Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

THA-A Hybrid Approach for Rule Induction System using Rough Set Theory, Genetic Algorithm and Boolean Algebra

Tribikram Pradhan[1], Harsh Anand[2] and Akul Goyal[3]

[1] Manipal Institute of Technology, Manipal University

Received: 14 December 2013 Accepted: 3 January 2014 Published: 15 January 2014

Abstract

The major process of discovering knowledge in database is the extraction of rules from classes of data. One of the major obstacles in performing rule induction from training data set is the inconsistency of information about a problem domain. In order to deal with this problem, many theories and technology have been developed in recent years. Among them the most successful ones are decision tree, fuzzy set, Dempster-Shafer theory of evidence. Unfortunately, all are referring to either prior or posterior probabilities. The rough set concept proposed by Pawlak is a new mathematical approach to inconsistent, vagueness, imprecision and uncertain data. In this paper we have proposed a hybridized model THA (Training dataset on hybrid approach) which combines rough set theory, genetic algorithm and Boolean algebra for discovering certain rules and also induce probable rules from inconsistent information. The experimental result shows that the projected method induced maximal generalized rules efficiently. The hybridized model was validated using the data obtained from observational study.

Index terms— rough set theory, genetic algorithm, mutation, crossover, boolean algebra.

1 Introduction

he major process of discovering knowledge in database is the extraction of rules from classes of data. In order to automate this problem, many inductive learning methods are introduced and applied to extract knowledge from databases such as decision tree learning [1], the technology employs a decision tree to achieve the learning function. The rough sets concept proposed by Pawlak in 1982 [2] is a new mathematical approach to imprecision, vagueness and uncertain. The rough sets philosophy is founded on the assumption that with every object of the universe of discourse we associated, some information objects characterized by the same information are indiscernible in the view of the available information about them. The indiscernibility relation is the mathematical basis of rough sets theory. Any set of all indiscernible objects is called an elementary set and forms a basis granule of knowledge about the universe [14] [18]. Any union of elementary Author ? ? ?: Department of Information and Communication Technology (ICT) Manipal University, Manipal, Karnataka, India. e-mails: tribikram.pradhan@manipal.edu, harshanand007@yahoo.co.in, akulgoyal4@gmail.com sets is referred to as a precise set, otherwise the set is rough. Each rough set has boundary line cases. With any rough sets a pair of precise sets-called the lower and the upper approximation of the rough sets is associated [7]. In recent 20 years, rough sets approach seems to be of fundamental importance to artificial intelligence (AI) and cognitive sciences and has been successfully applied many real life problems in medical diagnosis engineering [3], banking [3], finances [4] and others. By coupling rough sets theory with genetic algorithms (GA's), it is able to enhance search speed, induce decision rules from inconsistent information and this paper presents a hybrid approach that integrated rough sets theory, GA's and Boolean algebra for rule induction.

2 II.

Genetic Algorithms (ga's) GA's have been established as a viable technique for search, optimization, machine learning, and other problems. Theoretical developments by Holland and De Jong have laid the foundation of GA's [9]. GA's have been theoretically and empirically proven to provide robust search in complex space. The genetic algorithm consisting of a number of iteration process to make the population evolve [6]. Each iteration consists of the following steps:

? Selection: The first step consists of selecting individuals for reproduction [12]. This selection is done randomly with a probability depending on the relative fitness of the individuals so that best ones are often chosen for reproduction than poor ones.

? Reproduction: In the second step, offspring are bred by the selected individuals. For generating new chromosomes, the algorithm can use both recombination and mutation.

? Evaluation: Then the fitness of the new chromosomes is evaluated.

? Replacement: During the last step, individuals from the old population are killed and replaced by the new ones. Then, the genetic algorithm loops over an iteration process to make the population evolve. Figure 1 depicts the life cycle of GA's.

3 T Figure 1 : Genetic algorithm cycle

Finally, the reproduction step involves the creation of offspring chromosomes by using two genetic operators-mutation and cross-over. The most important part of the genetic algorithm is cross-over where we have to randomly select a cross site and swaps the genes of two parent chromosomes to produce two new offspring chromosomes. This can be easily represented using a pair of chromosomes encoded with two binary strings and cross site is denoted by "|". V is the domain of each attribute of A. F is a total function that defines the following application U x A? V.

Definition 2 : In an information system, S the attributes A are further classified into disjoint sets of condition attributes C and decision attributes D.

i.e. ? = C?D and Ø = C?D.

