Intelligent Detection of Information Outliers Using Linguistic Summaries with Non-monotonic Quantifiers

Agnieszka Duraj[(⊠)] , Piotr S. Szczepaniak , and Lukasz Chomatek

Institute of Information Technology, Lodz University of Technology,
ul. Wolczanska 215, 90-924 Lodz, Poland
{agnieszka.duraj,piotr.szczepaniak,lukasz.chomatek}@p.lodz.pl

Abstract. In the processing of imprecise information, principally in big data analysis, it is very advantageous to transform numerical values into the standard form of linguistic statements. This paper deals with a novel method of outlier detection using linguistic summaries. Particular attention is devoted to examining the usefulness of non-monotonic quantifiers, which represent a fuzzy determination of the amount of analyzed data. The answer is positive. The use of non-monotonic quantifiers in the detection of outliers can provide a more significant value of the degree of truth of a linguistic summary. At the end, this paper provides a computational example of practical importance.

Keywords: Intelligent data analysis · Linguistic summaries · Monotonic and non-monotonic quantifiers · Intelligent outlier detection

1 Introduction

Outliers represent objects whose attributes (or certain attributes) exhibit abnormal behavior in a particular or examined context. Outliers may include unexpected values for all the parameters that describe the object. They may additionally express unexpected values for a particular feature, attribute, or parameter. The customarily used definitions and recent concepts are the following:

– The formal proposition of Hawkins [20] is as follows: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism".
– Barnett and Lewis [3]: "An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data".
– A collection of objects – the subjects of a linguistic summary, is be called outliers if Q objects having the feature S is a true statement in the sense of fuzzy logic, where Q is a selected relative quantifier (e.g. *very few*), and S is a finite, non-empty set of attributes (features) of the set of examined objects, cf. [12,13].

– In the field of Knowledge Discovery in Databases (KDD), or more specific in KDD in Data Mining, outliers are detected as a degree of a deviation from a specified pattern.

The decision to identify outliers is considered when developing decision-making systems, performing intelligent data analysis, and in other situations wherever any impurity or noise affects the proper functioning of systems and may or may not lead to application errors. Therefore, they must be detected and checked to determine if they will be a significant factor or not. Predominantly, there are two distinct approaches to detecting outliers. One way is the case when an object detected as an outlier can be eliminated and deleted at the data preparation stage [18,39]. The second approach assumes the "unique" objects are identified as distinct, retaining an unclear meaning for the processed data [23], and therefore they are not removed. When using artificial intelligence or soft-computing, the methods of detecting outliers are considered to be a part of Intelligent Data Analysis (IDA).

In this paper, the authors show in detail how the use of linguistic summaries given in natural language becomes a method for detecting outliers. The basis is Yager's [40–42] idea of linguistic summaries and some of the numerous extensions and modifications introduced by Kacprzyk and Zadrozny [25,26,29–32]. The innovative aspect of this work lies in the use and examination of non-monotonic quantifiers what reflects situations appearing in practice.

The paper is organized as follows. In Sect. 2, the scopes of related works are briefly presented. Basic definitions of a linguistic variable and non-monotonic quantifiers related to classic fuzzy sets are given in Sect. 3. In the next section, the concept of a linguistic summary and the way of its generation is explained. The practical rules for determining the degree of truth for monotonic and non-monotonic quantifiers are given. In Sects. 5 and 6, the formal definition of an outlier based on the concept of linguistic summary is formulated and the practice of outliers' detection is presented. The ending of the work is constituted by conclusions.

2   Related Works

The scope of applicability of methods of outlier detection in applications is very wide and varied. Numerous works are strictly focus on aimed towards specific applications, e.g. detection of production defects [19], hacker attacks on the computer network [21], fraudulent credit card transactions on credit cards or their abuse [34], public monitoring systems [33], and climate change [2]. There are works that deal with the detection of outliers in networks [16], chat activity, text messages, and the identification of illegal activities [22] in this regard.

