Int. J. Fuzzy Syst. (2021) 23(3):878–889
https://doi.org/10.1007/s40815-020-00919-5

Detecting and Recognizing Outliers in Datasets via Linguistic Information and Type-2 Fuzzy Logic

Adam Niewiadomski[1] · **Agnieszka Duraj**[1]

Received: 25 July 2019 / Revised: 1 February 2020 / Accepted: 11 July 2020 / Published online: 22 September 2020

Abstract Uncertainty appearing in datasets (stochastic, linguistic, of measurements, etc.), if not handled properly, may negatively affect information analysis or retrieval procedures. One of possible methods of dealing with uncertain (rare, strange, unexampled) data is to treat them as "outliers" or "exceptions". Among different definitions and algorithms for detecting outliers, we are especially interested in those based on linguistic information represented with type-2 fuzzy logic. We introduce new definitions of outliers in datasets in terms fuzzy properties and linguistically expressed quantities of objects possessing them. Next, new algorithms for detecting outlying objects are presented, to answer whether outliers appear in a dataset or not. Finally, recognition algorithms are presented and exemplified to enumerate particular objects being outliers (e.g., to eliminate them for further considerations). The novelty of this contribution is that we define, detect and recognize outliers using linguistic information represented mostly by type-2 fuzzy sets and logic (if any other information like measures or distances is not accessible), and we supersede this way some earlier approaches based on similar but relatively limited assumptions.

Keywords Outliers in datasets · Detecting outliers · Recognizing outliers · Outliers defined via linguistic information · Type-2 linguistic quantification · Type-2 fuzzy logic

✉ Adam Niewiadomski
adam.niewiadomski@p.lodz.pl

[1] Institute of Information Technology, Lodz University of Technology, ul. Wólczańska 215, 90-924 Łódź, Poland

1 Introduction

Although it sounds like a truism, currently, an intensive development of data analysis methods applied to classification, grouping, machine learning, etc. is noticable. These methods refer to various tasks selected and targeted to purposes of different systems. What must be pointed out here is that in collecting and processing data, frequently from unknown sources, there is some uncertainty, mostly appearing as imprecise and/or incomplete information. Sources of uncertainty are commonly measurements, probability methods (stochastic uncertainty), lack of credibility of information (information uncertainty), and phenomena imprecise descriptions in natural language (linguistic uncertainty). One of manners of handling uncertain data is to look at them as at **outliers**. An "outlier" or "exception" (also anomaly, deviation, abnormality, aberration, etc.) in a natural language means something unique, rare, infrequent, special, specific, sensational, or unexampled. These terms suggest that some features of objects, situations, or phenomena are unobvious or unusual to recipients considering/observing them. Outliers, if occur, are especially possible to be noticed as highlighted or differing on a background of numerous objects or phenomena being similar one to another, typical, statistical, usual, frequent, obvious, normal, plain, common, or ordinary. In data mining and exploration, unrecognized outliers may influence reliability of analysis, cause noise and/or increase data uncertainty. In other words, outliers may blur or even distort the overall idea and/or "gist" of analyzed collections. On the contrary, properly detected and recognized outliers can be interpreted as unique information on intrusions into computer networks, change of activities and congestion in networks, illegal usage of credit cards,

damages of production lines, rapid changes of parameters of medical devices and patients' health status, etc.

Hence, detection and recognition of outliers are an important issue nowadays, in particular, to exclude found abbreviations from further analysis, or on the other hand, to interpret them properly as important, though rare data. The literature contains many different definitions of outliers, mostly intuitive, subjective, and related to different measures how much a given objects are out of analyzed sets. For instance, an outlier by Hawkins, the most often quoted, is [1] *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.* Also in [1], the author defines an outlier as "any object x in the space of consideration $\mathcal{X}$, which has some abnormal, unusual characteristics in comparison to other objects from $\mathcal{X}$". Aggarwal and Yu define outliers as "noise points outside the set which determine clusters", or alternatively, as "points outside the clusters but separated from the noise" [2]. Next, according to Knorr [3, 4], "a point p in a data set is an outlier with respect to parameters k and λ, if no more than k points in the data set are at a distance λ or less from p", where $k \in \mathbb{N}, \lambda \in \mathbb{R}$. The λ parameter used in this definition can be considered in terms of different relations for pairs of objects from $\{x_1, x_2,..., x_N\}$, $N \in \mathbb{N}$, not only as a distance or metric, but also similarity, a semantic connection in the sense of a binary fuzzy relation, a linear or partial ordering relation, etc. Parameter k can be interpreted as a fuzzy number or even a linguistic expression represented by a fuzzy set of any type. Besides, it is worth mentioning another outlier definitions [5–7], and concepts of local and global outliers [8–11]. Regarding applications of outlier detection techniques, interesting approaches combining outliers with clustering algorithms are given in [12, 13]. Outliers detected in linear structures (datastreams) are described in [14]. Another idea for detecting outliers defined via the so-called self-representations is presented in [15]. Detection and recognition of outliers are also considered in [3, 16, 17] and many other. It must be stressed that all the quoted references operate mainly on different and subjectively chosen (so not necessarily coherent one to another) definitions of outliers.

Thus, our inspiration to define outliers in terms of linguistic information represented with fuzzy sets is that literature hardly refers to such descriptions of uncertainty in datasets. In fact, no publications on outliers handled with linguistic terms or quasi-natural language can be found. The need for detecting and recognizing outliers via fuzzy logic appears when traditional quantitative terms or means are inaccessible (compare, e.g., the definition of outlier based on parameters k, λ, by Knorr [3, 4]) and the only information to detect anomalies is human experience and/or knowledge expressed linguistically (which is a very common reason to use fuzzy systems in many fields of data processing, too). In our previous papers, we made attempts to define and detect outliers with linguistic quantifiers and summaries evaluated via fuzzy logic [18, 19], besides in [20], we noticed that some occasional extensions of membership functions to the so-called high-order membership functions, cf. [21], allow us to represent imprecise knowledge and fuzzy queries in a more human-consistent way. Hence, in this paper, we generalize definitions of outliers and detection algorithms in terms of linguistic information represented by general type-2 fuzzy sets (not only interval type-2 fuzzy sets), that means that any secondary membership function, e.g., Gaussian, triangular, not only rectangular, can be used to represent uncertain and/or different linguistic data in detection methods; we also provide adequate arguments in Sect. 3 as the comments to Definition 1, and larger argumentation on using high-order fuzzy sets and their secondary membership functions is given in [21, 22]. Besides, we introduce new algorithms for recognizing the outliers detected with newly defined detection methods. In other words, earlier, we were able to answer "whether outliers do exist or do not in a dataset". Now, we propose a new and handy computational tool to recognize (determine, enlist) particular objects being outliers in a dataset.

The rest of the paper is organized as follows: Sect. 2 is a brief list of basic concepts and operations in type-2 fuzzy logic, besides the calculus of linguistically quantified statements in terms of type-2 fuzzy sets is reminded. Section 3 provides novel definitions of outliers in datasets based on linguistic information (quantified statements) represented by type-2 fuzzy sets, especially, by type-2 fuzzy quantifiers. Revisited algorithms for detection of outliers in datasets, on the base of linguistic information, are given in Sect. 4 and new algorithms for recognition (identification) of detected outliers—in Sect. 5. Two implemented examples of outliers detection and recognition are given to illustrate how the proposed definitions and method work on a real datasets (traffic events and patients suffering from old myocardial infarction) in Sect. 6. Finally, in Sect. 7, we conclude with comments on current and future works in the field discussed.

