DOI: 10.24143/2072-9502-2020-2-95-108
UDC 004.43:519.712

MATHEMATICAL METHODS AND ALGORITHMS FOR DATA MINING IN IT PROJECT INVESTMENT ATTRACTIVENESS ESTIMATION [1]

E. V. Chertina, A. E. Kvyatkovskaya, L. B. Aminul, K. I. Kvyatkovskii

*Astrakhan State Technical University,
Astrakhan, Russian Federation*

Abstract. The article is concerned with developing mathematical support and algorithms for solving the problem of economic diagnostics of enterprises. IT-companies and start-ups (IT projects) that have special characteristics during the growth period were selected as the object of research. Based on the system analysis of data domain there has been developed a system of quantitative and qualitative characteristics to identify the economic state of the IT companies and start-ups in the external and internal environment. Scales of indices of different nature have been determined. Methods to introduce order and equivalence relations for the found peer companies have been given in order to compare their proximity to the analyzed company. Metrics used for comparing the companies are considered taking into account the quantitative and qualitative characteristics. The possibilities of distributing innovative IT projects using fuzzy clustering algorithms are considered. The comparative analysis of two basic algorithms - Fuzzy Classifier Means algorithm and Gustafson - Kessel algorithm - has been given. The clustering procedure for each algorithm is shown, as well as the graphic results of their operation. There was done the clustering quality assessment using a distribution coefficient, entropy of classification, and Hie-Beni index. It has been inferred that using Gustafson - Kessel algorithm provides better results for solving the problem of splitting IT projects for their economic diagnostics.

Key words: IT start-up, case-based reasoning, precedents, peer company, comparative method, fuzzy clustering, Gustafson - Kessel algorithm, FCM.

For citation: Chertina E. V., Kvyatkovskaya A. E., Aminul L. B., Kvyatkovskii K. I. Mathematical methods and algorithms for data mining in IT estimation of project investment attractiveness. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*. 2020;2:95-108. (In Russ.) DOI: 10.24143/2072-9502-2020-2-95-108.

Introduction

The task of estimation is one of the low-formalized tasks of economic systems management under conditions of uncertainty. The results of various property objects estimation are the basis for most of decision making in the private and public sectors under current economic conditions. Analog method is one of the most effective estimation methods. It is based on comparing the company with the most suitable analog ones, choosing the relevant prototype and transferring its economic properties and trends to the object of research. Basing on the global trend towards digitalization of economic sectors, informatization process occupies a special place. Now there are about 5000 small IT companies in Russia. Taking into account interest of venture funds and large IT -companies in buying of start-ups the estimation task is very important both for business and information technologies development in Russia.

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

We can observe the rapid growth of start-ups which offer modern applied IT solutions accelerating the economic, technological, service and other processes both for business and people. The high concentration of start-ups in the IT industry led to the development of venture capital fund system, most investments of which are distributed to the 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 helps to shorten significantly 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.

At the same time, the use of the analog method for an estimation of IT - start-ups value using information technology is constrained by lack of information models and mechanisms that support this process:

– absence of constantly updated knowledge base about the peers required for comparison;

– absence of mechanisms for collecting information from any open source for supplementing the knowledge base;

– absence a frame of reference for peer companies that contain heterogeneous information;

– absence of justification of metrics for calculating the "proximity" of peer companies.

If we talk about an estimation of IT projects investment attractiveness, then, due to the uncertainty and risk, the application of investment analysis traditional methods to projects of this kind can lead to unreliable results, since traditional methods do not take into account the innovative component of projects.

In this regard, the development of mathematical methods and algorithms providing a qualitative IT projects estimation is an urgent scientific and practical task.

The purpose of the study is to develop and justify mathematical methods and algorithms providing a decision making support process for the value and investment attractiveness of IT companies (projects) estimation using data mining tools.

The purpose setting divides the research into 2 main stages:

1. Development of a mathematical device for the IT start-up estimation, using case-based reasoning method and comparing analogues.

2. Development of a procedure of estimation of IT projects investment attractiveness using cluster analysis tools.

The study is focused on an IT company or a start-up that has special values of economic characteristics during the growth period which are not specific for ordinary enterprises. In the course of the study we agree that an IT start-up and an IT-project are identical concepts.

The estimation of IT start-up value

Identifying start-up characteristics. The set of criteria required for economic diagnostics of an IT company was determined by the example of a startup. A startup estimation method depends on the stage of: preseed; seed; series A.

