Filomat 39:17 (2025), 6019–6033
https://doi.org/10.2298/FIL2517019Y

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

The rough topology for numerical data

Uğur Yiğit[a]

[a] *Department of Mathematics, Istanbul Medeniyet University, 34700 Istanbul, Türkiye*

Abstract. In this paper, we generalize the rough topology and the core to numerical data by classifying objects in terms of the attribute values. A new approach to finding the core for numerical data is discussed. A measurement criterion is introduced to determine whether an attribute belongs to the core. This new method for finding the core is used for attribute reduction. It is tested and compared by using eight different machine-learning algorithms. Also, it is discussed how this material is used to rank the importance of attributes in data classification. Finally, the algorithms and codes for data conversion and core determination are provided.

1. Introduction

Pawlak's rough set theory [14] handles the approximation of sets in terms of equivalence (indiscernibility) relations. The primary application of this theory is data analysis and decision-making processes. In Pawlak's work [15], the indiscernibility relations arise when one considers a given set of attributes. Two objects are equivalent if their values of all attributes in the data are the same. Thivagar et al.[20] introduce rough topology by means of rough sets and apply it to analyze real-life problems. They find the key attributes of some diseases to decide whether a patient has a disease or not.

Several generalizations of rough set theory have been proposed to extend its applicability. These generalizations are the following. Covering-based rough sets define approximations using coverings of the universe, as opposed to Pawlak's rough sets, which are based on equivalence relations. This method has been used in fields including data reduction and feature selection and offers greater flexibility when managing complicated data structures [22]. By introducing near concepts, the Generalized Covering Approximation Space model broadens the scope of rough set theory and enables more comprehensive classifications [1]. β-Basic Rough Sets present the idea of β-approximations, which modify the approximation boundaries according to a parameter β, enabling a more adaptable handling of uncertainty. This method has been used in medical diagnosis, where various levels of uncertainty are taken into account [7].

Rough set theory has found significant applications in the fields of medicine and decision-making. The theory has been applied to diagnosing heart failure, dengue fever, and other diseases using generalized rough sets [8], [3], [4]. To manage patient data uncertainty, rough sets have been utilized extensively in medical diagnosis. For instance, by offering more flexible approximations, β-basic rough sets have been used to increase the precision of medical diagnosis [7]. Similarly, by capturing the underlying structure of

2020 *Mathematics Subject Classification*. Primary 54A05, 54H30; Secondary 68T37, 68Q87.
Keywords. Rough Sets, rough topology, core for numerical data, machine learning
Received: 23 September 2024; Accepted: 06 April 2025
Communicated by Biljana Popović
Email address: ugur.yigit@medeniyet.edu.tr (Uğur Yiğit)
ORCID iD: https://orcid.org/0000-0002-6173-5727 (Uğur Yiğit)

patient records, rough topology has been utilized to improve medical data analysis [1]. Rough sets are also very useful in reducing data, aiming to simplify data while maintaining its key characteristics. To improve the analysis of rheumatic fever data, for example, tritopological approximation spaces have been utilized to reduce the dataset's dimensionality while preserving critical information [13].

The significance of precise data analysis in handling public health emergencies has been brought to light by the COVID-19 pandemic. COVID-19 variations have been analyzed using rough sets, which have shed light on the impact and dissemination of many variants. In this situation, the use of virtually initial-rough sets has been especially successful, providing a strong foundation for managing the uncertainty present in pandemic data [5].

Rough sets have also been applied to decision-making processes in healthcare, especially in diagnosing heart failure problems. By providing a systematic approach to handling uncertainty, rough sets enable more accurate and reliable decision-making, which improves patient outcomes [3]. Furthermore, using initial neighborhoods and ideals to increase diagnostic accuracy, the use of generalized rough sets in dengue fever diagnosis has demonstrated encouraging outcomes [8].

Several applications of the theory deal with the Boolean type of data in which attributes usually take values yes and no; or 0 and 1. A contribution of this work is to give a generalization of the rough topology and the core to numerical data by reducing data to a usable form by using the standard deviation of attributes. Also, this method provides a new model for attribute reduction for large-scale data processing. The main objective of this study is to use machine learning techniques, rough sets, and topology to create better models for handling uncertain data and to demonstrate how this theory can be used to create a better feature selection method.

2. Preliminaries

Let U be a non-empty set of objects called the universe. A relation R on U is a subset of the cartesian product $U \times U$. An element $(a, b) \in R$ is generally written as aRb. A relation on U is called reflexive if aRa for all $a \in U$. It is called symmetric if aRb implies bRa for all $a, b \in U$. It is called transitive if aRb and bRc then aRc for all $a, b, c \in U$. If R is reflexive, symmetric, and transitive, then R is said to be an equivalence relation on U.

Let R be an equivalence relation on a set U, and let $x \in U$. The set of all elements in U that is related to x is called the equivalence class of x under R and is denoted by $[x]_R$. That is, $[x]_R = \{y \in U | xRy\}$. The set of all equivalence classes U/R of R in U gives a partition of U, which means that all equivalence classes are disjoint, and the union of them is U.

Definition 2.1. *[15] Let U be a non-empty finite set and R be an equivalence relation on U. An approximation space is a pair (U, R). Let X be a subset of U.*

(i) The lower approximation of X with respect to R is

$$R_\star(X) = \cup_{x \in U}\{x | [x]_R \subset X\}.$$

(ii) The upper approximation of X with respect to R is

$$R^\star(X) = \cup_{x \in U}\{x | [x]_R \cap X \neq \emptyset\}.$$

(iii) The boundary region of X with respect to R is

$$B_R(X) = R^\star(X) - R_\star(X).$$

The set X is called a rough set with respect to R if $B_R(X) \neq \emptyset$.

3. Rough Topology

In this section, we first introduce the definition of a topology and a basis for a topology. We secondly present the rough topology which is given by Thivagar et al. in [20] in terms of the lower and the upper approximations.

Definition 3.1. *[11] A topology on a set U is a collection τ of subsets of U satisfying the following properties:*

(T1) $\emptyset, U \in \tau$.
(T2) The union of the elements of τ is in τ.
(T3) The intersection of the finite number of elements of τ is in τ.

The pair (U, τ) is called a topological space.

