The Concept of Detecting and Classifying Anomalies in Large Data Sets on a Basis of Information Granules

Adam Kiersztyn
Department of Computer Science
Lublin University of Technology
Lublin, Poland
adam.kiersztyn.pl@gmail.com

Paweł Karczmarek
Department of Computer Science
Lublin University of Technology
Lublin, Poland
pawel.karczmarek@gmail.com

Krystyna Kiersztyn
Department of Mathematical Modeling
The John Paul II Catholic University of Lublin
Lublin, Poland
krystyna.kiersztyn@gmail.com

Witold Pedrycz
Department of Electrical & Computer Engineering
University of Alberta
Edmonton, Canada
Department of Electrical and Computer Engineering
King Abdulaziz University
Jeddah, Saudi Arabia
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
wpedrycz@ualberta.ca

***Abstract*—Anomaly (outlier) detection is one of the most important problems of modern data analysis. Anomalies can be the results of database users' mistakes, operational errors or just missing values. The problem is important because of fast growth of the large data sets. Therefore, we present the initial results of work on a Granular Computing approach to data imputation and missing data analysis. Our proposal brings intuitive and interpretable solutions. Finally, in a series of experiments, we demonstrate its effectiveness for a large dataset in the area of transport.**

***Index Terms*—anomaly detection, anomaly classification, fuzzy clustering, granules of information, data transformation**

I. Introduction

In the current reality, more and more large data sets have been collected and processed. However, in many cases these files contain missing data or, from some point of view, significantly worse, erroneous data. Therefore, data integration and aggregation are two of the main challenges faced by researchers in the field of data analysis [1]–[5] and many others fields [6]–[8], in particular economy [9], [10] or engineering [11], [12]. There are a number of solutions to detect anomalies and outliers in Big Data [13]–[18] and time series [19], [20]. Traditionally, such tasks have been addressed by probability and statistical techniques [21], [22]. In recent years, more and more attention has been paid to the application of artificial intelligence methods [23], [24], in particular deep learning [25]. However, most of these solutions are closely coordinated with the type of variables analysed and work directly on individual elements without attempting to generalise the considerations and transfer them to a greater level of abstraction. Therefore, it is important to develop innovative methods for detecting anomalies in large data sets using information granules.

The task of detecting anomalies in the data, in particular by detecting outliers, is a very complex problem [26]. There are many approaches to obtaining these values [27]. However, as previously stated, most of these approaches are closely related to the type of analysed data [28]–[30]. Therefore, it is reasonable to develop methods operating at a certain level of abstraction in isolation from the values of individual elements of the analysed data sets or time series. The ideal solution here seems to be an application of broadly understood fuzzy set-based techniques or, in general, information granules, which allow a transfer of such considerations to a higher level of generality.

The concept of information granules was proposed in the 1970s, but it is currently undergoing a renaissance due to numerous publications promoting this idea [31]–[36]. Therefore, our main goal is to propose universal solutions to granular data representation as well as search and classify anomalies using this granular description. Only a few papers [37]–[41]

Funded by the National Science Centre, Poland under CHIST-ERA programme (Grant no. 2018/28/Z/ST6/00563)

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

present a general concept of using information granules to detect and classify anomalies, but the most of them contain only preliminary and general considerations. Therefore, it is justified to undertake research in this direction. The novelty of the proposed approach is to develop methods operating at a certain level of abstraction, in isolation from the data analysed. Thanks to this innovative approach, it will be possible to provide universal methods for detecting anomalies in various types of data sets. Moreover, our goal is to shed some light on the aspect of Granular Computing and fuzzy set-based computations to the data analysis. Therefore, the number of references is relatively large. However, as an initial work in the area, we focus on the problem and show the background of the topic, particularly within the framework of Granular Computing.

Finally, our motivation is to present the potential of our proposal in a series of experiments carried with the real data in the area of transportation. The results of the experiments show a high efficiency of the approach.

We meet the need to collect and process data in many areas of life. Very often we deal with large data sets in which outliers or other types of anomalies naturally appear. The occurrence of these artifacts may be caused by system errors, or in many cases human error, because man is very often the weakest link in the system. Therefore, the ability to detect and classify all types of anomalies is highly desirable.

So far, a number of methods for detecting anomalies have been developed in both large data sets [42] and in time series [43], but in most cases the proposed methods are very closely related to the nature of the analysed data. Studies on universal methods of detecting anomalies [44] continue to be conducted, but these methods are still based on the analysis of individual elements of the examined sets.

