# Linguistic Summaries using Interval-valued Fuzzy Representation of Imprecise Information - an Innovative Tool for Detecting Outliers

Agnieszka Duraj[1][0000−0002−3047−6662]
and Piotr S. Szczepaniak[1][0000−0002−9973−0673]

Institute of Information Technology, Łódź University of Technology, Poland
ul. Wólczańska 215, 90-924 Lodz, Poland
{agnieszka.duraj}@p.lodz.pl

**Abstract.** The practice of textual and numerical information processing often involves the need to analyze and test a database for the presence of items that differ substantially from other records. Such items, referred to as outliers, can be successfully detected using linguistic summaries. In this paper, we extend this approach by the use of non-monotonic quantifiers and interval-valued fuzzy sets. The results obtained by this innovative method confirm its usefulness for outlier detection, which is of significant practical relevance for database analysis applications.

**Keywords:** intelligent data analysis · outlier detection · linguistic summaries · fuzzy handling of linguistic uncertainty · interval-valued fuzzy sets · non-monotonic quantifiers · data mining · knowledge discovery

## 1 Introduction

Fuzzy sets (also known as uncertain sets) and fuzzy logic have been jointly used to deal with the uncertainty related to human perception and classification [1, 2]. This methodology enables the modeling of uncertainty features. It allows one to represent, in an integrated way, the uncertainty of qualitative terms based on quantitative aspects of various phenomena. Linguistic summarization based on fuzzy concepts has been found to be a useful method for the qualitative analysis of databases, including outlier detection [3].

The concept of an outlier, as related to data analysis, has been widely discussed in the literature, and consequently, many definitions of this term have been proposed, e.g. [4–6]. Information outliers represent information granules that are unexpected, occur rarely, or exhibit abnormal characteristics. Outlier detection is a crucial step in many data mining applications, as the presence of outliers in a data set may translate into serious errors. When using computational approaches, the representation of information hidden in the records of big databases must be extracted in a human-friendly form. In data mining and knowledge discovery tasks, outliers are understood as a degree of deviation from a specified information pattern. The present paper examines databases that contain both numeric and textual records.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

In general, there are two approaches to detecting outliers: they may be either eliminated at the stage of data preparation [7, 8] or saved [9]. In the latter case, the unique objects are defined as small clusters that are dissimilar to the other data. The basis for the approach presented is Yager's [10–12] idea of linguistic summaries and the selected extensions developed by Kacprzyk and Zadrozny [13–17]. This paper provides a new insight into how this idea may be applied to the problem of outlier detection. The innovative approach presented involves the use of two concepts: non-monotonic quantifiers and interval-valued fuzzy sets, both applicable to outlier detection problems arising in practice. Specifically, we examine the usefulness of non-monotonic quantifiers for outlier detection methods based on linguistic summarization of database contents. The present study endeavors to show that non-monotonic interval-valued fuzzy sets can provide a more significant value of the degree of truth of a linguistic summary.

The content of this work is split into several sections. In Section 2, a brief survey of related literature is presented. Basic definitions of non-monotonic interval-valued fuzzy quantifiers are given in Section 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 demonstrated in Section 4. Section 5 begins by introducing the concept of an outlier and presents the procedure of outlier detection using linguistic summaries. Section 6 demonstrates the practical application of the concept of outlier detection using linguistic summaries and two types of quantifiers, namely interval-valued monotonic and non-monotonic ones. Finally, our conclusions are drawn in Section 7.

## 2 Related works

The past decade has seen the rapid development of outlier detection methods. Often designed for a particular type of data, these methods have been used in many varied applications, for example, medical research, where the outlier is defined as an anomaly or pathogen [18–20].

Outlier detection is also often used in public monitoring systems [21], climate change research [22], computer networks — for identifying hacker attacks [4, 23], banking — for detecting fraud and fraudulent transactions on credit cards [24], or in manufacturing — for detection of defects [25]. The classic approach to outlier detection is based on distance or density. There are also other methods dedicated to specific types of data.

As observed by Kacprzyk and Zadrożny [17, 26], there is an ongoing trend towards natural language-based knowledge discovery systems. Many researchers have demonstrated the use of linguistic summaries in decision-making processes [15–17, 27–29]. In [30, 31], linguistic summaries based on classic and interval-valued fuzzy sets were successfully applied to outlier detection tasks. As demonstrated in [32, 33], outlier detection can also be performed using monotonic quantifiers.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

Non-monotonicity is closely related to default conclusions. Many papers have presented the usefulness of this kind of formalism for developing natural language-based systems, e.g. [34–37]. The new aspect of this work lies in the use of non-monotonic interval-valued fuzzy quantifiers. We demonstrate that linguistic summaries supported by non-monotonic interval-valued fuzzy quantifiers may prove even more useful than previous outlier detection solutions.

