mainly focus on three aspects: variation of objects [2, 8, 9, 14, 15], variation of attributes [3, 7, 12] and variation of attribute values [4, 10, 11].

In this paper, we focus on studying the coarsening and refining of attribute values. Instead of studying the incremental approach for updating lower approximation and upper approximation, we care more about the knowledge (rules) updating process. The rest of the paper is organized as follows: Section 2 provides the basic concepts of coarsening and refining of attribute values. The model of incremental learning when attributes value changes is given in Section 3. An algorithm for updating rules under coarsening and refining of attribute values is also presented. Section 4 shows an example to illustrate the proposed model, and experiment results are presented in Section 5. The paper ends with conclusions and further research topics in Section 6.

2. Preliminaries

The basic concepts, notations and results of coarsening and refining attribute values are briefly reviewed in this section [4, 10, 10, 13].

A complete information system is defined as a 4-tuple $S = (U, A, V, f)$, where $U = \{x_1, x_2, \cdots, x_p\}$ is a non-empty finite set of objects, $A = C \cup D$ is a non-empty finite set of attributes, C denotes the set of condition attributes and D denotes the set of decision attributes, $C \cap D = \emptyset$. $V = \bigcup_{a \in A} V_a$ and V_a is a domain of the attribute a, and $f : U \times A \to V$ is an information function such that $f(x, a) \in V_a$ for every $x \in U$, $a \in A$.

For an information system $S = (U, A, V, f)$, $B \subseteq A$, $a_l \in B$. $f(x_i, a_l)$ is the value x_i of on attribute a_l. $f(x_k, a_l)$ is the value of object $k(k \neq i)$ on attribute a_l, and $f(x_i, a_l) \neq f(x_k, a_l)$. Denote $U_{a_l} = \{x' \in U | f(x'_i, a_l) = f(x_k, a_l)\}$. Let $f(x'_i, a_l) = f(x_k, a_l) \cup f(x_i, a_l)$, $\forall x'_i \in U_{a_l}$, then $f(x_i, a_l)$ is coarsening to $f(x_k, a_l)$. For convenience, let $a_l^{\wedge}$ denote the attribute a_l after coarsening, $B^{\wedge}$ denote the attribute after coarsening B, $V_{a_l}^{\wedge}$ is the value domain of V_{a_l}.

For an information system $S = (U, A, V, f)$, $B \subseteq A$, $a_l \in B$. $f(x_i, a_l)$ is the value x_i of on attribute a_l. Denote $U_{a_l} = \{x' \in U | f(x'_i, a_l) = f(x_k, a_l)\}$. Let $f(x'_i, a_l) = v$ where $v \overline{\in} V_l$, $x'_i \in U_{a_l}$, then we call the attribute value $f(x_i, a_l)$ on object x'_i is refining to v. For convenience, let $a_l^{\vee}$ denote the attribute a_l after refining, $B^{\vee}$ denote the attribute after refining B, $V_{a_l}^{\vee}$ is the value domain of V_{a_l}.

3. An incremental rule learning model and its algorithm

Suppose there exists two different times: time t and time t+1 in our model. We denote $S = (U, A, V, f)$, $A = C \cup D$ as the information system at time t; we denote $S' = (U', A', V', f')$, $A' = C' \cup D'$ as the information system at time t+1. Obviously, we have $U = U'$, $A = A'$ and $V \neq V'$ by assuming the objects and attributes are not changing. Generally, the changes of attribute values can divide into two

categories: one is coarsening of attribute values, another is refining of attribute values. With the angle of granular computing, the coarsening of attribute values causes the coarsening of the knowledge granularity, and the partitions generated by the attributes also become coarser; in the same way, the refining of attribute values causes the refinement of the knowledge granularity, and the partitions generated by the attributes also become refiner. However, considering the definitions of coarsening and refining, there are four typical cases hold: (1). Coarsening of the condition attribute values; (2). Refining of the condition attribute values. (3). Coarsening of the decision attribute values. (4). Refining of the decision attribute values.

