Linguistic approach to the classification problem based on the multiset theory

L A Demidova* and Ju S Sokolova

1
MIREA – Russian Technological University, Vernadsky pr., 78, Moscow, 119454, Russia
2Ryazan State Radio Engineering University, Gagarin Str., 59/1, Ryazan, 390005, Russia

*E-mail: demidova.liliya@gmail.com

Abstract. The problem of developing generalizing decision rules for object classification, which arises under conditions of inaccurate knowledge about the values of objects’ attributes, and about the significance of the attributes themselves, has been considered. The approach to the binary classification of objects, which implements the representation of inaccurate knowledge based on linguistic variables and allows one to consider various strategies for the formation of generalizing decision rules for classification using the tools of multiset theory, has been proposed. The example of the formation of generalizing decision rules of binary classification for the set of competitive projects evaluated by the group of experts, confirming the effectiveness of the proposed approach, has been considered. A herewith, visualization of objects, the values of the features of which are the frequency of setting a certain score according to the a priori given rating scale by all experts, in a two-dimensional space using the non-linear dimensionality reduction algorithm named as the UMAP algorithm, has been implemented. Based on the results of visualization and cluster analysis of the initial set of competitive projects, the “noise” project, which negatively affects the results of the formation of generalizing decision rules of binary classification, was identified and removed from further analysis.

1. Introduction

Nowadays, data mining algorithms which allow building intelligent classifiers and regression models [1–7], performing discriminant analysis [8], searching for association rules [9], etc., are actively used to solve many applied problems. Machine learning algorithms such as SVM (Support Vector Machine) algorithm [1–3], RF (Random Forest) algorithm [6] and algorithms that involve artificial neural networks [7] can be implemented to solve data classification problems.

For many decomposition algorithms, it is important that the values of the characteristics of objects are represented by numerical values, i.e. converted to a scale of intervals or ratios. However, sometimes you can only define the intervals of membership for the values of the characteristics of objects. Such situations, for example, are possible when the values of the characteristics of objects are determined on the basis of readings from several measuring devices (groups of devices), according to the results of a group expert assessment or from several sources of information.

So, it is quite natural that when performing measurements, one should take into account the variation of the readings of the measuring devices, as well as the fact that quite often even highly qualified
specialists (experts) are able to determine only the intervals of belonging of the assessments of objects according to the characteristics of the assessment, score scale.

Despite the presence of various approaches to solving the problem of classification, ordering and selection of various objects based on the data of group measurements or group expert assessments [10–13], there is a need to develop a mathematical apparatus that makes it possible to take adequate and justified classification decisions using subjective qualitative data, presented, including in the form of interval estimates [14, 15].

One of the promising approaches to solving the classification problem, allowing to take into account all, including contradictory, data of group measurements, is based on the use of tools from the theory of multisets [10–13]. However, this approach presupposes the use of a classical assessment scale, on the basis of which unambiguous clear values of the characteristics of objects are set, which are used in the future to form the generalizing classification decision rules (GCDRs) of objects. In the case when it is supposed to work only with the intervals of belonging of the characteristic values, it is proposed to use a linguistic scale to represent the characteristic values. Using this scale, for each characteristic, a low, high, and average value can be determined, which can be called, for example, pessimistic, optimistic and neutral values, if we assume that the low value corresponds to the worst possible measurement, and the upper value corresponds to the best possible measurement. At the same time, it will be possible to talk about the formation of GCDRs using various classification strategies – pessimistic, neutral and optimistic.

In the case of a group expert assessment, it will be possible to speak, for example, about a conservative, neutral and risky assessment of an object for each assessment characteristic and, as a consequence, about the formation of GCDRs using conservative, neutral and risky classification strategies.

2. Representation of objects by the collection of multisets

Let \( Z = \{z_1, ..., z_s\} \) be the set of objects; \( P = \{P_1, ..., P_q\} \) be the set of characteristics, and the number of measurements for each characteristic is equal to \( m \).

The classical approach to the problem of developing GCDRs based on readings from several measuring devices (a group of devices) assumes that the values of the characteristics of objects are represented by clear integers or real numbers, and the worse the object is in terms of a certain characteristic, the less clear the value for this characteristic (if necessary, you can to ensure that this requirement is met for individual characteristics of objects by simple transformations).

The classical approach to the task of developing GCDRs based on group expert assessment assumes that the expert assessment of objects according to the assessment characteristics is carried out on a clear score scale (for example, on a five-score scale), while the lower the object is assessed according to some characteristic, the lower the clear numerical assessment (score). It is often difficult for an expert to give clear numerical estimates of objects according to the characteristics of the assessment – instead of it he can indicate some intervals to which these assessments belong.

To improve the quality of classification decisions in the presence of inaccurate, and often contradictory, data of group measurements of values of characteristics of objects and the uncertainty of information about the significance of the characteristics themselves, it is proposed to abandon the use of the traditional clear assessment scale and use a linguistic scale that allows to implement the principles of describing and processing inaccurate data based on linguistic variables.

