Outlier Detection in Medical Data Using Linguistic Summaries

Agnieszka Duraj
Lodz University of Technology
Institute of Information Technology
Lodz, Poland
Email: agnieszka.duraj@p.lodz.pl

***Abstract*—The main purpose of outlier detection algorithms is to find a new feature that is distinct from the other features of the vector in the analyzed data set. This paper concerns outlier detection in medical databases, and the supervised and unsupervised methods used in detection of outliers in medical data are discussed. Moreover, the author's original method for detecting outliers based on linguistic summaries is presented.**

***Index Terms*—Outlier detection, linguistic summaries.**

I. Introduction

Medical systems deal with many types of data, processing large collections of information, including both patients' personal information and the results of various laboratory tests. Taking into account the development of information technology, the supporting of healthcare management with automated patient history records seems natural. Moreover, it is possible to support the physician in making decisions, discovering new symptoms and new diagnostic features occurring in small groups of patients. No wonder that for several years there has been a revival of research in this field, see, for example [1], [2], [3], [4].

The task of outlier detection is to find such patterns in data vectors that are incompatible with the expected characteristics. Outlier detection research involves studies in a wide range of fields and offers a broad spectrum of possible applications, such as monitoring the activities of enemies, detecting traps or unidentified objects, detecting fraud on credit cards and bank transfers, fraudulent transactions, insurance fraud, or detecting hacker attacks. Each group of the above-mentioned applications deals with a different data type, see, e.g. [1], [4]. Initially, outliers were treated as noise, or incorrect data. They usually resulted from an error of a measurement device or from a human error. The entire record in which the abnormalities were detected was removed, or the attribute which was missing was given an averaged value. In such a case, outlier detection is carried out as a data preprocessing step, e.g. data preparation and cleaning performed before classification or grouping. The second group of outliers are those that possess different characteristics and thus are distinct from the rest of the data. They are considered as anomalies in the considered population. The formulation of the definition of an outlier depends heavily on the area of research and the type of data under examination. An outlier in medical applications may have various meanings depending on the specific area of study. In studies dealing with medical data, outlier detection is applied both as the preprocessing stage aiming to identify noise and errors or as the process of anomaly detection. The present research concerns data analysis in terms of medical diagnosis, with a particular focus on outlier detection.

978-1-5090-5795-5/17/$31.00

In this study, denition of an outlier in linguistic summaries is introduced and a new procedure aiming to detect outliers in a data set based on the concept of linguistic summaries is presented. Linguistic summaries are in the form introduced by Yager [37], [38], [39], [49] and the result is obviously in the form of a natural language sentences. The analysis can be conducted for both numeric and textual types of data. These responses are important for example Such approach is advantageous in decision making problems, when we are confronted with information that is incomplete, and this kind of situation occurs in medical systems. Rughly speaking, the set of linguistic variables replaces the numeric value of the detected outliers with a linguistic formulation e.g. "few", "little", "almost none" [5], [6]. The application of linguistic summaries to the analysis of a medical database provides the ability to determine whether that database contains outliers or not.

The remainder of the paper is structured as follows. Section II gives an overview of the literature, with a particular emphasis on the works concerning outlier detection in medical databases. Section III presents the basic definitions of an outlier proposed in related works. Next, in Section IV, selected outlier detection methods applied to numeric data are briefly discussed. In Section V, the application of linguistic summaries to the detection of abnormal data in medical records is proposed. The definition of an outlier based on the concept of linguistic summary is then presented. Section VI provides examples of practical application of the method proposed by the author. Section VII presents final conclusions.

II. Related Works

Outlier detection has been the subject of many publications, e.g. [4], [7], [8] and for a short review see [9], [10] Outlier detection using statistical approaches was presented by Barnett and Lewis [11], Rousseeuw and Leroy [12], Hawkinks [8], Markov's models - Xu and Cheng and Bayesian inference [13],

Otey [14]. In use are also algorithms based on distance, density and clusters such as distance-based-outlier (DB-outlier) [15], [16], the local outlier factor (LOF) [17], Connectivity-based Outlier Factor (COF) [18] and Density Based Spatial Clustering of Applications with Noise (DBSCAN) [19] and many others [20], [21], [22]. In [23], the use of data mining techniques to identify unusual objects in complex data sets, such as rule-based knowledge bases is proposed. Hierarchical clustering algorithms based on similarity analysis enable the identification of rare (outlier type) rules, which in turn can accelerate the process of efficient knowledge analysis in a given domain.