All the information represented by a table known as Information System which encompasses a number of rows and columns corresponding to the number objects and attributes. b) The set BNr (X) = R*(X) -R.(X) will be referred as the R-boundary region of X. If the boundary region of X is the empty set, i.e. BNr(X) = Ø, then the set X will be called crisp with respect to R; in the opposite case, i.e. if BNr(X) != Ø, the set X will be referred as rough with respect to R. c) In the same way, POSr (X) and NEGr (X) are defined as follows:

? NEGr (X) = U -R*(X) certainly non-member of X

? POSr (X) = R.(X) Year 2014 I THA-A Hybrid Approach for Rule Induction System Using Rough Set Theory, Genetic Algorithm and Boolean AlgebraObject U Condition Attribute Decision Attribute A B C x 1 1 0 0 x 2 0 1 1 x 3 1 1 0 x 4 0 0 0 x 5 1 0 1 x 6 0 0 1 x 7 1 1 1

IV.

4 Tha: The Hybridized Model

There are so many inductive learning systems such as ID3, ID4 and ID5 are not capable of handling inconsistent information about training data set effectively. After that Grzymala-Busse designed one system called as LERS which can deal with inconsistent information as well as training data set. But in LERS it's very difficult to maintain a huge training data set. And also the rules induced by LERS are very complicated and very difficult to understand. So in this paper we have proposed a hybridized approach known as THA (Training Data Set on Hybrid Approach), which a combination of Rough set theory, Genetic Algorithm and Boolean algebra.

? Rough Set Theory can handle inconsistent training data set and also missing values.

? Genetic Algorithm based search engine can induce probable decision rules.

? Finally, Boolean operations can simplify the probable decisions rules. The framework for the hybridized approach THA is depicted in Figure 3. Basically, it consists of 4 modules such as selection of raw data, rough set analyzer, performance evaluation of genetic algorithm, simplification by Boolean operation to generate rule.

The knowledge collected from the process or experts is forwarded to the hybridized model, THA for classification and generation of rules. After the completion of approximation analysis genetic algorithm will take both certain and possible data set. Generally genetic algorithm performs reproduction, cross-over and mutation to extract certain and possible rules from the training set. Finally, Boolean operations are used to simplify the probable decision rules generated by genetic algorithms. The Boolean operators such as union and intersection are used to simplify the rules. During these operations, redundant rules are to be removed, whereas related rules are to be clustered and generalized during simplification. For every possible rule we have to identify the reliability index, which is defined as the ratio of the number of observations that are correctly classified by possible rules and the number of observation whose condition attributes are covered by the same rule in the original training data set.

5 Reliability index =

Where, observation of possible rule is the number of observations that are correctly classified by a possible rule and observation of original data is the number of observations with the condition attribute covered by the same rule in the original data set. Now we can easily classify the inconsistent training data set correctly.

V.

6 Experimentation

THA is validated from the data obtained from a case study which monitors water quality parameters for drinkable water standards. The data in the Table 2 shows different parameters along with their range of values suitable, based on which we can determine the quality of water. For simplicity, we have taken two or three values for each water quality parameters. In the above table D stands for drinkable and UD stands for undrinkable.

In order to process the information, we need to depict the parameters in the form of integers. This is done using the following descriptor scheme show in Table4.

7 Global Journal of Researches in Engineering

8 Table 4 : Transformation scheme

The transformed result is shown in Table **??**.

9 Table 5 : Transformation water quality parameters for drinkable water standards

Observation Colour pH value Sulphur Turbidity Fluoride State1 1 1 1 1 0 0 2 0 2 0 0 1 1 3 0 2 2 0 1 1 4 1 2 2 1 1 0 5 0 0 1 1 0 0 6 0 2 0 0 1 0 7 1 0 2 0 0 0 8 1 1 2 0 0 1 9 1 2 2 1 1 1 10 0 1 0 1 1 0 11 1 1 1 1 0 0 12 0 2 2 0 0 1 13 0 0 2 1 1 0 14 1 1 1 1 1 0 15 0 0 0 0 1 1 16 1 2 2 1 1 1 17 1 1 1 1 0 1 18 0 2 0 0 1 0

As observed from the above table, the observation (1,11,17), (2,6,18) and (4, 9, 16) contradict each other, hence we need to perform approximation and concept forming through rough set analyzer. The two decision states, are characterized by C 1 (state=drinkable) and C 2 (state=non drinkable). Applying the rough set theory, we calculate the lower and upper approximations for the concept C 1 and C 2 . As observed from the above table 5, concepts C 1 and C 2 are represented by the following sets: Now we have to encode the coded data of table 5 by using such a scheme which is called Chromosomes encoding scheme.

