Soft measurements of qualitative integral indicators for monitoring quantitative dataset

S V Novikova, E S Kremleva, N L Valitova and A P Snegurenko

1 Kazan National Research Technical University named after A.N. Tupolev (KNRTU–KAI), Kazan, 420111, Russia
2 National Research Ogarev Mordovia State University, Saransk, 430005, Russia

E-mail: SVNovikova@kai.ru

Abstract. In many areas, the criteria for the adoption of certain decisions are not quantitative, but qualitative assessments of some crucial criteria. This applies, in particular, to the integral assessment of the health status of patients, or the assessment of the ecological state of the environment, etc. Therefore, an integral qualitative assessment of an object, based on the known numerical states of its individual elements, is an urgent non-trivial task.

Currently, there are methods for calculating the quantitative integral assessment, which is a kind of “code” for the qualitative assessment. A direct verbal evaluation is obtained, as a rule, by ranking the possible numerical values of the resulting integral code, and assigning to each given interval a certain qualitative definition. However, well-known methods for constructing analytical computational integrated assessment are not effective.

As an alternative, the authors of the article propose an approach consisting in a comprehensive assessment of the state of objects based on the allocation of "similar" groups, and analysis of the basic general properties of objects in the group. Such problems can be solved by cluster analysis methods. Cluster analysis allows you to group (decompose) data that has the property of "similarity" in a given sense and to separate them from "dissimilar" data. By analyzing data in groups, experts can give a high-quality interpretation of homogeneous data in a cluster.

However, when analyzing the obtained groupings by domain experts, a situation often arises when the data in some groups have a sufficient degree of homogeneity in order to give them a qualitative assessment, and in others the mathematical method of clustering does not allow to separate heterogeneous data from each other, and experts cannot give rating.

So the idea to apply step-by-step clustering when in each step it is decided whether every cluster has a sufficient degree of homogeneity or not. If not, all non-homogeneous clusters should be further decomposed into two or more clusters. It remained to decide which clustering method to choose for the data decomposition. To address this problem, reconnaissance experiments were conducted. As a result, the Kohonen Self-Organizing Map (SOM-cards) method has been proved to be the best.

The proposed algorithm for multilevel clustering was called the Cascade Neural Network Filtering Data. Its effectiveness was confirmed by numerical experiments.

1. Introduction

Recently, there has been a rapid development of data mining methods, among which one of the first places is, perhaps, occupied by neural network modeling methods [1-10]. What is usually meant by the term "multivariate data analysis"? As a rule, we mean various methods of cluster analysis, which make...
it possible to find patterns in the data, and, in addition, to determine general, or integral, indicators for their characteristics. [1, 11-13]. Cluster analysis is designed to group, in a sense, similar data, and separate dissimilar data from each other [14]. However, the number of groups, as a rule, cannot be known in advance. The most applicable clustering algorithms today are:

1. Algorithms for clear (hard) clustering:
   - McKean k-means algorithm;
   - Algorithm of weighted groups of Sokal and Michener;
   - Algorithm of self-organizing Kohonen maps [1]
   and etc.

2. Algorithms of fuzzy (soft) clustering:
   - Algorithm FCM (Fuzzy C-Means);
   - Algorithm PCM (Possibilistic C-Means)
   and etc.

However, upon further analysis of the obtained groupings by specialists of the subject area, a situation often arises when the mathematical method of clustering does not allow separating heterogeneous data from each other, and the data in other groups have a sufficient degree of homogeneity. This situation occurs when some of the datasets have much more subtle differences, and the other part of the data is sharply different from each other.

A similar situation arose among researchers, for example, when trying to analyze data from health gadgets (accelerometer, pedometer, etc.). Such analysis is necessary to develop recommendations for changing habits, lifestyle changes, etc. Initially, the data were grouped into three groups:

- No moving,
- Bus ride,
- Stroll.

In this case, the groups "No moving" and "Bus ride" had clear dividing boundaries, but the "Stroll" group contained mixed data about walking, data on movements around the apartment, data on slow cycling and running data. Thus, the detailing of this group was insufficient for a correct classification.

A similar situation arises in many other application areas.

To solve this problem, it is proposed to use the stepwise (cascade) clustering technique developed by the authors, called the cascade filtering technique.

