Advances and Applications in Mathematical Sciences
Volume 21, Issue 10, August 2022, Pages 5655-5664
© 2022 Mili Publications, India

IDENTIFICATION OF EFFICACIOUS ALGORITHM USING ROUGH SETS

NIRMALA REBECCA PAUL and R. SANGEETHA

[1]Department of Mathematics
Lady Doak College, Madurai, India
E-mail: nirmalarebeccapaul@ldc.edu.in

[2]Department of Computer Science
Lady Doak College, Madurai, India
E-mail: rsangeetha@ldc.edu.in

Abstract

Rough sets are defined to deal with vagueness. Rough set theory has been applied by many researchers which paved the way for its development. Rough sets are defined in terms of approximations. In this paper an algorithm is developed to find the core of an information system. Its efficiency is determined by comparing itself with already developed algorithm in terms of time complexity using Big O Notations. Two examples are tested using the algorithm to find the core of an information system and an incomplete information system.

1. Introduction

Rough sets introduced by Pawlak [5] kindled the interest of many researchers to apply the same in their research areas. Rough sets are defined in terms of approximations similar to interior and closure operations in a topology. Hence rough topology is introduced. Fuzzy sets and rough sets are introduced to deal with uncertainty. The difference between the fuzzy sets and rough sets is that the rough sets have precise boundaries. Rough sets are used for information system with incomplete and insufficient information. Any information system consists of several attributes and it is necessary to find the minimal attributes for the classification of objects. Thus rough sets

2020 Mathematics Subject Classification: Primary 05A15; Secondary 11B68, 34A05.

Keywords: Rough sets, rough topology, lower approximation, upper approximation, accuracy of approximation and Core.

Received April 3, 2022; Accepted April 13, 2022

are used for knowledge reduction problems in information systems. In this paper an algorithm is developed to find the minimum number of attributes called core. This algorithm makes use of the coefficient called accuracy of approximation. The time taken to find the core of the system using the new algorithm and already developed algorithm is compared and it is proved that the new algorithm called accuracy algorithm takes less time to find the core of an information system.

2. Preliminaries

Definition 2.1[5]**.** Let U be a non-empty finite set of objects called the universe and R be an equivalence relation on U named as the indiscernibility relation. Then U is divided into disjoint equivalence classes. Elements belonging to the same equivalence class are said to be indiscernible with one another. The pair (U, R) is said to be the approximation space. Let $X \subseteq U$.

(i) The lower approximation of X with respect to R is the set of all objects, which can be for certain classified as X with respect to R and it is denoted by $L_R(X)$. $L_R(X) = \cup_{x \in U}\{R(x) : R(x) \subseteq X\}$, where $R(x)$ denotes the equivalence class determined by x.

(ii) The upper approximation of X with respect to R is the set of all objects, which can be possibly classified as X with respect to R and it is denoted by $U_R(X)$. $U_R(X) = \cup_{x \in U}\{R(x) : R(x) \cap X \neq \phi\}$.

(iii) The boundary region of X with respect to R is the set of all objects, which can be classified neither as X nor as not X with respect to R and it is denoted by $B_R(X)$. That is $B_R(X) = U_R(X) - L_R(X)$.

Definition 2.2 [5]**.** In an information system, not all condition attributes depict the decision attribute. The decision depends not on the whole set of condition attributes but on a subset of it is called the CORE.

Definition 2.3 [2]**.** An information system is of the form $(U, A, \{V_a\}, f_a)$ where U is a non-empty finite set of objects, called the universe, A is a finite non-empty set of attributes, V_a is the attribute value set of an attribute $a \in A$ and $f_a : U \to V_a$ is called the information function. If $f_a(x)$ is equal to a missing value for some $x \in U$ and $a \in A$, then the information system is called an incomplete information system (IIS) otherwise it is a complete information system (CIS). A missing value is denoted by "*". That is an IIS is of the form $(U, A, \{V_a\}, f_a)$ where $a \in A$ and $* \in \cup V_a$. An IIS can also be denoted by (U, A).

Definition 2.4 [2]. Let U be a universe and A be a finite set of attributes. For any subset B of A, there is a binary relation on U corresponding to B given by $R(B) = \{\{x, y\} \in U \times U : f_a(x) = f_a(y)$ or $f_a(x) = *$ or $f_a(y) = *$ for any $a \in B\}$. Then $R(B)$ is a tolerance relation on U (reflexive and symmetric). $S_B(x)$ denotes the maximal set of objects which are possibly indiscrenible with x by the tolerance relation on U. That is $S_B(x) = \{y \in U : (x, y) \in R(B)\}, x \in U$.

Definition 2.5 [5]. The accuracy of approximation is defined as $\alpha = \frac{|L_R(X)|}{|U_R(X)|}$ where $|X|$ denotes the cardinality of X.

