Leonas Ustinovichius

Galina Shevchenko

THE CLARA METHOD – A NEW APPROACH TO EXPERT VERBAL CLASSIFICATION

Abstract

In project development it is hardly possible to get exhaustive and accurate information. As a result, the situations occur, the consequences of which can be very damaging to the project. Inaccurate evaluation of the strategy related to capital investment and project implementation is one of the reasons why such estimates are not required in practice. Instead, a classification approach may be used for this purpose. Classification is a very important aspect of decision making. In the present paper, a novel algorithm CLARA is offered for ordering multicriteria alternatives. It differs from the existing similar methods in the wider range of application allowing it to be used with various scales of criteria evaluation, a random number of the solution classes, incomplete order on the criteria scales as well as in the considerably rarefied space of the alternatives. The suggested algorithm is more effective in terms of the time spent by an expert. A comparative calculation of the efficiency of the algorithms used in classifying the objects in the order of their significance has shown that CLARA is much more effective than CLANSH and other algorithms. At the same time, its general effectiveness was found to be lower than that of the algorithms CYCLE which has a narrower scope of application.

Keywords

Expert system, decision-making, verbal analysis methods; methods of solving multi-criteria classification problems.

Introduction

In practical use the task of getting expert knowledge can often be formulated similarly to the task of classification, because experts sort objects (alternatives, states of object) through classes of decision [2,6,7]. For example an engineer entity to be classified may have different origin. They can be different physical objects, choice cases or conditions of some object.

Describing the method of assigning an object to a certain class of decisions is complicated because of non-verbality of the strategy used by the expert [5]. Anyway, these non-verbal skills are effectively and promptly used, when the expert solves a task of classification in his sphere of knowledge. Classification is a very important aspect in decision making [1, 10, 12, 13]. One of the tasks preparing a basis for classification is the setting of numerous criteria (attributes), which are capable of describing any object. The scale of all criteria is formed by defining a finite set of possible values. If in certain task the scale of values of one or more criteria is infinite, it can be changed to a finite one by limiting it to a finite set of intervals. Finally, on the basis of expert knowledge classification of definite intervals and its components must be organized i.e. rules must be formulated according which any object can be assigned to one of the predefined classes [11]. The projects classified are described by assessing various efficiency criteria that can be expressed both qualitatively and quantitatively [16].

The purpose of this paper is to demonstrate how multiple criteria can be used in the analysis of facility location problems. The paper begins with an overview, explains the most popular multiple objective analysis methods used in various countries (ORKLASS, DIFKLASS, CIKL etc.), and demonstrates their applications to real-life problems. To solve the classification problem, a method called CLARA (CLAssification of Real Alternatives) has been developed. This method can be used to classify a complete set, or a specified number of objects of the set, with minimal involvement of experts [10,13,18].

1. The data of the problem

The problem may be formally represented in the following way:

1. G is the property satisfying the target criterion of the problem.
2. $K = \{K_1, K_2, ..., K_Q\}$ is a set of evaluating criteria of an object.
3. $S_q = \left\{k_1^q, ..., k_{w_q}^q\right\}$ for q=1,...,Q is a set of estimates based on the criterion K_q, w_q is the number of graduation marks on the scale of the criterion K_q; the scales are arranged in the order of distinctness of the property G.
4. $Y = S_1 \times ... \times S_Q$ is the state space of the objects to be classified. Every object is described by a number of estimates based on the criteria $K_1, ..., K_Q$. In this way, a set of alternatives $\{y_1, y_2, ..., y_L\}$ is defined, where $L = |Y| = \prod_{q=1}^{Q} w_q$ is the cardinality of a set Y, (the number of alternatives).

5. $C = \{C_1, C_2, ..., C_M\}$ is a set of classes to be obtained by breaking down the set Y^a, which should be arranged in the ascending order of distinctness of the property G (in the class C_{n+1} this property is more distinct, while in the class C_n it is less distinct).
6. $Y^a \subseteq Y$ is a set of *admissible* real objects.

Since the estimates based on each criterion are ordered, then the scale showing the order of classes S_q can be compared with the numerical scale $B^q = \{1, 2, ..., w_q\}$, where $b_i^q < b_j^q$, if b_i^q is less preferable for a decision maker (DM) than b_j^q.

