ROUGH SET THEORY AND ITS APPLICATIONS IN DATA MINING

Ogba Paul Onu[1] and Bello Muriana[2]

[1]Department of Computer Science, Kogi State Polytechnic, Lokoja, Nigeria.
Email: ogbapaul@gmail.com

[2]Information Technology and Resource Center, Prince Abubakar Audu University, Nigeria.
Email: bmuriana685@gmail.com

Cite this article:

Ogba P. O., Bello M. (2024), Rough Set Theory and its Applications in Data Mining. British Journal of Computer, Networking and Information Technology 7(1), 84-92. DOI: 10.52589/BJCNIT-JAK93DUN

Manuscript History

Received: 19 Jan 2024

Accepted: 3 Apr 2024

Published: 1 May 2024

ABSTRACT: *One method for handling imprecise, ambiguous, and unclear data is rough set theory. Rough set theory offers a practical method for making decisions during data extraction. The practice of analyzing vast amounts of data to extract useful information from a larger collection of raw data is known as data mining. This paper discusses consistent data with rough set theory, covering blocks of attribute-value pairs, information table reductions, decision tables, and indiscernibility relations. It also explains the basics of rough set theory with a focus on applications to data mining. Additionally, rule induction algorithms are explained. The rough set theory for inconsistent data is then introduced, containing certain and potential rule sets along with lower and upper approximations. Finally, a presentation and explanation of rough set theory to incomplete data is given. This includes characteristic sets, characteristic relations, and blocks of attribute-value pairs.*

KEYWORDS: Rough set theory, imprecise data, uncertain data, data mining, Algorithm, attribute-value pairs

INTRODUCTION

In recent years, knowledge discovery (data mining, machine learning, extraction rule etc.) has gotten a lot of attention in the field of artificial intelligence thus, many types of information-finding strategies have arisen. Rough set theory (RST) is a widely used mathematics concept that was created to deal with uncertainties arising from data that contains ambiguity, inconsistencies, and errors. Rough set theory symbolizes an objective approach to inconsistencies in data, all computations are conducted directly on data sets. Rough set theory is a useful mathematical tool for dealing with inaccuracies, inconsistencies, and incomplete data. RST was proposed by Professor Pawlak in 1982 and the basic idea behind the theory can be divided into two parts. The first part discusses the concepts and rules from the classification of relational databases while the second part discovers knowledge from classifying equivalence relation and classification for the approximation of the target. After fuzzy set theory, evidence theory, and probability theory, the RST is a novel mathematical technique for handling imperfect data in data analysis. The fields in which the RST has been applied include data mining, image processing, pattern recognition, medical informatics, and expert systems. Numerous studies have combined RST with other artificial intelligence techniques, like neural networks and fuzzy logic, with encouraging results. The application of RST to a specific complex situation has spurred interest in additional research and development, broadening the scope of the original theory and its applications. Moreover, RST is a computationally efficient method that is crucial to a wide range of theoretical and practical computing and automation applications, particularly in the areas of machine learning, intelligent control and data mining (Vluymans et al., 2015).

Fundamentals of Rough Set Theory

The rough set concept can be described by set approximations. We now have a more detailed description of the problem. Let there be a finite set of objects Ω and a binary relation R$\subseteq \Omega \times \Omega$. The set R is called the indiscernibility relation while the set Ω is called the universe (Zhou & Shen, 2012). Assuming R is an equivalence relation, a pair (Ω, R) is called an approximation space.

Let Y be a subset of Ω, i.e. Y$\subseteq \Omega$, we can characterise the set Y with respect to R

R(y) denotes the equivalence class of R and is determined by element y.

i. R-lower approximation of a set Y with respect to R is the set of all objects that can be with **certainty** which are members of Y with respect to R is given by R_*(Y), i.e, $R_*(Y) = \{y : R(y) \subseteq Y\}$

ii. R-upper approximation of a set Y with respect to R is the set of all objects that are grouped as **possible** members of Y with respect to R and is given by $R^*(Y)$, i.e., $R^*(Y) = \{y : R(y) \cap Y \neq \emptyset\}$

iii. The boundary region of a set Y with respect to R is the set of all objects which cannot be grouped as members of Y or –Y and is given as $RN_R(Y)$, i.e. $RN_R(Y) = R^*(Y) - R_*$(Y)

Rough set theory can now defined thus;

i. A set Y is called **crisp** with respect to binary relation R if and only if the boundary