Definition 2.6 [3]. Let U be a universe and R be an equivalence relation on U and $\tau_R(X) = \{\phi,. U, B_R(X)\}$ where $X \subseteq U$. $\tau_R(X)$ forms a topology on U called the rough topology on U with respect to X.

Example 2.7. Let $U = \{a, b, c, d, e\}$, $U/R = \{\{a, c\}, \{b, e\}, \{d\}\}$, $X = \{a, b, d, e\}$, $L_R(X) = \{b, d, e\}$, $U_R(X) = \{a, b, c, d, e\}$, $B_R(X) = \{a, c\}$. The rough topology on X is $\tau_R(X) = \{\phi, U, \{a, c\}\}$.

Remark 2.8. The rough topology will have two elements ϕ, U when $B_R(X) = \phi$ otherwise it will have the three elements $\phi, U, B_R(X)$ when $B_R(X) \neq \phi$.

Algorithm 1 [3]. This algorithm was developed to find the deciding factors or core of an information system.

3. Accuracy Algorithm

In this section a new algorithm called accuracy algorithm is developed in terms of accuracy of approximations to find the core of an information system.

Step 1. Select a subset of U and find the equivalence relation corresponding to all condition attributes.

Step 2. Find the lower approximation, upper approximation and accuracy of approximation.

Step 3. Remove one condition attribute and find the equivalence relation corresponding to the remaining attributes. Find the accuracy of approximation.

Step 4. Repeat step 3 for all condition attributes.

Step 5. Find the subset of attributes in C for which the accuracy of approximations found in step 3 and 4 are equal to the accuracy of approximation found in step 2. The set of attributes for which the accuracy of approximations is equal to the accuracy of approximations in step 2 gives the core of the system.

Example 3.1. Corona virus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. People infected by the virus will experience mild to severe respiratory problems. It can spread from an infected persons mouth or nose when they cough, sneeze, speak, sing or breathe. The following table gives symptoms of eight of patients suffering from Covid. The most common symptoms of Covid are Loss of taste or smell (L), Breathing difficulty (B) and Temperature (T)

Patients	Loss of taste(L)	Breathing Difficulty(B)	Nausea(N)	Temperature(T)	Covid
1	Yes	Yes	Yes	High	Yes
2	Yes	Yes	No	High	Yes
3	Yes	No	No	High	No
4	No	No	No	Very High	No
5	Yes	Yes	No	Very High	Yes
6	Yes	Yes	No	Very High	Yes
7	No	Yes	Yes	High	No
8	Yes	Yes	No	No	No

Here "no" represents that the patient does not have the symptom and "yes" represents that the patient has the symptom. The columns of the table represent the symptoms for Covid and the rows represent the patients. The entries in the table are the attribute values. The given information system is given by (U, A) where $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $A = \{L, B, N, T, Covid\}$ which is divided into a set of C of condition attributes given by $C = \{L, B, N, T\}$ and the decision attribute $D = \{Covid\}$.

Step 1. Here U is the set of patients, X represents the set of patients with Covid and C represents the set of condition attribute. $X = \{1, 2, 5, 6\}$, $C = \{L, B, N, T\}$, $U/R(C) = \{\{1\}, \{2\}, \{5, 6\}, \{3\}, \{4\}, \{7\}, \{8\}\}$. Let the accuracy of approximation be denoted as A.

Step 2. $L_R(X) = \{1, 2, 5, 6\} = U_R(X)$, $A = 1$.

Step 3. If the attribute loss of taste is removed from the condition attributes then the equivalence relation corresponding to $C_1 = \{B, N, T\}$ is $U/R(C_1) = \{\{1, 7\}, \{2\}, \{3\}, \{5, 6\}, \{8\}\}$, $L_R(X) = \{2, 5, 6\}$, $U_R(X) = \{1, 2, 5, 6, 7\}$, $A = 3/5 = 0.6$.

Step 4. If the attribute breathing difficulty is removed from the condition attributes then the equivalence relation corresponding to $C_2 = \{L, N, T\}$ is $U/R(C_2) = \{\{1\}, \{2, 3\}, \{4\}, \{5, 6\}, \{7\}, \{8\}\}$, $L_R(X) = \{1, 5, 6\}$, $U_R(X) = \{1, 2, 3, 5, 6\}$, $A = 3/5 = 0.6$.

If the attribute nausea is removed from the condition attributes then the equivalence relation corresponding to $C_3 = \{L, B, T\}$ is $U/R(C_3) = \{\{1, 2\}, \{3\}, \{4\}, \{5, 6\}, \{7\}, \{8\}\}$, $L_R(X) = \{1, 2, 5, 6\} = U_R(X)$, $A = 4/4 = 1$. If the attribute temperature is removed from the condition attributes then the equivalence relation corresponding to $C_4 = \{L, B, N\}$ is $U/R(C_4) = \{\{1\}, \{2, 5, 6, 8\}, \{3\}, \{4\}, \{7\}\}$, $L_R(X) = \{1\}$, $U_R(X) = \{1, 2, 5, 6, 8\}$, $A = 1/5 = 0.2$.