In [4], Chen et al studied the properties for dynamic maintenance of upper and lower approximations under coarsening and refining of attribute values. Let $S = (U, A, V, f)$ be an information system, $\forall X \in U$, we have: $\underline{R}_{A^\wedge} \subseteq \underline{R}_A$, $\overline{R}_{A^\wedge} \supseteq \overline{R}_A$; $\underline{R}_{A^\vee} \supseteq \underline{R}_A$, $\overline{R}_{A^\vee} \subseteq \overline{R}_A$. That is, the coarsening of the attribute values causes the compression of the lower approximation and the expansion of the upper approximation. Conversely, the opposite cases happen when we refine of the attribute values. With the above analysis, we try to construct the model of the incremental learning process.

Suppose $S = (U, A, V, f)$ is an information system at time t. Where, $A = C \cup D$ and $C \cap D = \emptyset$. $U = \{x_1, x_2, \cdots, x_p\}$. The partitions U/C and U/D divide U into m condition equivalence classes and n decision equivalence classes, and we denote them as $U/C = \{X_1, X_2, \cdots, X_m\}$ and $U/D = \{D_1, D_2, \cdots, D_n\}$, respectively. At time t+1, four typical cases may happen, which are shown in the following.

(C1). Coarsening of the condition attribute values.

In this case, we have: $U/C \subseteq U'/C'$, $\underline{R}_{C^\wedge} \subseteq \underline{R}_C$, $\overline{R}_{C^\wedge} \supseteq \overline{R}_C$. Suppose $U'/C' = \{X'_1, X'_2, \cdots, X'_m\}$, $m' < m$ and $\Delta m' = m - m'$. The attribute values in $a_i \in C$ are coarsening, and these changes may affect their corresponding elements $\Delta U \subseteq U$. Hence, a heuristic strategy to estimate the knowledge granularity is comparing the partitions between $\Delta U/C' - \{a_i\}$ and $\Delta U/C'$, the system may keep steady; otherwise, we just recomputed the new partitions for $\Delta U/C'$. Specially, if a_i is a redundant attribute, $m' = m$ and $\Delta m' = 0$.

(C2). Refining of the condition attribute values.

In this case, we have: $U/C \supseteq U'/C'$, $\underline{R}_{C^\vee} \supseteq \underline{R}_C$, $\overline{R}_{C^\vee} \subseteq \overline{R}_C$. Suppose $U'/C' = \{X'_1, X'_2, \cdots, X'_{m''}\}$, $m < m''$ and $\Delta m'' = m'' - m$. The attribute values in $a_j \in C$ refining, and these changes may affect their corresponding elements $\Delta U' \subseteq U$. Hence, a heuristic strategy to estimate the knowledge granularity is comparing the partitions between $\Delta U'/C' - \{a_j\}$ and $\Delta U'/C'$, the system may keep steady; otherwise, we just recomputed the new partitions for $\Delta U'/C'$. Specially, if a_j is a redundant attribute, $m'' = m$ and $\Delta m'' = 0$.

(C3). Coarsening of the decision attribute values.

In this case, we have: $U/D \subseteq U'/D'$, $\underline{R}_{D^\wedge} \subseteq \underline{R}_D$, $\overline{R}_{D^\wedge} \supseteq \overline{R}_D$. Suppose $U'/D' = \{D'_1, D'_2, \cdots, D'_n\}$, $n' < n$ and $\Delta n' = n - n'$. These changes may affect their corresponding elements $\Delta U'' \subseteq U$. Considered there is only one decision attribute in the system, the coarsening of the decision attribute values causes the coarsening of the knowledge granularity.

(C4). Refining of the decision attribute values.

In this case, we have: $U/D \supseteq U'/D'$, $\underline{R}_{D^\vee} \supseteq \underline{R}_D$, $\overline{R}_{D^\vee} \subseteq \overline{R}_D$. Suppose $U'/D' = \{D'_1, D'_2, \cdots, D'_{n''}\}$, $n < n''$ and $\Delta n'' = n'' - n$. These changes may affect their corresponding elements $\Delta U''' \subseteq U$. Considered there is only one decision attribute in the system, the refining of the decision attribute values causes the refining of the knowledge granularity.

As stated above, we describe the updating rules strategies when attribute value changes. In order to build the incremental algorithm, the concept concern with interesting knowledge is introduced at first.

Suppose an information system $S = (U, C \cup D, V, f)$ with $C \cap D = \emptyset$. $U/C = \{X_1, X_2, \cdots, X_m\}$ is a partition of objects under the condition attributes of C, where X_i $(i = 1, 2, \cdots, m)$ is a condition equivalence class; $U/D = \{D_1, D_2, \cdots, D_n\}$ is a partition of objects under the decision attribute of D, where D_j $(j = 1, 2, \cdots, n)$ is a decision equivalence class. $\forall X_i \in U/C$, $\forall D_j \in U/D$, the support, accuracy and coverage of $X_i \to D_j$ are defined as follows, respectively [8, 9].