Chromosome coding for water parameters are given below.

10 quality parameters

The GA based search engine is then used to extract rules from certain possible training data set obtained by rough set analyzer then we have to randomly generate 120 chromosomes to form an initial population of possible solution i.e., chromosomes. These chromosomes are coded using the scheme depicted in Table 6.

For chromosome represented the corresponding hybrid approach follows traditional binary string representation and its corresponding crossover and mutation operators. Using this scheme each chromosome is expressed as a binary string i.e., a string containing ”0” and ”1”.

After using the scheme a classification rule can be easily represented by 16 bit chromosome. For example, If (colour>=15) and (pH value>=7.5) and (sulphate>=300) and (turbidity>=8) and (fluoride>=1) then state will be drinkable can be coded as 0110110110110010.

Other than choosing a good schema for chromosome representation, it is important to define a reasonable fitness function that rewards the kind of chromosomes. Basically the purpose of the GA’s is to extract rules that maximize the probability of classifying the objects correctly. Thus, the fitness value of a chromosome can be described by its reliability, or in other words, the probability to classify objects in a training data set correctly. Mathematically, the fitness function used in this work is expressed as( )**2**

For example, if a rule (representation of chromosome) can correctly classify five objects in a training data sets, an if there are six objects having the same condition attributes-value pairs as the said rules, then the fitness value of this chromosome is (5/6)2 =0.6944. The square operator appeared in the fitness function is to ensure rapid convergence. It is used to suppress bad chromosomes with low fitness scores and promotes the creation of good chromosomes with high fitness scores. Thus the above fitness function favours rules that can classify objects correctly. Furthermore, it also satisfies more consistency and completeness criteria, which are of great importance to the evaluation of the rule. A rule is said to be consistent if it covers no negative sample, this is, no object in the training data set violating the rule; and it is said to be complete if the rule is able to cover all the positive sample that satisfy the condition of the rule in the training data set. As previously mentioned, after evaluating the fitness values of chromosomes with above average fitness values are selected for reproduction. As for cross-over and mutation, the respective probabilities are fixed at 0.85 and 0.01. With a higher probability of cross-over, offspring chromosomes that maintain the genetic traits of the parent chromosomes can be generated easily. This allows chromosomes with higher fitness values, that is, better solutions, to be discovered. A lower probability of mutation prevents the search for optimal solutions todegenerate into a random one.

The rule set induced by GA based search engine may contain rules with identical fitness values. Some of these rules can be combined to form a more general or concise rule using Boolean algebra. The rule pruner is assigned to detect and solve the redundancy problem.

If colour>=1 and pH >=1 and fluoride>=1 then state will be drinkable.

If colour>=0 and pH >=1 and fluoride>=0 then state will be drinkable.

Here from the Boolean algebra point of view Rule1 is a subset of Rule2. And the resultant rule will be If colour>=0 and pH >=1 and fluoride>=0 then state will be drinkable. Two sets of rules are available. As already mentioned, the value recorded in the parentheses following each certain and possible rule represents the completeness and the reliability indices, respectively. All the indices are represented in fraction form, with the numerator corresponding to the number of correctly classified observations whose condition attributes are covered by the rule. Analysis shows that the rules induced by THA are simple, reasonable and logical.

11 VI.

12 Induced Rule Set

Possible Rules:

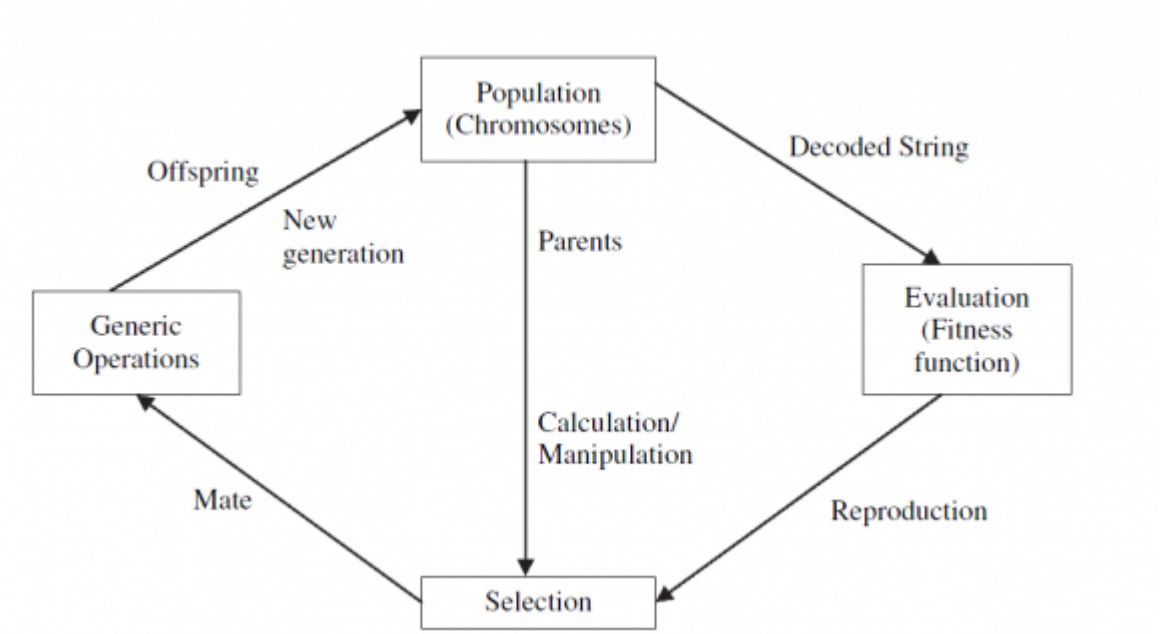
13 Discussion

The above experimentation shows that the hybridized approach THA is able to induce rule under uncertainty. We can infer that the hybrid approach can be used for inductive learning under uncertainty. This hybrid approach uses the strength of rough set theory along with efficient GA based search engine and Boolean algebra. In the above experimentation, the GA based search engine reaches its saturation within 60 generations for both possible and certain data sets. This hybrid approach is compared with other inductive learning technique and the result is shown in Table 7. The certain rules that are generated by this approach are identical to those produced by ID3. Finally the rules generated by this system are simple and concise as compared to those produced by LERS. In Figure 5, we have taken the population size as 120 and the number of generation as 60. In every generation we have identified the average fitness function and plotted the graph accordingly. And based on the average fitness function we have selected the chromosomes with higher fitness values. In this experiment, we have selected those chromosomes which have fitness function values more than 0.68. We have discarded all the chromosomes whose fitness function is less than 0.68. [1]

2

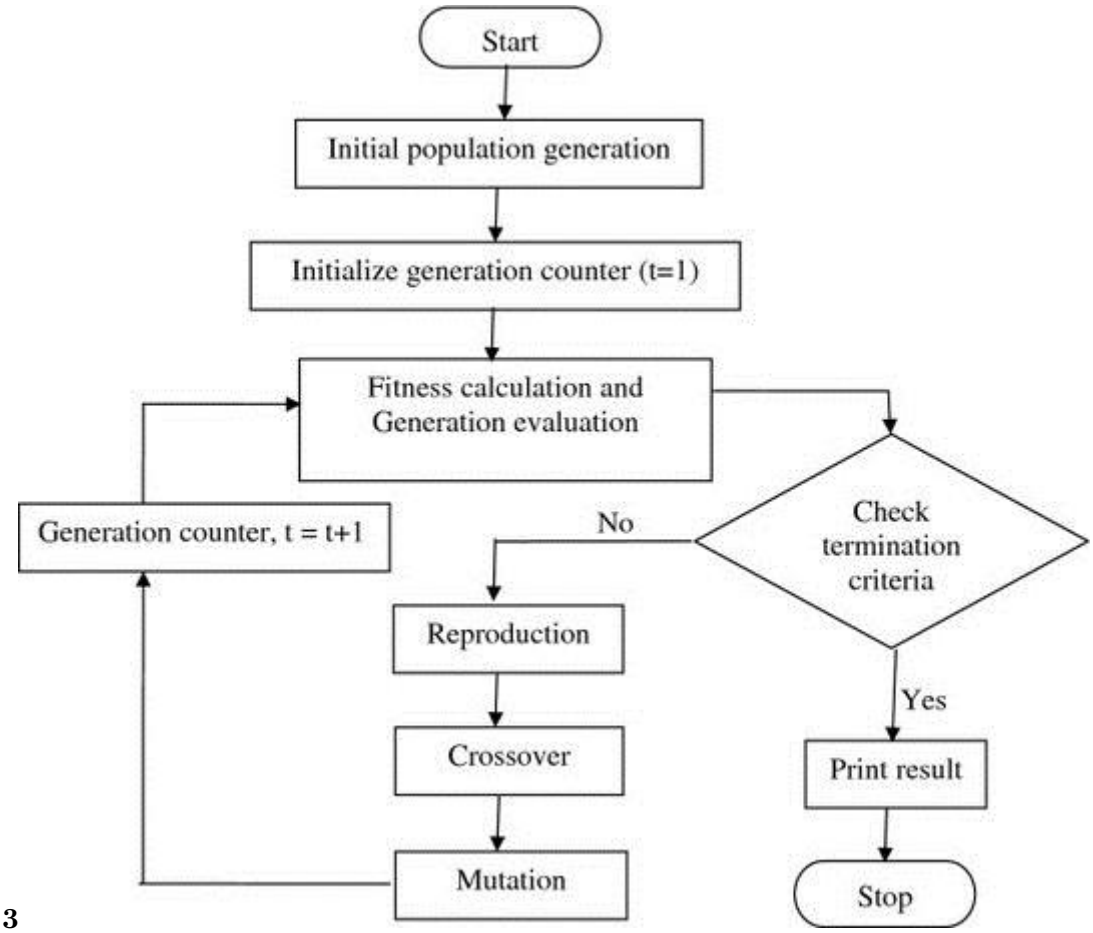
Figure 1: Figure 2 :