## 3 Non-monotonic quantifiers and interval-valued fuzzy implementation

The concept of a linguistic variable, introduced by Zadeh [38, 39], enables the description of complex and ill-defined phenomena, which are difficult to specify using quantitative methods.

The concepts used in the natural language, such as *less than, almost half, about, hardly, few*, can be interpreted as mathematically fuzzy linguistic concepts determining the number of objects that meet a given criterion. Note that relative quantifiers are defined on the interval of real numbers $[0; 1]$. They describe the relationship of objects that meet the summary feature for all objects in the analyzed data set. Absolute quantifiers are defined on a set of non-negative real numbers. They describe the exact number of objects that meet the summary feature.

A linguistic quantifier being a determination of the cardinality is then a fuzzy set or a single value of the linguistic variable describing the cardinality of objects that meet specific characteristics.

In practical solutions, monotone quantifiers are defined as classic fuzzy sets or interval fuzzy sets. For example, the linguistic variable $Q = \{few\}$ can be a trapezoidal or triangular fuzzy membership function. It can also be designated as a function in the form of a fuzzy interval set.

Obviously, not all quantifiers of practical significance meet the condition of monotonicity [28]. The *few* and *very few* quantifiers are of particular importance in the context of detecting abnormal objects.

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 $\lambda \in [0, 1]$, $\mu_Q(\lambda x_1 + (\lambda - 1)x_2) \geq \min(\mu_Q(x_1) + \mu_Q(x_2))$, where $Q$ is a chosen, relevant quantifier, e.g. $fe''$, and $x_1, x_2 \in X$ are the objects considered.

We use the $L-R$ fuzzy number to model the quantifiers with the membership function, where $L, R : [0, 1] \longrightarrow [0, 1]$ non-decreasing shape functions and $L(0) = R(0) = 0$, $L(1) = R(1) = 1$. If the term *few* is a non-monotonic quantifier, it can be defined as a membership function in the form of (1).

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

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

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

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

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

In interval-valued fuzzy sets, the degree of membership to the set is defined as an interval of real numbers [0; 1]. Thus, we obtain two membership functions: lower membership function $\underline{\mu}$, which determines the minimum degree of membership of an element, and upper membership function $\overline{\mu}$, which determines the maximum degree of the membership.

Having defined the linguistic variable $Q$ as a non-monotonic interval-valued fuzzy set, we make similar changes for equations (1) as (4) where $\mu_{\underline{Q}}(x)$ is calculated as (5) and $\mu_{\overline{Q}}(x)$ as (6).

$$\mu_Q(x) = [\mu_{\underline{Q}}(x), \mu_{\overline{Q}}(x)] \tag{4}$$

According to the definition of interval-valued fuzzy sets, we can define a non-monotonic quantifier $Q$ (1) as interval-valued fuzzy set as (5) and (6).

$$\mu_{\underline{Q}}(r) = \begin{cases} L(\frac{r-\underline{a}}{\underline{b}-\underline{a}}) & r \in [\underline{a},\underline{b}] \\ 1 & r \in [\underline{b},\underline{c}] \\ 0 & otherwise \\ R(\frac{r-\underline{d}}{\underline{d}-\underline{c}}) & r \in [\underline{c},\underline{d}] \end{cases} \tag{5}$$

$$\mu_{\overline{Q}}(r) = \begin{cases} L(\frac{r-\overline{a}}{\overline{b}-\overline{a}}) & r \in [\overline{a},\overline{b}] \\ 1 & r \in [\overline{b},\overline{c}] \\ 0 & otherwise \\ R(\frac{\overline{d}-r}{\overline{d}-\overline{c}}) & r \in [\overline{c},\overline{d}] \end{cases} \tag{6}$$

Additionally, it is known that $\mu_Q(x)$ as $Q$ non-monotonic quantifiers can be written as a combination of functions (2) and (3) as (7) and (8), where $\mu_{\underline{Q_L}}$, $\mu_{\overline{Q_L}}$, $\mu_{\underline{Q_R}}$ and $\mu_{\overline{Q_R}}$ are determined by (9), (10), (11) and (12).