There are works on detecting outliers in medical research and applications, e.g. personalized medicine, breast cancer, arrhythmia, monitoring of performance and endurance of athletes, or where outliers are pathogens or anomalies, e.g [1,8,35].

Outliers are distinct, they operate in separate dimensions. Outlier detection methods must, therefore, adapt to both the type of data they work on and the context in which they are operated. Numerous studies indicate an excessive interest in the issue of outlier detection, and the number of approaches increases. This is because we need utilizing a variety of methods adapted to the specific type of data we will analyze. Considering the aforementioned examples into consideration, it should be stated that tasks related to outlier to detection focus on the use of methods dedicated specific sets of data. For example, numerical and textual data - outliers are detected by using linguistic summaries based on classic and interval-valued fuzzy sets [12,13]. Another new approach is application of multiobjective genetic algorithms [7,11].

At present, the complexity of decision problems is constantly increasing. Therefore, authors of many works [6,24–30,32,38] describe not only the implementation and use of linguistic summaries but also emphasize the significance of linguistic summaries in decision-making processes. Moreover, according to Kacprzyk and Zadrozny [26,31] systems based on natural language will continue to develop.

3   Non-monotonic Quantifiers

The idea of a linguistic variable introduced by Zadeh [43,44]. The ideas used in natural language, such as *less than, almost half, about, hardly, few*, etc. can be interpreted as mathematically fuzzy linguistic concepts determining the number of items that fulfill a given criterion. It is worth noting that the relative quantifiers are defined on the interval of real numbers $[0;1]$. They describe the relationship of the objects that meet the summary feature for all items in the analyzed dataset. Absolute quantifiers are defined on the set of non-negative real numbers. They describe the exact number of objects that meet the summary feature. A linguistic quantifier represents a determination of the cardinality. This means that it is a fuzzy set or a single value of the linguistic variable describing the number of objects that meet specific characteristics.

In practical solutions, monotone quantifiers are defined as classic fuzzy sets. For example, the linguistic variable $Q = $ *"few"* can be defined as a membership function in the form of a fuzzy set in the classical form as a trapezoidal or triangular function, etc. However, monotonic quantifiers do not include all possible situations.

The monotonic logic follows an intuitive indication that new knowledge does not reduce the existing set of rules and conclusions. However, it is unable to cope in cases or tasks where some rules must be removed as a consequence of further reasoning and concluding. Problems of non-monotonic logic were introduced in the 1980s. Non-monotonic logic is used to represent formalism, to describe phenomena that cannot be calculated and clearly defined. It has been pointed out that non-monotonicity remain a property of consequence. The logic system is considered to be non-monotonic if its consequence relation possesses a non-monotonic property.

In other words, non-monotonic logic is designed to represent possible conclusions, while initial conclusions may be withdrawn based on other further evidence. Non-monotonicity is closely related to the default conclusions. Non-monotonic formalism is often used in systems based on natural language and many papers present its usefulness, e.g. [4,5,17,36]. In works [12,13], the detection of outliers for monotonic quantifiers was considered. It was observed that the determination of the number used to detect outliers may not always be based on monotonic logic. The *"few"* and *"very few"* quantifiers are of particular importance in this context. However, not all quantifiers meet the condition of monotonicity [37]. The quantifiers should be normal and convex. Normal, because the height of the fuzzy set representing the quantifiers is equal to 1. Convex, because for any $\lambda \in [0,1]$, $\mu_Q(\lambda x_1 + (\lambda - 1)x_2) \geq \min(\mu_Q(x_1) + \mu_Q(x_2))$ We will use the $L - R$ fuzzy number to model the quantifiers with the membership function, where $L, R : [0,1] \longrightarrow [0,1]$ nondecreasing shape functions and $L(0) = R(0) = 0$, $L(1) = R(1) = 1$. The term *"few"*, particularly, is a non-monotonic quantifier, so such linguistic variables can be defined as membership functions in the form of (1).

$$\mu_Q(r) = \begin{cases} L(\frac{r-a}{b-a}) & r \in [a,b] \\ 1 & r \in [b,c] \\ 0 & otherwise \\ R(\frac{d-r}{d-c}) & r \in [c,d] \end{cases} \tag{1}$$

