NOVEL INCREMENTAL ALGORITHMS FOR ATTRIBUTE REDUCTION FROM DYNAMIC DECISION TABLES WHEN ADDING OBJECT SET

Ho Thi Phuong[1*], Truong Duc Phuong[2]

[1]Tay Nguyen University, [2]Hanoi Metropolitan University

Abstract: *In recent years, attribute reduction methods following fuzzy rough set approach have attracted the attention of researchers because they improve the accuracy of the classification model. However, most of the proposed methods are performed on the unchanged decision table. In this paper, we build an incremental algorithm to find the approximate reduct according to the combined filter-wrapper approach. Experimental results on a number of sample datasets show that the proposed incremental algorithm is more efficient than some other incremental algorithms following the filter approach in terms of the number of reductive set attributes and classification accuracy.*

Keywords: *Fuzzy rough sets, fuzzy distances, incremental algorithms, decision tables, attribute reduction.*

Received 15 June 2022

Revised and accepted for publication 23 August 2022

(*) Email: htphuong@ttn.edu.vn

1. INTRODUCTION

In fact, decision tables are often large in size and are constantly changing and updating [1, 2]. The application of reductive set finding algorithms according to the traditional rough set approach and extended rough set models faces many challenges [3]. In case the decision tables are changed, these algorithms had to recompute the reduct on the entire decision table after the change, so the cost of computation time increases significantly. In case the decision table is large, the implementation of the algorithm on the entire decision table will be difficult in terms of execution time [4]. Therefore, splitting the decision table to find the reduct on each part is the proposed solution. However, calculating the reduct based on the reducts of each part is a problem to be solved. Therefore, the researchers proposed an incremental computational approach to find the reduct [5, 6, 7]. In case the decision table is changed, the incremental algorithm does not recompute the reduct on the entire decision table, but only updates the

existing reduct based on the changed data composition. In the case of a large decision table, the incremental algorithm finds the reduct on a fragmented component, then updates the reduct when adding the remaining components. In theory, the incremental algorithm is capable of minimizing the execution time and is capable of performing on large decision tables.

The main objective of the paper is to reduce the number of reductive set attributes and improve the classification accuracy compared to the published incremental algorithms. In this paper, Incremental Filter-Wrapper Algorithm for Fuzzy Partition Distance based Attribute Reduction When Add Objects, called IFW_FDAR_AdObj algorithm, is propose to find the approximate reduct of the decision table using fuzzy distance measure in the case of addition of the feature set. The proposed algorithm based on the combined filter-wrapper approach consists of two stages: the filter stage finds the candidates for the reduct each time the attribute with the greatest importance is added, called the approximate reduct, with the stopping condition that the fuzzy distance measure is preserved; Wapper stage finds the reduct with the highest classification accuracy. Experimental results on sample data sets show that the suggested incremental algorithm is more efficient than the non-incremental filter-wrapper algorithm in terms of execution time. Moreover, the proposed algorithm is more efficient than the published filter incremental algorithms in terms of the number of attribute sets and classification accuracy by selecting the candidate with the best classification accuracy in the wapper stage.

The rest of the paper is organized as follows. Section II develops IFW_FDAR_AdObj algorithm to find the approximate reduct of the decision table using fuzzy distance measure in the case of addition of the feature set. Section III presents result. Conclusions are drawn in section IV.

2. IFW_FDAR_ADOBJ ALGORITHM (INCREMENTAL FILTER-WRAPPER ALGORITHM FOR FUZZY PARTITION DISTANCE BASED ATTRIBUTE REDUCTION WHEN ADD OBJECTS)

In this section, we propose a filter-wrapper incremental algorithm by using FPD when **adding object set into the decision table**.

Algorithm IFW_FDAR_AdObj

Input:

1. A decision table $DS = (U, C \cup D)$ with $U = \{x_1, x_2, \ldots, x_n\}$, a FER $\tilde{R}$, the reduct $B \subseteq C$.
2. Fuzzy equivalent matrices

$$M_U(\tilde{R}_B) = [b_{ij}]_{n\times n}, M_U(\tilde{R}_C) = [c_{ij}]_{n\times n}, M_U(\tilde{R}_D) = [d_{ij}]_{n\times n}$$

3. Added set of objects $\Delta U = \{x_{n+1}, x_{n+2}, \ldots, x_{n+s}\}$

Output: The approximation reduct B_{best} of $DS' = (U \cup \Delta U, C \cup D)$ with highest classification accuracy.

