e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2021. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2021. № 1

UDC 519.876.2:336

METHOD OF SPECTRAL CLUSTERING OF PAYMENTS AND RAW MATERIALS SUPPLY FOR THE COMPLIANCE AUDIT PLANNING

Neskorodieva T. V. – PhD, Associate Professor, Head of the Department of Computer Science and Information Technology, Vasyl' Stus Donetsk National University, Vinnytsia, Ukraine.

Fedorov E. E. – Dr. Sc., Associate Professor, Professor of the Department of Robotics and Specialized Computer Systems, Cherkasy State Technological University, Cherkasy, Ukraine.

ABSTRACT

Context. The analytical procedures used in the audit are currently based on data mining techniques. The work solves the problem of increasing the efficiency and effectiveness of analytical audit procedures by clustering based on spectral decomposition. The object of the research is the process of auditing the compliance of payment and supply sequences for raw materials.

Objective. The aim of the work is to increase the effectiveness and efficiency of the audit due to the method of spectral clustering of sequences of payment and supply of raw materials while automating procedures for checking their compliance.

Method. The vectors of features are generated for the objects of the sequences of payment and supply of raw materials, which are then used in the proposed method. The created method improves the traditional spectral clustering method by automatically determining the number of clusters based on the explained and sample variance rule; automatic determination of the scale parameter based on local scaling (the rule of K-nearest neighbors is used); resistance to noise and random outliers by replacing the k-means method with a modified PAM method, i.e. replacing centroid clustering with medoid clustering. As in the traditional approach, the data can be sparse, and the clusters can have different shapes and sizes. The characteristics of evaluating the quality of spectral clustering are selected.

Results. The proposed spectral clustering method was implemented in the MATLAB package. The results obtained made it possible to study the dependence of the parameter values on the quality of clustering.

Conclusions. The experiments carried out have confirmed the efficiency of the proposed method and allow us to recommend it for practical use in solving audit problems. Prospects for further research may lie in the creation of intelligent parallel and distributed computer systems for general and special purposes, which use the proposed method for segmentation, machine learning and pattern recognition tasks.

KEYWORDS: audit planning, clustering, spectral decomposition, medoids, sequence of payment and supply of raw materials.

ABBREVIATIONS

NJW is a Ng, Jordan, Weiss method;

PAM is the partitioning around medoids;

EM is an expectation-maximization;

DBSCAN is a density-based spatial clustering of applications with noise;

OPTICS is an ordering points to identify the clustering structure;

DIANA is a divisive analysis;

SOM is a self-organizing map;

ART is a adaptive resonance theory;

TP is a true positive;

TN is a true negative;

FP is a false positive;

FN is a false negative.

NOMENCLATURE

A is a set of clustering objects;

a_i is an i-th object of clustering;

n is a number of objects of clustering;

X is an set of feature vectors from the space R^q;

x_i is a feature vector of i-th object of clustering from the space R^q;

q is a number of features in feature vector x_i;

$\breve{X}$ is a set of K-nearest feature vectors;

$\breve{x}_i$ is a feature vector of K-nearest to feature vector x_i;

δ is a threshold for determining the number of clusters;

K is a number of nearest neighbors;

σ_i is a scale parameter for the i-th feature vector;

$\mathbf{S}$ is a symmetric similarity matrix;

$\mathbf{D}$ is a diagonal degree matrix;

$\mathbf{L}$ is a normalized symmetric Laplace matrix;

$\mathbf{I}$ is a unit matrix;

λ_i is an i-th eigenvalue;

$\mathbf{w}_i$ is an i-th eigenvector;

c is a number of clusters;

R^2 is a coefficient of determination;

$\mathbf{V}$ is a principal component matrix;

$\tilde{X}$ is a set of feature vectors from the space R^c;

$\tilde{x}_i$ is an i-th feature vector from the space R^c;

$\check{X}$ is a set of feature vectors from the space R^c, which not corresponding to medoids;

A_k is a k-th cluster;

$\mathbf{\Lambda}$ is a set of indicator functions;

$\chi_{A_k}(\cdot)$ is an indicator function A_k (returns 1 or 0 depending on the belonging of the object to the k-th cluster);

D_{ik} is a square of distance between i-th object and medoid of k-th cluster;

$F(\cdot)$ is a target function;

DOI 10.15588/1607-3274-2021-1-13

y^* is a best target function value;

y is a target function value;

M is a set of cluster centroids;

$\mathbf{m}_k$ is a centroid of k-th cluster for the space R^q;

$\tilde{M}$ is a set of medoids of cluster;

$\tilde{\mathbf{m}}_i$ is a medoid of k-th cluster for the space R^c;

$\hat{\mathbf{m}}$ is a preserved medoid for space R^c;

$N(0,1)$ is a function, that returns standard normal distributed random number;

v^2 is a variance of Gaussian additive noise;

Accuracy is an accuracy;

Precision is a precision;

Recall is a recall;

F is a balanced F-measure;

θ_d is a types set of paid raw materials;

θ_k is a types set of raw materials obtained;

s_d is a type of paid raw materials;

s_k is a type of raw material received;

δ_{s_d} is a cost of paid raw material of type s_d;

v_{s_k} is a number of received raw material of type s_k.