$Supp(D_j|X_i) = |X_i \bigcap D_j|$;

$Acc(D_j|X_i) = |X_i \bigcap D_j|/|X_i|$;

$Cov(D_j|X_i) = |X_i \bigcap D_j|/|D_j|$.

where $|X_i|$ and $|D_j|$ denote the cardinality of set X_i and D_j, respectively. So, we can also define the support matrix, accuracy matrix and coverage matrix [8, 9].

$Supp(D|X) = (|X_i \bigcap D_j|)_{m \times n}$;

$Acc(D|X) = (|X_i \bigcap D_j|/|X_i|)_{m \times n}$;

$Cov(D|X) = (|X_i \bigcap D_j|/|D_j|)_{m \times n}$.

Simply, we set two thresholds α $(\alpha > 0.5)$ and β $(0 < \beta < 1)$, the rule $X_i \to D_j$ is called a interesting knowledge if it satisfies both $Acc(D_j|X_i) \geq \alpha$ and $Cov(D_j|X_i) \geq \beta$ for $\forall X_i$ $(i = 1, 2, \cdots, m)$, $\forall D_j$ $(j = 1, 2, \cdots, n)$ [16, 17].

4. An Illustration

In this section, an E-business example is given to show how to use the above approach and algorithm to maintain the interesting knowledge dynamically. In the information system at time t given in Table 1. $U = \{x_1, x_2, \cdots, x_{12}\}$ stands for 12 types of E-business on-line shops, the condition attributes $C = \{a_1, a_2, a_3\}$ stands for the 3 main characters and we denote them as security, credit standing and public praise, respectively. The decision attribute $D = \{d\}$ stands for the estimation level of the shop. And Num stands for the cardinal number of one certain type of E-business on-line shop.

U	a_1	a_2	a_3	d	N
x_1	1	1	1	0	10
x_2	1	2	1	0	15
x_3	1	2	1	1	25
x_4	1	2	1	2	5
x_5	1	2	2	1	3
x_6	2	2	1	1	42
x_7	2	2	2	2	3
x_8	2	3	3	1	20
x_9	2	3	3	2	47
x_{10}	2	3	3	3	5
x_{11}	3	3	3	2	20
x_{12}	3	3	3	3	25

Table 1: An E-business Information System

The meaning of the values for every attribute is shown as follows.

Security (a_1): 1 = Bad; 2 = Average; 3 = Good.

Credit standing (a_2):1 = Bad; 2 = Average; 3 = Good.

Public praise (a_3):1 = Bad; 2 = Average; 3 = Good.

Estimation Level (d): 0 = No Star; 1 = One Star; 2 = Two Star; 3 = Three Star.

From Table 1, we can calculate that $U/C = \{X_1, \cdots, X_7\} = \{\{x_1\}, \{x_2,$

$x_3, x_4\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8, x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$, $U/D = \{D_1, D_2, D_3, D_4\} = \{\{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}, \{x_7, x_8, x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$. We can compute the support matrix at time t as follows.

$$Supp^{(t)}(D|X) = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 15 & 25 & 5 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 42 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 20 & 47 & 5 \\ 0 & 0 & 20 & 25 \end{pmatrix}$$

Assume at time t+1, the attribute value 2 in condition attribute a_2 will refine to two parts: $V_{a_2}(x_2) = V_{a_2}(x_3) = V_{a_2}(x_4) = 2^-$, $V_{a_2}(x_5) = V_{a_2}(x_6) = V_{a_2}(x_7) = 2^+$, where the number 2^- stands for under average and 2^+ stands for above average. Furthermore, the attribute values $\{1, 2\}$ in decision attribute d will coarsen as $1'$, that is, the estimation level One Star and Two Star can combine as one level. So, we use (C1)-(C4) to update these classifications with U/C and U/D at time t.

