Clustering of inaccurate data using information on its precision

Vitaliy Bolschikov 1* and Konstantin Semenov 1

1 Peter the Great St. Petersburg Polytechnic University, Polytechnicheskaya, 29, St. Petersburg, 195251, Russia

* E-mail: vitaliy.bolschikov@gmail.com

Abstract. In this paper, an approach to clustering inaccurate data with taking into account information on their imprecision is proposed. The presented procedure allows to reasonably determining the maximum number of clusters, coordinated it with data accuracy: the low accuracy of the measurement results will not give grounds for separating some of the object states from others. Also taking into account the uncertainty of the initial data allows to simplify the clustering procedure by matching the complexity of the methods used in its implementation to the accuracy of the initial data. The proposed procedure allows not to create or specify clusters during the classification of inaccurate data beyond what is necessary, the experimental results of the application of described approach are presented in this paper.

1. Introduction

Clustering allows to split-up data into groups according to a given metric of their similarity or difference. The execution of this operation depends on the origin of processed information: the best results for specific data are achieved by applying special methods and approaches. The question on the influence of initial data inaccuracy on the results obtained during clustering is considered for a moment in scientific literature mainly from a formal point of view. If the processed information is represented in form of intervals of possible values, it will be necessary to modify classical methods of data clustering. Methods for clustering interval data [1-4], fuzzy variables [5-8], Dempster-Shafer sets [9-11] are described and developed in literature. At the same time, the issues related to the determination the number of clusters and the agreement of the final results of clustering with the inaccuracy degree of the initial data are not considered.

Often an assessment of the reliability of clustering or are of a private nature [12], although this directly concerns of the obtained results quality. results in practice reduces to examining influence exerted on their quality by incorrectly chosen number of clusters [13], excessively large dimension of initial data [14], robustness of chosen clustering method [15] and correctness of chosen metric in it [16].

Such an approach is satisfactory, since its results will carry reliable information only if the original data is accurate enough. For this reason, the studies listed above are only auxiliary and uncertainty of initial data provides primary influence on reliability of the final results of clustering. At the same time, accounting of this uncertainty will allow not only to naturally correspond to results of clustering with amount of useful information that is actually contained in them, but also to achieve final results faster and easier – this completely eliminates necessity for carrying out mentioned types of studies.
In this paper, a generalized view to clustering inaccurate data, taking into account their imprecision, is proposed. The experimental results of application of the proposed approach are presented. They show: if initial data are inaccurate enough, the justified results of clustering will not depend on which particular clustering algorithm will be applied, as it should be.

2. Description of generalized clustering algorithm that takes into account the inaccuracy of initial data

Traditional approach to clustering consists of the following implementation stages.

1) Preparing data for clustering.

The array \( \Omega = \{ \omega_1, \omega_2, \ldots, \omega_n \} \) of objects to be clustered is composed at this phase. Each element \( \omega_i \) is associated with a vector of values \( x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \) that is formed by a set of quantitative characteristics (attributes) of \( \omega_i \). The vectors \( x_i, i = 1, 2, \ldots, n \) form an array \( X \).

2) Choosing measure of distance between elements of the array \( \Omega \)

The measure \( \rho(x_i, x_j) \) that allows to quantify degree of proximity of vectors \( x_i \) and \( x_j \) towards each other is chosen at this phase. Since vectors \( x_i \) and \( x_j \) are associated with elements of the array \( \Omega \), the value \( \rho(x_i, x_j) \) allows to make a conclusion on the proximity degree of the corresponding elements \( \omega_i \) and \( \omega_j \).

3) Choosing clustering algorithm

On this stage, a rule \( f(x_i) \) should be defined that assigns to each vector \( x_i \) a number from array \( C = \{1, 2, \ldots, K\} \), which determined the cluster that contains \( x_i \) and, as a consequence, \( \omega_i \). Results of clustering by particular algorithm essentially depend on measure \( \rho(x_i, x_j) \) that is used for calculations (one algorithm can produce different results for different measures \( \rho \) – as well as different methods of clustering that use the same measure \( \rho \)). The choice of a particular clustering algorithm depends in practice on number of analyzed vectors \( n \) and their length \( m \).