2. General description of the technique

The method of cascading filtering is that a certain clustering algorithm for those groups of data in which the detailing is insufficient from the analyst's point of view will repeat the data decomposition procedure many times. At the first stage, all data a priori refer to a single cluster (group). Next, the group is divided into a given number of clusters, if the data in the cluster, according to the analyst, is in fact not a homogeneous group. Thus, clusters of the first level of the cascade are formed. After that, the data in each of the selected clusters (groups) is analyzed by a specialist. These groups will also be clustered in turn if the analyst discovers pronounced heterogeneity for any of the groups. As a result, clusters of the second level of the cascade are formed, the detail in which is higher than in the first level. The analysis of the groups obtained as a result of clustering and their further clustering (the formation of cascade levels with an increase in the degree of detail) continues until an acceptable (according to the analyst) degree of homogeneity is achieved for all the selected groups.

An analyst can be an automaton or an algorithm that evaluates the degree of homogeneity of data in a group using methods of mathematical statistical analysis, but also a person is a subject area specialist.

The step-by-step process of cascade filtration is graphically presented in Figure 1-3. Figure 1 shows the result of classical non-cascade clustering. Here the input data is grouped into 2 clusters.
The procedure for dividing clusters into subclasses can be schematically shown in Figure 2. You can see that at first the data is not grouped. Then they were divided into 2 groups (clusters), and then each cluster was further divided into 2 sub-clusters.

If more precise detailing is required, the number of cascade levels increases. So, in Figure 3, you can see the process of cascading clustering with four levels. On the first level, as you can see, two clusters are formed, on the second - four. Let the analysis of the homogeneity of these clusters show that only subcluster 2.2, belonging to the second level, is not sufficiently homogeneous. Therefore, further only this is divided, forming two subgroups at the third level and two at the fourth.

The procedure for the emergence of new levels can be continued. Only when it is determined for all subclusters of the next level that they are all sufficiently homogeneous, the process will end.

3. Analysis of various clustering algorithms

The technique presented by the authors in this article allows for different ways of splitting a set into subsets (branching). To find the most suitable method, the effectiveness of several cluster analysis algorithms was investigated, which represent three main types of applied cluster analysis methods:

1. Hierarchical clustering algorithm (the number of clusters is not known in advance);
2. Intelligent Kohonen algorithm (based on a neural network);
3. Algorithm k-means (and the number of clusters is predefined).

For the pilot study, two different datasets were generated:

1) Dataset obtained from health gadgets. There are 304 data lines in the dataset. Each line includes measurements for seven indicators. The data is normalized before clustering. It is assumed that the data should be divided into 4 types: walking, sitting, lying awake, sleeping. Initially, for all dataset rows, it is not specified what kind they belong to.

2) Dataset of measurements of the level of water salinity in rivers and lakes of one of the regions of the Russian Federation. The dataset will include 210 lines. Each line contains 27 parameters of water hardness. Before the start of clustering, as well as for the previous dataset, the data is normalized.
normalized. It is assumed that the data should be divided into 4 classes: low, medium, above average and high mineralization.

![Figure 3. Schema with four cascade level.](image)

Experts were invited to check the accuracy of each of the three clustering algorithms under study. They studied all the data lines in advance, and determined which class (group) it really belongs to.

Although there are no class labels in the clustering tuples, it is assumed that the clustering algorithm can correctly classify the data.

Based on these quality indicators [15] clustering, a comparative assessment of clustering is made:

- **Separability**: This condition means that we are trying to increase the distance between different clusters. We have three methods at our disposal to determine the distance: the distance between the nearest elements of two clusters, the distance between the farthest elements of two clusters and the distance between the conditional centers of two clusters.

- **Concentration**: This means that we want the elements of the cluster to be located as close to its center as possible.

- **Compactness**: This means that the elements of the cluster should be located as close to each other as possible. This can be verified by measuring the distance between all pairs of elements in the cluster or the density within the cluster.

For the quality metrics themselves, the classification is used [16]:

- **External** - you don't need to know the cluster structure in advance to use this. The metric is calculated based on the information originally contained in the dataset.

- **Internal** - you should have information about the cluster structure in the dataset.

- **Relative** - you should calculate such a metric based on a comparison of properties of multiple clusters. In this case, you must have information about the structure of the clusters and about the original dataset.