2 Type-2 Fuzzy Sets, Quantifiers and Linguistically Quantified Statements

This section is a brief review of basics in type-2 fuzzy logic and the calculus of linguistically quantified statements by Zadeh [23] generalized in terms of type-2 by Niewiadomski [22, 24]. A type-2 fuzzy set $\widetilde{A}$ in a finite non-empty $\mathcal{X}$ is denoted $\widetilde{A} = \sum_{x \in \mathcal{X}} \mu_{\widetilde{A}}(x)/x$, and $\mu_{\widetilde{A}}$:

$\mathcal{X} \to \mathcal{FS}([0,1])$ is its type-2 membership function. $\mathcal{FS}([0,1])$ is a set of all traditional fuzzy sets in [0, 1]. A membership degree of x to $\widetilde{A}$ is given by fuzzy set $\mu_{\widetilde{A}}(x) = \{\langle u_{\widetilde{A}}, \mu_x(u_{\widetilde{A}})\rangle : u_{\widetilde{A}} \in J_x\}$, $u_{\widetilde{A}}$ (or u for simplicity) is a primary membership degree and $\mu_x(u_{\widetilde{A}})$—a secondary membership degree of x to $\widetilde{A}$, $J_x \subseteq (0,1]$ is the set of all non-zero primary membership degrees of x, and $\mu_x : J_x \to [0,1]$. The intersection of $\widetilde{A}$, $\widetilde{B}$ in $\mathcal{X}$ is a type-2 fuzzy set in $\mathcal{X}$:

$$\mu_{\widetilde{A}\cap\widetilde{B}}(x) = \sum_{u_{\widetilde{A}}} \sum_{u_{\widetilde{B}}} T_1(\mu_x(u_{\widetilde{A}}), \mu_x(u_{\widetilde{B}}))/T_2(u_{\widetilde{A}}, u_{\widetilde{B}}), \tag{1}$$

where T_1 and T_2 are t-norms. Two of primary membership functions of $\widetilde{A}$ are distinguished: *lower*, $\mathrm{LMF}_{\widetilde{A}}$, and *upper membership function*, $\mathrm{UMF}_{\widetilde{A}}$:

$$\mathrm{LMF}_{\widetilde{A}} = \{\langle x, u\rangle : x \in \mathcal{X}, u = \inf J_x\}, \tag{2}$$

$$\mathrm{UMF}_{\widetilde{A}} = \{\langle x, u\rangle : x \in \mathcal{X}, u = \sup J_x\}. \tag{3}$$

A real-valued cardinality of $\widetilde{A}$ is defined [22, 25]

$$\mathrm{nf}\sigma\text{−count}(\widetilde{A}) =_{df} \sum_{x\in\mathcal{X}} \sup\{u \in J_x : \mu_x(u) = 1\}, \tag{4}$$

assuming $\sup\emptyset = 0$ if $u \in J_{x'} : \mu_{x'}(u) = 1$ does not exist for x'. In a particular case of interval or trapezoidal type-2 fuzzy set, its cardinality can be defined [26]

$$\mathrm{card}_I(\widetilde{A}) =_{\mathrm{df}} \frac{1}{2}\sum_{x\in\mathcal{X}} \left(\mathrm{LMF}_{\widetilde{A}}(x) + \mathrm{UMF}_{\widetilde{A}}(x)\right), \tag{5}$$

to take into account all $u : \mu_x(u) = 1$, not only the largest. A *relative cardinality* of a $\widetilde{A}$ with respect to a type-2 fuzzy set $\widetilde{B}$ is given:

$$\mathrm{card}(\widetilde{A}|\widetilde{B}) =_{\mathrm{df}} \frac{\mathrm{card}(\widetilde{A}\cap\widetilde{B})}{\mathrm{card}(\widetilde{B})}, \tag{6}$$

in which (4), (5), or another real-valued cardinality of $\widetilde{A}$ is taken. Other definitions of relative cardinality are given, e.g., in [27]. $\widetilde{A}$ is *normal* iff [22, 28]

$$\exists_{x\in\mathcal{X}}\ \sup\{u : u \in J_x\} = 1 \wedge \sup\{\mu_x(u) : u \in J_x\} = 1, \tag{7}$$

and this definition applies to type-2 fuzzy sets in an infinite $\mathcal{X}$, too. The set of all non-zero secondary membership degrees in $\widetilde{A}$ is denoted $\mathrm{SMD}(\widetilde{A}, \mathcal{X}) = \{r \in (0,1] : \exists_{x\in\mathcal{X}} \wedge \exists_{u\in J_x} \mu_x(u) = r\}$. $\widetilde{A}$ in $\mathbb{R}$ is *convex* iff $\forall_{c\in\mathrm{SMD}(\widetilde{A},\mathbb{R})}$ embedded type-1 fuzzy sets in $\widetilde{A}$ with membership functions:

$$\mu_{c,\min}(x) = \min\{u \in J_x : \mu_x(u) = c\}, \tag{8}$$

$$\mu_{c,\max}(x) = \max\{u \in J_x : \mu_x(u) = c\}, \tag{9}$$

are convex in $\mathbb{R}$ (i.e., each of their α-cuts is convex in the classic meaning in $\mathbb{R}$). The support of $\widetilde{A}$ in $\mathcal{X}$ is a traditional fuzzy set in $\mathcal{X}$ [28, 29]:

$$\mu_{\mathrm{supp}(\widetilde{A})}(x) = \sup_{u_{\widetilde{A}}\in J_x} \mu_x(u_{\widetilde{A}}). \tag{10}$$

Assume $\widetilde{S}_1$ and $\widetilde{S}_2$ are linguistic expressions (predicates) represented by type-2 fuzzy sets in a finite $\mathcal{X}$, describing some properties or characteristics possessed by objects $x \in \mathcal{X}$. Let $\widetilde{Q}$ be a relative linguistic quantifier that describes quantities of objects in relation to a cardinality of a larger set (superset, universe of discourse, etc.), e.g., *many of*, *very few (of)*, *almost all*, *about 2/3*, *less than 1/4*, so $\widetilde{Q}$ is represented by a normal and convex type-2 fuzzy set in [0, 1]. Now, the forms of linguistically quantified statements in the sense of Zadeh [23], but generalized with the use of type-2 fuzzy sets as $\widetilde{Q}$, $\widetilde{S}_1$, $\widetilde{S}_2$ [22], are presented:

$$\widetilde{Q}x\text{'s are } \widetilde{S}_1, \tag{11}$$

$$\widetilde{Q}x\text{'s being } \widetilde{S}_2 \text{ are } \widetilde{S}_1, \tag{12}$$

denoted also as Q^I, Q^{II}, and known as *the first form* and *the second form* of a linguistically quantified statement, respectively. Since they are sentences of type-2 fuzzy logic, their degrees of truth need to be evaluated to assess the represented information:

$$T\left(\widetilde{Q}x\text{'s are } \widetilde{S}_1\right) = \mu_{\widetilde{Q}}\left(\mathrm{card}(\widetilde{S}_1|\mathcal{X})\right), \tag{13}$$

$$T\left(\widetilde{Q}x\text{'s being } \widetilde{S}_2 \text{ are } \widetilde{S}_1\right) = \mu_{\widetilde{Q}}\left(\mathrm{card}(\widetilde{S}_1|\widetilde{S}_2)\right). \tag{14}$$

It must be noticed that the degrees of truth are values of a type-2 membership function, so they are fuzzy sets in [0, 1], or their special cases, e.g., real numbers if $\widetilde{Q}$ is represented by a type-1 fuzzy set, intervals $[\underline{T}, \overline{T}]$ for interval type-2 fuzzy sets, etc. Moreover, real cardinalities of $\widetilde{S}_1$, $\widetilde{S}_2$ in (13), (14) may be evaluated with (4), (5), or other real-valued cardinalities of type-2 fuzzy sets.