4) Clustering

The essence of clustering operation in introduced terms is execution of the mapping \( f: X \rightarrow C \). For each cluster its center’s coordinates \( c_k, k = 1, 2, \ldots, K \), should be determined. Number of clusters \( K \) is given by user and, as a rule, is based on a priori considerations. The evaluation of clustering quality is often the sum of measures of proximity between vectors \( x_i \) and the closest cluster centers:

\[
R(K) = \sum_{i=1}^{n} \rho(x_i, c_{f(x_i)}).
\]

It is proposed to supplement the presented clustering procedure so its results will be corresponded with initial data accuracy if values of the vectors \( x_i \) were uncertain or known with an error. For this, it is necessary at the first stage of clustering to involve information on the possible absolute error \( \Delta x_{ij} \) of values \( x_{ij} \) that are components of the vector \( x_i \). In the most frequent case this will be the error limits \( \Delta_x \):

\[ |\Delta x_{ij}| \leq \Delta_x, i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, m. \]

If values \( x_{ij} \) are measured then the values of \( \Delta x_{ij} \) are known from technical documentation for used measuring instruments. All further phases of clustering should be executed by implementing the approach of paper [17] for propagating the uncertainty of inaccurate initial data through all the computations during their processing or by using the Monte-Carlo method or by any other suitable method. Then the value \( R(K) \) together with the upper bound \( \Delta R(K) \) for absolute value of its error inherited from initial data will be automatically obtained on each step of the clustering.

This approach allows to automatically determining the maximum possible number of clusters \( K \), which would fully agree with accuracy of initial data [18]. The following iterative procedure should be implemented to apply the proposed approach.

1) Assign a value of 1 to the number of clusters \( K \).

2) Execute clustering by chosen method and get the value of \( R(1) \) with the limits \( \Delta R(1) \) of its possible absolute error inherited from the initial data.

3) Assign a value of 2 to the number of clusters \( K \).
4) Execute clustering by chosen method and get the value of $R(K)$ with the limits of its possible absolute error $\Delta_{R(K)}$.

5) If $|R(K) - R(K-1)| \geq \Delta_{R(K)}$, the value $(K+1)$ should be assigned to $K$, then go to step 4. Otherwise, the procedure should be interrupted.

Thus, the idea of the proposed approach is as follows. If the clarification of clustering results caused by increase in number of clusters turns out to be more significant than the results’ error, inherited from inaccuracy of initial data, then the number of clusters should be increased. Otherwise, the clustering procedure should be stopped because of reaching the maximum possible number of clusters that can be recognized in such an inaccurate data. This idea is a particular case of a more general idea: if iterative procedure deals with inaccurate initial data, then it should be stopped earlier in accordance with data accuracy [19]. The rule for stopping the presented sequential clustering procedure completely agrees with the rule presented in paper [20] for solving equations with inaccurate data, and in paper [21] for a well-founded choice of order of mathematical model used to formulate equation of indirect measurements.

3. Setup of a numerical experiment

The study of proposed approach to clustering of inaccurate data was conducted on specific examples. Clustering methods of different types were considered that are often used in practice but are fundamentally different in the principles of obtaining final results. The calculations were performed using a whole range of clustering approaches. This allowed to obtain conclusions that are valid for all types of methods and to confirm the reliability of the proposed generalized approach results. The list of used methods and their main characteristics are presented in Table 1.

The following notations are accepted in Table 1: $n$ is the number of elements of the array $\Omega$ that should be clustered; $K$ is the number of clusters; $N$ is the number of iterations in used clustering method. The second column of the table does not take into account complexity of tree construction algorithm (if it is used) and treats the calculation of distance measure between elements of the set array $\Omega$ as an elementary operation.

| Clustering algorithm                  | Computational complexity | Form of clusters | Results       |
|--------------------------------------|--------------------------|------------------|---------------|
| Single-linkage (hierarchical)        | $O(n^2)$                 | Arbitrary        | Binary tree   |
| $K$-means                            | $O(n \cdot K \cdot N)$   | Hypersphere      | Centroids     |
| Minimum panning tree                 | $O(n^2 \cdot \log n)$   | Arbitrary        | Tree structure|

All methods presented in Table 1 require indicating the number of clusters $K$ as an input parameter. The choice of these algorithms bases on the fact that these methods completely cover classical approaches to clustering data, however, ideas of grouping underlying these approaches are different.

When the mentioned algorithms were applied, two different measures of distances between vectors $x_i$ that should be clustered were used to confirm or disprove influence of the chosen measure on results of clustering inaccurate data. The Euclidean distance

$$\rho(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{i,k} - x_{j,k})^2},$$

which is the most common measure for vectors proximity in practice, and the Dynamic Time Warping (DTW) method [22], which is usually used to calculate distance between time series with possibly different time scales, were chosen.
The data clustered during the numerical experiments were time signals of measurement results of consumed electric power, taken from the database. Their preliminary processing involved performing the transformation $\omega_i \rightarrow x_i$ of the elements $\omega_i$ of the array $\Omega$ that should be divided into groups by degree of similarity into the attributes vector $x_i$.