![Figure 1. Assessment strategies (measurements).](image-url)
In the case of applying the linguistic approach to the formation of GCDRs, each object for each characteristic will not be compared with a clear numerical value, but a certain interval of the form $[\alpha, \beta]$ [14, 15]. A herewith, the left border of the interval $[\alpha, \beta]$ can be compared to a purely pessimistic (purely conservative) assessment strategy (and, accordingly, a pessimistic (conservative) strategy for making a classification decision), the right one – to a purely optimistic (purely risky), and the middle of this interval – to a neutral strategy (figure 1). Since certain different clear numerical values of characteristics belonging to intervals of the form $[\alpha, \beta]$ will be used for calculations in the formation of GCDRs, we can talk about the presence of one neutral assessment strategy (measurement) and a certain set of pessimistic (conservative) and optimistic (risky) assessment strategies (measurement) and, accordingly, similar strategies for making classification decisions (depending on which boundaries of interval assessments of the form $[\alpha, \beta]$ are closer to the clear numerical values of object characteristics used for calculations).

Let $G = \{g_\gamma \mid \gamma = -L, -L+1, \ldots, L-1, L\}$ be some discrete linguistic scale, where $g_\gamma$ is a linguistic variable; $L$ is some natural number ($L \in N$) [14, 15]. For example, if $L = 3$, the linguistic scale for the values of characteristics $G = (g_{-3}, g_{-2}, g_{-1}, g_0, g_1, g_2, g_3)$ can be defined as follows: $G = (g_{-3}, g_{-2}, g_{-1}, g_0, g_1, g_2, g_3)$ = (“extremely small”, “very small”, “small”, “medium”, “large”, “very large”, “extremely large”), where each linguistic term (for example, “extremely small”, “very small”, etc.) corresponds to a classical clear meaning (for example, $-3$, $-2$, etc.), which in case of presentation of inaccurate data of group measurements is one of the boundaries (left or right) of the estimation (measurement) interval. Moreover, if the data (and, therefore, the values of the characteristics) is accurate, then the left border will coincide with the right one.

In order to avoid the loss of linguistic information about a particular decision made, the discrete linguistic scale $G = (g_{-L}, g_{-L+1}, \ldots, g_0, g_{L-1}, g_L)$ can be extended to a continuous linguistic scale $\tilde{G} = \{g_\gamma \mid \gamma \in [-r, r]\}$, where $r$ is a sufficiently large positive number ($r \in R$). Then, if $g_\gamma \in G$, then $g_\gamma$ is the initial linguistic term, otherwise $g_\gamma$ is the extended (virtual) linguistic term [14]. In this case, the original linguistic terms are used both to represent the values of the characteristics of the objects themselves and the significance of these characteristics, and the extended (virtual) linguistic terms are used to implement calculations in order to classify objects [14].

The use of the linguistic approach to classification decision-making based on multisets allows to consider various strategies for the formation of GCDRs.

Regardless of which approach (classical or linguistic) is used to represent the values of the characteristics of objects, when developing decision rules for classification based on multisets, clear (crisp) numerical values of the characteristics of objects are used. Below are the main principles of the formation of GCDRs. Since each object is assessed by several experts, there are several different options for representing the values of the object's characteristics, while the values of the characteristics obtained from different measurements can be both similar and contradictory. The inconsistency of the individual values of the characteristics of objects can be caused by a variation in the readings of the measuring devices, and in the case of a group expert assessment, by the ambiguity of the experts' understanding of the problem being solved, errors and inaccuracies in the assessment of objects, the specificity of expert knowledge, which can also lead to inconsistency of individual assessments of the significance of the characteristics themselves. When deciding on the classification of objects, one should take into account all, even non-coinciding (contradictory) values of the characteristics of objects.
A collection of objects $z_i$ can be represented as a collection of multisets $Z_i$ [10, 11]. Let there be $u_j$ different values $p^{(j)}_i$ ($l_j = 1, u_j$) for each $j$-th characteristic of the object $z_i$, and the number of measurements leading to the value $p^{(j)}_i$ is $k_{Z_i}(p^{(j)}_i) = \sum_{l_j=1}^{u_j} k_{Z_j}(p^{(j)}_i) = m$; $i = 1, s$; $j = 1, q$.

Let us consider, based on the results of individual measurements, that object $z_i$ is preliminarily assigned to one of two classes $Y_c$ ($c = 1, 2$) based on the individual sorting rule $W = \{w_1, w_2\}$, which is another qualitative characteristic. Let the number of dimensions which have assigned the value $w_c$ to the object $z_i$ is $k_{Z_i}(w_c) = \sum_{c=1}^{2} k_{Z_i}(w_c) = m$; $i = 1, s$). Taking into account the introduced designations, we can say that:

- there are $m$ instances of each object $z_i$ which differ in sets of characteristic values $P = \{P_1, ..., P_q\}$;
- there are $m$ mismatched individual pre-sorts of a collection of objects $Z = \{z_1, ..., z_s\}$.

In this case, the set of characteristics can be formed: $U = \{P_1, ..., P_q, W\}$.