INTRODUCTION

The analytical procedures used in the audit are currently based on data mining techniques [1, 2]. In an automated audit system, the task of auditing expenses at the top level is decomposed into tasks of checking the sequence of displaying data of the middle level. First of which – display is paid-received. This is a mapping of the multidimensional data of payment for raw materials to suppliers to/in the set of multidimensional data for the delivery of raw materials. At the lower level, if there are no violations in accounting, this mapping should be one-to-one. In order to reduce the volume of checks at the lower level, the audit system analyzes the aggregated indicators of payment and delivery at the middle level or formed sets (clusters) of multidimensional data of the lower level. Also, when designing an IT audit, the goal is to automate the analysis to form recommended solutions. According to the method of generalized set mapping, at the middle level, generalized properties of data sets (condensation points, isolated points) are analyzed, that is, the density structure of each of the sets is determined, and then they are compared.

For analysis, pay and delivery sequence data can be aggregated over quantization periods:

1) for all suppliers;
2) by the nomenclature of raw materials.

Analysis of data on payment and supply of raw materials is carried out to form recommended solutions for the following audit tasks.

1. The task of the external audit is to check the completeness of accounting for settlements with suppliers.
2. Tasks of internal audit to verify compliance with contractual policies. The contractual policy of the enterprise is a set of rules characterizing the delivery time after payment, the nomenclature of raw materials, technical or physical characteristics, prices (discounts).
3. The task of the internal audit of pricing policy when concluding contracts (identifying a significant share of unfavorable contracts, which are features of "kick-backs" when concluding them).
4. The task of internal audit of receivables from suppliers of raw materials in terms of timing and amounts.

Clustering methods are used to audit the compliance of the sequence of payments for raw materials and the sequence of deliveries of raw materials at the stage of identifying characteristic properties.

Traditional clustering methods are:

1. Partition-based (partitioning-based) or center-based methods (e.g., methods k-means [3], PAM (k-medoids) [4], FCM [5]).
2. Mixture model or distribution-based or model-based methods (e.g., EM [5]).
3. Density-based methods (e.g., methods DBSCAN [6], OPTICS [7]).
4. Hierarchical methods:

– agglomerative or ascending (bottom up) (e.g. centroid communication methods, Vard, unit connection, full connection, group secondary) [8];

– divisive or descending (top down) (e.g., methods DIANA) [9].

Clustering methods can also be based on metaheuristics [10, 11] and artificial neural networks (e.g., SOM, ART) [12].

Object of study. Audit process for compliance with payment sequences and raw materials supply.

Subject of study. Spectral clustering method for auditing sequences of payment and supply of raw materials.

The aim of the work is to increase the effectiveness and efficiency of the audit by automating the analysis of data from sets of parallel-sequential operations of payment and supply of raw materials based on the spectral clustering method.

To achieve this goal, it is necessary to solve the following tasks:

1. Generate feature vectors for objects of sequences of payment and supply of raw materials.
2. Create a method for spectral clustering of sequences of payment and supply of raw materials.
3. Select characteristics for assessing the quality of spectral clustering.
4. Conduct a numerical study of the proposed spectral clustering method.

1 PROBLEM STATEMENT

The problem of increasing the efficiency of audit based on the method of spectral clustering of sequences of payment and supply of raw materials is presented as the problem of finding such a partition of the set of clustering

DOI 10.15588/1607-3274-2021-1-13

objects $A = \{a_1, ..., a_n\}$, represented by a set of feature vectors $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, per cluster $A_1, ..., A_c$ through a variety of indicator functions $\Lambda = \{\chi_{A_1}(\cdot), ..., \chi_{A_c}(\cdot)\}$, and with a set of cluster centroids $M = \{\mathbf{m}_1, ..., \mathbf{m}_c\}$, at which

$$\mathrm{F} = \sum_{i=1}^{n} \sum_{k=1}^{c} \chi_{A_k}(a_i) \left\| \mathbf{x}_i - \mathbf{m}_k \right\|^2 \to \min_{\Lambda, M}.$$

2 REVIEW OF THE LITERATURE

Existing clustering methods have one or more of the following disadvantages [6, 7]:

– have high computational complexity;

– do not allow the emission of noise and random emissions;

– clusters cannot have different shapes and sizes;

– require specifying the number of clusters;

– require the definition of parameter values.

In this regard, it is relevant to create a clustering method that will eliminate the indicated disadvantages.

One of these methods is spectral clustering [13, 14], which has already found application in the segmentation of signals of different physical nature [15]. Since initially the spectral clustering methods did not provide for the procedure for automating the determination of the parameters and the number of clusters, an attempt is being made to eliminate this drawback. [16], which will allow them to be used in IT audit of enterprises with different characteristics.

3 MATERIALS AND METHODS

Let's start by solving the first task formation of feature vectors for objects of sequences of payment and supply of raw materials.

The attributes for the objects of the sequence of payment and supply of raw materials are formed on the basis of the accounting variables of the lower level (Table 1), taking into account the possible options for generalizing their values at the average level for the periods of quantization. Clustering objects of payment (supply) for each supplier with which a long-term supply agreement is in force during the year for which the audit is carried out. Feature vector $\mathbf{x}_i = (x_{i1}, ..., x_{iq})$ objects of payment form indicators of the cost of paid raw materials δ_{s_d} by types $s_d \in \Theta_d$. Features vector $\mathbf{x}_i = (x_{i1}, ..., x_{iq})$ delivery objects form indicators of the amount of paid raw materials v_{s_k} by types $s_k \in \Theta_k$.

To assess the dimension of the vector of attributes and the number of objects of analysis, an analysis of the nomenclature of purchases of raw materials (components) of large engineering enterprises. So, based on this analysis, we can conclude that the sections of the nomenclature are on average from 8 to 12, the number of groups in each section is from 2–10. Analyzing the homogeneity of the procurement nomenclature, we can conclude that for continuous operation the plant can have long-term contracts with suppliers in the amount of 50 до 100.

