# Fuzzy Clustering Algorithms for Data Mining in the Economic Diagnostics of IT Projects

Elena V. Chertina [1], Varery F. Shurshev[2], Anastasia E. Kvyatkovskaya[3]
Institute of Information Technologies and Communications
Astrakhan State Technical University
Astrakhan, Russia
[1]saprikinae_1912@mail.ru, [2]v.shurshev@mail.ru, [3]zima00@list.ru

***Abstract*— The article is concerned with the possibilities of distribution of innovative IT projects using fuzzy clustering algorithms. Comparative analysis of two basic algorithms - the FCM algorithm and Gustafson-Kessel - are carried out. The clustering procedure for each algorithm is shown, as well as the graphic results of them. There was an assessment of the quality of clustering with the distribution coefficient, the entropy of the classification, and the Hie - Beni index. It is concluded that the usage of the Gustafson-Kessel algorithm allows to get better results for the problem solution of IT projects' classification for their economic diagnostics.**

***Keywords*— *IT project; fuzzy clustering; Gustafson-Kessel algorithm; FCM***

## I. Introduction

In conditions of dynamic development of the digital economy, the informatization of processes is important. We can observe the rapid growth of start-ups, which offer modern applied IT solutions that accelerate the economic, technological, service and other processes each of business and people. The high concentration of start-ups in the IT industry led to the development of a system of venture capital funds, most of whose investments are distributed to IT projects.

The reason for this is that to implement of R&D in the IT project the technology of rapid results is being used now. The technology allows to significantly shorten the period of output of the final product to the stage of commercialization. All this makes the IT sector the most attractive both from the point of view of developers and financial investments.

However, the process of developing and implementing a new IT project can be influenced by various external and internal factors that generate uncertainty of the final result and of the success of its commercial implementation. And for a venture investor, an important aspect is the investment risk profile acceptable to him.

In this context, venture funds have the task of careful economic diagnosis of projects aimed at determination of IT projects' level investment prospects and investment risk for decision-making on investment.

Practice and work review [1, 2] shows that the most frequently used investment indicators for economic diagnostics of deferent projects are net present value (NPV), profitability index (PI), internal rate of return (IRR), payback period (PP). The use of such indicators for economic diagnostics of an IT start-up is difficult, as for the decision-making on investment it is necessary to take into account not only the financial component of the project, but also risks, finance, marketing and others.

This means that the IT project needs to be evaluated according to certain groups of criteria. Multicriteria evaluation of projects is carried out by experts subject to consistency of options. Expert opinions have linguistic descriptions of the type "high", "medium", "low", which are expressed quantitatively on a scale of 0 to 1. The obtained aggregated expert opinions can be used as signs of classification of the set of IT projects. Thus, a selection of IT projects can be divided into groups of projects with a certain set of similar characteristics that allow one to judge the investment prospects of an IT project. Such a procedure can be carried out using the methods of cluster analysis.

## II. Problem statement

For the purposes of this study, we define the following. The is a set of IT projects $P = \{p_1, ..., p_n\}$, estimated by indicators $L_1 - L_6$ (*$L_1$ – novelty of the project relevance, $L_2$ – the degree of risk, $L_3$ – the characteristic of the scientific and technical product, $L_4$ – market potential, $L_5$ – , the evaluation of project feasibility, $L_6$ – economic efficiency).* The evaluation is carried out by an expert group at discrete instants of time $t_1, ..., t_l$ . The mathematical statement of the task is represented as follows.

1. It is required to distribute a set of IT projects $P$, each of which is characterized by six characteristics $\{L_1^j, ..., L_6^j\}$, into three non-overlapping clusters (groups on investment prospects (IP)) $K = \{K_1, ..., K_3\}$ ($K_1$ – IT projects with a high level of IP; $K_2$ – IT projects with a medium level of IP recommended for revision; $K_3$ – IT projects with a low level of IP recommended for refusal to finance)

2. Select the most appropriate clustering algorithm (1), by evaluating the quality of clustering:

The reported study was funded by RFBR according to the research project № 18-37-00130

978-1-5386-5824-6/18/$31.00 ©2018 IEEE

$$\forall P, L, K \exists \Lambda_C : P \to K \quad (1)$$

It should be noted that the fuzzy multivariate type of expert judgments in the implementation of the expert evaluation procedure generates uncertainty that will affect the structure of the cluster. In addition it will be difficult to range the *j*-th IT project only to one of the clusters $\{K_1,...,K_3\}$.

