Research Article

Tamilarasu Sangeetha and Amalanathan Geetha Mary*

Rough set-based entropy measure with weighted density outlier detection method

https://doi.org/10.1515/comp-2020-0228
received June 23, 2020; accepted May 10, 2021

Abstract: The rough set theory is a powerful numerical model used to handle the impreciseness and ambiguity of data. Many existing multigranulation rough set models were derived from the multigranulation decision-theoretic rough set framework. The multigranulation rough set theory is very desirable in many practical applications such as high-dimensional knowledge discovery, distributional information systems, and multisource data processing. So far research works were carried out only for multigranulation rough sets in extraction, selection of features, reduction of data, decision rules, and pattern extraction. The proposed approach mainly focuses on anomaly detection in qualitative data with multiple granules. The approximations of the dataset will be derived through multiequivalence relation, and then, the rough set-based entropy measure with weighted density method is applied on every object and attribute. For detecting outliers, threshold value fixation is performed based on the estimated weight. The performance of the algorithm is evaluated and compared with existing outlier detection algorithms. Datasets such as breast cancer, chess, and car evaluation have been taken from the UCI repository to prove its efficiency and performance.

Keywords: approximations, entropy, granules, outliers, rough sets

*** Corresponding author: Amalanathan Geetha Mary,** School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632 001, Tamil Nadu, India,
e-mail: geethamary.a@gmail.com
Tamilarasu Sangeetha: School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632 001, Tamil Nadu, India, e-mail: sangee_arasu05@yahoo.co.in

1 Introduction

Data may be in the form of text, numbers, or mixed types where a system can be easily identified and processed. Data have different structures and dimensions. Data mining is a technology that is used to obtain information from larger databases. Objects that deviate from others based on characteristics or behavior are anomalies [1]. When a machine fails to work properly, a system that does not respond properly to given inputs, manmade faults, a simple diversion in population, and fraudulent activities are the causes for arising outliers. Outliers are identified through a different kind of pattern generation in application areas like medical databases, cyber security systems, drastic effects of climatic variations, and also in military systems.

Some of the different types of outliers are point outliers, contextual outliers, and collective outliers. The point anomaly is also known as a global outlier where an object deviates from the rest of the objects [2]. For example, the broadcasting of packets in a very short period by a computer is identified as the victim of hacking. Contextual outliers are objects that deviate only from a particular situation [3]. For example, 28°C in Chennai is considered normal during summer, but the same will be taken as an outlier during winter. A group of objects, in particular, which deviates from the whole dataset, are called a collective outlier. A group of students who are irregular in their studies among the class is termed as outliers.

The proven theories, methods, and techniques for granular computing use granules in the form of classes, groups, or clusters in the universe. Some of the application domains such as analysis of intervals, clusters, retrieval of information from the databases, machine learning algorithms, Dempster–Shafer theories, and divide–conquer methods use granular computing technique. Sometimes, the available data are incomplete and unclear [4]. So granular computing is needed in this context to make the problem simple. Acquiring precise information costs very high, and granulation of data reduces the cost. The granularity of data can be achieved through proximity relation, the similarity between data. The granularity of data can be achieved through the

approximation concept. There exists a connection between granular computing and rough sets. Partitions can be made from the attributes, and approximations are defined.

This article provides a proposed method for detecting outliers in multigranular rough sets with a multiequivalence relation. The definitions for the positive and negative regions for multigranular rough sets are discussed in Section 2. A detailed literature review about the multigranular rough sets is discussed in Section 3. The proposed model and proposed algorithm are provided in Section 4. The empirical study and experimental analysis of a proposed system are presented in Sections 5 and 6.

2 Background

2.1 Rough set theory

Rough set concepts were developed by Pawlak [5] in 1980s. It is a powerful mechanism that used to handle the uncertainty and vagueness of data. It can be applied in all domains, particularly in artificial intelligence. The primary benefit of rough sets is that all the variables needed for computation are retrieved from the available dataset. The imprecise information can be obtained by the concept of approximations because there is no need of knowing preliminary information about the data. A dataset has been taken to represent knowledge. Columns of the dataset are labeled as an attribute, and rows of a dataset are labeled as objects. The subset formed from attributes associates a single or multiequivalence relation, whereas a group of objects forms a set. The knowledge of data can be derived from attributes, and the concepts can be derived from objects. The basic ideas of rough sets are as follows: a relational database has to be classified to generate rules and to make ideas, and knowledge can be obtained through equivalence classes and approximation concept.

The rough set theory provides a solution to all the major domains in artificial intelligence. Rough set concepts are used to determine the model and activities of the drug; attitude control of satellites, iron, and steel blast furnace; diagnosis of machines, in the area of neural network and also in decision support systems [6]. With the help of algorithms, the rough set theory detects unseen data, recognizes the link between data that cannot be analyzed by statistical methods, allows both measurable and quantifiable data, can obtain decision rules, leads to reduction of data, analyses data very significantly, can be learned easily, and the output can be interpreted directly without any prior knowledge.

In the classical rough set model, a lower approximation can be determined by equivalence classes that are a subset of the objective set, and in upper approximation, equivalence classes should be nonempty and overlapped with the objective set [7]. There is no error tolerance in this classical approach. In probabilistic rough sets, the approximation concepts are based on rough membership function and inclusion method. The decision-theoretic rough set model uses α and β for acceptance and rejection, and the values are defined between 0 and 1 [8].

