Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

11th International Conference on Information Technology and Quantitative Management (ITQM 2024)

Pattern Analysis Based on Interval Estimates: Ordinal-Interval Pattern Clustering

Alexey Myachin

National Research University Higher School of Economics, Myasnitskaya St., 20, Moscow, 101001, Russia
V.A. Trapeznikov Institute of Control Science of Russian Academy of Science, Profsoyuznaya St., 65, Moscow, 117997, Russia

Abstract

A new method for searching patterns in data using interval estimates is proposed. The computational complexity, algorithmic implementation, and some properties that allow the method to be applied to high-dimensional data are described, including the characteristics of ordering the studied indicators in groups obtained from the analysis, the unambiguous definition of such groups, and their lack of intersections. The possibility of using a generalization of Borda's rule for interval estimates in pairwise comparison of interval estimates in the practical implementation of this method has been studied. The term "ordinal-interval pattern clustering" is proposed. The potential for reducing computational complexity by introducing additional constraints is presented. Recommendations on the appropriateness of using ordinal-interval pattern clustering are discussed. An example of a practical application was studied using synthetic data. Recommendations for using ordinal-interval pattern clustering will help researchers and analysts optimize the data analysis process and obtain more accurate and interpretable results. Examples of practical application of the method on various types of data highlight its universality and effectiveness.

Keywords: pattern; pattern analysis; ordinal-interval pattern clustering

1. Introduction

Currently, there are numerous methods for extracting useful information from large datasets. One such method, which allows for the discovery of patterns in data, is pattern analysis [1, 9, 13]. Typically, these methods primarily employ numerical scales, presupposing the presence of precise values (with an acceptable level of error) for the features of the objects analyzed in a given study. The accumulation of measurement errors, inaccuracies in calculations, and mistakes in data entry can significantly impact the results of the analysis and the interpretation of the findings.

Consequently, in a number of practical tasks, it is advisable to transition to interval estimates, which presents some challenges: a limited variety of possible methods (compared to precise estimates); potential reduction in the quality of the final results; computational difficulties. There are several methods for discovering patterns in data based on interval estimates [2, 11]. Despite the complexities mentioned above, the significant increase in data volumes in recent years suggests a corresponding increase in error. Modern data analysis methods may also tolerate a certain degree of error in calculations. Employing methods that allow for computational errors with data that contains errors can lead to incorrect results. More about various types of errors and inaccuracies in data [5, 15].

Thus, the development of existing methods and the creation of new methods for working with interval data are currently highly relevant tasks. The significance of these tasks lies in improving the accuracy and efficiency of data analysis in situations where precise numerical values are difficult to obtain due to measurement errors or other uncertainties. Among these methods is a new pattern analysis technique based on interval data, proposed under the name 'ordinal-interval pattern clustering.' This paper details the algorithmic implementation, explores the individual properties and computational complexity of the proposed method, and demonstrates its practical application using synthetic data. The development of these new methods contributes to more robust and reliable data analysis processes, which are essential for making informed decisions in various fields.

In this work, the formal problem statement is addressed, the proposed algorithmic solution is detailed, specific properties are studied, and practical applications are considered.

2. Search for Pattern in Data Using Interval Estimates

In the context of increasing data uncertainty, methods for searching for patterns based on interval estimates are gaining popularity. For example, a search for "interval data clustering" on Google Scholar returns 4,680,000 results (with 18,300 in 2024 as of June 18, 2024). Part of the interest in these methods is described in the introduction to this work. Here, we provide a brief description of some methods that allow the identification of patterns in interval data.

One of the most well-known algorithms is fuzzy K-means clustering for interval data [2]. This method extends the classical K-means algorithm [6], where the objects under study are described not by exact values but by intervals. A distinctive feature of this method is the use of adaptive quadratic distances to measure the similarity between interval data, which allows for accounting for the internal structure of the data and reducing the influence of extreme values.

Another approach is the method of interval hierarchical clustering [4], which uses dendrograms for visualization and analysis of clusters. Unlike traditional hierarchical clustering, this method employs a special distance metric for interval data.