This problem can be solved by using of the fuzzy clustering method [3], which differs in determining the membership degree of the project $p_j$ to each cluster and based on the theory of fuzzy sets by Zade [4].

## III. Analysis of fuzzy clusterization algorithms

After analyzing the fuzzy clustering algorithms in the studies [5,6], we came to the conclusion that the presented algorithms can be conditionally divided into two main groups. The first group is the algorithms that form clusters of spherical shape. The second group is algorithms that form clusters in the form of hyperelipsoids of different orientations.

As the basic algorithms of these groups, we choose the fuzzy c-mean (FCM) algorithm and the Gustafson-Kessel algorithm, respectively. All other algorithms of fuzzy clustering are their derivatives [7].

If you use fuzzy clustering, the selected three groups $\{K_1,...,K_3\}$ will be fuzzy clusters, for convenience we will denote them by $\{\tilde{K}_1,...\tilde{K}_3\}$. Then, fuzzy clusters will be described by a fuzzy partition matrix of the following form (2) [8]:

$$F = \left[\mu_{k,i}\right], \quad (2)$$

where $\mu_{k,i} \in [0;1], k = \overline{1,n}$ – membership function of $k$-th IT project with a set of characteristics $\left(L_1^k,...,L_6^k\right)$ to clusters $\tilde{K}_1,...\tilde{K}_3, c = \overline{1,3}$.

So here it is a conclusion that every IT project having different membership degrees can be assigned to each of the three clusters. In this case, it is necessary to fulfill the following conditions (3):

$$\begin{cases} \sum_{i=1}^{l} \mu_{k,i} = 1, k = \overline{1,n} \\ 0 < \sum_{k=1}^{n} \mu_{k,i} < n, i = \overline{1,l} \end{cases}. \quad (3)$$

Now let us show the main distinguishing characteristics of the algorithms under consideration.

In the FCM method, the minimization of the functional has the form (4) [9]:

$$\Im = \sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m \left\|p_k - v_i\right\|_A^2 , \quad (4)$$

where $V = \left[v_1,...,v_l\right], v_i \in R^n$ - cluster center vector, and $D_{ikA}^2 = \left\|p_k - v_i\right\|_A^2 = \left(p_k - v_i\right)^T A\left(p_k - v_i\right)$ - distance matrix to cluster centers.

The quantities in (4) can be determined from expressions (5, 6):

$$\mu_{k_i} = \frac{1}{\sum_{j=1}^{l}\left(D_{ikA}/D_{jkA}\right)^{2/(m-1)}}; \quad (5)$$

$$v_i = \frac{\sum_{k=1}^{n}\mu_{k,i}^m p_k}{\sum_{k=1}^{n}\mu_{k,i}^m}, \quad (6)$$

where $m$ – exponential weight.

The condition for stopping this algorithm of fuzzy clustering is $\left\|F - F^*\right\| < \varepsilon$, where $\varepsilon$ - is given by decision maker.

The Gustafson-Kessel algorithm differs in that it has its own matrix $A_i$. In accordance with [10] we have the expression (7):

$$D_{ikA}^2 = \left\|p_k - v_i\right\|_{A_i}^2 = \left(p_k - v_i\right)^T A_i\left(p_k - v_i\right) \quad (7)$$

Then the functional $\Im$ will have the form (8):

$$\Im = \sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m\left(p_k - v_i\right)^T A_i\left(p_k - v_i\right) \quad (8)$$

The functional in the form (8) can not be minimized by $A_i$, since it is linear by $A_i$. Therefore, in order to obtain an acceptable solution, it is necessary that $\left\|A_i\right\| < \rho_i, \rho > 0$. It means, that should restrict the determinants of matrices $A_i$. Then the fuzzy covariance matrix for the i-th cluster will be determined as follows (9):

$$F_i = \frac{\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m\left(p_k - v_i\right)\left(p_k - v_i\right)^T}{\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m}. \quad (9)$$

For the next stage of the study, 50 IT projects were evaluated. The expert evaluations were made consistent, there was no affiliation between the experts. Given data for implementing the algorithms are as follows: $m=2$, $c=3$, $\varepsilon = 1\cdot e^{-6}$, matrix $P$ is an aggregated expert evaluation of the criteria considered above $\left\{L_1^j,\ldots,L_6^j\right\}$.

## IV. Implementation of fuzzy clustering algorithms

### A. The FCM algorithm

Formally, algorithm FCM (fuzzy c-average) can be represented in the form of a flowchart, which is shown in Fig. 1.

