The Classification of GIS Objects

Yu N Mironova¹
¹Kazan (Volga region) Federal University, Elabuga Institute, Elabuga, Russia

E-mail: mironovajn@mail.ru

Abstract. This paper discusses the current issues of the application of classification and data processing in geoinformation systems. The problems of classification of various objects have been studied in the works of many authors. These include a fairly wide range of problems: decryption of satellite images, pattern recognition, mathematical modeling, etc. In this paper, we study the methods and techniques for classifying objects listed in the literature, as well as preliminary data processing: feature normalization, feature weighting, aggregation, dimensionality reduction, etc. The result of finding spatial features in an attribute space is often a representation of spatial features in the form of an object-feature matrix that reflects the measurement of M features on N spatial features and contains N rows and M columns. To classify spatial objects, you must have a geographical map of these objects and an object-attribute matrix, the rows of which correspond to the spatial objects. In order to properly classify, you need to perform pre-processing of the data, including normalization, weighting, dimensionality reduction, aggregation, and identification. After preliminary data processing, the objects are classified. The paper lists and describes such classification methods as nuclear classification methods, hierarchical divisive classification methods, hierarchical agglomerative classification methods, near neighbor method, far neighbor method, centroid method, group mean method (mean link method) and other issues related to the classification of geoinformation objects.

1. Introduction

The development of computer technology and data analysis methods in the second half of the XX century brought classification to a qualitatively new level. Many sciences developed independently of each other. Therefore, in publications on classification issues, you can find special cases of methods that were already known and systematized in applied statistics at that time.

This paper discusses the current issues of the application of classification and data processing in geoinformation systems.

Materials from open sources were used in the work: textbooks and teaching aids [1], [2], [3], [4], the author's dissertation abstract [8], articles on geoinformatics by Russian authors [6], [7], the author's works on this topic [13] - [27], as well as works by foreign authors [9], [10], [11], [12]. Methods such as the analysis and comparison of data from various sources were used.

The problems of classification of various objects have been studied in the works of many authors. These include a fairly wide range of problems. So, if we consider recent work in this area, the problem of decoding space images are described in the works of Smirnov L. E. [3], Kapralov E. G., A. V. Koshkarev, Tikunov, V. S., etc. ([1], [5]), Bulgakov S. V. [6] the study method, interpretation of satellite images considered in the paper of A. I. Nazmutdinova [8], the use of mathematical methods in...
Geoinformatics and classification of dedicated work Mironova Yu. N. [13] - [27]. There are materials on this topic on the Internet, as well as in foreign sources [9] - [12].

Classifications in geoinformatics have some specific features peculiar to this science [1].

First, the object of classification in geoinformatics is spatially coordinated objects that have the characteristic of location. Its presence allows you to take into account the coordinates of objects and their relative positions when analyzing them.

Secondly, the map is a powerful tool for providing information in the form of a single image, which makes it possible to interactively set the initial conditions of the analysis, display its progress and results.

The purpose of the work is to study the methods of classification of geoinformation objects, which take into account the structural features of objects and with high accuracy make a decision about whether an object belongs to a certain class.

In this paper, a study of the methods and techniques for classifying objects listed in the literature, as well as pre-processing of data: normalization of features, weighting of features, aggregation, dimensionality reduction, etc.

2. Pre-Processing of Data

Spatial objects (or operational-territorial units – OTE [1]) are of the following types:

- administrative-territorial units;
- localities;
- cells of a regular or irregular grid superimposed on the study area;
- raster cells, etc.

Despite the geographical nature, spatial objects also have content characteristics - attributes. The result of finding spatial features in an attribute space is often a representation of spatial features in the form of an object-feature matrix that reflects the measurement of N features on M spatial objects and contains N rows and M columns [1].

\[
X = \begin{bmatrix}
O_1 & \ldots & O_M
\end{bmatrix} = (o_1^{(1)}, \ldots, o_1^{(m)})
\]

where \( O_i = (o_i^{(1)}, \ldots, o_i^{(m)}) \) - i-th spatial object in the N-dimensional feature space; \( o_j^{(j)} \) - j-th attribute, \( o_j^{(j)} = (o_{j_1}^{(j)}, \ldots, o_{j_N}^{(j)}) \); \( o_j^{(j)} \) - the value of the j-th attribute on the i-th spatial object, \( i = \{1, \ldots, N\}, j = \{1, \ldots, M\} \).

