CAAI Transactions on Intelligence Technology

Special Issue: Neutrosophic Decision Making and Applications in Knowledge Management

Outlier detection in neutrosophic sets by using rough entropy based weighted density method

ISSN 2468-2322
Received on 20th December 2019
Revised on 27th March 2020
Accepted on 10th April 2020
doi: 10.1049/trit.2019.0093
www.ietdl.org

Tamilarasu Sangeetha, Geetha Mary Amalanathan ✉

SCOPE, Vellore Institute of Technology, Vellore, India
✉ *E-mail: geethamary.a@gmail.com*

Abstract: Neutrosophy is the study of neutralities, which is an extension of discussing the truth of opinions. Neutrosophic logic can be applied to any field, to provide the solution for indeterminacy problem. Many of the real-world data have a problem of inconsistency, indeterminacy and incompleteness. Fuzzy sets provide a solution for uncertainties, and intuitionistic fuzzy sets handle incomplete information, but both concepts failed to handle indeterminate information. To handle this complicated situation, researchers require a powerful mathematical tool, naming, neutrosophic sets, which is a generalised concept of fuzzy and intuitionistic fuzzy sets. Neutrosophic sets provide a solution for both incomplete and indeterminate information. It has mainly three degrees of membership such as truth, indeterminacy and falsity. Boolean values are obtained from the three degrees of membership by cut relation method. Data items which contrast from other objects by their qualities are outliers. The weighted density outlier detection method based on rough entropy calculates weights of each object and attribute. From the obtained weighted values, the threshold value is fixed to determine outliers. Experimental analysis of the proposed method has been carried out with neutrosophic movie dataset to detect outliers and also compared with existing methods to prove its performance.

1 Introduction

The concepts of data mining techniques have been used to generate patterns and to retrieve useful information from the large databases. Some of the well-known tasks associated with data mining techniques are classification, clustering and outlier analysis [1]. Classification is a supervised learning technique, where the class label of objects should be predefined to predict its target class [2]. Clustering analysis is an unsupervised learning technique, where class labels of data are unknown, before processing the data. In cluster analysis, the grouping of objects has been made based on similar characteristics of the objects [3]. Even though objects, grouped under one category, some objects may deviate from other objects, when researchers focus their analysis on a particular or specific attribute are outliers. Causes of outliers are human errors, instrumental errors while taking measurements or doing experiments, and new patterns generated in the dataset.

Zadeh developed a concept of fuzzy set in 1965 [4]. From then, the logic of fuzzy and its sets have been used to handle uncertainty. On the universe, U, where X be a single fuzzy set and $\mu X(x)$ is to determine the degree of membership such as $\mu X(x)$ belongs to [0, 1]. Since $\mu X(x)$ by itself is uncertain, it is challenging to determine whether it belongs to a set or not. The calculation of uncertainties in degree of membership has been derived in interval value for the fuzzy set. In some of the application areas, such as an expert system, fusion of information and belief system not only the truth values are supported, but in some cases, they support falsity information also. The standard fuzzy set and interval-valued fuzzy set could not be able to handle this situation. Concept of the intuitionistic fuzzy set was first developed by Atanossov, which supports both fuzzy sets and interval-valued fuzzy sets by considering membership values of truth and falsity values [5]. It could not handle indeterminate or inconsistent information like incomplete information, generally available in systems of belief. By default, the intuitionistic fuzzy set uses the formula to calculate indeterminacy is $1-t-f$.

Suppose if we raise a question with the experts, the possibility of truth membership is 0.5 and falsity membership is 0.6. The indeterminate degree of membership is 0.2. But in the case of a neutrosophic set, the membership values such as truth, indeterminacy and falsity are independent. These constraints are essential in case of trying to do the fusion of information with different sensors. Smarandache introduced the concept of neutrosophy, which signifies its outset, scope and essence of neutralities [6]. It also generalises the concept of classical fuzzy, interval fuzzy and intuitionistic fuzzy sets.

As of now, fuzzy sets handle uncertainties whereas intuitionistic fuzzy sets handle incomplete information, but both the concepts failed to handle indeterminate information. Researchers require a powerful tool, namely neutrosophic sets which can handle information like incomplete, indeterminate and inconsistent information. Some of the objects or attributes which deviates from their characteristics are known as outliers. The presence of outliers slows down the performance of process execution which results in poor results.

A rough entropy-based weighted density outlier detection has been used to identify and remove outliers. There is no neutrosophic dataset available in repositories. So, a questionnaire was prepared, and a survey was conducted among college students to obtain the neutrosophic dataset. Hence, with the proposed method outliers are identified from neutrosophic dataset and comparison has been made with existing outlier methods in Section 5.

2 Basics of roughset theory

Pawlak (1982) developed a mathematical tool to provide solutions for data which are inconsistent. It follows the two concepts of approximation, such as lower approximation and upper approximation [7]. It also mainly involves in the decision-making process by extracting rules from data. Rough sets with data mining

are associated with reducts, information tables and indiscernible relation. It mainly contributes to the analysis of data, discovers knowledge from semantic data and self-reliant decision-making process.

