The metrologically reasonable clustering of measurement results

K K Semenov ¹, V A Bolschikov ¹

¹Department of measuring informational technologies, Institute of computer science and technologies, Peter the Great Polytechnic University, 29, Polytechnicheskaya str., St. Petersburg 195251, Russia.

E-mail: semenov.k.k@iit.icc.spbstu.ru, vitaliy.bolschikov@gmail.com

Abstract. This paper is dedicated to the metrological aspects of clustering procedure. It is shown that taking into account the information on uncertainty of the data to be clustered allows easily making reasonable decisions that are metrologically supported. The paper presents that the problems that are usually difficult to solve in practice of the clustering become simpler: the proposed approach allows determining the number of maximum recognizable clusters and thus protects from the unreasonable conclusions and allows performing the preconditioning of data to be clustered simpler.

1. Introduction

The clustering of inaccurate data is a problem that usually occurs in practice. Indeed, when some complex object is studied, its different characteristics are measured. The obtained results are used to predict its future behaviour. Examples of such objects can be different: turbines, engines, power units and other complex technological installations, distributed systems with numerous components, human health, environmental processes etc. In many cases, there is no theory that can describe the laws, under which the object of interest acts, or it is no sufficient amount of information to build the mathematical model that exhaustively describes the object. So, it is treated as “black-box” and we only can cluster its different states that can be recognized from the obtained data. After that, for any new observation, we can classify the current object state by assigning to one of the clusters. Such multi-stage cognitive procedure is usually used when the object under study is extremely difficult to be described – like human health, medicine conditions and others.

The inaccuracy of used initial data imposes restrictions on the possible conclusions that can be made: the accurate measurements results of characteristics of object under study allow differing much more object states than if the uncertainty is large. This paper proposes a quantitative method for describing this circumstance and fills the gap in the literature dedicated to clustering: there are enough papers on clustering the inaccurate data – like interval data [1-4], fuzzy variables [5-8], Dempster-Shafer sets [9-11] etc – and there is almost no papers on the determination the number of clusters and the agreement of the final results of clustering with the inaccuracy degree of the initial data [12, 13].

2. The metrological aspects of clustering procedures

Studying different objects, processes, phenomena in practice, all the objective data are obtained from the measurements results. So, all decisions based on it are performed from the uncertain initial
grounds and that’s why they may be unreasonable. The clustering procedure isn’t the exception – some of the discovered clusters can be parts of the one cluster in reality and look separate because of random errors in initial data; some of the identified clusters may consist several unrecognizable smaller subclusters we aren’t able to see – they merged into one since their boundaries are blurred and overlap because of errors of the clustered results of measurement. To detect such situations is extremely difficult problem in practice and is supposed to be solved using the only data to be clustered, which usually doesn’t contain any corresponding information in real-life situations. That’s why to accompany any decision with a characteristic of its quality, we need to take into account the metrological characteristics of data we process. Accounting the information on the initial data uncertainty brings benefits in many cases: we can easily answer some questions that are difficult if we deal with data treated as absolutely accurate. Applying to clustering, the algorithm shown below allows estimating the maximum reasonable quantity of recognizable clusters. To get such number is a hard problem in practice that is often solved subjectively based on some a priori information, but if we have information on metrological characteristics of data to be clustered then this quantity can be easily obtained and used as restriction or reinforcement of conclusions and decisions.

Traditional approach to clustering consists of the following main stages.

1) Collecting the initial data for clustering.

The array \( \Omega = \{ \omega_1, \omega_2, \ldots, \omega_n \} \) of states of studied object, process or phenomenon of our interest is composed at this stage. Each element \( \omega_i \) is described by a vector of values \( x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \) having the one structure for all \( i \). Its elements are a set of measured characteristics of studied object and are disturbed with measurement errors. All measured vectors \( x_i, i = 1, 2, \ldots, n \) form a set \( X \) of observations on studied object, process or phenomenon.

2) Preparing the data before clustering.

The vectors \( x_i \) can be inconvenient to deal with during clustering. So, they can be transformed to vectors \( y_i = (y_{i1}, y_{i2}, \ldots, y_{in}) \) of may be the smaller dimension or that are more comfortable to deal with. The elements of \( y_i \) inherit the uncertainty from the measurements results \( x_i \) and are inaccurate too. All vectors \( y_i \), form a set \( Y \) consisting of principal components of quantitative description of the studied object, process or phenomenon state.

The preconditioning of data can use different and extremely complex algorithms – including for example neural networks and self-organized maps [14].