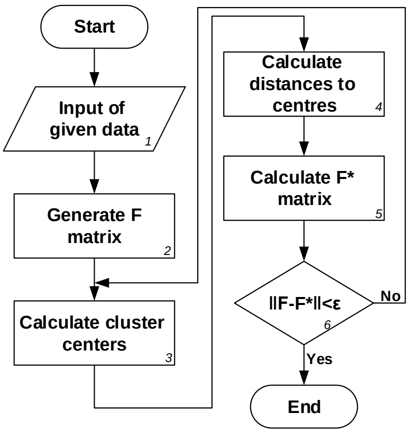
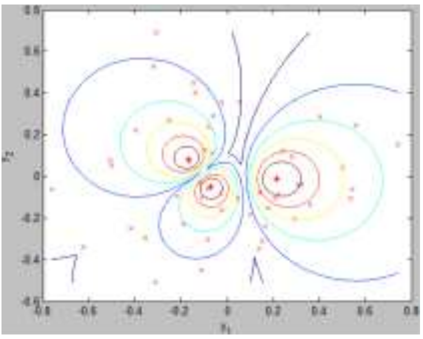


Fig. 1. Flowchart of the algorithm for clustering IT projects (FCM)

Fig. 2 shows the visualization of the results obtained using the Principal Component Analysis (PCA, implemented in the SOMToolbox of the Matlab engineering calculation environment) [11].



.

Fig. 2. Displaying FCM results using the PCA method

### B. The Gustafson-Kessel algorithm

After that, the Gustafson-Kessel algorithm is implemented, the block diagram of which is shown in Fig. 3.

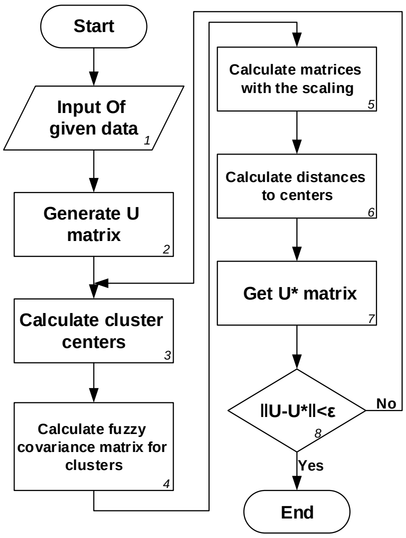

Start
Input Of given data
1
Generate U matrix
2
Calculate cluster centers
3
Calculate fuzzy covariance matrix for clusters
4
Calculate matrices with the scaling
5
Calculate distances to centers
6
Get U* matrix
7
||U-U*||<ε
8
No
Yes
End

Fig. 3. Flowchart of the algorithm for clustering IT projects (the Gustafson-Kessel algorithm)

It took 141 iterations (until the breakpoint of the algorithm stopped) to solve the task of fuzzy clustering by the Gustafson-Kessel method.

Fig. 4 shows the results of clustering by the Gustafson-Kessel method using PCA.

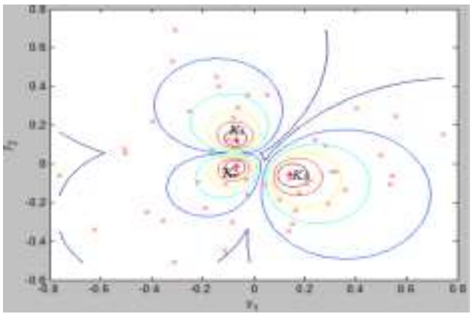


Fig. 4. Displaying the results of the Gustafson-Kessel algorithm by the PCA method

### C. Evaluation of clustering quality

Researches [12] propose to use the following indicators for evaluation of clustering quality.

1. The partition coefficient, calculated by the formula 10:

$$R_1 = \frac{1}{n}\sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^2 . \qquad (10)$$

It is used as a measure of fuzziness (the higher it is, the better assessment of fuzziness and clustering indirectly), but it does not take into account the pairwise distances needed to evaluate compactness and separation. Therefore, another indicator was proposed.

2. The classification entropy (11):

$$R_2 = \frac{1}{n}\sum_{i=1}^{l}\sum_{k=1}^{n}\mu_{k,i}\log\left(\mu_{k,i}\right). \qquad (11)$$

This indicator varies within $0 \le R_2 \le \ln l$. The main purpose of the application of indicators $R_1$ and $R_2$ – search for the most acceptable number of clusters in an unclear partition. But as both indicators depend on the number of clusters $l$, that are suitable for comparing partitions with only the same number of clusters.