The easiest way to cluster time series is to use as $x_i$ the values of the series itself, but the length of the vectors to be clustered becomes too large. This affects the laboriousness in the calculation of metrics $\rho$ and it leads sometimes to unreasonable results. Three different algorithms were used to reduce the dimensionality of vectors $x_i$ formed from the processed time series.

1) Singular Spectrum Analysis (SSA).

The main idea of SSA (the classical version of the algorithm is described in [23]) is to decompose a time series into oscillatory components and noise. This allows characterizing the underlying regular dynamical behavior which is the signature of the time signal under study.

The Figure 1 illustrates the described approach with the example of some measured time signal $\omega_i$ to be processed and with corresponding signal $\omega_i^*$ obtained with SSA.

2) Fast Fourier Transform (FFT) plus Principal Component Analysis (PCA).

The algorithms FFT and PCA are well known and widely used in practice. Their classical definitions are described in [24], [25]. In one variant of the performed numerical experiment, the time series $\omega_i$ were processed with the Fast Fourier Transformation, after that the obtained spectra were transformed into vectors $x_i$ of smaller dimension using PCA.

3) Statistical analysis.

In this approach, vectors $x_i$ present the main statistical features of processed signals $\omega_i$. As the components of $x_i$, signal mean value, its maximum and minimum values, standard deviation and coefficients of skewness and kurtosis were taken.

The considered combinations of the listed methods are shown in Table 2.

![Figure 1. The example of the time series $\omega_i$ and $\omega_i^*$.](image)

| Algorithm for reduction the dimension | Metrics          | Algorithm of clustering |
|--------------------------------------|------------------|------------------------|
| SSA                                  | Euclidean distance | K-means               |
|                                      |                  | Single-linkage         |
| FFT + PCA                            | Euclidean distance | Minimum spanning tree |
|                                      |                  | K-means               |
|                                      |                  | Single-linkage         |
Algorithm for reduction the dimension | Metrics | Algorithm of clustering
---|---|---
Statistical characteristics | Euclidean distance | Minimum spanning tree
| | | K-means
| | | Single-linkage
| | | Minimum spanning tree
| | | K-means
| | | Single-linkage
| | | DTW

4. Results

Program in Python for study of proposed generalized approach for clustering of inaccurate data was written using libraries of Numpy, Sklearn, Matplotlib, and Pandas.

To illustrate the actions of numerical experiment, a test calculation that allows convenient visualization of clustering results was performed. A set composed of points on the plane was considered, where their coordinates were taken as their attribute vectors \( \mathbf{x}_i \) as given in Figure 2.

As it can be seen in Figure 2, the processed data can be divided into 3, 4 or even 5 clusters. In order to reasonably determine which number \( K \) of groups is maximum allowable, information about their

Figure 2. Set to be clustered.

Figure 3. Involvement of information about the error of the data to be clustered (the range of possible values).

Figure 4. The values of \( R(K) \) and \( \Delta R(K) \) obtained with presented approach for clustering.

Figure 5. Clustering results.
accuracy should be engaged as given in Figure 3. As it can be noted, when the data are inaccurate, the natural boundaries between the clusters (observed in Figure 2) will be blurred. Thus, it is not necessary to select a large number of clusters. Figure 4 shows the values $R(K)$ of sum of the distances from the processed points to the centers of corresponded clusters. The limits of the error $\Delta R(K)$ obtained according to the procedure mentioned above are also set aside. As it can be seen, the accuracy and the variability of initial data is such that it is not reasonable to distinguish more than three clusters, which agrees with Figure 3. Clustering results obtained in this test example are shown in Figure 5.

The results obtained during the main numerical calculations described in Section 3 are summarized in Table 3.

