Dynamic Attribute Reduction from Multidimensional Data Based on Partitioning

Jin Zhou[1,2,a], Richard Irampaye[1,b], Xu E[1,2,c,*], Yunfeng Liu[1,d] and Yanhong Li[1,e]

[1]*College of information and technology, Bohai University, Jinzhou, PR China*
[2]*College of Food science and Engineering, Bohai University, Jinzhou, PR China*
a. zhoujin@bhu.edu.cn, b. 2117560583@qq.com, c. exu21@163.com
d. jzedulyf@bhu.edu.cn, e. zcfl613@126.com
**corresponding author: Xu E*

Keywords: Rough set, information update, multidimensional data, core attribute, partitioning, attribute reduction.

Abstract: In the real-world, some decision systems experience a dynamic variation of their attributes and attribute values over time. With this age of sensor technology and the Internet of Things (IoT), multidimensional data hard to process is generated. Some tools based on partitioning into small subsets relatively easy to process have been developed for the multidimensional data. However, they are statics and cannot be used in a changing environment. The attribute reduction is critical to handle. In this paper, we mainly focus on the update of the attribute reduction along with the time complexity improvement. The multidimensional data is divided into subsets, and then we compute the core for each subset and the global core is the union of all subsets' core in the system, after that the attribute reduction is computed. The new data type enters into the system as new subsets to fuse with the existing subsets. Any information for update will be processed on subset level. We may have scenarios such as adding, removing, adding and removing simultaneously along with the new data type to enter into the system. The updating process results in positive region change, which imply a computation of a new core and new attribute reduction, which will be done dynamically over time. We use the discernibility matrix method for computation. This algorithm avoids some re-computations by using existing subsets and core for those with the unchanged positive region. Some examples provided to illustrate the proposed algorithm.

1. Introduction

The rough set theory presented by Zdzislaw Pawlak[1,2,3] is a mathematical tool to deal with uncertainty. In the real world, they are many decision tables where attributes values may be incomplete and attributes and attributes values may change over time. As introduced by Pawlak, rough set theory allows us to process information without prior knowledge. M. Kryszkiewicz[4] proposed an extension of rough set theory using the tolerance relation for incomplete decision systems. Attribute reduction is an important step to eliminate irrelevant attributes and get the same representation of the knowledge. There are many approaches for attribute reduction in complete and incomplete decision tables in static and dynamic decision tables. Some reducts are based on entropy

[5,6], reduct based on positive region[7,8] reduct based on information quantity[9], reduct based on metric[10], reduct based on discernibility matrix[11,12]. A dynamic attribute reduction is ideal for the environment which varies over time such as financial evaluation, clinical decision, and some other real-time tasks. *Shu et al*.[13] worked on dynamic attribute reduction from incomplete decision systems with immerging and emerging objects. *Vu et al*.[14] presented a dynamic reduct computation from incomplete decision systems with an increasing and decreasing of attributes. As the volume data is increasing with the technology trend, the multidimensional data is generated and is hard to process .Based on the idea of portioning *Zhou et al*.[15] developed a multidimensional data fusion based on partitioning. However, this tool works in static environment and is not suitable for dynamic decision systems. In this paper we introduce the dynamic attribute reduction from multidimensional data based on partitioning. The dynamic update is performed on the subsets with scenarios such as adding, removing, adding and removing simultaneously and adding new data type. Using the discernibility matrix on subsets, we compute their core to get the global core which is the union of all subsets' core and then we compute the attribute reduction by adding most significant attribute from the remaining in order to keep the positive region of the information system. The partitioning method is effective to compute and process the multidimensional data.

2. Preliminaries

This section presents some concepts involved in this paper, and can be found in Refs [1,4,12].

Definition 1: An information system is the pair $IS = (U, A)$, where U is a universe of objects, A is a set of attributes, and for every $a \in A$, there is a mapping a, $a: U \rightarrow Va$, where Va is called the value set of a. In some situations, we may have Va which contains a missing value for at least one attribute $a \in A$, in that case, IS is called an incomplete information system, if not the case is complete. The missing value will be denoted by "*". An incomplete information system $IIS = (U, C \cup D)$ is an incomplete decision table where C is conditional attributes and D is the decision attributes set.

