ISSN NO: 2770-4491
December 2024

Method And Technology For Applying A Semi-Structured Data Classification Model In A Corporate Information System

Askaraliyev Odilbek Ulug‘bek o‘g‘li,
PhD on Technical Sciences, Rector of Sarbon University, Tashkent, Uzbekistan
e-mail: oasqaraliyev77@gmail.com

Annotation. A scheme for solving the problem of binary clustering of incoming data in the control system is proposed. Methods of sequential shortening and sequential merging of clusters, as well as the model of initial placement of clusters are considered. Estimating the number of clusters is given to solve the clustering problem. A binary clustering method has been proposed for points in a circle. Based on the results of the research, the clustering practice was performed and the developed model and algorithms were implemented

Keywords: management system, unstructured data, geometric clustering, artificial neural network, cluster, network model, classification.

Many theoretical issues of geometric clustering and classification of weakly structured information remain unresolved to date [1-6]. Semistructured data is data for which the exact structure is not known in advance and can change in the stream. Such data usually includes texts and pictures, which can be either separately or simultaneously present in documents presented in various formats.

On Fig. 1 presents the main stages, models and methods of cluster analysis of weakly structured information. Much attention is paid to metrics, issues of choosing the number and initial placement of clusters. Formal formulations of a number of problems are common both in the clustering of texts and images. Therefore, an urgent task is the integration of various models in a single processing system.

Models for representing semi-structured information. Models of data representation and knowledge have a significant impact on the choice of cluster analysis method. In the case of multimodal information, characterized by the presence of heterogeneous sources of information (texts, images, sounds), and in the presence of a large number of features, increased requirements are imposed on the form of data presentation.

Let us consider some data representation models that can be applied in cluster analysis problems [7-9].

Let the initial information for the objects $\omega_i, i = 1, \dots, m$ be given in the form of training samples. The corresponding data structure is presented in Table. 1. Classes Ω_1 *and* Ω_2 are represented by matrices of feature values X_1 *and* X_1 of dimensions $(m_1 \times p)$ *and* $(m_2 \times p)$, respectively, with $(m_1 + m_1 = m)$.

Thus, Tab. 1 contains data on m objects $\{\omega_1, \dots, \omega_m\}$, each of which is represented by a vector of values of information features $x = \{x_1, \dots, x_m\}$ and assigned by the expert to one of two classes $\{\Omega_1, \Omega_2\}$.

Tab. 1. Representation of the training sample

Objects	**Signs and their meanings**			**Class**
	x_1	…	x_p	
ω_1	x_{11}	…	x_{1p}	Ω_1
…	…	…	…	…
ω_{m_1}	$x_{m_1 1}$	…	$\omega_{m_1 p}$	Ω_1
ω_{m_1+1}	$\omega_{(m_1+1)1}$	…	$\omega_{(m_1+1)p}$	Ω_2
…	…	…	…	…
ω_m	x_{m1}	…	x_{mp}	Ω_2

The tabular method is a convenient representation of information for constructing decision rules for classification problems. In particular, it is well suited for solving classification problems by the algebraic method. Since the dimension of the attribute space changes in document processing tasks, the tabular method should work in conjunction with the method of variation of the attribute space [7].

A multiset or a set with repeating elements serves as a convenient mathematical model for describing objects that are characterized by many heterogeneous (quantitative and qualitative) features and can exist in several instances with different, in particular, contradictory values of features, the convolution of which is either impossible or mathematically incorrect. A multiset A generated by an ordinary set $U = \{x_1, x_2, \dots\}$, all elements of which are different, is a set of groups of elements of the form $A = \left\{k_A(x) \bullet x \,\middle|\, x \in U,\ k_A(x) \in Z+\right\}$. Here $k_A: U \to Z+ = \{0,1,2,\dots\}$ is called the function of the number of instances of the multiset, which determines the multiplicity of occurrence of the element $x_i \in U$ in the multiset A, which is denoted by the symbol • . If $k_A(x) = \chi_A(x)$ where $\chi_A(x) = 1$ for $x \in A$ and $\chi_A(x) = 0$ for $x \notin A$, then the multiset A becomes an ordinary many. If all multisets of the family $A = \{A_1, A_2, \dots\}$, are formed from elements of the set G, then G is called a domain for the family A, and the set $SuppA = \left\{x \middle| x \in G, \chi_{SuppA}(x) = \chi_A(x)\right\} -$ support set or carrier of the multiset A.