Clustering will make it possible to form subsets of payment and supply operations that are similar in terms of the features highlighted above, which will allow analyzing the set of operations when comparing and reducing the computational complexity of solving the matching problem.

To form the rules for matching the sequences of payments and deliveries after clustering, it is necessary to select the rules of relationships. Based on the analysis of the terms of payment agreements and the supply of raw materials, the rules for recording these transactions in the system, the following rules were identified:

1) Delivery operations are carried out after payment, in accordance with the contractual policy of the enterprise in accordance with payment orders.

2) Delivery under a new payment order is not carried out until the previous one is closed.

3) Low-level delivery data that corresponds to one payment order is aggregated before clustering.

Let's move on to solving the second problem – a method ***creating*** for spectral clustering of sequences of payment and supply of raw materials (Fig. 1).

1. Specifying multiple clustering objects $A = \{a_i\}$, $i \in \overline{1,n}$. Specifying a set of feature vectors $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and each object a_i corresponds to the feature vector $\mathbf{x}_i = (x_{i1}, ..., x_{iq})$. Setting the threshold for determining the number of clusters δ, $0 < \delta < 1$. Setting the number of nearest neighbors K.

2. Creation of a set of-nearest feature vectors $\breve{X} = \{\breve{\mathbf{x}}_i\}$, such that for each feature vector $\mathbf{x}_i$ K-nearest to it is the feature vector $\breve{\mathbf{x}}_i$.

3. Calculating scale options based on local scaling:

$$\sigma_i = \left\| \mathbf{x}_i - \breve{\mathbf{x}}_i \right\|, \ i \in \overline{1,n}.$$

4. Calculation of the symmetric similarity matrix

$$\mathbf{S} = [s_{ij}], \ i, j \in \overline{1,n},$$

$$s_{ij} = \begin{cases} \exp\left(-\dfrac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{2\sigma_i \sigma_j} \right), & i \neq j \\ 0, & i = j \end{cases}.$$

5. Calculating the diagonal degree matrix:

$$\mathbf{D} = diag(d_1, ..., d_n),$$

$$d_i = \sum_{j=1}^{n} s_{ij}.$$

6. Calculation of the normalized symmetric Laplace matrix:

$$\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{S})\mathbf{D}^{-1/2}.$$

DOI 10.15588/1607-3274-2021-1-13

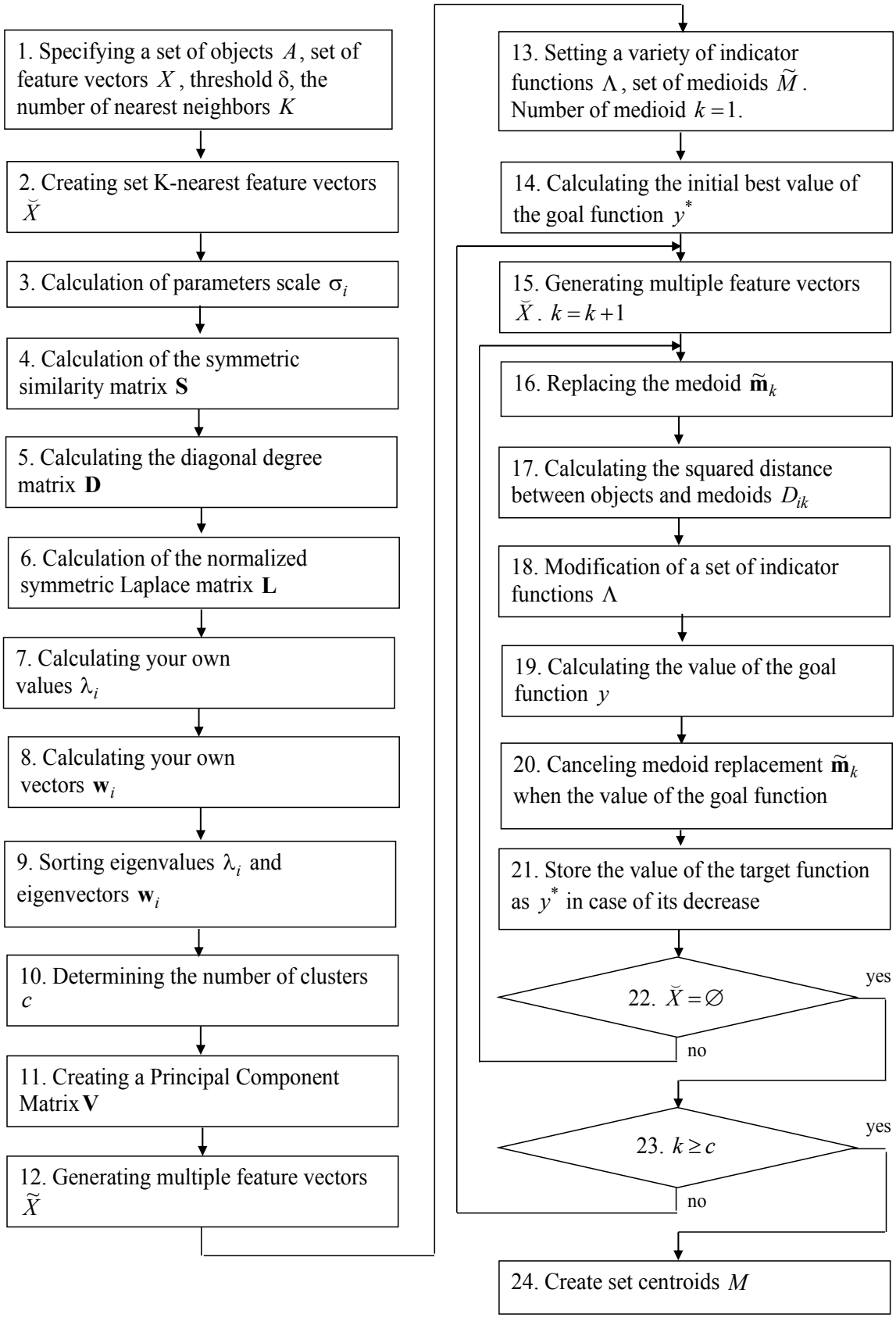

Figure 1 – The structure of spectral clustering of sequences of payment and supply of raw materials