Also, it does not need any prior information to process the data. It can be processed directly to provide solutions [8]. For example, take $G = (P, B)$. For every subset $F \subseteq P$ and equivalence relation $B \in Ind(G)$. The following functions define the subsets, which are lower and upper approximation concerning F and their difference ($\underline{B}F - \overline{B}F$) provides boundary region F

$$\underline{B}F = \left\{r \in P : [r]_B \subseteq F\right\}$$

$$\overline{B}F = \left\{r \in P : [r]_B \cap F \neq \emptyset\right\}$$

or

$$r \in \underline{B}F \text{ if and only if } [r]_B \subseteq F$$

$$r \in \overline{B}F \text{ if and only if } [r]_B \cap F \neq \emptyset$$

The lower approximation of m is classified based on full certainty which has been a member of set F, associated with attribute set B ($\underline{B}F$) and upper approximation of f is classified based on objects which may belong to the members of set F, along with attribute set B ($\overline{B}F$).

In this real world, there exists vagueness and uncertainty of data [9]. The rough set concepts are needed to handle uncertain data which follows the concept of approximation. Identify the outlier objects, by applying the proposed algorithm, rough entropy-based weighted density method on individual clusters. Hence, the proposed algorithm works for unsupervised data; it calculates weighted density value for both objects and conditional attributes (excluding decision attribute), so that identify outliers significantly in which existing methods fail [10].

3 Neutrosophic set

Consider points in space or objects, where the generic element Y, denoted as y. Neutrosophic set S in Y may be represented with membership function such as TS (truth), IS (indeterminate) and FS (falsity). Membership values ($TS(y)$, $IS(y)$, $FS(y)$) may be real standard or non-standard subsets of $]0^-, 1^+[$. The sum of truth $TS(y)$, indeterminacy $IS(y)$ and falsity $IS(y)$ can be of any value with no constraints [11]. The representation is as follows:

$T_S(Y) \rightarrow]0^-, 1^+[$ (i)
$I_S(Y) \rightarrow]0^-, 1^+[$ (ii)
$F_S(Y) \rightarrow]0^-, 1^+[$ (iii)

An object y represented as a neutrosophic set S such as $y = y\ (T, I, F) \in S$. The T, I, F might be the subsets of non-standard or real $]0^-, 1^+[$. T denotes truth membership; I denotes indeterminacy and F indicates falsity membership of the neutrosophic set S. From the analytical point of view, neutrosophy generalises the concept of classical fuzzy, interval fuzzy and intuitionistic fuzzy sets. But in the field of engineering and science, the neutrosophic sets are determined to be in specific. Then only we can apply it in real-time applications [12].

In recent trends, the neutrosophic set has been combined with rough sets to determine roughness and its interval as neutrosophic rough sets [13]. The fuzzy set approximation has been made based on a crisp approximation which results in the concept of fuzzy rough sets. So, the rough set concept is introduced with neutrosophic set to provide knowledge from various information system [14]. The neutrosophic is combined with rough sets to handle vague data with approximations such as lower and upper. So, the graph has been constructed by applying the concept of the hybrid model. It builds self-complementary digraph for rough neutrosophic sets in decision making cases [15].

Neutrosophic sets are simplified to get single-valued neutrosophic sets (SVNS) which is an enhancement of intuitionistic fuzzy sets, where three membership functions such as truth, indeterminacy and falsity are unrelated, and their values belong to unit closed interval [16]. With the help of correlation coefficients, decision making also is done. The crisp, complex proportional assessment is also extended with a name COPRAS-SVNS [17] to take multi-criterion decision making.

4 Related work

A new definition has been given to identifying outliers based on the local outlier factor, which shows the importance of behaviour of data, which is local [18]. Cluster-based local outlier factor was defined to measure and represent the natural quality of outliers.

Outlier objects, identified as abnormal behaviour of a single object or small clusters formed which are inconsistent than others. There is a chance of abnormal occurrences in spatial or temporal locality forms a cluster known as anomalies or outliers. They used LDBSCAN algorithm for clustering and LOF to find the inconsistency of a single object [19]. Detecting outliers is the primary step in the applications of data mining. They have proposed many outlier detection algorithms for parametric and non-parametric, univariate and multivariate [20]. Outlier detection techniques are also based on spatial, distance-based and density-based clustering methods. If outliers exist in the dataset, individual observation has taken to maintain the robustness by providing suitable estimators.

An object which is dissimilar from the rest of the objects is an anomaly. First, generate frequent patterns of a dataset. Items which are having lower frequent pattern are outliers. They have designed frequent pattern outlier factor (FPOF) to detect transactions which are outliers and to identify outliers alone use FindFPOF method [21]. Researchers show their interest when trying to find rare events than frequent patterns. Existing works shows that being an outlier object is a binary property. Each object is assigned with a degree of score to be an outlier. By using LOF, calculate the neighbourhood of an object with its surroundings, how much it is isolated from others. It is crucial to detect outliers in many application areas. The topic of determining outlier score was an extension of objects in terms of clusters [22]. The individual cluster has its outlier factor, which is the clustering-based outlier method. It has two stages: the first stage form clusters based on the clustering algorithm, and the second stage detect outliers based on outlier factor.