Let the values for each characteristic be ordered from worst to best: $p^{(1)}_j < p^{(2)}_j < ... < p^{(u_j)}_j$ (if necessary, it is possible to ensure that this requirement is met for individual characteristics of objects by means of simple transformations); $w_1 < w_2$, and data on the characteristics of classes and characteristics (importance, preference, etc.) are absent.

Each object $z_i$ can be assigned a multiset of the form [10, 11]:

$$Z_i = \{k_{Z_i}(p^{(1)}_1) \cdot p^{(1)}_1, ..., k_{Z_i}(p^{(u_1)}_1) \cdot p^{(u_1)}_1, ..., k_{Z_i}(p^{(1)}_q) \cdot p^{(1)}_q, ..., k_{Z_j}(p^{(u_j)}_q) \cdot p^{(u_j)}_q, k_{Z_i}(w_1) \cdot w_1, k_{Z_i}(w_2) \cdot w_2\},$$

(1)

where $k_{Z_i}(p^{(j)}_j)$ and $k_{Z_i}(w_c)$ are the numbers of measurements which have assigned the values $p^{(j)}_j$ and $w_c$ to the object $z_i$, respectively; the symbol “$\cdot$” is used to denote the relationship between $k_{Z_i}(p^{(j)}_j)$ and $p^{(j)}_j$, as well as between $k_{Z_i}(w_c)$ and $w_c$ ($i = 1, s$; $j = 1, q$; $c = 1, 2$; $l_j = 1, u_j$).

Representation of the object $z_i$ in the form (1) can be realized using, for example, rules of the form [10]:

$$\text{IF } <\text{conditions}> \text{, THEN } <\text{solution}>.$$  

(2)

The term <conditions> corresponds to various combinations of values for characteristics $p^{(j)}_j$ according to the object $z_i$. The term <solution> includes the set of individual conclusions on the preliminary sorting of objects $z_i$ and the certain rule that allows the object $z_i$ to be assigned to the certain class $Y_c$. As such the rule, the majority rule can be used, according to which the object $z_i$ belongs to the class $Y_c$ if $k_{Z_1}(w_{t}) > k_{Z_2}(w_{t})$ for all $t \neq c$ ($c = 1, 2$; $t = 1, 2$).

3. **Classification of objects represented by multisets**

GCDrs of objects should correspond to the greatest extent to all individual values of the characteristics of objects and ensure the decomposition of the set $Z = \{z_1, ..., z_s\}$ into two classes $Y_1$ and $Y_2$ in the best way (in the sense of closeness to preliminary individual sorts).
The formation of each class $Y_c \ (c = 1, 2)$ can be implemented by adding the corresponding multisets. In this case, all values of the characteristics of all objects included in the class $Y_c \ (c = 1, 2)$ must be taken into account. Elements $k_{Y_c}(p_j^1)$ and $k_{Y_c}(w_c)$ \ (j = 1, q; I_j = \{u_j; c = 1, 2\}) in multiset $\hat{Y}_c \ (c = 1, 2)$:

$$\hat{Y}_c = \{k_{Y_c}(p_j^1) \cdot p_j^1, ..., k_{Y_c}(w_c) \cdot w_c\}.$$  \hspace{1cm} (3)

where

can be calculated as the sums of the corresponding elements $k_{Z_i}(p_j^1)$ and $k_{Z_i}(w_c)$ for objects $z_i$ which fall into the class $Y_c \ (c = 1, 2)$ [10].

If we represent the multiset $\hat{Y}_c$ of the class $Y_c$ as $\hat{Y}_c = \sum_{j=1}^{q} P_{jc} + W_c \ (c = 1, 2; j = 1, q)$, where $P_{jc}$ and $W_c$ are the multisets, the elements of which are, respectively, the sums of the values of the $j$-th characteristic of objects $z_i$ which fall into the class $Y_c \ (c = 1, 2)$, and the sum of the values of belonging of objects $z_i$ which fall into the class $Y_c \ (c = 1, 2)$, then for the cardinalities of multisets $P_{jc}$, $W_c$ and $\hat{Y}_c \ (j = 1, q; c = 1, 2)$, the relations \ $|P_{j1}| + |P_{j2}| = s \ (j = 1, q); \ |W_{1}| + |W_{2}| = s; \ |\hat{Y}_1| + |\hat{Y}_2| = s(q + 1)$ \ where $s$ is the number of objects; $q$ is the number of characteristics, are true [10].

In the metric space of multisets, the Hamming metric can be used to calculate the distance between multisets $A$ and $B$ [10], which is defined by the expression:

$$d = |\text{metric}(AB)| = \sum_{j=1}^{q} \sum_{u_j=1}^{u_j} |k_A(p_j^1) - k_B(p_j^1)|.$$  \hspace{1cm} (4)

Objects $z_i \ (i = 1, s)$ in the decomposition \ $\{W_1, W_2\}$ based on the results of preliminary individual sorts form the best of all possible decompositions of the objects collection $Z = \{z_1, ..., z_s\}$ into 2 classes. The distance $d^* = d(W_1, W_2)$ between multisets $W_1$ and $W_2$ is the maximum possible distance in the space of multisets between objects belonging to different classes. With ideal preliminary individual sorts of objects in which there are no contradictions, the distance $d^*$ can be calculated as $d^* = s \cdot m$.