The function (1) can be written as a combination of functions L and R defined by Eqs. (2) and (3).

$$\mu_{Q_1}(r) = \begin{cases} 0 & r < a \\ L(\frac{r-a}{b-a}) & r \in [a,b] \\ 1 & r > b \end{cases} \tag{2}$$

$$\mu_{Q_2}(r) = \begin{cases} 0 & r < c \\ R(\frac{r-c}{d-c}) & r \in [c,d] \\ 1 & r > d \end{cases} \tag{3}$$

In the following section, the non-monotonic quantifiers defined above will be used in linguistic summaries. Both, the monotonic and non-monotonic quantifiers are applied to the detection of exceptions, and the results are compared.

4   Determining the Degree of Truth T_1 in a Linguistic Summary

The definition of linguistic summary introduce by R. Yager [41,42] is as follows.

Definition 1 *The ordered form of four elements [41,42], $<Q; P; S; T_1>$ is called a linguistic summary. Here*
Q - a linguistic quantifier, or quantity in agreement, which is a fuzzy determination of the amount. Quantifier Q determines how many records in an analyzed

database fulfill the following required condition - has the characteristic S.
P - the subject of the summary; it means the actual objects stored in the records of database;
S - the summarizer, the feature by which the database is scanned;
R - the subject's description of the summary;
T_1*- the degree of truth; it determines the extent to which the result of the summary, expressed in a natural language, is true.*

According to the definition of linguistic summaries, we get the response in the natural language of the form:
Q objects being P are (have a feature) S [the degree of truth of this statement is [T_1];
or the extended version:
Q P being R are/have S T_1
where R is the subject's description of the summary;
or in short:
Q P are/have the property S [T_1].
Generating natural language responses as Yager's summaries consist of creating all possible expressions for the predefined quantifiers and summarizers of the analyzed set of objects. The value of the degree of truth for each summary is determined according to $T_1 = \mu_Q(r)$, where $r = \frac{1}{n}\sum_{i=1}^{n}\mu(a_i)$. The value r is determined for each attribute $a_i \in A$. We determine the membership function $\mu_Q(a_i)$, thus defining how well attribute a_i matches the feature given in summarizer S.

Yager's basic linguistic summary takes into consideration only a simple feature that operates on the values of one attribute. The subject is then always a set of analyzed objects in the information system, and the summarizer S denotes that the objects belong to one of the classes of the linguistic variable. Nowadays we can observe numerous extensions of Yager's method. For example, the extension of George and Srikanth [15] which proposed a family of fuzzy sets for features S, R as (4). For multiple attributes (Kacprzyk and Zadrozny's modification [27]) r is defined as (5).

$$r = \frac{1}{n}\sum_{i=1}^{n}(\mu_R(a_i) \cdot \mu_S(b_i)) \tag{4}$$

$$r = \frac{\sum_{i=1}^{n}(\mu_R(a_i) \cdot \mu_S(b_i))}{\sum_{i=1}^{n}\mu_R(a_i)} \tag{5}$$

Example 1. Let's assume we're analyzing a set of data with the attributes: age, blood sugar. If we ask:
How many middle-aged patients have a blood sugar level above average?
The resulting summary could be:
Few middle-aged patients have a blood sugar level above average [0.60].
Many middle-aged patients have a blood sugar level above average [0.25].
Almost all middle-aged patients have a blood sugar level above average [0.15].
The numbers [0.60], [0.25], and [0.15] represent the obtained degrees of truth.

The degree of truth of the linguistic summary with the use of non-monotonic quantifiers is calculated as follows (6) or (7). (cf. Sect. 4 cf. Sect. 3):

$$T_1(Q \ P \ \text{is has} \ S) = T_1(Q_1 \ P \ \text{is (has)} \ S) - T_1(Q_2 \ P \ \text{is (has)} \ S) \qquad (6)$$

$$T_1 = \mu_Q(r) = \mu_{Q_1}(r) - \mu_{Q_2}(r) \qquad (7)$$

5   Detection of Outliers

Let us define the concept of an outlier using a linguistic summary.