Article DOI: 10.52589/BJCNIT-JAK93DUN
DOI URL: https://doi.org/10.52589/BJCNIT-JAK93DUN

region of X is empty.

ii. A set Y is called **rough** with respect to binary relation R if and only if the boundary region of X is nonempty

From these above, the following properties of approximations can be easily proven

i. $R_*(\mathrm{Y}) \subseteq Y \subseteq R^*(Y)$

ii. $R_*(\emptyset) = R^*(\emptyset); R_*(\Omega) = R^*(\Omega) = \Omega$

iii. $R_*(Y \cap Z) = R_*(Y) \cap R_*(Z)$

iv. $R^*(Y \cup Z) = R^*(Y) \cup R^*(Z)$

v. $R^*(Y \cap Z) \subseteq R^*(Y) \cap R^*(Z)$

vi. $R_*(Y \cup Z) \supseteq R_*(Y) \cup R_*(Z)$

vii. $R^*(-Y) = -R_*(Y)$

viii. $R_*(-Y) = -R^*(Y)$

The Inexactness and topological characterization of imprecision can be defined by these four basic classes of rough sets:

i. If $R_*(Y) \neq\emptyset\ and\ R^*(Y) \neq \Omega$ then a set Y is s roughly R-definable

ii. If $R_*(Y) =\emptyset\ and\ R^*(Y) \neq \Omega$, then a set Y is internally R-undefinable

iii. If $R_*(Y) \neq\emptyset\ and\ R^*(Y) = \Omega,$ then a set Y is s externally R-undefinable

iv. If $R_*(Y) =\emptyset\ and\ R^*(Y) = \Omega,$ then a set Y is s totally R-undefinable

This classification has the following basic connotations.

That a set y is roughly R-definable means that with respect to binary relation R, we can choose for some elements of Ω that belong to Y and for some elements of Ω that belong to –Y.

That a set y is internally R-undefinable means that with respect to binary relation R, we can choose for some elements of Ω that belong to -Y but we cannot choose for any element of Ω if it belongs to Y

That a set y is externally R-undefinable means that with respect to binary relation R, we can choose for some elements of Ω that belong to Y, but we cannot choose for any element of Ω whether it belongs to –Y

That a set Y is totally R-undefinable means that with respect to binary relation R, we cannot choose any element of U if it belongs to Y or –Y.

Article DOI: 10.52589/BJCNIT-JAK93DUN
DOI URL: https://doi.org/10.52589/BJCNIT-JAK93DUN

Rough Set Theory in Data Analysis

This study is built on the original rough set model's data-mining techniques. Some data mining methods and applications used with RST are discussed. RST is built on the notion that we associate some information with every element in the universe (data, knowledge), (Skowron, et al, 2018). Objects with the same information have a comparable view of the information available to them. The mathematical foundation of rough set theory is the similarity relation produced in this way. The term "elementary set" refers to any group of all related objects that make up an atom (basic granule) of information about the universe. Any union of any elementary set is said to have been nominated by a crisp or precise set, otherwise, the set is referred to as rough (inaccurate, unclear). Members of the set or its complement are objects contained in the available knowledge that cannot be categorized with certainty. Rough sets, unlike precise sets, cannot be described in terms of data about their elements (Xu. and Liu, 2013). A pair of precise sets known as the upper and lower approximation are associated with any rough set technique. The lower approximation is made up of all objects that belong to the set, whereas the upper approximation is made up of all objects that might belong to the set.

Information Table

The data is represented as a table, with each row representing an object and every column representing a measurable attribute (a variable, a property, etc.) for each object. Such a table is known as an information table. An information table is a pair $G = (\Omega, B)$ such that Ω is a non-empty finite set of objects otherwise known as the universe and B is a non-empty finite set of attributes. That is $b\colon \Omega \rightarrow V_b$ such that for every b ∈ B, the set V_b is called the value set of B (Meia, et al, 2015). The table presents data containing six (6) women who underwent pregnancy tests and five (5) sets of attributes were discovered: Mammary gland, Nausea, Bloating, Cramping, and Mood swing.

