RESEARCH ARTICLE

SPECIAL ISSUE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S067

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Outlier Detection in Single Universal Set using Intuitionistic Fuzzy Proximity Relation based on A Rough Entropy-Based Weighted Density Method

Geetha Mary A [1], *Sangeetha T* [1]

[1] *School of Computer Science and Engineering & Vellore Institute of Technology, Vellore, India*

Email: geethamary.a@gmail.com

Article History

Received: 26 February 2023
Accepted: 21 March 2023

Keywords:

Outliers;
Intuitionistic Fuzzy Proximity Relation;
Membership Relation;
Non-Membership Relation

Abstract

Data mining is a technique for analyzing larger datasets to identify patterns, information, and relationships that may be used to solve challenging problems. Identifying outliers has attracted the focus of researchers working on a variety of areas. Outliers are things that behave differently from other objects. With real-world data, rough set theory can cope with ambiguity and uncertainty. So far, the study has solely focused on spotting outliers using the membership function. Outliers may be recognized using membership and non-membership values, however, utilizing the principle of intuitionistic fuzzy proximity relation. At this step, the indiscernibility of objects is discovered, and the quantitative data is then converted to qualitative data. This article proposes outlier detection in single universal sets using an intuitionistic fuzzy proximity relation with a rough set based on complement entropy and weighted density approach. The empirical study has been considered for ranking the colleges based on the parameters evaluated.

1. Introduction

Data mining techniques may be used to uncover hidden patterns in datasets. When not specifically focused, data available in the real world may include uncertainties that lead to ambiguity. If the data is unclear, accuracy may be compromised, and failure may occur (García, Luengo, and Herrera). The majority of data mining research has concentrated on three techniques: object classification, object grouping, and spotting outliers among objects.

Clusters are formed when similar objects are gathered together. When the inner parts of the cluster are examined, however, some of the objects may vary depending on their characteristics, which are referred to as anomalies. Outliers are sometimes known as anomalies (Cios, W Pedrycz, and Swiniarsk).

In a dataset, for example, people, birds, and animals are classified into separate clusters. A man with six fingers is classified as an outlier in the human group based on a particular characteristic. Similarly, in each cluster, objects that deviate from other objects based on a certain property are referred to as outliers. Clustering's major purpose is to find the subgroups that exist within a dataset. When objects inside a cluster are compared to objects among clusters, they show significant similarity.

Outliers are items whose behavior deviates significantly from that of other objects (Hawkins). Outlier detection is essential because the presence of outliers causes the system to operate slowly. As a result, it is essential to remove outliers from the

dataset.

In 1965, Zadeh developed the notion of a fuzzy set to address and solve the issue of uncertainty (Zadeh). The membership part of fuzzy sets is simply focused on, but the information cannot be retrieved due to a lack of knowledge. In the actual world, non-membership value, which focuses on the deterministic aspect, correlates with membership value (S. K. Ghosh, Mitra, and A. K. Ghosh). Yet, the issue of membership and non-membership values must be addressed. As a result, Atanassov in the year 1986 presented membership and non-membership relationships through the intuitionistic fuzzy concept (Atanassov and Atanassov).

To study object indiscernibility, the fuzzy proximity relation is utilized. As a result, it has been enhanced to include the concept of intuitionistic fuzzy proximity relation, which outperforms fuzzy approximation on rough sets (Bello and Falcon). The sum of the membership and non-membership values ranges between 0 and 1 (Nanda and Majumdar). Also, the final relationship must be symmetric. The ordering relation may be applied to the dataset after identifying the equivalence classes.

This article provides a method for converting quantitative data to qualitative data by combining intuitionistic fuzzy proximity with an ordering relation. The rest of the article is structured as follows: Section 2 goes through the literature review. The proposed methodology is explained in Section 3. Section 4 demonstrates the notion of empirical study, and Section 5 concludes the chapter.