At the stage of Preseed, the estimation takes place at a fixed rate of a business angel or an accelerator, the main task of which is to speed up the delivery of early stage projects to the first investor, to refine and help them. It is rather difficult to structure the indicators at this stage, since the start-up does not have formal indicators that allow the construction of a financial model, but only meets the following requirements: an achievable market volume of at least 300 million rubles, deadline - 3-5 years; team of the project - at least two people; the presence of a working MVP (minimum viable product) - minimum viable product.

At the stage of Seed, the objective is to scale the business (increase the number of customers, customer segments, geography, etc.). The estimation can be viewed from two sides, determining how much investment is needed, based on the team's costs per month and the investor's expectations through a specific time period. It is possible to use the indicators accepted in the international practice for the analysis of investment projects, for example, - NPV (Net Present Value).

Stage A is the stage of active growth and increasing of the company. At this stage, the following indicators are highlighted: Cash-flow, multiplier, discount rate, scale-out limiters.

Comparing the formation of estimates in three stages, it must be taken into account that the accelerators note that the systematization of the estimation for the Preseed stage is an impossible task, since here the subjective assessment formed after personal communication with the creators is more significant. Therefore, we will consider the Seed & Series A stages. Therefore, we will consider the Seed & Series A stages.

The papers of B. Payne [1, 2] and S. Nasser [3] are the most popular papers in this area of research, which are much talked in online research. They are devoted to the valuation of companies, including start-ups to various stages of investment.

To analyze the selected stages, we use five commonly used estimation methods of startups, summarizing the indicators on which they are based. The methods were determinate after undertaken studies in the largest business incubators in Russia, which mark the feasibility and adaptability of the selected methods to the Russian conditions. It should be noted that most methods are based on data from comparable companies or basic estimates: the Berkus method, the method of summation of risk factors, the venture capital method, the discounted cash flow method, the comparison method.

The characteristics that generate the above methods are grouped as qualitative and quantitative, it was done for the subsequent structuring and scaling. In total, 15 quantitative and 14 qualitative indicators were selected, including 9 types of risk (Table 1).

Table 1

Characteristics of start-up

Quantitative characteristics	Qualitative characteristics
Customer Acquisition Cost (CAC), Rub.	Team evaluation
Cash-Flow, Rub.	Scaling drivers
Multiplier	Scaling limiters
Market capitalisation, Rub.	Strategic relationship
Backlog, Rub.	Product introduction or sales start
Operating profit, Rub.	Quality of the prototype
Sensible idea (cost base), Rub.	Managerial risks
ROI (Return On Investment), %	Risks at different stages of business development
Discount rate, %	Political risks
Expected growth rate, %	Marketing risks
Regular monthly income, Rub.	Risks related to financing / raising of capital
Number of persons employed, Piece	Litigation risks
EBITDA (Earnings before interest, taxes, depreciation and amortization), Rub.	International risks
Gross profit, Rub.	Reputational risks
–	Risks associated with a potentially profitable exit from a startup

The estimation system of an IT company under the given set of characteristics will determine a point set in the criteria space that have a formal criterion representation. In order one company to serve as a good analog for other evaluation, it is desirable that they resemble in many characteristics, at the same time it is possible to prioritize, reinforcing the weight significance of a particular characteristic.

Identifying a peer company selecting method

For the selection of peer companies, we apply one of the decision-making methods – the method of case-based reasoning, using knowledge of known situations or cases (precedents), which in our case are peer companies. We define the set (IT) of IT companies considered in the selection of analogues. The information about a set IT is represented in the form $IT = \{it_i, i = \overline{1,n}\}$. To determine the properties-characteristics of each IT company it_i we compare a set of characteristics $K = \{k_j\}, j = \overline{1,m}$. Then each IT company can be represented in a form $it_i = \{f_1(k_1), f_2(k_2), ..., f_m(k_m)\}$, where $f_j(k_j)$ is a characteristic function that defines a subset $k_j^* \subseteq K$ or the i-th IT company.

Once the it_i peer companies are extracted, you need to select the "similarity" to the it* precedent, describing the degree of proximity by the formula

$$R(it_i, it^*) = \frac{\sum_{j=1}^{m} \rho(f_j^i(k_j), f_j^*(k_j)) \cdot w_j}{\sum_{j=1}^{m} w_j},$$

where $p(f_j^i(k_j), f_j^*(k_j))$ – a metric is calculated by m characteristics of analog and precedent $f_j^i(k_j)$ and $f_j^*(k_j)$; w_j – a degree of importance of the j-th characteristics.

