Fuzzy Set Objects Clustering Method Using Evolution Technologies

Hryhorii Hnatiienko[1] and Oleh Suprun[1]

[1] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
g.gna5@ukr.net, oleh.o.suprun@gmail.com

Abstract. Presented in the article method allows an expert, or person, who makes decisions, to effectively evaluate, calculate and make the best decision on the large set of opinions. The biggest problem while performing the trustful quality assessment is to build an adequate comparison matrix, based on human opinions, since reviews may vary greatly, and it's impossible to appreciate them all, when dealing with big data. The number of these reviews may be up to hundreds, when working with the group of experts, or even thousands, operating with clients' opinions. Thus, a method to rank and cluster these results, for further operation and making decisions, is required. Analyzing the fuzzy set objects, most of classic algorithms can not be used, since it's impossible or hard to design a proper target function, or it can be discontinuous or impossible to evaluate by classic methods, so the evolution technology method is used. It allows not only to build a proper cluster around the given centers, but also to find these centers, to evaluate the best radius and the most important attributes. Besides that, large amount of parameters allows an expert to choose the most important aspects or vary the time, used to solve the problem. Another benefit is the possibility of an algorithm to adapt to the external impacts and improve the result. The algorithm and required preparations are described in the article.

Keywords: Complex Information Systems, Fuzzy Sets, Evolution Technologies, Objects Clustering.

1 Introduction

In the modern information society, more and more information require evaluating, investigating and ranking every day. It can be seen in every day life, when planning the day or the nearest future, choosing different products in the shop, or looking for the best options of summer vocation. Most of the decisions are made subconsciously, so the person doesn't pay big attention to this process, but it's much harder to choose the best option, if there are many different aspects, or the person is unfamiliar with the required topic. In all these situations the human has to overview large amounts of information and responds to make right decision. In most cases it is done, guided by pure experience, by the advices or topic analysis.

The same problem, but in much greater scale, appears while planning the future steps of big companies, or even countries, choosing the best products and options to

improve, and simulating the economical and financial models. The consequences are very important, the choice that has to be made affect the future of many people. But, to make right decisions, it's necessary to take into account the thoughts of many experts, their opinions and advises, the number of such experts in big companies may vary from dozens to hundreds.

Beside that, nowadays, when the Internet makes possible to communicate and share information among large groups of costumers from all over the world, it's very important for companies and selling structures to analyze the opinions and feedback from all of the buyers, to choose the best product, to improve their services and give answer to all questions and comments. It is necessary to maintain the company's authority among the opponents and costumers.

Dealing with large number of opinions is a complicated task by it's own, and even more, very often they differ not only by their content, but also the form, and a lot of time is required to build a normalized table of opinions. Doing this manually is impossible for big companies, since requires a lot of time, and new opinions may appear every minute. So, the normalization and first analysis must be automated to show an expert an approximate table. To handle this automatically the fuzzy set principles are used. Also, dealing with large amount of information requires the usage of big data and data mining functions.

But, at the same time, most algorithms of data mining are made to evaluate strict statistical data that has the same form and ranged values, can be put into a table and placed according to the current program, which is used. This can't be done with expert opinions that are presented as text massage and the form may be varying very differently. Thus, the non-classic method is required to simulate the work of a team that ranks and makes basic conclusions, based on large number of opinions. It must include both analyzing the text and building the initial table of opinions for their pairwise comparison.

2 Clustering the Fuzzy Set Objects

Cluster analysis is a method of grouping experimental data into classes [1]. Performing this method, the following condition must be fulfilled: the experimental values or data that are gathered into one class or group are closer to each other than to values from other classes, according to a certain parameter or attribute. The number of clusters can be arbitrary or constant.

The main purpose of data clustering methods is to ensure that the similarity of the data that is combined into one cluster is maximal. As a result of solving the clusterization problem, the values ranking in relatively similar groups is performed. The cluster analysis application is very diverse and common in many subject areas.