3) Choosing measure of distance between elements of the set \( Y \).

The measure \( \rho(y_i, y_j) \) that allows quantifying degree of proximity of vectors \( y_i \) and \( y_j \) towards each other is chosen at this phase. Since vectors \( y_i \) and \( y_j \) are associated with elements of the array \( \Omega \) by the chain of transformations \( \Omega \rightarrow X \rightarrow Y \), the value \( \rho(y_i, y_j) \) allows making a conclusion on the proximity of the corresponding elements \( \omega_i \) and \( \omega_j \).

4) Choosing clustering algorithm.

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

5) Clustering.

The clustering procedure in introduced terms has the meaning of the mapping \( f: Y \rightarrow C \). 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 available a priori information. The quantitative expression of quality of performed clustering is usually the sum of measures of proximity between vectors \( y_i \) and the closest cluster centers:
\[ R(K) = \sum_{i=1}^{n} \rho(y_i, c_{f(y_i)}) \]

The listed stages are the main for clustering procedure; in concrete situations it may differ only in some particular details.

We can conclude that the value \( R(K) \) computed from the measurements results \( x_i \) inherits its uncertainty and that’s why is uncertain too. This circumstance creates the possibility for determining the maximum possible number \( K_{\text{max}} \) of clusters that can be recognized from the initial data. To achieve this goal, it is necessary tometrologically accompany all preformed calculations from the list above: for any computation, we should obtain not only the calculation results but also the metrological characteristic of its accuracy. Then we will be able to connect the uncertainty of value \( R(K) \) as the quantitative quality estimation with the accuracy of the initial data – the measurement results \( x_i \). For this, it is necessary at the first stage of clustering procedure to involve information on the possible absolute errors \( \Delta x_i \) of values \( x_i \). In the most frequent practical case this will be the maximum possible error \( \Delta_j \): \( |\Delta i| \leq \Delta_j \), \( i = 1, 2, \ldots n \) and \( j = 1, 2, \ldots m \). If values \( x_i \) are measured then the values of \( \Delta_j \) are known from technical documentation for used measuring instruments. If not only maximum possible error values are available from the manuals, then all other metrological characteristics should be taken into account.

To estimate the metrological characteristics of the results of the calculations with values \( x_i \), we can use one from many approaches and methods that are developed to the present moment to propagate uncertainty of measurement results through the computational procedure. These methods can be grouped in two big sets: methods that use the randomization (like Monte-Carlo approach) and methods that perform automatic analysis by the overloading operations performed during calculations (a linearization of calculated functions with automatic differentiation or interval arithmetic). All of these techniques can be applied to any computational problem; the clustering isn’t the exception. We propose to use the approach of paper [15] for propagating the uncertainty of inaccurate initial data through all the computations because this approach has a metrological orientation, but any other suitable method can be also chosen.

If it is no need in initial data preliminary processing before clustering, then a computationally cheap simple procedure can be used in the case of small uncertainties of initial data. It can be described as follows. Let for each vector \( x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \) of initial data the vector \( \Delta x_i = (\Delta_1, \Delta_2, \ldots, \Delta_m) \) of maximum possible absolute errors of its components be known (from the technical documentation). Since no processing is provided, \( y_i = x_i \) and \( \Delta y_i = \Delta x_i \). Otherwise, we need to use one of the mentioned methods to propagate information on errors \( \Delta x_i \) during calculation the values \( y_i \) and, as a result, estimate their maximum possible absolute errors \( \Delta y_i \). Then the chosen clustering algorithm \( f \) should be applied to the set of vectors \( y_i \) and, as a result, the set of cluster centers coordinates \( c_k, k = 1, 2, 3, \ldots K \), should be obtained and the index \( f(y_i) \) of corresponded cluster should be determined for each vector \( y_i \). The area of possible actual values of vector \( y_i \) can be described as \( y_i \pm \Delta y_i \) and represents the polygon \( P_i \) in space \( R^m \). To estimate the error bound \( \Delta_{K_i} \) for the value \( R(K) \), we should compute the interval \([R_{\text{min}}, R_{\text{max}}]\) of possible values of \( R(K) \) using the following algorithm:

1) Assign a zero value to \( R_{\text{min}} \) and \( R_{\text{max}} \).

2) For each \( i \) from the set \( 1, 2, 3, \ldots n \), check if \( c_{f(y_i)} \in P_i \). If it is so, assign \( \rho(y_i, c_{f(y_i)}) = 0 \) and go to the next value of index \( i \). Otherwise, add to \( R_{\text{min}} \) the minimum value of measure \( \rho \) between \( c_{f(y_i)} \) and the vertices of \( P_i \).