Table 1: Information Table

Cases	Attributes Mammary gland	Nausea	Bloating	Cramping	Mood swing
1	Normal	No	Yes	Yes	Yes
2	Very big	Yes	Yes	Yes	No
3	Big	No	No	No	No
4	Big	Yes	Yes	Yes	Yes
5	Normal	Yes	No	No	No
6	Big	Yes	No	yes	No

Let Ω denotes the set of all cases, the set of all attributes denoted by B, and V the set of all attribute values. Such a table defines an information function I: Ω × B → V. For example, I(1, mammary gland) = normal.

Let $b \in B$, $v \in V$ and $\tau = (b, v)$ be an attribute-value pair. A block of τ, denoted by $[\tau]$, is a set of all cases from Ω which attribute b has value v. For the information table from Table 1, the block is defined as follows:

[(Mammary gland, normal)] = {1, 5},

[(Mammary gland, very big)] = {2},

Article DOI: 10.52589/BJCNIT-JAK93DUN
DOI URL: https://doi.org/10.52589/BJCNIT-JAK93DUN

[(Mammary gland, big)] = {3, 4, 6},

[(Nausea, no)] = {1, 3,},

[(Nausea, yes)] = {2, 4, 5, 6},

[(Bloating, yes)] = {1, 2, 4},

[(Bloating, no)] = {3, 5, 6},

[(Cramping, yes)] = {1, 2, 4, 6},

[(Cramping, no)] = {3, 5,},

[(Mood swing, yes)] = {1, 4},

[(Mood swing, no)] = {2, 3, 5, 6},

Let x∈ Ω and A ⊆ B. An elementary set of A containing x, denoted by [x]A, is the following set:

$\cap \{[(b, v)] I b \in A, \Omega(x, b) = v\}$

Elementary sets are a subset of U that consists of all U cases that are distinct from x while employing all B attributes. Information granules are the words used in soft computing to describe simple sets. Elementary sets are blocks of attribute-value pairs specified by that specific attribute when subset B is constrained to a single attribute. Therefore;

$[1]_{(mammary\ gland)} = [5]_{(mammary\ gland)}$ = [(Mammary gland, normal)] = {1, 5},

$[2]_{(mammary\ gland)}$ = [(Mammary gland, very big)] = {2},

$[3]_{(mammary\ gland)} = [4]_{(mammary\ gland)} = [6]_{(mammary\ gland)}$ = [(Mammary gland, big)] = {3, 4, 6},

Also; if A = {Mammary gland, Nausea},

$[1]_A = [(Mammary\ gland, normal) \cap [(Nausea, no)] = \{1\},$

$[2]_A = [(Mammary\ gland, very\ big) \cap [(Nausea, yes)] = \{2\},$

$[3]_A = [(Mammary\ gland, big) \cap [(Nausea, no)] = \{3\},$

$[4]_A = [6]_A[(Mammary\ gland, big\ ) \cap [(Nausea, yes)] = \{4{,}6\},$

$[5]_A = [(Mammary\ gland, normal) \cap [(Nausea, yes)] = \{5\},$

Another technique to define elementary sets is to use the concept of an indiscernibility relation. Once more Let A be a nonempty subset of the set B of all attributes. The indiscernibility relation IND(A) is a binary relation on Ω for $x, y \in \Omega$ as;

$(x, y) \in IND(A) if and\ only\ if\ I(x, b) = I(y, b) for\ all\ b \in A$

Article DOI: 10.52589/BJCNIT-JAK93DUN
DOI URL: https://doi.org/10.52589/BJCNIT-JAK93DUN

IND(A) is certainly an equivalence relation. Partitions are a convenient way to illustrate equivalence relations. A partition of *Ω* is a set of mutually disjoint nonempty subsets of *Ω* known as blocks, the union of which is *Ω*. The partition created by IND(A) will be indicated by A*. A* blocks are also referred to as elementary sets associated with A. For instance,

$$\{Mammary\ gland\}^{*} = \{\{1,5\},\{2\},\{3,4,6\}\}$$

$$\{Mammary\ gland, Nausea\}^{*} = \{\{1\},\{2\}.\{3\},\{4.6],\{5\}\}$$

So $IND(\{Mammary\ gland\}) = \{(1,1),(1,5),(2,2),(3,3),(3,4),(3,6),(4,3),(4,4),$

$(4,6),(5,1),(5,5),(6,3),(6,4),(6,6)\}$

$IND(\{Mammary\ gland, Nausea\} = \{(1,1),(2,2),(3,3),(4,4),(4,6),(5,5),(6,4),(6,6)\}$

B subset A of the set B is called a reduct if and only if

A* = B*; and A is minimal with this property, i.e., (A – {b})* ≠ B* for all b ∈ A.