$$\mu_{Q_L}(x) = [\mu_{\underline{Q_L}}(x), \mu_{\overline{Q_L}}(x)] \tag{7}$$

$$\mu_{Q_R}(x) = [\mu_{\underline{Q_R}}(x), \mu_{\overline{Q_R}}(x)] \tag{8}$$

$$\mu_{\underline{Q_L}}(r) = \begin{cases} 0 & r < \underline{a} \\ L(\frac{r-\underline{a}}{\underline{b}-\underline{a}}) & r \in [\underline{a},\underline{b}] \\ 1 & r > \underline{b} \end{cases} \tag{9}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

$$\mu_{\overline{Q_L}}(r) = \begin{cases} 0 & r < \overline{a} \\ L(\frac{r-\overline{a}}{\overline{b}-\overline{a}}) & r \in [\overline{a}, \overline{b}] \\ 1 & r > \underline{b} \end{cases} \tag{10}$$

$$\mu_{\underline{Q_R}}(r) = \begin{cases} 0 & r < \underline{c} \\ R(\frac{r-\underline{c}}{\underline{d}-\underline{c}}) & r \in [\underline{c}, \underline{d}] \\ 1 & r > \underline{d} \end{cases} \tag{11}$$

$$\mu_{\overline{Q_R}}(r) = \begin{cases} 0 & r < \overline{c} \\ R(\frac{r-\overline{c}}{\overline{d}-\overline{c}}) & r \in [\overline{c}, \overline{d}] \\ 1 & r > \overline{d} \end{cases} \tag{12}$$

In the following sections, the above definitions of non-monotonic quantifiers are used.

## 4 Linguistic summary

The collection of linguistic variables, referred to as linguistic quantifiers, is the expert knowledge used in linguistic summaries. This linguistic summary in a strictly structured form, expressed in a natural (or close to natural) language, is generated on the basis of the information contained in the information system and expert knowledge in the particular field. The definition of a linguistic summary is given by Def. 1.

**Def. 1** *Yager's linguistic summary*
*Yager's linguistic quantifier is in the form of ordered four elements*
$< Q; P; S; T >$
*where:*
$Q$ *- a linguistic quantifier, or quantity in agreement, which is a fuzzy determination of amount. Quantifier* $Q$ *determines how many records in the analyzed database fulfill the required condition - have the characteristic* $S$*;*
$P$ *- the subject of summary; the actual objects stored in the database;*
$S$ *- the summarizer, i.e., the feature by which the database is scanned;*
$T$ *- 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 $[T]$, or in short:
$Q$ $P$ are/have the property $S$ $[T]$.

Generating natural language responses as Yager's summaries consists 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

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

is determined according to formula $T = \mu_Q(r)$, where $r$ is defined in (13). The value $r$ is determined for each attribute $a_i \in A$. We determine the membership function $\mu_A(a_i)$, thus defining how well attribute $a_i$ matches characteristic $S$.

$$r = \frac{\sum_{i=1}^{n}(\mu_R(a_i)\cdot \mu_S(b_i))}{\sum_{i=1}^{n}\mu_R(a_i)} \tag{13}$$

For linguistic summaries that apply interval-valued fuzzy sets, we obtain the value of the degree of truth in the form of an interval, not a number. It is possible to introduce interval-valued fuzzy sets for sets of linguistic variables $Q$ or features $R$, $S$. We then get $T$ in the form (14).

$$T = [\underline{T}, \overline{T}] = [\underline{\mu}_Q(r), \overline{\mu}_Q(r)] \tag{14}$$

If features $R$ or $S$ are defined as an interval-valued fuzzy set, then $r = [\underline{r}, \overline{r}]$ is defined as follows (15).

$$[\underline{r}, \overline{r}] = [\frac{\sum_{i=1}^{n}(\underline{\mu}_R(a_i)\cdot \underline{\mu}_S(b_i)}{\sum_{i=1}^{n}\underline{\mu}_R(a_i)}, \frac{\sum_{i=1}^{n}(\overline{\mu}_R(a_i)\cdot \overline{\mu}_S(b_i)}{\sum_{i=1}^{n}\overline{\mu}_R(a_i)}] \tag{15}$$

The application of non-monotonic linguistic quantifiers to linguistic summaries affects the manner of calculating the degree of truth. For linguistic summaries with the defined interval-valued non-monotonic quantifiers, the degree of truth is determined as (16).

$$T = [(\mu_{\underline{Q_L}}(r) - \mu_{\underline{Q_R}}(r)), (\mu_{\overline{Q_L}}(r) - \mu_{\overline{Q_R}}(r))] \tag{16}$$

The membership function for the classifier is selected by the user, i.e. the expert. For non-monotonic quantifiers obtained in the linguistic summary, the value of the degree of truth could be higher. The summary becomes more reliable, which is important for detecting exceptions.