Definition 2: Let $IS = (U, A)$ be an incomplete information system. The tolerance relation on U for any attribute set $P \subset A$ can be defined as follows:

$SIM(P) = \{(u, v) \in U \times U \mid \forall a \in P, (a(u) = a(v)) \vee (a(u) = *) \vee (a(v) = *)\}$ (1)

$SIM(P)$ tolerance relation on U is reflexive and symmetric, but not necessarily transitive. $Sp(u)$ is called a tolerance class of U under P, which is the maximal set of objects that are possibly indistinguishable by P with u. $Sp(u) = \{v \in U \mid (u, v) \in SIM(P)$ (2)

$U/SIM(P)$ is the classification of U or the knowledge of U induced by P.

$U/SIM(P) = \{Sp(u) \mid u \in U\}$ (3)

It should be noted that the tolerance classes in $U/SIM(P)$ do not necessarily yield a partition of U. They form a cover of U in general.

Definition 3: Let $IIS = (U, C \cup D)$ be an incomplete decision system, regarded as subset (partition) from the multidimensional incomplete system. $P \subseteq C$, if $POS_P(D) = POS_C(D)$, and $POS_B(D) \neq POS_C(D)$ for any $B \subset P$, then P is a reduct (Red) of the IIS. The core attribute of the IIS is $Core_C = \bigcap Red(IIS)$ which contains all indispensable attributes of the IIS.

Example 1: The incomplete information as Table 1 showing several cars descriptions. $U = \{K_1, K_2, K_3, K_4, K_5, K_6\}$ is the universe of objects and $A = \{a_1, a_2, a_3, a_4\}$ is the condition attribute set with a_1,a_2,a_3 and a_4 stand for Price, Mileage, Size and Max-speed respectively.

Table 1: An incomplete information system.

Car	Price	Mileage	Size	Max-speed
K_1	High	High	Full	Low
K_2	Low	*	Full	Low
K_3	*	*	Compact	Low
K_4	High	*	Full	High
K_5	*	*	Full	High
K_6	Low	High	Full	*

By computing the positive region, one can get $POS_C(D)= \{K_1,K_2,K_3\}$, from the point of tolerance relation, $|S_C(K_1)/IND(A)= 1$,$|S_C(K_2)/IND(A)= 1$ and $|S_C(K_3)/IND(A)= 1$ which is not the case for the remaining objects. Using the discernibility matrix and discernibility function as proposed by A. Skowron [12] we get:

Table 2: Discernibility matrix.

	K_1	K_2	K_3
K_1			
K_2	a_1		
K_3	a_3	a_3	

The discernibility function *(DF), FSIS* $\{a_1, a_2, a_3, a_4\} = a_1 \wedge a_3 \wedge a_3 = a_1a_3$ after Boolean algebra simplification. This means we have $\{a_1, a3\}$ as one reduct and it is also the core from the positive region of the *IIS.* One can get a_4 more significant than a_2 for the *IIS* and then the union of the core with a_4 gives us $Red = \{a_1, a_3, a_4\}$ as a reduct of *IIS*, hence $POS_{Red}(A) = POS_C(A)$.

3. Dynamic Update on Subsets of the Incomplete Multidimensional Data System

In this section, we present some scenarios which may happen in the incomplete multidimensional data or multi-table incomplete information system. We have information arranged concerning the data type in Figure 1, and the dynamic attribute reduction computation process in Figure 2.

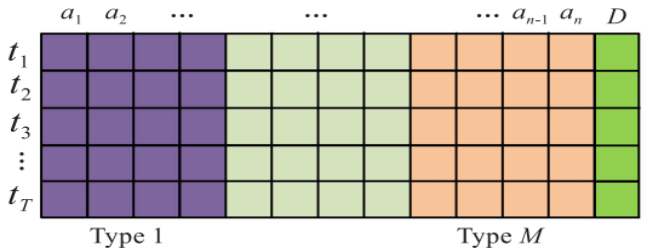

Figure 1: Problem modeling with regard to measurement types.