We will use the internal metric «FM index» (Folkes and Mallows) [17]. Let's calculate it in the following way. Consider pairs \((x_i, x_j)\) of elements \(X\). Let's count the number of pairs in which:
a) elements belong to one cluster and one class: SS;
b) elements belong to the same cluster, but to different classes: SD;
c) elements belong to different clusters, but to the same class: DS.

Then the FM index is defined as:

\[ FM = \sqrt{\frac{SS}{SS+SD}} \cdot \frac{SS}{SS+DS} \]  
(1)

This estimate changes in the range \([0,1]\). The lower the FM index value, the better. This means that the differences between the different clusters are large.

We will also use the relative SWC (silhouette rating index). Let's calculate this for each cluster as follows: let's say the element \(x_j\) belongs to the cluster \(c_p\). Let's denote the average distance from this object to other objects from the same cluster \(c_p\) through \(a_{pj}\). Now let us denote the average distance from \(x_j\) to objects from another cluster \(c_q, q \neq p\) by \(d_{qj}\). Put \(b_{pj} = \min_{q \neq p} d_{qj}\). The meaning of this value can be defined as a measure of the dissimilarity of an individual element with the elements of the nearest cluster. Thus, the “silhouette” of each individual element is defined as:

\[ S_{x_j} = \frac{b_{pj} - a_{pj}}{\max(b_{pj}, a_{pj})} \]  
(2)

Here division performs the function of normalization. You can see that the larger the value of the number \(S_{x_j}\) the more accurately the element \(x_j\) to the cluster \(p\). The assessment for the entire cluster structure is achieved by averaging the indicator for the elements:

\[ SWC = \frac{1}{N} \sum_{j=1}^{N} S_{x_j} \]  
(3)

The best split has the highest SWC value.

4. Computational experiments (preliminary)

In order to better understand the structure of the original datasets, visualization is helpful. To do this, we used the TSNE algorithm (t-distributed stochastic embedding of neighbors, stochastic embedding of neighbors with Student’s distribution). This algorithm allows you to represent large-scale data in two-dimensional format without losing the general properties of the data. The algorithm represents the proximity of each pair of points in the original multidimensional space as the probability that one data point is associated with another point as its neighbor.

In figure 4, you can see a two-dimensional T-SNE-simplification of the seven-dimensional human mobility data space. Here, data tuples related to the activity type "walking" are marked in green, "sitting" - in red, "lying awake" - in orange, and "sleep" - in blue:

![Figure 4. 2D-visualization of seven-dimensional dataset (people mobility).](image-url)
The Figure shows that the division of the elements of this set has a complex structure.

In Figure 5, you can see a two-dimensional projection of a multidimensional water salinity dataset. Here, the datasets with high salinity are colored green, above average is red, middle is orange, and low values is blue.

![Figure 5. Visualization of a dataset on surface water pollution.](image)

It can be seen that the data is initially distributed among the groups in a fairly ordered manner. The complexity of clustering such a combination of tuples is expectedly lower than in the previous case.

We tested each clustering algorithm on two datasets. Each time we wanted the data to be divided into 4 groups (clusters). Evaluation of the separation efficiency was carried out according to the following criteria:

- FM (Folkes and Mallows) coefficient — the lower the coefficient, the more different the clusters are from each other.
- SWC coefficient (silhouette coefficient) — the higher the value, the better the structure within the clusters;
- accuracy (percentage of correctly classified data) — the higher the value, the better the quality of the classifier;

The results are shown in table 1.

### Table 1. Results of testing models.

| Model                          | Data       | Mobility data | Mineralization data |
|--------------------------------|------------|---------------|---------------------|
|                                |            | FM  | SWC | accuracy | FM  | SWC | accuracy |
| Kohonen SOM networks           | Mobility   | 0.1 | 0.69| 90.78%    | 0.271| 0.69| 0.1     |
|                                | Mineralization | 0.271| 0.474| 78.32%    | 0.24 | 0.474| 0.274   |
| K – medium                     | Mobility   | 0.317| 0.318| 79.27%    | 0.559| 0.318| 0.317   |
|                                | Mineralization | 0.317| 0.474| 78.32%    | 0.24 | 0.474| 0.274   |
| Hierarchical agglomerative clustering | Mobility   | 0.274| 0.474| 78.32%    | 0.24 | 0.474| 0.274   |
|                                | Mineralization | 0.274| 0.474| 78.32%    | 0.24 | 0.474| 0.274   |