The problem of searching for GCDRs of objects is reduced to the $q$ problems of optimization by characteristics $P_j \ (j = 1, q)$ [10]:

$$d(P_{j1}, P_{j2}) \rightarrow \max d(P_{j1}, P_{j2}) = d(P_{j1}^*, P_{j2}^*).$$  \hspace{1cm} (5)

The solution to each of the problems (5) is expressed in terms of submultisets $P_{j1}^*$, $P_{j2}^*$, and determines the best binary decomposition \ $\{P_{j1}^*, P_{j2}^*\}$ of the set of objects $Z = \{z_1, ..., z_s\}$ by the group of values corresponding to the characteristic $P_j \ (j = 1, q)$.

Let $P_j^*$ be the boundary value of the characteristic that determines the boundary between the generated terms inside each pair $P_{jk}^{*1}$ and $P_{jk}^{*2}$. To construct GCDRs, it is advisable to use the boundary values $P_j^*$ of the characteristics, which occupy the first places in the ranking. The closer the distance value $d(P_{j1}^*, P_{j2}^*)$ is to the distance value $d^* = d(W_1, W_2)$, the more accurate the approximation of
preliminary individual sorting of objects will be. The quality of the approximation by the characteristic \( P_j ( j = 1, q ) \) can be estimated by the formula of the form [10]:

\[
\rho_j = \frac{d(P_{j1}^*, P_{j2}^*)}{d(W_1, W_2)}.
\]

The approximation index \( \rho_j \) characterizes the importance of the characteristic \( P_j ( j = 1, q ) \) in GCDRs.

As a result, GCDRs of objects are determined, showing how group classification decisions are made, that is, which assessment characteristics are really important (since they are present in the rules), and what are the boundary values of these characteristics that affect the assignment of an object to a particular class. In this case, the maximum possible number of GCDRs objects is equal to the number of characteristics.

An object is considered “correctly classified” if GCDR assigns it to the same class that was a priori determined for this object in the course of preliminary individual sorting. The estimation of the approximation accuracy provided by GCDR is calculated as the ratio of the number of objects “correctly classified” by this rule to the total number of objects. Obviously, if two GCDRs have the same approximation accuracy, then GCDR with the smaller number of characteristics should be chosen as the resultant one. The sought-for resultant GCDR should include the boundary values \( \rho^*_j ( j = 1, q ) \) of the characteristics that have values of the approximation index \( \rho_j \) which exceed the desired threshold level \( \rho_0 \) and provide the necessary accuracy of the approximation.

In the case of using a linguistic approach to evaluating objects, it is advisable to analyze the options for GCDRs, based on different assessment (measurement) strategies, in order to identify possible rearrangements of characteristics in GCDRs, as well as to identify changes in their significance when using different assessment strategies (measurements). If we compare the indicator \( \delta \) \((\delta \geq 0)\), to a certain strategy for evaluating (measuring) objects, then the estimate corresponding to this evaluation strategy can be calculated as [14]:

\[
(\beta + \delta \cdot \alpha)/(\delta + 1).
\]

When \( \delta = 0 \) the assessment (measurement) strategy is purely optimistic (or purely risky in the case of group expert assessment), when \( \delta \rightarrow +\infty \) the assessment strategy is purely pessimistic (or purely conservative in the case of group expert assessment), and when \( \delta = 1 \) it is neutral.

As noted earlier, when performing group measurements, each object for each characteristic using the initial linguistic scale is associated with the certain type interval \([\alpha, \beta]\). When analyzing various assessment strategies (measurements) in the computation process, an extended (virtual) linguistic scale is automatically generated for each strategy.

Based on the results of the analysis of the structure of GCDRs, compared to various assessment (measurement) strategies, it can be recommended for use, for example, those GCDRs (and, accordingly, strategies) that, with the same composition (number) of identified important (influencing decision-making) characteristics, have the maximum large values of the approximation index for these characteristics according to formula (6), and are also characterized by the maximum accuracy of approximation of the set of objects by this GCDR.

4. Ordering of objects represented by multisets

Using the multiset theory toolkit, objects can be ordered by distance from the “anti-ideal” (worst) object or by proximity to the “ideal” (best) object, which, on the basis of all measurements, would be compared, respectively, minimum (lowest) and maximum (highest) values for all characteristics. In this case, the “anti-ideal” and “ideal” objects can be associated with multisets:
Thus, the problem of ordering objects \( z_i \) \((i = 1,s)\) can be reduced to the problem of ordering the corresponding multisets \( Z_i \), for example, in increasing order of the distance \( d(Z_{min}, Z_i) \) calculated using the Hamming metric (4). In this case, it will be possible to order the objects from worst to best [13].