The choice of the metric is the most difficult problem. The inhomogeneity of the characteristics does not allow us to introduce an algebra of operations on the given set. The most famous is the mathematical method of nearest neighbor [4], which is able to measure the degree of proximity for any characteristic:

$$mnear(it^*) = \underset{it \in IT}{\text{argmax}} \sum_{j=1}^{m} [f_j^i(k_j) = f_j^*(k_j)] \cdot w_j,$$

where $[f_j^i(k_j) = f_j^*(k_j)]$ – is an error indicator that takes a logical value to a number by the rule [false] = 0, [true] = 1.

For quantitative characteristics it is also possible to use Euclidean distance or the Manhattan metric, provided that all characteristics are reduced to a single measurement scale or normalized.

If the exact match of characteristics is not required (or it is not attainable), it is possible to use the Zhuravlev metric

$$mzur(it^*) = \sum_{j=1}^{m} if((\left|f_j^i(k_j) - f_j^*(k_j)\right| < \varepsilon),\ then\ 1,\ else\ 0)$$

where ε is a given level of deviation of j-characteristics of the analogue and precedent from each other.

The number of characteristics has an effect on output error, since the curse of dimension may arise: according to the law of averages, the sums of a large number of deviations are very likely to have very close values. This fact subsequently leads up to the need to form a set of informative characteristics, but will require retrospective observations for them to form a sample of data, to reveal the dependence or multicollinearity.

For qualitative characteristics, it is possible to use the measure of Hamming's similarity by determining the maximum number of matching characteristics of a precedent and an analogue. If you cannot enter a metric, various proximity measures are used.

After the database of precedents is formed in any way - manual or automated, it is possible to allocate relationships of order and equivalence for the objects filling it [5]. Using a geometric approach to the solution of this problem, the importance of which was stressed by D. A. Pospelov [6], it is possible to represent analogs and precedents as independent information objects and, in the future, to compare them both by individual characteristics and in general.

Analyzing analogues using the equivalence relation, the original set is divided into equivalence classes $[it^*] \subset IT$ of element $it^* \in IT$ in the form of a subset of elements equivalent to $it^* : [it^*] = \{it \in IT \mid it \sim it^*\}$.

The classes of analogs can represent both nominal and ordinal scales. In the first case, they can be constructed in two ways: by clustering and using expert estimates. In the second case it is possible to use the partitioning of the original set into Pareto classes with subsequent ordering of these classes.

When analogues analyzed using the order relationship, precedents are arranged by rank in the absence of an accurate analog. Let's highlight the following decision-making tasks, using the ranking of analogues along the proximity to the precedent:

– the task of ranking analogs based on knowledge of their states at a given time $t^*(it_i, k_j), i = \overline{1,n}, j = \overline{1,m}$;

– the task of ranking analogues based on knowledge of their states at different times (for example, corresponding to the stages) $it^*(t_g, k_j), g = \overline{1,s}, j = \overline{1,m}$;

– the task of ranking analogues according to a given characteristic $k^*(it_i, t_g), i = \overline{1,n}, g = \overline{1,s}$;

– the task of ranking analogues on aggregate characteristics $k(it_i, t_g), i = \overline{1,n}, g = \overline{1,s}$.

In the latter case, the equal importance of characteristics is considered when the decision-maker can or cannot reliably establish priorities between them. In the case of equal characteristics, a set of incommensurable undominated alternatives are formed - the Pareto ITP set. Thus, in the case of the solution is selected not just one but many peers, which ultimately makes the final decision difficult. In this case, apply mathematical methods that narrow the Pareto set, for example, the method of median distributions [7, 8]. The advantage of the method is the combination of qualitative and quantitative assessments.

It is also possible to construct various functions for selecting C^K (*IT*) and C^D (*IT*) in case the absence of information about the relative importance of characteristics and the availability of characteristics of both quantitative and qualitative type. They narrow the Pareto set and take into account only the mutual relations between the estimates of the analogs without taking into account the absolute values of the differences in the estimates by characteristics.

For two analogues $it_i,\ it_l \in IT,\ i, l = \overline{1,n}$ we define the number of characteristics by which it_l has more proximity to it^* than it_i. For analogs whose maximum is this number, we define on the IT-set a numerical function $q_{i,l} = q(it_i, it_l)$ taking values corresponding to the maximal numbers found, where $q(it_i,\ it_l)$ is the number of characteristics over which it_l exceeds the variant it_i, in other words, is closer to the precedent.