In the Pawlak rough set model, the equivalence classes should be reflexive, symmetric, and transitive. If the transitive relation does not obtain equivalence relation, it has to be replaced with tolerance relation. Mostly, rough set concepts are used in the classification of data [5]. Pawlak's rough set model is very sensitive to noisy data. It can be identified by fixing a probabilistic threshold value β within the range of 0–0.5 based on the level of noisy data. This can be achieved by a variable precision rough set model. In the multigranulation rough set model, the multiequivalence classes are derived to achieve the goal [9]. To make intelligent decisions in critical situations, the game-theoretic rough set model has been used. Each player should possess the value of α or β based on the parameter region. Their probabilistic values should be either increased or decreased slightly. Decreased value of α indicates a positive region, and an increased value of β indicates a negative region. The granular structure has been introduced through equivalence relation, not by rough set data analysis. This article proposes outlier detection of categorical data in multigranular rough sets using the rough entropy-based weighted density method.

2.2 Multigranulation rough set

Any information system will have n number of attributes and objects. It can hold sometimes missing or null values, which are termed to be irregular. If a universe U contains regular objects and attributes, they are termed to be complete information system otherwise if they have irregular objects and attributes are known as incomplete information system [10]. Pawlak's rough set lower approximation was fully dependent on single binary relation, whereas, in multigranulation rough set, the lower approximation will be derived by using multiequivalence relation. In both cases, the upper approximation will be derived based on the complementary set of lower approximation.

Let us consider T to be the universe and $A \subseteq \hat{Y}$, $\hat{Y}$ be the partition of T. The approximation concept of SGRS is characterized as follows:

$$\underline{A} = \cup\{B \in \hat{Y} : B \subseteq A\}, \tag{1}$$

$$\bar{A} = \cup\{B \in \hat{Y} : B \cap A \neq \varnothing\}. \tag{2}$$

Also, optimistic multigranularity of the rough set model provides many individual granular structures that need minimum one granular structure to satisfy the inclusion condition between equivalence class and objective set, while in pessimistic multigranulation rough set model, the granular structure of minimum one should be a nonempty intersection of the objective. Let us consider a complete information system $T = (U, \text{Atr}, \text{fn})$ and $\hat{M}$, $\hat{N}$ be two segments over the universe U, $A \subseteq U$. The lower and upper approximation of MGRS is defined by the following formulae:

$$\underline{A} = \{a : \hat{M}(a) \subseteq A \text{ or } \hat{N}(a) \subseteq A\}, \tag{3}$$

$$\bar{A} = \sim(A)\,\hat{M} + \hat{N}\,. \tag{4}$$

In Figure 1, the small circles with shaded region $[a]_x$ and $[a]_y$ are lower approximation under MGRS and the big circles with shaded region $[b]_x$ and $[b]_y$ are lower approximation under SGRS.

2.3 Related work

Outlier objects are defined as a single object's anomalous behaviour or small groups of objects that are more inconsistent than the rest. There is a chance of abnormal occurrences in spatial or temporal locality that forms a cluster known as anomalies or outliers. They used the LDBSCAN algorithm for clustering and LOF to detect the inconsistency of a single object [11]. Detecting outliers is the primary step in the applications of data mining. They have proposed many outlier detection algorithms for parametric and nonparametric, univariate, and multivariate. Outlier detection techniques are also based on spatial, distance-based, and density-based clustering methods. If outliers exist in the dataset, individual observation should be taken to maintain robustness by providing suitable estimators.

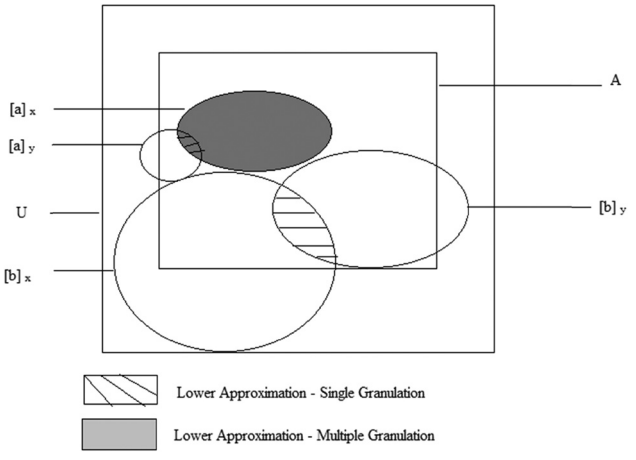

Figure 1: Difference between SGRS and MGRS [17].

An object that is dissimilar from the rest of the objects is an anomaly. First, frequent patterns of a dataset are generated. Items that are having lower frequent patterns are outliers. They have designed the frequent pattern outlier factor (FPOF) to detect transactions that are outliers, and to identify outliers alone, the find FPOF method needs to be used. Researchers show their interest when trying to find rare events than frequent patterns. Existing works show that being an outlier object is a binary property. Each object is assigned with a degree of score to be an outlier [12]. By using the local outlier factor (LOF), the neighborhood of an object with its surroundings and how much it is isolated from others are calculated. It is crucial to detect outliers in many application areas. The topic of determining outlier scores was an extension of objects in terms of clusters. The individual cluster has its outlier factor, which is the clustering-based outlier method [13]. It has two stages: the first stage forms clusters based on the clustering algorithm, and the second stage detects outliers based on the outlier factor.

A new definition was given to identify outliers based on the local outlier factor, which shows the importance of the behavior of data, which is local. Cluster-based local outlier factor (CBLOF) was defined to measure and represent the natural quality of outliers. Outliers were presented based on k nearest neighbor graph with outlier indegree factor [14]. Also, they have extended the work of k nearest neighbor clustering work. Existing outlier detection methods are not suitably fit for scattered real-world data due to parameter issues and data patterns, which are implicit. So, they had proposed the local density outlier factor (LDOF) to measure the distance around its neighbor [15]. If the distance is farther, then the isolated objects or small clusters are known as outliers. k means is the most popular clustering algorithm to form clusters on a dataset. However, it works only for a fixed data stream, which fails when data streams are dynamic [16]. The mean of previous clusters is compared with the current cluster to detect candidate outliers effectively. Neural network-based learning technique uses SOM and ART. The SOM algorithm builds to map a high-

dimensional input space to low-dimension output space by assuming that the topological structure exists in the input space.