Outlier detection is an important part of medical data investigation and various approaches are applied, mainly in relation to numeric data. Issues related to noise and outlier detection in medical data have been addressed by Gamberger [24]. Sirguson [25] applied outlier detection to skin cancer diagnostics. Ortega [26] described an effective medical claim on fraud/abuse of detection system. Santhanam and Padmavanthi in [27] compared the data reduction percentage performed by K-Means and Statistical Outliers. Bouarfa and Dankelman [28] used workflow mining to derive consensus workflow from multiple surgical activity logs using tree-guided multiple sequence alignment and detected outliers using global pairwise sequence alignment (NeedlemanWunsch) algorithm. Alba et al. [29] proposed an algorithm for the segmentation of highly abnormal hearts. See also the works by Laurikkala [30], Roberts [31], Kumar [1], Nielsen [32], Heimann [33], Hauskrecht [3], Shaari [34], Fritsch [35], Duraj [36] and many others.

On the other hand, linguistic summaries are helpful in decision making. Linguistic summarization was defined by Yager [37], [38], [39] and used fuzzy quantifiers which were presented by Zadeh [40], [41]. New forms of linguistic summaries were proposed in many works by Kacprzyk, Zadrozny and Wilbik, e.g. [42], [43], [44], [45], [46], and some new measures were introduced by Niewiadmski [47].

Application of linguistic summaries to outlier detection in databases containing textual and numeric attributes was reported by Duraj, et.al. in [5], [6].

III. Definition of Outlier

As stated in the introduction, patterns are treated as outliers in the following cases::

- if they do not meet the specified and defined normal standards;
- if the properties of the object (one or all) differ from the characteristics of the correct pattern.

In the case of two-dimensional data outliers will be those objects that are distant from the other two groups. For example, in Fig. 1 two groups of patients are indicated. It is easy to notice the two indicated groups and three objects that are outliers. An example of three designated groups of patients, the healthy ones and the ones with influenza (squares represent the patients with influenza, triangles the healthy patients, and wheels the outliers) is shown in Fig.1.

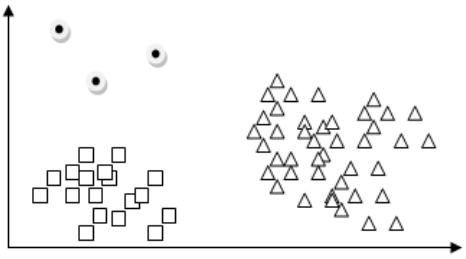

Fig. 1. An illustration of detection of outliers in a group of patients. Indication: round - outliers, squares-healthy patients, triangles - the ill patients.

In the given example, persons classified into the "patients infected with influenza" group reported high fever. The three objects, although they shared common attributes with the group of "patients infected with influenza", differed in one attribute the temperature. They were not classified into the "healthy patient group" because differed significantly in other attributes, except for temperature.

The algorithms used for detecting outliers in medical data must have a very high accuracy and sensitivity. A slight deviation in values may already be regarded as an outlier. Moreover, a slight deviation in attribute value in the case of medical data may be an indication of a new object. This may be a new variant of the virus, a new form of cancer cells, etc. In other fields, such as stock exchange, minor deviations can be considered as a normal situation and will not be classified as outliers.

The definition of an outlier for medical systems will depend on the type of data under examination. Most of the works are based on the definition of Barnett and Lewis.

- According to the definition of Barnett and Lewis [11], an outlier is "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data".

Another definition that is referred to in this paper is the one proposed by Hawkins.

- As Hawkins [8] states, "for any object $x \in X$, if x has some abnormal characteristics as compared to the other objects in X, we may consider x as an outlier".

It should also be emphasized that the very definition of an outlier is very important, because it translates into the detection of new anomalies in diseases, new symptoms and new forms or cases of diseases. There are also different extreme definitions of outliers, closely related to the specific area of medical research.