The cardinality of the multiset $|A|$ defined as the total number of instances of all its elements; dimension of the multiset $|SuppA|$ as the total number of distinct elements.

All initial multisets are grouped (summed up) into two multisets representing two classes of objects. Multisets-sums, in turn, are divided into several multisets-summands according to the number of features that characterize objects. For each group of features, for each pair of summands of multisets, a pair of new multisets is generated that are maximally distant from each other in the metric space. The boundary between new terms in each pair is determined by some value of the corresponding feature. Various combinations of such "boundary" feature values provide generalized decision rules for classifying objects.

Classification is carried out with the help of generalized decision rules, composed of different "boundary" values of features, which resembles the method of decision trees, providing the desired level of accuracy of object classification.

Phase trajectories for describing multimodal data. In [4], a method for classifying multimodal data for processing speech, images, and texts is proposed. The associativity of accessing information allows you to quickly obtain the necessary information, regardless of the sample size, and the structural approach to information processing allows you to automatically restore the structure and compactly store the information received.

The time complexity of the clustering algorithm is estimated in different sources as $O(m^2)$ or $O(m^3)$. We write the analytical dependence of the average duration of the solution of the problem with cubic complexity in the form [12]:

$$T(m) = am^3 + am(\tfrac{n}{m} + 1)^3 = am^3 + a\tfrac{(n+m)^3}{m^2}. \qquad (5)$$

Solving the optimization problem, we get:

$$3am^2 + a\frac{3(n+m)^2m^2 - 2m(n+m)^3}{m^4} = 0,$$

$$3m^5 + m^3 - 3n^2m - 2n^3 = 0. \qquad (6)$$

The results of the numerical solution of equation (6) are contained in Table. 2 (results rounded to whole numbers).

Tab. 2. Results of the numerical solution of the equation

n	10	50	100	200	300	400	500
m	4	10	15	22	29	34	39

7. Binary clustering**.** A solution to the problem of binary clustering of points located on a circle is proposed. Let there be n points (objects) ordered on the boundary of the circle L.

Without loss of generality, we divide the set of points located on a circle into two subsets (clusters) C_1 and C_2 so that the sum of intraclass distances is greater than the interclass distance. To do this, perform the following steps.

We number the points on the circle from 1 to n. Let a pair (l, k) characterize a tuple formed by a sequence of points, where $l(l = 1, \dots, n) -$ is the number of the initial point of the cluster, $k(2 \leq k \leq n - 2) -$is the number of cluster points located on the circle.

For each l, we construct a discrete function $F(l,k)$, which characterizes the quality of partitioning, depending on the number of elements in the cluster k. Best local solution:

$$F(l,k) = \frac{distance\ between\ classes}{sum\ of\ intra{-}class\ distances} \to_{k}^{max} \tag{7}$$

The optimal solution corresponds to the global maximum among all $\mathrm{F(l,k)} \to_{\mathrm{l,k}}^{\mathrm{max}}$ a distance, we will use the well-established Euclid-Mahalanobis metric d_{E-M}, which takes into account the statistical properties of clusters.

Let us define the components of the quality function. intra-class distance. Counting for each class

$C_1 = C_1(l,k)$ and $C_2 = C_2(l,k)$ covariance matrices $S_1 = S_2(l,k)$ and $S_2 = S_2(l,k)$, as well as intraclass distances as the sum of distances from each point of class $x \in C_1$ to the center of the corresponding class $\bar{x}_i$:

$$\sum_{x \in C_1} d_{E-M}(x, C_1) = \sum_{x \in C_1} \sqrt{(x - \bar{x}_1)^T S_1^{-1}\,(x - \bar{x}_1)},$$

$$\sum_{x \in C_2} d_{E-M}(x, C_2) = \sum_{x \in C_2} \sqrt{(x - \bar{x}_2)^T S_2^{-1}\,(x - \bar{x}_2)}.$$

Distance between classes. For each option (l,k), we calculate the interclass distance, i.e. distance between classes C_1 and C_2 To do this, we build the combined covariance matrix according to the formula:

$$S_{1,2} = \frac{1}{n-2}(S_1 + S_2).$$

The distance between classes is calculated as the distance between the centers of the classes with the merged matrix:

$$d_{E-M}(C_1, C_2) = \sqrt{(\bar{x}_1 - \bar{x}_2)^T S_{1,2}^{-1}\,(\bar{x}_1 - \bar{x}_2)}.$$

Quality control:

$$F(l,k) = \frac{d_{E-M}(C_1, C_2)}{\sum_{x \in C_1} d_{E-M}(x, C_1) + \sum_{x \in C_2} d_{E-M}(x, C_2)} \to_{l,k}^{max}.$$

The procedure has time complexity (by the number of iterations $O(n(n-3)/2)$.