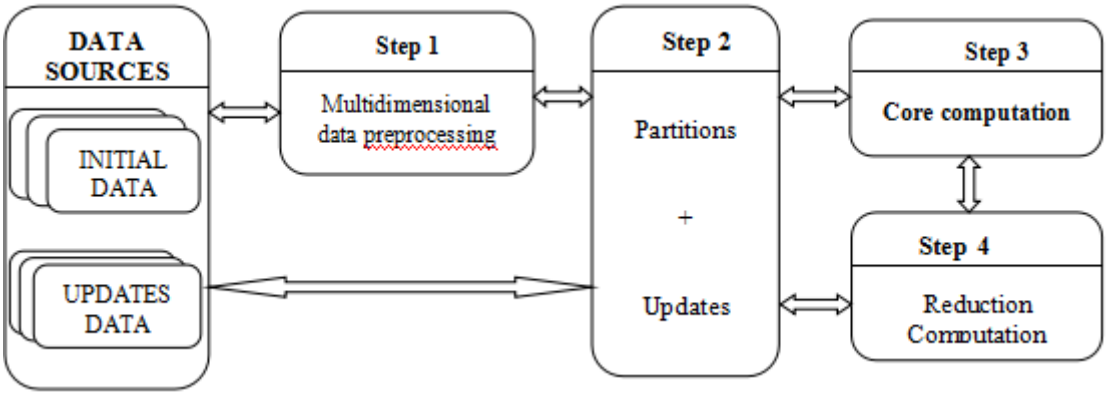

Figure 2: Flow diagram of the updating algorithm.

We have four steps with specifics tasks along the way until the new attribute reduction is computed. Below are the steps:

Step 1: The first step mainly focus on the identification and indexation of the data in the information system, according to some criteria the information system arranged in the way that similar data type and objects are classified together. On this step, the missing values are replaced with "*" to compute successfully the missing data. In the updating process, the new data identified to be similar with existing data will make change in the step 2 according to their indexes. The new data type will be passed on step 2 as a subset.

Step 2: The second step is to divide the multidimensional information system into relatively small subset by certain rules. If there is some new information or some new data enter the system after the initial input, all updating operations including the new data type in the system are performed on this level. Existing partitions get updated and new data type is computed as new data subset, which implies the creation of new partitions for new data type. As shown in Figure 1, the information system keep track of any change in order to get at different times an effective computation and then accurate results.

Step 3: On this step, we compute the core of every partition (granule) obtained on step 2 by computing the discernibility matrix and the discernibility function; we get the core by performing the intersection of all attributes reduction gotten on that partition. The Algorithm 1 shows the dynamic global core of the information system which is the union of all cores from all partitions.

Step 4: This step is the attribute reduction computation using the global core gotten in step 3. The goal is to get the attribute reduction of the conditional attribute *C* relative to *D*. The significance measure on remaining attribute allow us to select to most relevant attribute to add to the core and keep the classification ability of the information system with fewer attributes (As Shown in Algorithm 2). This level keeps a consistent checking for a new updated core on step 3 in order to compute the new attribute reduction.

The multidimensional incomplete information system *MIS* is partitioned in to sub- incomplete information systems *SIS*, $MIS = (Q_1, Q_2, Q_M)$, then $SIS = Q_j$ for convenience. Q_{ad}, Q_{rm} and Q_{nw} are respectively linked to information to add, to remove and new information as well. The ...*Up* is to denote the updated information, so Q_j becomes Q_{jUp}.

Proposition 1: Let $SIS = (Qj, C \cup D)$ be a sub-incomplete information system, for $B \subseteq C$, Q_{ad} is the multiple immigrating objects or attributes. Then the positive region $POS_B^{Qj \cup Qad}(D) = POS_B^{Qj}(D) \cup POS_B^{Qad}(D) - \{x_p \in X || S_B^{Qj \cup Qad}(x_p)/IND(D) \neq 1\} - \{u_k \in X || S_B^{Qj \cup Qad}(u_k)/IND(D) \neq 1\}$, where $X = \{x_p \in Q_j, u_k \in Qad\ (x_p, u_k) \in TR_B\}$, $S_B^{Qj \cup Qad}(x_p) = S_B^{Qj}(x_p) \cup \{u_k\}$, $S_B^{Qj \cup Qad}(u_k) \cup \{u_k\}$, $S_B^{Qj \cup Qad}(u_k) = S_B^{Qj}(u_k) \cup \{x_p\}$.

Example 2: (Continued from Example 1 with Table 1 as the original *SIIS*). Let $B \subseteq C$, two objects O_1, O_2 are immigrating into the system; simultaneously a missing value is available for the object K_4 which becomes K_{4Up}. $Qad = \{O_1, O_2, K_{4Up}\}$, where $O_1 = \{High, Low, Full, High\}$, $O_2 = \{High, Medium, Full, Low\}$ and $K_{4Up} = \{High, Low, Full, High\}$.

By computing, the positive region becomes $POS_C(U)=\{K_1,K_2,K_3,O_2\}$. One can get $a_1a_2a_3$, which means a_1, a_2, a_3 is the only reduct and core of the *SIS* in this case.