Firstly, the refining of a_2 affects the elements $\Delta U' = \{x_2, x_3, x_4, x_5, x_6, x_7\}$, and $\Delta U'/C' - \{a_2\} = \{x_2, x_3, x_4\}$, $\{x_5\}, \{x_6\}, \{x_7\}\}$. So, we have $\Delta U'/C' = \Delta U'/C' - \{a_2\}$, $\Delta m' = 0$, the system may keep steady. Secondly, the coarsening of d affects the elements $\Delta U'' = \{x_4, x_5, x_6, x_7, x_8, x_9, x_{11}\}$, and $\Delta U''/d' = \{x_4, x_5, x_6, x_7, x_8, x_9, x_{11}\}$. Comparing with $\Delta U''/d'$ and $\Delta U''/d$, the decision equivalence classes D_2 and D_3 at time t will coarse to D'_2 at time t+1, $\Delta n' = 1$. The columns in the updating matrices are change as $n' = n - \Delta n' = 4 - 1 = 3$, so we denote the decision equivalence classes as $\{D'_1, D'_2, D'_3\}$. Due to the first column and forth column in support matrix $Supp^{(t)}(D|X)$ at time t is not changed at time t+1, we merely need to update the second and third columns in $Supp^{(t)}(D|X)$, which recalculate and denote as $Supp^{(t+1)}(D'|X')$.

$$Supp^{(t+1)}(D'|X') = \begin{pmatrix} 10 & 0 & 0 \\ 15 & 30 & 0 \\ 0 & 3 & 0 \\ 0 & 42 & 0 \\ 0 & 3 & 0 \\ 0 & 67 & 5 \\ 0 & 20 & 25 \end{pmatrix}$$

The accuracy matrix and coverage matrix at time t and t+1 can be directly generated from the support matrix. If we set $\alpha = 0.6$ and $\beta = 0.4$, we can easily get that: $X_1 \rightarrow D_1$, $X_4 \rightarrow D_2$ and $X_6 \rightarrow D_3$ satisfy the condition $Acc(D_j|X_i) \geq 0.6$ and $Cov(D_j|X_i) \geq 0.4$, and there are the interesting knowledge at time t. However, $X'_1 \rightarrow D'_1$ and $X'_6 \rightarrow D'_2$ satisfy the condition $Acc(D'_j|X'_i) \geq 0.6$ and $Cov(D'_j|X'_i) \geq 0.4$, and there are the interesting knowledge at time t+1.

5. Experimental Evaluations

In this section, we design an experiment to estimate the effectiveness of the proposed knowledge incremental updating algorithm when attribute values are coarsened and refined.

Experiments were performed on a computer with Inter(R) CPU E5520 2.27G (16CPUs), 16 GB of memory, running Microsoft Window Server 2003. Methods of incremental updating proposed in our paper (Algorithm 2) and the non-incremental updating (Algorithm 1) were developed in VC++ 6.0. We chose four data sets named "IRIS", "CPU", "Bank-data" and "Segment", which listed in Table 2, and are available from the well-known machine learning webs Weka and UCI (http://archive.ics.uci.edu/ml/datasets.html). Since the proposed incremental updating method is discussed based

on the complete information system, we delete the object where one of its attribute values is null or missed. The time unit of incremental updating is second.

Name	*IRIS*	*CPU*	*Bank*	*Seg*
$\|U\|$	150	209	600	1500
$\|U/C\|$	4	6	10	18
$\|U/D\|$	1	1	1	1

Table 2: The basic information of the four databases

The strategy of our experiments is to compare the computing speed for the four databases by using algorithm 1 and algorithm 2. The threshold values α and β are fixed at first. Then, we randomly choose 5% (10%) data from the original information database at time t as the change data, and we randomly coarsen or refine the values of these 5% (10%) data. The average elapsed times calculating by repeating the computing process for 100 times are used to estimate the efficiency of the two algorithms. The experimental results are shown in Table 3.

Batabase	*Alg.*1 (5%)	*Alg.*2 (5%)
IRIS	0.0017	0.0006
CPU	0.0039	0.0026
Bank	0.0527	0.0089
Seg	0.2582	0.0123
Batabase	*Alg.*1 (10%)	*Alg.*2 (10%)
IRIS	0.0035	0.0009
CPU	0.0074	0.0013
Bank	0.0950	0.0097
Seg	0.4992	0.0137

Table 3: The average elapsed times between the two algorithms (Unit: Seconds)

From Table 3, we discover the algorithm 2 is more effective for the dynamic information system, especially for the complex and massive database. These tables and figures give us an intuitive understanding for the incremental learning process, which can help the decision makers do quicker and easier choices in practical dynamic decision problems.