The justification of the research problem is the need to develop a universal system for detection and classification of anomalies, which will operate at a higher level of abstraction and will not be strongly dependent on the data analysed. The proposed application of the information granules paradigm allows for conducting the reflection in a way apart from the level of raw data. The use of information granules to detect and classify anomalies in large data sets and time series is an innovative approach that has the potential to introduce significant added value. Introducing considerations to a higher level of abstraction will allow the use of more general methods of data analysis that are not dependent on the nature of the data being analysed. It is difficult to compare with each other the individual elements of multidimensional data sets, whose particular components often have different values and even different types of data. By detaching from individual data and specific set structures, it will be possible to introduce new, universal solutions that will be able to work properly for various data from different sources.

Generally, the proposed concept is based on recoding available data into vectors in the n-dimensional space (anomaly detection stage) and examining the mutual position of such new entities (anomaly classification stage) in isolation from the source data [41]. We will call the newly created vectors information granules.

The structure of the paper is as follows. In section II the proposed method is discussed. Section III covers the results of experiments, while the last section IV is devoted to conclusions and future work.

II. Method description

The main idea of transferring considerations into abstract elements is to use the recoding of the value of a single column of the data set into a fixed number range. In the approach described here, a discrete range of numbers is considered, namely $\{-N, -(N-1), \ldots, -1, 0, 1, N-1, N\}$, where encoding a single value by zero means no anomalies, the sign specifies the type of anomaly (underestimation or overestimation of values), while the descriptor module indicates the strength of the anomaly. For data with a distribution similar to the normal distribution, the thresholds for the range of descriptors can be associated with quantiles. In other cases, the thresholds are set arbitrarily by an expert based on his/her knowledge. In the analysed data set, each record (row) is transformed into a numerical vector whose components are these descriptors. Data transformation is carried out using the function

$$T(x) = \begin{cases} -5, & \text{if } x < x_0 \\ -4, & \text{if } x_0 \leq x < x_1 \\ -3, & \text{if } x_1 \leq x < x_2 \\ -2, & \text{if } x_2 \leq x < x_3 \\ -1, & \text{if } x_3 \leq x < x_4 \\ 0, & \text{if } x_4 \leq x < x_5 \\ 1, & \text{if } x_5 \leq x < x_6 \\ 2, & \text{if } x_6 \leq x < x_7 \\ 3, & \text{if } x_7 \leq x < x_8 \\ 4, & \text{if } x_8 \leq x < x_9 \\ 5, & \text{if } x_9 \leq x \end{cases} \tag{1}$$

At the stage of recoding values into a fixed numerical set, the process of detecting anomalies is carried out. It is possible to use fuzzy decision rules here, but in this case two vectors are obtained for each: one of them contains the values of the descriptors and the other the degree of membership to the given descriptors. At this stage, only crisp values are considered and only the degrees of membership in individual classes of anomalies are blurred.

After the transformation of the analysed dataset using carefully selected decision rules, considerations are transferred to the level of abstract entities described by vectors of integers.

For Z^n sub-space elements, classification can be done using well-known fuzzy grouping techniques, e.g. Fuzzy C-Means. At the same time, it is suggested to define the centres of individual clusters on the edge of the considered collection and ideally in its centre. For example, in the case of three-dimensional space $\{N, -N+1, \ldots, -1, 0, 1, \ldots, N-1, N\}^3$ cluster centres should be points with coordinates of the form $(x, 0, y)$, where $x, y \in \{-N, 0, N\}$.

It is obvious that in many practical examples, the data analysed contain various types of information, not all of them

which need to be at risk of anomalies. In this case, only those features (columns) in which anomalies may occur are considered for analysis. In addition, information calculated on the basis of available raw data can be analysed.

Membership in individual clusters can be designated as standardised reciprocals of the distance of a given point (its descriptor) from the imposed cluster centres.

III. Experimental study

Experiments were carried out using Knime analitics platforms. Part of the data set available in the set provided in [45] was analysed, limited to the last part of the collection containing a total of 15,004,556 rows. Initially, the rows with missing values were removed and only 15,003,820 rows remained after this operation. For each line describing one taxi journey, the following information is available: hack_license, vendor_id, rate_code, pickup_datetime, dropoff_datetime, passenger_count, trip_time_in_secs, trip_distance, pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude. The results of preliminary research on the use of such an approach are presented in the experimental section, where the data describing the New York taxi [45] courses were tested.