8. An example of solving the problem of binary clustering. Let us consider an example of joint application of the feature space variation method [7] and the method based on the network model. 1) Initial table of precedents (Table 3).

Tab. 3. Source cases

	x_1	x_2	x_3	x_4	x_5	x_6	Class
W_1^1	1	2	0	1	-1	0	1
W_2^1	-1	0	-1	-2	-2	1	1
W_3^1	1	-1	2	-1	-1	0	1
W_1^2	0	2	1	-1	0	1	2
W_2^2	2	-1	0	-1	-1	0	2

2) Let's construct the covariance matrix C .

$$C = \begin{pmatrix} 1.3 & -0.55 & 0.45 & 0.60 & 0.25 & -0.55 \\ -0.55 & 2.30 & -0.20 & 0.90 & 0.50 & 0.30 \\ 0.45 & -0.20 & 1.30 & 0.15 & 0.50 & -0.20 \\ 0.60 & 0.90 & 0.15 & 1.20 & 0.25 & -0.35 \\ 0.25 & 0.50 & 0.50 & 0.25 & 0.50 & 0.00 \\ -0.55 & 0.30 & -0.20 & -0.35 & 0.00 & 0.30 \end{pmatrix} \tag{8}$$

3) Find the eigenvectors and eigenvalues for the covariance matrix (8) by solving the determinant (Table 4):

$$\begin{vmatrix} c_{11} & -\lambda_1 & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & c_{22} & -\lambda_2 & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & c_{33} & -\lambda_3 & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & c_{44} & -\lambda_4 & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & -\lambda_5 & c_{56} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & c_{56} & -\lambda_6 \end{vmatrix} = 0 \qquad (9)$$

Tab. 4. Vector of eigenvalues

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0	2,96	1,18	2,45	0	0,29

4) The resulting Table,5 vectors in the feature system $Y = (y_1, \dots, y_n)$.

Tab. 5. Vectors in the new coordinate system

Vectors in the new system.	y_1	y_2	y_3	y_4	y_5	y_6	Class
V_1^1	-0.33	0.27	-0.01	-0.43	-0.02	-0.85	1
V_2^1	-0.16	0.87	-0.06	0.39	0.18	0.08	1
V_3^1	-0.32	0.07	0.75	-0.38	0.34	0.2	1
V_1^2	0.64	-0.06	0.47	0.45	0.24	-0.29	2
V_2^2	-0.45	-0.36	-0.22	0.31	0.66	-0.26	2

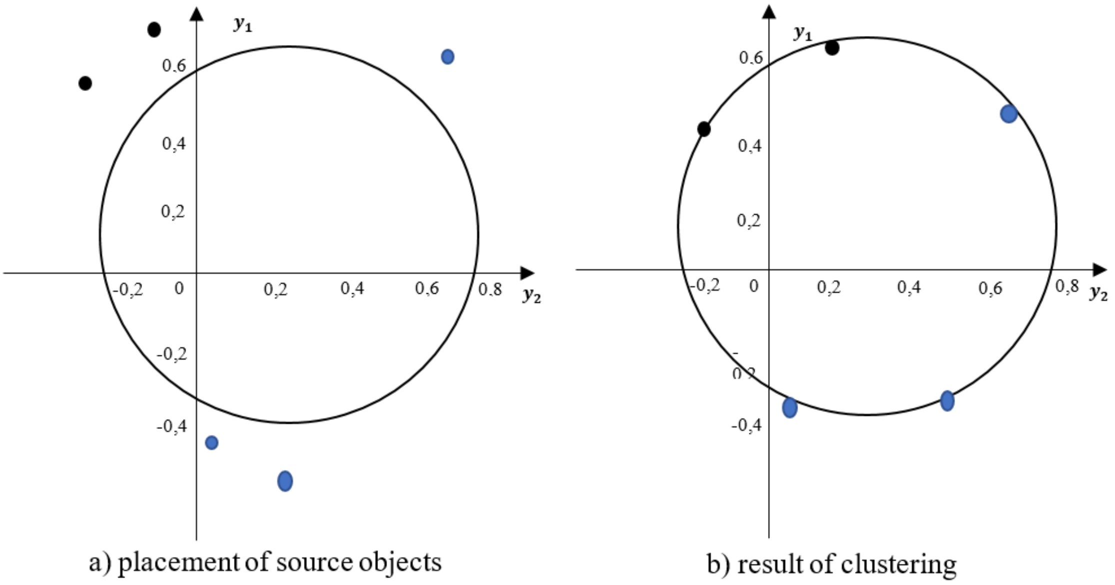
For the further course of research, it is important for us to use the largest eigenvalues 2,96, $\lambda_4 = 2{,}45$ and the eigenvectors corresponding to them. After the dimension reduction, the original objects were placed on the plane as shown in Fig. 4a).