The information of the DM preferences determines the relationships of rigorous preference (or dominance) P^0 in the set Y:

$$P^0 = \{(y_i, y_j) \in Y \times Y \mid \forall q \in K \; b_i^q \geq b_j^q \; \wedge \; \exists q^0 : b_i^{q^0} > b_j^{q^0}\}$$

implying that the alternative $x \in Y$ is dominant over the alternative $y \in Y$.

On the other hand, it is known that the classes of solutions are ordered for the DM. It means that any alternative from the class *n+1* is more preferable for the DM than any alternative from the class *n*. This is shown by the following binary preference relationship in the set Y:

$$P^1 = \{(y^i, y^j) \in Y \times Y / y^i \in Y^k, y^j \in Y^l, \quad k > l\}$$

It can be assumed that none of the vector estimates in the set Y, dominating over the given one, should be assigned to a less preferable class. This statement is known as the “hypothesis of distinctness”. It can be formally expressed as follows:

$$(y_i, y_j) \in P^0 \Rightarrow (y_j, y_i) \notin P^1 \tag{1}$$

Definition. Partition of a set of vector estimates Y into the M ordered classes is consistent if the condition (1) is *satisfied* for any $y_i, y_j \in Y$.

Based on the preferences of the decision maker, a consistent representation of $F: Y^a \rightarrow \{Y_l\}, l = 1, 2, ..., M$, has to be constructed, such that:

$$Y^a = \bigcup_{l=1}^{M} Y_l ; \; Y_l \bigcap Y_k = 0$$

where $k \neq l$, Y_l is a set of the vector estimates from Y, assigned to the class C_l.

2. The analysis of verbal decision methods for classification of alternatives

Many widely known methods for solving multi-criteria classification problems are presented in Table 1 [2, 6, 7, 8, 9]. In this chapter some most frequently used verbal ordinal classification methods are considered. All these methods belong to the Verbal Decision Analysis group and have the following common features:

1. The attribute scale is based on verbal description unchanged in the process of solution, when verbal evaluation is not converted into the numerical form or score.
2. An interactive classification procedure is performed in steps, where the DM is offered an object of analysis (a course of treatment, for example). The object is presented as a small set of rankings. The DM is familiar with this type of description, therefore he/she can make the classification based on his/her expertise and intuition.
3. When the DM has decided to assign an object to a particular class, the decisions are ranked on the dominance basis. This provides the information about other classes of objects related to it by the relationship of dominance. Thus, an indirect classification of all the objects can be made based on a single decision of the DM.
4. A set of objects dominating over an object considered is referred to as a domination cone. A great number of objects have been classified many times. This ensures error-free classification. If the DM makes an error, violating this principle, he/she is shown the conflicting decision on the screen and is prompted to adjust it.
5. In general, a comprehensive classification may be obtained for various numbers of the DM decisions and phases in an interactive operation. The efficiency of multi-criteria classification technique is determined on the basis of the number of questions for the DM needed to make the classification. This approach is justified because it takes into consideration the cost of the DM's time and the need for minimizing classification expenses [2, 9, 11].

Let us consider several most commonly used methods in more detail.

ORCLASS [4, 6]. This method (Ordinal CLASSification) allows us to build a consistent classification, to verify the information and to obtain general decision rules. The method relies on the notion of the most informative alternative, allowing a great number of other alternatives to be implicitly assigned to various classes. ORCLASS takes into account the possibilities and limitations of the human information processing system.

Method assessment: The main disadvantage of the method is its low effectiveness due to the great number of questions to DM needed for building a comprehensive classification.

Table 1

Verbal analysis methods

The type of method	The purport of method	Notes
ORCLASS	This method is used for classifying different types of loans [6,7,8]	By deficiency algorithm appears its the large number of questions for DM to do the classifications [6]
DIFCLASS	This method was the first to use dynamic construction of chains covering alternative space for selecting questions to DM (decision maker) [7 9]	The area of DIFCLASS application is restricted to tasks with binary criteria scales and two decision classes [7]
CYCLE	The CYCLE algorithm makes it possible to effectively build the complete non--contradictory bases of expert knowledge for the subject areas by complete order of the scales of criteria [8, 9]	The methods can be successfully applied to classify investment projects when the decision classes and the criteria used are thoroughly revised [9]
CLANSH	The CLANSH method makes it possible to build the wheel bases of expert know-ledge, when assumption about the presence of linear order of many estimations with respect to each of the criteria is substituted by assumption about the presence of the incoherent transitive binary relation [8,13]	
STEPCLASS	System realizes technological approach to the structurization subject area and to the development of the decisive rules of expert and guarantees completeness and consistency [9]	

DIFCLASS [4]. This method was the first to use dynamic construction of chains covering *Y* space for selecting questions to DM. However, the area of DIFCLASS application is restricted to tasks with binary criteria scales and two decision classes.