Although the definitions of an outlier in scientific research are based largely on those by Barnett or Hawkins, there are also definitions which differ in single words from the indicated definitions. Changing a single word is very important in light of the conducted research and analysis of medical databases. See for example works by Laurikkala et al. [30].

IV. Selected methods of outlier detection

As far as numerical records are concerned, the methods applied to medical data analysis are no different from the standard methods of data mining. Similarly, the outlier detection methods applied to medical databases are no different to those applied in other fields. The situation is different when the research involves non-numeric data, such as image or clinical signal data. These include the results of medical examinations, such as mammography, ultrasound, CT (Computer Tomography), ECG (Electrocardiography), EEG (Electroencefalography), magnetic resonance, EMG (ElectroMyoGraphy), EEA (ElectroOculoGraphy), SC / GSR (skin conductance - Galvanic Skin Response) - electrical conductivity of the skin, BVP (Blood Volume Pulse) - measurement of blood flow, SCP (slow cortical potentials), etc. Another challenging area in the field of detection and analysis of outliers in medical data is the interpretation of medical research results expressed in a natural language. The current methods proposed for this purpose suffer from a lack of standardization (especially in the description of examinations in a natural language). A major difficulty is also the lack of possibility to record a specific medical case in the mathematical form.

A. Statistical methods

The first group of methods used to detect outliers are statistical methods, including all types of regression. In many studies, regression has been considered as the first step - the so-called preprocessing of data.

While detecting outliers using linear regression, the high leverage point observations are of exceptional importance. They assume a very high or low value of the predictor variable (independent variable) x in comparison to the dependent variable y. In addition, the influential observations are also important, since they can lead to a change in slope of the regression line. It is assumed that there are equations (1) and (2) where h_i is the impact of the $i-th$ observation determined as the difference between the actual value of the variable x_i and the predicted value $\overline{x}$, and p is the number of influential observations and dependent variables. Additionally, it is specified that h_i is influential if for the $i-th$ observation the value h_i is above the value of $2p/n$.

$$\sum_{i=1}^{n} h_i = p \tag{1}$$

$$1 \geq h_i \geq 1/n \tag{2}$$

Additionally, it is specified that h_i is influential if for the $i-th$ observation the value h_i is above the value of $2p/n$.

B. Supervised methods

The functioning of these methods is based on the knowledge of the data set and, above all, the division into classes. It is known how many and what classes there are in the analyzed data set. Physicians determine precisely which patients and with which ailments became ill with which disease. However, in outlier detection unusual cases are of interst. Frequently, two types of classes are created in the analyzed data set: the class of diseases and the class of outliers. Unusual objects, as their very name suggests, occur very rarely. Hence, the problem of inequality of classes arises, which will be discussed briefly in the next section.

This group of supervised methods includes popular probabilistic methods, such as Bayesian models, e.g. see [48]. Probabilistic inference based on the Bayes theorem is widely applied in the classification of medical data. The process of discovering outliers is, in this case, composed of three steps, the first two of which relate to the typical process of data classification. These steps are:

- Building a classification model, or, in other words, classification function, classifier.
- The classification of new objects created on the basis of the classifier, which determines the membership of a new object in one of the many defined classes based on the determined probability.

The third step is directly related to the specifying of outliers. It is known that the vector describing the new event will be classified as belonging to class B if the posterior probability $P(A/B)$ is the highest. However, there may be extreme values, for which the degree of membership will be considerably low, different from the other probabilities. The observations detected in this way, after an additional analysis performed by an expert, can be classified as outliers. Another way is to introduce the limit value. Then the objects (observations) with the degree of membership probability lower than the limit value are automatically considered as outliers.

The big popularity of Bayesian methods in the classification and detection of outliers in medical data is associated with the possibility to use Bayesian classifiers for different types of data, including numeric and categorical data.

C. Unsupervised methods

Unsupervised learning methods do not require the involvement of a human expert in the classification of objects. First of all, they analyze the whole set and extract repeatable characteristics of the objects in order to obtain homogenous groups. This may also lead to discovering new, unknown characteristics or data structures. The choice of method depends on the area of application and the type of data under examination. Unsupervised learning methods, also called patternless classification, are very popular in the detection of outliers in medical databases, because they lead to the identification of objects with different characteristics, or groups of characteristics. They can also indicate several groups of outliers. This group of methods includes the k-means algorithm, hierarchical data clustering algorithms, and many more.