6. Conclusions

The core idea of incremental strategy in machine learning is decreasing the computing complexity and avoiding relearning the whole data in updating datasets. Observed by these opinions, a new rule induction incremental approach under coarsening and refining of attribute values is proposed in this paper. An incremental model as well as its algorithm is also presented. The illustration and experimental results validate the rationality and efficiency of the proposed method. Our future research will focus on extensions of the current approach to incomplete systems and the real-life applications of our approach.

Acknowledgement

This work is partially supported by the National Science Foundation of China (No.60873108), the Scientific Research Foundation of Graduate School of Southwest Jiaotong University and the Doctoral Innovation Foundation of Southwest Jiaotong University (No. 200907), China.

References

[1] Available from: http://www.dba-oracle.com/oracle_news/news_ebay_petafiles.html.

[2] Bang. W., Bien. Z.: A new incremental learning algorithm in the framework of rough set theory. *In-*

ternational Journal of Fuzzy Systems, 1, pp. 25-36, 1999.

[3] Chan. C.: A rough set approach to attribute generalization in data mining. *Information Sciences*, 107, pp. 177-194, 1998.

[4] Chen. H., Li. T., Qiao. S., Ruan. D.: A rough set based dynamic maintenance approach for approximations in coarsening and refining attribute values. *International Jounral of Intelligence System*, 25(10), pp. 1005-1026, 2010.

[5] Hu. F., Wang. G., Huang. H., Wu. Y.: Incremental attribute reduction based on elementary sets. In: *RSFDGrC2005*, LNAI, 3641, pp. 185-193, 2005.

[6] Jerzy. B., Slowinski. R.: Incremental induction of decision rules from domi-nance-based rough approximations. *Electronic Notes in Theoretical Computer Science*, 82, pp. 40-51, 2003.

[7] Li. T., Ruan. D., Wets. G., et.al.: A rough sets based characteristic relation ap-proach for dynamic attribute generaliza-tion in data mining. *Knowledge-Based Systems*, 20, pp. 485-494, 2007.

[8] Liu. D., Li. T., Ruan. D., Zou. W.: An incremental approach for inducing knowledge from dynamic information systems. *Fundamenta Informaticae*, 94, pp. 245-260, 2009.

[9] Liu. D., Li. T., Ruan. D., Zhang. J.: Incremental Learning Optimization on Knowledge Discovery in Dynamic Business Intelligent Systems. *Journal of Global Optimization*. DOI: 10.1007/s10898-010-9607-8, 2010.

[10] Liu. D., Li. T., Chen. H., Ji. X.: Approaches to knowledge incremental learning based on the changes of attribute values. In: Proceedings of the 4th International Conference on Intelligent Systems and Knowledge Engineering (*ISKE 2009*), pp. 94-99, 2009.

[11] Liu. D., Li. T., Liu. G., Hu. P.: An approach for inducing interesting incremental knowledge based on the change of attribute values. In: Proceedings of 2009 IEEE International Conference on Granular Computing (*GRC 2009*), pp. 415-418, 2009.

[12] Liu. D., Zhang. J., Li. T.: A probabilistic rough set approach for incremental learning knowledge on the change of attribute. In: Proceedings of 2010 International Conference on Foundations and Applications of Computational Intelligence (*FLINS 2010*), pp. 722-727 (2010)

[13] Pawlak. Z.: Rough sets. *International Journal of Computer and Information Science*, 11, pp. 341-356, 1982.

[14] Shan. L., Ziarko. W.: Data-based acquisition and incremental modification of classification rules. *Computational Intelligence*, 11, pp. 357-370, 1995.

[15] Tong. L., An. L.: An Incremental learning of decision rules based on rough set theory. In: Proceedings of the World Congress on Intelligent Control and Automation (*WCICA2002*), pp. 420-425, 2002.

[16] Tsumoto. S.: Accuracy and coverage in rough set rule induction. In: Proceedings of 3rd International Conference on Rough Sets and Current Trends in Computing (*RSCTC 2002*), LNAI, 2475, pp.373-380, 2002.

[17] Wong. S., Ziarko. W., Pawlak. Z.: Algorithm for inductive learning, *Bulletin of the Polish Academy of Sciences, Technical Sciences*, 34, pp. 271-276, 1986.