For instance, {Mammary gland} is not a reduct since

$\{Mammry\ gland\}^{*} = \big\{\{1,5\},\{2\},\{3,4,6\}\big\} \neq B^{*} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$

Consequently, $\{Mammary\ gland, Nausea\}$ is not a reduct since

$\{Mammary\ gland, Nausea\}^{*} = \big\{\{1\},\{2\}.\{3\},\{4,6\},\{5\}\big\} \neq B^{*} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$

Conversely, $\{Mammary\ gland, Nausea, Bloating\}$ is a reduct because $\{Mammary\ gland, Nausea, Bloating\}^{*} = \big\{\{1\},\{2\}.\{3\},\{4\},\{5\},\{6\}\big\} = B^{*} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$

So

$\{Mammary\ gland, Nausea\}^{*} \neq B^{*}$

$\{Mammary\ gland, Bloating\}^{*} \neq B^{*}$

$\{Mammary\ gland, Cramping\}^{*} \neq B^{*}$

$\{Mammary\ gland, Mood\ swing\}^{*} \neq B^{*}$

$\{Mammary\ gland, Nausea, Bloating\}^{*} = B^{*}$

$\{Mammary\ gland, Nausea, Cramping\}^{*} \neq B^{*}$

$\{Mammary\ gland, Nausea, Mood\ swing\}^{*} = B^{*}$

$\{Nausea, Bloating, Cramping\}^{*} \neq B^{*}$

$\{Nausea, Bloating, Mood\ swing\}^{*} \neq B^{*}$

$\{Bloating, Cramping, Mood\ swing\}^{*} \neq B^{*}$

Therefore, reducts are: $\{Mammary\ gland, Nausea, Bloating\}$ and $\{Mammary\ gland, Nausea, Mood\ swing\}$

Decision Tables

A decision table is any information table of the form $I = (\Omega, B, V, f)$, where

Ω Represents the universe, B is the set of attributes which include (Mammary gland, Nausea, Bloating, cramping and mood swing. V represents the union of the set of values of an attribute b included in B which can be represented as $\cup_{b\in B} V_b$, f is a decision function which can be represented as $f: \Omega \times B \rightarrow V$, such that f(y,b) belongs to V_b for every y that belongs to Ω and every attribute value that belongs to A (Ehrenfeucht, et al., 2017).

Table 2: Decision table

Cases	Attributes					Decision
	Mammary gland	Nausea	Bloating	Cramping	Mood swing	Pregnancy
1	Normal	No	Yes	Yes	Yes	No
2	Very big	Yes	Yes	Yes	No	Yes
3	Big	No	No	No	No	No
4	Big	Yes	Yes	Yes	Yes	Yes
5	Normal	Yes	No	No	No	No
6	Big	Yes	No	yes	No	Yes

Table 2 contains data concerning six patients that were subjected to pregnancy tests. The condition attributes displayed in the table show the attributes of the patients which are Mammary gland, Nausea, Bloating, Cramping, and mood swing respectively, and the results of the test are displayed. In the decision table Ω ={1,2,…,6}, Cond={Mammary gland, Nausea, Bloating, Cramping, Mood swing}, Dec={Pregnancy} and all attributes domains are equal V={Normal, Very big, Big, Yes, No}.

In Rough Set Theory, the dependency between condition and decision attributes is determined by approximations. Given that the state of the patients cannot be determined exactly by the attributes possessed by the patients, it is possible to use approximations to identify the state of the patients by identifying the functional relationship between decision attributes and values of condition (Slezak and Eastwood, 2019).

The degree of dependency between condition and decision attributes may define the consistency factor of the decision rule conflicting. This means using rules with the same conditions but different decisions. As an example, the Table 2 consistency factor is 3/6, this factor means that three out of six (50%) patients can be appropriately classified as being pregnant or not pregnant based on their attributes (Jankowski, et al, 2015)

Assuming A is a subset of B. It is possible to assign to every subset Y of the universe Ω two sets $\underline{A(Y)}$ and $\underline{A(Y)}$ which are called, the A-upper and the A-lower approximation of Y denoted as;

$$\underline{A(Y)} = \{y \in \Omega: A(y) \sqcap Y \neq 0\}$$

$\underline{A(Y)} = \{y \in \Omega : A(y) \subseteq Y\}$

A function rule also known as a decision rule f_y included in Ω is consistent or deterministic if for every x included in $\Omega, x \neq y$ $(f_y/Cond = f_x/cond) \longrightarrow (f_y/Dec = f_x/Dec,$ otherwise, the decision rule f_y is inconsistent or nondeterministic.