Among the unsupervised learning methods that are applied effectively to medical data there are also the algorithms which are closely related to outlier detection. These include: Local Outlier Factor (LOF), Connectivity Outlier Factor (COF), Density Based Spatial Clustering (DBSCAN).

Local Outlier Factor (LOF) introduces the so-called outlier factor. Each object is assigned a degree of uniqueness, taking

into account local neighborhood of the object. If the value of LOF is close to 1, this object belongs to the group. If the value of LOF changes abruptly in respect to the local neighbor, the object is identified as a local outlier.

Connectivity Outlier Factor (COF) takes into account the distance between objects. In contrast to LOF, COF takes into account also the density of objects in the set. It divides the data into two sets: a set of typical objects and a set of outliers of far smaller density. The following situations are possible:

- detection of a cluster of low density and a small number of subjects;
- detection of an isolated cluster, which is often of low density and at a large distance from the other objects;

Low density of the outlier is due to the deviation of the object (samples) of a cluster of high density, and isolation of the outlier is due to the deviation of the object from the samples located within its vicinity.

Density Based Spatial Clustering of Applications with Noise (DBSCAN) belongs to the group of unsupervised learning methods based on the concept of the density of the set of objects. The classification of objects is performed on the basis of the maximum radius of neighborhood of objects and the minimum number of objects in the neighborhood.

More information on these methods and their application to medical data can be found in the works of Duraj and Szczepaniak [6].

V. Detection of Outliers Using Linguistic Summaries

The denition of the linguistic summary, as introduced by Yager in [49], [39], is given in Def.1. It is the basic starting point for the innovative outlier detection method using linguistic summaries described by Duraj and Szczepaniak [5], [6].

Def. 1: Linguistic summary
A linguistic summary is an ordered four in the form of $< Q, P, S, T >$, where:
$Q-$quantity in agreement, linguistic quantifier,
$P-$subject of summary,
$S-$summarizer,
$T-$degree of truth.

This method allows searching the database under a determined criterion. The answer is given in a natural language, which is very important from the point of view of decision support systems. Database search can be carried out for each type of data (numeric, text). The authors introduced a new definition of an outlier in linguistic summaries.(Def.2).

Def. 2: An outlier via a linguistically quantified statement. Let $X = \{x_1, x_2, ..., x_N\}$ for $N \in \mathbb{N}$ be a finite, non-empty set of objects. Let S be a finite, non-empty set of attributes (features) of the set of objects X. $S = \{s_1, s_2, ..., s_n\}$ Let Q be a fuzzy relative quantifier monotonically decreasing in the following space $\{0, \frac{1}{N}, \frac{2}{N}, ..., \frac{N}{N}\}$.
A collection of objects – the subjects of a linguistic summary are called outliers if Q objects having S is a true statement in the sense of fuzzy logic.
If the linguistic summary of Q objects in the P are/have S, $[T]$ has $T > 0$. Therefore it is true in the sense of fuzzy logic.

The essence of the method lies in proper defining of the set of linguistic values $X = \{Q_1, Q_2, ..., Q_n\}$. Since the outlier is defined as a small collection, the value of Q_1 should contain synonyms of the words "little, almost none, very little, hardly anyone and so on in order to detect outliers.

The next steps of the procedure are the same as in the case of the linguistic summary and can be described as follows:

1. Specify the set of linguistic variable $X = \{Q_1, ..., Q_n\}$
2. Enter the question: How many P being R are S?
3. Determine the membership functions for the S, R
4. Calculate the value of r according to (3)
5. Determine T as (4).

$$r = \frac{\sum_{i=1}^{n}(\mu_R(x) * \mu_S(x))}{\sum_{i=1}^{n} \mu_R(x)} \tag{3}$$

$$T = \mu_Q(r) \tag{4}$$

If the degree of truth T is bigger than 0 in the generated linguistic sentences for the linguistic variable Q_1, Q_2, the statement is true in the sense of Zadeh's fuzzy logic. We can also determine the existence of outliers. If for Q_1, Q_2, the value of T is 0, then this sentence is not a true sentence, and, therefore, exceptions are not found.