Let the object \( z_h \) be worse than the object \( z_g \) \((z_h < z_g)\), if \( d(Z_{min}, Z_h) < d(Z_{min}, Z_g) \). Let the object \( z_h \) and the object \( z_g \) be equivalent and the ordering of the objects is loose if \( d(Z_{min}, Z_h) < d(Z_{min}, Z_g) \).

Distance can be calculated as:

\[
d(Z_{min}, Z_i) = \sum_{j=1}^{q} \chi_j \cdot \sum_{l=1}^{u_j} |k_{Z_{min}}(p_j^l) - k_{Z_i}(p_j^l)| = 2 \cdot \sum_{j=1}^{q} \chi_j \cdot |m - k_{Z_i}(p_j^1)|,
\]

where \( \chi_j \) is the value of the coefficient of relative importance of the \( j \)-th characteristic \((j = 1,q)\); \( \chi_j > 0 \). In the case of using the classical rating scale, it is assumed that characteristics may have different relative importance, but the values related to one and the same characteristic are equivalent. A herewith, additional restrictions may be imposed on the values \( \chi_j \) of the coefficients of relative importance, for example, \( \sum_{j=1}^{q} \chi_j = 1 \). In the case when all the characteristics are equal, the values of all coefficients \( \chi_j \ (j = 1,q) \) are considered equal to 1. When using a linguistic scale of estimation, the values of the coefficients of relative importance \( \chi_j \ (j = 1,q) \) can be determined, for example, in accordance with the conclusions obtained during the formation of the GCDRs.

Taking into account (7), the object \( z_h \) is worse than the object \( z_g \) \((z_h < z_g)\) if \( \sum_{j=1}^{q} \chi_j \cdot k_{Z_h}(p_j^1) > \sum_{j=1}^{q} \chi_j \cdot k_{Z_g}(p_j^1) \).

The problem of ordering objects \( z_i \) \((i = 1,s)\) is reduced to comparing the weighted sums of the first (worst) values of the characteristics of objects \( H_{Z_i}^1 = \sum_{j=1}^{q} \chi_j \cdot k_{Z_i}(p_j^1) \). The worst object is the \( z_j \) for which the sum \( H_{Z_i}^1 \) is the largest. The procedure for calculating and comparing the sums of the second, third, etc. values of the characteristics of objects must be performed before the complete ordering of all objects (and the corresponding multisets) [13]. With this approach to ordering objects in the ordering list, the worst objects will take the first place and the best objects last. Therefore, when ranking objects from best to worst, the lowest rank, equal to 1, should be assigned to the object that took the last place in the ordering list, and the highest rank should be assigned to the object that took the first place in the ordering list (the higher the rank, the worse the object).

When using the linguistic scale of assessment, it is advisable to analyze the options for GCDRs based on different assessment strategies in order to identify possible rearrangements of characteristics, as well as to identify changes in their significance when using different assessment strategies. Also, the choice of one or another assessment strategy can significantly affect the results of ordering objects assigned on the basis of GCDRs to one of the classes.
5. Experimental part

The proposed linguistic approach to making classification decisions based on the multiset theory was applied to the analysis of competitive projects (CP). The problem of binary classification and ordering of CPs based on the results of an independent preliminary individual sorting performed by 7 experts for the group of 16 CPs according to 4 characteristics was considered (s = 16, q = 4, m = 7).

When performing preliminary individual sorting, each expert assessed CP according to 4 characteristics:
- $P_1$ – “social and economic significance”;
- $P_2$ – “competitiveness”;
- $P_3$ – “financial level of the applicant”;
- $P_4$ – “the relevance and novelty”,

setting up interval marks on the linguistic scale at $L = 3$, and, in addition, attributed CP to one of two classes: “To accept CP for implementation” and “To reject CP”. The resulting belonging of CP to the class was determined on the basis of the data of preliminary individual sorting of CP according to the simple majority rule.

**Table 1.** Division of competitive projects into classes “To accept the project for implementation” and “To reject the project”.