Definition 2 *Let $X = \{x_1, x_2, ..., x_N\}$ for $N \in \mathbb{N}$ be a finite, non-empty set of objects. Let S be a finite, non-empty set of attributes (features) of the set of objects X. $S = \{s_1, s_2, ..., s_n\}$.*
Let Q be relative quantifiers.
A collection of objects, which are the subjects of a linguistic summary, will be called outliers if Q objects having the feature S is a true statement in the sense of fuzzy logic.
If the linguistic summary of Q objects in P are/have S, $[T_1]$ and $T_1 > 0$ (therefore, it is true in the sense of fuzzy logic), than outliers were found.

The procedure for detecting outliers using linguistic summaries according to Definition 2 begins with defining a set of linguistic values $X = \{Q_1, Q_2, ..., Q_n\}$. The next step is to calculate the value of r according to the procedure of generating linguistic summary described in Sect. 4. We determine T_1 for classic fuzzy sets. If we used non-monotonic quantifiers in the form of classic fuzzy sets, the degree of the truth T_1 can be determined according to (7).

One obtains
$Q_1 \ P \ \text{is (has)} \ S \ [T_1]$
$Q_2 \ P \ \text{is (has)} \ S \ [T_1]$
...
$Q_N \ P \ \text{is (has)} \ S \ [T_1]$

It is known that if $T_1 > 0$, one obtains a true sentence in the Zadeh's sense. Outliers are found if $T_1 > 0$ for q_i, where q_i is defined as: *very few, few, almost none*, and the like. For example, if for the linguistic variable $Q_1 = \{few\}T_1 > 0$ then one can expect that outliers are present.

If the set of linguistic variables is composed of several values like *"very few"*, *"few"*, *"almost none"*, then all summaries generated for those variables whose values are $T_1 > 0$ should be taken into consideration. In practical applications [10,12,14], the authors take into account the maximum to variables characterizing outliers. There exist four sets of possible responses, which are given in Table 1. Consequently, the use of linguistic summaries enables to generate information if outliers exist in the data bases under consideration. Note, that for the companies' management, the information provided in the linguistic form is of preferable form. The non-trivial example is examined in the following section.

Table 1. Types of responses of a linguistic summary indicating the existence of outliers.

Q_1 P is (has) S $[T_1]$	$T_1 = 0$	Outliers have not been found
Q_2 P is (has) S $[T_1]$	$T_1 > 0$	Outliers have been found
Q_1 P is (has) S $[T_1]$	$T_1 > 0$	Outliers have been found
Q_2 P is (has) S $[T_1]$	$T_1 = 0$	Outliers have not been found
Q_1 P is (has) S $[T_1]$	$T_1 > 0$	Outliers have been found
Q_2 P is (has) S $[T_1]$	$T_1 > 0$	Outliers have been found
Q_1 P is (has) S $[T_1]$	$T_1 = 0$	Outliers have not been found
Q_2 P is (has) S $[T_1]$	$T_1 = 0$	Outliers have not been found

6   The Practice of Outliers' Detection

Let us consider a set describing the activities of enterprises. The dataset was composed of publicly available data from Statistics Poland [9]. The examined set consists of many attributes which allow to reason about the accounting liquidity of enterprises. Attributes include, among others: company size, short-term liabilities, long-term liabilities, company assets, number of employees, financial liquidity ratio and bankruptcy risk.

Example of the data is presented in the Table 2. Current Ratio measures whether resources owned by a company are enough to meet its short-term obligations. All the calculations were performed in Java and R environment.

Table 2. Sample records of the dataset analyzed in the paper

ID	Current ratio	$\cdots$	Bankruptcy risk
0001	0.148	$\cdots$	0.2
0002	3.65	$\cdots$	1.45
$\cdots$	$\cdots$	$\cdots$	$\cdots$
0298	2.44	$\cdots$	0.54
0299	4.39	$\cdots$	0.12
$\cdots$	$\cdots$	$\cdots$	$\cdots$
1587	1.74	$\cdots$	2.13
1588	0.43	$\cdots$	0.73
$\cdots$	$\cdots$	$\cdots$	$\cdots$

Let us consider the two following questions.
Query 1: How many enterprises with a high current ratio are in the high risk of bankruptcy group?
Query 2: How many enterprises with low profitability are in the high-risk group?