A choice function for the C^K was constructed, considering the number of dominant characteristics of the analogue, which are close to the precedent, choosing the maximum values of the row of the matrix $Q_{IT} = \{q_{i,l}\}$ and then separating the minimal of them:

$$C^K(IT) = \{it_i \in IT \mid i \in Arg \min_i q_i, i = \overline{1,n}\},$$

where $q_i = \max_l q_{i,l}$.

As a result, a subset of analogues is formed, which have a greatest number of characteristics close to the precedent. The resulting subset has less potency than the Pareto set, and $C^K(IT) \subseteq IT^P$.

Consider the second method of generating analogues, closed to precedent using Q_{IT} matrix. The dominant index of the set *IT* was defined, equal to($\min_{it \in IT} Q_{IT}(it)$. The value of the choice function C^D (*IT*) is a subset of all variants of $it \in IT$ with a minimum *IT* dominant index:

$$C^D(IT) = \{it^* \in ALT \mid Q_{IT}(it^*) = \min_{it \in IT} Q_{IT}(it)\}.$$

A circular n-tournament selection function C^T was constructed:

$$C^T(IT) = \{it^* \in IT \mid QM_{IT}(it^*) = \min_{it \in IT} QM_{IT}(it)\},$$

where $QM_{ALT}(it_i) = \sum_{l=1}^{n} q_{i,l}$.

This function also narrows the Pareto set, forming a subset of analogues close to the precedent, with $C^T(IT) \subseteq C^K(IT) \subseteq IT^P$.

The next stage is the investment attractiveness estimation of formed IT start-ups set.

Investment attractiveness estimation

Cluster approach to IT projects of investment attractiveness estimation. Let us consider IT project investment attractiveness task in detail.

Practice and work review [9, 10] shows that the most frequently used investment indicators for economic diagnostics of investment attractiveness 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 [11]. 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.

There 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 estimation 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, by evaluating the quality of clustering:

$$\forall P, L, K \exists \Lambda_C : P \to K.$$

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 [12], 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 [13].

Analysis of fuzzy clustering algorithms

After analyzing the fuzzy clustering algorithms in the studies [14, 15], 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 [16].

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 [17]:

$$F=\left[\mu_{k,i}\right],$$

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

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

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 [18]:

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

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 (1) can be determined from expressions

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

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

where m – exponential weight.

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

The Gustafson - Kessel algorithm differs in that it has its own matrix A_i. In accordance with [19] we have the expression

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

Then the functional $\Im$ will have the form

$$\Im = \sum_{i=1}^{l} \sum_{k=1}^{n} \left(\mu_{k,i}\right)^{m} \left(p_k - v_i\right)^T A_i \left(p_k - v_i\right). \tag{2}$$

The functional in the form (2) cannot 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

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

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 $\left\{L_1^j, \ ..., \ L_6^j\right\}$.

Implementation of fuzzy clustering algorithms

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

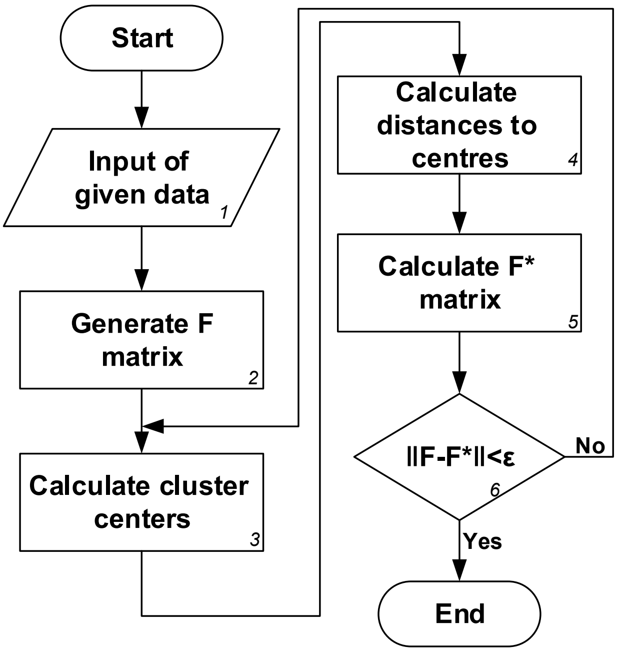

Fig. 1. Flowchart of the algorithm for clustering IT projects (FCM)

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) [20].