Granular Computing [12] also deserves attention. In this method, data is considered as a set of granules, each representing an interval. Clustering is performed by combining similar granules based on their distances and similarities. This approach allows for efficient handling of large volumes of data and improves the interpretability of results by working with higher-level abstractions.

Thus, clustering methods based on interval estimates provide powerful tools for data analysis under conditions of uncertainty. However, despite their advantages, interval-based clustering methods have several drawbacks. First, the computational complexity of these methods can increase significantly with the size of the data and the number of parameters, which often makes them less suitable for large datasets. Second, adapting distance metrics for interval estimates may require specific tuning, which can be a complex and time-consuming process. Third, most existing methods are not well studied in terms of their robustness to noise and outliers, which can lead to a decrease in the accuracy of the final results.

In light of these drawbacks, this work proposes a new method for searching for patterns in data based on interval estimates: "ordinal-interval pattern clustering."

3. Ordinal-Interval Pattern Clustering

Let us provide a formal description of the task. The input data consists of a finite set of objects $W = \{w_1, w_2,\ldots, w_n\}$. Each object in the set W is represented in vector form as $w_i = ([w_{i1}-\varepsilon_{i1}, w_{i1}+\varepsilon_{i1}],[w_{i2}-\varepsilon_{i2}, w_{i2}+\varepsilon_{i2}],\ldots,[w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}],\ldots,[w_{im}-\varepsilon_{im}, w_{im}+\varepsilon_{im}])$. It is assumed that preliminary data processing has been carried out (for more details, see [7, 14]). The tasks to be solved are as follows:

- Visualize multidimensional interval data.
- Identify patterns in the data being studied.
- Divide the set under study into groups so that objects within the same group are similar (according to a chosen proximity measure), while objects from different groups exhibit significant differences (based on the indicators studied).

In utilizing pattern analysis methods, visualization typically employs a parallel coordinates system [3], with additional consideration of the interval values of the input data. This may involve using the actual intervals or averaging them to simplify visual perception. An example of visualizing three objects with 4-dimensional interval data is presented in Fig. 1.

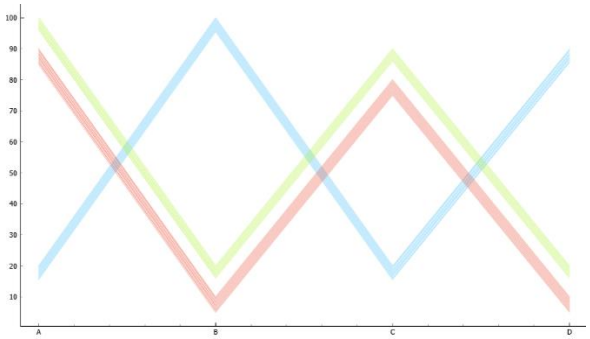

Fig. 1. Example of visualizing interval data in a parallel coordinates system

The search for patterns in the data will be conducted using a generalization of Borda's rule for interval estimates. According to this rule, the interval value $[w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}]$ dominates the interval value $[w_{ik}-\varepsilon_{ik}, w_{ik}+\varepsilon_{ik}]$ (typically denoted as $[w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}] \succ [w_{ik}-\varepsilon_{ik}, w_{ik}+\varepsilon_{ik}]$), if and only if $w_{ij}-\varepsilon_{ij} > w_{ik}+\varepsilon_{ik}$. Each object from the set W is assigned a code s_i according to the formula:

$$s_i = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} 10^{j-1} r_{ij}^{q}, \tag{1}$$

where:

$$\begin{cases} r_{ij}^{q} = 2, if\ [w_{ij} - \varepsilon_{ij}, w_{ij} + \varepsilon_{ij}] \succ [w_{iq} - \varepsilon_{iq}, w_{iq} + \varepsilon_{iq}] \\ r_{ij}^{q} = 1, if\ [w_{ij} - \varepsilon_{ij}, w_{ij} + \varepsilon_{ij}] \prec [w_{iq} - \varepsilon_{iq}, w_{iq} + \varepsilon_{iq}] \\ r_{ij}^{q} = 0, if\ [w_{ij} - \varepsilon_{ij}, w_{ij} + \varepsilon_{ij}] \infty [w_{iq} - \varepsilon_{iq}, w_{iq} + \varepsilon_{iq}] \end{cases} \tag{2}$$

For objects to be grouped together, their codes must be identical, i.e. $w_k, w_p \in v_{int}$ (a unified group formed based on the proposed method), if $s_k - s_p = 0$.