Classification based on an optimistic multigranulation rough set model was proposed for the medical diagnosis system [18]. From the generated patterns, the initial cause for the disease and its symptoms can be diagnosed. They proved that a single granular method provides effective results when compared with the multigranular method [19]. The approximation concept of the classical approach was based on a single binary relation on the universe. But multiple granularities of approximations were based on multiequivalence relation. The decision rule of SGRS follows the "AND" rule, and MGRS follows the "OR" rule [20]. The tolerance of incomplete rough sets could not be determined with the help of single granulation, but it can be achieved with the help of multiple granularities. Plenty of algorithms are available for attribute reduction and for making decision rules. In both data mining and machine learning, the test cost is not taken into consideration, which can be solved by the multigranulation rough set [21]. This test cost method is a generalization of three methods such as optimistic, pessimistic, and β MGRS [22]. The approximation concept of optimistic MGRS forms a lattice. The formed lattice should not be distributive and complemented, and it is equivalent to a single granulation model. But pessimistic multigranulation forms a clopen topology on the available dataset, which forms a normal Boolean algebra. The MGRS model has been generalized into fuzzy sets. The single relationship of the fuzzy rough set and optimistic and pessimistic MGRS model was discussed. By combining the idea of SGRS and MGRS, a Bayesian decision-theoretic rough set [23] was developed using a probabilistic theory that converts the parameters into rough sets.

Outlier detection, as well as background knowledge about the domain, was obtained by applying the scheme of multilevel approximation. With the help of data, table outliers were detected by using the granularity method. Outliers are detected by assigning scores to local outlier factors and class outlier factors. Also, the rough membership function is used to determine outliers [24]. A multigranulation rough set model was developed based on SCED (seeking common ground while eliminating differences), which was also termed as pessimistic multigranulation rough set model. From this, attribute reduction, approximations, and decision rule were induced.

On multiple granulations, a characteristic function and parameter, which is called the level of information, are added to determine an object, to support the precise information. When the size of the neighborhood is zero, a new rough set model has degenerated to normal multiple granularities [25]. The neighborhood-based multiple granular approaches extend the application in different domains. The approximation was built on combined relation, and the groups of equivalence relations is brought into an equivalence relation through the union and intersection sets.

The intuitionistic fuzzy multiple granularities were developed to generalize the existing three intuitionistic fuzzy rough set model and their extensions to remove redundancy in multigranular structures [26]. The differences and relationships between multiple granularities and multiple granular covering rough set models are determined [13]. The constraints of two MGCRS form a lattice. In real life, there may exist two different universes map under a multigranulation rough set, and a decision has been made with optimistic and pessimistic multiple-granulation method [27]. The standard rough set model's idea of approximation is based on a single equivalence relation, but the multigranulation rough set model uses several equivalence relations, and multi granulation fuzzy approximation spaces (MGFAS) identifies six types of rough approximations. Hence, they proved that fuzzy binary relations are the nearest pair to the undefined set, and pessimistic-based multigranulation upper and lower approximations are the farthest pair to the undefined set [28].

A new approach called rough topology has been developed to analyze many medical problems. The rough lower and upper approximation, boundary region, and core reduct are made to find out the key element, which is the cause of disease occurrence [29]. By using the concept of lower approximation reduct, a matrix with discernibility and theorem for judgment has been developed to make a fuzzy decision. For a fuzzy-based incomplete information system, the multigranulation rough set has been applied based on dominance relation. Hence, the dominance multigraulation rough set [17] has been established, and the three kinds of "OR" decision rules are obtained.

3 Attribute reduction

The attribute reduct concepts have been given more importance in rough sets. To maintain a good accuracy level, the original dataset is reduced into different subfragments. The attribute selection process uses a reduct method to remove attributes that are considered to be weak (less strength) or unnecessary [30]. The indispensable attribute of the dataset defined in Table 1 is determined based on the association rule strength or

Table 1: Hiring dataset

Objects	Degree	Experience	Reference	Decision
*E*1	MTech	High	Big	Yes
*E*2	MSc	High	Big	Yes
*E*3	MSc	Medium	Small	No
*E*4	MSc	Medium	Small	No
*E*5	MSc	High	Small	No
*E*6	MSc	Medium	Medium	Yes
*E*7	MTech	Low	Medium	No
*E*8	ME	Low	Medium	Yes

confidence [13]. The rule is defined as the ratio of several samples E_i, which contains E_i *U decision* to the number of samples that contain E_i. Table 1 presents the hiring dataset [31], with conditional attributes such as {*degree*, *experience*, *reference*} and one {*decision*} attribute.

The strength of rules for attribute degree is as follows:

- (Degree = MTech) → (Decision = Yes), rule strength → 50%.
- (Degree = MSc) → (Decision = Yes), rule strength → 40%.
- (Degree = MSc) → (Decision = No), rule strength → 60%.
- (Degree = MTech) → (Decision = No), rule strength → 50%.
- (Degree = ME) → (Decision = Yes), rule strength → 100%.

The strength of rules for attribute experience is as follows:

- (Experience = High) → (Decision = Yes), rule strength → 66%.
- (Experience = Medium) → (Decision = No), rule strength → 66%.
- (Experience = High) → (Decision = No), rule strength → 33%.
- (Experience = Medium) → (Decision = Yes), rule strength → 33%.
- (Experience = Low) → (Decision = No), rule strength → 50%.
- (Experience = Low) → (Decision = Yes), rule strength → 100%.

The strength of rules for attribute reference is as follows:

- (Reference = Big) → (Decision = Yes), rule strength → 100%.
- (Reference = Small) → (Decision = No), rule strength → 100%.
- (Reference = Medium) → (Decision = Yes), rule strength → 66%.
- (Reference = Medium) → (Decision = No), rule strength → 33%.