Step 5. The accuracy of approximation in step 2 is equal to the accuracy of approximation found in step 4 for the set $C = \{L, B, T\}$. Hence the loss of taste and smell, breathing difficulty and temperature are the key attributes for the patients with Covid 19 disease. Thus core $= \{L, B, T\}$.

Note. The same result is obtained by applying the algorithm (1) in terms of rough topology. That is Core $= \{L, B, T\}$.

Example 3.2. This example gives information about patients with different symptoms of breast cancer namely lump in breast, inverted nipple, rashes, nipple discharge and swelling in the armpit and they are represented shortly by L, I, R, D and S

Patients	L	I	R	D	S	Breast Cancer
P_1	Yes	Yes	$*$	Yes	No	Yes
P_2	Yes	Yes	Yes	$*$	$*$	Yes
P_3	No	Yes	No	$*$	Yes	No
P_4	Yes	No	$*$	No	$*$	No
P_5	No	Yes	$*$	Yes	Yes	No
P_6	Yes	$*$	No	Yes	$*$	Yes

The columns of the table represent the symptoms for breast cancer and the rows represent the patients. The entries in the table are the attribute values. The given information system is incomplete and is given by (U, A) where $U = \{P_1, P_2, P_3, P_4, P_5, P_6\}$ and $A = \{L, I, R, D, S, \text{Breast cancer}\}$ which is divided into a set of C of condition attributes given by $C_1 = \{L, I, R, D, S\}$ and $D = \{\text{Breast cancer}\}$. The attribute Inverted Nipple generates the tolerance classes $\{P_1, P_2, P_3, P_4, P_5, P_6\}$ and $\{P_4, P_6\}$, since the missing attribute value for P_6 can be 'Yes' or 'No'. Similarly, the maximal tolerance classes for other combination of attributes can be formed. Considering all condition attributes together, the maximal tolerance classes are $\{P_1, P_2\}, \{P_1, P_6\}, \{P_3, P_5\}, \{P_4\}$ and $U/R(C) = \{\{P_1, P_2\}, \{P_1, P_6\}, \{P_3, P_5\}, \{P_4\}\}$.

Step 1. Considering all condition attributes together, the maximal tolerance classes are $\{P_1, P_2\}, \{P_1, P_6\}, \{P_3, P_5\}, \{P_4\}$ and $U/R(C) = \{\{P_1, P_2\}, \{P_1, P_6\}, \{P_3, P_5\}, \{P_4\}\}$. Let $X = \{P_1, P_2, P_6\}$ be the set of patients with breast cancer.

Step 2. $L_R(X) = \{P_1, P_2, P_6\} = U_R(X)$, $A = 1$.

Step 3. If the attribute Lump in breast is removed from the condition attributes then the equivalence classes corresponding to $C_1 = \{I, R, D, S\}$ is given by $U/R(C_1) = \{\{P_1, P_2\}, \{P_3, P_5, P_6\}, \{P_4\}, \{P_2, P_5\}, \{P_1, P_6\}\}$. $L_R(X) = \{P_1\}, U_R(X) = \{P_1, P_2, P_3, P_5, P_6\}$ and $A = 1/5 = 0.2$.

Step 4. If Inverted Nipple is removed from the condition attributes then the equivalence class corresponding to $C_2 = \{L, R, D, S\}$ is given by $U/R(C_2) = \{\{P_1, P_2\}, \{P_1, P_6\}, \{P_2, P_4\}, \{P_3, P_5\}\}$. The lower and upper approximations are $L_R(X) = \{P_1, P_6\}, U_R(X) = \{P_1, P_2, P_6, P_4\}, A = 2/4 = 0.5$. If rashes is removed from the set of condition attributes then the equivalence classes corresponding to $C_3 = \{L, I, D, S\}$ is given by $U/R(C_3) = \{\{P_1, P_2, P_6\}, \{P_3, P_5\}, \{P_4\}\}$. The lower and upper approximations are $L_R(X) = \{P_1, P_2, P_6\}, U_R(X) = \{P_1, P_2, P_6\}, A = 3/3 = 1$. If Nipple discharge is removed from the set of condition attributes then the equivalence classes corresponding to $C_4 = \{L, I, R, S\}$ is given by $U/R(C_4) = \{\{P_1, P_2\}, \{P_1, P_6\}, \{P_3, P_5\}, \{P_4, P_6\}\}$. The lower and upper approximations are $L_R(X) = \{P_1, P_2\}, U_R(X) = \{P_1, P_2, P_6, P_4\}, A = 2/4 = 0.5$ If swelling in the arm pit is removed from the set of condition attributes then the equivalence classes corresponding to $C_5 = \{L, I, R, D\}$ is given by $U/R(C_5) = \{\{P_1, P_2\}, \{P_1, P_6\}, \{P_3, P_5\}, \{P_4\}\}$. The lower and upper approximations are $L_R(X) = \{P_1, P_2, P_6\}, U_R(X), A = 3/3 = 1$. The accuracy of approximation in step 1 is equal the sets of condition attributes $\{L, I, D, S\}, \{L, I, R, D\}$.