Besides, quality of the generated linguistically quantified statements can be assessed with a quality measure of type-2 fuzzy quantifier, presented in [22, 24]. It is based on fuzzy support (10) of a type-2 fuzzy set $\widetilde{Q}$ representing the quantifier:

$$\widetilde{T}_{\mathrm{supp}}(\widetilde{Q}) = 1 - |\mathrm{supp}(\widetilde{Q})|. \tag{15}$$

$\widetilde{T}_{\mathrm{supp}}$ depends on chosen characteristics of $\widetilde{Q}$ and its meaning is the closer to 1, the more informative (more

precise, more specific) quantifier $\widetilde{Q}$.

Obviously, choosing representations for quantifier $\widetilde{Q}$ and predicates $\widetilde{S}_1$, $\widetilde{S}_2$ (so their membership functions, cardinalities, etc.) depends on character of analyzed data, especially linguistic information, accessible knowledge, and other presumes for modeling expressions and statements. In the next sections, the presented linguistically quantified statements are crucial for defining formally "an outlier", and for detecting and recognizing outliers in datasets.

3 Outliers in Terms of Linguistic Information

In the context of this study, definitions of outliers are introduced in relation to linguistic information represented by type-2 fuzzy sets and linguistically quantified statements with their degrees of truth evaluated via (13) and (14).

Definition 1 (*Outlier in terms of linguistically expressed quantities*) Let $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, $N \in \mathbb{N}$, be a finite non-empty set of objects, and S—a linguistic expression (predicate, feature, property) describing by objects in $\mathcal{X}$. Let Q be a regular non-increasing relative linguistic quantifier (like "few", "several", "almost none" or synonymous). Let $\mathcal{X}_{\text{out}} \subseteq \mathcal{X}$ contain only and all x's being/having S. An $x \in \mathcal{X}_{\text{out}}$ is *outlier* iff the relative cardinality of $\mathcal{X}_{\text{out}}$ can be intuitively expressed with Q.

The regular relative linguistic quantifier Q is in this context represented by a fuzzy set in $\{0, \frac{1}{N}, \frac{2}{N}, ..., \frac{N-1}{N}, 1\}$ with the monotonically non-increasing μ_Q:

$$\mu_Q(0) = 1, \quad \mu_Q(1) = 0, \tag{16}$$

$$\forall_{x_1, x_2 \in [0,1]} x_1 \leq x_2 \longrightarrow \mu_Q(x_1) \geq \mu_Q(x_2). \tag{17}$$

Generalization of properties (16), (17) for linguistic quantifiers represented by interval-valued fuzzy sets (now known as interval type-2 fuzzy sets), and properties of normality and convexity have been given previously in [24, 30]. Besides, having defined normal and convex type-2 fuzzy sets (7–9), an adequate ordering relation for fuzzy membership degrees $\mu_{\widetilde{Q}}(x_1)$, $\mu_{\widetilde{Q}}(x_1)$ in (17) would be required to reconsider it in terms of type-2 fuzzy sets.

Unfortunately, Definition 1 does not provide any method for assessing how much the statement *x IS OUTLIER* (or *x IS NOT OUTLIER*) is true or reliable. This drawback was partially met with the definition of an outlier in terms of a linguistic summary of a dataset $\mathcal{X}$ (in which outliers are detected), in the sense of Yager, cf. [19, 20]. In this paper, we supersede that definition with new Definitions 2 and 3 based on the first form and the second forms of linguistically quantified statements (11), (12), respectively, while the older version was based on a linguistic summary, so related rather to an algorithm of aggregating datasets than to determining truth degrees of logical sentences in type-2 fuzzy logic.

Definition 2 (*Outlier in terms of the first form of linguistic quantified statement*) Let $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, $N \in \mathbb{N}$, be a finite non-empty set of objects. Let $\widetilde{S}$ be a linguistic expression describing objects in $\mathcal{X}$ and represented by a type-2 fuzzy set in $\mathcal{X}$. Let $\widetilde{Q}$ be a regular non-increasing relative linguistic quantifier (like "few", "several", "almost none" or synonymous) represented by a type-2 fuzzy set, and $\alpha \in [0, 1]$. An $x \in \mathcal{X}$ being $\widetilde{S}$ is *outlier* iff

$$T(\widetilde{Q}x\text{'s are } \widetilde{S}) >^* \alpha. \tag{18}$$

The degree of truth of (18) can be evaluated, using (13), as

$$T(\widetilde{Q}x\text{'s are } \widetilde{S}) = \mu_{\widetilde{Q}}\left(\frac{\text{nf}\sigma\text{−count}(\widetilde{S})}{N}\right), \tag{19}$$

and this means that only x's such that $\exists u_{\widetilde{S}} \in J_x : \mu_x(u_{\widetilde{S}}) = 1$ are taken into account as outliers. Using other forms of real cardinalities in (19) is also possible, e.g., (5). Details on evaluating interval membership degrees for interval type-2 fuzzy sets are given in [24, 30] and applied to detecting outliers in [20]. Also, the ordering relation $>^*$ in (18) depends on type of fuzzy sets used to represent $\widetilde{Q}$ and/or $\widetilde{S}$. For instance, in terms of interval type-2 fuzzy sets, T is an interval $[\underline{T}, \overline{T}] \subseteq [0, 1]$ and its comparison to a real α is defined as

$$T >^* \alpha \text{ iff } \alpha \leq \underline{T}, \tag{20}$$

$$T \geq^* \alpha \text{ iff } \underline{T} \leq \alpha \leq \overline{T}. \tag{21}$$

If $\widetilde{S}$, $\widetilde{Q}$ are traditional fuzzy sets, the relation $>^*$ in is the linear order in $\mathbb{R}$, so (18) takes the form of

$$\mu_Q\left(\frac{\sum_{i=1}^{N} \mu_{\widetilde{S}}(x_i)}{N}\right) > \alpha. \tag{22}$$

Now the definition of outliers in terms of (12), i.e., based on two properties $\widetilde{S}_1$, $\widetilde{S}_2$, is introduced:

Definition 3 (*Outlier in terms of the second form of linguistically quantified statement*) Let $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, $N \in \mathbb{N}$, be a finite non-empty set of objects and $\widetilde{S}_1$, $\widetilde{S}_2$—linguistic expressions describing objects $x \in \mathcal{X}$ and represented by type-2 fuzzy sets in $\mathcal{X}$. Let $\alpha \in [0, 1]$, and $\widetilde{Q}$—a

regular non-increasing relative linguistic quantifier as in Definition 2. An $x \in \mathcal{X}$ being $\widetilde{S}_1$ and $\widetilde{S}_2$ is *outlier* iff

$$T(\widetilde{Q}x'\text{s being } \widetilde{\text{S}}_2 \text{ are } \widetilde{\text{S}}_1) >^* \alpha. \tag{23}$$

The degree of truth in (23) is evaluated via (14) as

$$T\big(\widetilde{Q}x'\text{s being } \widetilde{S}_2 \text{ are } \widetilde{S}_1\big) = \mu_{\widetilde{Q}}\left(\frac{\text{nf}\sigma\text{−count}(\widetilde{S}_1 \cap \widetilde{S}_2)}{\text{nf}\sigma\text{−count}(\widetilde{S}_2)}\right). \tag{24}$$

andss also (5) or another formula may express cardinalities of type-2 fuzzy sets in (24), compare (19). Hence, outliers are these x's for which

$$\exists u_{\widetilde{S}_1}, u_{\widetilde{S}_2} \in J_x : \mu_x(u_{\widetilde{S}_1}) = 1 \wedge \mu_x(u_{\widetilde{S}_2}) = 1, \tag{25}$$

and relation $>^*$ in (23) is interpreted analogously to (18), see e.g., (20–21), or (22).