The values of the basic input statistics are presented in Table I.

Analysing the basic statistics for the available data (Table II), we can see that the database is burdened with outliers, which are probably incorrect, as evidenced by extreme values of longitude and latitude.

Based on the available data, the values of auxiliary variables such as calculated_trip_time, which is the difference between dropoff_datetime and pickup_datetime are calculated. In addition, the following variables are calculated: distance in a straight line, distance difference, percentage of added road, time difference, speed based on available data, speed based on calculated data and distance from Times Square.

Analyzing the values in the table, it can be seen that there are outliers for all of the newly determined variables. For example, in the database there are taxi rides for which the average speed was 102 960 mph. In addition, some taxi routes began at a point more than 10,000 miles from Times Square. Due to justified suspicion of anomalies, some of the variables were transformed into discrete values describing the direction and strength of the anomalies in the analysed data by using the function given by the formula (1). Due to the fact that the analysed values do not have normal distributions, the thresholds $x_0, \ldots, x_9$ were set arbitrarily in accordance with the values in the Table III.

After introducing a unified scale determining the type and strength of anomalies, one can visualise and compare individual points. Figs. 1-3 show the locations of the start points of the course by a taxi marked with the colour of the descriptor value describing the percentage of the overpowered road.

At first glance, it can be seen that points located on the ocean and at a significant distance from New York have a value of 5 for the descriptor describing the percentage of the added road, i.e., the declared length of the route is more than twice the distance in a straight line between the beginning and end point of the analysed taxi route.

A more detailed analysis (zoom of the image) shows that also for points located in the vicinity of New York, as well as in the city itself, there is a significant difference between the declared route length and the distance in a straight line.

The visualisation of the distance from Times Square (Figs. 4 - 5) presents the concept of introducing the anomaly descriptor described above very well. Points located at most 25 miles from Times Square are marked in blue, while points located at least 100 miles from the square in red. Intermediate colours

TABLE I
Based statistics of available data

Attribute	Min	Max	Mean	Std. deviation
Passenger_count	0	9	1.7	1.5
Trip_time_in_secs	0	10800	784.2	582.1
Trip_distance	0	100	2.9	3.4
Pickup_longitude	-2150.2	83.4	-73.2	7.8
Pickup_latitude	-49.5	500.6	40.3	4.3
Dropoff_longitude	-2130.8	160.6	-73.1	8.0
Dropoff_latitude	-3114.4	494.8	40.3	4.5

TABLE II
Based statistics of calculated data

Attribute	Min	Max	Mean	Std. deviation
Distance in a straight line	0	10324.1	10.3	208.5
Distance difference	-10321.1	99.4	-7.3	208.6
Percentage of added road	1.72E-05	88828596	2153.8	133474.3
Time difference	-317174	2147735	1.4	686.5
Speed based on available data	0	102960	13.9	141.4
Speed based on calculated data	-2769176	3875174	27.3	6983.9
Distance from Times Square	8.8E-04	10324.8	65.8	582

TABLE III
Thresholds taken into account when transforming empirical values into a fixed scale for classifying anomalies

Threshold / Column	Percentage of overpowered road	Relative time difference	Distance from Times Square	Speed based on available data
x_0	-15	-15	0	0
x_1	-10	-10	0	0
x_2	-5	-5	0	0
x_3	-1	-1	0	0
x_4	0	0	0	0
x_5	10	10	25	25
x_6	20	20	30	27
x_7	30	30	40	30
x_8	50	50	50	40
x_9	100	100	100	50

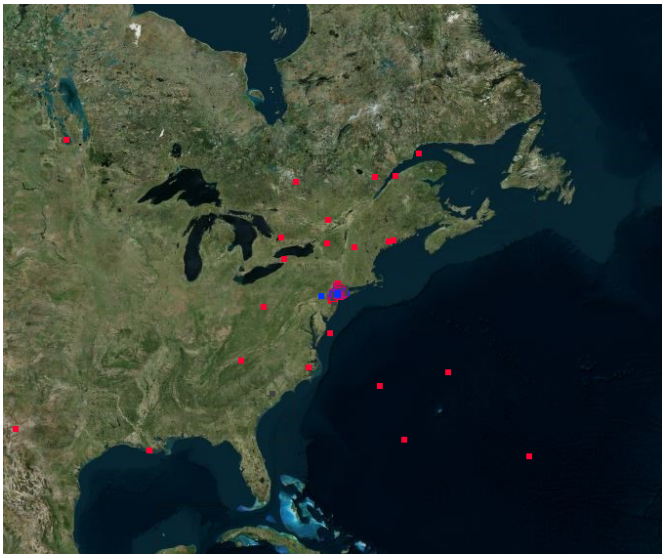

Fig. 1. The results with the use of a descriptor presentining the percentage of added road, i.e. the difference between the distance declared and the distance in a straight line normed by the distance in a straight line. The red value is 5 for the descriptor and the blue value is 0 for the descriptor, intermediate values are the intermediate colours.