Sep 1: Initialization

1. $T := \emptyset$; // T contains the candidates for best reduct
2. Compute fuzzy equivalent matrices on $U \cup \Delta U$

$M_{U\cup\Delta U}(\tilde{R}_B) = [b_{ij}]_{(n+s)\times(n+s)}$, $M_{U\cup\Delta U}(\tilde{R}_D) = [d_{ij}]_{(n+s)\times(n+s)}$;

Step 2: Check the added set of objects

3. Set $X := \Delta U$;
4. For $i = 1$ to s do
5. If $[x_{n+i}]_{\tilde{B}} \subseteq [x_{n+i}]_{\tilde{D}}$ then $X := X - \{x_{n+i}\}$;
6. If $X = \emptyset$ then Return B_0; //Approximation reduct does not change
7. Set $\Delta U := X$; $s := |\Delta U|$; //Reset the object set

Step 3: Finding the best reduct

8. Compute original FPDs

$FPD_U\left(\Phi(\tilde{R}_B), \Phi(\tilde{R}_{B\cup D})\right)$; $FPD_U\left(\Phi(\tilde{R}_C), \Phi(\tilde{R}_{C\cup D})\right)$;

9. Compute FPDs using incremental formulas:

$$FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_B), \Phi(\tilde{R}_{B\cup D})\right);\ FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_C), \Phi(\tilde{R}_{C\cup D})\right)$$

//Filter stage: finding candidates for reduct

10. While $FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_B), \Phi(\tilde{R}_{B\cup D})\right) \neq FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_C), \Phi(\tilde{R}_{C\cup D})\right)$ do
11. Begin
12. For each $a \in C - B$ do
13. Begin
14. Compute $FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_{B\cup\{a\}}), \Phi(\tilde{R}_{B\cup\{a\}\cup D})\right)$ by using cremental formulas;
15. Compute $SIG_B(a) = FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_B), \Phi(\tilde{R}_{B\cup D})\right) - FPD_{U\cup\Delta U}\left(\Phi(\tilde{R}_{B\cup\{a\}}), \Phi(\tilde{R}_{B\cup\{a\}\cup D})\right)$;
16. End;
17. Select $a \in C - B$ satisfying $SIG_B(a_m) = \underset{a\in C-B}{Max}\{SIG_B(a)\}$;
18. $B := B \cup \{a_m\}$;
19. $B_0 := B_0 \cup \{a_m\}$;
20. $T := T \cup B_0$;
21. End;

//**Wrapper stage: Finding the reduct with the highest classification accuracy**

22. Set $t := |T|$ //t is the number of T, $T = \{B_0 \cup \{a_1\}, B_0 \cup \{a_1, a_2\}, \ldots, B_0 \cup \{a_1, a_2, \ldots, a_t\}\}$;

23. Set $T_1 := B_0 \cup \{a_1\}; T_2 := B_0 \cup \{a_1, a_2\}; \ldots; T_t := B_0 \cup \{a_1, a_2, \ldots, a_t\}$;

24. For j:= 1 to t do

25. Calculate the classification accuracy on T_j by using 10-fold classifier;

26. $B_{best} := T_{jo}$ in which T_{jo} has the highest classification accuracy;

Return B_{best};

3. EXPERIMENTAL RESULTS

Compared with two fuzzy rough set based incremental algorithm (IV-FS-FRS-2 , IARM) and two rough set based incremental algorithm (ASS-IAR , IFSA). Specifically, IV-FS-FRS-2 is a filter algorithm based on fuzzy discernibility matrix, while IARM is a filter algorithm based on relative discernibility relation. ASS-IAR is a filter algorithm based on active sample selection, while IFSA is a filter algorithm based on dependency function.

3.1 Data set

This subsection introduces experiments for evaluating the classification accuracy of IFW_FDAR_AdObj algorithm compared with two fuzzy rough set based incremental algorithm (IV-FS-FRS-2 [5,8], IARM [6,9]) and two rough set based incremental algorithm (ASS-IAR [7,10], IFSA [8,11]). Specifically, IV-FS-FRS-2 is a filter algorithm based on fuzzy discernibility matrix, while IARM is a filter algorithm based on relative discernibility relation. ASS-IAR is a filter algorithm based on active sample selection, while IFSA is a filter algorithm based on dependency function. Experiments are deployed on some benchmark datasets from UCI [9,12] as in Table 1.