By rule generation, it can be easily identified that attribute Degree and reference have maximum strength when compared with attribute experience. Table 2 shows the indispensable attributes of the hiring dataset.

4 Proposed model

Outlier detection plays a key role in all application domains. The missing values and incomplete data presented in the data table provide ambiguousness, while compiling the data results in the erroneous output [32]. To avoid such kind of scenario, outlier detection is needed. Several methods are employed for outlier detection in the qualitative, quantitative, and mixed types of data. The proposed model detects outlier in the multigranulation rough set with lower and upper approximated values. Approximations are derived through multiequivalence relations with segments of attributes. The given input should be categorical.

In the preprocessing stage, through multiequivalence relation, the lower and upper approximation of the dataset is derived. Then, at the postprocessing stage, the rough set-based entropy measure outlier detection method is applied to the approximation sets. By fixing the appropriate value for the threshold, outliers are identified. The steps are clearly shown in Figure 2.

4.1 Rough set-based entropy measure with weighted density outlier detection method

A dataset may incorporate missing information, some negative and invalid data. So the dataset is characterized to be unclear and deficient. To handle this context, a rough

Table 2: Indispensable attributes

Objects	Degree	Reference
*E*1	MTech	Big
*E*2	MSc	Big
*E*3	MSc	Small
*E*4	MSc	Small
*E*5	MSc	Small
*E*6	MSc	Medium
*E*7	MTech	Medium
*E*8	ME	Medium

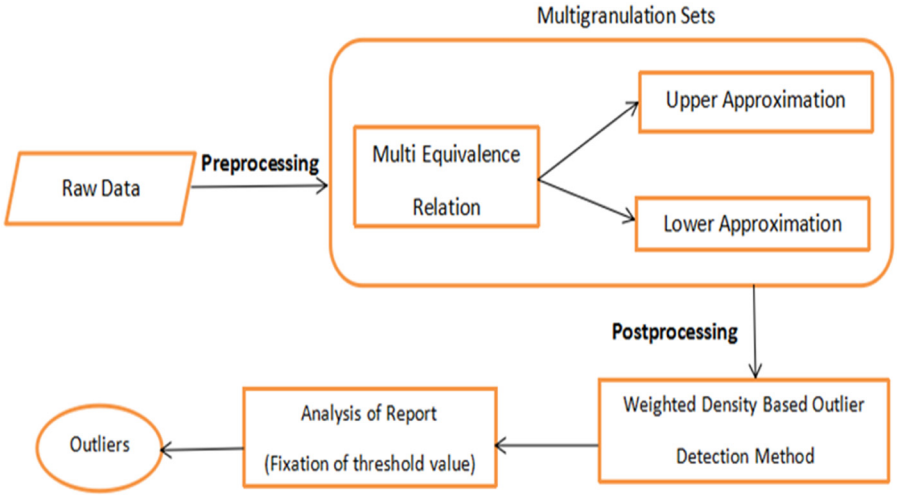

Figure 2: The proposed model for outlier detection in MGRS.

set-based entropy measure with a weighted density outlier detection method is proposed for multigranulation rough sets [33]. Based on the multiequivalence relation, the upper and lower approximation will be derived. In the postprocessing stage, the indiscernibility relation concerning objects is determined, the objects that are having uncertain values are calculated by complement entropy measure, and then the weighted density values will be calculated for every attribute and object [34]. The threshold value will be fixed from obtained values. Values lower than the threshold value are denoted as outliers. A higher threshold value is fixed for stable data, and the lower threshold value for unstable data is required to detect outliers. Sometimes we need prior knowledge from experts to fix proper values for the parameter [35]. It is not easy to maintain uniformity in fixing threshold values that can be applied to all data sets. The following definitions will be used to detect outliers, which are discussed below:

Definition 1. A dataset DST is defined by the triplet DST = (T, P, Q), where T represents the universe, P represents the objects, and Q represents the attributes in a dataset.

Definition 2. Let DST = (T, P, Q) and RY$\subseteq Q$. The indiscernible relation of RY for p_i in P or q_i in Q is represented as follows:

$$\{T|\text{ind(RY)}\} = \{[p_i]_{\text{RY}}|p_i \in T\}.$$

Definition 3. Let DST = (T, P, Q), and RY$\subseteq Q$ and $\frac{T}{\text{ind(RY)}}\{Q_1, Q_2 \ldots Q_m\}$. The complement entropy (CPME) with respect to RY is defined as follows:

$$\text{CPME(RY)} = \sum_{j=1}^{n} \frac{|Q|}{|T|} \frac{|Q|_j^q}{|P|},$$

where Q_j^q denotes complement set of Q_j, which is $Q_j^q = P - Q$.

Definition 4. Let DST = (T, P, Q), the weight of every attribute for Q is defined as follows:

$$\text{Weight of attribute}(Q) = \frac{1 - \text{CPME(RY)}}{\sum_{j=1}^{n}(Q_j)}.$$

Definition 5. The average density of each attribute will be determined as follows:

$$\text{Average density of each attribute}(P_j) = \frac{|[P_j]_Q|}{|T|}.$$

From that, the weighted density of each object will be determined as follows:

$$\text{Weighted density of object } (P) = \sum_{p_i \in P} (\text{Avg density}(P_j) \cdot T(Q)).$$

Definition 6. Let us consider the dataset DST = (T, P, Q), and $\varnothing$ is a fixed threshold value from the weighted density objects. If the value of Weighted density $(P) < \varnothing$, then p is termed to be an outlier.

The algorithm for the proposed model has been shown below:

Input: Dataset DST (T, P, Q) and $\varnothing$ be threshold value.

Output: Set Y has outlier data.

Step 1: Start.

Step 2: Input the dataset of categorical type.

Step 3: Apply multiequivalence relation over the dataset to determine upper and lower approximation.