In the next section, we introduce two novel algorithms for detecting outliers in datasets, based on the presented definitions and measures.

4 Detecting Outliers with Type-2 Linguistically Quantified Statements

In [18–20], the authors proposed algorithms for detecting outliers using linguistic summaries based on computing degrees of truth of linguistic summaries on analyzed datasets. Efficiency of those algorithms was checked for mixed data (text and numbers). Now, we intend to present generalized versions of those algorithms using newly presented definitions of outliers (Definitions 2 and 3).

There are two algorithms for outliers detection presented. Both are designed to detect exceptional data or aberrations in dataset when only imprecise and linguistically expressed knowledge on them is accessible. In particular, objects x in an analyzed dataset $\mathcal{X}$ are considered to be outliers, if they can be intuitively described by expressions like "small", "big", "hot", "very expensive", etc., represented by type-2 fuzzy sets $\widetilde{S}_1$, $\widetilde{S}_2$ (as given in Sect. 3), and their small quantity is either not determined with any real number (or precise value), but expressed linguistically with "very few", "almost none", etc., represented by $\widetilde{Q}$. Hence, both algorithms will confirm that outliers are found in a dataset, iff statements "$\widetilde{Q}$ x's are $\widetilde{S}_1$", and "$\widetilde{Q}$ x's being $\widetilde{S}_2$ are $\widetilde{S}_1$" are of sufficiently large (close to 1) degree of truth. The assumptions for both Algorithms 1 and 2 are

1. $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, $N \in \mathbb{N}$—a finite non-empty dataset,
2. $\widetilde{S}_1, \widetilde{S}_2$—linguistic expressions for properties (features) of objects $x \in \mathcal{X}$, represented by type-2 fuzzy sets in $\mathcal{X}$,
3. $\{\widetilde{Q}_1, \widetilde{Q}_2, \ldots, \widetilde{Q}_K\}$, $K \in \mathbb{N}$—a set of regular monotonically non-increasing relative linguistic quantifiers, as in Definition 2, represented by type 2 fuzzy sets in [0, 1],
4. $\alpha \in [0, 1]$—an arbitrarily chosen minimal value of degree of truth for (18), (23) to detect outliers.

Algorithm 1 for detecting outliers is related to Definition 2;

$$\text{Algorithm1} : \mathcal{X} \times \mathcal{T}2\mathcal{FS}(\mathcal{X}) \times \mathcal{T}2\mathcal{FS}([0,1])^K \times [0,1] \rightarrow \{\text{true}, \text{false}\}, \tag{26}$$

where $\mathcal{T}2\mathcal{FS}(\cdot)$ is a set of all type-2 fuzzy sets in a given space, $K \in \mathbb{N}$, and $\{\text{true}, \text{false}\}$—the binary set of logical values interpreted as: true = "THERE ARE OUTLIERS IN $\mathcal{X}$", false = "NO OUTLIERS IN $\mathcal{X}$". To start Algorithm 1, an entry query is needed

$$\text{How many x's are } \widetilde{\text{S}}_1\text{?} \tag{27}$$

that determines the property S_1 with respect to which outliers are to be detected using linguistic quantifiers $\widetilde{Q}_1, \widetilde{Q}_2, \ldots, \widetilde{Q}_K$.

Algorithm 1 Detecting outliers via the first form of linguistically quantified statement

for all $k = 1, 2, \ldots, K$ **do**
  $T_k \leftarrow 0$, $r \leftarrow 0$
  for all $n = 1, 2, \ldots, N$ **do**
    $r \leftarrow r + \mu^*_{\widetilde{S}_1}(x_n)$
  $T_k \leftarrow \mu_{\widetilde{Q}_k}(r/N)$
if not $T_1 >^* \alpha$ **and not** $T_2 >^* \alpha$ **and** ... **and not** $T_K >^* \alpha$ **then return** „NO OUTLIERS IN $\mathcal{X}$"
return „THERE EXIST OUTLIERS IN $\mathcal{X}$"

Comments to Algorithm 1:

1. The side effects of Algorithm 1, important in the algorithms for recognizing outliers (if detected) in $\mathcal{X}$, see Sect. 5, are K linguistic expressions:

$$\begin{aligned} &\widetilde{Q}_1 x'\text{s are } \widetilde{S}_1[T_1], \\ &\quad \ldots \\ &\widetilde{Q}_K x'\text{s are } \widetilde{S}_1[T_K], \end{aligned} \tag{28}$$

   and their degrees of truth $T_1, T_2, \ldots, T_K$ evaluated in Step 5.
2. It is possible to detect outliers in Step 5 (i.e., before the algorithm stops), if only $T_k > {}^*\alpha$ for any $k = 1,2,\ldots, K$. Nevertheless, the presented version of the algorithm provides information that is total and complete, that means it takes into account all assumed linguistic quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2, \ldots, \widetilde{Q}_K$.
3. Symbol $\mu^*_{\widetilde{S}_1}(x_n)$ in Step 4 denotes a real value representing the membership degree of x_n to $\widetilde{S}_1$, and not a membership degree itself, e.g., fuzzy or interval, to provide the representation of linguistic quantifier Q_k, $k = 1, 2, \ldots, K$, with a real $r/N \in [0, 1]$ in Step 5.

Algorithm 1 is now enhanced to Algorithm 2 with respect to Definition 3. That means that detecting outliers in $\mathcal{X}$ is now based on two, possibly overlapping, predicates $\widetilde{S}_1$, $\widetilde{S}_2$, according to the second form of linguistically quantified statements (11). Algorithm 2 is a function

$$\begin{aligned} &\text{Algorithm 2} \ : \\ &\mathcal{X} \times \mathcal{T}2\mathcal{FS}(\mathcal{X}) \times \mathcal{T}2\mathcal{FS}(\mathcal{X}) \times \mathcal{T}2\mathcal{FS}([0,1])^K \times [0,1] \\ &\rightarrow \{\text{true}, \text{false}\} \end{aligned} \tag{29}$$

with symbols analogous to (26). Algorithm 2 needs an entry query in the form of

$$\text{How many x}'\text{s being } \widetilde{\text{S}}_2 \text{are} \widetilde{\text{S}}_1? \tag{30}$$

pointing at properties $\widetilde{S}_1$, $\widetilde{S}_2$ necessary to detect outliers in the sense of Definition 3.

Comments to Algorithm 2:

1. As in Algorithm 1, the side effects of Algorithm 2 are statements and their degrees of truth

$$\begin{aligned} &\widetilde{Q}_1 x'\text{s being } \widetilde{S}_2 \text{ are } \widetilde{S}_1[T_1], \\ &\quad \ldots \\ &\widetilde{Q}_K x'\text{s being } \widetilde{S}_2 \text{ are } \widetilde{S}_1[T_K]. \end{aligned} \tag{31}$$

2. It is possible to answer that outliers exist in $\mathcal{X}$, in Step 5 (i.e., before the algorithm stops), compare comment 2 to Algorithm 1.
3. Symbols $\mu^*_{\widetilde{S}_2}(x_n)$, $\mu^*_{\widetilde{S}_1 \cap \widetilde{S}_2}(x_n)$ in Steps 4, 5—as in comment 3 to Algorithm 1.

The presented detecting algorithms are able to answer whether outliers exist or not in dataset $\mathcal{X}$, but they are cannot show which objects are outliers in $\mathcal{X}$. Thus, we introduce two another algorithms for recognizing x's in $\mathcal{X}$ possessing properties $\widetilde{S}_1$, $\widetilde{S}_2$ determining them to be outlying (exceptional), or in other words, for showing explicitly which particular objects are exceptions in $\mathcal{X}$.