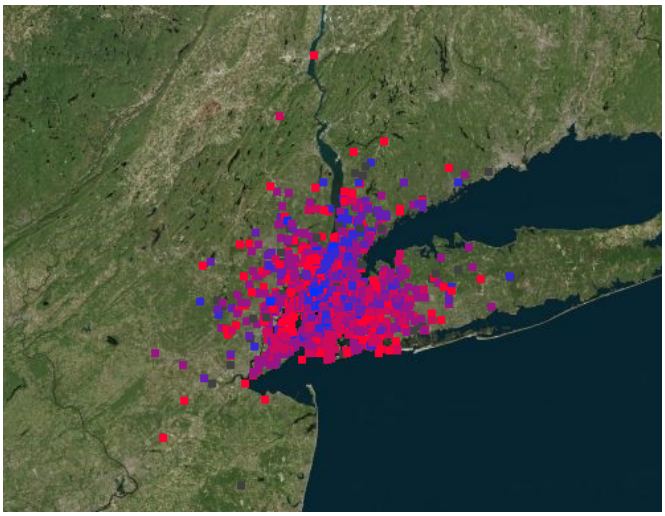

Fig. 2. The results with the use of a descriptor presenting the percentage of added road, i.e. the difference between the distance declared and the distance in a straight line normed by the distance in a straight line. The red value is 5 for the descriptor and the blue value is 0 for the descriptor, intermediate values are the intermediate colours (The area bounded by the neighbourhoods of New York City)

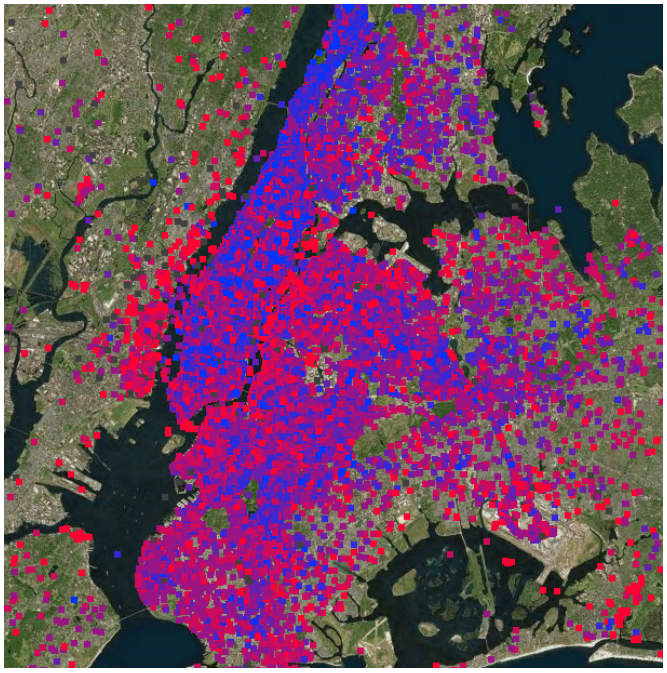

Fig. 3. The results with the use of a descriptor presenting the percentage of added road, i.e. the difference between the distance declared and the distance in a straight line normed by the distance in a straight line. The red value is 5 for the descriptor and the blue value is 0 for the descriptor, intermediate values are the intermediate colours (zoomed picture)

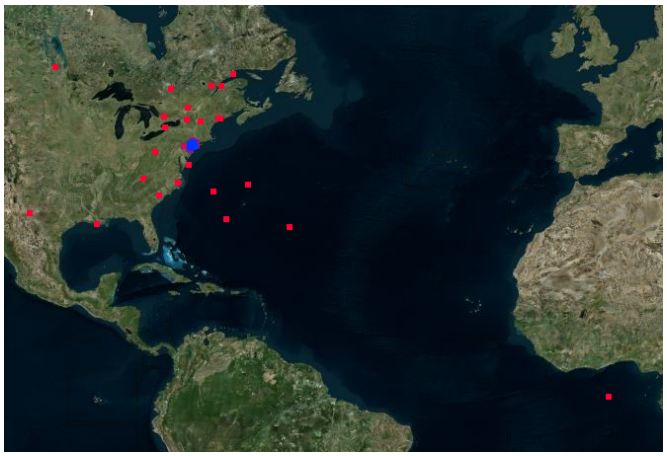

Fig. 4. Visualisation of the descriptor value describing the distance from Times Square.

mean intermediate descriptor values.