The same figure shows neurons evenly spaced on a circle. Next, the clustering problem is solved based on the network model. At the same time, a network model (based on the Kohonen neural network) is necessary for iterative “pulling” of objects to a circle. After that, the simplified problem of binary clustering (7) is solved. As can be seen from Fig.4, b), as a result, the objects are located on the circle as they were classified by experts in the original table.

Thus, the joint application of methods allows us to solve the problem of binary clustering.

It is expedient to take the tabular method as the basis for representing the semi-structured initial data of the clustering problem, which is widely used in solving classification problems by the algebraic method, neural networks, decision trees, and in constructing the reference set of clusters. The considered methods for choosing the initial number of clusters and the model of their placement allow us to solve the problem of binary clustering. Moreover, the solution of the problem of placing points on a circle based on a network model is considered as a step preceding clustering.

ISSN NO: 2770-4491
December 2024



a) placement of source objects

b) result of clustering

Fig. 4. An example of solving the clustering problem

The expediency of using different methods of cluster analysis together is shown in the example of solving the problem of binary clustering. The binary cluster method was found to be the most optimal for the study object. Based on the experiments performed, it is applied to the data array included in the management system.

REFERENCES

[1]. Askaraliyev O.U., Sharipov Sh.O., Akbarova N.R.: "Implementation of Decision Support Procedures using an Expert System in Integrated Management (On the Field of Tax Authorities)" // Design Engineering: Y 2021 Issue 9, -P 4048-406, https://www.scopus.com/sourceid/28687

[2]. Averkin A.N., Gaaze-Rapoport M.G., Pospelov D.A. Explanatory Dictionary of Artificial Intelligence. M .: Radio and communication, 1992.

[3]. [7]. Bolotova L.S. Artificial intelligence systems: models and technologies based on knowledge: textbook. / FGBOU VPO RGUITP; FGAU GNII ITT "Informika". Moscow: Finance and Statistics, 2012.

[4]. Хачумов М.В. Задача кластеризации текстовых документов. — Информационные технологии и вычислительные системы, № 2, 2010, с.42-49.

[5]. Петровский А.Б. Пространства множеств и мультимножеств. – М.: УРСС, 2003. – 248 с. - http://www.raai.org/about/persons/petrovsky/pages/Petrov sky_2003.pdf

[6]. Osipov G. Strategies for Stabilization Behaviour of Intelligent Dynamic Systems. – Proc. of 20th European Meeting on Cybernetics and Systems, Vienna, 2010, pp.195-197.

[7]. Larichev OI, Mechitov A.I., Moshkovich E.M., Furems E.M. Systems for identifying expert knowledge in classification problems // Izv. USSR Academy of Sciences. - Ser. Technical cybernetics. - 2005. - No. 2.

[8]. Larichev OI, Moshkovich E.M. Qualitative decision making methods. - M .: Science. Fizmatlit. - 1996.

[9]. J.B.Elov, U.R.Khamdamov, Dj.B.Sultanov, O.Q.Makhmanov, "Organizing functional processes of information system for the advanced training of medical personnel on the basis of IDEF methodology", *International Journal of Advanced Research in Science, Engineering and Technology,* vol. 6, Issue 12, December 2019. India. pp. 12085-12090.

[10]. Gavrilova T.A., Khoroshevsky V.F. Knowledge base of intelligent systems. SPb .: Peter, 2000.