Step 4: Let $Y = \varnothing$.

Step 5: For every attribute $q_i \in Q$.

Step 6: The indiscernibility relation $U/IND\ (P_i)$ according to definition 2 will be calculated.

Step 7: The complement entropy function according to definition 3 will be calculated.

Step 8: For every attribute $q_i \in Q$, the weighted density method will be applied by Definition 4.

Step 9: For each object $p_i \in T$, the weighted density method will be applied by Definition 5.

Step 10: If (weighted density $(p_i) < \varnothing$).

Step 11: $Y = Y \cup \{p_{ii}\}$.

Step 12: Return Y.

Step 13: Stop.

5 An empirical study on hiring dataset

Let us consider, the hiring data set taken by Komorowski, which has been used for classification purposes. The proposed algorithm is explained briefly by taking eight samples from the dataset, which is presented in Table 1.

The multigranulation rough set method uses the multiequivalence relation [36] to derive a lower and upper approximation for the attribute degree and reference, which are denoted as M and N, respectively. From Table 1, consider $A = \{E1, E2, E6, E8\}$, and $\widehat{M \cup N} = \{E1\}, \{E2\}, \{E3, E4, E5\}, \{E6\}, \{E7\}, \{E8\}$. The lower approximation is $\underline{A}_{\hat{M}+\hat{N}} = \{E1, E2, E8\}$, and upper approximation is $\bar{A}^{\hat{M}+\hat{N}} = \{E1, E2, E6, E7, E8\}$. While applying the proposed method, object $E8$ is detected as an outlier, and when we extend our approach to the upper approximation level, object $E7$ is detected as an outlier. The obtained values are clearly explained in Sections 6.1 and 6.2.

5.1 Concept of approximation under multigranulation rough set

Based on decision = "yes," let $A = \{E1, E2, E6, E8\}$ and consider the attributes Degree and Reference, which are represented as $\hat{M}$ and $\hat{N}$, respectively. Then, the three segments are procured from Table 2 as follows:

$$\hat{M} = \{E1,\ E7\}\{E2,\ E3,\ E4,\ E5,\ E6\}\{E8\},$$

$$\hat{N} = \{E1,\ E2\}\{E3,\ E4,\ E5\}\{E6,\ E7,\ E8\}.$$

$\widehat{M \cup N} = \{E1\}, \{E2\}, \{E3, E4, E5\}, \{E6\}, \{E7\}, \{E8\}$.Then, by applying equation (3), the lower approximation of the dataset with the multiequivalence relation is derived.

$$\hat{M} = \{E1, E7\}\{E2, E3, E4, E5, E6\}\{E8\},$$

$$A = \{E1,\ E2,\ E6,\ E8\} = \{E8\},$$

$$\hat{N} = \{E1,\ E2\}\{E3,\ E4,\ E5\}\{E6,\ E7,\ E8\},$$

$$A = \{E1,\ E2,\ E6,\ E8\} = \{E1,\ E2\},$$

$\underline{A}_{\hat{M}+\hat{N}} = \{E8\} \cup \{E1, E2\} = \{E1, E2, E8\}$.The upper approximation with multiequivalence relation based on equation (4) is as follows:

$$\hat{M} = \{E1, E7\}\{E2, E3, E4, E5, E6\}\{E8\},$$

$$A = \{E1,\ E2,\ E6,\ E8\},$$

$$= \{E1,\ E2,\ E3,\ E4,\ E5,\ E6,\ E7,\ E8\},$$

$$\hat{N} = \{E1,\ E2\}\{E3,\ E4,\ E5\}\{E6,\ E7,\ E8\},$$

$$A = \{E1,\ E2,\ E6,\ E8\},$$

$$= \{E1,\ E2,\ E6,\ E7,\ E8\},$$

$$\bar{A}^{\hat{M}+\hat{N}} = \{E1,\ E2,\ E3,\ E4,\ E5,\ E6,\ E7,\ E8\} \cap$$

$$= \{E1,\ E2,\ E6,\ E7,\ E8\},$$

$$= \{E1,\ E2,\ E6,\ E7,\ E8\}.$$

5.2 Outlier detection in multigranulation rough set

Through multiequivalence relations, lower and upper approximations are derived. Then, the rough set-based entropy measure with weighted density outlier detection method has been applied on the lower approximation set values to detect outliers that are presented in Table 3.

The indiscernibility relation for each attribute is calculated. Objects with similar values based on attributes are defined as follows:

$$\frac{U}{\text{Degree}} = \{E1,\ E2\}\{E8\}.$$

$$\frac{U}{\text{Experience}} = \{E1,\ E2\}\{E8\}.$$

$$\frac{U}{\text{Reference}} = \{E1,\ E2\}\{E8\}.$$

The complement entropy function is calculated for each attribute with the obtained indiscernible relation.

$$\text{CE (Degree)} = \frac{2}{3}\left(1 - \frac{2}{3}\right) + \frac{1}{3}\left(1 - \frac{1}{3}\right) = \frac{4}{9}.$$

Table 3: Lower approximation

Objects	Degree	Experience	Reference
$E1$	MTech	High	Big
$E2$	MTech	High	Big
$E8$	ME	High	Medium

$$\text{CE (Experience)} = \frac{3}{9},\ \text{CE (Experience)} = \frac{4}{9}.$$

The weight of each attribute should be calculated by adding the total number of attributes with the complement entropy function.

$$\text{Weight of attribute (Degree)} = \frac{5}{12}.$$

$$\text{Weight of attribute (Experience)} = \frac{6}{12}.$$

$$\text{Weight of attribute (Reference)} = \frac{5}{12}.$$

The weight of each object should be calculated by the summation of the product of the weight of attributes with indiscernible objects.

$$W(E1) = \frac{2}{3} \times \frac{5}{12} + 1 \times \frac{6}{12} + \frac{2}{3} \times \frac{5}{12} = 1.05,$$

$$W(E2) = 1.05,\ W(E8) = 0.91,$$

$$\text{If } \theta < 1.05, \text{ then object } E8 \text{ is an outlier.}$$

The same method has to be followed upon the upper approximation set to detect outliers, which are shown in Table 4. The rough set-based entropy measure with weighted density value will be calculated for each object and attribute to detect outlier. Then, object $E7$ is detected as an outlier.