3) For each \( i \) from the set \( 1, 2, 3, \ldots n \), add to \( R_{\text{max}} \) the maximum value of measure \( \rho \) between \( c_{f(y_i)} \) and the vertices of \( P_i \).

The obtained bounds \([R_{\text{min}}, R_{\text{max}}]\) allow easily computing the value of \( \Delta_{K_i} \).

So, using the presented algorithm or using other suitable technique of propagating the uncertainty through the computational procedure, we will automatically obtain the value \( R(K) \) together with the
upper bound $\Delta R_{(K)}$ for absolute value of its error inherited from initial data.

To automatically determine the maximum possible number of clusters $K_{\text{max}}$, the following iterative procedure should be applied.

1) Assign a zero value to the number of clusters $K$ and assign $R(0)=+\infty$.
2) Assign $K$ the value ($K+1$), perform the clustering procedure by chosen method and get the value $\Delta R_{(K)}$ of its possible absolute error inherited from the initial data.
3) If $|R(K)-R(K-1)|<\Delta R_{(K)}$, then go to step 2. Otherwise, the procedure should be finalized and $K_{\text{max}}=K$.

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 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 [16].

The obtaining the value $K_{\text{max}}$ protects us from unreasonable conclusions on clusters structure. The dealing at any step of clustering procedure with uncertainty characteristics creates the opportunity for better applying clustering results for classification of further observations of the state of the studied object, process or phenomenon: we will get the error characteristics $\Delta \xi_i$ for clusters center’s coordinates $\xi_i$ and use it to find out if there are reasons to include the new observation $x_{n+1}$ to several clusters at once. The usual practice is to attribute the vector to the closest cluster, but in reality the distances to clusters’ centers are distorted by errors inherited from initial data used during clustering and this brings us in some cases to situation when we aren’t find out what cluster is the closest one.

Taking into account the uncertainty of measurement results allows us to detect this situation and draw right conclusion.

The other great possibility that is the consequence from the taking into account the uncertainty of initial data is that the proposed approach produces the reasonable and objective way to compare different measures $\rho$ (or clustering algorithms at whole) and to choose the most suitable. Really, the best measure $\rho$ from the set $\{\rho_1, \rho_2, \rho_3, \ldots\}$ is one that for the same data to be processed produces the largest value $K_{\text{max}}$: this means that this measure allows recognizing the maximum number of clusters. At the moment, the choosing of measure $\rho$ is poorly formalized procedure and usually is based on some subjective grounds or is ill-founded. The proposed approach produces the objective reasons for choice. The presented method can be applied for any dataset that should be clustered and allows individual determining the measure of proximity what is the new word in clustering.

3. Using example

Let us examine some metrological situation as an example of using the proposed approach. Let the physical process $\Omega$ be studied that takes place during electrochemical reaction in substance. Let the flow of this process can be described only with two electrical potentials $x_1$ and $x_2$ that can be measured directly. Different modes of process $\Omega$ flow can be recognized varying the external conditions, which $x_1$ and $x_2$ depend on. The measurements of $x_1$ and $x_2$ are performed independently but using the same voltmeter. All obtained values form the vectors $x_i^T = (x_{i1}, x_{i2})$ that are processed in the following manner: the measurements results are scaled to fit the diapason $[-1; 1]$ V. So, using the notation defined in previous section of the paper, the vectors $y_i^T = (y_{i1}, y_{i2})$ are calculated from preliminary data processing performed before clustering procedure applying to determine different states of process $\Omega$.

Let the quantity of measurements results be equal to $n = 40$ and, as before, index $i = 1, 2, \ldots, n$.

Let the described physical process $\Omega$ tend to have $K_{\text{true}} = 4$ separate modes of its flow. In figure 1a, the all obtained vectors $y_i$ are shown. Different colors are used to mark data that are corresponded to different states of the process $\Omega$ and consequently to different clusters. In practice, we almost always
don’t know the actual value $K_{\text{true}}$, so we should assign the maximum quantity of clusters $K_{\text{max}}$ and its value potentially can be equal to any number from 1 to $n$. The concrete value depends on additional information and on some subjective factors – intuition and personal experience of the researcher. If we can’t take them into account, the only way to produce the reasonable results of clustering is to use information on the accuracy of the data to be processed. The figure 1b contains the graphical representation of data and its potential absolute errors’ bounds inherited from used measurement instruments.