| Characteristics assessments | Classes |
|-----------------------------|---------|
| CP $p_1^{-1} p_0^1 p_1^1 p_2^3 p_3^{-1} p_2^{-1} p_0^2 p_2^2 p_3^{-1} p_3^{-1} p_1^1 p_3^{-1} p_3^1 p_2^3 p_3^{-1} p_4^{-1} p_4^0 p_4^1 p_4^2 p_4^3 W_1 W_2$ |
| **1-st class** |
| $z_1$ | 0 0 3 2 2 0 1 2 3 1 0 0 1 2 4 0 0 1 3 3 1 6 |
| $z_3$ | 0 0 1 6 0 0 0 1 2 4 0 0 2 3 2 0 0 1 1 5 0 7 |
| $z_4$ | 0 0 0 0 2 5 0 0 1 3 3 0 0 1 3 3 0 1 0 2 4 1 6 |
| $z_5$ | 0 1 1 3 2 0 1 2 3 1 0 1 0 4 1 0 0 2 1 4 2 5 |
| $z_6$ | 0 0 0 0 4 3 0 0 1 1 5 0 0 0 3 4 0 0 1 2 4 0 7 |
| $z_7$ | 0 0 3 2 2 1 0 1 3 2 1 0 2 3 1 0 1 2 2 2 2 5 |
| $z_{11}$ | 0 1 1 3 2 0 1 0 5 1 0 2 0 3 2 0 0 1 1 4 1 6 |
| $z_{12}$ | 0 1 1 2 3 0 0 2 3 2 0 1 1 0 4 0 0 2 2 3 2 5 |
| $z_{13}$ | 0 1 1 0 5 0 0 1 1 5 0 0 1 1 5 0 0 2 0 0 5 1 6 |
| $z_{14}$ | 0 0 0 0 2 3 3 0 0 3 3 1 0 1 1 3 2 0 0 2 2 3 2 5 |
| $z_{15}$ | 0 0 1 3 3 0 0 0 1 3 3 0 0 1 4 2 0 0 2 2 3 0 7 |
| **2-nd class** |
| $z_2$ | 0 0 4 3 0 1 3 2 1 0 0 0 5 2 0 1 1 3 1 1 5 2 |
| $z_8$ | 0 1 6 0 0 2 0 3 2 0 0 1 5 1 0 0 0 5 1 1 5 2 |
| $z_9$ | 1 0 6 0 0 0 1 1 5 0 0 0 3 4 0 0 0 2 5 0 0 7 0 |
| $z_{10}$ | 0 1 6 0 0 0 0 1 5 1 0 0 2 3 1 1 1 4 0 1 6 1 |

| Classes | Sums by characteristics and solutions |
|---------|--------------------------------------|
| $X_1$ | 0 4 14 29 30 1 4 15 34 23 1 6 11 34 25 0 5 14 18 40 12 65 |
| $X_2$ | 1 2 22 3 0 4 5 15 4 0 0 6 17 4 1 2 4 17 2 3 23 5 |
| $d_1$ | 0.944 0.803 0.859 0.831 |
| $\rho$ | 0.67 0.57 0.61 0.59 0.71 |

Noise

| $z_{16}$ | 0 0 0 6 1 1 6 0 0 0 0 0 6 1 1 6 0 0 0 4 3 |
To improve the quality of GCDRs, the visualization of the original set of objects associated with the competition projects in the two-dimensional space using the nonlinear dimensionality reduction algorithm named as the UMAP algorithm [16] was performed. The values of the attributes of objects in the initial space are the frequency of assigning the certain score according to a priori given rating scale by all experts. [16]. In addition, a cluster analysis of the initial set of competitive projects was carried out using EM algorithm [17]. The results of cluster analysis showed that the optimal number of clusters is 3 (it is with this number of clusters that the maximum value of the cluster silhouette index [18], equal to 0.373) is achieved. At the same time, 11 objects got into the first cluster, 4 objects got into the second cluster, and only one object got into the third cluster.

Figure 2,a shows the results of visualization of the initial set of objects in two-dimensional space with the indication of object numbers and different color fill for different clusters. As can be seen from figure 2, the only object of the third cluster is object number 16 (z₁₆), located at the border separating the other two clusters.

![Figure 2. Data visualization: a – the original dataset; b – the reduced dataset.](image)

It is obvious that the removal of this object from consideration should provide a higher quality of the division of objects into clusters, and, therefore, a higher quality of GCDRs. The results of cluster analysis for a set of 15 objects showed that the optimal number of clusters is 2 (it is with this number of clusters that the maximum value of the cluster silhouette index, equal to 0.406, is achieved). Figure 2,b shows the results of visualization of a reduced set of objects in two-dimensional space with the indication of object numbers and different color shading for different clusters.

The reduced set of objects was used in further experiments.

Figure 3 shows the interval assessments of 15 CPs for 4 characteristics, set by 7 experts (in the each image, the assessments of the experts are located from bottom to top, starting from the 1-st). Table 1 shows the results of CPs dividing into classes W₁ and W₂ based on the preliminary individual CPs sorts, as well as the results of CPs classification for the purely risky estimation strategy (δ = 0), when the numbers β which are the upper (right) boundaries of the intervals [α, β] are chosen as CPs characteristics assessments. A herewith, the scoring scale is formed in accordance with what CPs assessments are actually used when using the purely risky evaluation strategy.

Also, Table 1 shows the characteristics of the object z₁₆ which was recognized as the noise.

Figure 4 shows boxes with whiskers for the reduced set of objects, by analyzing which we can determine by what scores the characteristics had the largest outliers.

Here, the “green” triangular markers represent the median, and the vertical lines represent the mean for the feature, “red” square markers represent the outliers for the feature.
Also in the figure 4 we can see the lower and upper quartiles, the minimum and maximum values of the sample and outliers. The distances between different parts of the box allow you to determine the degree of scatter (variance) and skewness of the data and to identify outliers.

**Figure 3.** Interval expert assessments of CPs according to assessment characteristics.
The ideal distance between the classes for analyzed CPs turned out to be 105, and the real distance, according to the calculation results, was 71.