A fragment of the feature object matrix may look like this:

To classify spatial objects, you must have a geographical map of these objects and an object-attribute matrix, the rows of which correspond to the spatial objects. In order to properly classify, you need to perform pre-processing of the data, including normalization, weighting, dimensionality reduction, aggregation, and identification.

During the pre-processing, you need to determine the type of each attribute. Usually, there are quantitative and qualitative signs. Qualitative characteristics are divided into ordinal (rank) and nominal (columns 3 and 2 of Table 1, respectively).
Table 1. Subjects of the Russian Federation.

| Name of the subject of the Russian Federation | Status of the subject of the Russian Federation | Level of education | Population (people) |
|---------------------------------------------|-----------------------------------------------|-------------------|---------------------|
| Moscow city                                 | high                                          | 12 615 88         |
| Omsk region                                | average                                       | 1 944 225         |
| Buryatia Republic                          | low                                           | 983 276           |

2.1. Normalization of features

Normalization is used to bring the analyzed features to a single scale and reference point. There are the following types of rationing:

- **Normalization by a given indicator**:

  \[ o^{(j)}_i = \frac{o^{(j)}_i}{o^{(j)}_{i,\text{norm}}} \quad \forall i = \{1, \ldots, N\} \]

  where \( o^{(j)}_i \) - source attribute, \( j = \{1, \ldots, N\} \), \( o^{(j)}_{i,\text{norm}} \) - the fiducial sign.

The normalizing feature can be the total population of a spatial object, the area of the territory occupied by the spatial object, and other criteria. Examples of the use of such normalization: population density, the country’s GDP per capita, etc.

- **Нормировка по дисперсиям и математическим ожиданиям**:

  \[ o^{(j)}_i = \frac{\bar{o}^{(j)}_i - \bar{o}^{(j)}}{\sqrt{\text{var}(o^{(j)}_i)}} \quad \forall i = \{1, \ldots, N\}, j = \{1, \ldots, M\} \]

  where \( \bar{o}^{(j)}_i = \frac{1}{N} \sum_{i=1}^{N} o^{(j)}_i \) - estimation of the mathematical expectation of the \( j \)-th indicator,

  \( \text{var}(o^{(j)}_i) = \frac{1}{N} \sum_{i=1}^{N} (o^{(j)}_i - \bar{o}^{(j)})^2 \) - estimation of the variance of the \( j \)-th indicator.

This normalization is used in the case when the analyzed features have a different nature or are incommensurable.

- **Normalization by best or worst values**:

  \[ \bar{o}^{(j)}_i = \frac{o^{(j)}_i - c}{o^{(j)}_{\max} - o^{(j)}_{\min}} \quad \forall i = \{1, \ldots, N\}, j = \{1, \ldots, M\} \]

  where \( o^{(j)}_{\min} = \min \{o^{(j)}_i \mid i = 1, \ldots, N\} \), \( o^{(j)}_{\max} = \max \{o^{(j)}_i \mid i = 1, \ldots, N\} \), \( c = o^{(j)}_{\min} \).

The numbers obtained as a result of this normalization are bounded by the segment \([0,1]\), which is often convenient for further data processing.
2.2. Weighting of characteristics

The process of finding weights for indicators, or weighing indicators, is necessary for the correct implementation of geoinformation and other classifications. The weights are usually set in an expert way and are numbers that do not exceed one in modulus. If the weights are set for a certain feature system by the vector \( \mathbf{\omega} = \{\omega_1, \omega_2, \ldots, \omega_M\} \), then the values of the features are recalculated according to the formula

\[
\overline{\sigma}_i^{(j)} = \omega_j \sigma_i^{(j)} \quad \forall i = \{1, \ldots, N\}, \ j = \{1, \ldots, M\}.
\]

2.3. Aggregation

Aggregation in the simplest form is one of the methods of moving from a set of initial indicators to a single one, by which you should distinguish between spatial objects.