5 Recognizing Outlying Objects in Datasets

The outlier detection algorithms (Algorithm 1 and 2) presented in Sect. 4 and corresponding with definitions Definitions 2 and 3, respectively, provide only the binary information that some outliers did appear in the analyzed set $\mathcal{X}$ (*true*) or did not appear (*false*). However, subsets of outliers $\mathcal{X}_{\text{out}} \subseteq \mathcal{X}$ remain unspecified as results of the algorithms. Therefore, we now deal with an algorithm that accomplishes the task: recognition of particularly these objects in $\mathcal{X}$ that are outliers with respect to given properties $\widetilde{S}_1$, $\widetilde{S}_2$, or in other words, thanks to algorithms presented in this section, subset of outliers $\mathcal{X}_{\text{out}} \subseteq \mathcal{X}$ with respect to properties $\widetilde{S}_1$ and $\widetilde{S}_2$ is determined.

At first, we consider the concept of an outlier via Definition 2, and the assumptions and denotations are as for

Algorithm 2 Detecting outliers via the second form of linguistically quantified statement

1: **for all** $k = 1, 2, \ldots, K$ **do**
2: $T_k \leftarrow 0$, $rn \leftarrow 0$, $rd \leftarrow 0$
3: **for all** $n = 1, 2, \ldots, N$ **do**
4: $rn \leftarrow rn + \mu^*_{\widetilde{S}_2 \cap \widetilde{S}_1}(x_n)$
5: $rd \leftarrow rd + \mu^*_{\widetilde{S}_2}(x_n)$
6: $T_k \leftarrow \mu_{\widetilde{Q_k}}(rn/rd)$
7: **if not** $T_1 >^* \alpha$ **and not** $T_2 >^* \alpha$ **and** ... **and not** $T_K >^* \alpha$ **then return** „NO OUTLIERS IN $\mathcal{X}$"
8: **return** „THERE EXIST OUTLIERS IN $\mathcal{X}$"

Algorithm 1. Hence, for given dataset $\mathcal{X} = \{x_1, x_2,..., x_N\}$, $N \in \mathbb{N}$, properties $\widetilde{S}_1$, $\widetilde{S}_2$, linguistic quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$,..., $\widetilde{Q}_K$, $K \in \mathbb{N}$, and parameter $\alpha \in [0, 1]$, Algorithm 3 is a function

$$\text{Algorithm3} : \mathcal{X} \times \mathcal{T}2\mathcal{FS}(\mathcal{X}) \times \mathcal{T}2\mathcal{FS}([0,1])^K \times [0,1] \to \mathcal{X}_{\text{out}} \tag{32}$$

where $\mathcal{X}_{\text{out}}$ is a set of outliers in $\mathcal{X}$. As the consequence of using Definition 2 and Algorithm 1, recognizing a subset of outliers $\mathcal{X}_{\text{out}}$ in $\mathcal{X}$ is based on the query (27).

Algorithm 3 Recognizing outliers detected with Algorithm 1

1: declare $\mathcal{X}out = \emptyset$
2: **for all** $n = 1, 2, \ldots, N$ **do**
3: **if** $\mu^*_{\widetilde{S}_1}(x_n) >^* \alpha$ **then** $\mathcal{X}out \leftarrow \mathcal{X}out \cup \{x_n\}$
4: **return** $\mathcal{X}out$

With symbols $\mu^*_{\widetilde{S}_1}(x_n)$—as in comment 3 to Algorithm 1, relation $>^*$—as explained in Definition 2. Of course, Algorithm 3 is fired if only Algorithm 1 detected existing outliers in $\mathcal{X}$ before, because otherwise there is no point to determine objects being outliers for $\mathcal{X}_{\text{out}} = \emptyset$. The result of Algorithm 3 is the set of outliers in $\mathcal{X}$ selected with property $\widetilde{S}_1$ and query (27).

Analogously, subsets of outliers $\mathcal{X}_{\text{out}} \subset \mathcal{X}$ can be detected according to Definition 3, with assumptions for Algorithm 2. The Algorithm 4 for recognizing outliers in $\mathcal{X}$ on the base of properties $\widetilde{S}_1$, $\widetilde{S}_2$, is a function:

$$\text{Algorithm4} : \mathcal{X} \times \mathcal{T}2\mathcal{FS}(\mathcal{X}) \times \mathcal{T}2\mathcal{FS}(\mathcal{X}) \times \mathcal{T}2\mathcal{FS}([0,1])^K \times [0,1] \to \mathcal{X}_{\text{out}} \tag{33}$$

and the entry query to recognize outliers is in the form of (30). Symbols $\mu^*_{\widetilde{S}_1 \cap \widetilde{S}_2}(x_n)$, and $>^*$ analogous to those used in Algorithm 3. The result, i.e., the set of $x \in \mathcal{X}$ recognized as outliers, is evaluated and then returned as an array of indices of several objects $x_n \in \mathcal{X}$, $n \in \{1, 2,..., N\}$, selected from $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$.

6 Application Examples

We illustrate now how the algorithms for detection and recognition of outliers, introduced in Sects. 3 and 4, work on real datasets. We present two examples of detecting and recognizing outliers with linguistic information represented by type-2 fuzzy sets: outliers in a database on traffic events (Example 1) and outliers in a group of patients with old myocardial infarction, I25.2 via International Classification of Diseases, ICD10 [31] (Example 2). Example 1 combines type-1 and type-2 fuzzy sets to represent information on outlying situations on roads. Example 2 shows how detection of outliers may help to identify patients with rare symptoms of ischemia caused by myocardial infarctions; besides, it presents how using type-2 fuzzy sets may improve detection and recognition of outliers.

6.1 Example 1: Detecting and Recognizing Outliers in a Set on Traffic Events

Here, we concentrate on outliers appearing in the data on traffic events, a fragment of which is given in Table 1. The fields taken into account, apart form ID—the main key of the table, are driver's age in years (Age), time of the event (Time), and visibility in the sense of "air transparency" (Visibility). The "visibility" attribute takes text values (description of air transparency in the moment of event) and is taken into account as an argument for a membership function representing similarity to a value given by query (27) or (30) (details of evaluating that similarity are given in [20]).

Algorithm 4 Recognizing outliers detected with Algorithm 2

1: $i \leftarrow 0$
2: declare $outlierIndices[N]$
3: **for all** $n = 1, 2, \ldots, N$ **do**
4: **if** $\mu^*_{\widetilde{S}_1 \cap \widetilde{S}_2}(x_n) >^* \alpha$ **then** $\{i \leftarrow i + 1;\ outlierIndices[i] \leftarrow n\}$
5: **if** $i = 0$ **then return** „NO OUTLIERS IN $\mathcal{X}$"
6: **return** $outliersIndices$

The test is performed in two separated variants, I and II, with different representations for quantifiers and properties in linguistically quantified statements. The variants are to illustrate detection and recognition of outliers, and also possibilities of using different types of fuzzy sets in representations of fuzzy expressions in Definitions 2 and 3 and corresponding algorithms, so partially illustrate the profits of generalizing detection and recognition methods to type-2 fuzzy sets, as we declare in the introduction. The database analyzed in both variants contains nearly 2000 records, and some of them are shown in Table 1. The relative linguistic quantifiers chosen to fire Algorithm 2, acc. to Definition 3, are $\widetilde{Q}_1$ = "almost none", $\widetilde{Q}_2$ = "few", $\widetilde{Q}_3$ = "many"; however, only $\widetilde{Q}_1$ fulfills Definition 3, i.e., it is regular monotonically non-increasing; $\widetilde{Q}_2$, $\widetilde{Q}_3$ are shown only to point at the contrast between detected set of outliers and other, more numerous subsets of common objects not being outliers in $\mathcal{X}$. The membership functions for quantifiers are (34–36) (for $r \in [0,1]$), shown in Fig. 1.

