Detection and recognition of outliers by the use of IF-THEN rules

Marcin Kacprowicz[0000−0003−4381−1359], Monika Bartczak[0000−0002−1361−3805], and Adam Niewiadomski[0000−0001−7346−5472]

Institute of Information Technology
Lodz University of Technology
Wolczanska 215, 90-924 Lodz, Poland
{adam.niewiadomski,marcin.kacprowicz}@p.lodz.pl,
monika.bartczak@dokt.p.lodz.pl

Abstract. In data mining and exploration, outliers are strange, specific, rare data in the database that cannot be ignored because they can provide potentially dangerous information. Outlier detection can contribute to detecting an illegal usage of a credit card, breaking into transaction systems at the bank, hacking into a computer system, etc. The paper presents an original fuzzy solution to the issue of outliers detection in data sets. The following novelties are introduced: definition of an outlier in terms of fuzzy logic and methods for detecting and recognizing outliers in databases. Finding outliers is done by using fuzzy rules (IF-THEN rules) based on fuzzy sets and fuzzy logic. This method is useful when linguistic knowledge rather than crisp data that are accessible. Based on the proposed new definition of the outlier and the method, the research work was conducted. Calculations and the test were based on data with the database which were expressed in a nonprecise, lingual way (similar to natural, human language).

Keywords: Outliers in databases · fuzzy rules · detection of outliers · outlying objects.

1 An outlier in terms of fuzzy rules definition

The article is a continuation of research on artificial intelligence systems using fuzzy logic. The fuzzy implications proposed by Niewiadomski and Kacprowicz [1] are applied to the general concept of outliers search presented in [2]. The application of the designed methods to the analysis of data in graph databases [3] confirmed the effectiveness of the method and allowed to formulate an outlier in terms of fuzzy rules definition. Linguistically quantfied statements based on fuzzy logic are applied to detection recognition by the authors in [4,5].
Let $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, $N \in \mathbb{N}$, be a finite non-empty set of objects. Let $\mathcal{R} = \{R_1, R_2, \ldots, R_K\}$, $K \in \mathbb{N}$, be a set of fuzzy rules IF d_i is A_k THEN d_i is B_k, $i = 1, 2, \ldots, N$, and A_k, B_k are the antecedent and the consequent of R_k, respectively, represented by fuzzy sets in $\mathcal{D}$ (so R contains traditional IF-THEN rules). For a given $k \leq K$, *the degree of the outlier of* R_k is defined [6,7]:

$$O(R_k) = \begin{cases} \min\{\max\{T, 1-T\}, 1-C\}, & T > 0 \\ 0, & T = 0, \end{cases} \tag{1}$$

where T is *the degree of truth* (aka *conditional and unqualified proposal* [8]) evaluated as:

$$T = \frac{\sum_{i=1}^{N} \mu_{A_k \to B_k}(d_i)}{\sum_{i=1}^{N} \mu_{A_k}(d_i)}, \tag{2}$$

for $A_k \to B_k$ – a fuzzy implication, e.g. t-norm, and C – *the degree of sufficient coverage* [9], determines if the rule is activated for sufficiently large number of $d \in \mathcal{D}$ objects:

$$C = f(r_c), \tag{3}$$

where r_c is the coverage ratio:

$$r_c = \frac{1}{N} \sum_{i=1}^{N} t_i \tag{4}$$

with $t(i)$ computed as:

$$t_i = \begin{cases} 1, & \text{if } \mu_{A_K}(d_i) > 0 \text{ and } \mu_{B_K}(d_i) > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

Finally, the S-shape function $f\colon [0,1] \to [0,1]$

$$f(r) = \begin{cases} 0, & r < r_1 \\ g(r) & r_1 \le r \le r_2 \\ 1 & r > r_2 \end{cases} \tag{6}$$

where $0 \le r_1 < r_2 \le 1$ and g is a non-decreasing and continuous function on $[r_1, r_2]$.