First, consider the results for datasets on water salinity (the data is initially linear). In this case, we obtained the highest accuracy by applying intelligent clustering using neural networks (Kohonen networks). The FM and SWC coefficients have the best values for the hierarchical clustering algorithm. Next, let’s look at data from health gadgets. This data has a complex non-linear structure. For this dataset, the Kohonen network algorithm performed best in all respects. Given that the difference between the FM and SWC coefficients for agglomerative clustering and Kohonen networks for the water salinity dataset is less than 2% for FM and 7% for SWC, while the accuracy of the SOM
algorithm is 11% higher, it can be argued that the highest efficiency in this experiment was demonstrated by the clustering algorithm by Kohonen's self-learning neural networks. This algorithm will be chosen for the final formulation of the developed algorithm.

5. Algorithm for constructing a cascade neural network filtering model

We introduce the following notation:

- \( N \) is the dimension of the input vectors of data tuples, on the basis of which the model is built (clustered data).
- \( l \) - level number of the filtering cascade
- Each cluster of the cascade at level \( l \) is characterized by:
  - Own identifier, i.e. the sequence number \( v \);
  - Ancestor cluster number. An ancestor cluster is specified by a pair of indices \( pt \), where \( p \) is the number of the cluster generating the ancestor cluster, \( t \) is the number of the cluster generating the current one.

- \( P^l \) is the set of identifiers of clusters of level \( l \) for which there exist descendants.
- \( K^l_{pt}(v) \) is the cluster of the \( l \)-th level of the cascade with number \( v \) generated by the ancestor cluster \( pt \) of the previous level.

So, the algorithm can be formulated as follows.

![Algorithm for cascade constructing](image-url)

**Figure 6.** Algorithm for cascade constructing.
While running this algorithm, a cascade filter model will be built. The input vector is \( X = (x_1, x_2, \ldots, x_N) \). It contains the values of attributes for an object. Subsequently, the constructed model will be able to identify each new object to the corresponding cluster. We expect that the model will be able to say at the output that the given vector \( X \) belongs to the cluster \( K_{p(v)} \).

6. Computational experiments (main)

We have implemented several experiments using the developed algorithm to investigate the efficiency and adequacy of the cascade neural network filtering model. We used the same datasets that we used in previous amplification experiments:

1) Data from wearable health gadgets.
2) Data of measurement for water salinity level.

In order to investigate the capabilities of the model as strictly as possible, we made an assumption:
- We will assume that the cluster is heterogeneous (the cluster granularity is not determined), it needs to be subdivided into subclusters in any case.
- During branching, each cluster is always divided into two subclusters of the lower level of the cascade.
- The number of cascade levels is set in advance: first layer - 2 clusters, second layers - 4 clusters, third 8 clusters and fourth - 16 clusters.

These assumptions allowed us to more accurately track the quality of generated clusters and the adequacy of the calculations.

The results of studying the properties of the obtained clusters are shown in Table 2 (averaged values for all clusters).

| The number of layers SOM-cascade | Datasets | Mobility data | Mineralization data |
|---------------------------------|----------|---------------|---------------------|
|                                 |          | accuracy | SWC | FM | accuracy | SWC | FM |
| 1                               |          | 61.75%   | 0.419 | 0.466 | 60%       | 0.407 | 0.481 |
| 2                               |          | 85.2%    | 0.489 | 0.21  | 85%       | 0.472 | 0.273 |
| 3                               |          | 88.59%   | 0.53  | 0.13  | 86%       | 0.477 | 0.271 |
| 4                               |          | 90.78%   | 0.69  | 0.1   | 85.83%    | 0.472 | 0.273 |

The table shows that with an increase in the number of layers in a cascade, the accuracy of classification by dataset from health gadgets increases. This confirms the effectiveness of the developed model.

At the same time, for a dataset about water salinity after the third level, the quality indicators begin to fall (accuracy, FM and SWC indices). Realizing this, we performed additional experiments, increasing the number of cascade levels to seven.

Based on the results obtained, a graph of the dependence of the percentage of correctly classified data on the number of cascade layers was built for the two considered data sets (Figure 7).