$$\mu_{\text{almost none}}(r) = \begin{cases} 1 & r < 0.06 \\ \dfrac{0.18 - r}{0.12} & 0.06 \le r < 0.18 \end{cases} \tag{34}$$

$$\mu_{\text{few}}(r) = \begin{cases} \dfrac{r - 0.03}{0.12} & 0.03 \le r < 0.15 \\ 1 & 0.15 \le r < 0.25 \\ \dfrac{0.45 - r}{0.2} & 0.25 \le r < 0.45 \end{cases} \tag{35}$$

$$\mu_{\text{many}}(r) = \begin{cases} \dfrac{r - 0.3}{0.2} & 0.3 \le r < 0.5 \\ 1 & 0.5 \le r < 0.6 \\ \dfrac{0.8 - r}{0.2} & 0.6 \le r < 0.8 \end{cases} \tag{36}$$

The quantifiers can also be represented by interval-type-2 fuzzy sets, and their lower and upper membership functions for "almost none", "few" and "many" are given by (37–42), respectively, see Fig. 2.

Table 1 Sample records of the dataset analysed in the experiment

ID	Age	Time	Visibility for driver
0001	18	8:00	Visibility—quite good
0002	21	4:30	Bad visibility
...	...	...	...
0254	29	5:00	Visibility rather good
0255	24	7:30	Good visibility
...	...	...	...
1483	18	9:15	Visibility very good
...	...	...	...
1876	23	4:50	Visibility rather good

$$\text{LMF}_{\text{almost none}}(r) = \begin{cases} 1 & r < 0.08 \\ \dfrac{0.23 - r}{0.15} & 0.08 \le r \le 0.23 \end{cases} \tag{37}$$

$$\text{UMF}_{\text{almost none}}(r) = \begin{cases} 1 & r < 0.1 \\ \dfrac{0.25 - r}{0.15} & 0.1 \le r \le 0.25 \end{cases} \tag{38}$$

$$\text{UMF}_{\text{few}}(r) = \begin{cases} \dfrac{r - 0.1}{0.1} & 0.1 \le r < 0.2 \\ 1 & 0.2 \le r \le 0.36 \\ \dfrac{0.48 - r}{0.12} & 0.36 \le r \le 0.48 \end{cases} \tag{39}$$

$$\text{UMF}_{\text{few}}(r) = \begin{cases} \dfrac{r - 0.08}{0.1} & 0.08 \le r < 0.18 \\ 1 & 0.18 \le r \le 0.38 \\ \dfrac{0.5 - r}{0.12} & 0.38 \le r \le 0.5 \end{cases} \tag{40}$$

$$\text{LMF}_{\text{many}}(r) = \begin{cases} \dfrac{r - 0.3}{0.2} & 0.3 \le r < 0.5 \\ 1 & 0.5 \le r \le 0.6 \\ \dfrac{0.8 - r}{0.2} & 0.6 \le r \le 0.8 \end{cases} \tag{41}$$

$$\text{UMF}_{\text{many}}(r) = \begin{cases} \dfrac{r - 0.25}{0.2} & 0.25 \le r < 0.45 \\ 1 & 0.45 \le r \le 0.65 \\ \dfrac{0.85 - r}{0.2} & 0.65 \le r \le 0.85 \end{cases} \tag{42}$$

Interval-valued fuzzy quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$ and other are illustrated in Fig. 2.

The query in the form of (27) is

$$\text{How many } (\widetilde{Q}) \text{ events caused by young drivers } (\widetilde{S}_2) \text{ took place in good visibility } (\widetilde{S}_1)? \tag{43}$$

For property $\widetilde{S}_1$ = "good visibility", the membership function is evaluated using the so-called n-grams (the measure is defined [24] and computational details shown in [20]). Property $\widetilde{S}_2$ = "young (driver)" is represented by real-valued membership function (44)

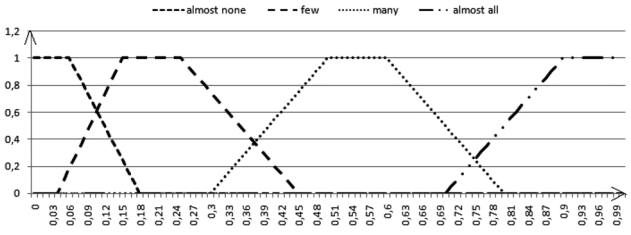

Fig. 1 Traditional fuzzy sets representing linguistic quantifiers "almost none" in Variant I

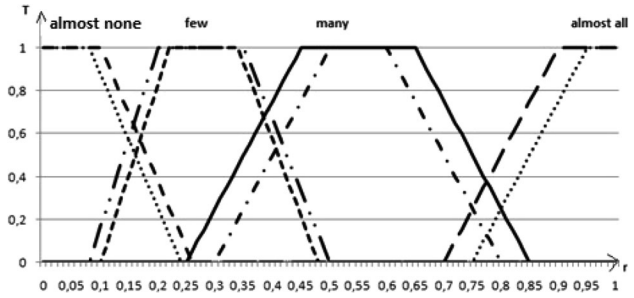

Fig. 2 Interval type-2 fuzzy sets for quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$, $\widetilde{Q}_3$

$$\mu_{\text{youngdrivers}}(x) = \begin{cases} \dfrac{x-16}{3} & 16 \le x < 19 \\ 1 & 19 \le x < 22 \\ \dfrac{26-x}{4} & 22 \le x < 26 \end{cases}, \tag{44}$$

or by interval type-2 fuzzy set (45–46)

$$\text{LMF}_{\text{youngdrivers}}(x) = \begin{cases} \dfrac{x-18}{3} & 18 < x \le 22 \\ 1 & 22 < x \le 23 \\ \dfrac{26-x}{3} & 23 < x \le 26 \end{cases}, \tag{45}$$

$$\text{UMF}_{\text{youngdrivers}}(x) = \begin{cases} \dfrac{x-16}{4} & 16 < x \le 20 \\ 1 & 20 < x \le 26 \\ \dfrac{28-x}{2} & 26 < x \le 28 \end{cases}. \tag{46}$$

Two variants of the test are presented, both using Definition 3 of outliers (so in the meaning of the second form of linguistically quantified statement), Algorithm 2 for detection, and Algorithm 4 for recognition.

Variant I In variant I, the membership function (44) is chosen to represent $\widetilde{S}_2$ and linguistic quantifiers are (37–42). According to (24), $r = 0,14$ is evaluated as the argument for the $\widetilde{Q}_1, \widetilde{Q}_2, \widetilde{Q}_3$ membership functions. Hence, the following linguistically quantified statements are obtained:

"Almost none" of events caused by young drivers took place in good visibility [0.62, 0.75],

"Few" of events caused by young drivers took place in good visibility [0.33, 0.50],

"Many" of events caused by young drivers took place in good visibility [0.00, 0.00].

Thus, for query (43) and fixed value of $\alpha = 0.6$, outliers are detected in the database for $\widetilde{Q}_1$ = "almost none".

The result of recognition performed via Algorithm 4; the following objects are recognized as elements of the subset of outliers $X_{\text{out1}} \subset \mathcal{X}$:

$$\mathcal{X}_{\text{out1}} = \{0001, 0047, 0430, 1493, 1780, 1803\}. \tag{47}$$

The graphical interpretation of evaluating interval degree of truth for $\widetilde{Q}_1, \widetilde{Q}_2$ is shown in Fig. 3.