[1] © 2014 Global Journals Inc. (US)



314

Figure 2: Definition 3 : 1 Definition 4 :



3

Figure 3: Figure 3 :

4

Figure 4: DataFigure 4 :

1

Figure 5: C 1 =

Figure 6: I

1

Figure 7: Rule 1 :

5

Figure 8: Figure 5 :?

1

In Table 1 there are 7 objects, 2 conditions attribute A and B and 1 decision attribute C. For example, in object x 5 the condition attributes are 1 and 0, and its decision attribute is 1. Imprecise information causes indiscernibility of objects. This indiscernibility relation is called equivalence relation on the set of object U. In our above example a conflict (inconsistency) exists between object x 1 and x 5 because they are indiscernible by condition attribute A and B and have different decision attribute (C).
Equivalence classes of relation are called Elementary set in S. Any finite union of elementary set is known as Definable Set. The decision elementary set are called concepts. For example, the above table shows the decision attribute is having 2 types of values which is 0 and 1. Hence, 2 types of concept will be shaped. For example, C 1 and C 2 .
C 1 is having the decision attribute 0 and C 2 is having values 1.
C 1 ={x1, x3, x4}
C 2 ={x2, x5, x6, x7}

Figure 9: Table 1 :

2

Condition	Range	Values
Colour	[5-25]	(12, 15)
P.H. Value	[6.5-8.5]	(7, 7.5, 8)
Sulphate	[200-400]	(220, 300, 375)
Total Hardness	[300-600]	(300, 400)
Turbidity	[5-10]	(6, 8)
Fluoride	[1.0-1.5]	(1.0, 1.2)

Figure 10: Table 2 :

3

Observation	Colour	pH value	sulphur turbidity	Fluoride	state	
1	15	7.5	300	8	1	D
2	12	8	200	6	1.2	UD
3	12	8	375	6	1.2	UD
4	15	8	375	8	1.2	D
5	12	7	300	8	1	D
6	12	8	200	6	1.2	D
7	15	7	375	6	1	D
8	15	7.5	375	6	1	UD
9	15	8	375	8	1.2	UD
10	12	7.5	200	8	1.2	D
11	15	7.5	300	8	1	D
12	12	8	375	6	1	UD
13	12	7	375	8	1.2	D
14	15	7.5	300	8	1.2	D
15	12	7	200	6	1.2	UD
16	15	8	375	8	1.2	UD
17	15	7.5	300	8	1	UD
18	12	8	200	6	1.2	D

Figure 11: Table 3 :

6

Bit 3,6,9,12,15; operator	0=Less than or equal to(?) 1=Greater than or equal to (?)
Bit 1 and 2:Colour	00=12 01=15
Bit 4 and 5: PH value	00=7 01=7.5 10=8
Bit 7 and 8: Sulphate	00=200 01=300 10=375
Bit 10 and 11:Turbidity	00=6 01=8
Bit 13 and 14:Flouride	00=1 10=1.2
Bit 16:State	0=D 1=UD

Figure 12: Table 6 :

7

Technique	Dealing with uncertainty and inconsistency	Simple and concise rule induction	Extracting complete rules
ID3	No	Yes	For consistent data set
LERS	Yes	No	Not evaluated
RClass	Yes	Yes	No
THA	Yes	Yes	Yes

Figure 13: Table 7 :

[Global Journal of Researches in Engineering] , *Global Journal of Researches in Engineering*

[Springer ()] , Springer . 2000. (to appear)

[Grant ()] *Churn modeling by rough set approach*, J Grant . 2001. (manuscript)

[Wang et al. ()] 'Comparison of rough-set and statistical methods in deduce learning'. S K M Wang , W Ziarko , Y R Li . *international Journal of Man-Machine Studies* 1986. 24 p. .