VI. Practical Use of The Method

The problem of outlier detection in medical databases was investigated by the author in several earlier works. The paper by Duraj [36] discussed the detection of outliers using the K-means algorithm and DBSCAN. The performance of the method was illustrated by examples from three medical databases. In [6] the results of outlier detection using linear regression, as well as COF and DBSCAN algorithms were presented.

It should be emphasized that the method based on linguistic summaries does not specify the number of detected outliers in a numerical form, but the advantage is that the information is obtained in the form of natural language sentences. For example:
Few patients have a pacemaker $T[0.25]$.
Almost no patient in middle age belongs to the group at high risk of ischemic chest $T[0.47]$.
Assuming that the degree of truth is greater than zero, it is the quantifiers "little", "almost none" that indicate that there are outliers in the analyzed data set. Additionally, it should be noted that the proposed method can be used for both numeric and text data.

Consider, therefore, a medical data set consisting of 895 records with measurements of diagnostic cases of patients with diabetes. In addition to age, body weight, and BMI, there are

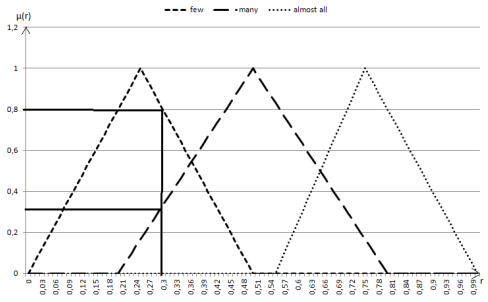

Fig. 2. Graphical interpretation of outliers.

also attributes associated with diastolic blood pressure, blood glucose, pregnancy history and so on.

According to the procedure of detection of outliers in linguistic summaries, the set of linguistic values is defined. Let us define a set of values of linguistic variables as (5).

$$Q = \{Q_1, Q_2, Q_3\} = \{\textit{"few"}, \textit{"many"}, \textit{"almost all"}\} \quad (5)$$

For definitions see, respectively: Q_1="few" – (6), Q_2="many" – (7), Q_3="almost all" –(8).

$$\mu_{\textit{few}}(r) = \begin{cases} \frac{r-0.25}{0.25} & 0 \leq r < 0.25 \\ \frac{0.5-r}{0.25} & 0.25 \leq r < 0.50 \\ 0 & r \geq 0.5 \end{cases} \quad (6)$$

$$\mu_{\textit{many}}(r) = \begin{cases} 0 & r \leq 0.2, r \geq 0.8 \\ \frac{r-0.2}{0.3} & 0.2 \leq r < 0.5 \\ \frac{0.8-r}{0.3} & 0.5 \leq r < 0.8 \end{cases} \quad (7)$$

$$\mu_{\textit{almost all}}(r) = \begin{cases} 0 & r \leq 0.55 \\ \frac{r-0.75}{0.2} & 0.55 \leq r < 0.75 \\ \frac{1-r}{0.25} & 0.75 \leq x < 1 \end{cases} \quad (8)$$

Let us analyze the detection of outliers for the following query:
How many young pregnant women have a high level of glucose in the blood?
Let us define the membership function for young age in the form of (9).

$$\mu_{\textit{young}}(x) = \begin{cases} 0 & x \leq 18, x > 32 \\ \frac{x-18}{8} & 18 < x \leq 26 \\ \frac{32-x}{8} & 26 < x \leq 34 \end{cases} \quad (9)$$

We proceed to determine the degree of truth of the generated linguistic summary. According to formula (3) we determine the value of r (r is 0.29). Similarly, we verify r for every linguistic variable Q_i. The graphical interpretation of the degree of truth for $r = 0.29$ is shown in Fig.2.

The following linguistic summary is obtained:
Few young pregnant women have a high level of glucose in the blood T [0.80].
Many young pregnant women have a high level of glucose in the blood T [0.30].
Almost all young pregnant women have a high level of glucose in the blood T [0.00].
All of the generated statements are true in Zadeh' sense, because the degree of truth is greater than zero. According to Def.2, only the quantifier *"few"* is taken into account for the analysis of the existence of outliers. For this quantifier, the sentence
Few young pregnant women have a high level of glucose in the blood T [0.80]
is a true statement, which proves the presence of outliers in the analyzed data set. Outlier detection in textual data is performed in a similar way, see e.g. [5], [6].