For analyzed CPs, the set of approximating boundary values of assessments \( p_j \) by characteristics, ordered in descending order of distance values \( d(P_{j1}, P_{j2}) \), can be written as:

\[
\{ p_1^2, p_1^3, p_2^3, p_3^3, p_4^3, p_2^2, p_2^3 \},
\]

and, therefore, the most important characteristic for classifying CP to the class “To accept the project for implementation” is the characteristic \( P_1 \) (“social and economic significance”), and the next in importance are characteristics \( P_3 \) (“financial level of the applicant”), \( P_4 \) (“relevance and novelty”), \( P_2 \) (“competitiveness”).

In accordance with the set of approximating boundary values of the \( p_j \) for the characteristics of GCDRs have the following form (table 2).

1. If the value of the assessment for the characteristic \( P_1 \) is equal to 2 or 3, “To accept CP for implementation” with the approximation index of 0.944.
2. If the value of the assessment for the characteristic \( P_1 \) is 2 or 3; the value of the assessment for the characteristic \( P_3 \) is equal to 2 or 3, “To accept CP for implementation” with an approximation index of 0.859.
3. If the value of the assessment for the characteristic \( P_3 \) is 2 or 3; the value of the assessment \( P_3 \) is 2 or 3; the value of the assessment for the characteristic \( P_4 \) is equal to 2 or 3, “To accept CP for implementation” with the approximation index of 0.831.
4. If the value of the assessment for the characteristic \( P_1 \) is 2 or 3; the value of the assessment \( P_3 \) is 2 or 3; the value of the assessment \( P_4 \) is 2 or 3; the value of the assessment \( P_2 \) for the
characteristic is equal to 2 or 3, “To accept CP for implementation” with the approximation index of 0.803.

Table 2. GCDRs, allowing to assign the class “To accept the project for implementation” for the project.

| Rule | $P_1$ | $P_3$ | $P_4$ | $P_2$ | Approximation index |
|------|-------|-------|-------|-------|---------------------|
| 1    | 2 or 3|       |       |       | 0.944               |
| 2    | 2 or 3| 2 or 3|       |       | 0.859               |
| 3    | 2 or 3| 2 or 3| 2 or 3|       | 0.831               |
| 4    | 2 or 3| 2 or 3| 2 or 3| 2 or 3| 0.803               |

Analysis of the values of the features of a noise object in the context of the obtained rules, which approve the next rating in descending order of importance of the characteristics $P_1$, $P_3$, $P_4$, $P_2$ allows us to conclude that the corresponding CP was highly rated by experts for more important features and low for less significant ones, which ultimately led to assignment him to the category of noise.
All 4 characteristics of the assessment were involved in the formation of GCDRs, that is, for all characteristics, there are approximating boundary values of the assessments. A herewith, there are no exact GCDRs and all 4 approximate GCDRs provide the same approximation of the preliminary expert division of the CPs into two classes: only the CP \( z_2 \), previously referred to the class “To reject the project”, was erroneously assigned as the result of the approximation to the class “To accept the project for implementation”. It is advisable to take the 1st GCDR as the final GCDR, since the approximation results for all GCDRs are the same.

Figures 5 – 8 show dependencies and diagrams characterizing the process of formation of GCDRs for various strategies for assessing CPs.

It can be seen from figures 5 – 8, a change in the assessment strategy can lead to a change in the number of GCDRs (that is, to a change in the number of assessment characteristics, the values of which must be taken into account when performing the classification) and the accuracy of the approximation (the number of classification errors) in the final GCDR.

Figure 7 presents a diagram showing the number of classification errors in each group of 4 GCDRs (when displaying rules from left to right, starting with the 1st), corresponding to the certain assessment strategy.

Based on the diagram in figure 8, recommendations on the use of certain strategies for assessing CPs can be formed, in particular, it can be recommended to abandon strategies with \( \delta = 0; 0.5; 1; 2 \), characterized by lower values of the approximation indicators for the assessment characteristics of CP (the characteristics in each group are located from left to right, starting from the 1st), especially in the case when it is supposed to perform the subsequent ordering and selection of CP for funding.

Table 3 shows the results of ordering (ranks) of 12 CPs according to the distance from the “anti-ideal” (worst) CP for different strategies for assessing CPs under the condition of equilibrium of the assessment characteristics (the higher the rank, the worse the CP). A herewith, with assessment strategies, the values \( \delta \) of indicators of which lie in the ranges: 0.1 – 0.4; 0.6 – 0.9; 1.1 – 1.5; 1.6 – 1.9; 2 – 3, the results of ordering CPs remain unchanged, but differ significantly from each other when changing the range or choosing assessment strategies with \( \delta = 0; 0.5; 1 \). Thus, the choice of one or another assessment strategy can have a significant impact on the results of ordering CPs.