Variant II In Variant II, quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$, $\widetilde{Q}_3$ are given by membership functions (34–36), and property $\widetilde{S}_2$—by (45–46). Hence, again one can notice the capability of the introduced detection and recognition methods for mutual use of different types of fuzzy sets, here interval type-2 and traditional. $\widetilde{S}_2$ = "good visibility" is as in Variant I. For query (43) and $\alpha = 0.70$, via Algorithm 2, $r = [0.11, 0.14]$ is obtained. Obviously, to be an argument of a linguistic quantifier membership function, its mean is taken as 0.125. Hence, the following statements are obtained via (24):

"Almost none" of events caused by young drivers took place in good visibility [0.75, 0.93],

"Few" of events caused by young drivers took place in good visibility [0.25, 0.50],

"Many" of events caused by young drivers took place in good visibility [0.00, 0.00].

For $\widetilde{Q}_1$ = "almost none" and the ordering relation (20–21), the occurrence of outliers in $\mathcal{X}$ is confirmed. Moreover, we evaluate the chosen representation of $\widetilde{Q}_1$ with quality measure (15) as $\widetilde{T}_{\text{supp}}(\widetilde{Q}_1) = 1 - |[0, 0.15]| = 0.85$, for the $\text{supp}(\widetilde{Q}) = [0, 0.15]$, see Fig. 1, and that means that $\widetilde{Q}_1$ is fairly precise to determine the occurrence of outliers in $\mathcal{X}$ in terms of Definitions 2 or 3.

The recognition process is performed via Algorithm 4 and the subset of outliers $\mathcal{X}_{\text{out2}} \subset \mathcal{X}$ is

$$\mathcal{X}_{\text{out2}} = \{0001, 0027, 0047, 0430, 0845, 1088, 1493, 1780, 1803\}. \tag{48}$$

Notice that though the same query (43) is being handled in Variants I and II, the subsets of recognized outliers are different, $\mathcal{X}_{\text{out1}} \neq \mathcal{X}_{\text{out2}}$; this is because of different, type-1 or type-2, representations of quantifiers $\widetilde{Q}_1, \widetilde{Q}_2$ and properties $\widetilde{S}_1, \widetilde{S}_2$. The explanation is as follows: thanks to more flexible (i.e., type-2) representation of property $\widetilde{S}_2$ more objects in a data are recognized as outliers.

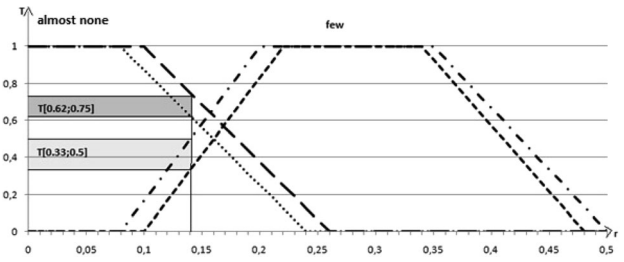

Fig. 3 Graphic interpretation of the calculation of the degree of truth for $r = 0.14$ and for the interval-valued fuzzy quantifier $\widetilde{Q}_1$, $\widetilde{Q}_2$

6.2 Example 2: Recognizing Outlying Patients with Chronic Ischemic Heart Disease

In this example, we show how detection of outliers may help to identify patients with rare symptoms of ischemia caused by old myocardial infarctions, I25.2 via ICD10 [31]. We operate on the database of anonymous patients suffering from ischemic chest symptoms being result of myocardial infarction in the past. One of the symptoms is the so-called Central Venous Pressure (CVP) expressed in medical terminology as integers, and normal values vary between 4 and 12 cm H_2O [32]. Besides, the patients are classified to one of the four so-called Forrester classes (FC) (or Forrester hemodynamic subsets)[1] [33]. The database contains about 1700 records and its fragment is given by Table 2.

From the point of view of statistics, patients suffering from old myocardial infarction are characterized by normal or high values of CVP, i.e., ≥ 10. However, few patients with this disease may not complete this criterion, or even be characterized by very low CVP values, e.g., ≤ 6, and the aim of the presented analysis is to recognize those patients as outliers, in order to provide them a proper treatment.

As an entry to Algorithm 2, the following statement is given via (30):

$$\text{How many } (\widetilde{Q}) \text{ middle-aged } (\widetilde{S}_3) \text{ patients have low CVP } (\widetilde{S}_4)? \tag{49}$$

The properties $\widetilde{S}_3$ and $\widetilde{S}_4$ are represented with (50) and (51), respectively:

$$\mu_{\text{middle-aged}}(x) = \begin{cases} \dfrac{x-32}{8} & 32 < x \leq 40 \\ \dfrac{48-x}{8} & 40 < x \leq 48 \\ 0 & \text{otherwise} \end{cases} \tag{50}$$

$$\mu_{\text{lowCVP}}(x) = \begin{cases} 1 & x \leq 4 \\ \dfrac{6-x}{2} & 4 < x \leq 6 \\ 0 & x > 6 \end{cases} \tag{51}$$

Variant III The linguistic quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$ and $\widetilde{Q}_3$ are represented by traditional fuzzy sets in this variant, and their membership functions are given by (34–36), respectively. For query (49) and $\alpha = 0.4$ we evaluate $r = 0.11$ via (24), and the following statements are obtained:

"Almost none" middle-aged patients have low CVP [0.4].

"Few" middle-aged patients have low CVP [0.42].

"Many" middle-aged patients have low CVP [0.0].

Hence, according to Algorithm 2, the occurrence of outliers in the set is detected. Recognition via Algorithm 4 identifies the following subset of outliers $\mathcal{X}_{\text{out3}}$:

$$\mathcal{X}_{\text{out3}} = \{0057, 0164, 0523, 0763, 1322\}. \tag{52}$$

Variant IV In this variant, linguistic quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$, $\widetilde{Q}_3$ are represented by type-2 fuzzy sets (37–42), $\alpha = 0.4$, and $r = 0.11$. We obtain the following statements via (24):

"Almost none" middle-aged patients have low CVP [0.0, 0.0].

"Few" middle-aged patients have low CVP [0.42, 1].

"Many" middle-aged patients have low CVP [0.0, 0.0].

Thus, outliers are detected (via Algorithm 2). Algorithm 4 is run to recognize which particular objects belong to the subset of outliers $\mathcal{X}_{\text{out4}}$:

$$\mathcal{X}_{\text{out4}} = \{0057, 0164, 0385, 0523, 0763, 1322, 1522\}. \tag{53}$$

As a conclusion, an important observation must be stressed; here, the recognized subsets of outliers differ, $\mathcal{X}_{\text{out3}} \neq \mathcal{X}_{\text{out4}}$. Similarly to the conclusion of Example 1 in Sect. 6.1, this is because quantifiers $\widetilde{Q}_1$, $\widetilde{Q}_2$, $\widetilde{Q}_3$ in Variant IV are represented by type-2 fuzzy sets, and in Variant III—by traditional fuzzy sets. The result was confirmed by an expert (medicine doctor) and hence we may claim some majority of type-2 representations.