DOI 10.15588/1607-3274-2021-1-13

7. Calculation of eigenvalues $\lambda_i, i \in \overline{1,n}$, matrix $\mathbf{L}$ as roots of the characteristic equation $\det(\mathbf{L}-\lambda\mathbf{I})=0$.

8. Computing eigenvectors $\mathbf{w}_i, i \in \overline{1,n}$, dimensions n from the equation $(\mathbf{L}-\lambda_i\mathbf{I})\mathbf{w}_i=0$, which is obtained from the relation $\mathbf{L}\mathbf{w}_i=\lambda_i\mathbf{w}_i$.

9. Sorting eigenvalues λ_i and eigenvectors $\mathbf{w}_i$ in descending eigenvalues λ_i.

10. Determining the number of clusters c as the number of selected eigenvalues and eigenvectors by means of a rule based on the coefficient of determination

$$0 < R^2 < \delta,\ R^2 = \frac{\sum_{i=1}^{c}\lambda_i}{\sum_{i=1}^{n}\lambda_i},$$

at that $\sum_{i=1}^{c}\lambda_i$ interpreted as a fraction of the variance explained, and $\sum_{i=1}^{n}\lambda_i$ interpreted as a fraction of the total variance.

11. Creating a Principal Component Matrix $\mathbf{V}=[v_{ij}]$ dimensions $n \times c$, whose columns are selected eigenvectors $\mathbf{w}_i$, which have eigenvalues λ_i that are the greatest.

12. Generating multiple feature vectors $\tilde{X}=\{\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_n\}$, and each object a_i corresponds to the feature vector $\tilde{\mathbf{x}}_i=(\tilde{x}_{i1},...,\tilde{x}_{ic})$

$$\tilde{x}_{ij}=\frac{v_{ij}}{\sqrt{\sum_{j=1}^{c}\left(v_{ij}\right)^2}},\ i \in \overline{1,n},\ j \in \overline{1,c}.$$

13. Setting randomly the initial partition of a set of clustering objects $A=\{a_1,...,a_n\}$, represented by a set of feature vectors $\tilde{X}=\{\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_n\}$, per clusters $A_1,...,A_c$ through a variety of indicator functions $\Lambda=\{\chi_{A_1}(\cdot),...,\chi_{A_c}(\cdot)\}$ (return 1 or 0 depending on whether the object belongs to the *k*-th cluster). From set $\tilde{X}$ a set of medoids are randomly selected $\tilde{M}=\{\tilde{\mathbf{m}}_1,...,\tilde{\mathbf{m}}_c\}$. Medoid number is $k=0$.

14. Calculating the initial best value of the goal function

$$y^*=\sum_{i=1}^{n}\sum_{k=1}^{c}\chi_{A_k}(a_i)\left\|\tilde{\mathbf{x}}_i-\tilde{\mathbf{m}}_k\right\|^2.$$

15. Creation of a set of vectors of features that do not correspond to medoids $\breve{X}=\tilde{X}\setminus\tilde{M}$. Incrementing the medoid number, i.e. $k=k+1$

16. Replacing the medoid $\tilde{\mathbf{m}}_k$.

16.1. Saving the replaceable medoid $\tilde{\mathbf{m}}_k$, i.e. $\hat{\mathbf{m}}=\tilde{\mathbf{m}}_k$.

16.2. Extract from the set $\breve{X}$ next feature vector and assigning it the vector $\tilde{\mathbf{m}}_k$.

17. Calculating the squared distance between objects and medoids

$$D_{ik}=\left\|\tilde{\mathbf{x}}_i-\tilde{\mathbf{m}}_k\right\|^2,\ i \in \overline{1,n},\ k \in \overline{1,c}.$$

18. Modification of a set of indicator functions

$$\chi_{A_k}(a_i)=\begin{cases}1, & k=\arg\min\limits_{j\in\overline{1,c}} D_{ij}\\ 0, & k\neq\arg\min\limits_{j\in\overline{1,c}} D_{ij}\end{cases},\ i \in \overline{1,n},\ k \in \overline{1,c}.$$

The following conditions must be met for indicator functions

$$\sum_{k=1}^{c}\chi_{A_k}(a_i)=1,\ i \in \overline{1,n},$$

$$\sum_{i=1}^{n}\chi_{A_k}(a_i)>0,\ k \in \overline{1,c},$$

$$\chi_{A_k}(a_i) \in \{0,1\},\ k \in \overline{1,c},\ i \in \overline{1,n}.$$

19. Calculating the value of the goal function

$$y=\sum_{i=1}^{n}\sum_{k=1}^{c}\chi_{A_k}(a_i)\left\|\tilde{\mathbf{x}}_i-\tilde{\mathbf{m}}_k\right\|^2.$$