It is necessary to perform the initial clustering preparation, dealing with opinions and expert reviews, first of all, to find similar ones, to analyze the overall dynamic and make initial conclusions, for example, is the new technology, used in production, effective or not, or are the customers satisfied with the made improvements. Furthermore, it is much easier to deal with clusters on the next steps, analyzing their common

features instead of each opinion separately, that can require a lot of time, when the decision must be made quickly. It is especially important in economics and financial sphere.

Of course, dealing with clusters instead of separated opinions, some detalization will be lost, since assumptions are made and the overall opinion is analyzed. But, choosing the right coefficients, this detalization loss can be lowered to the acceptable level that can be different for each situation. For example, if the mistake is found, it is more important to deal with it as fast as possible, and leave detailed analysis for the further steps.

Clustering methods can be used to construct a fuzzy set membership function [2,3]. By definition, membership function $\mu_A(x)$ quantitatively calibrates the membership of the fundamental set elements of the considerations space $x \in X$ to the fuzzy set $\tilde{A}$. The value $\mu_A(x) = 0$ means that the fuzzy set does not include an element x. The value $\mu_A(x) = 1$ means the full membership of an element to a fuzzy set. The values of the membership function from the interval $(0,1)$ numerically characterize fuzzy elements.

The designing of membership functions (belonging function, F-functions) is one of the most important stages in solving decision-making problems in the fuzzy statement. The uncertainty of measuring the attribute intensity of an object can consist of the complexity of measurement, inaccurate intensity measurement, different perception of the objects properties by experts, etc. The membership function must fulfil the following requirements [4]:

- continuity, that is a formalization of the following intuitive statement: when two solutions of the set X do not differ much from each other, the values of membership functions for these solutions are also closeto each other;
- consistency with the ratio of advantages, i.e. $\mu_D(x_1) \geq \mu_D(x_2)$ then and only when $x_1 \succ x_2$.

Let a series of empirical data in a range $(0,1)$, obtained as a result of measuring a certain value x is given. According to the data analysis results, it is necessary to formulate a conclusion acceptable to the researcher, what values the indicated variable acquires. One-dimensional analysis suggests describing the distribution of one variable, including its central trend (including average values, median and mode), and dispersion (including the range and quintile of the data set, and dissemination measures such as variance and standard deviation). Since for many practical tasks the determination of the mode, the median, or the average mean of a given series is insufficient, a cluster analysis can be applied to the specified series. It is necessary to define clusters that can be used to build a membership function for the value x, which best classify the results of measuring the value x and allow to design an membership function for the value x in the interval $(0,1)$.

The membership function of a fuzzy set is a generalization of the indicator function of the classical set. In fuzzy logic, it is a degree of truthfulness. The degrees of truth are sometimes confused with probability, but they are fundamentally different, since truth indicates that the value belongs to a given set that is not similar to any phenomenon or condition. Algorithms for determining the membership function by analyzing the frequency of values were considered in [5, 6].

Fuzziness appears when the expert tries to quantify the subject area. Obtaining fuzzy knowledge is an extremely difficult task, such as experts, as a rule, are not able to adequately design a membership function. Therefore, so-called standard membership functions, which have a given form and are described by well-known analytic functions, are often used. Standard membership function features are easy to apply to many practical tasks.

Usually, there are two groups of methods for constructing a membership function: direct and indirect. In direct methods, the expert directly sets the rules for determining the values of the function: a table, defined by formulas or with examples. Direct methods are used to describe the concepts that are characterized by features that can be measured: height, weight, volume, etc.

Using the direct methods is the best option when dealing with more or less standard opinions, for example, when the experts or customers are asked the same questions with direct answers. Of course, this is rather partial option, but at the same time it must be used when possible, to save time. The indirect methods require more time and more knowledge about domain area, thus, more preparation, but they allow to make better and more precise function for current situation. Besides, such function may be more complicated, so it is impossible to use classic analyzing methods, or their performance will take a lot of time.

3 Clusterization Algorithm

When solving multicriteria optimization problems, the problem of determining the Pareto domain is strictly objective and solved without the use of any heuristics [7]. Narrowing the area of effective objects requires the use of additional information from experts, since effective set of parameters can not be compared formally with each other.