In geoinformation research, the following methods are usually used to obtain a single resulting feature \( o_{\text{agr}}^{(agr)} = (o_1^{(agr)}, \ldots, o_N^{(agr)}) \):

- **Summing up** the values of pre-normalized and weighted indicators:

\[
o_i^{(agr)} = \sum_{j=1}^{M} \overline{\sigma}_i^{(j)}
\]

- **Calculation of distances** in a multidimensional feature space to the best or worst (conditional) spatial object \( o' \):

\[
o_i^{(agr)} = d(o_i, o')
\]

2.4. Reducing the dimensionality of

When analyzing multidimensional data, we usually consider not one, but many tasks, in particular, choosing dependent and independent variables in different ways. Therefore, we consider the problem of reducing the dimension in the following formulation [2]:

A multidimensional sample is given. It is necessary to move from it to a set of vectors of a smaller dimension, preserving the structure of the original data as much as possible, without losing the information contained in the data. The problem is specified in the framework of each specific method of reducing the dimension.

- **The method of principal components**

The principal component method is one of the most commonly used dimensionality reduction methods. Its main idea is to consistently identify the directions in which the data has the greatest variation. In the principal component method, first of all, the direction of the maximum spread is found. Then, in terms of linear algebra, we consider a hyperplane in \( n \)-dimensional space that is perpendicular to the first principal component, and project all the elements of the sample onto this hyperplane. The dimension of the hyperplane is 1 less than the dimension of the original space.

In the hyperplane under consideration, the procedure is repeated. It finds the direction of the largest spread, i.e. the second main component. Then a hyperplane perpendicular to the first two principal components is distinguished. Its dimension is 2 less than the dimension of the original space. Next-the next iteration.

From the point of view of linear algebra, we are talking about the construction of a new basis in an \( n \)-dimensional space, the orthonormality of which are the main components.

The variance corresponding to each new principal component is less than for the previous one. Usually stop when it is less than the specified threshold. If \( k \) principal components are selected, it means that it was possible to move from the \( n \)-dimensional space to the \( k \)-dimensional one, i.e. to reduce the dimension from \( n \) to \( k \), practically without distorting the structure of the original data.

- **Factor analysis**
The principal component method is one of the methods of factor analysis. The different algorithms of factor analysis are united by the fact that in all of them there is a transition to a new basis in the original n-dimensional space. A new idea compared to the principal component method is that the factors are divided into groups based on the loads. The factors that have a related influence on the elements of the new basis are combined into one group. Then it is recommended to leave one representative from each group. Sometimes, instead of choosing a representative by calculation, a new factor is formed, which is central to the group under consideration. The reduction in dimension occurs when switching to a system of factors that are representatives of groups. The remaining factors are discarded.

2.5. Identification
Before you start classifying multidimensional data in GIS, you need to establish a relationship between the external table and the map layer in order to uniquely map the OTE-point of the multidimensional feature space to the OTE-object, that is, some spatial object. Most often, the identification problem is solved manually. Physically, this means establishing a relationship between an attribute table and an external table, which is a standard database operation.

3. Classification methods
According to the degree of consideration of geographical space, classification methods are divided into those that use and those that do not use geographical space.

By the presence of a training sample:
• with a teacher;
• without a teacher.

By the presence of a priori information about the statistical properties of classes:
• parametric;
• non-parametric.

The purpose of parametric methods is to find unknown parameters of known distribution functions of spatial objects within the boundaries of each class and the probability of occurrence of these classes.

Nonparametric methods are usually divided into hierarchical (agglomerative and divisive) and non-hierarchical (most often nuclear).

It is practically important to know the specifics of a particular classification method, the conditions for its application, as well as the specified parameters.

3.1. Calculating distances between spatial objects
A common parameter for hierarchical and nuclear algorithms is the way the metric is set. The need to measure the distance between spatial objects is dictated by the definition of the classification problem.