Definition 1 (An outlier in terms of fuzzy rules). *Let $\kappa \in (0,1]$. An object $d_i \in D$, $i = 1, 2, \ldots, N$ is* an outlier *if it activates any rule R_k, $k = 1, 2, \ldots K$, such that*

$$O(R_k) \ge \kappa. \tag{7}$$

2 Example

Now, we are going to show the detection of outliers via fuzzy rules in practice. The fuzzy rules are created on the base of expert knowledge collected from bank and financial analytics, including personnel responsible for transfer security and recognition of cyberattacks on bank systems (especially phishing and unauthorized access), here, we exemplify fuzzy rules and implications as follows: let A be a fuzzy set representing linguistic label *summer* in $X = \{1, 2, \ldots, 366\}$ – numbers of days in a year in which the complaint is submitted, with μ_A given:

$$\mu_A(x) = \begin{cases} \frac{x-123}{47}, & x \in [123, 170] \\ \frac{-x+217}{47}, & x \in [170, 217] \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Let B represent label *middle county* in $Y = [5, 70]$ – per capita income in county (representing in thousands) from which come of a person who submits a complaint with $\mu_B(y)$:

$$\mu_B(y) = \begin{cases} \frac{y-35}{9}, & y \in [35, 45] \\ \frac{-y+54}{9}, & y \in [45, 54] \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Let C represent label *average time* in $Z = [0, 30]$ – numbers of days between receiving and sending the complaint to the company by CFPB (Consumer Financial Protection Bureau), with $\mu_C(z)$:

$$\mu_C(z) = \begin{cases} \frac{z-2}{4}, & z \in [2, 6] \\ \frac{-z+10}{4}, & z \in [6, 10] \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Hence the sample fuzzy rule is:

IF complaint is submitted in summer
AND submitter comes from middle county (per capita income)
THEN in average time CFPB sends complaint (11)

For example, a complaint is submitted on May 4 (the 124th day of the year) and the submitting person comes from the county where per capita income is 53 000 $. Hence, $\mu_{A_K}(124) \simeq 0.02$, $\mu_{B_K}(53) \simeq 0.11$, so via the product implication the value of the rule is $\simeq 0.0022$.
Next, for each fuzzy rule, perform calculations according to the formulas (1)-(7). A fuzzy rule for which the degree of sufficient coverage $C >= 0.1$ and the degree of outlier $O(R_k) >= 0.9$ is assumed to be an outlier. For fuzzy rule: IF complaint is submitted in summer AND submitter comes from the middle county (per capita income) THEN in average time CFPB sends complaint we received $O(R_k) = 0.91$ and $C = 0$.

3 Results and comments

The paper proposes and presents a new definition of outlier detection and recognition in databases. The definition is universal and can be applied to relational as well as non-relational datasets. A graph database was used in the study. To detect the outlier used the new definition and the IF-THEN method was used. The IF-THEN method based on different implications: product, Lukasiewicz, K_1, K_2, and K_3, and subsumed under three different defined S-shape functions. The tests were performed on 648 fuzzy rules. The application of the above implication enriched our experiment. We detected new outliers. Not all of the implications used detected, classified a given fuzzy rule as an outlier. Therefore, they were looked at in detail. It was assumed that an outlier is a fuzzy rule for which the degree of coverage $C >= 0.1$ and the degree of outliers $O(R_k) >= 0.9$. Analysis, interpretation of the results showed that the detected outliers by using other implications e.g. K2 had a high degree of uniqueness $O(R_k)$ (e.g. 0.89) for other implications e.g. PROD.

As a result, we obtained the following six unique fuzzy rules R_{85}, R_{95}, R_{121}, R_{137}, R_{149}, R_{175} (see Table 1) which are associated with 613 objects.
The definition introduced allows detection and, more importantly, recognition of specific objects (which are outliers). We detected 6 outlying rules - 613 objects. The method has been tested on real data available in the data set, where outliers may indicate incorrect data input, or even, anomalous data properties.