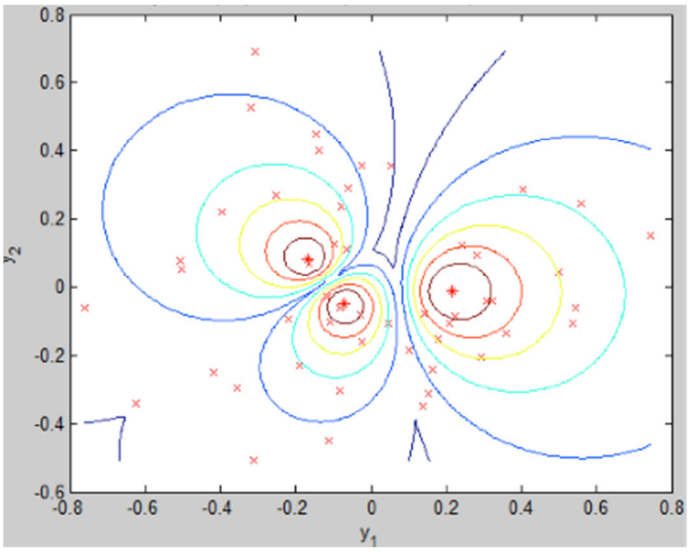

Fig. 2. Displaying FCM results using the PCA method

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

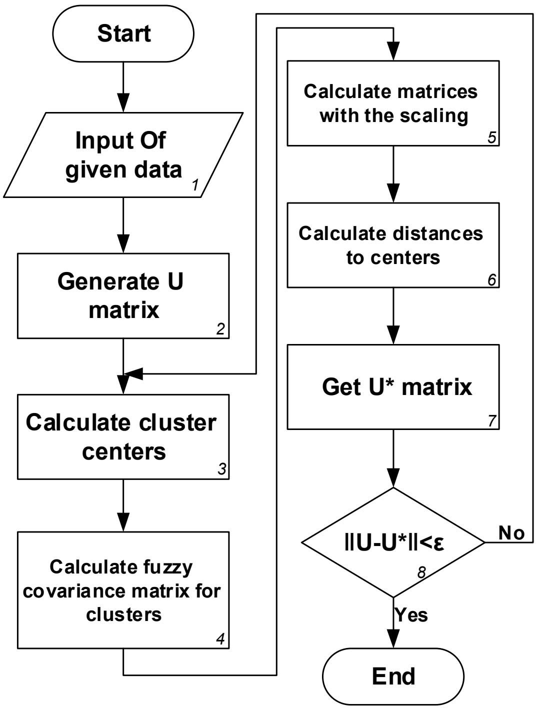
Start
Input Of given data
1
Generate U matrix
2
Calculate cluster centers
3
Calculate fuzzy covariance matrix for clusters
4
Calculate matrices with the scaling
5
Calculate distances to centers
6
Get U* matrix
7
||U-U*||<ε
8
No
Yes
End

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.

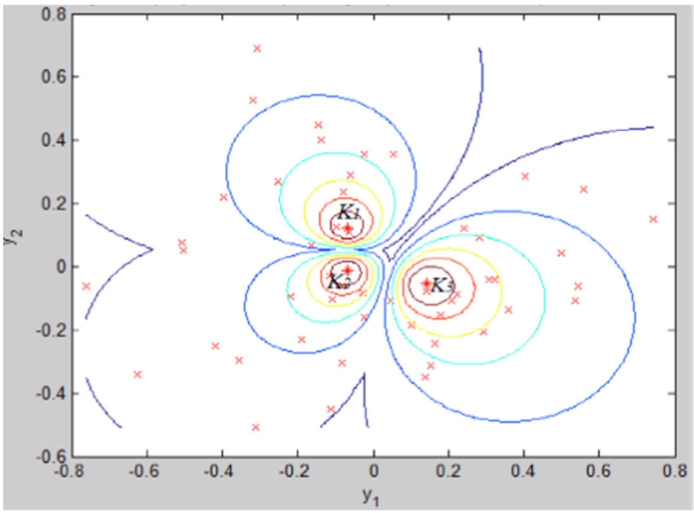

Fig. 4. Displaying the results of the Gustafson - Kessel algorithm by the PCA method

Clustering Quality Assessment. Researches [21] propose to use the following indicators for evaluation of clustering quality.

1. The partition coefficient, calculated by the formula

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

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

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

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

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

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 2

The 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 2 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 of dividing IT projects into groups according to the degree of investment attractiveness, the most preferred 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.