Outliers were presented based on k-nearest neighbour graph with outlier indegree factor. Also, they have extended the work of k-nearest neighbour clustering work [23]. Compare the proposed method with the benchmark datasets. 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 local density outlier factor to measure the distance around its neighbour. 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 [24]. 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 the topological structure exists in the input space [25].

ART is an incremental algorithm, used to generate neurons at run time if the existing neurons are not enough to create a new pattern [26]. In density and distance-based clustering, the number of objects surrounded nearer to the cluster with some scope is large enough, or centre of the cluster may be far away from objects. The objects local density value and minimum distance among them or objects having higher local density value are needed to construct a

decision graph. The clusters centres are identified based on decision graph so that remaining objects placed to the nearest cluster where objects with higher local density may place at last. Construction of waste incineration plant requires weighted aggregated sum product assessment with SVNS and also used in the floatation circuit design of lead and zinc [27].

Attribute weight evaluation usually carried out in multi-decision criteria. The usual attribute weight calculation methods such as statistics based on fuzzy, grading based on experts, and comparison on binary method also used. But these methods require more experience of decision makers.

Later, the rough set concepts have been used to compute the attribute weights [28]. In the perspective of algebraic theory, the rough degree has used to compute attribute weights, but it produces wrong results when weights become zero. Information entropy has used in case of information theory which avoids the situation of weights becoming zero. But in this method, the redundant attributes are more significant than the non-redundant attributes. The disadvantage has overridden by rough entropy method, by determining attribute weights.

Fuzzy sets mainly use the concepts of similarity and entropy method. The similarity between two entities has been measured by similarity method [29]. Extension of Hamming distance, Euclidean distance to intuitionistic fuzzy set has made [30]. The Hausdorff distance also extended to intuitionistic fuzzy set and similarity measures also defined. In fuzzy sets, Zadeh first time introduced the concept of entropy measures to identify the degree of fuzziness. The concept of entropy and similarity measures in decision making are expanded widely [31] and soft entropy based on a distance measure also introduced.

The concept of similarity and entropy measure is used to solve the problem of multi-criterion decision making under the environment of SVNS. Based on Shannon's inequality, cross-entropy has used to solve discriminant information between entities. It is challenging to define the degree of truth, falsity and indeterminacy in real-world situations. The SVNS is generalised to neutrosophic interval sets which can be of intervals but not real numbers [11].

The sustainable market of Croydon University Hospital uses the definition of multi-attribute market value assessment with SVNS [32]. The heronian mean operator was used with neutrosophic sets to take multiple attributes for group decision-making system. To select the location of a garage in a residential house, NS is demonstrated. Under the environment of neutrosophic conditions, similarity measures had made between interval neutrosophic sets [33]. A theoretical study has made to study the distance, similarity and entropy of SVNS. Ye applied three-vector similarity measures in a multi-criterion decision-making system with necessary neutrosophic information. For two universal sets, a single-valued neutrosophic multi-granulation was determined. The problem of multi-period medical diagnosis, weighted aggregation of multi-period information and similarity distance based on tangent were determined. Study of different types of kernels and closure of SVNR was developed by Yang [31].

For a simple neutrosophic set, harmonic averaging projection with its multi-attribute decision making also is defined [34]. MULTIMOORA has extended with neutrosophic sets for optimisation of multi-objective context. The most crucial social happening is about love dynamics. More predictions have been drawn based on the behaviour of Romeo and Juliet concerning love impact factor. The analysis of this concept is drawn based on neutrosophic logic [35].

The implicator (I) and t-norm based on (T) defines a standard neutrosophic system with rough sets. Also, it defines the approximation space for neutrosophic sets [36]. Decision making based on soft sets is most popular nowadays. Fuzzy set decision making on soft sets had done by level soft sets, whereas decision making gradually extended to fuzzy set with interval-valued soft set [37].

The earlier statistical method decides by considering the product quality or without considering its sampling plan. If the sampling plan is an acceptable one, it has considered being determinate. But most of the cases, data may be indeterminate or imprecise, where neutrosophic interval method has used for sampling method [38]. The triplet extended neutrosophic loop of Abel Grassman's properties have been analysed in detail. It proves that it satisfies commutative and disjoint when its subgroups are maximal [39].

The noise occurred during signal transmission can be handled by neutrosophic logic. It has functions like confidence (truth), falsehood (falsity) and dependency (determinacy). With the help of several neutrosophic systems, the degree of falsehood function is reduced [40]. The algebraic properties of neutrosophic sets duplets, triplets and multisets are presented [41]. The existing statistical methods are not able to apply for reliable censored failure test. Weibull distribution has used to determine its failure and optimisation problem of producer and consumer risk also determined by combining fuzzy with neutrosophic properties [42].

The neutrosophic logic has applied for handling sinos river basin management by considering factors such as ecology, technology and society. The maps of cognitive have been designed with neutrosophy logic for PESTEL analysis [43]. The final set of neutrosophic α^n such as continuous, strongly continuous and uncertainty was defined [44]. Membership function plays an important role to get the output of any system. The trapezoidal function was used with a membership function to handle uncertain data in fuzzy and neutrosophic logic [45].