CYCLE [6]. CYCLE (Chain Interactive Classification) algorithm overcomes the restrictions of DIFCLASS, generalizing the idea of dynamic chain construction to the area of ordinal classification task with arbitrary criteria scales and any number of decision classes. The chain here means an ordered sequence of vectors $\langle x_1, \ ..., x_d \rangle$, where $(x_{i+1}, x_i) \in P$ and vectors x_{i+1} and x_i differ in one of the components.

Method assessment: As comparisons demonstrate, the idea of dynamic chain construction allows us to get an algorithm close to optimal by a minimum number of questions to DM necessary to build a complete classification. The application of ordinal classification demonstrates that problem formalization as well as introduction of classes and criteria structuring allow to solve classification problems by highly effective methods.

The method can be successfully applied to classification of investment projects when the decision classes and the criteria used are thoroughly revised.

3. A method of constructing a comprehensive order classification

At the first stage, the alternatives of the set Y are numbered in the specified order. In this case, $y_i > y_j \Rightarrow i < j$. This preliminary numbering ensures that a particular alternative is considered when all the alternatives dominant over it had been already analysed.

The use of the hypothesis of distinctness (1) allows us to considerably reduce the number of questions to an expert, required to make the classification.

Let us denote by G^i a set of class numbers $Y_l (1 \le l \le M)$, admissible for the vector estimate $y_i \in Y$. Before questioning the DM (an expert), $G^i = \{1,2, \ldots, \mathrm{M}\}$ is assumed for $\forall y_i \in Y$, because we do not have any information about the expert's preferences. Finally, it is required that all G^i consist of only one element.

Suppose that the expert decided that the vector estimate $y_i \in Y$ should belong to the class $Y_l (1 \le l \le M)$ in accordance with its global quality. Following the hypothesis of distinctness, in this case a vector estimate – described by a number of the criteria values, which are not less preferable for an expert – cannot belong to a less preferable class.

Similarly, a vector estimate, described by a number of the criteria values which are not more preferable that those of y_i, cannot belong to a more preferable class.

Consequently, the data, related only to one vector estimate of *Y,* which were elicited from an expert, can result in the reduction of the sets G^i, corresponding to other vector estimates. In this way, in a particular case, vector estimates can be assigned to a particular class of vector estimates without being submitted to an expert.

It is necessary to take into consideration the possibility of assigning a particular vector to a particular class. The indicator p_{il} (assessing the possibility of assigning the vector y_i to the class Y_i) shows the proximity of the vector considered to the members of this class because the vectors of the same class usually form compact groups in multidimensional space. To calculate p_{il}, the normalized distance between the vector y_i and the center of the class C_k can be used.

Relying on two indicators, p_{il} and G^i, a unified quantitative estimate of the informativity of any not estimated state Φ can be obtained:

$$\Phi_i = f(\{p_{il}, g_{il} \mid l \in G^i\}) \tag{3}$$

where *f* is a certain real function, g_{il} is the number of vectors from Y whose membership in a particular class becomes known (i.e. the cardinal number of the corresponding set of the class numbers G^i is equal to one) if the expert assigns the vector y_{il} to the class Y_l

This concept underlies a multistage procedure of carrying on a dialogue which can be generally described in the following way. A subset of the alternatives Y_g for which the set G^i of the admissible classes contains more than one element is determined. If Y_g is empty, go to stage 7.

1. The indicator p_{il} is calculated for all the alternatives from Y_g and g_{il} is determined for $\forall l \in G^i$.
2. The indicators p_{il} are found from the formula.
3. Based on the above indicators, the amount of information of the vector $y_i - \Phi_i$ is determined.
4. $y_i \in Y_g : \Phi_i = \max\limits_{y_j \in Y_g} \Phi_j$ is determined.

5. The above vector is submitted to an expert to be assigned to one of the classes.
6. The sets G^i are modified in accordance with the class of the vector as specified by the expert. Go to stage 1.
7. The procedure is completed.