## 5 Detection of Outliers

An outlier is treated as a single element or a very small group of objects which, in comparison with other objects in the database, differ in the values of the analyzed feature. Let us define the concept of an outlier using a linguistic summary.

**Def. 2** *Let:*
$X = \{x_1, x_2, \ldots, x_N\}$ *for* $N \in \mathbb{N}$ *be a finite, non-empty set of objects;*
$S$ *be a finite, non-empty set of attributes (features) of the set of objects* $X$, $S = \{s_1, s_2, \ldots, s_n\}$;
$Q$ *be non-monotonic interval-valued quantifiers defining the requested low cardinality.*
***A collection of objects (the subjects of a linguistic summary) are called outliers if*** $Q$ ***objects having the feature*** $S$ ***is a true statement in the sense of interval-valued fuzzy logic****. If the linguistic summary of* $Q$ *objects in the* $P$ *is/has* $S$, $[\underline{T}, \overline{T}]$ *has* $[\underline{T}, \overline{T}] > 0$ *(therefore, it is true in the sense of fuzzy logic) which means that outliers have been found.*

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

The procedure for detecting outliers using linguistic summaries according to Def. 2 begins with defining a set of linguistic values $Q = \{Q_1, Q_2, ..., Q_n\}$. For example, $Q_1$=*very few*, $Q_2 = few$, $Q_3 = many$, $Q_4$=*almost all*. The next step is to calculate the value of $r$ according to the procedure for generating the linguistic summary described in Section 4.

As for detecting outliers using linguistic summaries, according to Def. 1, the most important elements are linguistic variables defining cardinality as *very few, few,little, almost none*, etc. If for the linguistic variable $Q_i$ (e.g. $Q_1$=*very few*, $Q_2 = few$) defined according to Def. 1, the value of a measure $[\underline{T}, \overline{T}] > 0$, then the resulting sentence is true in the sense of Zadeh's fuzzy logic, and thus, according to Def. 2, outliers have been detected.

In the practical applications [30, 32, 3, 40], the authors took into account a maximum of two variables characterizing the outliers. In such a case, four responses are possible.

The same assumptions should be made when the set of linguistic variables is determined by interval-values fuzzy sets. According to the definition of linguistic summary, in which the set of linguistic variables $Q$ is defined as interval-valued fuzzy sets, we obtain the degree of truth for each generated sentence in the form of (14) for monotonic quantifiers and (16) for non-monotonic quantifiers. Outliers are found if the interval-valued fuzzy $T$ contains values greater than 0. Then, four responses are possible, as shown in Table 1.

**Table 1.** Types of responses of the system based on interval-valued fuzzy sets, with two quantifiers $Q_1$ and $Q_2$: *very few* and *few*.

| Response | Degree of truth | Result |
|---|---|---|
| $Q_1$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}] = [0, 0]$ | Outliers were found |
| $Q_2$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}] >$ [0,0] | for the linguistic variable $Q_2$ |
| $Q_1$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}] >$ [0,0] | Outliers were found |
| $Q_2$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}] =$ [0,0] | for the linguistic variable $Q_1$ |
| $Q_1$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}] >$ [0,0] | Outliers were found |
| $Q_2$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}] >$ [0,0] | for the linguistic variables $Q_1$ and $Q_2$ |
| $Q_1$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}]$=[0,0] | Outliers were not found |
| $Q_2$ $P$ is (has) $S$ $[\underline{T}, \overline{T}]$ | $[\underline{T}, \overline{T}]$ =[0,0] | |

## 6 Practical examples

The dataset used for the present analysis is composed of publicly available data from Statistics Poland [41]. It is a collection of 20 attributes on the basis of which we may reason about the financial liquidity of enterprises. The attributes are: company size, short-term liabilities, long-term liabilities, company assets, number of employees, financial liquidity ratio, bankruptcy risk, etc. The novelty

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

of the present approach and its practical evaluation lies in the introduction of non-monotonic quantifiers based on interval-valued fuzzy sets.