Let for the linguistic variables describing the risk of bankruptcy, the considered values are: *low, medium and high.* For the current ratio of a company, the assumed values are: *very low, low, medium, and high.*

For the each values (*low, medium, high*) the risk of bankruptcy is determined using trapezoidal membership functions

$Trap[x, a, b, c, d] = 0 \vee (1 \wedge \frac{x-a}{b-a} \wedge \frac{d-x}{d-c}), a < b \le c < d, x \in X$:

$Trap_{low}[0, 0, 0, 2, 0.4]$, $Trap_{medium}[0.3, 0.5, 0.7, 0.9]$ and $Trap_{high}[0.6, 0.8, 1, 1]$.

Similarly, the membership functions of the current liquidity indicator can be defined.

6.1   Monotonic Quantifiers

According to the procedure for detecting outliers using linguistic summaries, the set of linguistic values must be defined, here $Q=\{$ *"very few", "few", "many", "almost all"*$\}$ and the trapezoidal form is chosen:

$Trap_{veryfew}[0, 0.1, 0.2, 0.3]$, $Trap_{few}[0.15, 0.3, 0.45, 0.6]$,

$Trap_{many}[0.5, 0.65, 0.8, 0.95]$, $Trap_{almostall}[0.75, 0.9, 1, 1]$.

On the basis of Eq. (5), the values of the coefficient of r for the two queries of interest are calculated as (8), where cls is the current liquidity indicator and $risk$ indicates the risk of bankruptcy.

$$r_{Query1} = \frac{\sum_{i=1}^{n}(\mu_{risk}(a_i) \cdot \mu_{cli}(b_i))}{\sum_{i=1}^{n}\mu_{risk}(a_i)} = 0.28 \tag{8}$$

$$r_{Query2} = \frac{\sum_{i=1}^{n}(\mu_{risk}(a_i) \cdot \mu_{prof}(b_i))}{\sum_{i=1}^{n}\mu_{risk}(a_i)} = 0.34 \tag{9}$$

The obtained linguistic summaries are of the form: Query No. 1:

Very few enterprises with a high current ratio are in the high risk of bankruptcy group; $T_1[0.2]$.

Few enterprises with a high current ratio are in the high risk of bankruptcy group; $T_1[0, 86]$.

Many enterprises with a high current ratio are in the high risk of bankruptcy group; $T_1[0]$.

Almost all enterprises with a high current ratio are in the high risk of bankruptcy group; $T_1[0]$.

According to the Definition 2 outliers were detected – see the values of the degree of truth T_1 for *few* and *very few.*

Query No. 2:

Very few enterprises with low profitability are in the high risk group; $T_1[0]$.

Few enterprises with low profitability are in the high risk group; $T_1[1]$.

Many enterprises with low profitability are in the high risk group; $T_1[0]$.

Almost all enterprises with low profitability are in the high risk group; $T_1[0]$.

Outliers were not detected because $T_1 = 0$ for all quantifiers.

6.2   Non-monotonic Quantifiers

Let the linguistic variables $Q_1=$ *"very few"* and $Q_2=$ *"few"* now be non-monotonic classic fuzzy sets. According to the Eq. (1), in this case the membership function Q_1 is transformed into two functions (2) and (3), and one obtains (10) and (13). Similarly for Q_2 we have (12) and (11).