Conclusion

The conducted research allowed to achieve the following results:

– there have been considered the issues of IT companies and startups economic diagnostics in the task of business value estimation based on the use of case- based reasoning method and comparing analogues were considered;

– there have been selected characteristics and considered the issues of metrics and proximity measures choice for quantitative and qualitative characteristics of peer companies;

– there have been presented mathematical methods that arrange set of peer companies by proximity to a precedent;

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

REFERENCES

1. Payne B. *Methods for Valuation of Seed Stage Startup Companies*. Available at: www.angelcapitalassociation. org/blog/methods-for-valuation-of-seed-stage-startup-companies/ (accessed: 21.01.2020).

2. Payne B. *Startup Valuations: The Risk Factor Summation Method*. Available at: http://billpayne. com/2011/02/27/startup-valuations-the-risk-factor-summation-method-2.html (accessed: 21.01.2020).

3. Nasser S. *Valuation For Startups – 9 Methods Explained*. Available at: http://medium.com/parisoma-blog/valuation-for-startups-9-methods-explained-53771c86590e/ (accessed: 24.01.2020).

4. Anand S. S., Hughes J. G., Bell D. A., Hamilton P. *Utilising Censored Neighbours in Prognostication. Workshop on Prognostic Models in Medicine.* Denmark, Aalborg, 1999. Pp. 15-20.

5. Karpov L. E., Iudin V. N. *Metody dobychi dannykh pri postroenii lokal'noi metriki v sistemakh vyvoda po pretsedentam* [Data mining methods for constructing local metrics in systems of deduction by precedents]. Moscow, Izd-vo ISP RAN, preprint № 18, 2006. 21 p.

6. Pospelov D. A. *Modelirovanie rassuzhdenii. Opyt analiza myslitel'nykh aktov* [Modeling of reasoning. Practice in analysis of mental acts]. Moscow, Radio i sviaz' Publ., 1989. 184 p.

7. Kosmacheva I., Kvyatkovskaya I. Y., Sibikina I., Lezhnina Y. Algorithms of Ranking and Classification of Software Systems Elements. *Knowledge-Based Software Engineering: Proceedings of 11th Joint Conference, JCKBSE 2014.* Volgograd, Springer International Publishing, 2014. Pp. 400-409.

8. Pham Quang Hiep, Kvyatkovskaya I. Y., Shurshev V. F., Popov G. A. Methods and Algorithms of Alternatives Ranging in Managing the Telecommunication Services Quality. *Journal of Information and Organizational Sciences*, 2015, vol. 39, no. 1, pp. 65-74.

9. Kulikov D. L., Kucherov A. A. Stanovlenie i razvitie metodov otsenki effektivnosti innovatsionnykh proektov [Formation and development of methods for evaluating effectiveness of innovative projects]. *Sovremennye problemy nauki i obrazovaniia*, 2015, no. 1. Available at: https://www.science-education.ru/ru/article/view?id=19451 (accessed: 30.01.2020).

10. Malova O. T. Podkhody k otsenke innovatsionnykh investitsionykh proektov [Approaches to the assesment of innovative investment projects]. *Mezhdunarodnyj nauchnyj institut «Educatio»*, 2015, no. 3 (10), pp. 140-142.

11. Popov G. A., Kvyatkovskaya I. Y., Zholobova O. I., Kvyatkovskaya A. E., Chertina E. V. Making a choice of resulting estimates of characteristics with multiple options of their evaluation. *Proceedings of 3rd Conference on Creativity in Intelligent Technologies and Data Science, CIT and DS 2019 (Volgograd, Russia, September 16–19, 2019). Part of the Communications in Computer and Information Science book series (CCIS, volume 1083).* Springer, 2019. Part I. Pp. 89-104.

12. 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.

13. Zade L. A. *Poniatie lingvisticheskoi peremennoi i ego primenenie k priniatiiu priblizhennykh reshenii* [Concept of linguistic variable and its application to approximate decision making]. Moscow, Mir Publ., 1976. 165 p.

14. Neiskii I. M. *Klassifikatsiia i sravnenie metodov klasterizatsii* [Classification and comparison of clustering methods]. Available at: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (accessed: 05.02.2020).

15. Jain A. K., Murty M. N., Flynn P. J. Data Clustering: A Review. *ACM Computing Surveys*, 1999, vol. 31, no. 3, pp. 264-323.