VII. Conclusion

This study has provided examples of outlier detection in medical data. The proposed method for the detection of outliers using linguistic summaries was verified for numerical medical data, although the definition of the outlier in linguistic summaries used in this work may also be applied to textual data. In medicine, most of the patient's records, including diagnosis and treatment, are formulated in medicine specic but natural language. The standard practice is also to describe and to perform analysis of single cases. Abnormal (outlier), single or rare cases are of particular interest in this context. For this reasons, it is reasonable to formulate summaries of medical records in natural language. Further works will include modication of the described method taking into account other types of classiers, e.g. monotonic non-decreasing quantiers. The outlier detection algorithm presented in this paper can be used in medical expert systems and medical decision-making systems.

Acknowledgment

This work was supported by a grant of the Dean of the Faculty of Technical Physics, Information Technology and Applied Mathematics, Lodz University of Technology, within the Program No. I-1/501/17-1-1/716.
I would like to thank Piotr Szczepaniak and Adam Niewiadomski for motivating, very interesting and valuable discussions.

References

[1] V. Kumar, D. Kumar, and R. Singh, "Outlier mining in medical databases: an application of data mining in health care management to detect abnormal values presented in medical databases," *IJCSNS International Journal of Computer Science and Network Security*, pp. 272–277, 2008.
[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.
[3] M. Hauskrecht, I. Batal, M. Valko, S. Visweswaran, G. F. Cooper, and G. Clermont, "Outlier detection for patient monitoring and alerting," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 47–55, 2013.
[4] C. C. Aggarwal, *Outlier Analysis*. Springer Science and Business Media, 2013.
[5] A. Duraj, P. S. Szczepaniak, and J. Ochelska-Mierzejewska, "Detection of outlier information using linguistic summarization." in *Flexible Query Answering Systems 2015; Advances in Intelligent Systems and Computing 400, (Eds.: Andreasen T., et al.), Proceedings of the 11th International Conference FQAS 2015, Cracow, Poland; Springer 2016*, 2015, pp. 101–113.

[6] A. Duraj and P. S. Szczepaniak, "Information outliers and their detection," in *M. Burgin and W. Hofkirchner (Eds.): Information Studies and the Quest for Transdisciplinarity*, vol. 9, Chapter 15. World Scientific Publishing Company, 2017, pp. 413–437.

[7] C. C. Aggarwal and P. S. Yu, *Outlier detection for high dimensional data*. ACM Sigmod Record, vol. 30, no. 2.

[8] D. M. Hawkins, *Identification of outliers*. Springer, 1980, vol. 11.

[9] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.

[10] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[11] V. Barnett and T. Lewis, *Outliers in statistical data*. Chichester: John Wiley, 1995. 584p, 1964.

[12] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John wiley & sons, 2005, vol. 589.

[13] H. Galicia, Q. He, and J. Wang, "A bayesian supervisory approach of outlier detection for recursive soft sensor update," in *CPC VIII Conference*, vol. 54, 2012.

[14] M. E. Otey, A. Ghoting, and S. Parthasarathy, "Fast distributed outlier detection in mixed-attribute data sets," *Data mining and knowledge discovery*, vol. 12, no. 2-3, pp. 203–228, 2006.

[15] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB JournalThe International Journal on Very Large Data Bases*, vol. 8, no. 3-4, pp. 237–253, 2000.

[16] E. M. Knox and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," pp. 392–403, 1998.

[17] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[18] W. Jin, A. K. Tung, and J. Han, "Mining top-n local outliers in large databases," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 293–298.

[19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[20] G. H. Orair, C. H. Teixeira, W. Meira Jr, Y. Wang, and S. Parthasarathy, "Distance-based outlier detection: consolidation and renewed bearing," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1469–1480, 2010.

[21] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg, "Outlier detection under interval uncertainty: algorithmic solvability and computational complexity," *Reliable Computing*, vol. 11, no. 1, pp. 59–76, 2005.

[22] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 190–237, 2014.