There is no such simple relationship between the location of the taxi start point and the descriptor value in the case of the descriptor presenting the average speed at which the course was carried out.

As can be seen in Fig. 6, the beginnings of routes for which the average speed exceeding 100 mph, calculated on the basis of the declared time and length of travel, are located both in the city and its surroundings.

It should also be noted that in the heart of New York some drivers are able to travel at speeds exceeding 100 miles per hour.

Analysing the results of Table IV, we find that the vast majority of data are not classified as anomalies (value 0 for individual descriptors). However, within each descriptor there are points that have a significant deviation from the standard, i.e. an anomaly. In addition, it turns out that four courses are characterised by the maximum degree of anomaly within each descriptor. The occurrence of missing values is caused by an incorrect range of raw data (geographical coordinates out of range).

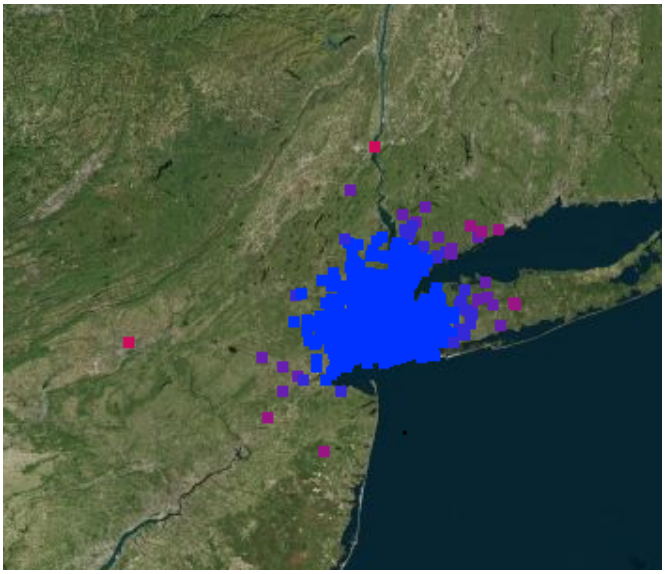

Fig. 5. Visualisation of the descriptor value describing the distance from Times Square (zoomed).

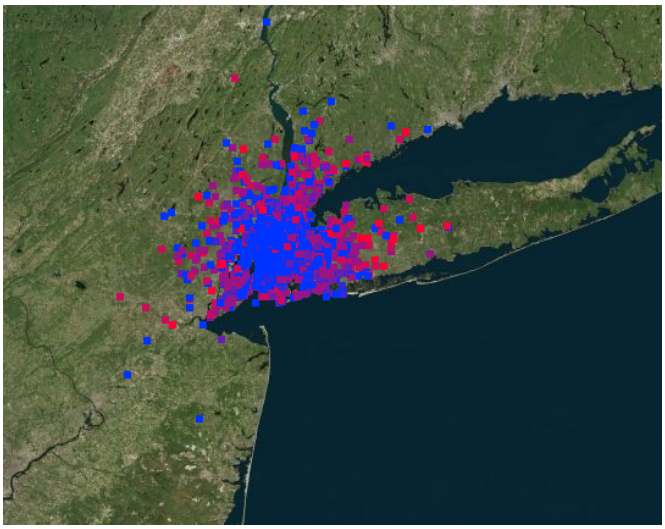

Fig. 6. Speed descriptor

TABLE IV
THE DISTRIBUTION OF THE SENSITIVITY OF INDIVIDUAL DESCRIPTORS

Descr. value	**Percent. of over-powered road granule**	**Time difference granule**	**Times Square granule**	**Speed granule**	**Speed 2 granule**
0	1733487	14946643	14824132	14071171	14656424
1	2189247	2804	816	219221	96980
2	2435743	1721	564	236565	89545
3	4465728	1421	212	349897	79233
4	2358264	1504	111	71155	10520
5	1575226	7813	177985	13870	25099
no data	246125	41914	0	41941	46019