3. Xie and Beni's Index (12):

$$R_3 = \frac{\sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m \left\|p_j - v_i\right\|^2}{n\min_{i,j}\left\|p_j - v_i\right\|^2}. \qquad (12)$$

This coefficient is most suitable for estimating the compactness and separability of clusters in a fuzzy partition. It allows to judge the adequacy of the results obtained

The table shows the results of assessing the quality of clustering using two algorithms with the help of the considered indicators.

TABLE I. RESULTS OF THE EVALUATION OF THE QUALITY OF CLUSTERING

| Indicator | FCM algorithm | The Gustafson-Kessel algorithm |
|---|---|---|
| $R_1$ | 0,405 | 0,623 |
| $R_2$ | 1,022 | 0,751 |
| $R_3$ | 1,038 | 1,183 |

Table 1 shows that FCM has a smaller value $R_1$, the large value of entropy and its coefficient Hie-Beni $R_3$ exceeds the analogous indicator of the Gustafson-Kessel algorithm. Thus, to solve the task posed in the study, the most preferred algorithm is Gustafson-Kessel's fuzzy clustering.

In addition, the advantage of the Gustafson-Kessel algorithm is that it forms an adaptive form for each cluster, which makes it possible to order objects on clusters more correctly.

## V. Conclusion

The conducted research allowed to achieve the following results:

- there has been proved the necessity of fuzzy clustering using for solving the problem of economic diagnostics of IT projects in particular of determining the level of investment prospects;
- there has been carried out the analysis of two basic fuzzy clustering algorithms Gustafson-Kessel and FCM and also the features of its functional were considered;
- there was carried out the practical implementation of the considered algorithms for 50 IT projects with aggregated expert estimates;
- there was carried out an evaluation of clustering quality and was made a conclusion about the preference for using the Gustafson-Kessel algorithm.

The proposed approach will allow to formalize the uncertainty and risk in the economic diagnostics of IT projects, as well as improve the effectiveness of the financial decisions made by venture investment funds, as well as other investment companies.

## References

[1] Kulikov D., Kucherov A.A. Formation and development of methods for assessing the effectiveness of innovation projects [Electronic resource] Modern problems of science and education. 2015. No. 1. Available at: https://www.science-education.ru/ru/article/view?id=19451 (Accessed 14 May 2018)

[2] Malova O.T. Approaches to the evaluation of innovative projects. *Zhurnal Educatio* [Educatio Journal]. 2015, No. 3 (10), pp.140- 142 (in Russian)

[3] Bezdek J.C., Ehrlich R., Full W. FCM: The Fuzzy c-Means Clustering Algorithm. Computers & Geoscience. 1984, Vol. 10. No. 2-3, pp. 191-203.

[4] Zadeh L. *Ponyatie lingvisticheskoj peremennoj i ego primenenie k prinyatiyu priblizhennyh reshenij* [The concept of a linguistic variable and its application to the adoption of approximate solutions]. Moscow. Peace Publ. 1976. 165 p.

[5] Neysky I.M. Classification and comparison of clustering methods [Electronic resource]. Available at: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (Accessed 14 May 2018).

[6] Jain A.K., Murty M.N., Flynn P.J. Data Clustering: A Review. ACM Computing Surveys. 1999, Vol. 31, no. 3, pp. 264–323.

[7] Rozilawati Binti Dollah, Aryati Binti Bakri, Mahadi Bin Bahari, Pm Dr. Naomie Binti Salim. Feasibility Study Of Fuzzy Clustering Techniques In Chemical Database For Compound Classification. 2006.

[8] Shtovba S.D. *Proektirovanie nechetkih sistem sredstvami MATLAB* [Designing fuzzy systems using MATLAB]. Moscow. Goryachaya liniya – Telecom Publ. 2007. 288 p.

[9] Bezdek J.C., Dunn J.C. Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Dustrubutions. IEEE Transactions on Computers. 1985, pp. 835-838.

[10] Gustafson D.E., Kessel W.C. Fuzzy clustering with fuzzy covariance matrix. In Proceedings of the IEEE CDC, San Diego. 1979, pp. 761-766.

[11] Jolliffe I.T. Principal Component Analysis. Springer Series in Statistics, 2nd ed. NY. Springer. 2002, XXIX, 487 p.

[12] Xie X.L., Beni G.A. Validity measure for fuzzy clustering. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. 1991,Vol. 3(8), pp. 841-846.