The method of detecting outliers using linguistic summaries is demonstrated using two queries.
Query 1:
How many enterprises with a high current liquidity ratio are in the group with a high risk of bankruptcy?
Query 2:
How many enterprises with low profitability are in the high-risk group?

A bankruptcy risk score is a number that indicates whether a company or an individual has a high probability of becoming insolvent. There is no single, universally agreed-upon index of measurement [42, 43]. For example, the Altman Z-score [44] relies on five financial factors: profitability, leverage, liquidity, solvency and activity. In the banking sector, it is common practice to employ various bankruptcy risk scoring methodologies as a tool for assessing people's creditworthiness [45]. Below, we present a linguistically motivated approach to bankruptcy risk estimation. The exact definition of linguistic expressions such as low, medium, high etc. depends on the policy adopted by a given financial institution. A similar linguistic approach can be applied to the liquidity, profitability, leverage, solvency, and activity ratios used in the Altman Z-score. The analysis of bankruptcy risk involves two key steps: definition of uncertainty levels and estimation of the impact of uncertainty. One also needs a bridge between the qualitative and quantitative analysis. This bridge is provided by the fuzzy sets theory where the shapes of membership functions and their parameters are defined by the users or domain experts.

The linguistic variables describing the risk of bankruptcy are expressed as *low, medium* and *high* values. The current liquidity ratio of the company is expressed as *very low, low, medium*, or *high*. All values of bankruptcy risk *(low, medium, high)* are determined as trapezoidal membership functions:

$Tr[a, b, c, d]$;
$Tr_{low}[0, 0, 0, 2, 0.4]$;
$Tr_{medium}[0.3, 0.5, 0.7, 0.9]$;
$Tr_{high}[0.6, 0.8, 1, 1]$.

The membership functions of the current liquidity indicator are defined in a similar way.

The value of the coefficient $r_{query1}$ is calculated as (17), where $cls$ is a current liquidity indicator and $risk$ is the risk of bankruptcy for Query 1 and (18) for Query 2. The definition of monotonic quantifiers is not change for Query 2.

$$r_{query1} = \frac{\sum_{i=1}^{n}(\mu_{risk}(a_i) \cdot \mu_{cli}(b_i))}{\sum_{i=1}^{n} \mu_{risk}(a_i)} = 0.28 \tag{17}$$

$$r_{query2} = \frac{\sum_{i=1}^{n}(\mu_{risk}(a_i) \cdot \mu_{c}(b_i))}{\sum_{i=1}^{n} \mu_{risk}(a_i)} = 0.34 \tag{18}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

### 6.1 Quantifier $Q$ as a monotonic interval-valued fuzzy set

Let us now consider a set of linguistic variables defined by interval-valued fuzzy sets. For the variable $Q_1$ there are two membership functions, namely $\underline{Q_1}$ and $\overline{Q_1}$, defined by the trapezoidal membership functions. Fig. 1 shows a graphical representation of the membership function of a set of linguistic variables $Q_1$=*very few* and $Q_2$=*few* defined by interval-valued fuzzy sets.

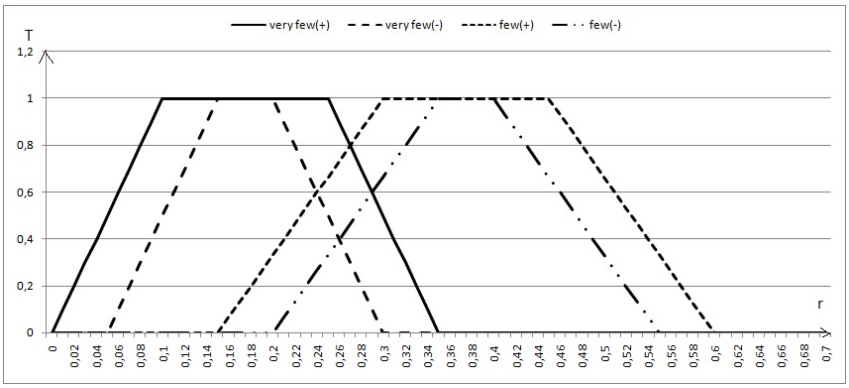


**Fig. 1.** Graphical presentation of the membership function of linguistic variables *very few* and *few* defined by interval-valued fuzzy sets.