$$\mu_{Q_{vf_1}}(r) = \begin{cases} 0 & r < 0 \\ L(\frac{r}{0.1}) & r \in [0, 0.1] \\ 1 & r > 0.1 \end{cases} \tag{10}$$

$$\mu_{Q_{vf_2}}(r) = \begin{cases} 0 & r < 0.2 \\ R(\frac{r-0.2}{0.1}) & r \in [0.2, 0.3] \\ 1 & r > 0.3 \end{cases} \tag{11}$$

$$\mu_{Q_{f_1}}(r) = \begin{cases} 0 & r < 0.15 \\ L(\frac{r-0.15}{0.15}) & r \in [0.15, 0.3] \\ 1 & r > 0.3 \end{cases} \tag{12}$$

$$\mu_{Q_{f_2}}(r) = \begin{cases} 0 & r < 0.45 \\ R(\frac{r-0.45}{0.15}) & r \in [0.45, 0.6] \\ 1 & r > 0.6 \end{cases} \tag{13}$$

The next step in the procedure of detecting outliers is to calculate the value of the coefficient of r. We use the Eq. (5) for Query No. 1 (8), Eq. (9) for Query No. 2. In the case of non-monotonic quantifiers T_1, we designate with (7).

We received the following generated sentences:

Query No. 1:

Very few enterprises with a high current ratio are in the high risk of bankruptcy group. T_1 [0.7]
because:

$$T_1(\mu_{Q_{veryfew}}) = T_1(\mu_{Q_{vf1}}) - T_1(\mu_{Q_{vf2}}) = 1 - 0.3 = 0.7$$

Few enterprises with a high current ratio are in the high risk of bankruptcy group. $T_1[0.86]$
because:

$$T_1(\mu_{Q_{few}}) = T_1(\mu_{Q_{f1}}) - T_1(\mu_{Q_{f2}}) = 0.86 - 0 = 0.86$$

Many enterprises with a high current ratio are in the high risk of bankruptcy group. $T_1[0]$
Almost all enterprises with a high current ratio are in the high risk of bankruptcy group. $T_1[0]$

Query No. 2:

Very few enterprises with low profitability are in the high risk group T_1 [0.1]. because:

$$T_1(\mu_{Q_{veryfew}}) = T_1(\mu_{Q_{vf1}}) - T_1(\mu_{Q_{vf2}}) = 1 - 0.9 = 0.1$$

Few enterprises with low profitability are in the high risk group $T_1[1]$. because:

$$T_1(\mu_{Q_{few}}) = T_1(\mu_{Q_{f1}}) - T_1(\mu_{Q_{f2}}) = 1 - 0 = 1$$

Many enterprises with low profitability are in the high risk group $T_1[0]$.
Almost all enterprises with low profitability are in the high risk group $T_1[0]$.

Table 3. The results of the degree of truth for the monotonic and non-monotonic quantifiers *very few* and *few*

Query	Quantificator	Monotonic	Non-monotonic
No. 1	Very few	0.2	0.7
No. 1	Few	0.86	0.86
No. 2	Very few	0.0	0.1
No. 2	Few	1	1

In Table 3, illustration of the degree of truth obtained for both monotonic and non-monotonic quantifiers is given. Application of non-monotonic quantifiers also indicates the existence of outliers but the value of the degree of truth is bigger. This fact can be interpreted that non-monotonic quantifiers give higher reliability of the result.

7   Conclusions

The aim of this study was to present a non-standard approach to the detection of outliers using linguistic summaries. It is a practical solution to the mentioned problem when a dataset is of numeric, or both numeric and linguistic character. However, the text attributes should be partially standardized. The presented idea is based on the summaries introduced by Yager. Other well-known standard approaches cannot directly be used for the analysis of textual or mixed data, and this is a significant advantage of the method which can operate in the case of big data evaluation as well. The results obtained in the form of sentences in a natural language are understandable and user friendly. This paper has introduced an algorithm for detecting outliers using non-monotonic quantifiers in linguistic summaries based on classic fuzzy sets. The non-monotonic quantifiers has not been considered in any of the previous studies on outlier detection with the use of linguistic summaries. In Sect. 6, the performance of the algorithm was

illustrated. The conducted research and experiments confirm, that it is possible to detect outliers using linguistic summaries. To be specific, the work verified the correct functioning of the proposed method for non-monotonic quantifiers. This method enhances database analysis and decision-making processes, and it is useful for managers and data science experts.