Table 4: Upper approximation

Objects	Degree	Experience	Reference
$E1$	MTech	High	Big
$E2$	MTech	High	Big
$E6$	MSc	High	Medium
$E7$	MTech	High	Medium
$E8$	M.E	High	Medium

6 Experimental analysis

The benchmark datasets, from the UCI repository, have been taken to illustrate the working procedure of the proposed method. The breast cancer dataset has 286 objects and 9 attributes with one class attribute. Among the 286 objects, 9 missing values exist, and so to make the dataset balanced, some of the majority classes are removed randomly and made equal to the minority classes by the undersampling method. Also, data sets such as chess and car have been taken for the analysis. The chess dataset has 3,196 objects and 36 attributes with no missing values. The car dataset has 1,728 objects and 6 attributes with no missing values and is compared with other machine learning outlier detection algorithms. The implementation was carried out with Intel Pentium Processor, 1GigaByte RAM, and Windows10 operating system. The rough set-based entropy measure with weighted density outlier detection

Table 5: Comparison between proposed and existing method

Dataset	Feature type	No. of features considered	Feature extraction and selection method	Classification algorithm	Accuracy (RSBEMOD) (%)	Accuracy (LOF) (%)
Breast cancer	Categorical	1. Nodecap 2. Breast 3. Quad 4. Irradiat	Reduct (rough sets)	Support vector machine (SVM)	95.71	93.72
Chess	Categorical	1. bkspr 2. Bkxbq 3. Bkxcr 4. bxqsq	Reduct (rough sets)	SVM	93.24	91.23
Car	Categorical	1. Buying 2. Maint 3. lug_boot 4. Safety	Reduct (rough sets)	SVM	93.55	92.67

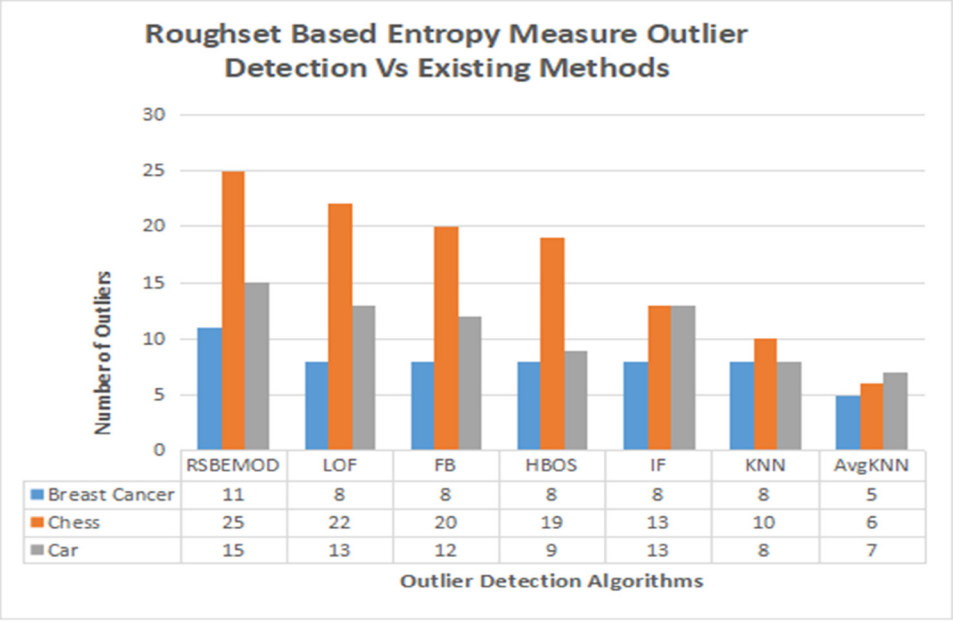

Figure 3: Comparison chart for proposed and existing outlier detection methods.

method is applied on the dataset and compared with existing outlier detection methods to prove its efficiency.

The performance of rough set-based entropy measure with weighted density values has been compared with several existing outlier detection algorithms like *k* nearest neighbor (KNN), average KNN (Avg KNN), histogram-based outlier sequence (HBOS), feature bagging (FB), isolation forest (IF), and local outlier factor (LOF). The local outlier factor algorithm detects outliers by calculating the distances of the neighbors with their density. It forms a group based on proximal values, and deviated values are considered as outliers. In feature bagging, the base estimator is fixed and divides the dataset into subsamplings. The accurate prediction can be calculated by taking an average of all base estimators. In most cases, the local outlier factor is used as the base estimator. In an isolation forest, the dataset is divided into multiple subtrees. The isolated objects from others are considered outliers. The algorithm particularly suits well for multidimensional data. By constructing histograms, outliers are easily identified by applying an unsupervised histogram-based outlier sequence algorithm. The regression and classification problems are handled by the *k* nearest neighbor algorithm. Based on the distance measure, calculate the vote of each neighbor. Average *knn* creates a super sample for all classes and a particular class average is calculated by its training samples. Rough set-based entropy measure outlier detection algorithm (RSBEMOD) determines all objects and attributes weighted density value by considering its indiscernible relation, complement entropy, and an average weight of attributes and objects. Table 5 shows the performance comparison between the proposed method and the local outlier factor (existing method).

Also, Figure 3 shows the comparison chart for a rough set-based entropy measure weighted density over existing methods.