Uncertainty occurs in image processing when data are not available properly. Missing data of an image are handled by fusing the concept of dice coefficient with neutrosophic sets [46]. Neutrosophic concepts are carried out when it outreaches the value of [0, 1] through neutrosophic inequalities, equalities, infimum, supremum and standard intervals [47].

5 Proposed method

In the preprocessing stage, raw data has been taken as an input which is in the form of (T, I, F). By applying cut relation (α, β, γ) values, the binary relation of the dataset has achieved by truth, indeterminacy and falsity membership values. At the post-processing stage, the Boolean dataset has converted to categorical data by ordering. Then proposed algorithm weighted density outlier detection based on rough entropy method has implemented, and a threshold value has fixed to detect outliers as shown in Fig. 1.

5.1 Weighted density outlier detection method based on rough entropy

The neutrosophic dataset has been represented with a degree of truth membership α, indeterminacy relationship β and falsity relationship γ values which have been transformed to binary relation by applying cut relation on it. The defined value is not less than α and not more than that β and γ. If the condition is true, it should mark as 1 otherwise 0, represented in the form (α, β, γ). Obtain categorical data by ordering, and then indiscernibility, entropy, the average weight of attributes and objects have defined to identify outliers based on the following definitions shown below.

Definition 1: A dataset is defined as DTS = (P, Q, R) where P denotes universe, Q denotes objects and R denotes attributes. For example,

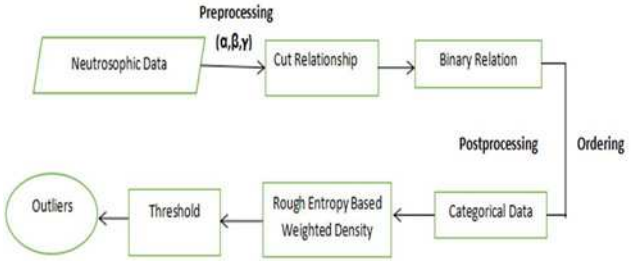

Fig. 1 *Proposed method for outlier detection in neutrosophic sets*

Table 1 Movie dataset

P	Q_1	Q_2	Q_3	Q_4
P_1	(0.05, 0.85, 0.05)	(0.10, 0.25, 0.65)	(0.50, 0.40, 0.10)	(0.10, 0.05, 0.85)
P_2	(0.05, 0.90, 0.05)	(0.10, 0.25, 0.65)	(0.50, 0.40, 0.10)	(0.10, 0.05, 0.85)
P_3	(0.05, 0.85, 0.10)	(0.05, 0.15, 0.80)	(0.10, 0.85, 0.05)	(0.40, 0.05, 0.55)
P_4	(0, 0.10, 0)	(0.70, 0, 0)	(0.80, 0, 0)	(0.30, 0, 0.70)
P_5	(0.05, 0.90, 0.05)	(0.05, 0.90, 0.05)	(0.10, 0.80, 0.10)	(0.15, 0.15, 0.70)
P_6	(0.10, 0.20, 0.70)	(0.90, 0.05, 0.05)	(0.10, 0.75, 0.75)	(0.30, 0.20, 0.50)
P_7	(0.10, 0.80, 0.10)	(0.20, 0.80, 0.10)	(0.10, 0.80, 0.10)	(0.10, 0.10, 0.90)
P_8	(0.50, 0.25, 0.25)	(0.70, 0.10, 0.20)	(0.70, 0.20, 0.10)	(0.50, 0.25, 0.25)

movie dataset which is shown in Table 1 has eight objects with four attributes. Apply cut relationship, and obtain Boolean values (Table 2). Then convert Boolean values to categorical data which is shown in Table 3.

Definition 2: Let DTS = (P, Q, R) and RT $\subseteq R$. RT represent indiscernible relation with respect to q_i in Q or r_i in R is represented as

$$\{P|ind(RT)\} = \{[q_i]_{RT}]|q_i \in P\}$$

The indiscernible relation for Table 3 will be obtained by identifying similar data in each attribute.

Definition 3: Let DTS = (P, Q, R) and RT $\subseteq R$ and P/ind(RT) = $\{R_1, R_2 \ldots R_m\}$. *Complement entropy* based on RT is defined as

$$\text{Complement entropy (RT)} = \sum_{j=1}^{m} \frac{|q_i|}{|P|}\left(1 - \frac{|q_i|}{|P|}\right)$$

where R_j^k indicates complement set of R_j, $(R_j^k = Q - R)$.

For the obtained indiscernible relation in Table 3, the complement entropy values should be calculated.

Definition 4: Let DTS = (P, Q, R), the attribute weight based on R is defined as

$$\text{Wght of Attibuter}\,(R) = \frac{1 - \text{CompEntrpy}(RT)}{\sum_{j=1}^{n} R_j}$$

From the values of Table 3 complement entropy, individual attribute weights should be calculated.