The coefficient $r$ is calculated according to equations (17) and (18) for Query 1 and Query 2, respectively. The value of the degree of truth for $Q$ defined as an interval-valued fuzzy set is calculated as (14).

The obtained linguistic summaries have the following values of measure $T$.
Query 1:
*Very few* enterprises with a high current liquidity ratio are in the group with a high risk of bankruptcy $T[0.2; 0.67]$.
*Few* enterprises with a high current liquidity ratio are in the group with a high risk of bankruptcy $T[0.53; 0.86]$.
*Many* enterprises with a high current liquidity ratio are in the group with a high risk of bankruptcy $T[0; 0]$.
*Almost all* enterprises with a high current liquidity ratio are in the group with a high risk of bankruptcy $T[0; 0]$

For the linguistic variable $Q_1$ = *very few* and $Q_2$=*few*, defined in accordance with Def. 2, the sentences are true in the sense of Zadeh's logic. Thus, the outliers have been found. A graphical interpretation of the determined degree of truth for the variable Q defined as an interval-valued fuzzy set is given in Fig.2.

Query 2:
*Very few* enterprises with low profitability are in the high-risk group $T[0.0; 0.1]$.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

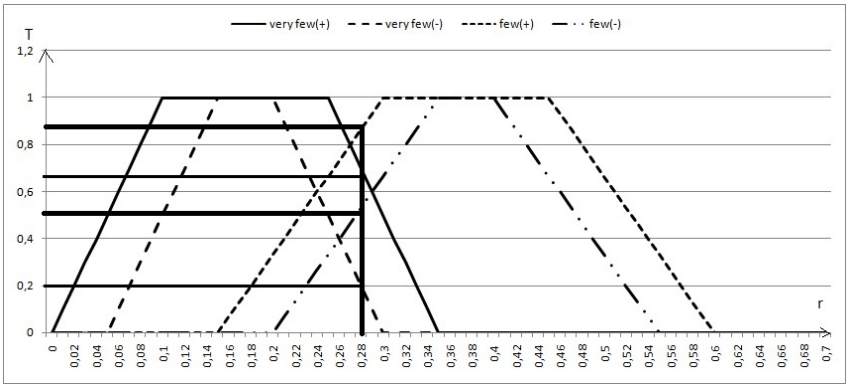


**Fig. 2.** Graphical presentation of the determined degree of truth for $Q_1$=*very few* and $Q_2$=*few*.

*Few* enterprises with low profitability are in the high-risk group $T[0.93; 1]$.
*Many* enterprises with low profitability are in the high-risk group $T[0; 0]$.
*Almost all* enterprises with low profitability are in the high-risk group $T[0; 0]$.
Outliers were found for Query 2 for linguistic variables $Q_1$ and $Q_2$.

### 6.2 Quantifier $Q$ as non-monotonic interval-valued fuzzy set

For quantifiers $Q$ defined as non-monotonic interval-valued fuzzy sets, the relevant membership functions are defined according to (9), (10), (11), (12). Consequently, the linguistic variable $Q_1$= *very few* is defined using four membership functions $\underline{\mu_Q}_{L1}$, $\underline{\mu_Q}_{R1}$, $\overline{\mu_Q}_{L1}$, $\overline{\mu_Q}_{R1}$ and $r = 0.28$. The degree of true $T$ is determined using the equation (16).

The generated linguistic summaries are:
Query 1:
*Very few* enterprises with a high current liquidity ratio are in the group of a high risk of bankruptcy $T[0.2; 0.7]$.
*Few* enterprises with a high current liquidity ratio are in the group of a high risk of bankruptcy $T[0.53; 0.86]$.
*Many* enterprises with a high current liquidity ratio are in the group of a high risk of bankruptcy $T[0; 0]$.
*Almost all* enterprises with a high current liquidity ratio are in the group of a high risk of bankruptcy $T[0; 0]$

Query 2:
*Very few* enterprises with low profitability are in the high-risk group $T[0.0; 0.1]$.
*Few* enterprises with low profitability are in the high-risk group $T[0.95; 1]$.
*Many* enterprises with low profitability are in the high-risk group $T[0; 0]$.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

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

The conducted research and experiments confirm that it is possible to detect outliers using linguistic summaries. In addition, the work verified the functioning of the proposed method for non-monotonic quantifiers. The functioning of the method for monotonic classifiers was also shown in [33, 3, 32, 30]. It was found that the increase of the degree of truth for non-monotonic quantifiers as normal fuzzy sets is of particular importance as compared to the monotonic quantifiers used as normal fuzzy sets.