To calculate the distance on M quantitative features, use:

• euclidean distance

\[ d_e(o_i, o_j) = \sqrt{\sum_{x=1}^{M} (o_{ix} - o_{jx})^2} \]

• manhattan distance

\[ d_h(o_i, o_j) = \sum_{x=1}^{M} |o_{ix} - o_{jx}| \]

When conducting a practical analysis, the following rules must be followed:
• the application of the Euclidean distance is theoretically justified in the case of a normal distribution of features, independence of features, and equality of feature variances;
• the Manhattan distance is logical to apply in the case of discrete feature values;
• the use of corellia-type distances is used for a large number of features.

3.2. Nuclear classification methods
It is assumed that there are "cores" - some real or hypothetical spatial objects that are typical representatives of their class across the entire set of features. The initial cores can be found using a training sample, selected randomly, or using heuristic algorithms. At each iteration of the algorithm, the selected cores are sequentially joined by the nearest spatial objects from the point of view of the selected metric.

The criterion for stopping the nuclear classification algorithm based on the minimality of distances is to view and assign to the cores all spatial objects that are not cores.

The "k-means" kernel algorithm, after classification, recalculates the kernels by finding a new most typical feature in each generated class, and continues classification until the kernels stop changing.

An example is the well-known "Trout" algorithm, in which the position of the cluster center gradually improves. In this algorithm, at each step, build a ball of certain radius in advance, select the elements klasterizatsii together within the globe, and the new center of cluster is built as the centre of gravity of the selected items [2].

Note that the cluster analysis algorithms can be modified in a variety of ways. For example, describing the "Trout" algorithm in the style of statistics of objects of non-numeric nature, we note that calculating the center of gravity for a set of multidimensional points is finding the empirical mean for a measure of proximity equal to the square of the Euclidean distance. If we take a more natural measure of proximity - the Euclidean distance itself, we get the "Median" cluster analysis algorithm, which differs from the "Trout" in that the new center is constructed not using the arithmetic mean coordinates of the elements that fall into the cluster, but using the medians.

An additional parameter of the nuclear algorithms is the number of classes that need to be generated.

3.3. Hierarchical divisive classification methods
Hierarchical divisive classification methods consist in dividing one of the previously obtained classes into several new classes (usually two) at the next iteration. The stop criterion is either that the specified number of classes is reached, or that each feature corresponds to a separate class.

3.4. Hierarchical agglomerative classification methods
Hierarchical agglomerative classification methods consistently combine two or more existing classes into a larger class at each iteration. The criterion for combining classes is their proximity. After getting a new class, all distances from it to other classes are recalculated.

The choice of inter-class distance is very important here and is often included in the name of the algorithm. Most often, the following types of inter-class distances are used:
- **Near Neighbor method**
  The distance between two classes is calculated as the distance between the two closest spatial features of these classes:
  \[ D_{\text{min}}(S_i, S_j) = \min \{d(o_{ix}, o_{iy}) | o_{ix} \in S_i, o_{iy} \in S_j \} \]

  The disadvantage of the method is that if there are anomalous observations in the sample (i.e., those that differ significantly in their values of indicators from other spatial objects of this class) they will be placed in separate classes, and the main group of features will merge into one class. Thus, the method should be used in the absence of anomalous observations.
- **Far Neighbor method**
  The distance between two classes is calculated as the distance between the two farthest spatial objects of these classes:
This method is more resistant to the formation of a single class, and tends to form classes that are commensurate in the number of spatial objects.

- **Centroid method**
  The distance between two classes is calculated as the distance between the centers of the classes:
  \[
  D_{\text{cen}}(S_i, S_j) = d(\bar{o}_i, \bar{o}_j)
  \]

- **Group Average method (average link method)**
  The distance between the two classes is calculated using the formula:
  \[
  D_{\text{avg}}(S_i, S_j) = \frac{1}{N_i N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} d(o_{ix}, o_{iy})
  \]

Hierarchical classification methods are more time-consuming than nuclear algorithms. The advantage of such methods is that they can be applied without having a priori information about the properties of classes (for example, class kernels or training samples).