In the ORCLASS method, mathematical expectation of the number of classified vectors is used as a function of informativity (2) for developing a comprehensive classification:

$$\Phi_i^{abcde} = \sum_{i=G^l} p_{il} g_{il} \tag{5}$$

To classify a specified subset, the number of actual alternatives, whose classes become known when a particular choice is made by an expert, should be maximized. This implies that, in calculating the indicators, g_{il}, only the alternatives belonging to Y^a should be taken into account. Thus, to achieve the specified aim, the way of determining the informativity should be changed. Similarly to the formula (3), in the following expression:

$$\Phi_i = \sum_{i \in G^i} p_{il} g_{il}^a \tag{7}$$

g_{il}^a is the number of vectors from Y^a, whose membership of a particular class becomes known when an expert refers the vector y_i to the class Y_l. The coefficients g_{il} in the informativity formula are considered a random quantity Γ_i with the probabilities of realizing the l-th value of p_{il}. Then, $\Phi_i = M\Gamma_i$, where $M\Gamma_i$, is mathematical expectation of the random quantity Γ_i. The spread of the random quantity Γ_i about its mean value $M\Gamma_i$ is the mean square deviation $\sigma_i = \sqrt{D\Gamma_i} = \sqrt{M(\Gamma_i - M\Gamma_i)^2} = \sqrt{M\Gamma_i^2 - (M\Gamma_i)^2}$, where $D\Gamma_i$ is the variance Γ_i However, relative rather than absolute deviation is important for this analysis. In fact, large deviations may be allowed for large values of Φ_i. Therefore, the following function can be used to express the informativity:

$$\tilde{\Phi}_i = \frac{\Phi_i}{1+n\dfrac{\sigma}{\Phi_i}} = \frac{\Phi_i}{1+n\dfrac{\sqrt{\sum_{l \in G_i} p_{il}(g_{il}^a)^2 - \Phi_i^2}}{\Phi_i}}, \quad n \geq 0 \tag{5}$$

The following notation is used in this formula:

Φ_i is the informativity in the sense of mathematical expectation g_{il}^{a} (4),

σ / Φ_i indicates relative deviation of the indicators g_{il}^{a} from their mean value,

A unity is added to the denominator for it to be not less than one (i.e. to be more than zero in all cases), n is an empirical multiplier in the case of relative deviation referred to as *the significance level of variance*. This multiplier allows us to specify the effect of deviation on informativity. It is clear that, when n=0, $\widetilde{\Phi}_i = \Phi_i$, i.e. informativity is determined without taking into account the variance. This helps to avoid some "risky" situations, when an alternative is offered to the expert for evaluation with the varying numbers of indirectly classified alternatives (depending on the expert's decision).

4. CLARA. Classification algorithms

The CLARA algorithm (Classification of Real Alternatives) is based on the dichotomy of the chains of alternatives, beginning with the longest chain. This concept, first used in the DIFCLASS algorithm [6] and then in CLANSH [18], has been adapted for rarefied spaces *Y*. Moreover, the CLARA algorithm uses a new idea of the adaptive dichotomy allowing us to determine the boundaries between classes of solutions and perform classifications much faster.

A general block-diagram of the CLARA algorithm is presented in Figure 1.