For the first query and the *very few* quantifier defined by a normal fuzzy set, the degree of truth equal to 0.2 was obtained. In the case where *very few* was defined as a non-monotonic normal fuzzy set, the result was 0.7. There is no doubt that outliers were detected in both cases. It should be emphasized that the increase in $T$ for $Q$ non-monotonic normal fuzzy set results in the validation of our research. We have received a greater degree of truth. For the second query, no outliers were detected for the quantifier $Q$=*very few* defined as a normal fuzzy set. However, after using the non-monotonic quantifier, the value of truth was 0.1. Thus, the degree of truth increased.

The results of the degree of truth for the monotonic and non-monotonic quantifiers *very few* and *few* are given in Table 2. In the case of interval-valued fuzzy sets, the degree of truth obtained for monotonic and non-monotonic quantifiers was similar or the same. An increase in the degree of truth was observed for both Query 1 and Query 2.

**Table 2.** The results of the degree of truth for the monotonic and non-monotonic quantifiers *very few* and *few*.

| Query 1 | monotonic | non-monotonic |
|---|---|---|
| very few (normal fuzzy set) | 0.2 | 0.7 |
| few (normal fuzzy set) | 0.86 | 0.86 |
| very few (interval-valued fuzzy set) | [0.2;0.67] | [0.2;0.7] |
| few (interval-valued fuzzy set) | [0.53;0.86] | [0.55;0.86] |
| Query 2 | | |
| very few (normal fuzzy set) | 0.0 | 0.1 |
| few (normal fuzzy set) | 1 | 1 |
| very few (interval-valued fuzzy set) | [0.0;0.1] | [0.0;0.1] |
| few (interval-valued fuzzy set) | [0.93;1] | [0.95;1] |

## 7 Conclusion

This paper has proposed a novel method for outlier detection. The solution presented provides the user with human-understandable natural language responses, in the form of fuzzy numerical values given as linguistic variables. The response

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

generated by the system concerns a linguistic variable, which constitutes a linguistic specification of the records found, e.g., *about a half, not many, a lot, almost all,* etc.

As demonstrated by practical examples, the application of non-monotonic interval-fuzzy sets, which characterize the least numerous groups of objects (*very few, few*), thus corresponding to the definition of an outlier, improves the reliability of the results, as it leads to an increase in the degree of truth of a linguistic summary. This proves the usefulness of our method for outlier detection.

## References

1. Shareef, D.M.A.M., Aminifar, S.A.: Uncertainty handling in big data using fuzzy logic-literature review. (2021)
2. Ross, T.J., et al.: Fuzzy logic with engineering applications. Volume 2. Wiley Online Library (2004)
3. Duraj, A., Szczepaniak, P.S.: Information outliers and their detection. In: M. Burgin and W. Hofkirchner (Eds.): Information Studies and the Quest for Transdisciplinarity. Volume 9, Chapter 15., World Scientific Publishing Company (2017) 413–437
4. Hawkins, D.M.: Identification of outliers. Volume 11. Springer (1980)
5. Hawkins, S., He, H., Williams, G., Baxter, R.: Outlier detection using replicator neural networks. In: International Conference on Data Warehousing and Knowledge Discovery, Springer (2002) 170–180
6. Barnett, V., Lewis, T.: Outliers in statistical data. Volume 3. Wiley New York (1994)
7. 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)
8. 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. Springer (2011) 789–797
9. Jayakumar, G., Thomas, B.J.: A new procedure of clustering based on multivariate outlier detection. Journal of Data Science **11**(1) (2013) 69–84
10. Yager, R.R.: A new approach to the summarization of data. Information Sciences **28**(1) (1982) 69–86
11. Yager, R.R.: Linguistic summaries as a tool for database discovery. In: FQAS. (1994) 17–22
12. Yager, R.: Linguistic summaries as a tool for databases discovery. Workshop on Fuzzy Databases System and Information Retrieval (1995)
13. Kacprzyk, J., Wilbik, A., Zadrozny, S.: Linguistic summaries of time series via a quantifier based aggregation using the sugeno integral. In: Fuzzy Systems, 2006 IEEE International Conference on, IEEE (2006) 713–719
14. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets and Systems **159**(12) (2008) 1485–1499
15. Kacprzyk, J., Yager, R.R., Zadrozny, S.: Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support. In: Knowledge Discovery for Business Information Systems. Springer (2002) 129–152