Classification of spatial objects by a single attribute is a visualization tool and is built into most vector GIS (ArcView, Mapinfo, etc.). Functions for classifying spatial objects simultaneously by several attribute features are less common (GIS Cluster 2.0, etc.).

### 4. Conclusion

Thus, we have considered the various classification methods currently known, as well as the data processing methods necessary for conducting classifications. It is especially important to choose the correct classification option when classifying objects that are characterized by a high degree of uncertainty, such as geoinformation objects or the results of statistical studies.

### 5. References

[1] Kapralov E G, Tikunov V S, Zavarzin A V et al 2009 Collection of tasks and exercises in geoinformatics: textbook. manual for students. higher. studies. institutions ed by V S Tikunov 2nd ed., reprint. and add M.: Publishing center "Academy" p 512

[2] Orlov A I 2004 Applied statistics M.: Publishing House "Exam" URL: http://www.aup.ru/books/m163/

[3] Smirnov L E 2005 Aerospace methods of geographical research: Textbook SPb.: St. Petersburg University press p 348

[4] Yakhyaeva G E 2006 Fuzzy sets and neural networks: a Textbook Moscow: Internet University of Information technologies; BINOM p 316

[5] Kapralov E G 2010 Geoinformatics: in 2 books. kN. 1: textbook for students. higher. studies'. institutions' under the editorship of V S Tikunova 3rd ed., Rev. and DOP M.: Publishing center "Academy" p 400

[6] Bulgakov S V 2013 Fundamentals of geoinformation modeling Izvestiya vysshikh uchebnykh zavedenii Geodesy and aerial photography 3 pp 77-80

[7] Skvortsov A V, Sarychev D S 2002 Technology of construction and analysis of topological structures for geoinformation systems and computer-aided design systems Vestnik of Tomsk state University 275 pp 60-63

[8] Nazmutdinova A I 2016 Development and research of method of interpretation of satellite images of area landmarks based on wavelet analysis The author's abstract Diss. for the degree of Cand. Techn. sciences' (Izhevsk)

[9] Jing Zhang, Jia Zhang, Xiangyang Du, Kang Hou, Minjuan Qiao 2017 An overview of ecological monitoring based on geographic information system (GIS) and remote sensing (RS) technology in China IOP Conf. Series: Earth and Environmental Science 94 012056
[10] Risky Yanuar S, Wahyu Nurbandi, Raden Ramadhani Yudha A, Brada Irmaning T, Artanti Prisma Z, Rosyita Alifaya, Wisudawan Putra D and Sudaryatno 2018 Using Remote Sensing and Geographic Information System (GIS) for Peak Discharge Estimating in Catchment of Way Ratai, Pesawaran District, Lampung Province *IOP Conf. Series: Earth and Environmental Science* **165** 012032 DOI :10.1088/1742-6596/165/1/012032

[11] Lubis M Z, Taki H M, Anurogo W, Pamungkas D S, Wicaksono P, Apriliyanti T 2017 Mapping the Distribution of Potential Land Drought in Batam Island Using the Integration of Remote Sensing and Geographic Information Systems (GIS) *IOP Conf. Series: Earth and Environmental Science* **98** 012012 DOI :10.1088/1755-1315/98/1/012012

[12] Neema Nicodemus Lyimo, Zhenfeng Shao, AllyMgelwa Ally, Nana Yaw Danquah Twumasi, Orhan Altan and Camilius A Sanga 2020 A Fuzzy Logic-Based Approach for Modelling Uncertainty in Open Geospatial Data on Landfill Suitability Analysis *International Journal of Geo-Information. ISPRS Int. J. Geo-Inf.* **9** 737 URL: www.mdpi.com/journal/ijgi doi:10.3390/ijgi9120737

[13] Mironova Yu N 2018 Use of fuzzy sets in modeling of GIS objects *Journal of Physics: Conference Series* Vol 1015 1015 032094 URL: http://iopscience.iop.org/article/10.1088/1742-6596/1015/3/032094 doi:10.1088/1742-6596/1015/3/032094

[14] Mironova Yu N 2018 Some aspects of Geoinformatics. Monograph (Moscow: Pero Publishing House) p 98