[Pawlak ()] 'Decision rules, Bayes' rule and rough sets'. Z Pawlak . *New Di-rection in Rough Sets, Data Mining, and Granular-Soft Computing*, A Zhong, S Skowron, Ohsuga (ed.) 1999. Springer. p. .

[Zadeh et al. ()] 'Discovering Attribute Relationships, Dependencies and Rules by using Rough Sets'. L A Zadeh , W Ziarko , N Shan . *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, (the 28th Annual Hawaii International Conference on System SciencesWailea-USA; New Jersey-USA) 1965. 1995. Jan. 3-6, 1995. IEEE Press. p. . (Fuzzy Sets, Information and Control)

[Francis et al. ()] 'Economic and financial prediction using rough set model'. E H Francis , Lixiang Tay , Shen . *European Journal of Operational Research* 2002. 141 p. .

[Mckee and Lensberg ()] 'Genetic programming and rough sets: A hybrid approach to bankruptcy classification'. Thomas E Mckee , Teje Lensberg . *European Journal of Operational Research* 2002. 138 p. .

[Quinlan ()] 'Improved use of continuous attributes in C4.5'. I Ross Quinlan . *Journal of artificial intelligence research* 1996. 4 p. .

[N. Zhong, A. Skowron, and S. Ohsuga (ed.) ()] *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing*, N. Zhong, A. Skowron, and S. Ohsuga (ed.) 1999. Springer.

[Pawlak (2001)] 'New look Bayes' theorem -the rough set outlook'. Z Pawlak . *Proc. Int. RSTGC-2001*, (Int. RSTGC-2001Matsue Shimane, Japan) May 2001. 1/2, 2001. 5 p. .

[Polkowski et al.] L Polkowski , S Tsumoto , T Y Lin . *Rough Set Methods and Applications -New Developments in Knowledge Discovery in Information Systems*,

[Guan and Bell ()] 'Rough computational methods for information system'. J W Guan , D A Bell . *Artificial intelligence* 1998. 105 p. .

[S. K. Pal and A. Skowron (ed.) ()] *Rough Fuzzy Hybridization*, S. K. Pal and A. Skowron (ed.) 1999. Springer.

[Wei ()] 'Rough Set based Approach to Selection of Node'. J M Wei . *International Journal of Computational Cognition* 1542-8060. 2003. 1 (2) p. .

[Swiniarski and Skowron ()] 'Rough set methods in feature selection and recognition'. Roman W Swiniarski , Andraej Skowron . *Pattern Recognition Letters* 2003. 24 p. .

[Pawlak ()] 'Rough Sets'. Pawlak . *international Journal of information and computer sciences* 1982. 11 p. .

[Pawlak ()] *Rough Sets -Theoretical Aspects of Reasoning about Data*, Z Pawlak . 1991. Boston, London, Dordrecht: Kluwer.

[L. Polkowski and A. Skowron (ed.) ()] *Rough Sets and Current Trends in Computing. Lecture Notes in Artificial Intelligence 1424*, L. Polkowski and A. Skowron (ed.) 1998. Springer.

[Pawlak ()] 'Rough sets theory and its application to data analysis'. Pawlak . *Cybernetic And Sysrems* 1998. 29 p. .

[Shen and Jensen ()] 'Rough Sets, Their Extensions and Applications'. Q Shen , R Jensen . *Set Theory and Logic*, (Mineola-USA) 2007. Jul. 2007. 1979. Dover Publications. 4 p. . (Stoll, R.R.)

[Xie et al. (2004)] 'RST-Based System Design of Hybrid Intelligent Control'. G Xie , F Wang , K Xie . *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics*, (the 2004 IEEE International Conference on Systems, Man and CyberneticsThe Hague-The Netherlands; New Jersey-USA) 2004. Oct. 10-13, 2004. IEEE Press. p. .

[Wu et al. ()] C Wu , Y Yue , M Li , O Adjei . *The Rough Set Theory and Applications, Engineering Computations*, 2004. 21 p. .