16. 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*. Available at: http://eprints.utm.my/id/eprint/4402/ (accessed: 17.12.2019).

17. Shtovba S. D. *Proektirovanie nechetkikh sistem sredstvami MATLAB* [Designing fuzzy systems using MATLAB software]. Moscow, Goriachaia liniia – Telekom Publ., 2007. 288 p.

18. 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.

19. Gustafson D. E., Kessel W. C. Fuzzy clustering with fuzzy covariance matrix. *Proceedings of the IEEE CDC.* San Diego, 1979. Pp. 761-766.

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

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

The article submitted to the editors 13.03.2020

INFORMATION ABOUT THE AUTHORS

Chertina Elena Vitalievna – Russia, 414056, Astrakhan; Astrakhan State Technical University, Candidate of Technical Sciences; Assistant Professor of the Department of Higher and Applied Mathematics; saprikinae_1912@mail.ru.

Kvyatkovskaya Anastasia Evgenievna – Russia, 414056, Astrakhan; Astrakhan State Technical University; Assistant of the Department of Higher and Applied Mathematics; zima00@list.ru.

Aminul Lubov Borisovna – Russia, 414056, Astrakhan; Astrakhan State Technical University; Candidate of Pedagogical Sciences; Assistant Professor of the Department of Higher and Applied Mathematics; aminul.25@mail.ru.

Kvyatkovskii Kirill Igorevich – Russia, 414024, Astrakhan; LLC “Digital water and wastewater treatment plant”; director; k-i-k@mail.ru.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ПРИ ОЦЕНКЕ ИНВЕСТИЦИОННОЙ ПРИВЛЕКАТЕЛЬНОСТИ IT-ПРОЕКТОВ

Е. В. Чертина, А. Е. Квятковская, Л. Б. Аминул, К. И. Квятковский

Астраханский государственный технический университет,
Астрахань, Российская Федерация

Рассматриваются вопросы создания математического обеспечения и алгоритмов для задачи оценки инвестиционной привлекательности компаний. Объектом исследования выбраны IT-компании, в том числе стартапы (IT-проекты), обладающие в период роста особенными характеристиками. На основе системного анализа предметной области разработана система количественных и качественных характеристик для идентификации экономического состояния IT-компаний и стартапов во внешней и внутренней среде. Определены шкалы показателей различной природы. Приведены методы, позволяющие ввести отношения порядка и эквивалентности для найденных компаний-аналогов в целях сравнения их близости к анализируемой компании. Рассмотрены метрики, используемые для сравнения компаний, с учетом количественных и качественных характеристик. Рассмотрены возможности распределения инновационных IT-проектов с использованием алгоритмов нечеткой кластеризации. Приведена сравнительная характеристика двух базовых алгоритмов – алгоритма FCM и Густафсона – Кесселя. Представлена процедура кластеризации по каждому алгоритму, а также графически изображены результаты работы каждого алгоритма. Проведена оценка качества кластеризации с использованием коэффициента распределения, энтропии классификации и показателя Хие – Бени. Сделан вывод, что использование алгоритма Густафсона – Кесселя позволяет достичь более качественных результатов в решении задачи разбиения IT-проектов для цели их экономической диагностики.

Ключевые слова: ИТ стартап, прецедентный подход, прецеденты, одноранговая компания, сравнительный метод, нечеткая кластеризация, алгоритм Густафсона – Кесселя, метод нечеткой кластеризации.

Для цитирования: *Чертина Е. В., Квятковская А. Е., Аминул Л. Б., Квятковский К. И.* Математические методы и алгоритмы интеллектуального анализа данных при оценке инвестиционной привлекательности IT-проектов // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2020. № 2. С. 95–108. DOI: 10.24143/2072-9502-2020-2-95-108.

СПИСОК ЛИТЕРАТУРЫ

1. *Payne B.* Methods for Valuation of Seed Stage Startup Companies. URL: www.angelcapitalassociation.org/blog/methods-for-valuation-of-seed-stage-startup-companies/ (дата обращения: 21.01.2020).

2. *Payne B.* Startup Valuations: The Risk Factor Summation Method. URL: http://billpayne.com/2011/02/27/startup-valuations-the-risk-factor-summation-method-2.html (дата обращения: 21.01.2020).

3. *Nasser S.* Valuation For Startups – 9 Methods Explained. URL: http://medium.com/parisoma-blog/valuation-for-startups-9-methods-explained-53771c86590e/ (дата обращения: 24.01.2020).