2. Literature Review

Many exceptional set and minimal exceptional set instances have been studied to identify outliers by computing the exceptional degree of each object in minimal sets (Jiang, Sui, and C. Cao). Outlier detection is widely used in statistics, however, well-known distribution values are limited to univariate data. Due to this constraint, it cannot be used in real-world data that contains multivariate data.

Outliers are identified using the distance-based outlier identification approach, which measures the unusualness of their neighbors. Although it is a non-parametric technique, the computation time is lengthy (Chandola, Banerjee, and Kumar). The use of intuitionistic fuzzy sets for multiattribute decision-making is investigated. Various mathematical programming models are built to provide the optimal weights for the attributes, as well as the associated decision-making methods were suggested.

According to fuzzy set theory, an element's membership in a fuzzy set is represented by a single value between zero and one. Even so, since there may be some hesitation degree, the degree of non-membership of an element in a fuzzy set may not always be equal to 1 minus the membership degree (Ejegwa et al.). As a result, intuitionistic fuzzy sets (IFS), a fuzzy set extension, which includes the amount of hesitation known as the hesitation margin (which is defined as 1 minus the total of membership and non-membership degrees, respectively).

Some researchers proposed a solution to decision-making problems by employing an intuitionistic fuzzy soft set in two universal sets. Accuracy and rough degrees were also used to acquire the optimal solutions. They also built a binary relationship with an intuitionistic fuzzy relation between the two non-empty universal sets U and V (Liu).

3. Proposed Approach

Consider a single universal set that contains quantitative data. In the preprocessing stage, the quantitative data has been converted to qualitative data using intuitionistic fuzzy proximity relation (Geetha, Acharjya, and Iyengar). The membership function (μ) can be obtained using equation (1) and the non-membership (v) value can be obtained by using equation (2).

$$\mu(O_i, O_j) = 1 - ((|O_i - O_j|)/ (Maximum\ Value\ of\ the\ Parameter)) \quad (1)$$

$$v(O_i, O_j) = ((|O_i - O_j|)/ (2 * Maximum\ Value\ of\ the\ Parameter)) \quad (2)$$

Based on the indiscernible values obtained, the quantitative data is converted to qualitative data by using ordering relation. Now, in the post-processing stage, the indiscernible function, complement entropy values, and weighted density values of

attributes and objects should be calculated (Zhao, Liang, and F. Cao). Based on the computed weighted density values, the threshold value will be fixed. The weighted density value of objects will be compared with the threshold value to identify