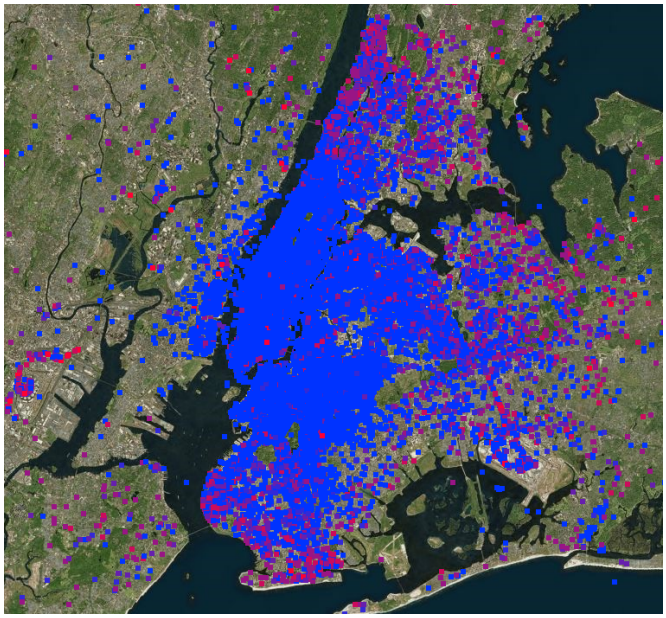

Fig. 7. Speed descriptor (zoomed)

In the analysed example, in which five dimensions of the information grain were distinguished, cluster centres were limited to points where one coordinate equals 5 and to points with coordinates (0;0;0;0;0) and (5;5;5;5;5) corresponding to elements showing an extremely high anomaly within one coordinate, elements free from anomalies and elements characterised by a high anomaly within each grain dimension.

As degrees of membership to individual clusters, as mentioned earlier, one can use normalised inverses of distance from individual clusters. It is obvious that with such a membership function, one element belongs to several clusters at the same time, and the closer the element is to the centre of a given cluster, the greater the value of belonging. The proposed solution for determining the degree of membership to individual clusters (see Table V) is an alternative to using other fuzzy clustering methods, such as Fuzzy C-Means, which with such a large number of tested elements and any large dimension of the space under consideration do not work properly.

Thus, the most of the analysed taxi routes are characterised by a relatively large percentage of additional roads, which due

TABLE V
PERCENTAGE OF ITEMS MOST BELONGING TO A GIVEN CLUSTER

Cluster center	**Percentage**
“0;0;0;0;0”	0.425427
“0;0;0;0;5”	0.003568
“0;0;0;5;0”	0.00279
“0;0;5;0;0”	0.00458
“0;5;0;0;0”	0.0101
“5;0;0;0;0”	0.547752
“5;5;5;5;5”	0.005784

to the specificity of the city and the density of roads may suggest that drivers specially choose longer routes. Analysing the values in Fig. 8, we can conclude that the most numerous group of anomalies is the additional route.

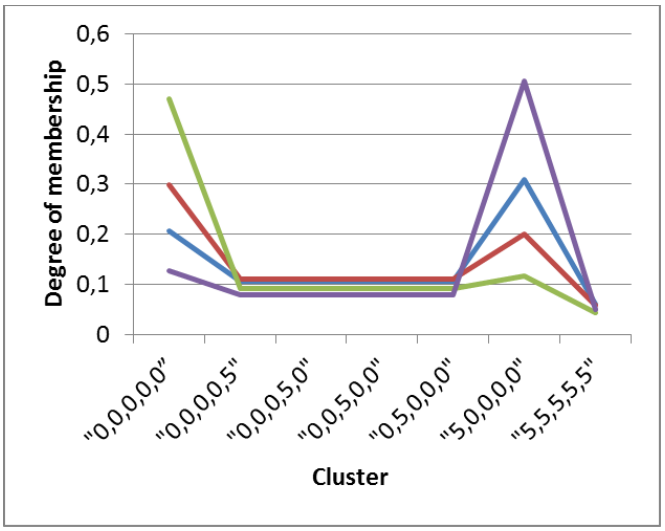

Fig. 8. Distribution of belonging of the 4 most numerous groups of elements constituting nearly 0.75 of all objects.