7 Conclusions

In this paper, we introduce new definitions of outliers and algorithms for detecting them in datasets, when only linguistic imprecise information is given to differ them from typical, common, or regular data, which means that any other information on objects, situations, or phenomena, are not accessible (and quantitative terms, e.g., in definitions of outliers by Aggarwal, Knorr, etc. cannot be applied). Besides, we show representations of linguistic information with type-2 fuzzy sets, that, generally speaking, cover most of types of fuzzy sets, like traditional, interval-valued, and in a broader sense, even intuitionistic fuzzy sets, *L*-fuzzy sets, etc. (so we claim this approach supersedes our initial ideas based on type-2 fuzzy sets only, cf. [18, 19]). That is why the presented methods can be considered as sufficiently universal and flexible, e.g., to join within one method different types of fuzzy sets representing various pieces of information (see application examples in Sect. 6). Besides, novel algorithms of recognizing outliers detected in datasets are the contribution worth noticing, since, till now, no linguistic information have been applied in the

[1] The Forrester classes are not taken into account in this particular example; however, we quote them to provide the coherence of description of the analyzed data.

Table 2 Sample records identifying patients with old myocardial infarction in Example 2

ID	Age	···	Forrester class (FC)	CVP
0001	48	···	II	8
0002	35	···	III	14
···	···	···	···	···
0354	41	···	III	16
···	···	···	···	···
1259	42	···	III	11
1260	43	···	III	14
···	···	···	···	···
1690	47	···	III	12

field, or initially, dealt mostly with detection only. Now, the possibility of answering the question "which objects are outliers in $\mathcal{X}$?" but not only "are there outliers in $\mathcal{X}$ or not?" is the profit provided by recognition algorithms proposed in this article.

As results of Algorithms 1 or 2, outliers are detected, and the subsets of outliers $\mathcal{X}_{\text{out}}$ in the analyzed $\mathcal{X}$ are recognized by Algorithms 3 or 3 based on the degrees of truth of linguistically quantified statements generated as side effects of detection, see (28) and (31). Besides, it should be underlined that the presented approach to the problem of outliers in datasets is based on linguistically quantified statements interpreted in terms of type-2 fuzzy sets, and they were not used before in detecting and recognizing atypical data or anomalies, though the issue is initially (based on rectangular secondary membership only) addressed in [18, 20].

Currently, research on detecting and recognizing outliers using multi-subject forms of linguistically quantified statements, cf. [34], and analyzing non-relational (non-sequential) datasets, mostly graph databases, cf. [35], are in progress.

References

1. Hawkins, D.M.: Identification of outliers, vol. 11. Springer, Berlin (1980)
2. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: ACM sigmod record, vol. 30, pp. 37–46, ACM (2001)
3. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. VLDB J. **8**(3–4), 237–253 (2000)
4. Knox, E.M., Ng, R.T.: Algorithms for mining distancebased outliers in large datasets. In: Proceedings of the international conference on very large data bases, pp. 392–403, Citeseer (1998)
5. Aggarwal, C.C.: Outlier analysis. Springer, Berlin (2013)
6. Barnett, V., Lewis, T.: Outliers in statistical data, vol. 3. Wiley, New York (1994)
7. Knorr, E.M., Ng, R.T.: A unified notion of outliers: properties and computation. In: KDD, pp. 219–222 (1997)
8. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM sigmod record, vol. 29, pp. 93–104, ACM (2000)
9. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. KDD **96**, 226–231 (1996)
10. Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1649–1652, ACM (2009)
11. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: ACM SIGMOD record, vol. 29, pp. 427–438, ACM (2000)
12. Jiang, F., Liu, G., Du, J., Sui, Y.: Initialization of k-modes clustering using outlier detection techniques. Inf. Sci. **332**, 167–183 (2016)
13. Flanagan, K., Fallon, E., Connolly, P., Awad, A.: Network anomaly detection in time series using distance based outlier detection with cluster density analysis. In: 2017 internet technologies and applications (ITA), pp. 116–121, IEEE (2017)
14. Tran, L., Fan, L., Shahabi, C.: Distance-based outlier detection in data streams. VLDB Endow **9**, 1089–1100 (2016)
15. You, C., Robinson, D.P., Vidal, R.: Provable self-representation based outlier detection in a union of subspaces. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp. 3395–3404 (2017)
16. Aggarwal, C.C.: Outlier detection in categorical, text, and mixed attribute data. In: Outlier analysis, pp. 249–272, Cham, Springer (2017)
17. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif Intell Rev **22**(2), 85–126 (2004)
18. Duraj, A.: Outlier detection in medical data using linguistic summaries. In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE international conference on, pp. 385–390, IEEE (2017)
19. Duraj, A., Niewiadomski, A., Szczepaniak, P.S.: Outlier detection using linguistically quantified statements. Int. J. Intell. Syst. **33**(9), 1858–1868 (2018)
20. 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)
21. Mendel, J.M.: Uncertain rule-based fuzzy logic systems: introduction and new directions. Prentice-Hall, Upper Saddle River (2001)
22. Niewiadomski, A.: A type-2 fuzzy approach to linguistic summarization of data. IEEE Trans. Fuzzy Syst. **16**(1), 198–212 (2008)

23. Zadeh, L.A.: A computational approach to fuzzy quantifiers in natural languages. Comput Maths Appl **9**, 149–184 (1983)
24. Niewiadomski, A.: Methods for the linguistic summarization of data: applications of fuzzy sets and their extensions. Academic Publishing House EXIT, London (2008)
25. Jang, L.-C., Ralescu, D.: Cardinality concept for type-two fuzzy sets. Fuzzy Sets Syst. **118**, 479–487 (2001)
26. Wu, D., Mendel, J.M.: A vector similarity measure for interval type-2 fuzzy sets. In: Proceedings of FUZZ-IEEE 2007 international conference, 23–26 July. England, London (2007)
27. Wu, D., Mendel, J.M.: Uncertainty measures for interval type-2 fuzzy sets. Inf. Sci. **177**, 5378–5393 (2007)
28. Niewiadomski, A.: On finity, countability, cardinalities, and cylindric extensions of type-2 fuzzy sets in linguistic summarization of databases. IEEE Trans. Fuzzy Syst. **18**(3), 532–545 (2010)
29. Niewiadomski, A.: Interval-valued and interval type-2 fuzzy sets: a subjective comparison. In: Proceedings of FUZZ-IEEE'07, 23–26.07.2007, London, pp. 1198–1203 (2007)
30. Niewiadomski, A., Ochelska, J., Szczepaniak, P.S.: Interval-valued linguistic summaries of databases. Control Cybern. **35**(2), 415–444 (2006)
31. International classification of diseases, the web's free 2019/2020 ICD-10-cm/pcs medical coding reference. https://www.icd10data.com/. Accessed 27 Jan 2020
32. Goers, T. A.: W. U. S. of Medicine Department of Surgery. In: Klingensmith, M.E., Li, E.C., Glasgow, S.C. (eds.) The Washington manual of surgery (2008)
33. Forrester, J.S., Diamond, G., Chatterjee, K., Swan, H.J.: Medical therapy of acute myocardial infarction by application of hemodynamic subsets. N. Engl. J. Med. **295**(24), 1356–1362 and 1404–1413 (1976)
34. Niewiadomski, A., Superson, I.: Multi-subject type-2 linguistic summaries of relational databases. In: Sadeghian, A., Tahayori, H. (eds.) Frontiers of higher order fuzzy sets, pp. 167–181. Springer, Berlin (2015)
35. Strobin, Ł., Niewiadomski, A.: Linguistic summaries of graph datasets using ontologies: an application to semantic web. J. Intell. Fuzzy Syst. **32**(2), 1193–1202 (2017)

Adam Niewiadomski is an Associate Professor in the Institute of Information Technology in Lodz University of Technology, Lodz, Poland. He published over 100 books and papers on artificial intelligence, soft computing and fuzzy logic. He was the supervisor of 5 PhD thesis.

Agnieszka Duraj is an Assistant Professor in the Institute of Information Technology in Lodz University of Technology, Lodz, Poland. She published about 50 papers on artificial intelligence and outlier search and recognition.