Let (U, τ) be a topological space. A basis for (U, τ) is a collection $\beta \subset \tau$ such that for each $A \in \tau$ and each $x \in A$, there exists $B \in \beta$ such that $x \in B \subset A$.

Definition 3.2. *[20] Let U be a non-empty finite set and R be an equivalence relation on U. For $X \subset U$, $\tau_R = \{U, \emptyset, R_\star(X), R^\star(X), B_R(X)\}$ forms a topology on U, which is called a rough topology on U with respect to X.*

Lemma 3.3. *[20] The set $\beta_R = \{U, R_\star(X), B_R(X)\}$ is a basis for the rough topology τ_R on U with respect to X.*

Example 3.4. *Let $U = \{1, 2, 3, 4, 5\}$ and*

$$R = \{(1,1), (2,2), (3,3), (4,4), (5,5), (1,2), (2,1), (3,5), (5,3)\}$$

be an equivalence relation on U. Then the set of equivalence classes of U by the equivalence relation R is $U/R = \{\{1,2\}, \{3,5\}, \{4\}\}$. For $X = \{1,2,3\}$, $R^\star(X) = \{1,2,3,5\}, R_\star(X) = \{1,2\}$ and $B_R(X) = \{3,5\}$. Therefore, the rough topology $\tau_R = \{U, \emptyset, \{1,2\}, \{1,2,3,5\}, \{3,5\}\}$. The basis for τ_R is $\beta_R = \{U, \{1,2\}, \{3,5\}\}$.

Definition 3.5. *[20] A subset M of the set of attributes is called the core of R if $\beta_M \neq \beta_{(R-\{r\})}$ for every r in M. That is, the elements of the core cannot be removed without affecting the classification power of attributes.*

Example 3.6. *Consider the Table 1, which is taken from Pawlak [16].*

Table 1: Sample Data 1

Patient	*Headache(H)*	*Muscular pain(M)*	*Temperature(T)*	*Flu*
1	*No*	*Yes*	*High*	*Yes*
2	*Yes*	*No*	*High*	*Yes*
3	*Yes*	*Yes*	*Very High*	*Yes*
4	*No*	*Yes*	*Normal*	*No*
5	*Yes*	*No*	*High*	*No*
6	*No*	*Yes*	*Very High*	*Yes*

Let $U = \{1,2,3,4,5,6\}$ be the set of patients, and $X = \{1,2,3,6\}$ be the set of patients having flue. Let R be an equivalence (indiscernibility) relation, in which two patients are equivalent if their values of all attributes are the same. Then the set of equivalence classes of U by the equivalence relation R is $U/R = \{\{1\}, \{2,5\}, \{3\}, \{4\}, \{6\}\}$. For $X = \{1,2,3,6\}$, the upper approximation $R^\star(X) = \{1,2,3,5,6\}$, the lower approximation $R_\star(X) = \{1,3,6\}$, and $B_R(X) = \{2,5\}$. Therefore the rough topology $\tau_R = \{U, \emptyset, \{1,3,6\}, \{1,2,3,5,6\}, \{2,5\}\}$. The basis for τ_R is $\beta_R = \{U, \{1,3,6\}, \{2,5\}\}$.

If we remove the attribute "Headache" from the set of condition attributes, the family of equivalence classes with the resulting set of attributes is given by $U/(R-\{H\}) = \{\{1\},\{2,5\},\{3,6\},\{4\}\}$. *Then, the lower and upper approximations of X with respect to* $R-\{H\}$ *are given by* $(R-\{H\})^{\star}(X)$ =*{1,2,3,5,6} and* $(R-\{H\})_{\star}(X)$=*{1,3,6}, respectively. Therefore,* $\tau_{R-\{H\}}$=$\{U, \emptyset$, *{1,3,6}, {1,2,3,5,6}, {2,5}}. The basis for this topology* $\tau_{(R-\{H\})}$ *is given by* $\beta_{(R-\{H\})} = \{U, \{1,3,6\}, \{2,5\}\}$. *Therefore,* $\beta_R = \beta_{(R-\{H\})}$, *which means "Headache" is not in the Core*(X).

If the attribute "Muscular pain" is omitted, then $U/(R-\{M\})$=*{ {1}, {2,5}, {3}, {4}, {6}}, which is the same with* U/R. *Hence,* $\tau_{(R-\{M\})} = \tau_R$ *and* $\beta_{(R-\{H\})} = \beta_R$, *which means "Muscular pain" is not in the Core*(X).

If we remove the attribute "Temperature" from the set of condition attributes, the family of equivalence classes with the resulting set of attributes is given by $U/(R-\{T\}) = \{\{1,4,6\},\{2,5\},\{3\}\}$. *Then, the lower and upper approximations of X with respect to* $R-\{T\}$ *are given by* $(R-\{T\})^{\star}(X) = \{1,2,3,4,5,6\}$ *and* $(R-\{T\})_{\star}(X) = \{3\}$. *Therefore,* $\tau_{R-\{T\}} = \{U, \emptyset, \{3\}\}$. *The basis for this topology* $\tau_{(R-\{T\})}$ *is given by* $\beta_{(R-\{T\})} = \{U, \{3\}\}$. *Therefore,* $\beta_R \neq \beta_{(R-\{H\})}$, *which means "Temperature" is in the Core*(X).

Therefore, Core$(X) = \{$*Temperature*$\}$. *If we take* $X = \{4,5\}$ *as the set of patients not having flu, then similarly Core*$(X) = \{$*Temperature*$\}$.

Observation: *We conclude that "Temperature" is the key attribute to decide whether a patient has flu or not.*

4. The Rough Topology for numerical data

The rough topology and the core can be used to analyze many real-life problems like diseases, electrical transmission lines, decision-making problems, etc. in the literature [2, 12, 18]. However, the values of attributes are like "yes or no"; or "high, normal, very high". In this section, we give a generalization of the method to analyze numerical data by converting them to pertinent data utilizing standard deviations of the attributes. We also provide algorithms and Python codes to find equivalence classes, the lower and the upper approximations, topologies, bases, and the core.