For the algorithms IV-FS-FRS-2 and IARM by fuzzy rough set approach, all real value attributes are normalized into the values in interval [0, 1] on each dataset [8]:

$$a'(x_i) = \frac{a(x_i) - min(a)}{max(a) - min(a)} \qquad (1)$$

in which, max(a) and min(a) are the maximum and minimum of a. FER $\tilde{R}_a$ [8] on a is defined as:

$$\tilde{R}_a(x_i, x_j) = 1 - |a(x_i) - a(x_j)| \text{ where } x_i, x_j \in U \qquad (2)$$

For each attribute $a \in C$ with nominal or binary value, the FER R_a is defined where $x_i, x_j \in U$:

$$R_a = \begin{cases} 1, & a(x_i) = a(x_j) \\ 0, & otherwise \end{cases} \qquad (3)$$

On decision attribute $\{d\}$, we use FER $R_{\{d\}}$. For $x_i, x_j \in U$

$$R_{\{d\}} = \begin{cases} 1, & d(x_i) = d(x_j) \\ 0, & otherwise \end{cases} \qquad (4)$$

The partition $U/R_{\{d\}} = \{[x_i]_{\{d\}}\}$, where $x_i \in U$ and $[x_i]_{\{d\}} = \{x_j \in U | R_{\{d\}}(x_i, x_j) = 1\}$ is an equivalent class. Then, equivalent class $[x_i]_d$ is considered as a fuzzy equivalent class, denoted by $[x_i]_{\tilde{d}}$. The membership function is defined as $\mu_{[x_i]_{\tilde{d}}}(x_j) = 1$ if $x_j \in [x_i]_d$ and $\mu_{[x_i]_{\tilde{d}}}(x_j) = 0$ if $x_j \notin [x_i]_d$.

For the algorithms ASS-IAR and IFSA by traditional rough set approach, we use a fuzzy C-mean clustering (FCM) to discretize real-valued data before attribute reduction.

Each data set is divided into two approximately equal parts: original data set (Column 5 in TABLE 3.1) and incremental data set (Column 6 in TABLE 3.1). Original data set is denoted as U_0. Incremental data set is randomly separated into 5 equal parts, each part is denoted by U_1, U_2, U_3, U_4, U_5 respectively. To applying incremental algorithm IFW_FDAR_AdObj, IV-FS-FRS-2, IARM, ASS-IAR and IFSA, at first we perform this algorithm on original data set. Next, this algorithm is executed when sequentially adding from the first part to the fifth part of incremental dataset.

Table 3.1. Description of data sets when adding object set

ID	Data	Description	Number of objects	Original number of objects	Incremental number of objects	Number of condition attributes			Number of decision class
						All	Nominal value	Real-valued	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	Libra	Libras movement	360	180	180	90	0	90	15
2	WDBC	Wisconsin diagnostic breast cancer	569	284	285	30	0	30	2
3	Horse	Horse colic	368	183	185	22	15	7	2
4	Heart	Statlog (heart)	270	135	135	13	7	6	2
5	Credit	Credit approval	690	345	345	15	9	6	2
6	German	German credit data	1000	500	500	20	13	7	2
7	Cmc	Contraceptive Method Choice	1473	733	740	9	7	2	3
8	Wave	Waveform	5000	2500	2500	21	0	21	3

3.2. Computation time of IFW_FDAR_Adobj, IV-FS-FRS-2 IARM, ASS-IAR AND IFSA

Figure 3.1 shows that the execution time of IFW_FDAR_AdObj is higher than the execution time of IV-FS-FRS-2 and IARM on all datasets. Although the calculation of fuzzy distances in IFW_FDAR_AdObj is simpler than others measures in IV-FS-FRS-2, IARM,

ASS-IAR and IFSA, the algorithm IFW_FDAR_AdObj needs more time to run the classifier. The execution time of ASS-IAR is smallest because of the elimination of useless incomming samples in incremental computation.