[15] Mironova Yu N 2015 Mathematical aspects of Geoinformatics *Internet journal "NAUKOVEDENIE"* Vol. 7 5 p 143 URL: http://naukovedenie.ru/PDF/93TVN515.pdf (access is free). DOI: 10.15862/93TVN515

[16] Mironova Yu N 2016 Virtual modeling in geoinformation technologies Materials of conferences (Munich, Germany, 1-6 November 2016): «EDUCATION AND SCIENCE WITHOUT BORDERS», «FUNDAMENTAL AND APPLIED RESEARCH IN NANOTECHNOLOGY» *International Journal Of Applied And Fundamental Research* **2** URL: www.science-sd.com/464-24988

[17] Mironova Yu N 2016 The classification of geoinformation objects Materials of conferences (Munich, Germany, 1-6 November 2016): «EDUCATION AND SCIENCE WITHOUT BORDERS», «FUNDAMENTAL AND APPLIED RESEARCH IN NANOTECHNOLOGY» *International Journal Of Applied And Fundamental Research* **2** URL: www.science-sd.com/464-25151

[18] Mironova Yu N 2016 New methods of virtual modeling in geoinformation technologies Internet journal "NAUKOVEDENIE" Vol. 8 5 http://naukovedenie.ru/PDF/03TVN516.pdf

[19] Mironova Yu N 2017 Difficulties of classification of geoinformation objects *Internet journal "NAUKOVEDENIE"* Vol 9 1 http://naukovedenie.ru/PDF/09TVN117.pdf

[20] Mironova Yu N 2017 Decryption of space images by using GIS-technologies Materials of conferences (Munich, Germany, 31 October - 5 November 2017): «EDUCATION AND SCIENCE WITHOUT BORDERS» «FUNDAMENTAL AND APPLIED RESEARCH IN NANOTECHNOLOGY» *International Journal Of Applied And Fundamental Research* **3** URL: www.science-sd.com/471-25226

[21] Mironova Yu N 2018 Fuzzy information in Geoinformatics Materials of conferences (Munich, Germany, 31 October - 5 November 2018): «EDUCATION AND SCIENCE WITHOUT BORDERS», «FUNDAMENTAL AND APPLIED RESEARCH IN NANOTECHNOLOGY» *International Journal Of Applied And Fundamental Research* **6** URL: www.science-sd.com/478-25434

[22] Mironova Yu N 2018 Using fuzzy set theory in Geoinformatics Materials of conferences (Munich, Germany, 31 October - 5 November 2018): «EDUCATION AND SCIENCE WITHOUT BORDERS», «FUNDAMENTAL AND APPLIED RESEARCH IN NANOTECHNOLOGY» *International Journal Of Applied And Fundamental Research* **6** URL:
www.science-sd.com/478-25462

[23] Mironova Yu N 2019 The Use of Mathematical Logic in the Automated Decoding of Images 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) 1-4 Oct. 2019 (Vladivostok, Russia) URL: https://ieeexplore.ieee.org/document/8933838 DOI: 10.1109/FarEastCon.2019.8933838

[24] Mironova Yu N 2020 Mathematical Methods in Geoinformatics 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) 6-9 Oct. 2020, (Vladivostok, Russia) URL: https://ieeexplore.ieee.org/document/9271594 DOI: 10.1109/FarEastCon50210.2020.9271594

[25] Mironova Y N 2001 On tau-pseudocompact mappings Siberian Mathematical Journal 42(3) 537–545 342233

[26] Mironova Yu N 2019 Generalized Euler-Venn diagrams for fuzzy sets Geometry and graphics Vol. 7 4 DOI: 10.12737/2308-4898-2020-34-43 URL: https://naukaru.ru/ru/nauka/article/35508/view

[27] Mironova Yu N 2020 Place of Geoinformatics in the system of Sciences Bulletin of Syktyvkar University. Series 1: Mathematics. Mechanics. Informatics 1(34) pp 35-46 URL: https://www.syktsu.ru/_fac/math/vestnik/article/v34.pdf