The pattern analysis method based on the above algorithm is referred to as "ordinal-interval pattern clustering." The groups of objects formed by this method are accordingly called "ordinal-interval pattern clusters."

4. Properties of Ordinal-Interval Pattern Clustering

The proposed method possesses several properties important for practical use and interpretation of final results:

Property 1. No object can belong to two different ordinal-interval pattern clusters.

Proof: Consider objects with respective interval representations $w_1, w_2 \in W$: $w_1 = ([w_{11}-\varepsilon_{11}, w_{11}+\varepsilon_{11}],[w_{12}-\varepsilon_{12}, w_{12}+\varepsilon_{12}],\ldots,[w_{1j}-\varepsilon_{1j}, w_{1j}+\varepsilon_{1j}],\ldots,[w_{1m}-\varepsilon_{1m}, w_{1m}+\varepsilon_{1m}])$, $w_2 = ([w_{21}-\varepsilon_{21}, w_{21}+\varepsilon_{21}],[w_{22}-\varepsilon_{22}, w_{22}+\varepsilon_{22}],\ldots,[w_{2j}-\varepsilon_{2j}, w_{2j}+\varepsilon_{2j}],\ldots,[w_{2m}-\varepsilon_{2m}, w_{2m}+\varepsilon_{2m}])$. Suppose two different ordinal-interval pattern clusters v_{int} and v^*_{int} are formed such that $v_{int} \neq v^*_{int}$, $w_1 \in v_{int}$, $w_2 \in v^*_{int}$.

If Property 1 is not true, it would imply the existence of at least one object belonging to both ordinal-interval pattern clusters v_{int} and v^*_{int} ($v_{int} \cap v^*_{int} = w_3$: $w_3 = ([w_{31}-\varepsilon_{31}, w_{31}+\varepsilon_{31}],[w_{32}-\varepsilon_{32}, w_{32}+\varepsilon_{32}],\ldots,[w_{3j}-\varepsilon_{3j}, w_{3j}+\varepsilon_{3j}],\ldots,[w_{3m}-\varepsilon_{3m}, w_{3m}+\varepsilon_{3m}])$).

Since $w_1, w_3 \in v_{int}$, implies $s_1 - s_3 = 0$, and $w_2, w_3 \in v^*_{int}$ also implies $s_2 - s_3 = 0$, it follows that $s_1 - s_2 = 0$, indicating w_1 and w_2 belong to the same cluster, thus $v_{int} = v^*_{int}$. Property 1 is proven.

Property 2. All indicators of objects belonging to the same ordinal-interval pattern cluster can be ordered in a consistent manner according to the rules: $[w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}] \backsim [w_{ij+1}-\varepsilon_{ij+1}, w_{ij+1}+\varepsilon_{ij+1}]$ if $w_{ij}+\varepsilon_{ij} \in [w_{ij+1}-\varepsilon_{ij+1}, w_{ij+1}+\varepsilon_{ij+1}]$ (particularly $w_{ij}+\varepsilon_{ij} = w_{ij+1}+\varepsilon_{ij+1}$); $[w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}] \prec [w_{ij+1}-\varepsilon_{ij+1}, w_{ij+1}+\varepsilon_{ij+1}]$ if $w_{ij+1}-\varepsilon_{ij+1} > w_{ij}+\varepsilon_{ij}$; $[w_{ij+1}-\varepsilon_{ij+1}, w_{ij+1}+\varepsilon_{ij+1}] \prec [w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}]$ if $w_{ij}-\varepsilon_{ij} > w_{ij+1}+\varepsilon_{ij+1}$. This order will be different for different ordinal-interval pattern clusters.