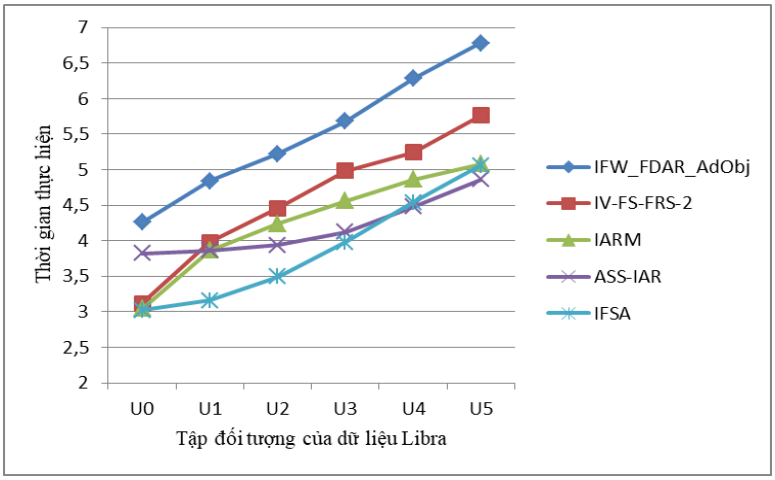

Figure 3.1. Time of IFW_FDAR_AdObj, IV-FS-FRS-2 IARM, ASS-IAR và IFSA on Libra (second)

3.3 Classification accuracy and reduct cardinality of IFW_FDAR_Adobj, IV-FS-FRS-2, IARM, ASS-IAR AND IFSA

We use CART classifier (CART – Classification And Regression Tree) to compute the classification accuracy in the wrapper stage of IFW_FDAR_AdObj. We also use CART classifier to compute the classification accuracy for IFW_FDAR_AdObj, IV-FS-FRS-2, IARM, ASS-IAR after attribute reduction. The 10-fold cross-validation technique is also used. We divide the dataset into 10 approximately equal parts. One part is selected randomly for testing, the others are used for training. This progress is repeated 10 times. We denote the accuracy of classification as $v \pm \sigma$ where v is the mean of 10 runs and σ is standard error. All experiments are installed on PC Core(TM) Intel (R) i7-3770CPU, 3.40 GHz, Windows 7 using Matlab.

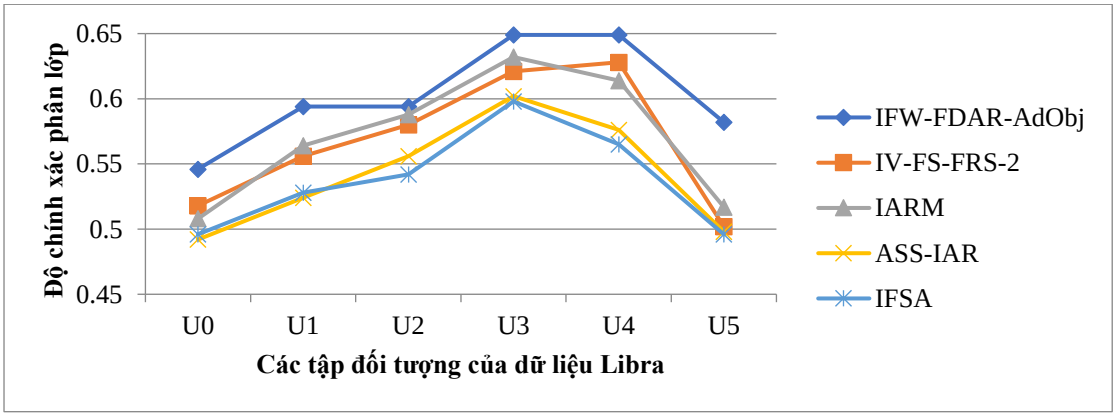

Figure 3.2. Classification accuracy of IFW_FDAR_AdObj, IV-FS-FRS-2, IARM, ASS-IAR và IFSA

The result of classification accuracy obtained by our algorithm are presented in Figure 3.2. As in this figure, for the cardinality of reduct at each incremental step, the proposed filter-wrapper algorithm IFW_FDAR_AdObj is much smaller than IV-FS-FRS-2, IARM, ASS-IAR and IFSA. As a result of this paper, the accuracy and the generality of classification rule set on reduct of IFW_FDAR_AdObj are better than those of IV-FS-FRS-2, IARM, ASS-IAR and IFSA. Moreover, because of the selection of the reduct with highest classification accuracy in wrapper stage, the classification accuracy of IFW_FDAR_AdObj is higher than IV-FS-FRS-2, IARM, ASS-IAR and IFSA on all data sets. The classification accuracy of IV-FS-FRS-2, IARM by fuzzy rough set approach is higher than that of ASS-IAR, IFSA by traditional rough set approach. For each data set, we can see that the classification accuracy does not increase when adding incremental data set. This is because there are some noise objects in incremental data sets that decrease the classification accuracy of learning algorithms.