Example 3: (Continued from Example 1 with Table 1 as the original *SIIS*) Let $B \subseteq C$, two attribute a_5,a_6 stand for Internet and GPS respectively are simultaneously immigrating into the system, $Qad = \{a_5, a_6\}$, where $a_5 = \{No, Yes, No, No, Yes, No\}$ and $a_6 = \{No, *, No, *, Yes, No\}$.

We assume that $Qad < Qj$ to avoid getting subsets bigger than the normal (standard) partitions of *MIS*. In this case, the positive region in this case is $POS_C\ (U) = \{K_1, K_2, K_3, K_5\}$.we get three reducts $\{a_1, a_3, a_4\}$, $\{a_3, a_4, a_5\}$ and $\{a_3, a_4, a_6\}$ and $\{a_3, a_4\}$ is the core of the *SIS*.

Proposition 2: Let $SIS = (Qj, C \cup D)$ be a sub-incomplete information system, for $B \subseteq C$, $Qrm = \{e1, e2, ez\} \subset Qj$ is the multiple emigrating objects, then the positive region $POS_B^{Qj-Qrm}(D) = POS_B^{Qj}(D) - Qrm \cup \{x \in Y || S_B^{Qj-Qrm}(x) / IND\ (D)\ |1\}$, where $Y = \bigcup_{i=1}^{z} S_B\ (e_i) - Qrm$, and $S_B^{Qj-Qrm}(x) = S_B^{Qj}(x) - Qrm$.

Here, we assume that the existing subset is bigger than subset of objects to remove (emigrating) from it. $Qrm \subseteq Qj$.

Example 4: (Continued from Example 1 with Table 1 as the original *SIIS*) Let $B \subset C$, two objects *K4, K5* are simultaneously removed from the system, $Qrm = \{K_4, K_5\}$. In this case all remaining objects are discernible, then $POS_C(U)=\{K_1, K_2, K_3, K_6\}$. By computing, we get {a1, a3, a4} as the only reduct and hence the core of *SIS*. From Proposition 1 and Proposition 2, we get a mixed case with objects immigrating into the system along with objects emigrating from the system simultaneously.

Example 5: Consider Table 1 as the original sub-incomplete information system ***SIS***. Here we are adding the objects set $Qad = \{O_1 = \{High, Low, Compact, High\}, O_2 = \{High, Medium, Full, Low\}, O_3 = \{High, Low, Full, High\}\}$ and simultaneously removing objects set $Qrm = \{K_5, K_6\}$ from the system. The computation in the same way as in Example1 and Example 3 give us results as, the positive region is $POS_C(U) = \{K_1, K_2, K_3, O_1, O_2\}$, and $\{a_1, a_2, a_3, a_4\}$ is the reduct and core of the *SIS*.

Proposition 3: Let $MIS = (Qj, C \cup D)$ be a multidimensional incomplete information system. The new partitions (subsets) are created if new data type is coming into the system. $Qj \leftarrow Qnw$. In the example, *Qnw* contains several car drivers' descriptions as shown in Table 3.

Table 3: New data type regarded as new *SIS*.

Driver	D-License	Experience	French	Reference
X_1	B	Medium	Yes	Excellent
X_2	B	Low	Yes	Neutral
X_3	C	Low	Yes	Good
X_4	D	High	Yes	Neutral
X_5	D	Medium	Yes	Neutral
X_6	D	High	Yes	Excellent
X7	B	High	*	Good
X8	C	Low	No	Excellent

Example 6: Let this Table 6 be the *SIIS*, and with C_1, C_2, C_3, C_4 standing for Driving-License, Experience, French and Reference respectively. After computation we have the positive region

$POS_C(U)=\{X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8\}$ and the core is from the intersection of the two reducts $\{C_1,C_2,C_3,C_4\}$ and $\{C_2,C_4\}$ which give us $\{C_2,C_4\}$ as the core.

4. Dynamic Attribute Reduction Algorithm for Multidimensional Data based on Partitioning

In this section, we present our algorithm for the dynamic attribute reduction computation. The first phase focuses on global core computation and subsets update, the second phase compute the attribute reduction using the global core gotten from phase 1 and add some relevant attributes according to their significant measure. If a new global core is computed, the phase 2 is triggered to compute a new attribute reduction dynamically.

Phase 1

Algorithm 1: Dynamic global core computation algorithm from the multidimensional data

Input: Information system S = (Q_1, Q_2, …, Q_M), and Qad, Qrm, Qnw for data to add, remove or new data in the system.