IV. Conclusions and Future Work

In this study, we have presented the concept of moving data data analysis to a higher level of abstraction and detachment from the analysis of raw data, which has great potential for applications. We have shown that it is possible to fully develop the proposed concept and introduce other, more sophisticated methods of data analysis carried out on abstract entities being a granular description of raw data. Moreover, we have thoroughly considered the analysis of the literature related to the Granular Computing problems in the area of data analysis, in particular, anomaly or outlier detection.

Future work should concentrate on in-depth analysis of Granular Computing possibilities to cover the problems discussed. In addition, it is planned to combine the proposed solutions with other techniques, such as those described in [46]–[48].

Acknowledgment

Funded by the National Science Centre, Poland under CHIST-ERA programme (Grant no. 2018/28/Z/ST6/00563).

References

[1] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, "Big data challenge: a data management perspective," *Front. Comput. Sci.*, vol. 7, no. 2, pp. 157–164, 2013.

[2] A. Alyass, M. Turcotte, and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC Medical Genom.*, vol. 8, no. 1, p. 33, 2015.

[3] A. Cuzzocrea, I.-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!" in *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP*. Association for Computing Machinery, 2011, pp. 101–104.

[4] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, 2015.

[5] A. Kiersztyn and P. Karczmarek, "Fuzzy approach for detection of anomalies in time series," in *Artificial Intelligence and Soft Computing*. Cham: Springer International Publishing, June 2019, pp. 397–406.

[6] K. Chromiński and M. Tkacz, "Comparison of outlier detection methods in biomedical data," *J. Med. Inform. Technol.*, vol. 16, pp. 89–94, 2010.

[7] A. Bartkowiak, "Outliers in biometrical data: what's old, what's new," *Int J Biom*, vol. 2, no. 1, pp. 2–18, 2010.

[8] L. Rettig, M. Khayati, P. Cudré-Mauroux, and M. Piórkowski, *Online anomaly detection over big data streams*. Springer, 2019, pp. 289–312.

[9] K. Golmohammadi and O. R. Zaiane, "Time series contextual anomaly detection for detecting market manipulation in stock market," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, October 2015, pp. 1–10.

[10] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, 2018.

[11] M. Sanayha and P. Vateekul, "Fault detection for circulating water pump using time series forecasting and outlier detection," in *2017 9th International Conference on Knowledge and Smart Technology (KST)*. IEEE, February 2017, pp. 193–198.

[12] H. N. Akouemo and R. J. Povinelli, "Probabilistic anomaly detection in natural gas time series data," *Int J Forecast*, vol. 32, no. 3, pp. 948–956, 2016.

[13] A. Razaq, H. Tianfield, and P. Barrie, "A big data analytics based approach to anomaly detection," in *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, ser. BDCAT 16. Association for Computing Machinery, December 2016, pp. 187–193.

[14] L. Stojanovic, M. Dinic, N. Stojanovic, and A. Stojadinovic, "Big-data-driven anomaly detection in industry (4.0): An approach and a case study," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, December 2016, pp. 1647–1652.

[15] Z. Hasani, "Robust anomaly detection algorithms for real-time big data: Comparison of algorithms," in *2017 6th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, June 2017, pp. 1–6.

[16] M. Ahmed, N. Choudhury, and S. Uddin, "Anomaly detection on big data in financial markets," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, July 2017, pp. 998–1001.

[17] R. Alguliyev, R. Aliguliyev, and L. Sukhostat, "Anomaly detection in big data based on clustering," *Stat. Optim. Inf. Comput.*, vol. 5, no. 4, p. 325, 2017.

[18] P. Casas, F. Soro, J. Vanerio, G. Settanni, and A. D'Alconzo, "Network security and anomaly detection with Big-DAMA, a big data analytics framework," in *2017 IEEE 6th International Conference on Cloud Networking (CloudNet)*. IEEE, September 2017, pp. 1–7.

[19] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[20] A. Zhang, S. Song, J. Wang, and P. S. Yu, "Time series data cleaning: From anomaly detection to anomaly repairing," *Proceedings of the VLDB Endowment*, vol. 10, no. 10, pp. 1046–1057, 2017.

[21] B. Abraham and G. E. Box, "Bayesian analysis of some outlier problems in time series," *Biometrika*, vol. 66, no. 2, pp. 229–236, 1979.

[22] J. Li, W. Pedrycz, and I. Jamal, "Multivariate time series anomaly detection: A framework of Hidden Markov Models," *Appl. Soft Comput.*, vol. 60, pp. 229–240, 2017.