Table 2 Cut relationship (α, β, γ)

P	Q_1	Q_2	Q_3	Q_4
P_1	0	1	1	1
P_2	1	1	1	1
P_3	1	1	1	1
P_4	0	1	1	1
P_5	0	0	1	1
P_6	1	0	1	1
P_7	1	1	1	1
P_8	1	1	1	1

Table 3 Conversion of Boolean to categorical data

Participants	Q_1	Q_2	Q_3	Q_4
P_1	no	yes	good	high
P_2	yes	yes	good	high
P_3	yes	yes	good	high
P_4	no	yes	good	high
P_5	yes	no	bad	low
P_6	yes	no	good	high
P_7	yes	yes	good	high
P_8	yes	yes	bad	low

Definition 5: For every attribute, average density should be calculated based upon

$$\text{Attr density}(q_i) = \frac{|[q_i]|_R}{|P|}$$

Then for every object, weighted density should be calculated by applying the formula such as

$$\text{Object weighted density}\,(Q) = \sum_{q_i \in Q} (\text{Avg density of attr}(q_i) \cdot P(R))$$

From the obtained average attribute weights, weighted density value for each object (Table 3) should be calculated.

Definition 6: Let the dataset be DTS = (P, Q, R) and from the objects weighted density value, Φ be a threshold value which has to be fixed. Suppose the values of *weighted density of object(Q)* $< \Phi$, then q is identified as an outlier.

From the obtained weighted density values of each object (Table 3), fix the threshold values. Values which are lesser than the threshold value are outliers.

The algorithm for the proposed model is shown in Fig. 2.

5.2 Empirical study

A general survey has been conducted through general movie questionnaire form with 12 questions. Out of these, four questions were designed in the form of neutrosophy with truth membership, indeterminacy membership and falsity membership. For our analysis, we have shown eight objects with four attributes. Let the participants be represented as $P = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8\}$ and attributes as $Q = \{Q_1, Q_2, Q_3, Q_4\}$ where Q_1 is download, Q_2 is internet, Q_3 is remake and Q_4 denotes preference.

The representation of the degree of membership for truth, indeterminacy and falsity are α, β and γ. The cut relationship of (α, β, γ) are fixed as (0.05, 0.90, 0.85). It indicates that the truth value should not be lesser than 0.05 and β; γ value should not be higher than 0.90 and 0.85. If the condition is satisfied, denote the data has 1 or marked as 0, as shown in Table 2.

Then by ordering, the Boolean values are converted to categorical data (Table 3). By applying, weighted density outlier detection method based on rough entropy, anomalies are identified from the dataset.

Attribute download and internet is ordered to Yes or Order attributes such as download and internet to Yes, or No, attribute remake is ordered to Good or Bad, attribute preference is ordered to High or Low.

Now, the proposed algorithm weighted density outlier detection method based on rough entropy has been implemented over the dataset to identify outliers. For every attribute indiscernible values should be calculated as follows:

Indiscernibility (download) = $\{P_1, P_4\}$ $\{P_2, P_3, P_5, P_6, P_7, P_8\}$
Indiscernibility (internet) = $\{P_1, P_2, P_3, P_4, P_7, P_8\}$ $\{P_5, P_6\}$
Indiscernibility (remake) = $\{P_1, P_2, P_3, P_4, P_6, P_7\}$ $\{P_5, P_8\}$
Indiscernibility (preference) = $\{P_1, P_2, P_3, P_4, P_6, P_7\}$ $\{P_5, P_8\}$

Input: Dataset *DTS(T,I,F)* and Φ is a threshold value.

Output: Set *K* holds outlier data.

Step 1: Start

Step 2: Input the dataset with (*T,I,F*)values.

Step 3: Apply cut relationship(α,β,γ) to obtain Boolean values.

Step 4: Order Boolean values to get categorical data.

Step 5: Let *K* be void.

Step 6: For every attribute $q_i \in Q$.

Step 7: By definition 2, the indiscernible relation $P / ind\ (q_i)$ is determined.

Step 8: By definition 3, complement entropy function of a dataset to be calculated.

Step 9: By definition 4, weighted density of each attributer$_i \in R$ to be calculated.

Step 10: By definition 5 , weighted density value for each object should be calculated.

Step 11: If (*Object Weighted density* $(q_i) <$ Φ)

Step 12: $K = K \cup \{q_i\}$.

Step 13: Return *K*.

Step 14: Stop.

Fig. 2 *Algorithm for the proposed model*

The second step is, for each attribute, calculate complement entropy from the obtained indiscernible values

$$\text{Complement entropy (download)} = \frac{2}{8}\left(1-\frac{2}{8}\right)+\frac{6}{8}\left(1-\frac{6}{8}\right)=\frac{3}{8}$$

$$\text{Complement entropy(internet)} = \frac{6}{8}\left(1-\frac{6}{8}\right)+\frac{2}{8}\left(1-\frac{2}{8}\right)=\frac{3}{8}$$

$$\text{Complement entropy(remake)} = \frac{3}{8};$$

$$\text{Complement entropy}\left(\text{preference}\right) = \frac{3}{8}$$

The third step is to find the average for each attribute from the calculated complement rough entropy value