But, such additional information may be found far not always. Gathering the information, such as opinions, takes time, and during assembling of a new expert team and getting new opinions the situation may change. And customers far not always are ready and willing to answer more questions, also taking into account anonymous replies, which also must be considered for making the services better. Expanded reply system could help with it, but bigger reply far not always means better reply. So, the system has to work in conditions of lack of information. It is possible only making some assumptions, that may have negative influence on the result, but their usage is necessary in general.

As a rule, three heuristics are used to determine a single solution to a multicriteria problem:

- one of the allowed transformations is used to transform all values of object parameters to dimensionless form in a given value range;
- the vector of the criteria relative importance is determined;
- it is assumed that the multicriteria problem solution is the point of intersection of the normalized weight coefficients of criteria relative importance beam with the field of effective alternatives to the problem.

It is known that designing a structured table of benefits in a formalized form is a complex task for a human, since it is rather hard to present the opinion as a set of numbers for different categories. Research in the field of expert evaluation tasks and the practice of building decision support systems show that far not always experts and decision makers have a clear idea of the structure of preferences for a plurality of objects. In most cases, a person can not adequately determine the weight factors, or allocate in the obvious case the heuristics that are used in a decision-making situation. This is especially important, when dealing not with experts and professionals in the current domain, but analyzing the customers' opinions.

A common method of presenting the weight coefficients values for n objects with index $i, i \in I = \{1,...,n\}$ are valid numbers taking into account the condition of normalization is the following:

$$\sum_{i \in I} \rho_i = 1,$$

where

$$\rho_i > 0, \ i \in I.$$

Let the result of a calculation series the weight coefficients set is formed:

$$\rho_i \in \left\{ \rho_i^1, ..., \rho_i^L \right\}, i \in I,$$

where L is the number of values' indexes of normalized weight coefficients obtained as a computations result. Based on the obtained values, the membership functions of the weight coefficients values in the fuzzy set (0,1) are determined. Approaches to the determination of membership functions and algorithms for constructing membership functions based on the analysis of the values frequency are given in [5]. That is, each weighted coefficient as a result of the described procedure application will be characterized by its function of membership to the fuzzy set.

4 Evolution Technologies Implementing

Using the elements of evolutionary technologies for clustering fuzzy sets objects is logical and can give practically useful results, since they allow to avoid most problems, that the classic methods meat [8]. For example, in most cases the target function

is undifferentiated, it may not have a certain interpretation, since it's built upon the human opinions that are rather hard to normalize.

Although evolutionary methods do not necessarily have a concrete result, the resulting data obtained as a result can be effectively used in practice. It is especially useful when calculating the overall customers opinion and improving the service – it is impossible to reach the perfect state, since it depends on customers desires and always change, and the evolution technologies allow to fend a better solution, then implemented, at each current situation.

The most common situation while evaluating the expert's conclusions is the need to analyse their thoughts on the set of objects according to more then one attribute.

Thus, the clusterization problem can be stated as follows: to split the set Ω, that consists of n objects into m clusters $\Omega = \{S_1, S_2, ..., S_m\}$. Every object S_i has the set of characteristics, $i, i \in I = \{1, ..., n\}$.

The objects data are presented in the "object-attribute" type table:

TABLE I. OBJECTS DATA

X / S	X_1	X_2	…	X_k
S_1	ρ_{11}	ρ_{12}	…	ρ_{1k}
S_2	ρ_{21}	ρ_{22}	…	ρ_{2k}
…	…	…	…	…
S_n	ρ_{n1}	ρ_{n2}	…	ρ_{nk}

where, $X_i, i \in I = \{1, ..., \mathrm{k}\}$ are the characteristics or attributes of each object, that is analyzed.

Let the set $C^* = \{C_1^*, C_2^*, ..., C_m^*\}$ be the solution of the clustering problem. Considering that the objects are points located inside a k - dimensional hyperparallelepiped, the following can be obtained:

$$C^* = \arg\min_{C \in \theta} F(c) = \arg\min_{C \in \theta} \min_{(C_1, C_2, ... C_m)} \sum_{i=1}^{n} \sum_{j=1}^{m} \chi\{S_i \in R_j\} \cdot d(S_i, C_j),$$

with restriction that all potential cluster centers lie inside the hyperparallelepiped, $C = (C_1, C_2, ..., C_m)$.