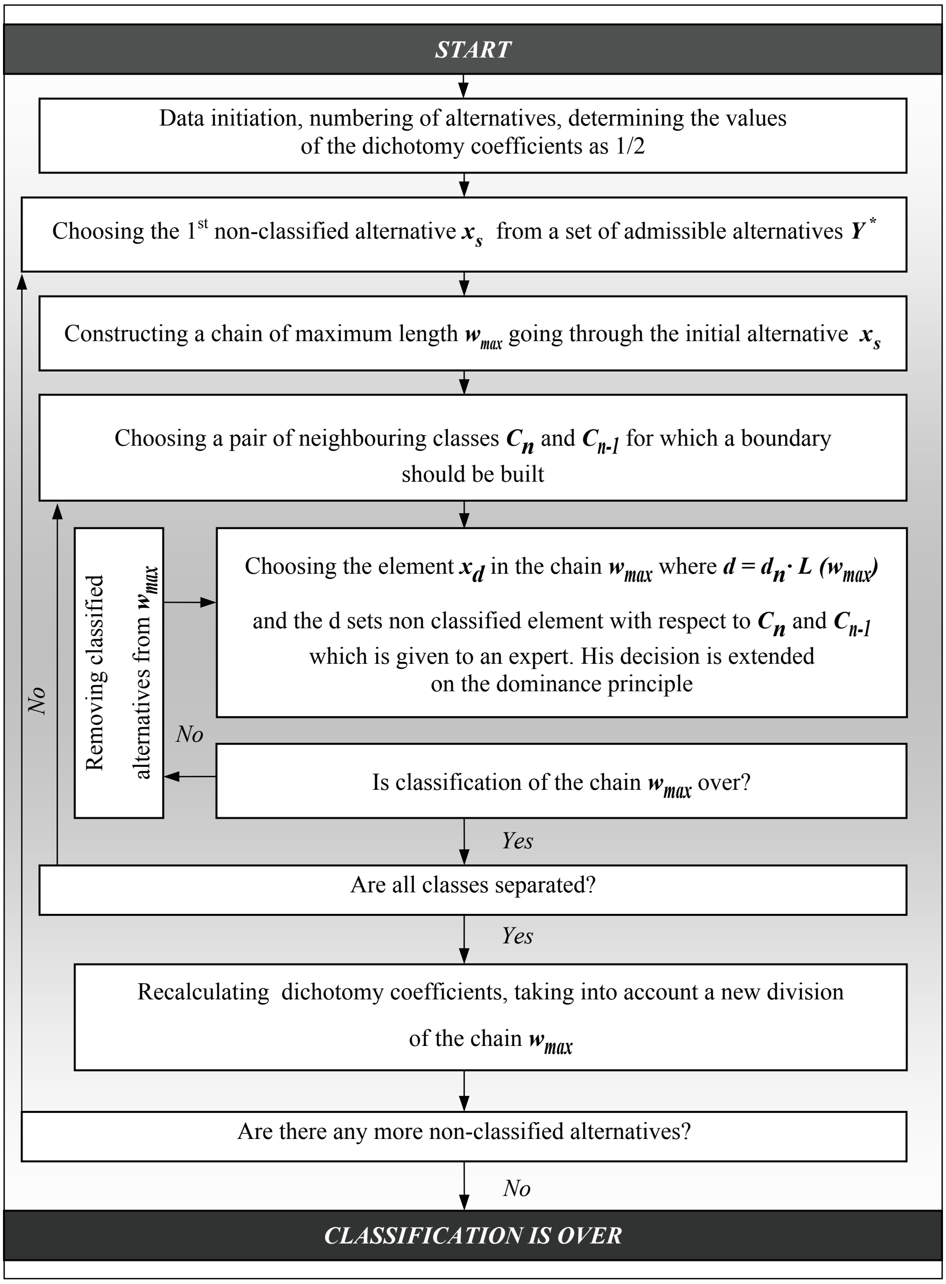

Figure 1. A general block-diagram of the CLARA algorithm

4.1. The main stages of gaining expert knowledge by using the CLARA method

The knowledge is gained by carrying on a dialogue with an expert. First, the main operations are outlined:

1. Discussing the statement of the problem. Defining the properties of G.
2. Generating a set of criteria K by an expert.
3. Constructing the scales for criteria evaluation. Preliminary analysis: checking if the estimates are (partially) arranged in the descending order of the distinctness of the property G.
4. Defining a set of ordered classes of solutions C by an expert [12].

The second stage of applying the method – expert-made classification – involves submitting to an expert the possible combinations of the attribute values for analysis. This is a time-consuming procedure because the number of combinations is usually large. This may entail expert's errors. Therefore, the method allows to define some simple problems within the original classification problem by considering only two values of any attribute. Then, the results obtained are included in the original problem, and the expert solves this partially solved problem on the full scale [13].

In the process of classification it may become clear that some combinations of the criteria values cannot be practically realized. In this case, the objects to which they refer are excluded from the analysis by an expert.

The classification is over, when all the objects included in the analysis (the set Y*) are assigned to a particular classes.

At the third stage of analysis the boundaries of classes are verified again because the mistakes the expert could made during previous stages. Since class boundaries are the key factors in making classifications, every class specified by the expert should be verified. For this purpose, every boundary element is offered to the expert again for checking. At the fourth stage, the boundaries of the classes are converted into the expert rules of solution of the form:

$$ab{*}{*}{*} + p_n^{k_1[x_1]}, \text{ except } \{abcde, ..., abpqr\} \quad (6)$$

So that every alternative follow one rule, where $ab{*}{*}{*}$ is *a fixed part* of the rule, while $p_n^{k_1[x_1]}$ is the *rearrangeable part* of the rule. Here, n is equal to the number of asterisks, k_i is the number of estimates x_i involved in the rearrangement. The third part is activated if a set of alternatives described by a template is not completely rearrangeable, and to achieve this, the number of elements should be small. Then, the missing elements are simply listed.