It is noteworthy that the maximum accuracy for the data set on water salinity was obtained already at the third level of the cascade. For the physical activity dataset, it is fifth. A further increase in the levels of clustering does not lead to an increase in accuracy.

The reason for this behavior was revealed when analyzing the composition of the clusters: as the cascade grows with the number of clusters, the number of data tuples in each cluster decreases, this fact leads at the next stage to their lack and the generation of empty clusters. Since the number of tuples in the set with data on the salinity of water bodies is less than in the set of data on physical activity - 210 and 304, respectively, the drop in accuracy with increasing cascade in the first set is observed faster.
Figure 7. Changes in the neural filtering model accuracy of the with an increase in the cascade levels number.

Thus, the effectiveness of the developed model of cascade neural network filtering was confirmed by the results of numerical experiments, which, in turn, determine the area of such efficiency: the number of cascade layers, and hence the number of clusters for splitting the initial data set, should not lead to degeneration of clusters and the appearance of empty groups.

7. Conclusion
To cluster data with the required level of detail, the developed algorithm for constructing a cascade neural network filter model can be successfully applied. When using the model, a restriction will be imposed — the power of the training data set, according to which the model is built according to the developed algorithm: the number of filter stages, and, as a consequence, the number of partitioning clusters, should not allow the existence of empty clusters.

References
[1] Kohonen T 2013 Essentials of the self-organizing map Neural Networks 37 52–65
[2] Novikova S V 2016 Structural optimization of the neural network model for the gas turbine engine monitoring Russian Aeronautics 2 263-70
[3] Tunakova Y, Novikova S, Ragimov A, Faizullin R and Valiev V 2017 A method for assessing the retention of trace elements in human body using neural network technology Journal of Healthcare Engineering 3471616 7
[4] Tunakova Y, Novikova S, Krasnyuk I, Faizullin R and Gabdrakhmanova G 2018 Neural network self-learning model for complex assessment of drinking water safety for consumers BioNanoSci 8 504-510
[5] Valitova N and Novikova S 2013 Environmental emergencies prediction based on time series measurements of pollutants concentration in the atmosphere Actual Problems of Computer Science 1(3) 7-18
[6] Emaletdinova L Y, Matveev I V and Kabirova A N 2015 Method of constructing a neural regulator for the automatic one-dimensional control of a technical object Russian Aeronautics 2 227-232
[7] Anikin I V, Makhmutova A Z and Gadelshin O E 2016 Symmetric encryption with key distribution based on neural networks 2nd International Conference on Industrial Engineering, Applications and Manufacturing 2 7911640
[8] Emaletdinova L Y, Matveev I V and Kabirova A N 2017 Method of designing a neural controller for the automatic lateral control of unmanned aerial vehicles Russian Aeronautics 3 365-373
[9] Katasev A S, Kataseva D V and Emaletdinova L Y 2016 Neuro-fuzzy model of complex objects approximation with discrete output 2nd International Conference on Industrial Engineering, Applications and Manufacturing 2 7911653
[10] Ismagilov I I, Khasanova S F, Katasev A S and Kataseva D V 2018 Neural network method of
dynamic biometrics for detecting the substitution of computer Journal of Advanced Research in Dynamical and Control Systems 10 1723-28

[11] Hou J, Gao H and Li X 2017 Feature Combination via Clustering Transactions on Neural Networks and Learning Systems 99 1-12

[12] Robert J 2016 Lecture notes on cluster algebras Bulletin (New Series) of the american mathematical society 2 325–30

[13] Ulrike von L, Belkin M and Bousquet O 2008 Consistency of spectral clustering Ann. Statist 2 555–86

[14] Kaufman L and Rousseau P J 2005 Finding groups in data: An introduction to cluster analysis (John Wiley & Sons, Inc., Hoboken, New Jersey) 355

[15] Ultsch A and Loetsch J 2017 Machine-learned cluster identification in high-dimensional data Journal of biomedical informatics 66 95-104

[16] Desgraupes B 2013 Clustering Indices University of Paris Ouest - Lab Modal X 1-34

[17] Gan G, Chaoqun M and Jianhong W 2007 Data clustering: theory, algorithms, and applications ASA-SIAM Series on Statistics and Applied Probability 20 488