$$\text{Weight of first attribute (download)} = \frac{5}{12};$$

$$\text{Weight of second attribute (internet)} = \frac{5}{12};$$

$$\text{Weight of third attribute (remake)} = \frac{5}{12};$$

$$\text{Weight of fourth attribute (preference)} = \frac{5}{12}$$

Then for each object, the weighted density value should be calculated. From those values, fix the threshold value to identify outliers

$$\text{Weight of Obj}\ (P_1) = \frac{2}{8}\times\frac{5}{12}+\frac{6}{8}\times\frac{5}{12}+\frac{6}{8}\times\frac{5}{12}+\frac{6}{8}\times\frac{5}{12} = 1.04;$$

$$\text{Weight of Obj}(P_2) = 1.4;\ \text{Weight of Obj}(P_3) = 1.4;$$

$$\text{Weight of Obj}(P_4) = 1.4;\ \text{Weight of Obj}(P_5) = 1.6;$$

$$\text{Weight of Obj}(P_6) = 1.04;\ \text{Weight of Obj}(P_7) = 1.4;$$

$$\text{Weight of Obj}(P_8) = 1.2;$$

From this, fix the threshold value as 1.4. Object P_5 is detected as an outlier because its weighted density value is higher than 1.4.

5.3 Experimental analysis

A general survey has been done on movies among college students [48], and the fabricated neutrosophic type of data has been taken into consideration. The implementation has been carried out with Intel Pentium Processor, 1GigaByte RAM and Windows10 operating system. For our analysis, we have taken 121 objects with 4 attributes such as download, internet, remake and preference. The first attribute refers to the frequency of downloading movies, the second attribute refers to watching movies on the internet, the third attribute refers to the idea of remaking movies and the fourth attribute refers to the preference for watching movies. With this, achieve membership degree for truth, indeterminacy and falsity.

With the help of cut relationship (α, β, γ) values, the neutrosophic data have been converted to Boolean values based on the condition, α should not be lesser and β, γ should not be higher than the fixed cut relationship value. Then by ordering, we can convert the Boolean values to categorical data. For outlier detection, the proposed algorithm has been implemented using C language with rough set concepts. C is a powerful and structured language to develop mathematical and complex models. Rapid Miner 7 was used to detect outliers using distance-based, density-based, local outlier factor and class outlier factor method. The obtained results have been compared with the proposed method, as shown in Fig. 3.

5.4 Comparison of proposed method with existing methods

In the distance-based outlier detection method, the distance of objects to its *k*th neighbour has been calculated, then the objects which are significantly distant from their neighbours are identified as outliers. In other words, the outlier score is generated by the objects which are distant to its centre. If the distance is small, it is not an outlier. But for a significant distance, the object is considered to be an outlier. For the neutrosophic movie dataset,10 outlier objects are identified by this method.

The density has been compared around the objects to its neighbours at the local level. The non-outlier object density is

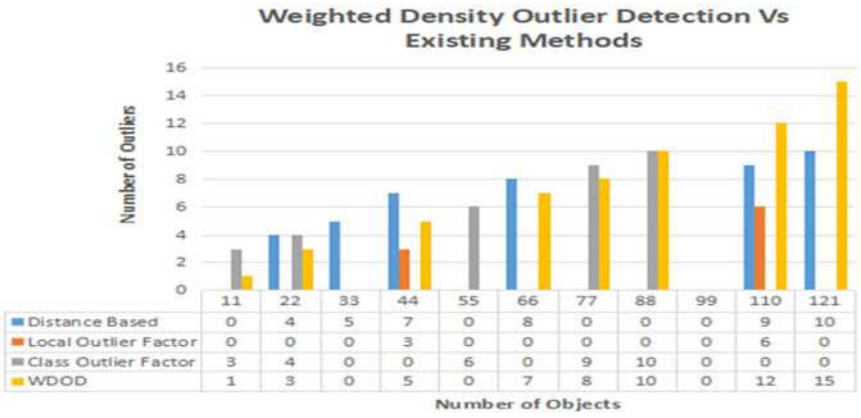

Fig. 3 *Comparison chart for proposed and existing methods*

similar to its neighbour density and outlier object density significantly deviates from its neighbours. Based on the distance D value and proportion p-value, outliers have been detected in the density-based method. For the neutrosophic movie dataset, no outliers have been detected by this method.

Based on nearest neighbours, assign a rank for each object in a dataset and determine top-class outliers. The high-level N outlier classes are generated based on the distance of kth nearest neighbour in class outlier factor method. Ten outliers are detected by this method for neutrosophic movie dataset. Objects with lower density values in the local outlier factor method have been identified as outliers. Calculate the local density value to its neighbours. The lowest density value is identified as outliers when compared to high-density value. Six outliers are detected by this method for neutrosophic movie dataset.

The proposed method for outlier detection detects outliers based on rough entropy weighted density values which have been calculated for each row and column. From the weighted values, fix the threshold value. Values which are higher than the threshold value are detected as outliers. Fifteen outliers are detected by this method for neutrosophic movie dataset. Fig. 3 shows the comparison chart for the efficiency of proposed method with existing outlier detection methods.