References

1. Aggarwal, C.C.: Toward exploratory test-instance-centered diagnosis in high-dimensional classification. IEEE Trans. Knowl. Data Eng. **19**(8), 1001–1015 (2007)
2. Angiulli, F., Basta, S., Pizzuti, C.: Distance-based detection and prediction of outliers. IEEE Trans. Knowl. Data Eng. **18**(2), 145–160 (2006)
3. Barnett, V., Lewis, T.: Outliers in Statistical Data, 584 p. Wiley, Chichester (1964)
4. Benferhat, S., Dubois, D., Prade, H.: Nonmonotonic reasoning, conditional objects and possibility theory. Artif. Intell. **92**(1–2), 259–276 (1997)
5. van Benthem, J., Ter Meulen, A.: Handbook of Logic and Language. Elsevier, Amsterdam (1996)
6. Boran, F.E., Akay, D., Yager, R.R.: A probabilistic framework for interval type-2 fuzzy linguistic summarization. IEEE Trans. Fuzzy Syst. **22**(6), 1640–1653 (2014)
7. Chomatek, L., Duraj, A.: Multiobjective genetic algorithm for outliers detection. In: 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pp. 379–384. IEEE (2017)
8. Cramer, J.A., Shah, S.S., Battaglia, T.M., Banerji, S.N., Obando, L.A., Booksh, K.S.: Outlier detection in chemical data by fractal analysis. J. Chemom. **18**(7–8), 317–326 (2004)
9. Databases: Statistic Poland. https://stat.gov.pl/en/databases/
10. Duraj, A.: Outlier detection in medical data using linguistic summaries. In: 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pp. 385–390. IEEE (2017)
11. Duraj, A., Chomatek, L.: Supporting breast cancer diagnosis with multi-objective genetic algorithm for outlier detection. In: Kościelny, J.M., Syfert, M., Sztyber, A. (eds.) DPS 2017. AISC, vol. 635, pp. 304–315. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-64474-5_25
12. Duraj, A., Niewiadomski, A., Szczepaniak, P.S.: Outlier detection using linguistically quantified statements. Int. J. Intell. Syst. **33**(9), 1858–1868 (2018)
13. Duraj, A., Niewiadomski, A., Szczepaniak, P.S.: Detection of outlier information by the use of linguistic summaries based on classic and interval-valued fuzzy sets. Int. J. Intell. Syst. **34**(3), 415–438 (2019)
14. Duraj, A., Szczepaniak, P.S.: Information outliers and their detection. In: Burgin, M., Hofkirchner, W. (eds.) Information Studies and the Quest for Transdisciplinarity, vol. 9, Chapter 15, pp. 413–437. World Scientific Publishing Company (2017)
15. George, R., Srikanth, R.: Data summarization using genetic algorithms and fuzzy logic. In: Genetic Algorithms and Soft Computing, pp. 599–611 (1996)
16. Giatrakos, N., Kotidis, Y., Deligiannakis, A., Vassalos, V., Theodoridis, Y.: In-network approximate computation of outliers with quality guarantees. Inf. Syst. **38**(8), 1285–1308 (2013)
17. Giordano, L., Gliozzi, V., Olivetti, N., Pozzato, G.L.: A non-monotonic description logic for reasoning about typicality. Artif. Intell. **195**, 165–202 (2013)

18. Guevara, J., Canu, S., Hirata, R.: Support measure data description for group anomaly detection. In: ODDx3 Workshop on Outlier Definition, Detection, and Description at the 21st ACM SIGKDD International Conference On Knowledge Discovery And Data Mining (KDD2015) (2015)

19. Guo, Q., Wu, K., Li, W.: Fault forecast and diagnosis of steam turbine based on fuzzy rough set theory. In: Second International Conference on Innovative Computing, Information and Control, ICICIC 2007, pp. 501–501. IEEE (2007)

20. Hawkins, D.M.: Identification of Outliers, vol. 11. Springer, Heidelberg (1980). https://doi.org/10.1007/978-94-015-3994-4