Output: $Core_C^U$

1. Begin ;
2. $Core_C^U \leftarrow \varnothing$;
3. Compute POS_{Qj}^U;
4. If ∃ data or information for Update Qj , and $Q_j \leftarrow Q_{jUp}$ /* The update of Qj, if new data type, create new partitions Q_{M+f}, $and\ f \in \{1,2,\ldots,N\}$*/
5. Switch {data for update } do;
6. {Case 1{data to add}{$Q_{jUp} \leftarrow$ Qj ∪ Qad; break;}
7. Case 2{data to remove}{ $Q_{jUp} \leftarrow$ Qj - Qrm ; break;}
8. Case 3{data to add and data to remove}{$Q_{jUp} \leftarrow$ ((Qj ∪ Qad) - Qrm); break;}
9. Case 4{new data type}{ $Q_{jUp} \leftarrow$ Qnw ; break;}
10. Otherwise {Nothing to update}$Q_{jUp} \leftarrow$ Qj ; }
11. Compute POS_{QjUp}^U/*the updated subset positive region computed by Proposition 1 or Proposition 2. */
12. If $POS_{QjUp}^U = POS_{Qj}^U$, then $Core_{QjUp}^U \leftarrow POS_{Qj}^U$/*Use the existing core for that subsets, no change.*/
13. Else for $POS_{QjUp}^U \neq POS_{Qj}^U$ /*Change in the subset, compute again the core for the subset.*/
14. For every POS_{QjUp}^U, compute{Discernibility matrix DM and discernibility function DF */Method described in example1 and example 3*/
15. {If we get only one reduct Red_{QjUp} from DF, then is the core for the subset, $Core_{QjUp}^U = Red_{QjUp}$
16. else $Core_{QjUp}^U = \bigcap_{i=1}^{n} redQjup$ }}
17. Compute global updated core, $Core_C^U = U_{i=1}^N Core_{QjUp}^U$/*the global Core is the union of the core of the all subsets. */
18. Return $Core_C^U$;
19. End

Phase 2

Algorithm 2: Computation of the updated attribute reduction using the updated core

Input: Updated conditional attribute *C'* and the new $Core_{C'}^{U}$

Output: Attribute reduction

1. Begin;
2. $Core_{C}^{U} \leftarrow Core_{C'}^{U}$;
3. $Reduct \leftarrow Core_{C'}^{U}$;
4. Compute $POS_{C'}^{U}$(D) and POS_{Reduct}^{U}(D);
5. While POS_{Reduct}^{U}(D) $\neq POS_{C'}^{U}$(D) do
6. { For every attribute k ϵ (C' - Reduct);
7. {Compute *Sig (k, (C' - Reduct);* }
8. If Sig*(k, (C' - Reduct) > 0*, then *Reduct = Reduct* $\cup \{k\}$, /*Where k =(sig(k, (C' - Reduct)) > 0) until POS_{Reduct}^{U}(D) = $POS_{C'}^{U}$(D) */
9. Compute $POS_{Reduct}^{U}(D)$;}
10. Return *Reduct;*
11. End

5. The Algorithms Computational Time Comparison

The classical approach is to compute a discernibility matrix from the system and then process all computations to get the attribute reduction. For this perspective, the classical attribute reduction *(CAR)* algorithm has the time complexity O ($|C|^2|U|^2$). However, as nowadays we have an increasing of data in an unpredictable way, we have systems with high dimensions, big enough and hard to process with serial computation method. The idea of dividing into subsets, easy to process to get the global computation result is good for multidimensional data. As this approach of attribute reduction for the multidimensional data using partitions *(ARP)* shows his effectiveness, and the time complexity becomes O$\left(\frac{|C|2|U|2}{N}\right)$ because the conditional attribute *C* is segmented into sections. For the dynamic attribute reduction from multidimensional data based on partitioning algorithm *(DARP),* we have advantage on the fact that the partitioning process is done once which is not the case for the static algorithm *(ARP).* The dynamic global core computation time complexity is improved by a factor k, $0 < k \leq 1$ in comparison with the A*RP*. Then the *DARP* time complexity is O $\left(\frac{|C|2|U|2}{N}\right)$ k. The serial computation of this algorithm concerning the time complexity may have a no improvement, but has significant improvement for the parallel computation approach.