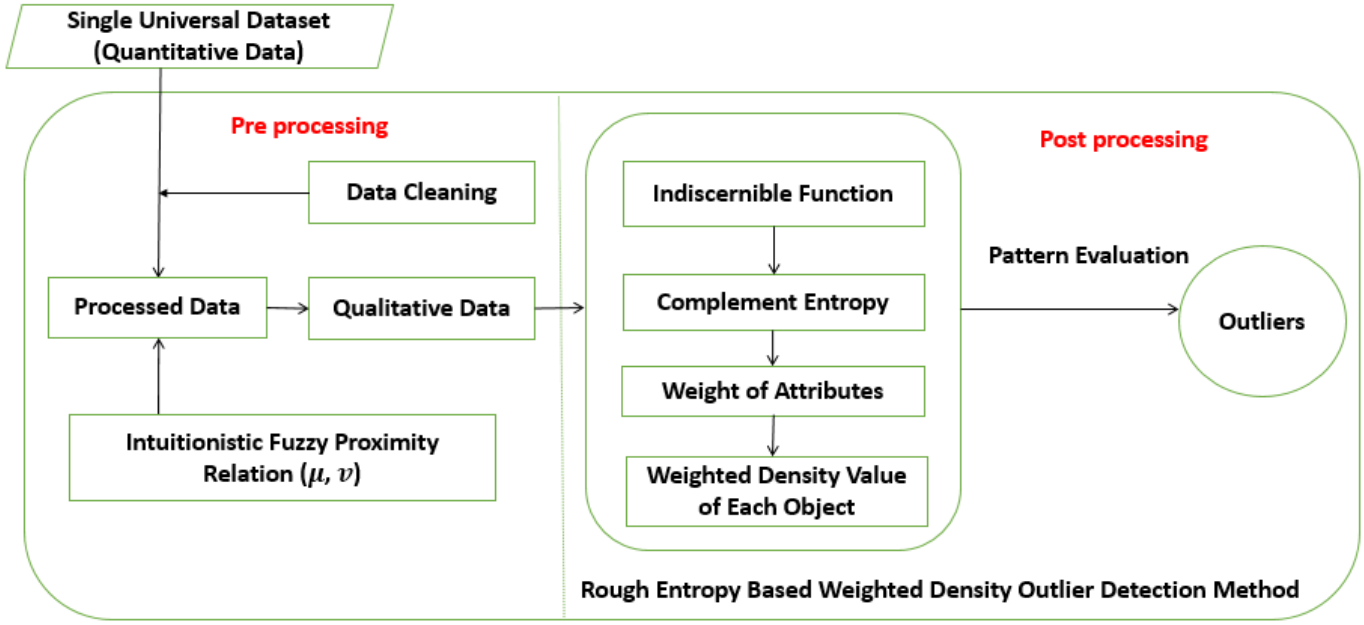

FIGURE 1. Working Model of the Proposed Methodology

the outlier object. Table 1 shows the information about the dataset considered for evaluation. Figure 1 shows the working model of the proposed methodology.

4. Empirical Study

Let us consider the ranking of colleges based on the attributes of faculty, education system, placement, infrastructure, and collaboration with the industry. The colleges will be represented as $C = (C_1, C_2, C_3, C_4, C_5\}$ and the attributes are represented as

{Faculty,Education System,Placement,Infrastructure,Industry Collaboration}

The computed IFPR values for the attribute faculty, education system, infrastructure, and industry collaboration are shown in Tables 2,3,4,5, and 6 respectively.

Let us consider the similarity value greater than 0.83 and dissimilarity values lesser than 0.125, the equivalence classes obtained are shown below:

$R^1 = \{\{C_1, C_2, C_5\}, \{C_3, C_4\}\}$
$R^2 = \{\{C_1, C_2, C_3\}, \{C_4, C_5\}\}$
$R^3 = \{\{C_1\}, \{C_2, C_3, C_4, C_5\}\}$
$R^4 = \{\{C_1, C_2\}, \{C_3, C_4, C_5\}\}$
$R^5 = \{\{C_1, C_2\}, \{C_3, C_4, C_5\}\}$

Apply ordering relation and convert the quantitative data to qualitative data. The ordering relation is as follows:

$\prec_{Faculty}$: Good$\prec$Average
$\prec_{Education\ system}$: High$\prec$Low
$\prec_{Placement}$: Good$\prec$Average
$\prec_{Infrastructure}$: High$\prec$Low
$\prec_{Industry\ collaboration}$: Good$\prec$Average

4.1. Identify indiscernible relations among objects

U/IND (Faculty) = $\{\{C_1, C_2, C_5\}, \{C_3, C_4\}\}$
U/IND (Education System) = $\{\{C_1, C_2, C_3\}, \{C_4, C_5\}\}$
U/IND (Placement) = $\{\{C_1\}, \{C_2, C_3, C_4, C_5\}\}$
U/IND (Infrastructure) = $\{\{C_1, C_2\}, \{C_3, C_4, C_5\}\}$
U/IND (Industry Collaboration) = $\{\{C_1, C_4, C_5\}, \{C_2, C_3\}\}$