[23] A. Nowak-Brzezińska, "Mining rule-based knowledge bases inspired by rough set theory," *Fundamenta Informaticae*, vol. 148, no. 1-2, pp. 35–50, 2016.

[24] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in data preprocessing: experiments in medical domains," *Applied Artificial Intelligence*, vol. 14, no. 2, pp. 205–223, 2000.

[25] S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka, and H.-C. Wulf, "Detection of skin cancer by classification of raman spectra," *IEEE transactions on biomedical engineering*, vol. 51, no. 10, pp. 1784–1793, 2004.

[26] P. A. Ortega, C. J. Figueroa, and G. A. Ruz, "A medical claim fraud/abuse detection system based on data mining: A case study in chile." *DMIN*, vol. 6, pp. 26–29, 2006.

[27] T. Santhanam and M. Padmavathi, "Comparison of k-means clustering and statistical outliers in reducing medical datasets," in *Science Engineering and Management Research (ICSEMR), 2014 International Conference on*. IEEE, 2014, pp. 1–6.

[28] L. Bouarfa and J. Dankelman, "Workflow mining and outlier detection from clinical activity logs," *Journal of biomedical informatics*, vol. 45, no. 6, pp. 1185–1190, 2012.

[29] X. Albà, M. Pereañez, C. Hoogendoorn, A. J. Swift, J. M. Wild, A. F. Frangi, and K. Lekadir, "An algorithm for the segmentation of highly abnormal hearts using a generic statistical shape model," *IEEE transactions on medical imaging*, vol. 35, no. 3, pp. 845–859, 2016.

[30] J. Laurikkala, M. Juhola, E. Kentala, N. Lavrac, S. Miksch, and B. Kavsek, "Informal identification of outliers in medical data," in *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, vol. 1, 2000, pp. 20–24.

[31] S. J. Roberts, "Extreme value statistics for novelty detection in biomedical data processing," *IEE Proceedings-Science, Measurement and Technology*, vol. 147, no. 6, pp. 363–367, 2000.

[32] F. Å. Nielsen and L. K. Hansen, "Modeling of activation data in the brainmap database: Detection of outliers," *Human brain mapping*, vol. 15, no. 3, pp. 146–156, 2002.

[33] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: a review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[34] F. Shaari, A. A. Bakar, and A. R. Hamdan, "A predictive analysis on medical data based on outlier detection method using non-reduct computation," in *International Conference on Advanced Data Mining and Applications*. Springer, 2009, pp. 603–610.

[35] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion, "Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators," *Medical image analysis*, vol. 16, no. 7, pp. 1359–1370, 2012.

[36] A. Duraj and A. Krawczyk, "Finding outliers for large medical datasets," *Przeglad Elektrotechniczny*, vol. 86, pp. 188–191, 2010.

[37] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, no. 1, pp. 69–86, 1982.

[38] R. Yager, "Linguistic summaries as a tool for databases discovery." *Workshop on Fuzzy Databases System and Information Retrieval*, 1995.

[39] R. R. Yager, "Database discovery using fuzzy sets," *International Journal of Intelligent Systems*, vol. 11, no. 9, pp. 691–712, 1996.

[40] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[41] ——, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy sets and systems*, vol. 90, no. 2, pp. 111–127, 1997.

[42] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 813–834, 2000.

[43] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General System*, vol. 30, no. 2, pp. 133–154, 2001.

[44] J. Kacprzyk, A. Wilbik, and S. Zadrozny, "Linguistic summaries of time series via a quantifier based aggregation using the sugeno integral," in *Fuzzy Systems, 2006 IEEE International Conference on*. IEEE, 2006, pp. 713–719.

[45] J. Kacprzyk and S. Zadrozny, "Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 3, pp. 461–472, 2010.

[46] J. Kacprzyk and S. Zadrożny, "Bipolar queries: Some inspirations from intention and preference modeling," in *Combining Experimentation and Theory*. Springer, 2012, pp. 191–208.

[47] A. Niewiadomski, *Methods for the Linguistic Summarization of Data: Aplications of Fuzzy Sets and Their Extensions*. Akademicka Oficyna Wydawnicza" Exit", 2008.

[48] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.

[49] R. R. Yager, "Linguistic summaries as a tool for database discovery," in *FQAS*, 1994, pp. 17–22.