6 Conclusion

We have a proposed system to detect outliers in neutrosophic sets. The representation of data is truth α, indeterminacy β and falsity membership γ. By applying cut relationship, draw binary relation between objects. At this stage, the ordering of data has been done to obtain categorical data. Then weighted density outlier detection method based on rough entropy has been implemented over the dataset to detect outliers. Our proposed method calculates weighted density values for each tuple and attributes to detect outliers, but existing methods compute a distance between neighbours. Movie dataset has been taken for analysis and calculated weighted density values for both attributes and objects. Based on the threshold value fixed, outliers are detected. A performance comparison chart has been made between the existing and proposed system to show its efficiency. The proposed work has some limitations because the membership values such as truth, indeterminacy and falsity are independent. So, they do not influence, or any communication exists between the degrees of membership values. If the sum of $t+i+f>1$, the information is contradictory, when $t+i+f<1$, the information is incomplete and $t+i+f=1$, the information is complete. However, the membership values do not affect one another in decision making. This method validates its performance and also enhanced with dynamic categorical inputs by applying mathematical concepts in future.

7 References

[1] Kantardzic, M.: 'Data mining: concepts, models, methods, and algorithms' (John Wiley & Sons, Hoboken, NJ, USA, 2011)

[2] Breunig, M.M., Kriegel, H.-P., Ng, R.T., *et al.*: 'LOF: identifying density-based local outliers'. Proc. of the 2000 ACM SIGMOD int. Conf. on Management of data, Dallas, TX, USA, 2000, pp. 93–104

[3] Cadez, I.V., Smyth, P.: 'Probabilistic clustering using hierarchial models'. Information and Computer Science, University of California, Irvine, 1999

[4] Zadeh, L.A.: 'Fuzzy sets', *Inf. Control*, 1965, **8**, pp. 338–353

[5] Atanassov, K.T.: 'Intuitionistic fuzzy sets', in 'Intuitionistic fuzzy sets' (Physica, Heidelberg, 1999), pp. 1–137

[6] De Luca, A., Termini, S.: 'A definition of a non-probabilistic entropy in the setting of fuzzy sets theory', *Inf. Control*, 1972, **20**, pp. 301–312

[7] Smarandache, F.: 'About nonstandard neutrosophic logic: answers to imamura's', Note on the Definition of Neutrosophic Logic, Infinite Study, 2019

[8] Pawlak, Z.: 'Rough sets', *Int. J. Comput. Inf. Sci.*, 1982, **11**, (5), pp. 341–356

[9] Duan, L., Xu, L., Liu, Y., *et al.*: 'Cluster-based outlier detection', *Ann. Oper. Res.*, 2009, **168**, (1), pp. 151–168

[10] Dong, G., Xie, M.: 'Color clustering and learning for image segmentation based on neural networks', *IEEE Trans. Neural Netw.*, 2005, **16**, (4), pp. 925–936

[11] Smarandache, F.: 'A unifying field in logics', in 'Neutrosophy: neutrosophic probability, set and logic' (American Research Press, Rehoboth, DE, USA, 1999), pp. 1–141

[12] Goguen, J.A.: 'L fuzzy sets', *J. Math. Anal. Appl.*, 1963, **18**, pp. 145–174

[13] Pasha, E.: 'Fuzzy entropy as cost function in image processing'. Proceeding of the 2nd IMTGT Regional Conf. on Mathematics, Statistics and Applications, Universiti Sains Malaysia, Penang, Malaysia, 2006

[14] Thao, N.X., Cuong, B.C., Smarandache, F.: 'Rough standard neutrosophic sets', An application on standard neutrosophic information systems. Infinite Study, 2016

[15] Sayed, S., Ishfaq, N., Akram, M., *et al.*: 'Rough neutrosophic digraphs with application', *Axioms*, 2018, **7**, (1), p. 5

[16] Kosoko, B.: 'Fuzzy entropy and conditioning', *Inf. Sci.*, 1986, **40**, (2), pp. 165–174

[17] Yagar, R.R.: 'On the measure of fuzziness and negation, part I: membership in the unit interval', *Int. J. Gen. Syst.*, 1979, **5**, pp. 189–200

[18] Luhr, S., Lazarescu, M.: 'Incremental clustering of dynamic data streams using connectivity based representative points', *Data Knowl. Eng.*, 2009, **68**, (1), pp. 1–27

[19] Ester, M., Kriegel, H.-P., Sander, J., *et al.*: 'A density-based algorithm for discovering clusters in large spatial databases with noise', *KDD*, 1996, **96**, (34), pp. 226–231

[20] Pedrycz, W., Waletzky, J.: 'Fuzzy clustering with partial supervision', *IEEE Trans. Syst. Man Cybernet. B, Cybernet.*, 1997, **27**, (5), pp. 787–795

[21] Han, J., Pei, J., Kamber, M.: 'Data mining: concepts and techniques' (Elsevier, Waltham, MA, USA, 2011)

[22] Ganji, V.R., Prasad Mannem, S.N.: 'Credit card fraud detection using anti-k nearest neighbor algorithm', *Int. J. Comput. Sci. Eng.*, 2012, **4**, (6), pp. 1035–1039

[23] Atanassov, K., Stoeva, S.: 'Intuitionistic L fuzzy sets', *Cybernet. Syst. Res.*, 1984, **2**, pp. 539–540