Proof: Assuming a set W with $|W| > 1$, two different clusters v_{int} and v^*_{int} ($v_{int} \neq v^*_{int}$) are formed. Each ordinal-invariant pattern cluster belongs to at least one object. Consider two objects $w_1, w_2 \in W$: $w_1 = ([w_{11}-\varepsilon_{11}, w_{11}+\varepsilon_{11}],[w_{12}-\varepsilon_{12}, w_{12}+\varepsilon_{12}],\ldots,[w_{1j}-\varepsilon_{1j}, w_{1j}+\varepsilon_{1j}],\ldots,[w_{1m}-\varepsilon_{1m}, w_{1m}+\varepsilon_{1m}])$, $w_2 = ([w_{21}-\varepsilon_{21}, w_{21}+\varepsilon_{21}],[w_{22}-\varepsilon_{22}, w_{22}+\varepsilon_{22}],\ldots,[w_{2j}-\varepsilon_{2j}, w_{2j}+\varepsilon_{2j}],\ldots,[w_{2m}-\varepsilon_{2m}, w_{2m}+\varepsilon_{2m}])$. Here, $w_1 \in v_{int}$, and $w_2 \in v^*_{int}$.

Next, assume that Property 2 is false, i.e., there exists a unified way to order the indicators of two different ordinal-interval pattern clusters. In other words, for the indicators of objects belonging to v_{int} and v^*_{int} there exists a unified ordering P. However, based on the method of forming object codes s_i (Equation 1) it is known that $r^q_{ij} = 2$, if $[w_{iq}-\varepsilon_{iq}, w_{iq}+\varepsilon_{iq}] \prec [w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}]$; $r^q_{ij} = 1$, if $[w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}] \prec [w_{iq}-\varepsilon_{iq}, w_{iq}+\varepsilon_{iq}]$; $r^q_{ij} = 0$, if $[w_{iq}-\varepsilon_{iq}, w_{iq}+\varepsilon_{iq}] \backsim [w_{ij}-\varepsilon_{ij}, w_{ij}+\varepsilon_{ij}]$ (Equation 2). Thus, if a unified P exists for v_{int} and v^*_{int} at least one object w_3: $s_3 = s_1$ and $s_3 = s_2$. Consequently, $v_{int} \cap v^*_{int} \neq \emptyset$, which contradicts Property 1. Property 2 is proven.

Furthermore, let us assess the computational complexity of ordinal-interval pattern clustering. With n objects and m parameters the complexity amounts to n^3m^3. However, analogous to other pattern analysis methods based on pairwise comparison of indicators [8-10], the complexity can be estimated from $O(mlog(n))$ to $O(n^2)$.

5. Practical Applications

Consider a practical application example of ordinal-interval pattern clustering. For simplicity, 1,000,000 objects with 10 indicators are generated, each defined in an interval form with conditions:

- $w_{ij} \in [0;100]$;
- $\varepsilon_{ij} < 40$;
- $\min(w_{ij}-\varepsilon_{ij}) = 0$;
- $\max(w_{ij}+\varepsilon_{ij}) = 100$.

For some data, precise estimates ($\varepsilon_{ij} = 0$) are included. The visualization, shown in Fig. 2a, uses piecewise linear functions based on average values. Examples of formed ordinal-interval pattern clusters are displayed in Fig. 2b.