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

16. Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. Information Sciences **173**(4) (2005) 281–304
17. Kacprzyk, J., Wilbik, A., Zadrożny, S.: An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. International Journal of Intelligent Systems **25**(5) (2010) 411–439
18. Ng, R.: Outlier detection in personalized medicine. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, ACM (2013) 7–7
19. Aggarwal, C.C.: Toward exploratory test-instance-centered diagnosis in high-dimensional classification. IEEE transactions on knowledge and data engineering **19**(8) (2007) 1001–1015
20. 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. Journal of chemometrics **18**(7-8) (2004) 317–326
21. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. The VLDB Journal—The International Journal on Very Large Data Bases **8**(3-4) (2000) 237–253
22. Angiulli, F., Pizzuti, C.: Fast outlier detection in high dimensional spaces. In: PKDD. Volume 2., Springer (2002) 15–26
23. Giatrakos, N., Kotidis, Y., Deligiannakis, A., Vassalos, V., Theodoridis, Y.: In-network approximate computation of outliers with quality guarantees. Information Systems **38**(8) (2013) 1285–1308
24. Last, M., Kandel, A.: Automated detection of outliers in real-world data. In: Proceedings of the second international conference on intelligent technologies. (2001) 292–301
25. Guo, Q., Wu, K., Li, W.: Fault forecast and diagnosis of steam turbine based on fuzzy rough set theory. In: Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on, IEEE (2007) 501–501
26. Kacprzyk, J., Zadrozny, S.: Protoforms of linguistic database summaries as a human consistent tool for using natural language in data mining. International Journal of Software Science and Computational Intelligence (IJSSCI) **1**(1) (2009) 100–111
27. Kacprzyk, J., Yager, R.R.: Linguistic summaries of data using fuzzy logic. International Journal of General System **30**(2) (2001) 133–154
28. Wilbik, A., Keller, J.M.: A fuzzy measure similarity between sets of linguistic summaries. IEEE Transactions on Fuzzy Systems **21**(1) (2013) 183–189
29. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. red **30**(2) (2008) 3
30. Duraj, A., Niewiadomski, A., Szczepaniak, P.S.: Outlier detection using linguistically quantified statements. International Journal of Intelligent Systems **33**(9) (2018) 1858–1868
31. 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. International Journal of Intelligent Systems **34**(3) (2019) 415–438
32. Duraj, A.: Outlier detection in medical data using linguistic summaries. In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on, IEEE (2017) 385–390
33. Duraj, A., Szczepaniak, P.S., Ochelska-Mierzejewska, J.: Detection of outlier information using linguistic summarization. In: Flexible Query Answering Systems 2015. Springer (2016) 101–113

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38

34. van Benthem, J., Ter Meulen, A.: Handbook of logic and language. Elsevier (1996)
35. Benferhat, S., Dubois, D., Prade, H.: Nonmonotonic reasoning, conditional objects and possibility theory. Artificial Intelligence **92**(1-2) (1997) 259–276
36. Giordano, L., Gliozzi, V., Olivetti, N., Pozzato, G.L.: A non-monotonic description logic for reasoning about typicality. Artificial Intelligence **195** (2013) 165–202
37. Schulz, K., Van Rooij, R.: Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. Linguistics and philosophy **29**(2) (2006) 205–250
38. Zadeh, L.A.: Fuzzy sets. Information and control **8**(3) (1965) 338–353
39. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-iii. Information sciences **9**(1) (1975) 43–80
40. Niewiadomski, A., Duraj, A.: Detecting and recognizing outliers in datasets via linguistic information and type-2 fuzzy logic. International Journal of Fuzzy Systems (2020) 1–12
41. Databases: Statistic Poland https://stat.gov.pl/en/databases/
42. Arora, N., Kaur, P.D.: A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing **86** (2020) 105936
43. Kaur, S.: Comparative analysis of bankruptcy prediction models: An indian perspective. CABELL'S DIRECTORY, USA 19
44. Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K., Suvas, A.: Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. Journal of International Financial Management & Accounting **28**(2) (2017) 131–171
45. Greco, S., Matarazzo, B., Slowinski, R.: A new rough set approach to evaluation of bankruptcy risk. In: Operational tools in the management of financial risks. Springer (1998) 121–136

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_38