Recall that an ordered pair (U, R) where U is a non-empty set and R is an equivalence relation defined on U is called an approximation space. Let (U, R) be an approximation space with indiscernibility (equivalence) relation R. Two objects are equivalent if and only if their values of all attributes are the same. However, most of the values are different from each other in the numerical data. One needs a method for converting numerical data into formats that help you analyze by using the rough topology and the core. We use standard deviations of attributes to do this. We assume that two objects take the same value if they are close to each other as near as the standard deviation of the attribute. The algorithm of the procedure is as follows Algorithm 1:

Algorithm 1:

Step 1: Given a data table, columns of which are attributes, rows of which are objects, and entries of the table are attribute values, pick one of the attributes.

Step 2: Take the maximum (*Max*) and the standard deviation (*St*) of the chosen attribute.

Step 3: Assign 1 to the values between (*Max*) and (*Max* − *St*) and discard these rows. Find the next maximum after discarding rows assigned as 1. Assign 2 to the values between the new (*Max*) and new (*Max* − *St*) and discard them. Repeat this process until every value is assigned to a new value.

Step 4: Repeat Step 2 and Step 3 for every attribute.

Step 5: Generate the new data table.

(see the acknowledgment for the link for Python codes).

Example 4.1. *Consider the Table 2, which is taken from Järvinen [9].*

Table 2: Sample Data 2

Patient	*Temperature(T)*	*Blood Pressure(BP)*	*Hemoglobin(HB)*	*Results*
1	*39.3*	*103/65*	*125*	*No*
2	*39.1*	*97/60*	*116*	*No*
3	*39.2*	*109/71*	*132*	*No*
4	*37.1*	*150/96*	*139*	*Yes*
5	*37.3*	*145/93*	*130*	*Yes*
6	*37.8*	*143/95*	*121*	*Yes*
7	*36.7*	*138/83*	*130*	*No*

By applying the algorithm 1 to Table 2, we get the following Table 3:

Table 3: Converted Sample Data 2

Patient	*Temperature(T)*	*Blood Pressure(BP)*	*Hemoglobin(HB)*	*Results*
1	*1*	*2*	*2*	*No*
2	*1*	*1*	*3*	*No*
3	*1*	*2*	*1*	*No*
4	*2*	*2*	*1*	*Yes*
5	*2*	*2*	*2*	*Yes*
6	*2*	*3*	*3*	*Yes*
7	*3*	*1*	*2*	*No*

Let $U = \{1, 2, 3, 4, 5, 6, 7\}$ be the set of patients, and $X = \{4, 5, 6\}$ be the set of patients having a positive result. Let R be an equivalence (indiscernibility) relation, in which two patients are equivalent if their values of all attributes are the same. Then the set of equivalence classes of U by the equivalence relation R is $U/R = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}\}$. For $X = \{4, 5, 6\}$, the upper approximation $R^\star(X) = \{4, 5, 6\}$, the lower approximation $R_\star(X) = \{4, 5, 6\}$, and $B_R(X) = \emptyset$. Therefore the rough topology $\tau_R = \{U, \emptyset, \{4, 5, 6\}\}$. The basis for τ_R is $\beta_R = \{U, \{4, 5, 6\}\}$.

If we remove the attribute "Temperature" from the set of condition attributes, the family of equivalence classes with the resulting set of attributes is given by $U/(R - \{T\})$=$\{\{1,5\},\{2\},\{3,4\},\{6\},\{7\}\}$. Then, the lower and the upper approximations, and boundary of X with respect to $R - \{T\}$ are given by $(R - \{T\})^\star(X) = \{1, 3, 4, 5, 6\}$ and $(R - \{T\})_\star(X) = \{6\}$, and $B_{(R-\{T\})}(X) = \{1, 3, 4, 5\}$, respectively. Therefore, $\tau_{R-\{T\}} = \{U, \emptyset, \{1, 3, 4, 5, 6\}, \{6\}, \{1, 3, 4, 5\}\}$. The basis for this topology $\tau_{(R-\{T\})}$ is given by $\beta_{(R-\{T\})} = \{U, \{1, 3, 4, 5\}, \{6\}\}$. Therefore, $\beta_R \neq \beta_{(R-\{T\})}$, which means "Temperature" is in the Core(X).

If the attribute "Blood Pressure" is omitted, then $U/(R - \{BP\})$=$\{\{1\}, \{2\}, \{3\}, \{4\},\{5\}, \{6\}, \{7\}\}$, which is the same with U/R. Hence, $\tau_{(R-\{BP\})} = \tau_R$ and $\beta_{(R-\{BP\})} = \beta_R$, which means "Blood Pressure" is not in the Core(R).

If we remove the attribute "Hemoglobin" from the set of condition attributes, the family of equivalence classes with the resulting set of attributes is given by $U/(R - \{HB\})$=$\{\{1,3\},\{2\}, \{4,5\}, \{6\},\{7\}\}$. Then, the lower and upper approximations of X with respect to $R - \{HB\}$ are given by $(R - \{HB\})^\star(X) = \{4, 5, 6\}$ and $(R - \{HB\})_\star(X) = \{4, 5, 6\}$,respectively. Hence, $\tau_{(R-\{HB\})} = \tau_R$ and $\beta_{(R-\{HB\})} = \beta_R$, which means "Hemoglobin" is not in the Core(R).

Therefore, Core(X) = {Temperature}. If we take $X = \{1, 2, 3, 7\}$ as the set of patients having a negative result, then similarly Core(X) = {Temperature}.

Observation: *We conclude that "Temperature" is the key attribute to decide whether a patient has a positive result or not.*

5. A topological method for big numerical data

For big numerical data, the rough topology is restrictive since the topology could be changed by only one object (row). In other words, even if only one object (row) is in the boundary and the lower approximation is different from the boundary or lower approximation for the equivalence relation which is omitted one attribute, then topologies are different. As a result, most attributes are at the core of big data. However, we can give a measurement for this change (see definition 5.2). For example, let's consider the data with 1000 objects. If only 10 objects change the boundary and the lower approximation, these differences are possibly insignificant, so it could be tolerated for big data. This tool also provides the order of importance of attributes in the dataset.