4. CONCLUSION

The paper proposed a solution to find the reduct of the decision table according to the combined filter-wrapper approach in the case of adding object sets to minimize the number of reductive set attributes and improve improve the accuracy of the classification model. In this paper, IFW_FDAR_AdObj algorithm was introduced to solve the problem. The experimental results are compared with the other algorithms, that has shown that IFW_FDAR_AdObj algorithm is efficient. We will continue to study the incremental algorithms that find the reduct of the decision table in the case of adding and removing the attribute set in the future.

REFERENCES

1. Demetrovics, J., Thi, V.D., & Giang, N.L. (2014), *Metric Based Attribute Reduction in Dynamic Decision systems*, Annales Univ. Sci. Budapest., Sect. Comp, Vol. 42, 157-172.
2. Huong, N. T. L., &Giang, N. L. (2016), "Incremental algorithms based on metric for finding reduct in dynamic decision systems", *Journal on Research and Development on Information & Communications Technology*, Vol.E-3, No.9, 26-39.
3. Z. Pawlak (1991), *Rough sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publisher, London.
4. Q.H. Hu, D.R. Yu, Z.X. Xie (2016), *Information-preserving hybrid data reduction based on fuzzy-rough techniques*, Pattern Recognit. Lett. 27(5), pp. 414-423, 2016.
5. J.H. Dai, H. Hu, W.Z. Wu,Y.H. Qian, D.B. Huang (2018), "Maximal Discernibility Pairs Based Approach to Attribute Reduction in Fuzzy Rough Sets", *IEEE Transactions on Fuzzy Systems*, Vol. 26, Issue 4, pp. 2174-2187.
6. J.H. Dai, Q.H. Hu, H. Hu, D.B.Huang (2017), "Neighbor inconsistent pair selection for attribute reduction by rough set approach", *IEEE Transactions on Fuzzy Systems*, Vol. 26, Issue 2, pp. 937-950.
7. L.J.Ping, Z. W. Xia, T.Z. Hui, X.Y. Fang, M. T. Yu, Z.J. Jing, Z. G. Yong, J. P. Niyoyita (2020), "Learning with fuzzy rough set-based attribute selection", *Expert Systems with Applications*, Vol. 139, pp. 1- 17.
8. Y.Y. Yang, D.G. Chen, H. Wang, X.H. Wang (2017), "Incremental perspective for feature selection based on fuzzy rough sets", *IEEE Transactions on Fuzzy Systems*, Vol. 26, Issue 3, pp. 1257-1273,

9. Y.Y. Yang, D.G. Chen, H. Wang, Eric C.C.Tsang, D.L. Zhang, "Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving", Fuzzy Sets and Systems, Volume 312, pp. 66-86, 2017.
10. Y.Y. Yang, D.G. Chen, H. Wang (2017), "Active Sample Selection Based Incremental Algorithm for Attribute Reduction With Rough Sets", *IEEE Transactions on Fuzzy Systems*, Vol. 25, Issue 4, pp. 825-838.
11. W.H. Shua, W.B. Qian, Y.H. Xie (2019), "Incremental approaches for feature selection from dynamic data with the variation of multiple objects", *Knowledge-Based Systems*, Vol. Vol. 163, pp. 320-331.

VỀ THUẬT TOÁN GIA TĂNG RÚT GỌN THUỘC TÍNH KHI BỔ SUNG TẬP ĐỐI TƯỢNG TRONG BẢNG QUYẾT ĐỊNH THAY ĐỔI

Tóm tắt: *Trong mấy năm gần đây, các phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ đã thu hút sự quan tâm của các nhà nghiên cứu vì chúng nâng cao độ chính xác của mô hình phân lớp. Tuy nhiên, phần lớn các phương pháp đề xuất đều thực hiện trên bảng quyết định không thay đổi. Trong bài báo này, chúng tôi xây dựng thuật toán gia tăng tìm tập rút gọn xấp xỉ theo hướng tiếp cận kết hợp filter-wrapper. Kết quả thử nghiệm trên một số bộ số liệu mẫu cho thấy, thuật toán gia tăng đề xuất hiệu quả hơn một số thuật toán gia tăng khác theo tiếp cận filter về số lượng thuộc tính tập rút gọn và độ chính xác phân lớp.*

Từ khoá: *Tiếp cận tập thô mờ thuật toán gia tăng đề xuất, bảng quyết định, rút gọn thuộc tính.*