The rules described are introduced into the system when solving the problem on a large scale. They simplify the solution considerably by reducing the classification space.

The decision rules of a particular class can be represented as a two-level tree (Figure 2):

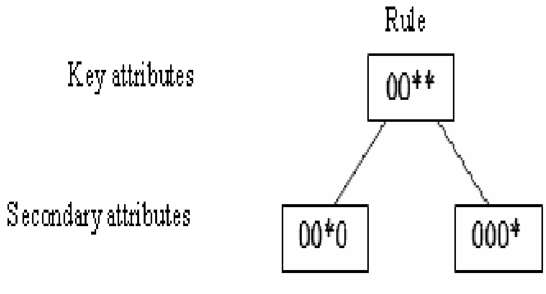

Figure 2. Decision rules of a particular class

Here the values of key attributes are found at the higher level, while the combinations of the values of secondary attributes are found at the lower level [12].

The rules described comply with inexplicit expert knowledge. The rules are submitted to an expert for approval. Some rules may be too complicated. In this case, the procedure of identifying the zone of superficial knowledge might be needed because complicated rules often indicate that knowledge is not stable [12]. For this purpose, it is necessary to go back to the second stage of method application (Figure 3).

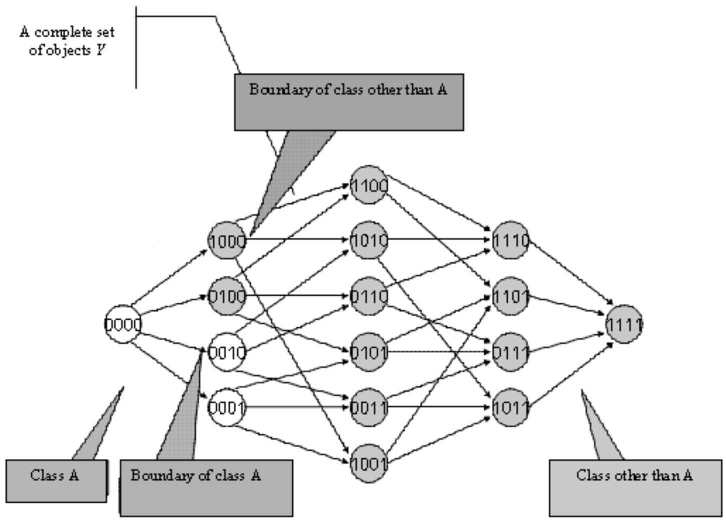

Figure 3. Example of method application

Conclusion

In project development, it is hardly possible to get exhaustive and accurate information. As a result, the situations occur, the consequences of which can be very damaging to the project.

Very often investment decision-making and research planning are referred to as non-structured problems. Since the essential characteristics of such problems are qualitative, they can hardly be used in the analysis. On the other hand, the quantitative models are not sufficiently reliable.

Non-structured problems have the following common characteristics. They are unique decision-making problems, i.e. every time the decision-maker is faced with an unknown problem or with one having new features compared to the previously considered case. These problems are associated with the uncertainty of the alternatives to be evaluated, caused by the lack of information for making a decision. The evaluation of the alternatives is of qualitative nature, being usually expressed verbally (in statements). Very often, experts cannot measure qualitative variables against an absolute scale where the level of quality does not depend on the alternatives. When the uncertainty is high, experts can only compare the alternatives qualitatively, based on particular criteria.

The CLARA algorithm (Classification of Real Alternatives) is based on the dichotomy of the alternatives chains, beginning with the longest chain. This concept was first used in the DIFCLASS algorithm and then in CLANSH.

Investment risk in construction can be evaluated efficiently enough using the CLARA method. This method allows to classify all possible construction investment projects presented by evaluations on the predefined criteria into several accurately defined classes reflecting the project risk level. CLARA method contains an algorithm to achieve the minimal amount of the DM questions. Moreover, the CLARA algorithm uses a new idea of the adaptive dichotomy allowing us to determine the boundaries between classes of solutions and to make classifications much faster.