21. Hawkins, S., He, H., Williams, G., Baxter, R.: Outlier detection using replicator neural networks. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds.) DaWaK 2002. LNCS, vol. 2454, pp. 170–180. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-46145-0_17

22. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. **24**(9), 1641–1650 (2003)

23. Jayakumar, G., Thomas, B.J.: A new procedure of clustering based on multivariate outlier detection. J. Data Sci. **11**(1), 69–84 (2013)

24. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets Syst. **159**(12), 1485–1499 (2008)

25. Kacprzyk, J., Wilbik, A., Zadrozny, S.: Linguistic summaries of time series via a quantifier based aggregation using the Sugeno integral. In: 2006 IEEE International Conference on Fuzzy Systems, pp. 713–719. IEEE (2006)

26. Kacprzyk, J., Wilbik, A., Zadrożny, S.: An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. Int. J. Intell. Syst. **25**(5), 411–439 (2010)

27. Kacprzyk, J., Yager, R.R.: Linguistic summaries of data using fuzzy logic. Int. J. Gen. Syst. **30**(2), 133–154 (2001)

28. Kacprzyk, J., Yager, R.R., Zadrożny, S.: A fuzzy logic based approach to linguistic summaries of databases. Int. J. Appl. Math. Comput. Sci. **10**(4), 813–834 (2000)

29. Kacprzyk, J., Yager, R.R., Zadrozny, S.: Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support. In: Abramowicz, W., Zurada, J. (eds.) Knowledge Discovery for Business Information Systems. SECS, vol. 600, pp. 129–152. Springer, Heidelberg (2002). https://doi.org/10.1007/0-306-46991-X_6

30. Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. Inf. Sci. **173**(4), 281–304 (2005)

31. Kacprzyk, J., Zadrozny, S.: Protoforms of linguistic database summaries as a human consistent tool for using natural language in data mining. Int. J. Softw. Sci. Comput. Intell. (IJSSCI) **1**(1), 100–111 (2009)

32. Kacprzyk, J., Zadrożny, S.: Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation. IEEE Trans. Fuzzy Syst. **18**(3), 461–472 (2010)

33. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. VLDB J. Int. J. Very Large Data Bases **8**(3–4), 237–253 (2000). https://doi.org/10.1007/s007780050006

34. Last, M., Kandel, A.: Automated detection of outliers in real-world data. In: Proceedings of the Second International Conference on Intelligent Technologies, pp. 292–301 (2001)

35. Ng, R.: Outlier detection in personalized medicine. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, p. 7. ACM (2013)
36. Schulz, K., Van Rooij, R.: Pragmatic meaning and non-monotonic reasoning: the case of exhaustive interpretation. Linguist. Philos. **29**(2), 205–250 (2006). https://doi.org/10.1007/s10988-005-3760-4
37. Wilbik, A., Kaymak, U., Keller, J.M., Popescu, M.: Evaluation of the truth value of linguistic summaries – case with non-monotonic quantifiers. In: Angelov, P., et al. (eds.) Intelligent Systems 2014. AISC, vol. 322, pp. 69–79. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-11313-5_7
38. Wilbik, A., Keller, J.M.: A fuzzy measure similarity between sets of linguistic summaries. IEEE Trans. Fuzzy Syst. **21**(1), 183–189 (2013)
39. Xiong, L., Póczos, B., Schneider, J., Connolly, A., Vander Plas, J.: Hierarchical probabilistic models for group anomaly detection. In: International Conference on Artificial Intelligence and Statistics 2011, pp. 789–797 (2011)
40. Yager, R.: Linguistic summaries as a tool for databases discovery. In: Workshop on Fuzzy Databases System and Information Retrieval (1995)
41. Yager, R.R.: A new approach to the summarization of data. Inf. Sci. **28**(1), 69–86 (1982)
42. Yager, R.R.: Linguistic summaries as a tool for database discovery. In: FQAS, pp. 17–22 (1994)
43. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
44. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-iii. Inf. Sci. **9**(1), 43–80 (1975)