Table 1. Generated IF AND THEN fuzzy rules with an evaluated degree of outlier $O(R_k)$ and degree of coverage C for different implications: $PROD$, $LUKASIEWICZ$ (LUK), K_1, K_2, and K_3.

Rule No.	Fuzzy rules	$PROD$		LUK		K_1		K_2		K_3	
		$O(R_k)$	C	$O(R_k)$	C	$O(R_k)$)	C	$O(R_k$	C	$O(R_k)$	C
85.	IF complaint is submitted in the middle of spring AND submitter comes from the rich county (median household) THEN in an average time CFPB send a complaint.	0.91	0	0.92	0	0.85	0	0.98	0	0.85	0
92.	IF complaint is submitted in summer AND submitter comes from the middle county (median household) THEN in a short time CFPB send a complaint.	0.84	0	0.85	0	0.74	0	0.92	0	0.39	0
95.	IF complaint is submitted in summer AND submitter comes from the middle county (median household) THEN in an average time CFPB send a complaint.	0.93	0	0.85	0	0.74	0	0.92	0	0.39	0
121.	IF complaint is submitted in early winter AND submitter comes from the rich county (median household) THEN in an average time CFPB send a complaint.	0.95	0	0.96	0	0.88	0	0.97	0	0.9	0
137.	IF complaint is submitted in the middle of spring AND submitter comes from the rich county (per capita income) THEN in an average time CFPB send a complaint.	0.91	0	0.92	0	0.81	0	0.96	0	0.89	0
148.	IF complaint is submitted in the summer AND submitter comes from the middle county (per capita income) THEN in a short time CFPB send a complaint.	0.89	0	0.9	0	0.79	0	0.95	0	0.75	0
149.	IF the complaint is submitted in the summer AND the submitter comes from the middle county (per capita income) THEN in an average time CFPB sends a complaint.	0.91	0	0.92	0	0.9	0	0.98	0	0.87	0
175.	IF the complaint is submitted in early winter AND the submitter comes from the rich county (per capita income) THEN in an average time CFPB sends a complaint.	0.96	0	0.96	0	0.89	0	0.99	0	0.82	0

Acknowledgement This publication was completed while the second author was the Doctoral Candidate in the International Doctoral School at Lodz University of Technology, Lodz, Poland.

References

1. A.Niewiadomski and M. Kacprowicz:Type-2 Fuzzy Logic Systems in Applications: Managing Data in Selective Catalytic Reduction for Air Pollution Prevention. Journal of Artificial Intelligence and Soft Computing Research. 8597 (2021)
2. M. Kacprowicz: Search for outliers by fuzzy logic systems general concepts. TEKI(Technology, Education, Knowledge, Innovation). 6773 (2021)
3. A.Niewiadomski and M. Kacprowicz and M. Bartczak:Outliers Detection In Graph-Represented Databases Using Fuzzy Rules. Pacific Asia Conference on Information Systems, PACIS 2021 (2021).
4. A. Niewiadomski and A. Duraj and M. Bartczak: Outliers Recognition Via Linguistic Aggregation of Graph Databases. Applied Sciences, (2021), Tom 11(16), 7434, MDPI, ISSN: 2076-3417, Doi: 10.3390/app11167434, pp. 1-13,
5. A. Niewiadomski and A. Duraj: Detecting and Recognizing Outliers in Datasets via Linguistic Information and Type-2 Fuzzy Logic. International Journal of Fuzzy Systems, , nr , str. 878889. (2020).
6. B. Kosko: Fuzziness vs. probability. International Journal of General Systems (17). 11240 (1990)
7. J. van den Berg and U. Kaymak and W.-M. van den Bergh: Fuzzy classification using probability-based rule weighting. in Proc. IEEE Intl Conf. on Fuzzy Systems. 991-996 (2002)
8. G. J. Klir and B. Yuan: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Upper Saddle River, NJ:Prentice-Hall (1995)
9. D. Wu and J. M. Mendel, Linguistic summarization using IF-THEN rules. IEEE International Conference on Fuzzy Systems. 1-8 (2010)