Therefore the $CORE = \{L, I, D, S\} \cap \{L, I, R, D\} = \{L, I, D\}$. The key attributes for the breast cancer are lump in the breast, inverted nipple and nipple discharge.

Note. The core of the above example by applying algorithm 1 is also $\{L, I, D\}$.

4. Data Preparation

The proposed mathematical model is validated using association rule mining. In order to identify common symptoms and describe best patterns in the rules discovered the researcher has applied association rule mining, a widely known rule-based machine learning technique. The association rule mining based on decision making rule is incorporated to validate the proposed approach. The equivalence relation was validated using apriori algorithm which involves three parameters support, confidence and lift. Support represents the frequent occurrence of the attributes in equivalence relation. The number of times the attributes in the equivalence relation participated is represented by support, probability of the attribute in the equivalence relation participated is represented by confidence and the ratio of confidence to support is represented by lift. The goal is to identify the symptoms that produce covid disease and calculate the performance in terms of time. Use a correlation technique to determine the frequency of the item sets from a trained data set covid19. For the research, a set of 544 patient records with 6 attributes were used, all features selected without resize the dimension of the given feature and data preparation has done in the early phase to deal with missing values. Finally the model was built and evaluated using Association rule.

5. Results and Discussion

As shown in the output, the proposed algorithm retrieves the following attributes as input to the Apriori algorithm: frequent set with Loss of Taste (L), Breathing Difficulty (B), Nausea (N) and Temperature (T).

These symptom pattern mining techniques can be used in combination with other options for better understanding of the covid19 disease patterns in clinical settings.

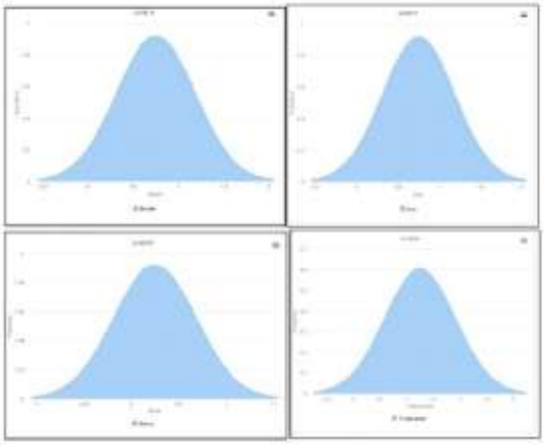

Figure 1. Probability of the Covid symptoms.

The frequent set with Loss of taste (L), Breathing Difficulty (B) and Temperature (T) plays an important part in determining the patients affected by covid. According to the rule priority indicated in the best rules derived using association rule mining. As a result, the property retrieved using the rough set approaches and the attribute retrieved using association rule mining are the same. The above clinical pattern rule found as a best rule.

Using Big O Notation, time complexity of given two algorithms were calculated and obtain the time complexity of the accuracy algorithm (1) required 1.1ms per loop

```
1000 loops, best of 5: 1.1 ms per loop
```

The time complexity of accuracy algorithm (2) was 1.09ms per loop.

Thus, it is proved that the efficiency of data obtaining and the accuracy of the covid symptoms in patients has been improved with the help of the accuracy algorithm.

6. Conclusion

A new algorithm is developed to find the core of an information system in terms of accuracy of approximation. The core of two information systems is obtained by applying the same. Its efficiency is tested in terms of time and proved that the accuracy algorithm is more efficient than the already developed algorithm.

References

[1] S. Gayathri Devi, K. Selvam and S. P. Rajagopalan, An abstract to calculate big o factors of time and space complexity of machine code, International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011), (2011), 844-847, doi: 10.1049/cp.2011.0483.

[2] M. Kryszkiewiecz, Rules in incomplete information systems, Inform. Sci. 113 (1999), 271-292.

[3] Nirmala Rebecca Paul, Decision Making in an Information system via new topology, Annals of Fuzzy Mathematics and Informatics 12(5) (2016), 591-600.

[4] Nirmala Rebecca Paul, S. Pichumani Angayarkanni and S. Jayachandra, Rough Topology and its Applications, JETIR 6(2) (2019), 275-278.

[5] Z. Pawlak, Rough Sets, Int. J. Inf. Syst. Sci. 11 (1982), 341-356.