[24] Cheng, C.C., Liao, K.H.: 'Parameter optimization based on entropy weight and triangular fuzzy number', *Int. J. Eng. Ind.*, 2011, **2**, (2), pp. 62–75

[25] Jiang, Y., Tang, Y., Chen, Q., *et al.*: 'Interval-valued intuitionistic fuzzy soft sets and their properties', *Comput. Math. Appl.*, 2010, **60**, (3), pp. 906–918

[26] Wang, H., Smarandache, F., Zhang, Y.Q., *et al.*: 'Interval neutrosophic sets and logic', Theory and Applications in Computing, Hexis, AZ, 2005

[27] Kaufmann, A.: 'Introduction to the theory of fuzzy subsets' (Academic Press, New York, NY, USA, 1975)

[28] Sangeetha, T., Geetha Mary, A.: 'A rough entropy-based weighted density outlier detection method for two universal sets'. Proc. of the 2nd Int. Conf. on Data

Engineering and Communication Technology, Symbiosis International University, Pune, Maharashtra, India, 2019, pp. 509–516
[29] Majumdar, P., Samanta, S.: 'On similarity and entropy of neutrosophic sets', *J. Intell. Fuzzy Syst.*, 2014, **26**, pp. 1245–1252
[30] Szmidt, E., Kacprzyk, J.: 'Entropy for intuitionistic fuzzy sets', *Fuzzy Sets Syst.*, 2001, **118**, pp. 467–477
[31] Hu, J., Yang, Y., Zhang, X., *et al*.: 'Similarity and entropy measures for hesitant fuzzy sets', *Int. Trans. Oper. Res.*, 2018, **25**, (3), pp. 857–886
[32] Tuskan, I.: 'Interval valued fuzzy sets based on normal forms', *Fuzzy Sets Syst.*, 1986, **20**, pp. 191–210
[33] Majumdar, P.: 'A study of several types of sets expressing uncertainty and some applications on them'. Ph.D. Thesis, Visva Bharati University, India, 2013
[34] Wang, H., Smarandache, F., Zhang, Y., *et al*.: 'Single valued neutrosophic sets'. Proc. of 10th Int. Conf. on Fuzzy Theory & Technology, Salt Lake City, UT, USA, 2005
[35] Patro, S.K.: 'On a model of love dynamics: a neutrosophic analysis'. Infinite Study, 2016
[36] Thao, N.X., Smarandache, F.: '(T, τ)-standard neutrosophic rough set and its topologies properties', *Neutrosophic Sets Syst.*, 2016, **14**, pp. 341–356
[37] Qin, H., Ma, X., Herawan, T., *et al*.: 'An adjustable approach to interval-valued intuitionistic fuzzy soft sets based decision making'. Asian Conf. on Intelligent Information and Database Systems, Daegu, Republic of Korea, 2011, pp. 80–89
[38] Aslam, M.: 'A new attribute sampling plan using neutrosophic statistical interval method', *Complex Intell. Syst*, 2019, **5**, (4), pp. 1–6
[39] Jha, S., Kumar, R., Chiclana, F., *et al*.: 'Neutrosophic approach for enhancing quality of signals', *Multimedia Tools Appl.*, 2019, **67**, (3), pp. 1–32
[40] Wu, X., Zhang, X.: 'The decomposition theorems of AG-neutrosophic extended triplet loops and strong AG-(l, l)-loops', *Mathematics*, 2019, **7**, (3), p. 268
[41] Smarandache, F., Zhang, X., Ali, M.: 'Algebraic structures of neutrosophic triplets, neutrosophic duplets, or neutrosophic multisets' (MDPI, Basel, Switzerland, 2019), p. 171
[42] Aslam, M.: 'A new failure-censored reliability test using neutrosophic statistical interval method', *Int. J. Fuzzy Syst.*, 2019, **21**, (4), pp. 1214–1220
[43] Ortega, R.G., Rodríguez, M., Leyva Vázquez, M., *et al*.: 'Pestel analysis based on neutrosophic cognitive maps and neutrosophic numbers for the sinos river basin management', *Neutrosophic Sets Syst.*, 2019, **26**, (1), p. 16
[44] Dhavaseelan, R., Devi, R., Jafari, S., *et al*.: 'Neutrosophic alpha-m-continuity', *Neutrosophic Sets Syst.*, 2019, **27**, (1), p. 16
[45] Broumi, S., Nagarajan, D., Bakali, A., *et al*.: 'Implementation of neutrosophic function memberships using MATLAB program', *Neutrosophic Sets Syst.*, 2019, **27**, (1), p. 5
[46] Jha, S., Kumar, R., Priyadarshini, I., *et al*.: 'Neutrosophic image segmentation with dice coefficients', *Measurement*, 2019, **134**, pp. 762–772
[47] Qin, H., Luo, D.: 'New uncertainty measure of rough fuzzy sets and entropy weight method for fuzzy-target decision-making tables', *J. Appl. Math.*, 2014, **2014**, pages 7
[48] Sangeetha, T., GeethaMary, A.: 'Moviedataset', Mendeley Data, v1, 2019, doi:10.17632/v526j384hx.1