4.2. Calculate complement entropy for the attributes

CE(Faculty)= $\frac{3}{5}(1-\frac{3}{5}) + \frac{2}{5}(1-\frac{2}{5}) = \frac{3}{5} \times \frac{2}{5} + \frac{2}{5} \times \frac{3}{5} = \frac{12}{25}$
CE(Education System)= $\frac{12}{25}$;
CE(Placement)= $\frac{8}{25}$;
CE(Infrastructure)=$\frac{12}{25}$;
CE(IndustryCollaboration= $\frac{12}{25}$

4.3. Calculate the attribute's weight

$$Weight\,(Faculty)\,\frac{13}{30};$$

$$Weight\,(Education\ System) = \frac{13}{30};$$

TABLE 1. Ranking of the Colleges

Colleges	Faculty	Education System	Placement	Infrastructure	Industry Collaboration
C1	60	40	40	30	60
C2	60	40	30	30	50
C3	50	40	30	40	50
C4	50	30	30	40	60
C5	60	30	30	40	60

TABLE 2. The IFPR for the attribute faculty

R1	60	60	50	50	60
60	1	1	0.833	0.833	1
	0	0	0.083	0.083	0
60	1	1	0.833	0.833	1
	0	0	0.083	0.083	0
50	0.833	0.833	1	1	0.833
	0.083	0.083	0	0	0.083
50	0.833	0.833	1	1	0.833
	0.083	0.083	0	0	0.083
60	1	1	0.833	0.833	1
	0	0	0.083	0.083	0

TABLE 3. The IFPR for the attribute education system

R2	40	40	40	30	30
40	1	1	1	0.750	0.750
	0	0	0	0.125	0.125
40	1	1	1	0.750	0.750
	0	0	0	0.125	0.125
40	1	1	1	0.750	0.750
	0	0	0	0.125	0.125
30	0.750	0.750	0.750	1	1
	0.125	0.125	0.125	0	0
30	0.750	0.750	0.750	1	1
	0.125	0.125	0.125	0	0

TABLE 4. The IFPR for the attribute placement

R3	40	30	30	30	30
40	1	0.750	0.750	0.750	0.750
	0	0.125	0.125	0.125	0.125
30	0.750	1	1	1	1
	0.125	0	0	0	0
30	0.750	1	1	1	1
	0.125	0	0	0	0
30	0.750	1	1	1	1
	0.125	0	0	0	0
30	0.750	1	1	1	1
	0.125	0	0	0	0

TABLE 5. The IFPR for the attribute infrastructure

R4	30	30	40	40	40
30	1	1	0.750	0.750	0.750
	0	0	0.125	0.125	0.125
30	1	1	0.750	0.750	0.750
	0	0	0.125	0.125	0.125
40	0.750	0.750	1	1	1
	0.125	0.125	0	0	0
40	0.750	0.750	1	1	1
	0.125	0.125	0	0	0
40	0.750	0.750	1	1	1
	0.125	0.125	0	0	0

$$Weight\,(Placement)\,\frac{13}{30};$$

$$Weight\,(Infrastructure) = \frac{17}{30};$$

$$Weight\,(Collaboration) = \frac{13}{30};$$

4.4. Calculate the Weight of Objects

$$W\ (C_1) = (\tfrac{3}{5} \times \tfrac{13}{30}) + (\tfrac{3}{5} \times \tfrac{13}{30}) + (\tfrac{1}{5} \times \tfrac{13}{30}) + (\tfrac{2}{5} \times \tfrac{17}{30}) + (\tfrac{3}{5} \times \tfrac{13}{30}) = 1.09;$$

$W\ (C_2) = 1.26$; $W\ (C_3) = 1.29$;
$W\ (C_4) = 1.26$; $W\ (C_5) = 1.38$;

If the threshold value is 1.26, then object C_1 which is lesser than 1.26 is determined to be an outlier.