Definition 5.1. *Let (U, R) be an approximation space and M be the set of attributes. For $r \in M$ and $X \subset U$, the accuracy of an attribute and the accuracy of the core with respect to X are defined respectively by*

$$\mu_r(X) = \frac{|B_R(X)|}{|B_{(R-r)}(X)|}$$

$$\mu_{RT}(X) = Sup_{r \in M}\{\frac{|B_R(X)|}{|B_{(R-r)}(X)|}\}.$$

Definition 5.2. *Let (U, R) be an approximation space and M be the set of attributes. For $r \in M$ and $X \subset U$, the accuracy of the boundary for r with respect to X are defined by*

$$\nu_r(X) = |B_R(X) - B_{(R-r)}(X)|.$$

Note that $\mu_{RT}(X) = \mu_{RT}(U - X)$ and $\nu_r(X) = \nu_r(U - X)$ for any $r \in M$. Obviously, $\mu_{RT}(X) \leq 1$.

Remark 5.3. *In essence, the variable $\nu_r(X)$ indicates how many rows (objects) move from the upper approximation to the lower approximation, or vice versa, when the attribute r is ignored. In other words, it determines the number of rows that affect how close the set X is to being a rough set. Therefore, it is a crucial variable in this method that indicates the extent to which the attribute r influences the data's classification. Consequently, $\nu_r(X)$ is a measuring tool used to lay out insignificant objects that can be tolerated.*

Remark 5.4. *In machine learning, the percentage of data that can be removed from a dataset without significantly altering the classification distribution depends on the dataset size, class balance, and the reason for removal. For small datasets (e.g., a few thousand rows), it is recommended to remove $\leq 5\%$ of the data to avoid impacting the class distribution. For large datasets (e.g., hundreds of thousands or millions of rows), a higher percentage of $5 - 20\%$ can often be removed, provided the removal is done in a stratified manner to preserve the class balance. When dealing with imbalanced datasets, rows from the majority class can be removed to balance the dataset. The acceptable percentage of removal typically ranges from $1 - 10\%$ for outliers, $5 - 30\%$ for missing data, and $5 - 15\%$ for noisy or inconsistent data, depending on the dataset and application. The key is to ensure that the removal process maintains the original class distribution, often achieved through stratified sampling, and to evaluate the impact on model performance to avoid degrading results.*

As a result, $\nu_r(X)$ is a measuring tool used to lay out insignificant objects that can be tolerated. To determine which $\nu_r(X)$ value is tolerable for any specific data, a certain margin of error can be selected depending on the structure of the data, by generally accepted methods and taking into account the distribution of $\nu_r(X)$ values. In a dataset consisting of 1000 rows (objects), attributes with $\nu_r(X) \leq 10$ values can be ignored with a 1% marjin of error. Then one can consider that an attribute r is not in the core because it could be tolerated for the data with 1000 objects. It is a method for determining which attributes are more important than others in this classification. It also lists the relative importance of these attributes.

5.1. Applications

Example 5.5. *The dataset was obtained from the UCI Machine Learning Repository. This dataset contains information about wart treatment results of 90 patients using cryotherapy [6]. Seven attributes are sex, age, time, number of warts, type, area, and result of treatment.*

Table 4: Cryotherapy Treatment

Objects	*Sex*	*Age*	*Time*	*Number of Warts*	*Type*	*Area*	*Results*
1	*1*	*35*	*12*	*5*	*1*	*100*	*0*
2	*1*	*29*	*7*	*5*	*1*	*96*	*1*
3	*1*	*50*	*8*	*1*	*3*	*132*	*0*
4	*1*	*32*	*11.75*	*7*	*3*	*750*	*0*
5	*1*	*67*	*9.25*	*1*	*1*	*42*	*0*
6	*1*	*41*	*8*	*2*	*2*	*20*	*1*
7	*1*	*36*	*11*	*2*	*1*	*8*	*0*
8	*1*	*59*	*3.5*	*3*	*3*	*20*	*0*
9	*1*	*20*	*4.5*	*12*	*1*	*6*	*1*
10	*2*	*34*	*11.25*	*3*	*3*	*150*	*0*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

By applying the algorithm 1 to Table 4, we get the following Table 5:

Table 5: Converted Cryotherapy Treatment

Objects	*Sex*	*Age*	*Time*	*Number of Warts*	*Type*	*Area*	*Results*
1	*2*	*3*	*1*	*2*	*3*	*2*	*0*
2	*2*	*3*	*2*	*2*	*2*	*2*	*1*
3	*2*	*2*	*2*	*3*	*1*	*2*	*0*
4	*2*	*3*	*1*	*2*	*1*	*1*	*0*
5	*2*	*1*	*1*	*3*	*3*	*2*	*0*
6	*2*	*2*	*2*	*3*	*2*	*3*	*1*
7	*2*	*3*	*1*	*3*	*3*	*3*	*0*
8	*2*	*1*	*3*	*3*	*1*	*3*	*0*
9	*2*	*4*	*3*	*1*	*3*	*3*	*1*
10	*1*	*3*	*1*	*3*	*1*	*2*	*0*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Let $U = \{1, 2, 3, \cdots, 88, 89, 90\}$ *be the set of objects, and* X *and* Y *be the set of objects having result* 0 *and* 1*, respectively. Let* R *be an equivalence (indiscernibility) relation, in which two objects are equivalent if their values of all attributes are the same. If attributes which have* $v_r(X) \leq 2$ *are eliminated, then the* $Core(X) = \{time, number\ of\ warts, area\}$ *(See Table 6). Similarly,* $Core(U - X) = Core(Y) = \{time, number\ of\ warts, area\}$.

Table 6: The Core of the Dataset

Column	*1*	*2*	*3*	*4*	*5*	*6*
Attribute(r)	*Sex*	*Age*	*Time*	*N. of Warts*	*Type*	*Area*
$v_r(X)$	*1*	*2*	*21*	*5*	*0*	*6*

As a result, key attributes to decide the result of the treatment are time, number of warts, and area. The order of significance of these attributes for the classification is time, area, and number of warts.

In this part, we apply 8 different machine learning algorithms, which are Support Vector Classifier (SVC), Random Forest Classifier (RFC), Linear Regression (LR), Gradient Boosting Classifier (GBC), Extreme Gradient Boosting (XGBC), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), and Hybrid (HYB), to the data by using all attributes and attributes in the core, respectively. Here, the Hybrid algorithm is the algorithm that is created based on whether 4 out of 7 algorithms predict correctly or not. Then, the results are compared for each method and each class. Looking at average classification accuracy, we get better results by using the core.