20. Cancellation of medoid replacement in case of increasing the value of the goal function

$$\text{if } y > y^*, \text{ then } \tilde{\mathbf{m}}_k=\hat{\mathbf{m}}.$$

21. Keeping the value of the target function as best as it increases.

$$\text{if } y < y^*, \text{ then } y^*=y.$$

22. If set $\breve{X}$ not empty then go to step 16.

23. If not all medoids are viewed, i.e. $k < c$, then go to step 15.

24. Creating multiple centroids

DOI 10.15588/1607-3274-2021-1-13

$$M = \{\mathbf{m}_1, ..., \mathbf{m}_c\},$$

$$m_{kj} = \frac{\sum_{i=1}^{n} \chi_{A_k}(a_i) x_{ij}}{\sum_{i=1}^{n} \chi_{A_k}(a_i)}, \; k \in \overline{1,c}, \; j \in \overline{1,q}.$$

The result of the method is a set of indicator functions $\Lambda = \{\chi_{A_1}(\cdot), ..., \chi_{A_c}(\cdot)\}$ and set of cluster centroids $M = \{\mathbf{m}_1, ..., \mathbf{m}_c\}$.

Fig. 1 shows the structure of spectral clustering of sequences of payment and supply of raw materials.

Let's move on to solving the third task – characteristics selecting for assessing the quality of spectral clustering. In the work, the following characteristics were chosen for assessing the quality of spectral clustering:

In the work, the following characteristics were chosen for assessing the quality of spectral clustering:

1. $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$,
2. $Precision = \frac{TP}{TP + FP}$,
3. $Recall = \frac{TP}{TP + FN}$,
4. Balanced F-measure

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

4 EXPERIMENTS

A numerical study of the proposed spectral clustering method was carried out in the package MATLAB.

The work used a standard database of handwritten numbers digit1000 (http://www.stat.washington.edu/spectral/datasets.html). There were 100 objects for each of the 10 digits, i.e. number of clustering objects $n = 1000$. For each object, the length of the feature vector was $q = 64$. Objects were noisy with additive Gaussian noise, i.e. added noise component $v^2 N(0,1)$, $v^2 = 0.05$. Threshold for determining the number of clusters $\delta = 0.05$, number of nearest neighbors $K = 7$.

5 RESULTS

The function reflecting the dependence of the determination coefficient on the number of clusters is presented in the form

$$R^2(c) = \frac{\sum_{i=1}^{c} \lambda_i}{\sum_{i=1}^{n} \lambda_i}.$$

Function part satisfying inequality $0 < R^2(c) < \delta$, shown in Fig. 2.

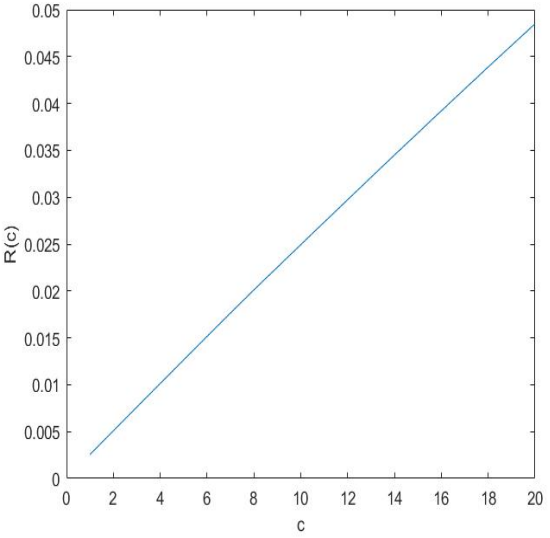

Figure 2 – Function reflecting the dependence of the coefficient of determination on the number of clusters

The dependence (Fig. 2) of the determination coefficient on the number of clusters shows that the determination coefficient increases with an increase in the number of clusters.

The results of comparison of the qualitative characteristics of the proposed method with the NJW method described in [13] are presented in Table 1.

Table 1 – Comparison of the qualitative characteristics of the proposed spectral clustering method with the existing NJW method

№	Method characteristics	Spectral clustering methods	
		This method	NJW
1	automatic determination of the number of clusters	+	–
2	automatic determination of the scale parameter	+	–
3	resistance to noise and accidental emissions	+	–
4	data can be sparse	+	+
5	clusters can have different shapes and sizes	+	+

The results of comparison of the quantitative characteristics of the proposed method with the NJW method described in [13] are presented in table 2.

Table 2 – Comparison of the quantitative characteristics of the proposed spectral clustering method with the existing NJW method

№ p/p	Method characteristics	Spectral clustering methods	
		This method	NJW
1	Accuracy	0.97	0.82
2	Precision	0.97	0.73
3	Completeness	0.97	0.82
4	Balanced F-measure	0.96	0.76

DOI 10.15588/1607-3274-2021-1-13

6 DISCUSSION

The selected values of the parameters of the proposed spectral clustering method provide high accuracy of clustering.

Traditional NJW Spectral Clustering Method [13]:

– requires specifying the number of clusters;

– scale parameter required;

– is not robust to noise and random outbursts (instead of the k-means method, a modified PAM method is used, i.e. centroid clustering is replaced by medoid clustering).

The proposed method eliminates the indicated disadvantages (table 2).

In terms of accuracy, precision, completeness, balanced F-measure, the proposed method is more effective than the NJW method (table 2).

CONCLUSIONS

The urgent task of increasing the effectiveness and efficiency of the audit was solved by creating a method of spectral clustering of sequences of payment and supply of raw materials.