6.1 Metrics used to evaluate the performance

To measure the performance of the algorithm, precision (P), recall (R), accuracy (A), and $F1$ measure are

Table 6: Presence of outliers

Measures	Precision	Recall	Accuracy (%)	*F*1 measure
Breast cancer	1.0	0.9167	91.67	0.9565
Chess	1.0	0.9155	91.55	0.9559
Car	1.0	0.9294	92.94	0.9634

Table 7: Removal of outliers

Measures	Precision	Recall	Accuracy (%)	*F*1 measure
Breast cancer	1.0	0.9571	95.71	0.9781
Chess	1.0	0.9324	93.24	0.9650
Car	1.0	0.9355	93.55	0.9667

calculated. The formula used to calculate these measures are as follows:

$$P = \frac{\text{True positive}}{\text{True positive} + \text{False positive}},$$

$$R = \frac{\text{True positive}}{\text{True positive} + \text{False negative}},$$

$$A = \frac{\text{True positive} + \text{True negative}}{\text{Frue positive} + \text{False positive} + \text{False negative} + \text{True negative}}.$$

$F1$ measure clearly labels valid objects without any false alarms. It lowers the threat caused by false positive and false negative values.

$$F1 \text{ measure} = \frac{2 \times P \times R}{P + R}.$$

Precision or positive predictive value represents the percentage of relevant objects from the total objects, whereas recall does the same function of sensitivity. $F1$ score clearly labels valid objects without any false alarms. It lowers the threat caused by false positive and false negative values. Table 6 shows the performance of the algorithm over the datasets in the presence of outliers, and Table 7 shows the performance after the removal of outliers.

7 Conclusion

In this article, outlier detection for categorical data using multiple granules has been developed. The classical rough set concept uses single binary relation, and the multigranulation rough set uses multiequivalence relation to derive approximations over the universe. Then, rough set-based entropy measure with a weighted density outlier detection method has been applied to detect outliers. So far, the single granular method uses the "AND" rule, whereas the multiple granulations use the "OR" rule. The proposed method applies multiequivalence relation to derive approximations, and then, a rough set-based entropy measure with weighted density value for objects and attributes is calculated. From that, a threshold value will be fixed. The values that are smaller than the threshold value are identified as anomalies. So, a proper object will not be detected as an outlier anymore. Datasets, which are taken from UCI repositories such as breast cancer, chess, and car evaluation datasets, are compared with rough set-based entropy measure weighted density outlier detection method and the existing outlier detection algorithms. The proposed method is very accurate in detecting outliers when compared with other existing methods. In the future, outlier detection for a mixed dataset using the multigranulation rough set and also for dynamic inputs can be developed.

Funding information: The authors state no funding involved.

Author contributions: Tamilarasu Sangeetha: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, roles/writing – original draft, and writing. Amalanathan Geetha Mary: writing – review and editing.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

[1] R. J. Beckman and R. D. Cook, "Outlier ... ... s," *Technometrics*, vol. 25, no. 2, pp. 119–149, 1983, DOI: 10.1080/00401706.1983.10487840.
[2] D. M. Hawkins, "Monographs on applied probability and statistics," *Identification of Outliers*, Chapman and Hall, London, 1980.
[3] V. Barnett, T. Lewis, *Outliers in Statistical Data*, Wiley and Sons, New York, 1994.
[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Comput. Surveys, ACM Comput Surv.*, vol. 41, no. 3, pp. 58–66, 2011, DOI: 10.1145/1541880.1541882.
[5] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982, DOI: 10.1007/BF01001956.
[6] F. Jiang, Y. Sui, and C. Cao, "Outlier detection based on rough membership function," *Rough. Sets Curr. Trends Comput.*, vol. 4259, pp. 388–397, Nov. 2006, DOI: 10.1007/11908029_41.
[7] J. Liang, F. Wang, C. Dang, and Y. Qian, "An efficient rough feature selection algorithm with a multi-granulation view," *Int. J. Approximate Reasoning*, vol. 53, no. 6, pp. 912–926, Sep. 2012, DOI: 10.1016/j.ijar.2012.02.004.
[8] T. Feng and J. Mi, "Variable precision multi granulation decision-theoretic fuzzy rough sets," *Knowl. Syst.*, vol. 91, pp. 93–101, Jan. 2016, DOI: 10.1016/j.knosys.2015.10.007.
[9] W. Xu, Q. Wang, and X. Zhang, "Multi-granulation rough sets based on tolerance relations," *Soft Comput.*, vol. 17, no. 7, pp. 1241–1252, Jul. 2013, DOI: 10.1007/s00500-012-0979-1.
[10] W. Xu, W. Li, and X. Zhang, "Generalized multigranulation rough sets and optimal granularity selection," *Granul. Comput.*, vol. 2, no. 4, pp. 271–288, Dec. 2017, DOI: 10.1007/s41066-017-0042-9.
[11] M. M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, LOF: identifying density-based local outliers, *Proc. of the 2000 ACM SIGMOD Int. Conf. on Mgmt. of data*, 2000, pp. 93–104. DOI: 10.1145/342009.335388.