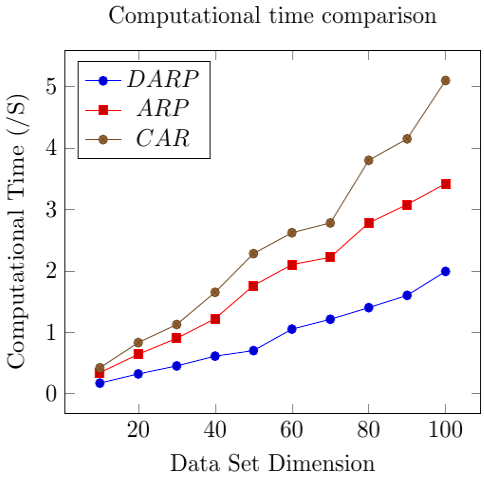

Figure 3: Computational time comparison.

6. Conclusions

In this paper we have shown a method for a dynamic attribute reduction computation from multidimensional data based on partitioning. The updating process is performed on subsets with for both conditional attribute and objects set in a multidimensional incomplete information system. Five scenarios have been covered in this paper: adding information, removing, adding and removing simultaneously, the new data type and no update case. The dynamic attribute reduction from multidimensional data based on partitioning algorithm avoids some re-computations and there is an improvement on the time complexity and the update issue has been tackled because the existing methods are statics and work only with the initial data. Further research is to develop an algorithm for extracting global rules directly from local rules for some multidimensional incomplete decision systems.

Acknowledgments

This work is funded by: National Key R & D Program of PR China under Grant No. 2019YFD0901605, and by National Natural Science Foundation of China under Grant No. 61602056, and by Natural Science Foundation of Liaoning Province of China under Grant No. 20170540005, and by Social Science Foundation of Liaoning Province of China under Grant No. L19BGL016, and by Doctoral Research Initiation Foundation of Liaoning Province of China under Grant No. 201601353, and by Scientific Research Foundation of Department of Education of Liaoning Province of China under Grant No. LZ2016005 & LQ2017002.

References

[1] Pawlak, Z. (1982) Rough Sets. International Journal of Computer and Information Science, 11, 341-356.
[2] Z. Pawlak, Rough sets and intelligent data analysis, Information Sciences, vol. 147, no. 124, pp. 1-12, 2002.
[3] Z. Pawlak, Rough sets theory and its applications, Journal of Communications and Information Technology, no. 3,pp. 7-10, Mar. 2002.
[4] M. Krysikiewicz, Rough set approach to incomplete information system, Information Sciences, vol. 112, nos. 1-4, pp. 39-49, 1998.
[5] D. Tian, X.J. Zeng, J. Keane, Core-generating approximate minimum entropy discretization for rough set feature selection in pattern classification, Int. J. Approx. Reason. 52 (2011) 863-880.

[6] J.Y. Liang, Z.Z. Shi, D.Y. Li, M.J. Wireman, The information entropy, rough entropy and knowledge granulation in incomplete information systems, Int. J. Gen. Syst.
[7] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, An efficient accelerator for attribute reduction from incomplete data in rough set framework, Pattern Recogn. 44 (2011)1658-1670.
[8] Z.Q. Meng, Z.Z. Shi, A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets, Inf. Sci. 179 (2009) 2774-2793.
[9] Guan Y. Y., H. K. Wang. Set-valued information systems. Information Sciences, 176 (2006), 2507–2525.
[10] Qian Y. H., C. Y. Dang, J. Y. Liang, D. W. Tang. Set-valued ordered information systems. Information Sciences, 179(2009), 2809-2832.
[11] Y. Zhao, Y.Y. Yao, F. Luo, Data analysis based on discernibility and indiscernibility, Inf. Sci. 177 (2007) 4959-4976.
[12] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, Intelligent Decision Support, 1992.
[13] W. Shu, W. Qian, (2015). An incremental approach to attribute reduction from dynamic incomplete decision systems in rough set theory. Data and Knowledge Engineering, 100, 116-132.
[14] Vu Van Dinh, Nguyen Long Giang, Vu Duc Thi, Generalized Discernibility Function Based Attribute Reduction in Incomplete Decision Systems.
[15] Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, Inf. Sci. 177 (2007) 41-73.
[16] J. Zhou, Hu, L. Wang, F. Lu, H. Zhao, K. (2013). An efficient multidimensional fusion algorithm for IoT data based on partitioning. Tsinghua Science and Technology, 18(4), 369-378.
[17] M. Kryszkiewicz, Rough set approach to incomplete information systems, Inf. Sci. 112 (1998) 39-49.
[18] Y.H. Qian, C.Y. Dang, J.Y. Liang, H.Y. Zhang, J.M. Ma, On the evaluation of the decision performance of an incomplete decision table, Data Knowl. Eng. 65 (3) (2008) 373-400.