The scientific novelty of obtained results is that the method of spectral clustering. It improves the quality of clustering due to:

– automatic determination of the number of clusters based on the explained and sample variance rule;

– automatic scaling parameter based on local scaling;

– resistance to noise and random outliers by replacing the k-means method with a modified PAM method, i.e. replacing centroid clustering with medoid clustering.

The practical significance of obtained results is that the proposed method makes possible to expand the scope of clustering methods based on spectral decomposition, which is confirmed by its adaptation for the audit task, and contributes to increasing the efficiency of intelligent computer systems for general and special purposes.

Prospects for further research are the study of the proposed method for a wide class of artificial intelligence tasks, as well as the creation of a method for matching payment and delivery sequences after clustering to solve audit problems.

ACKNOWLEDGEMENTS

The research was carried out in accordance with the priority direction of the development of science and technology in Ukraine “Information and communication technologies” and contain some results of research “Methods, models for the processing of intellectual, information technologies for highly efficient computational and local control systems in problem-based systems” (state registration number 0106U004501) and “Development of models and methods in biometric identification of people” (state registration number 0119U002860).

REFERENCES

1. Neskorodieva T., Fedorov E., Izonin I. Forecast Method for Audit Data Analysis by Modified Liquid State Machine, *The 1st International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2020), Khmelnytskyi, Ukraine, 10–12 June, 2020: proceedings,* 2020, CEUR-WS, Vol. 2623, pp. 25–35.
2. Neskorodieva T., Fedorov E. Method for Automatic Analysis of Compliance of Expenses Data and the Enterprise Income by Neural Network Model of Forecast, *The 2nd International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT&DS-2020), Lviv-Shatsk, Ukraine, 2–3 June, 2020: proceedings. CEUR-WS, Volume I: Main Conference.* 2020, Vol. 2631, pp. 145–158.
3. Brusco M. J., Shireman E , Steinley D. A Comparison of Latent Class, K-means, and K-median Methods for Clustering Dichotomous Data, *Psychological Methods*, 2017, Vol. 22 (3), pp. 563–580. DOI: 10.1037/met0000095.
4. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York, Plenum Press, 1981, 256 p. DOI: 10.1007/978-1-4757-0450-1.
5. Fu Z., Wang L. Color Image Segmentation Using Gaussian Mixture Model and EM Algorithm, *Multimedia and Signal Processing,* 2012, pp. 61–66. DOI: 10.1007/978-3-642-35286-7_9.
6. Ester M., Kriegel H.-P., Sander J., Xu X. Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Second International Conference on Knowledge Discovery and Data Mining (KDD), Portland, Oregon, August 2–4, 1996: proceedings*. AAAI Press, pp. 226–231.
7. Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering Points to Identify the Clustering Structure, *International Conference on Management of Data and Symposium on Principles of Database Systems. Philadelphia, Pennsylvania, USA, May, 1999: proceedings, Association for Computing Machinery*. New York, NY, United States, 1999, pp. 49–60.
8. Mirkin B. G. Clustering for Data Mining: A Data Recovery Approach. Boca Raton, FL, CRC Press, 2005, 277 p. DOI: 10.1201/9781420034912.
9. Aggarwal C. C., Reddy C. K. Data Clustering. Boca Raton, FL: CRC Press, 2014, 620 p. DOI:10.1201/9781315373515.
10. Subbotin S., Oliinyk A. , Levashenko V. , Zaitseva E. Diagnostic Rule Mining Based on Artificial Immune System for a Case of Uneven Distribution of Classes in Sample, *Communications*, 2016, Vol. 3, pp. 3–11.
11. Fedorov E., Utkina T., Nechyporenko O., Korpan Y. Development of technique for face detection in image based on binarization, scaling and segmentation methods, *Eastern-European Journal of Enterprise Technologies*, Vol. 1/9, 2020, pp. 23–31. DOI: 10.15587/1729-4061.2020.195369.
12. He J., Tan A.-H., Tan Ch.-L. Modified ART 2A Growing Network Capable of Generating a Fixed Number of Nodes, *IEEE Transactions on Neural Networks,* 2004, Vol. 15(3), pp. 728–37. DOI: 10.1109/TNN.2004.826220.
13. Andrew Y. Ng., Jordan I. M., Weiss Y. On spectral clustering: Analysis and an algorithm, *In Advances in neural information processing systems*, 2002, pp. 849–856.
14. Ulrike V. L. A tutorial on spectral clustering, *Statistics and computing*, Vol. 17(4): 2007, pp. 395–416. DOI: 10.1007/s11222-007-9033-z.
15. Fabien L., Schnörr C. Spectral clustering of linear subspaces for motion segmentation, *12th International Conference on Computer Vision (ICCV’09), Sep 2009, Kyoto*. Japan, proceedings, IEEE, pages to-appear, 2009. DOI: 10.1109/iccv.2009.5459173.
16. Tao Xiang, Shaogang G. Spectral clustering with eigenvector selection, *Pattern Recognition*, 2008, Vol. 41(3), pp. 1012–1029. DOI: 10.1016/j.patcog.2007.07.023.

DOI 10.15588/1607-3274-2021-1-13

17. Tao Xiang, Shaogang G. Spectral clustering with eigenvector selection, *Pattern Recognition,* 2008, Vol. 41(3), pp. 1012–1029. DOI: 10.1016/j.patcog.2007.07.023.
18. Feng Z., Licheng J., Hanqiang L., Xinbo G., Maoguo G. Spectral clustering with eigenvector selection based on entropy ranking, *Neurocomputing*, 2010, Vol. 73(10–12), pp. 1704–1717 DOI: 10.1016/j.neucom.2009.12.029.
19. Feng Zhao, Licheng J., Hanqiang L., Xinbo G., Gong M. Spectral clustering with eigenvector selection based on entropy ranking, *Neurocomputing*, 2010, 73(10–12):1704–1717. DOI: 10.1016/j.neucom.2009.12.029.