Machine learning algorithms are used by default setting. It is carried out without any optimizations or parameters using a random selection procedure. We split the data set so that 80% is used to train the model and 20% is used to test with a fixed random selection. We consider running the ML algorithms multiple times, comparing the average result, and then reporting metrics.

The bar chart Figure 1 illustrates the accuracy rates of several machine learning algorithms based on two sets of features: "Core Attributes" and "All Attributes." Across the majority of algorithms, using the "Core Attributes" yields a higher accuracy compared to "All Attributes," with notable examples including the SVC (0.556 vs 0.500) and HYB (0.889 vs 0.722) models. However, in the case of the RFC algorithm, the difference in accuracy between the two feature sets is marginal (0.833 vs 0.778), suggesting that adding more attributes did not significantly improve or harm its performance. For other algorithms such as LR, GBC, XGBC, and LDA, the "Core Attributes" consistently outperform "All Attributes" with a difference of about 0.1 in some instances, suggesting the core features contribute more meaningfully to the predictive power of these models. Overall, the "Core Attributes" feature set appears to perform better across most models, which may indicate a more focused and optimized feature selection. However, certain models like LDA and GNB show lower overall performance compared to others.

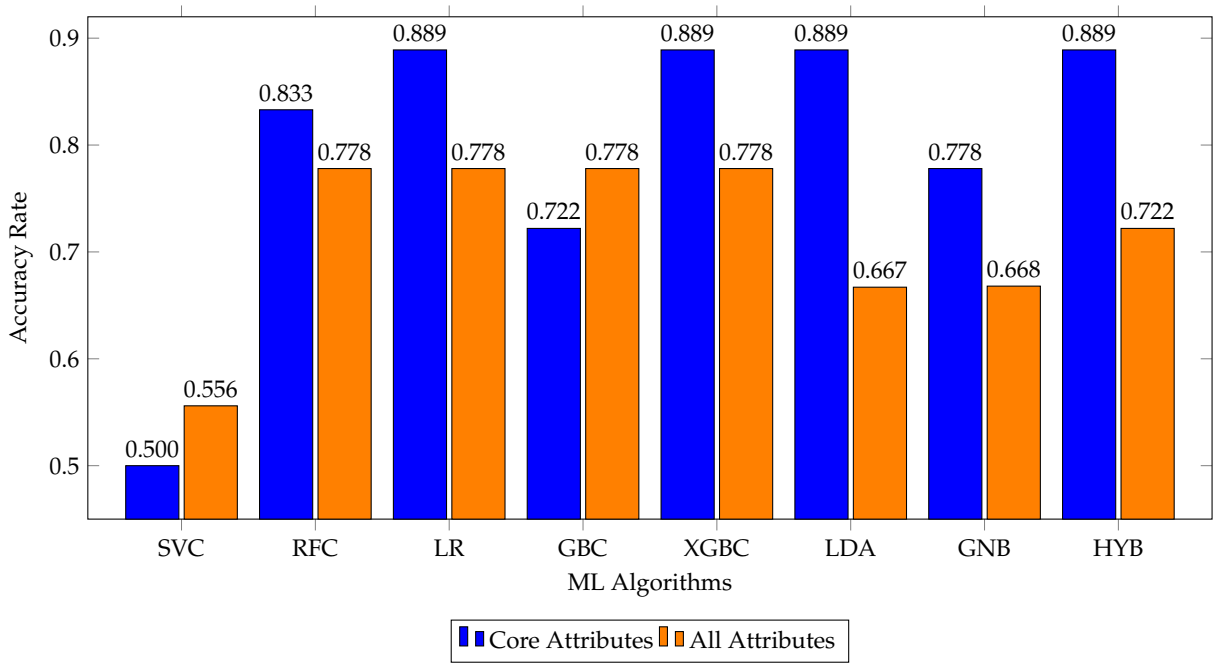

Figure 1: Comparison of Core Attributes and All Attributes for ML Algorithms

The following Table 7 demonstrates the classification outcomes of the eight ML algorithms.

Table 7: ML Algorithm Results

Method	*Use all attributes*	*Use the core*
SVC	*Accuracy:0.556* *Precision:0.529* *Recall:1* *F1:0.692*	*Accuracy:0.5* *Precision:0.5* *Recall:1* *F1:0.667*
RFC	*Accuracy:0.778* *Precision:0.727* *Recall:0.889* *F1:0.8*	*Accuracy:0.889* *Precision:1* *Recall:0.778* *F1:0.875*
LR	*Accuracy:0.778* *Precision:0.857* *Recall:0.667* *F1:0.75*	*Accuracy:0.889* *Precision:1* *Recall:0.778* *F1:0.875*
GBC	*Accuracy:0.778* *Precision:0.727* *Recall:0.889* *F1:0.8*	*Accuracy:0.833* *Precision:0.8* *Recall:0.889* *F1:0.842*
XGBC	*Accuracy:0.778* *Precision:0.727* *Recall:0.889* *F1:0.8*	*Accuracy:0.889* *Precision:1* *Recall:0.778* *F1:0.875*
GNB	*Accuracy:0.667* *Precision:0.636* *Recall:0.778* *F1:0.7*	*Accuracy:0.778* *Precision:0.778* *Recall:0.778* *F1:0.778*
LDA	*Accuracy:0.667* *Precision:0.667* *Recall:0.667* *F1:0.667*	*Accuracy:0.889* *Precision:1* *Recall:0.778* *F1:0.875*
HYB	*Accuracy:0.722* *Precision:0.667* *Recall:0.889* *F1:0.762*	*Accuracy:0.889* *Precision:1* *Recall:0.778* *F1:0.875*

Recall that the class-wise distribution of a classification model's predicted performance is called a confusion matrix. A table used to assess a model's performance in classification tasks is called a 2×2 *confusion matrix. It consists of four cells: False Positives (FP), where the model predicts a positive class for a negative instance; False Negatives (FN), where it predicts a negative class for a positive instance; True Positives (TP), where the model predicts the positive class correctly; and True Negatives (TN), where it predicts the negative class correctly. This matrix aids in the computation of important metrics such as F1-score, recall, accuracy, and precision. The confusion matrices we generated for the hybrid algorithm (which achieves the highest accuracy with Core attributes) employing all of the features and the features in the core are shown below in Figures 3 and 2, respectively.*