Let there be two decision rules D1 and D2 as follows;

D1: IF (Mammary gland=Very big AND Nausea= Yes AND Bloating=Yes AND Cramping=Yes AND Mood swing=No) Then (Pregnancy=Yes)

D2: IF (Mammary gland=Normal) Then (Pregnancy= Yes) OR (Pregnancy=No)

The following are some of the most specific definitions of a decision rule:

Rule Strength is the count of objects in the data set with the property described by the decisions and the rule conditions. The rule strength of D1=6.

Rule Length is the count of objects in the data set with the property described by the rule conditions. The rule length of D1=5.

An exact rule is the outcome of an exact rule that corresponds to one or more different conditions. The set of objects in the lower approximation is used to generate exact rules. D1 is an exact Rule.

Approximate rule: An approximate rule's similar condition corresponds to more than one outcome. For the boundary, approximate rules are created. D2 is an approximate rule

Rule support: is the number of objects in the data set that have the property described by the rule's conditions. The rule support of D1=5

Rule acceptance: this is the count of a rule's condition that may be used to express the rule acceptance measure. It's a subjective metric that expresses the user's confidence in the extracted rules. It's a broadening of the rule support and rules coverage concepts.

Discrimination level (DL): The precision of a rule that represents the corresponding objects is measured by the Discrimination level.

CONCLUSION

In recent years, data mining applications based on the original concept of rough set theory have been tried as valuable approaches to produce decision rules. More research is needed on the obtained results, especially when quantitative features are involved. Due to space constraints, this paper only covers data representation using rough set theory (information and decision tables dealing with consistent data), as well as a few basic data mining applications. This paper presents the LERS (Learning from Examples based on Rough Sets) data mining system as an example of a successful rough set theory data mining application (Ehrenfeucht and Rozenberg, 2014).

FUTURE RESEARCH

We intend to continue working on the data representation so that it may be applied to inconsistent data.

REFERENCES

Ehrenfeucht, A., Rozenberg, G. (2014). Zoom structures and reaction systems yield exploration systems. Int J Found Comput Sci 25:275–306

Ehrenfeucht, A., Petre, I., Rozenberg, G. (2017). Reaction systems: a model of computation inspired by the functioning of the living cell. In: Konstantinidis S, Moreira N, Reis R, Shallit J (eds) The role of theory in computer science—essays dedicated to Janusz Brzozowski. World Scientific, Singapore, pp 1–32

Jankowski, A., Skowron, A., Dutta, S. (2015). Toward problem solving support based on big data and domain knowledge: interactive granular computing and adaptive judgement. In: Japkowicz N, Stefanowski J (eds) Big data analysis: new algorithms for a new society, series big data, vol 16. Springer, Heidelberg, pp 44–90

Meia, S., Zarrabi, N., Lees, M., Sloot, P. (2015). Complex agent networks: an emerging approach for modeling complex systems. Appl Soft Comput 37:311–321

Skowron, A., Jankowski, A., Wasilewski, P. (2018). Rough sets and sorites paradox. Fundam Inf 157(4):371–384

Slezak, D., Eastwood, V. (2019). Data warehouse technology by Infobright. In: Çetintemel U, Zdonik SB, Kossmann D, Tatbul N (eds) Proceedings of the ACM SIGMOD international conference on management of data, SIGMOD 2019, Providence, Rhode Island, USA. ACM, pp 841–846

Vluymans, S, D'eer L., Saeys, Y., Cornelis, C. (2015). Applications of fuzzy rough set theory in machine learning: a survey. Fundam Inf 142(1–4):53–86

Xu Y., and Liu C. (2013). A rough margin-based one class support vector machine, Neural Computing & Applications, 22(6):1077-1084.

Zhou Z., and Shen G. (2012). Application of the theory of rough set in intelligence analysis for index weight determination [J]. Information Studies Theory & Application, 35(9):61-65.

Article DOI: 10.52589/BJCNIT-JAK93DUN
DOI URL: https://doi.org/10.52589/BJCNIT-JAK93DUN