[23] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognit*, vol. 58, pp. 121–134, 2016.

[24] T. Kieu, B. Yang, and C. S. Jensen, "Outlier detection for multidimensional time series using deep neural networks," in *2018 19th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2018, pp. 125–134.

[25] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Comput*, pp. 1–13, 2017.

[26] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Min Knowl Discov*, vol. 29, no. 3, pp. 626–688, 2015.

[27] E. de la Hoz, E. de la Hoz, A. Ortiz, J. Ortega, and A. Martínez-Álvarez, "Feature selection by multi-objective optimisation: Application

to network anomaly detection by hierarchical self-organising maps," *Knowl Based Syst*, vol. 71, pp. 322–338, 2014.
[28] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *Int J Inf Manage*, vol. 45, pp. 289–307, 2019.
[29] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 194–207, 2012.
[30] A. Duraj, P. S. Szczepaniak, and J. Ochelska-Mierzejewska, "Detection of outlier information using linguistic summarization," in *Flexible query answering systems 2015*. Springer, 2016, pp. 101–113.
[31] A. Bargiela and W. Pedrycz, *Human-centric information processing through granular modelling*. Springer Science & Business Media, 2009, vol. 182.
[32] ——, "Granular computing," in *Handbook on Computational Intelligence: Volume 1: Fuzzy Logic, Systems, Artificial Neural Networks, and Learning Systems*. World Scientific, 2016, pp. 43–66.
[33] W. Pedrycz, *Knowledge-based clustering: from data to information granules*. John Wiley & Sons, 2005.
[34] W. Pedrycz and S.-M. Chen, *Granular computing and intelligent systems: design with information granules of higher order and higher type*. Springer Science & Business Media, 2011, vol. 13.
[35] ——, *Information granularity, big data, and computational intelligence*. Springer, 2014, vol. 8.
[36] S. K. Pal, S. K. Meher, and A. Skowron, "Data science, big data and granular mining," *Pattern Recognit Lett*, vol. 67, no. 2, pp. 109–112, 2015.
[37] Y. Chen, D. Miao, and R. Wang, "Outlier detection based on granular computing," in *International Conference on Rough Sets and Current Trends in Computing*, October 2008, pp. 283–292.
[38] F. Jiang and Y.-M. Chen, "Outlier detection based on granular computing and rough set theory," *Appl. Intell.*, vol. 42, no. 2, pp. 303–322, 2015.
[39] X. Zhu, W. Pedrycz, and Z. Li, "Granular models and granular outliers," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3835–3846, 2018.
[40] M. R. Mashinchi, A. Selamat, S. Ibrahim, and H. Fujita, "Outlier elimination using granular box regression," *Inf Fusion*, vol. 27, pp. 161–169, 2016.
[41] C. Zhong, W. Pedrycz, D. Wang, L. Li, and Z. Li, "Granular data imputation: a framework of granular computing," *Appl. Soft Comput.*, vol. 46, pp. 307–316, 2016.
[42] H. Izakian and W. Pedrycz, "Anomaly detection and characterization in spatial time series data: A cluster-centric approach," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 6, pp. 1612–1624, 2014.
[43] ——, "Anomaly detection in time series data using a fuzzy C-Means clustering," in *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, June 2013, pp. 1513–1518.
[44] P. Karczmarek, A. Kiersztyn, W. Pedrycz, and E. Al, "K-Means-based isolation forest," *Knowl Based Syst*, vol. 195, p. 105659, February 2020, in press. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705120301064
[45] B. Donovan and D. Work, "New York City taxi trip data (2010-2013)," 2016. [Online]. Available: https://doi.org/10.13012/J8PN93H8
[46] H. Fujita, A. Gaeta, V. Loia, and F. Orciuoli, "Resilience analysis of critical infrastructures: a cognitive approach based on granular computing," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1835–1848, 2018.
[47] T. Lai, R. Chen, C. Yang, Q. Li, H. Fujita, A. Sadri, and H. Wang, "Efficient robust model fitting for multistructure data using global greedy search," *IEEE Trans. Cybern.*, 2019.
[48] T. Lai, H. Fujita, C. Yang, Q. Li, and R. Chen, "Robust model fitting based on greedy search and specified inlier threshold," *IEEE Trans. Ind. Electron.*, vol. 66, no. 10, pp. 7956–7966, 2018.