Received 17.11.2020.
Accepted 15.01.2021.

УДК 519.876.2:336

МЕТОД СПЕКТРАЛЬНОЇ КЛАСТЕРИЗАЦІЇ ПЛАТЕЖІВ І ПОСТАВКИ СИРОВИНИ ДЛЯ ПЛАНУВАННЯ АУДИТУ ВІДПОВІДНОСТІ

Нескородєва Т. В. – канд. техн. наук, доцент, Донецький національний університет імені Василя Стуса, Вінниця, зав. кафедри комп'ютерних наук та інформаційних технологій, Вінниця, Україна.

Федоров Є. Є. – д-р техн. наук, доцент, професор кафедри робототехніки та спеціалізованих комп'ютерних систем, Черкаський державний технологічний університет, Черкаси, Україна.

АНОТАЦІЯ

Актуальність. В даний час аналітичні процедури, які використовуються в ході аудиторської перевірки, базуються на методах інтелектуального аналізу даних. В роботі вирішується завдання підвищення результативності та ефективності аналітичних процедур аудиту шляхом кластеризації на основі спектрального розкладання. Об'єктом дослідження є процес аудиту відповідності послідовностей оплати і поставок сировини.

Мета. Метою роботи є підвищення результативності та ефективності аудиту за рахунок методу спектральної кластеризації послідовностей оплати і поставок сировини при автоматизації процедур перевірки їх відповідності.

Методи. Сформовано вектори ознак для об'єктів послідовностей оплати і поставок сировини, які потім використовуються в запропонованому методі. Створений метод вдосконалює традиційний метод спектральної кластеризації за рахунок автоматичного визначення кількості кластерів на основі правила поясненої і вибіркової дисперсії; автоматичного визначення параметра масштабу на основі локального масштабу (використовується правило K-найближчих сусідів); стійкості до шуму і випадковим викидів за рахунок заміни методу *k*-середніх модифікованим методом PAM, тобто заміни центроїдної кластеризації медоідною кластерізацією. Як і в традиційному підході дані можуть бути розріджені, а кластера можуть мати різну форму і розмір. Обрані характеристики оцінювання якості спектральної кластеризації.

Результати. Запропонований метод спектральної кластеризації був програмно реалізований в пакеті MATLAB. Отримані результати дозволили досліджувати залежність значень параметрів на якість кластеризації.

Висновки. Проведені експерименти підтвердили працездатність запропонованого методу і дозволяють рекомендувати його для використання на практиці при вирішенні завдань аудиту. Перспективи подальших досліджень можуть полягати в створенні інтелектуальних паралельних і розподілених комп'ютерних систем загального і спеціального призначення, які використовують запропонований метод для задач сегментації, машинного навчання та розпізнавання образів.

КЛЮЧОВІ СЛОВА: планування аудиту, кластеризація, спектральне розкладання, медоіди, послідовності оплати і поставок сировини.

УДК 519.876.2:336

МЕТОД СПЕКТРАЛЬНОЙ КЛАСТЕРИЗАЦИИ ПЛАТЕЖЕЙ И ПОСТАВКИ СЫРЬЯ ДЛЯ ПЛАНИРОВАНИЯ АУДИТА СООТВЕТСТВИЯ

Нескородева Т. В. – канд. техн. наук, доцент, Донецкий национальный университет имени Васыля Стуса, Винница, зав. кафедры компьютерных наук и информационных технологий, Винница, Украина.

Федоров Е. Е. – д-р техн. наук, доцент, профессор кафедры робототехники и специализированных компьютерных систем, Черкасский государственный технологический университет, Черкассы, Украина.

АННОТАЦИЯ

Актуальность. В настоящее время аналитические процедуры, используемые в ходе аудиторской проверки, базируются на методах интеллектуального анализа данных. В работе решается задача повышения результативности и эффективности аналитических процедур аудита путем кластеризации на основе спектрального разложения. Объектом исследования является процесс аудита соответствия последовательностей оплаты и поставок сырья.

Цель. Целью работы является повышение результативности и эффективности аудита за счет метода спектральной кластеризации последовательностей оплаты и поставок сырья при автоматизации процедур проверки их соответствия.

Методы. Сформированы вектора признаков для объектов последовательностей оплаты и поставок сырья, которые затем используются в предложенном методе. Созданный метод усовершенствует традиционный метод спектральной кластеризации за счет автоматического определения количества кластеров на основе правила объясненной и выборочной дисперсии; автоматического определения параметра масштаба на основе локального масштабирования (используется правило K-ближайших соседей); устойчивости к шуму и случайным выбросам за счет замены метода k-средних модифицированным методом PAM, т.е. замены центроидной кластеризации медоидной кластеризацией. Как и в традиционном подходе данные могут быть разрежены, а кластера могут иметь разные форму и размер. Выбраны характеристики оценивания качества спектральной кластеризации.

DOI 10.15588/1607-3274-2021-1-13

Результаты. Предложенный метод спектральной кластеризации был программно реализован в пакете MATLAB. Полученные результаты позволили исследовать зависимость значений параметров на качество кластеризации.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного метода и позволяют рекомендовать его для использования на практике при решении задач аудита. Перспективы дальнейших исследований могут заключаться в создании интеллектуальных параллельных и распределенных компьютерных систем общего и специального назначения, которые используют предложенный метод для задач сегментации, машинного обучения и распознавания образов.