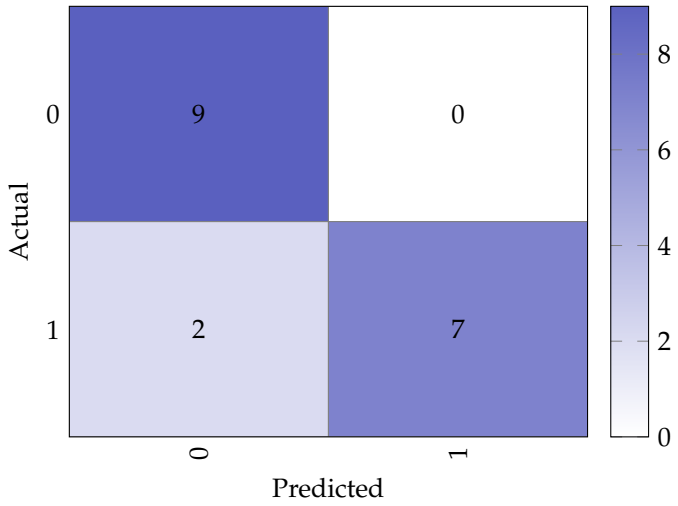

Figure 2: Confusion Matrix for the Hybrid Algorithm (The Core)

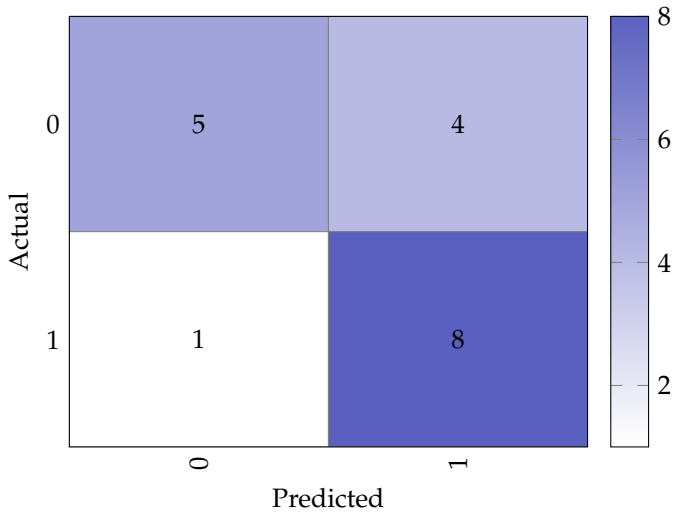

Figure 3: Confusion Matrix for the Hybrid Algorithm (All Attributes)

5.1.1. Comparison with PCA

The algorithm introduced, which we call as topological dimension reduction (TDR) is essentially a dimensionality reduction technique. Unlike other methods in this area, it maintains the original data set's properties without actually erasing any data. The user can also select the number of dimensions to eliminate, since it indicates the qualities' degree of importance. We also shared cases where the entire data set was used during comparisons. However, our main rival in this area is Principal Component Analysis (PCA), the most well-known dimensionality reduction algorithm. As seen in Figure 4, we generally performed better than PCA but not SVC. It functions remarkably fast even though it is totally developed in Python.

The figure 4 compares the accuracy rates of different machine learning (ML) algorithms using Core Attributes (blue bars) and Principal Component Analysis (PCA) (green bars). The results indicate that Core Attributes consistently outperform PCA across all models, but no SVC. The highest accuracy of 0.889 is observed in LR, XGBC, LDA, and HYB when using Core Attributes, while PCA generally results in lower accuracy, often around 0.5. The only exceptions where PCA performs moderately well are GBC (0.667) and LDA (0.667). Notably, models like RFC and GNB show significant drops in accuracy when using PCA compared to Core Attributes. This suggests that Core Attributes retain more valuable information for classification than PCA, which may discard critical features during dimensionality reduction.

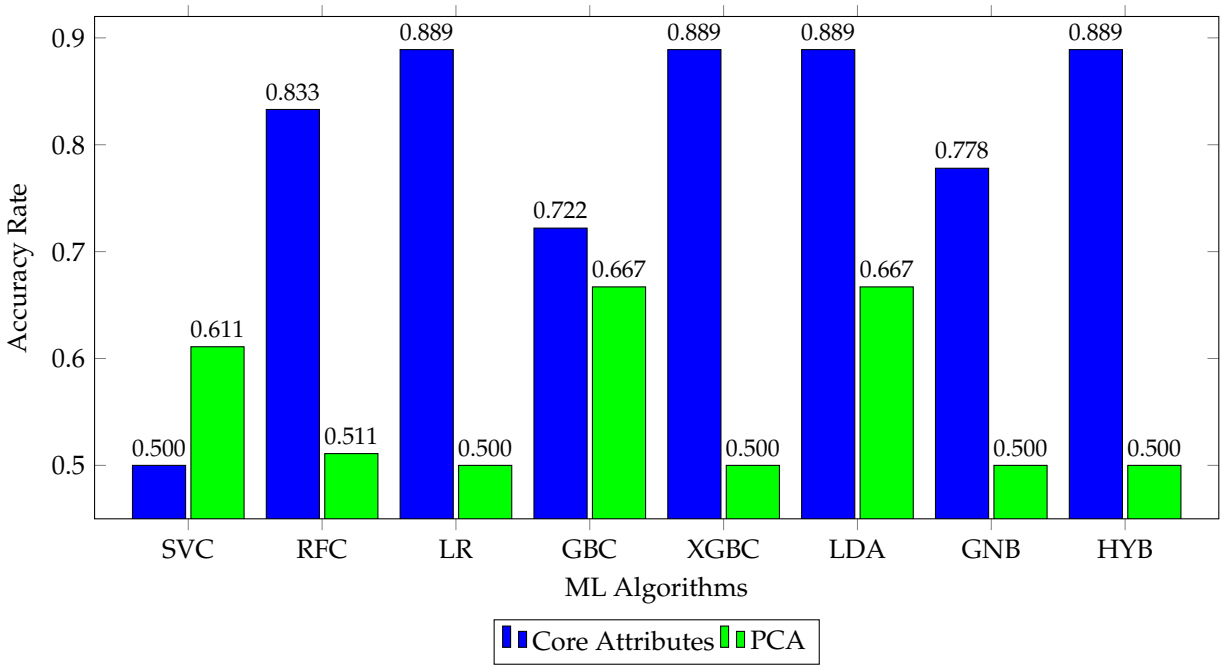

Figure 4: Comparison of Core Attributes and PCA for ML Algorithms

5.2. An Application to Dataset with Categorical Decision Attribute

It can also be applied to data with that a decision attribute takes more than two values. One can get better classification results by using the core rather than using all attributes by applying machine learning algorithms to most numerical data. Also, with $\nu_r(X)$, in the classification, the most important attributes and their importance ranking in deciding which class an object is in can be given.