Table 3. Results of the numerical experiments.

| Algorithm for reduction the dimension | Metrics       | Algorithm of clustering       | Number of clusters determined by proposed algorithm |
|--------------------------------------|---------------|-------------------------------|---------------------------------------------------|
| SSA                                  | Euclidean     | K-means                       | 3                                                 |
| distance                             |               | Single-linkage K-means        |                                                   |
| FFT + PCA                            | Euclidean     | Minimum spanning tree         | 4                                                 |
| distance                             |               | Single-linkage K-means        |                                                   |
| Statistical characteristics          | Euclidean     | Minimum spanning tree         | 6                                                 |
| distance                             |               | Single-linkage K-means        |                                                   |
| –                                    | DTW           | Minimum spanning tree         | 2                                                 |

Table 3 shows that algorithm for extracting feature vectors $x_i$ strongly affects to further clustering and final number of clusters determined by the proposed generalized algorithm. However, it also shows that with the same approach to formation of feature vectors using different clustering algorithms, the same cluster value is detected, which confirms the assumption in paper [18]; results of the complex algorithms performing will coincide with the results of the work of simpler methods, and, therefore, the latter will be preferable for using, if accuracy of the measured signals to be clustered is not high enough.

5. Conclusions

In this paper, the approach to clustering inaccurate data taking into account information on their precision is proposed. The algorithm that allows to determine the maximum possible number of clusters, which can be recognized, and to provide the best agreement with initial data and their accuracy, is proposed. The numerical experiments for applying of the proposed algorithm are presented. It is demonstrated that the proposed generalized approach allows to reasonably determine the number of clusters in analyzed data.

The presented approach corresponds to principle of methodological reductionism of scientific conclusions commonly known as “Occam’s razor”: the proposed procedure allows not to create or specify clusters during the classification of inaccurate data beyond what is necessary.

References

[1] de Souza R M C R and De Carvalho F A T 2004 Pattern Recognition Letters 25 353–365
[2] Asharaf S, Murty M N and Shevade S K 2006 Pattern Recognition Letters 27 515–519
[3] Peng W and Li T 2006 Proc. of ICTAI’06 – 18th IEEE Int. Conf. on Tools with Artificial
Intelligence 11 355–362

[4] Irpino A and Verde R 2008 Pattern Recognition Letters 29 1648–1658
[5] De Carvalho F A T and Tenório C P 2010 Fuzzy Sets and Systems 161 2978–2999
[6] D’Urso P and Leski J M 2016 Pattern Recognition Fuzzy 58 49–67
[7] Ji Z, Xia Y, Sun Q and Cao G 2014 Fuzzy Sets and Systems 253 138–156
[8] Hwang C and Rhee F C H 2007 IEEE Transactions on Fuzzy Systems 15 107–120
[9] Le Hegarat-Masrelle S, Bloch I and Vidal-Madjar D 1997 IEEE transactions on geoscience and remote sensing 35 1018–1031
[10] Denoeux T A 1995 IEEE transactions on systems, man, and cybernetics 25 804–813
[11] Schubert J 2000 Information, uncertainty and fusion Fast Dempster-Shafer clustering using a neural network structure (Boston: Springer)
[12] Parschutin S V 2007 Clustering time series using self-organization maps (Kolomna: Phizmalit)
[13] Jain A, Murty M and Flynn P 1999 ACM Computing Surveys 31 264–323
[14] Ayvazyan S A, Buhschtaber V M, Enukov I S and Meschalkin L D 1989 Applied statistics: classification and reduction of dimension (Moscow: Finance and statistics)
[15] Kotov A and Krasil’nikov N 2006 Data clustering [Klasterizaciya dannykh] (SPB: SPbGU ITMO)
[16] Mandel’ I D 1988 Cluster analysis (Moscow: Finance and statistics)
[17] Semenov K K and Solopchenko G N 2011 Measurement Techniques 54 378–386
[18] Bolschikov V A and Semenov K K 2017 Proc. of 6th All-Russian Conf. “Measurements in the modern world – 2017” 130–134
[19] Semenov K K 2015 Metrological aspects of stopping iterative procedures for inverse problems in static-mode measurements, https://doi.org/10.1142/9789814678629_0039
[20] Semenov K K and Tselishcheva A A 2018 Measurement Techniques 61 203–209
[21] Semenov K K, Maximov V V and Titova N D 2013 Proc. of 4th All-Russian Conf. “Selection of the order of the surface wave model for indirect measurements in hydrodynamics” 119–120
[22] Berndt D J and Clifford J 1994 Using Dynamic Time Warping to Find Patterns in Time Series, https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf
[23] Ustinov S M and Zimnickiy V A 2009 Computational Mathematics (Saint Petersburg: BKHV-Petersburg)
[24] Brigham E O 1988 The Fast Fourier Transform And Its Applications. Signal Processing Series (New Jersey: Prentice-Hall)
[25] Jolliffe I T 2002 Principal Component Analysis, vol 2 (New York: Springer)