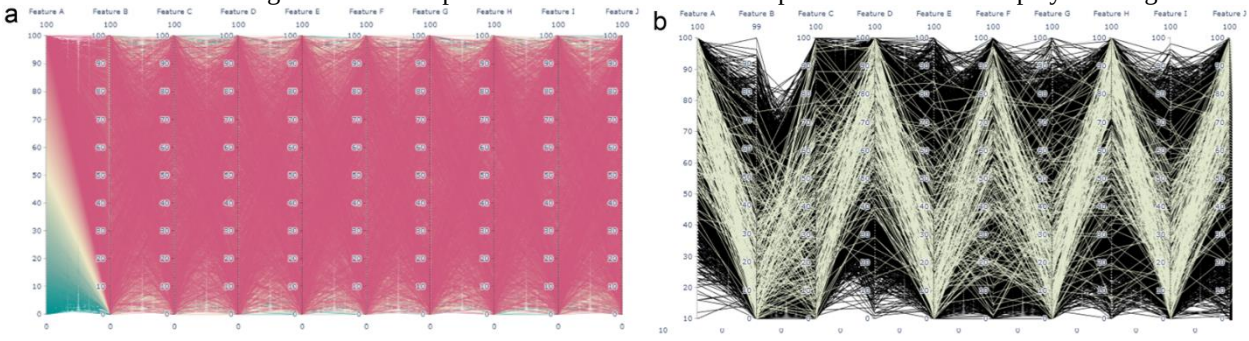

Fig. 2. (a) Visualization of average values of the initial data; (b) Examples of formed ordinal-interval pattern clusters.

Due to its relatively low computational complexity, this methodology can be effectively applied to high-dimensional data.

6. Conclusion

Pattern analysis methods, which are contemporary approaches to discovering data patterns, have significantly developed in recent years. This work introduces a new method for working with interval data, termed "ordinal-interval pattern clustering." The algorithmic implementation features relatively low computational complexity, enabling pattern discovery in large-scale data. The properties studied ensure the uniqueness and stability of the final partitioning of the object set based on ordinal-interval pattern clustering, also helping to reduce computational complexity. The practical implementation has been demonstrated with 1,000,000 generated objects. Further research will focus on adjusting results for varying values of ε_{ij}.

Acknowledgements

This work has been supported by the grants the Russian Science Foundation, RSF No. 24-61-00030, https://rscf.ru/project/24-61-00030/

References

[1] Aleskerov, Fuad, Veronika Belousova, Maria Serdyuk and Vasily Solodkov. (2008) “Dynamic analysis of the behavioural patterns of the largest commercial banks in the Russian federation.” *International Centre for Economic Research Working Paper* **12**.

[2] De Carvalho, Francisco de AT, and Camilo P. Tenório. (2010) “Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances.” *Fuzzy Sets and Systems* **161(23)**: 2978-2999.

[3] Inselberg, A., & Dimsdale, B. (1990) “Parallel coordinates: a tool for visualizing multi-dimensional geometry”, in *Proceedings of the first IEEE conference on visualization: visualization90*: 361-378.

[4] Gowda, K. Chidananda, and Edwin Diday. (1991) “Symbolic clustering using a new dissimilarity measure”. *Pattern recognition* **24.6**: 567-578.

[5] Klein, Barbara D., Dale L. Goodhue, and Gordon B. Davis. (1997) “Can humans detect errors in data? Impact of base rates, incentives, and goals.” *Mis Quarterly*: 169-194.

[6] MacQueen, James. (1967) “Some methods for classification and analysis of multivariate observations”, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **14**: 281-297.

[7] Mirkin, Boris. (2013) “Mathematical classification and clustering.” *Springer Science & Business Media*.

[8] Myachin, Alexey. (2016) “New methods of pattern analysis in the study of Iris Anderson-Fisher Data”, in *2016 6th International Conference on Computers Communications and Control (ICCCC)*: 97-102.

[9] Myachin, Alexey. (2019) “Pattern analysis in parallel coordinates based on pairwise comparison of parameters.” *Automation and Remote Control* **80**: 112–123.

[10] Myachin, Alexey. (2022) “Finding Structurally Similar Objects Based on Data Sorting Methods”, in *Science and Information Conference*: 826-835.

[11] Nagpal, Arpita, Arnan Jatain, and Deepti Gaur. (2013) “Review based on data clustering algorithms”, in *2013 IEEE conference on information & communication technologies*: 298-303.

[12] Pedrycz, Witold. (2018) “Granular computing: analysis and design of intelligent systems”. *CRC press.*

[13] Shawe-Taylor, John, and Nello Cristianini. (2004) “Kernel methods for pattern analysis”. *Cambridge university press*.

[14] Wickham, Hadley, and Hadley Wickham. (2016) “Data analysis.” *Springer International Publishing*

[15] Wolff, Hendrik, Howard Chong, and Maximilian Auffhammer. (2011) “Classification, detection and consequences of data error: evidence from the human development index.” *The Economic Journal* **121(553)**: 843-870.