Let U be the set of objects, and $X_1, X_2, \cdots, X_n$ be the set of objects having result $1, 2, \cdots, n$, respectively. Let R be an equivalence (indiscernibility) relation, in which two objects are equivalent if their values of all attributes are the same. Then, we have $Core(X_i) = Core(U - X_i)$ for all $i = 1, 2, \cdots, n$. However, $Core(U - X_i)$ does not need to equal to $Core(U - X_j)$ for all $1 \leq i, j \leq n$. Having said that, the most important attributes

and their relative weights in determining the class an object belongs to can be determined using $v_r(X_i)$ in the classification process for all $i = 1, 2, \cdots, n$.

Example 5.6. *The dataset was obtained from the UCI Machine Learning Repository about measurements of geometrical properties of three different types of wheat kernels: Kama, Rosa, and Canadian, 70 objects each [19]. In the data, seven attributes of wheat kernels were measured:*

A: area,
B: perimeter,
C: compactness,
D: length of kernel,
E: width of kernel,
F: asymmetry coefficient,
G: length of kernel groove,
Result: kernel type (Kama=1, Rosa=2, Canadian=3).

Table 8: Wheat Kernel Data

Objects	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*Results*
1	*15.26*	*14.84*	*0.871*	*5.763*	*3.312*	*2.221*	*5.22*	*1*
2	*14.88*	*14.57*	*0.8811*	*5.554*	*3.333*	*1.018*	*4.956*	*1*
3	*14.29*	*14.09*	*0.905*	*5.291*	*3.337*	*2.699*	*4.825*	*1*
4	*13.84*	*13.94*	*0.8955*	*5.324*	*3.379*	*2.259*	*4.805*	*1*
5	*16.14*	*14.99*	*0.9034*	*5.658*	*3.562*	*1.355*	*5.175*	*1*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
71	*17.63*	*15.98*	*0.8673*	*6.191*	*3.561*	*4.076*	*6.06*	*2*
72	*16.84*	*15.67*	*0.8623*	*5.998*	*3.484*	*4.675*	*5.877*	*2*
73	*17.26*	*15.73*	*0.8763*	*5.978*	*3.594*	*4.539*	*5.791*	*2*
74	*19.11*	*16.26*	*0.9081*	*6.154*	*3.93*	*2.936*	*6.079*	*2*
75	*16.82*	*15.51*	*0.8786*	*6.017*	*3.486*	*4.004*	*5.841*	*2*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
141	*13.07*	*13.92*	*0.848*	*5.472*	*2.994*	*5.304*	*5.395*	*3*
142	*13.32*	*13.94*	*0.8613*	*5.541*	*3.073*	*7.035*	*5.44*	*3*
143	*13.34*	*13.95*	*0.862*	*5.389*	*3.074*	*5.995*	*5.307*	*3*
144	*12.22*	*13.32*	*0.8652*	*5.224*	*2.967*	*5.469*	*5.221*	*3*
145	*11.82*	*13.4*	*0.8274*	*5.314*	*2.777*	*4.471*	*5.178*	*3*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

By applying the algorithm 1 to the Table 8, we get the following Table 9:

Table 9: Converted Wheat Kernel Data

Objects	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*Results*
1	*2*	*2*	*2*	*3*	*2*	*4*	*3*	*1*
2	*3*	*3*	*2*	*3*	*2*	*5*	*4*	*1*
3	*3*	*3*	*1*	*4*	*2*	*4*	*4*	*1*
4	*3*	*3*	*1*	*4*	*2*	*4*	*4*	*1*
5	*2*	*2*	*1*	*3*	*2*	*5*	*3*	*1*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
71	*2*	*1*	*3*	*2*	*2*	*3*	*1*	*2*
72	*2*	*2*	*3*	*2*	*2*	*3*	*2*	*2*
73	*2*	*2*	*2*	*2*	*2*	*3*	*2*	*2*
74	*1*	*1*	*1*	*2*	*1*	*4*	*1*	*2*
75	*2*	*2*	*2*	*2*	*2*	*3*	*2*	*2*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
141	*3*	*3*	*3*	*3*	*3*	*2*	*3*	*3*
142	*3*	*3*	*3*	*3*	*3*	*1*	*3*	*3*
143	*3*	*3*	*3*	*3*	*3*	*2*	*3*	*3*
144	*3*	*3*	*3*	*4*	*3*	*2*	*3*	*3*
145	*4*	*3*	*4*	*4*	*4*	*3*	*3*	*3*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Let $U = \{1, 2, 3, \cdots, 208, 209, 210\}$ *be the set of objects, and* X=*{1, 2, 3,* $\cdots$*, 68, 69, 70},* Y=*{71, 72 , 73,* $\cdots$*, 138, 139, 140} and* Z=*{141, 142, 143,* $\cdots$*, 208, 209, 210} be the set of objects having result* 1, 2, *and* 3, *respectively. Let R be an equivalence (indiscernibility) relation, in which two objects are equivalent if their values of all attributes are the same.*

Table 10: The Core of Dataset

Column	1	2	3	4	5	6	7
Attribute (r)	A	B	C	D	E	F	G
$v_r(X)$	0	1	25	1	11	67	55
$v_r(Y)$	0	0	4	1	0	23	52
$v_r(Z)$	0	1	21	0	11	44	3

As a consequence, firstly, $Core(X) = \{C, E, F, G\}$. *The order of importance of the attributes for the objects classified as in X is* F, G, C, E. *Secondly,* $Core(Y) = \{C, F, G\}$. *The order of importance of the attributes for the objects classified as in Y is* G, F, C. *Lastly,* $Core(Z) = \{C, E, F, G\}$. *The order of importance of the attributes for the objects classified as in Z is* F, C, E, G. *Finally, key attributes to decide the type of seeds are* C, E, F, *and* G *columns.*

6. Conclusion and Future Directions

In this work, we give the rough topology and core for numerical data. By using the core, one can also get better results by using fewer attributes (attributes in the core) for machine learning algorithms. As a result, the rough topology is a useful model for attribute reductions for numerical data. This method could

be applied to big numerical data for future selection problems in future research. We will use this method to make predictions of qualities such as maintenance cost overruns, work accidents, and the severity of construction quality failures for datasets.