**Table 3. Ranks of CP in the ordering list.**

| Number of CP | Assessment strategy \( \delta \) |
|--------------|---------------------------------|
|              | 0  | 0.1 – 0.4 | 0.5 | 0.6 – 0.9 | 1  | 1.1 – 1.5 | 1.6 – 1.9 | 2 – 3 |
| 1            | 7  | 4          | 4   | 4          | 4  | 4          | 4          | 4     |
| 2            | 8  | 12         | 12  | 12         | 12 | 11         | 11         | 11    |
| 3            | 4  | 2          | 2   | 2          | 3  | 3          | 3          | 3     |
| 4            | 1  | 10         | 8   | 10         | 7  | 5          | 5          | 5     |
| 5            | 12 | 11         | 11  | 11         | 11 | 10         | 9          | 10    |
| 6            | 2  | 1          | 1   | 1          | 1  | 1          | 1          | 1     |
| 7            | 6  | 7          | 10  | 7          | 10 | 12         | 12         | 12    |
| 8            | 11 | 8          | 9   | 8          | 9  | 9          | 7          | 9     |
| 9            | 10 | 6          | 6   | 6          | 5  | 6          | 8          | 6     |
| 10           | 9  | 9          | 7   | 9          | 8  | 8          | 6          | 8     |
| 11           | 5  | 5          | 5   | 5          | 6  | 7          | 10         | 7     |
| 12           | 3  | 3          | 3   | 3          | 2  | 2          | 2          | 2     |
6. Conclusion
The experimental results confirm the effectiveness of the proposed linguistic approach to object classification. At the same time, the involvement of the nonlinear dimensionality reduction algorithm (UMAP algorithm) and cluster analysis algorithms for additional data analysis makes it possible to identify and exclude noise objects from consideration and, as a result, to ensure the construction of more accurate classification rules.

We aim to develop a methodology for detecting outliers and novelty in data presented on the basis of multisets in the further research, and consider using the Isolation Forest algorithm [19] as a single-class classification algorithm for that purpose.

References
[1] Vapnik V 1998 Statistical Learning Theory John Wiley NY 732 p
[2] Lean Yu, Shouyang Wang, Kin Keung Lai and Ligang Zhou 2008 Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines (Springer-Verlag Berlin Heidelberg)
[3] Demidova L, Sokolova Yu, Klyueva I, Stepanov N and Tyart N 2016 Intellectual Approaches to Improvement Of the Classification Decisions Quality On the Base Of the SVM Classifier XII International Symposium eIntelligent Systems-2016 (INTELS 2016) pp 222-30
[4] Awad M and Khanna R 2015 Support Vector Regression Efficient Learning Machines 67-80 https://doi.org/10.1007/978-1-4302-5990-9_4
[5] Dahan H, Cohen S, Rokach L, and Maimon O 2014 Proactive Data Mining with Decision Trees
[6] Saffari A, Leistner C and Jakob Santner J 2009 On-line Random Forests 3rd IEEE ICCV Workshop on On-line Computer Vision pp 112-27
[7] Aydn Olgun and Guldamlasioglu Seren 2017 Using LSTM networks to predict engine condition on large scale data processing framework 2017 4th International Conference on Electrical and Electronic Engineering (ICEEE) pp 281-5
[8] Silkina U I and Balandin V A 2020 Discriminant analysis of variational pulsometry parameters Russian Technological Journal 8(3) 81–91 (in russ) https://doi.org/10.32362/2500-316X-2020-8-3-81-91
[9] Bakheet Aldosari, Ghada Almodaifer, Alaaeldin Hafez and Hassan Mathkour 2012 Constrained Association Rules for Medical Data Journal of Applied Sciences 12 1792-800
[10] Petrovsky A B. 2004 Multi-Attribute Classification of Credit Cardholders: Multiset Approach International Conference on Multiple Criteria Decision Making pp 6-11
[11] Petrovsky A B 2003 Cluster Analysis in Multiset Spaces Information Systems Technology and its Applications, International Conference ISTA'2003 pp 109-19
[12] Petrovskiy A B 2004 Multicriteria decision making on conflicting data: an approach of multiset theory Information technologies and computing systems 2 56–66
[13] Petrovskiy A B 2003 Ordering and classification of objects with conflicting features Artificial intelligence news 4 34–43
[14] Liu P, Zhang X and Liu W 2011 A Risk Evaluation Method for the High-Tech Project Investment Based on Uncertain Linguistic Variables Technological Forecasting and Social Change 78(1) 40-50
[15] Wei Y 2009 Risk Evaluation Method of High-technology Based on Uncertain Linguistic Variable and TOPSIS Method Journal of Computers 4(3) 276-82
[16] McInnes L and Healy J 2018 UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction ArXiv e-prints 1802.03426
[17] Hastie T, Tibshirani R and Friedman J 2001 The EM algorithm The Elements of Statistical Learning pp 236-43
[18] Rousseeuw P 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of Computational and Applied Mathematics 20(1) 53-65
[19] Liu F T, Ting K M and Zhou Z 2008 Isolation Forest 2008 Eighth IEEE International Conference on Data Mining 413–22 https://doi: 10.1109/ICDM.2008.17