Figure 1. Measured initial data to be clustered:

a – without taking into account information on measurements errors
b – with bounds of possible values in accordance with data uncertainty characteristics.

Figure 2. The dependence of maximum number of recognizable clusters $K_{\text{max}}$ on used metric in clustering procedure obtained in three different launches of ‘kmeans’ method.

Let us apply the proposed approach to the data shown in Fig. 1 and take into account measurements uncertainties. As the clustering procedure, the ‘kmeans’ method was chosen because of its popularity in practice. Earlier we suppose that there is no additional information on the nature of the process $\Omega$. So, we need to use the random guesses as the initial approximation of clusters centers’ positions. In
the performed modelling, we use the uniform distribution when covering the data to be clustered with such start points. In this situation, the results are to a certain extent random. For simplicity, the procedure was limited to the number of clusters – no more than 6 could be returned.

In the figure 2, three different results of applying proposed method are shown for different initial guesses for clusters centers depending on the maximum possible error of used voltmeter recalculated to the scale of \( y_1 \) and \( y_2 \) values (marked \( \Delta \) in figure the 2). Three different metrics of proximity for clustered vectors were used: in the figure 2, the red line is corresponded to popular Euclidian metric \( (L_2) \), the blue line – to ‘cityblock’ metric (the sum of absolute values of components of the comparing vectors difference, \( L_1 \)), the black line – to ‘cosine’ metric (the proximity degree is calculated as one minus the cosine of the included angle between comparing vectors).

We see that ‘cityblock’ metric \( (L_1) \) better corresponds to the data to be clustered: it allows recognizing larger quantity of clusters that can be potentially highlighted in the data. The metric ‘cosine’ produces the lowest sensitivity to find out clusters. The important feature of the proposed approach and this model application is that performed analysis is very simple, it doesn’t require significant amount of computations and can produce results that are individual for each concrete practical situation.

4. Conclusions.
The presented paper is dedicated to the metrological aspects of clustering procedure and the consequences of outlined viewpoint. It is shown that taking into account the information on uncertainty of the data to be clustered allows easily making reasonable decisions that are metrologically supported: it becomes possible to determine the maximum number of recognizable clusters; the indicator arises that can be used for reasonable comparing the different metrics used in clustering methods or for comparing clustering algorithms at whole. The paper presents that the problems that are usually difficult to solve in clustering practice become simpler in this case.

References
[1] De Souza R M C R and De Carvalho F A T 2004 Pattern Recogn. Lett. 25 353
[2] Asharaf S, Murty M N and Shevade S K 2006 Pattern Recogn. Lett. 27 515
[3] Peng W, Li T 2006 Proc. ICTAI’06 – 18th IEEE Int. Conf. on Tools with Artificial Intelligence 11 355
[4] Irpino A and Verde R 2008 Pattern Recogn. Lett. 29 1648
[5] De Carvalho F A T and Tenório C P 2010 Fuzzy Set. Syst. 161 2978
[6] D’Urso P and Leski J M 2016 Pattern Recogn. 58 49
[7] Ji Z, Xia Y, Sun Q and Cao G 2014 Fuzzy Set.Syst. 253 138
[8] Hwang C and Rhee F C H 2007 IEEE Trans. on Fuzzy Systems 15 107
[9] Le Hegarat-Mascle S, Bloch I and Vidal-Madjar D 1997 IEEE T. Geosci. Remote Sensing 35 1018
[10] Denoeux T A 1995 IEEE T.Syst. Man Cyb. 25 804
[11] Schubert J 2000 Information, uncertainty and fusion 419
[12] Benderskaya E N 2017 Proc. 2017 XX IEEE Int. Conf. on Soft Computing and Measurements (SCM) 162
[13] Parschutin S V 2007 Integrated models and soft computing in artificial intelligence: collection of scientific papers [Integrirovannye modeli i myagkiye vychisleniya v iskusstvennom intellekte: sbornik nauchnykh trudov] 465 (in Russian).
[14] Yakovenko A A and Malykhina G F 2018 Procedia Comput. Sci. 123 547
[15] Semenov K K and Solopchenko G N 2010 Meas. Tech. 53 592
[16] Semenov K K 2015 Advanced Mathematical and Computational Tools in Metrology and Testing X (AMCTM X) (F Pavese, W Bremser, A G Chunovkina, N Fischer, A B Forbes eds) 10 Series on Advances in Mathematics for Applied Sciences 86 (Singapore: World Scientific) 320-9