[12] W. Xu, Q. Wang, and S. Luo, "Multi-granulation fuzzy rough sets," *J. Intell. Fuzzy Syst.*, vol. 26, no. 3, pp. 1323–1340, Jan. 2014, DOI: 10.3233/IFS-130818.
[13] R. Vashist and M. L. Garg, "Rule generation based on reduct and core: a rough set approach," *Int. J. Comput. Appl.*, vol. 29, no. 9, pp. 0975–8887, Sep. 2011.
[14] J. Li, C. Mei, W. Xu, and Y. Qian, "Concept learning via granular computing: a cognitive viewpoint," *Inf. Sci.*, vol. 298, pp. 447–467, Mar. 2015, DOI: 10.1016/j.ins.2014.12.010.
[15] P. Ashok and G. M. K. Adharnawaz, "Outlier detection method on UCI repository dataset by entropy-based rough K-means," *Def. Sci. J.*, vol. 66, no. 2, pp. 113–121, Mar. 2016, DOI: 10.14429/dsj.66.9463.
[16] W. Xu and W. Li, "Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 366–379, Feb. 2016, DOI: 10.1109/TCYB.2014.2361772.
[17] Y. Qian, J. Liang, Y. Yao, and C. Dang, "MGRS: a multi-granulation rough set," *Inf. Sci.*, vol. 180, no. 6, pp. 949–970, Mar. 2010, DOI: 10.1016/j.ins.2009.11.023.
[18] S. S. Kumar and H. H. Inbaran, "Optimistic multi-granulation rough set based classification for medical diagnosis," *Proc. Comput. Sci.*, vol. 47, pp. 374–382, Jan. 2015, DOI: 10.1016/j.procs.2015.03.219.
[19] M. A. Geetha, D. P. Acharjya, and N. C. S. Iyengar, "Algebraic properties of rough set on two universal sets based on multi-granulation," *Int. J. Rough. Sets Data Anal.*, vol. 1, no. 2, pp. 49–61, Jul. 2014, DOI: 10.4018/ijrsda.2014070104.
[20] X. B. Yang, X. N. Song, H. L. Dou, and J. Y. Yang, "Multi-granulation rough set: from crisp to fuzzy case," *Ann. Fuzzy Math. Inf.*, vol. 1, no. 1, pp. 55–70, Jan. 2011.
[21] M. I. Petrovskiy, "Outlier detection algorithms in data mining systems," *Program. Comput. Softw.*, vol. 29, no. 4, pp. 228–237, Jul. 2003, DOI: 10.1023/A:1024974810270.
[22] S. S. Kumar and H. H. Inbarani, "Optimistic multi-granulation rough set-based classification for medical diagnosis," *Proc. Comput. Sci.*, vol. 47, pp. 374–382, 2015, DOI: 10.1016/j.procs.2015.03.219.
[23] W. Yu, Z. Zhang, and Q. Zhong, "Consensus reaching for MAGDM with multi-granular hesitant fuzzy linguistic term sets: a minimum adjustment-based approach," *Ann. Oper. Res.*, vol. 300, no. 2, pp. 1–24, May 2021, DOI: 10.1007/s10479-019-03432-7.
[24] F. Jiang and Y. M. Chen, "Outlier detection based on granular computing and rough set theory," *Appl. Intell.*, vol. 42, no. 2, pp. 303–322, 2015, DOI: 10.1007/s10489-014-0591-4.
[25] H. Liu, A. Gegov, and M. Cocoa, "Rule-based systems: a granular computing perspective," *Granul. Comput.*, vol. 1, no. 4. pp. 259–274, Dec. 2016, DOI: 10.1007/s41066-016-0021-6.
[26] B. Apolloni, S. Bassis, J. Rota, G. L. Galliani, M. Gioia, and L. Ferrari, "A neuro-fuzzy algorithm for learning from complex granules," *Granul. Comput.*, vol. vol.1, no. 4, pp. 225–246, Dec. 2016, DOI: 10.1007/s41066-016-0018-1.
[27] M. A. Geetha, D. P. Acharjya, and N. C. S. N. Iyengar, "Privacy preservation in fuzzy association rules using rough computing and DSR," *Cybern. Inf. Technol.*, vol. 14, no. 1, pp. 52–71, 2014.
[28] X. Zhu, W. Pedrycz, and Z. Li, "Granular models and granular outliers," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3835–3846, Dec. 2018, DOI: 10.1109/TFUZZ.2018.2849736.
[29] W. Li and W. Xu, "Multigranulation decision-theoretic rough set in ordered information system," *Fund. Inform.*, vol. 139, no. 1, pp. 67–89, Jan. 2015, DOI: 10.3233/FI-2015-1226.
[30] W. H. Xu, X. Y. Zhang, J. M. Zhong, and W. X. Zhang, "Attribute reduction in ordered information systems based on evidence theory," *Knowl. Inf. Syst.*, vol. 25, no. 1, pp. 169–184, Oct. 2010, DOI: 10.1007/s10115-009-0248-5.
[31] J. Komorowski, Z. Pawlak, L. Polkowski, and A. Skowron, "Rough sets: a tutorial," *Rough Fuzzy Hybridization: A N Trend Decision-Making*, pp. 3–98, Dec. 1999.
[32] C. Liu, D. Miao, and J. Qian, "On multi-granulation covering rough sets," *Int. J. Approx. Reasoning*, vol. 55, no. 6, pp. 1404–1418, Sep. 2014, DOI: 10.1016/j.ijar.2014.01.002.
[33] J. Li, Y. Ren, C. Mei, Y. Qian, and X. Yang, "A comparative study of multigranulation rough sets and concept lattices via rule acquisition," *Knowl. Syst.*, vol. 91, no. 1, pp. 152–164, Jan. 2016, DOI: 10.1016/j.knosys.2015.07.024.
[34] F. Jiang, H. Zhao, J. Du, Y. Xue, and Y. Peng, "Outlier detection based on approximation accuracy entropy," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 9, pp. 2483–2499, Sep. 2019, DOI: 10.1007/s13042-018-0884-8.
[35] Z. Zhang, W. Yu, L. Martínez, and Y. Gao, "Managing multigranular unbalanced hesitant fuzzy linguistic information in multiattribute large-scale group decision making: a linguistic distribution-based approach," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2875–2889, Nov. 2020, DOI: 10.1109/TFUZZ.2019.2949758.
[36] G. Lin, J. Liang, and Y. Qian, "An information fusion approach by combining multigranulation rough sets and evidence theory," *Inf. Sci.*, vol. 314, pp. 184–199, 2015, DOI: 10.1016/j.ins.2015.03.051.