References

[1] Ustinovichius L.; Zavadskas E.K.: *Assessment of Investment Profitability in Construction from Technological Perspectives.* VGTU, Publishing House "Technika", Vilnius 2004, pp. 134-150 (in Lithuanian).

[2] Larichev O., Mechitov A., Moshovich E., Furems E.: *Revealing of Expert Knowledge*. Publishing House "Nauka", Moscow 1989 (in Russian).

[3] Nedzveckas J., Rasimavichius G.: *Investment Currency Risk and Ways of Its Decrease.* "Economy" 2000, Vol. 51, 2, pp. 63-74.

[4] Ashihmin I.V., Furems E.M.: *UNICOMBOS – Intellectual Decision Support System For Multicriteria Comparison And Choice*. Artificial Intelligence (Искуственный интелект) Proceedings of the Ukrainian Academy of Sciences, Kiev 2004, 2, pp. 243-247 (in Russian).

[5] Zavadskas E.K., Ustinovichius L., Stasiulionis A.: *Multicriteria Valuation of Commercial Construction Projects for Investment Purposes*. "Journal of Civil Engineering and Management" 2004, Vol. 10, 2, pp. 151-166.

[6] Larichev O., Bolotov A.: *System DIFKLASS: Construction of Full and Consistent Bases of Expert Knowledge in Problems of Differential Classification. The Scientific and Technical Information*. "Informacionie Procesi i Sistemi" (Information processes and systems) 1996, (9), a series 2, pp. 276-285 (in Russian).

[7] Asanov A., Borisenkov P., Larichev O., Nariznij E., Rozejnzon G.: *Method of Multicriteria Classification CYCLE and Its Application for the Analysis of the Credit Risk*. "Economy and Mathematical Methods" (Ekonomika i Matematicheskije metodi) 2001, Vol. 37(2), pp. 14-21 (in Russian).

[8] Korhonen P., Larichev O., Moshkovich H., Mechitov A., Wallenius J.: *Choice Behavior in a Computer-Aided Multiattribute Decision Task.* "Journal Multicriteria Decision Analysis" 1997, Vol. 2, No. 2(2), pp. 43-55.

[9] Larichev O., Kochin D., Kortnev A.: *Decision Support System for Classification of a Finite Set of Multicriteria Alternatives.* "Decision Support Systems" 2002, Vol. 33, No. 1, pp. 13-21.

[10] Ustinovichius L., Ševčenko G., Kochin D.: *Classification of Real Alternatives and its Application to the Investment Risk in Construction.* In: *Multiple Criteria Decision Making '05*. T. Trzaskalik (ed.). Karol Adamiecki University of Economics, Katowice 2006, pp. 299-318.

[11] Naryzhny Y.V.: *The Development of Intellectual Training Systems Based on Expert Knowledge*. D.Th. ISA RAN, Moscow 1998 (in Russian).

[12] Kochin D.Yu.: *Building Expert Knowledge Bases to Be Used for Intellectual Training Systems.* D.Th. ISA RAN, Moscow 2006.

[13] Ustinovichius L., Kochin D.: *Verbal Analysis of the Investment Risk in Construction*. "Journal of Business Economics and Management" 2003, Vol. 4(4), pp. 228-234.

[14] *Multiple Criteria Decision Analysis: State of the Art Surveys.* J. Figueira, S. Greco, M. Ehrgott (eds). Springer, Boston 2005.

[15] Zavadskas E.K., Ustinovichius L., Turskis Z., Peldschus F., Messing D.: *LEVI 3.0 – Multiple Criteria Evaluation Program for Construction Solutions*. "Journal of Civil Engineering and Management" 2002, Vol. 8, 3, pp. 184-191.

[16] Zavadskas E.K., Kaklauskas A., Raslanas S., Malienė V.: *The Application of Multi-Criterion Methods for Valuation Recreation Property*. "STATYBA (Civil Engineering)" 2001, Vol. 7, 4, pp. 327-333.

[17] Ustinovichius L.: *Determination of Efficiency of Investments in Construction*. "International Journal of Strategic Property Management" 2004, Vol. 8, 1, pp. 25-44.

[18] Kutut V., Ustinovichius L., Kochin D., Turskis Z.: *Verbal Analysis as an Effective Tool of the Investment Strategy Management*. The Fourth International Conference on Cooperative Design, Visualization and Engineering, CDVE2007, Shanghai, China September, 16-20, 2007.