Another project is to apply the rough topology to solve the missing values problem in incomplete numerical information tables. The rough topological method for numerical data could be given like Salama's work for Boolean Type of data [17, 21].

The method constructed in this article can be generalized by combining with or applying techniques such as covering approximation spaces, initial-rough sets, β-based rough sets, and generalized rough sets as another future project.

Finally, this new rough topological method can be given for expansions of rough sets in incomplete information systems by taking tolerance relations rather than indiscernibility relations to apply missing value problems for numerical data [10, 21].

Acknowledgements

I would like to thank Kenan Evren Boyabatlı for helping me to write Python codes. Also, the author sincerely thanks the referees and editors for their helpful suggestions and comments that improved the paper.

Code Availability Statement: Some or all models, or codes that support the findings of this study are available from the corresponding author upon reasonable request. See the link URL:

`https://github.com/EvReN-jr/TDR-Topological-Dimensional-Reduction`

for Python and C++ codes to find the converted data, the approximations, topology, and the Core.

References

[1] M. E. Abd El-Monsef, A. M. Kozae, and M. K. El-Bably, Generalized covering approximation space and near concepts with some applications, Appl. Comput. Inform. **12** (2016), 51–69. doi: 10.1016/j.aci.2015.02.001

[2] A. Abd Atik and M. El-Bably, Soft β-rough sets and its application to determine COVID-19, Turk. J. Math. **45** (2021). doi: 10.3906/MAT-2008-93

[3] D. I. Taher, R. Abu-Gdairi, M. K. El-Bably, and M. A. El-Gayar, Decision-making in diagnosing heart failure problems using basic rough sets, AIMS Math. **9** (2024), 21816–21847. doi: 10.3934/math.20241061

[4] D. I. Taher, R. Abu-Gdairi, M. K. El-Bably, and M. A. El-Gayar, Correction: Decision-making in diagnosing heart failure problems using basic rough sets, AIMS Math. **9** (2024), 34270–34271. doi: 10.3934/math.20241632

[5] R. Abu-Gdairi and M. K. El-Bably, The accurate diagnosis for COVID-19 variants using nearly initial-rough sets, Heliyon **10** (2024), Art. e31288. doi: 10.1016/j.heliyon.2024.e31288

[6] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, An expert system for selecting wart treatment method, Comput. Biol. Med. **81** (2017), 167–175. doi: 10.1016/j.compbiomed.2017.01.001

[7] M. K. El-Bably, R. Abu-Gdairi, K. K. Fleifel, and M. A. El-Gayar, Exploring β-Basic Rough Sets and Their Applications in Medicine, Eur. J. Pure Appl. Math. **17** (2024), 3743–3771. doi: 10.29020/nybg.ejpam.v17i4.5545

[8] R. A. Hosny, R. Abu-Gdairi, and M. K. El-Bably, Enhancing Dengue fever diagnosis with generalized rough sets: Utilizing initial-neighborhoods and ideals, Alex. Eng. J. **94** (2024), 68–79. doi: 10.1016/j.aej.2024.03.028

[9] J. Järinen, Approximations and rough sets based on tolerances, in: W. Ziarko and Y. Yao (Eds.), Rough Sets and Current Trends in Computing, Lecture Notes in Comput. Sci., vol. 2001, Springer, Berlin, 2001, pp. 182–189. doi: 10.1007/3-540-45554-X_21

[10] M. Kryszkiewicz, Rough set approach to incomplete information systems, Inform. Sci. **112** (1998), 39–49.

[11] J. R. Munkres, Topology, Prentice Hall, Upper Saddle River, NJ, 2000.

[12] A. A. Nasef, M. Shokry, and O. M. Mukhtar, Some methods to reduction on electrical transmission lines by using rough concepts, Filomat **34** (2020), 111–128.

[13] A. S. Nawar, R. Abu-Gdairi, M. K. El-Bably, and H. M. Atallah, Enhancing Rheumatic Fever Analysis via Tritopological Approximation Spaces for Data Reduction, Malays. J. Math. Sci. **18** (2024), 321–341. doi: 10.47836/mjms.18.2.07

[14] Z. Pawlak, Rough sets, Int. J. Comput. Inform. Sci. **11** (1982), 341–356. doi: 10.1007/BF01001956

[15] Z. Pawlak, Rough set theory and its applications, J. Telecommun. Inform. Technol. **3** (2002), 7–10.

[16] Z. Pawlak, Some issues on rough sets, Lect. Notes Comput. Sci. **3100** (2004), Springer, Berlin, Heidelberg. doi: 10.1007/978-3-540-27794-1_1

[17] A. Salama, Topological solution of missing attribute values problem in incomplete information tables, Inform. Sci. **180** (2010), 631–639. doi: 10.1016/j.ins.2009.11.010

[18] M. El Sayed, M. A. El Safty, and M. K. El-Bably, Topological approach for decision-making of COVID-19 infection via a nano-topology model, AIMS Math. **6** (2021), 7872–7894. doi: 10.3934/math.2021457

[19] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak, Complete gradient clustering algorithm for features analysis of X-ray images, in: Information Technologies in Biomedicine, Lect. Notes Comput. Sci. **69** (2010), 15–24. doi: 10.1007/978-3-642-13105-9_2

[20] M. L. Thivagar, C. Richard, and N. R. Paul, Mathematical innovations of a modern topology in medical events, Int. J. Inform. Sci. **2** (2012), 33–36. doi: 10.5923/j.ijis.20120204.01

[21] X. Yang and J. Yang, Incomplete information system and rough set theory: Models and attribute reductions, Springer, Berlin, 2013. doi: 10.1007/978-3-642-25935-7

[22] W. Zhu and F.-Y. Wang, Reduction and axiomization of covering generalized rough sets, Inform. Sci. **152** (2003), 217–230.