4. *Anand S. S., Hughes J. G., Bell D. A., Hamilton P.* Utilising Censored Neighbours in Prognostication // Workshop on Prognostic Models in Medicine. Eds. Ameen Abu-Hanna and Peter Lucas. Denmark, Aalborg, (AIMDM'99), 1999. P. 15–20.

5. *Карпов Л. Е., Юдин В. Н.* Методы добычи данных при построении локальной метрики в системах вывода по прецедентам. М.: Изд-во ИСП РАН, препринт № 18, 2006. 21 с.

6. *Поспелов Д. А.* Моделирование рассуждений. Опыт анализа мыслительных актов. М.: Радио и связь, 1989. 184 с.

7. *Kosmacheva I., Kvyatkovskaya I. Y., Sibikina I., Lezhnina Y.* Algorithms of Ranking and Classification of Software Systems Elements // Knowledge-Based Software Engineering: Proceedings of 11th Joint Conference, JCKBSE 2014. Volgograd: Springer International Publishing, 2014. P. 400–409.

8. *Pham Quang Hiep, Kvyatkovskaya I. Y., Shurshev V. F., Popov G. A.* Methods and Algorithms of Alternatives Ranging in Managing the Telecommunication Services Quality // Journal of Information and Organizational Sciences. 2015. V. 39. N. 1. P. 65–74.

9. *Куликов Д. Л., Кучеров А. А.* Становление и развитие методов оценки эффективности инновационных проектов // Современные проблемы науки и образования. 2015. № 1. URL: https://www.science-education.ru/ru/article/view?id=19451 (дата обращения: 30.01.2020).

10. *Малова О. Т.* Подходы к оценке инновационных инвестиционных проектов // Международный научный институт «Educatio». 2015. № 3 (10). С. 140–142.

11. *Popov G. A., Kvyatkovskaya I. Y., Zholobova O. I., Kvyatkovskaya A. E., Chertina E. V.* Making a choice of resulting estimates of characteristics with multiple options of their evaluation // Proceedings of 3rd Conference on Creativity in Intelligent Technologies and Data Science, CIT and DS 2019 (Volgograd, Russia, September 16–19, 2019). Part of the Communications in Computer and Information Science book series (CCIS, volume 1083). Springer, 2019. Part I. P. 89–104.

12. *Bezdek J. C., Ehrlich R., Full W.* FCM: The Fuzzy c-Means Clustering Algorithm // Computers & Geoscience. 1984. V. 10. N. 2-3. P. 191–203.

13. *Заде Л. А.* Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976. 165 с.

14. *Нейский И. М.* Классификация и сравнение методов кластеризации. URL: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (дата обращения: 05.02.2020).

15. *Jain A. K., Murty M. N., Flynn P. J.* Data Clustering: A Review // ACM Computing Surveys. 1999. V. 31. N. 3. P. 264–323.

16. *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. URL: http://eprints.utm.my/id/eprint/4402/ (дата обращения: 17.12.2019).

17. *Штовба С. Д.* Проектирование нечетких систем средствами MATLAB. М.: Горячая линия – Телеком, 2007. 288 с.

18. *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. P. 835–838.

19. *Gustafson D. E., Kessel W. C.* Fuzzy clustering with fuzzy covariance matrix // In Proceedings of the IEEE CDC. San Diego, 1979. P. 761–766.

20. *Jolliffe I. T.* Principal Component Analysis // Springer Series in Statistics, 2nd ed. NY: Springer, 2002. XXIX. 487 p.

21. *Xie X. L., Beni G. A.* Validity measure for fuzzy clustering // In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. 1991. V. 3 (8). P. 841–846.

Статья поступила в редакцию 13.03.2020

ИНФОРМАЦИЯ ОБ АВТОРАХ

Чертина Елена Витальевна – Россия, 414056, Астрахань; Астраханский государственный технический университет; канд. техн. наук; доцент кафедры высшей и прикладной математики; saprikinae_1912@mail.ru.

Квятковская Анастасия Евгеньевна – Россия, 414056, Астрахань; Астраханский государственный технический университет; ассистент кафедры высшей и прикладной математики; zima00@list.ru.

Аминул Любовь Борисовна − Россия, 414056, Астрахань; Астраханский государственный технический университет; канд. пед. наук; доцент кафедры высшей и прикладной математики; aminul.25@mail.ru.

Квятковский Кирилл Игоревич **–** Россия, 414024, Астрахань; ООО «Цифровой водоканал»; директор; k-i-k@mail.ru.