КЛЮЧЕВЫЕ СЛОВА: планирование аудита, кластеризация, спектральное разложение, медоиды, последовательности оплаты и поставок сырья.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Neskorodieva T. Forecast Method for Audit Data Analysis by Modified Liquid State Machine / T. Neskorodieva, E. Fedorov, I. Izonin // The 1st International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2020), Khmelnytskyi, Ukraine, 10–12 June, 2020: proceedings. – CEUR-WS. – 2020 – Vol. 2623. – P. 25–35.
2. Neskorodieva T. Method for Automatic Analysis of Compliance of Expenses Data and the Enterprise Income by Neural Network Model of Forecast. / T. Neskorodieva, E. Fedorov // The 2nd International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT&DS-2020), Lviv-Shatsk, Ukraine, 2–3 June, 2020: proceedings. – CEUR-WS, Volume I: Main Conference. – 2020. – Vol. 2631. – P. 145–158.
3. Brusco M. J. A Comparison of Latent Class, K-means, and K-median Methods for Clustering Dichotomous Data. / M. J. Brusco, E Shireman, D. Steinley // Psychological Methods. – 2017. – Vol. 22 (3). – P. 563–580. DOI: 10.1037/met0000095.
4. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – New York : Plenum Press, 1981. – 256 p. DOI: 10.1007/978-1-4757-0450-1.
5. Fu Z. Color Image Segmentation Using Gaussian Mixture Model and EM Algorithm. / Z. Fu, L. Wang // Multimedia and Signal Processing, 2012. – P. 61–66. DOI: 10.1007/978-3-642-35286-7_9.
6. Ester M. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // Second International Conference on Knowledge Discovery and Data Mining (KDD), Portland, Oregon, August 2–4, 1996: proceedings. – AAAI Press. – P. 226–231.
7. OPTICS: Ordering Points to Identify the Clustering Structure / [M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander] // International Conference on Management of Data and Symposium on Principles of Database Systems. Philadelphia, Pennsylvania, USA, May, 1999: proceedings. – Association for Computing Machinery : New York, NY, United States, 1999. – P. 49–60.
8. Mirkin B. G. Clustering for Data Mining: A Data Recovery Approach / B. G. Mirkin // Boca Raton, FL: CRC Press, 2005. – 277 p. DOI: 10.1201/9781420034912.
9. Aggarwal C. C. Data Clustering: / C. C.Aggarwal, C. K. Reddy // Boca Raton, FL: CRC Press , 2014. – 620 p. DOI: 10.1201/9781315373515.
10. Diagnostic Rule Mining Based on Artificial Immune System for a Case of Uneven Distribution of Classes in Sample / [S. Subbotin, A. Oliinyk, V. Levashenko, E. Zaitseva] // Communications. – 2016. – Vol. 3. – P. 3–11.
11. Fedorov E. Development of technique for face detection in image based on binarization, scaling and segmentation methods. /[E. Fedorov, T. Utkina, O. Nechyporenko, Y. Korpan] // Eastern-European Journal of Enterprise Technologies, Vol. 1/9, 2020. – P. 23–31. DOI: 10.15587/1729-4061.2020.195369.
12. He J. Modified ART 2A Growing Network Capable of Generating a Fixed Number of Nodes / J. He, A.-H. Tan, Ch.-L. Tan // IEEE Transactions on Neural Networks. 2004. Vol. 15(3). – P. 728–37. DOI: 10.1109/TNN.2004.826220.
13. Andrew Y. Ng. On spectral clustering: Analysis and an algorithm / Y Ng. Andrew, Jordan I. M., Y. Weiss // In Advances in neural information processing systems. – 2002. – P. 849–856.
14. Ulrike V. L. A tutorial on spectral clustering. / V. L. Ulrike// Statistics and computing. – 2007. – Vol. 17(4). – P. 395–416. DOI: 10.1007/s11222-007-9033-z.
15. Fabien L. Spectral clustering of linear subspaces for motion segmentation / L. Fabien, C Schnörr // 12th International Conference on Computer Vision (ICCV'09), Sep 2009, Kyoto, Japan: proceedings. – IEEE, pages to-appear, 2009. DOI: 10.1109/iccv.2009.5459173.
16. Tao Xiang. Spectral clustering with eigenvector selection. / X. Tao, G. Shaogang // Pattern Recognition. – 2008. – Vol. 41(3). – P. 1012–1029. DOI: 10.1016/j.patcog.2007.07.023.
17. Tao Xiang. Spectral clustering with eigenvector selection / X. Tao, G. Shaogang // Pattern Recognition. – 2008. – Vol. 41(3). – P. 1012–1029. DOI: 10.1016/j.patcog.2007.07.023.
18. Feng Z. Spectral clustering with eigenvector selection based on entropy ranking / [Z. Feng, J. Licheng, L. Hanqiang et all] // Neurocomputing. – 2010. – Vol. 73(10–12). – P. 1704–1717 DOI: 10.1016/j.neucom.2009.12.029.
19. Feng Zhao. Spectral clustering with eigenvector selection based on entropy ranking. / [Feng Zhao, J. Licheng, L. Hanqiang et al.] // Neurocomputing. – 2010. – 73(10–12). – P. 1704–1717. DOI: 10.1016/j.neucom.2009.12.029.

DOI 10.15588/1607-3274-2021-1-13