TABLE 6. The IFPR for the attribute industry collaboration

R5	60	50	50	60	60
60	1	0.833	0.833	1	1
	0	0.083	0.083	0	0
50	0.833	1	1	0.833	0.833
	0.083	0	0	0.083	0.083
50	0.833	1	1	0.833	0.833
	0.083	0	0	0.083	0.083
60	1	0.833	0.833	1	1
	0	0.083	0.083	0	0
60	1	0.833	0.833	1	1
	0	0.083	0.083	0	0

TABLE 7. Qualitative Data

College	Faculty	Educa System	Placeme	Infrasti	Industry Collaboration
C1	Good	High	Average	High	Good
C2	Good	High	Good	High	Average
C3	Average	High	Good	Low	Average
C4	Average	Low	Good	Low	Good
C5	Good	Low	Good	Low	Good

5. Conclusion

This article suggests a method for finding outliers in a single universal dataset utilizing intuitionistic fuzzy proximity relations with weighted density values of objects and attributes. The quantitative data is transformed into qualitative data by calculating membership and non-membership values, followed by ordering relations. Then by finding indiscernibility, computing complement entropy and weighted density values of objects and attributes, employ the threshold value. The threshold value is compared with the computed weighted density value of objects, to determine outliers. The empirical study shows that the proposed methodology detects outliers accurately. The implementation of the proposed idea will be further investigated in future work with two universal sets.

References

Atanassov, Krassimir T and Krassimir T Atanassov. "Intuitionistic fuzzy sets". *Physica-Verlag HD* (1999): 80034–80037.

Bello, Rafael and Rafael Falcon. "Rough Sets in Machine Learning: A Review". *Thriving Rough Sets* 708 (2017): 87–118.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". *ACM computing surveys (CSUR)* 41 (2009): 1–58.

Cios, K J, W Pedrycz, and R M Swiniarsk. "Data Mining Methods for Knowledge Discovery". *IEEE Transactions on Neural Networks* 9.6 (2012): 1533–1534.

Ejegwa, P A, et al. "An overview on intuitionistic fuzzy sets". *Int. J. Sci. Technol. Res* 3 (2014): 142–145.

García, Salvador, Julián Luengo, and Francisco Herrera. "Data Preprocessing in Data Mining". 72 (2015).

Geetha, Mary A, D P Acharjya, and N Ch. S N Iyengar. "Algebraic properties and measures of uncertainty in rough set on two universal sets based on multi-granulation". *Proceedings of the 6th ACM India Computing Convention* (2014).

Ghosh, Swarup Kr, Anirban Mitra, and Anupam K Ghosh. "A novel intuitionistic fuzzy soft set entrenched mammogram segmentation under Multigranulation approximation for breast cancer detection in early stages". *Expert Systems with Applications* 169 (2021): 114329–114329.

Hawkins, D. "Identification of Outliers". *obituary symposium on instructions to authors university of Chicago* 35 (1980): 129–129.

Jiang, Feng, Yuefei Sui, and Cungen Cao. "Outlier Detection Using Rough Set Theory". *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005. 79–87.

Liu, Guilong. "Rough set theory based on two universal sets and its applications". *Knowledge-Based Systems* 23.2 (2010): 110–115.

Nanda, S and S Majumdar. "Fuzzy sets and systems 45". 2 (1992): 249–251.

Zadeh, Lotfi A. "Fuzzy sets". *Information and Control* 8.3 (1965): 338–353.

Zhao, Xingwang, Jiye Liang, and Fuyuan Cao. "A simple and effective outlier detection algorithm for categorical data". *International Journal of Machine Learning and Cybernetics* 5.3 (2014): 469–477.

Embargo period: The article has no embargo period.

To cite this Article: , Geetha Mary A, and Sangeetha T . "**Outlier Detection in Single Universal Set using Intuitionistic Fuzzy Proximity Relation based on A Rough Entropy-Based Weighted Density Method**." International Research Journal on Advanced Science Hub 05.05S May (2023): 501–506. http://dx.doi.org/10.47392/irjash.2023.S067