Obtaining the cluster centers, it will be possible to operate not with all the objects-opinions, but only with their centers, that can be presented as averaged object. Number of these objects will be much lower and easier to analyze.

To find the problem solution, genetic algorithm can be used, since it is one of the most flexible evolution methods. Also, the evolution strategies usage is presented as example [9].

The solution search can be presented as a sequence of the following steps:

Step 1. Generate q sets, that consist of m elements $C^i = \left(C_1^i, C_2^i, ..., C_m^i\right), i = \overline{1,q}$ and, as usual, q is a number from the set. The values C_j^i are uniformly distributed in the hypercube, $i = \overline{1,q}, j = \overline{1,m}$,

Step 2. Calculate the distance from each object to each cluster center:

$$d_{jp}^i = d\left(S_j, C_p^j\right) = \left(\sum_{l=1}^{k}\left(x_{jl} - c_{pl}^j\right)^2\right)^{1/2},$$

where $i = \overline{1,q}, j = \overline{1,n}, p = \overline{1,m}$.

In order to determine which cluster the objects belong to, the search problem must be solved: for $\forall i = \overline{1,q}, \forall j = \overline{1,n}$ find

$$\arg\min_p d(S_j, C_p^i).$$

Table 2 is obtained, where p_{ij} is a cluster number, which the object S_j belongs to, for i-th potential problem solution.

TABLE II. OBJECT-CLASTER ATTACHMENT

S \ i	S_1	S_2	…	S_k
1	p_{11}	p_{12}	…	p_{1n}
2	p_{21}	p_{22}	…	p_{2n}
…	…	…	…	…
q	p_{q1}	p_{q1}	…	p_{qn}

Table 2 must be built anew for each set of clusters centers, since, using the evolution algorithm, a lot of false results will be found, and only working with them, the resulting answer, according to the parameters, can be received.

Step 3. Calculate the distance form every object to the center of appropriate cluster that is the potential problem solution:

$$d_i = \sum_{j=1}^{n} d(S_j, C_{p_{ij}}^i), \forall i = \overline{1,q}.$$

Step 4. If the Genetic Algorithm is chosen to solve the problem, then, considering the value d_i, the following operations are made with the set C^i: crossover operation, mutation, and elite selection into the new population of potential solutions is made.

If the Evolution Strategy method is used, the new potential population is generated, where every new solution is obtained from "parent" solution, by adding a normally distributed random displacement $X_{new} = X_{parent} + \xi(N(0, \delta^2))$. The amount of new potential solutions, as usual, is 7 times bigger, than the amount of "parent" solutions. The best solutions from the "parent" and intermediate populations are selected to the new population.

The choice of the used evolution algorithm is based on each current problem and situations, since almost every problem is unique.

Step 5. Steps 3-5 are repeated until the criteria for iterative process stop, that are set by the expert before the algorithm starts and can be changes for result improving, are not achieved.

Such criteria may be:

– the priori number of iterations;

– for a given ε:

$$\left| \max_i d_i^{(it)} - \max_i d_i^{(it+1)} \right| < \varepsilon ,$$

or

$$\left| avg_i\, d_i^{(it)} - avg d_i^{(it+1)} \right| < \varepsilon ,$$

or

$$\max_i d(C^{i(it)}, C^{i(it+1)}) < \varepsilon ,$$

where it is iteration number.

These criteria can be used in different situations. The number of iterations allows to get a result in a certain time period, when the solution must be found fast, and with set accuracy it is possible to find the result, good enough for the current problem.

Also, other criteria can be used. For example, the results in different iterations can be compared between each other, it allows to control the convergence rate, and, hence, not to waste additional time, if improving the current result will be ineffective and take too long.

Step 6. The clustering problem solutions, after the criteria for iterative process stop are achieved, are the following:

$$\arg\min_{C^{i(it)}} d_i^{(it)} .$$

The proposed clustering method is a parametric method and it's efficiency depends on researcher's qualification and efficiency of parameters setting for each specific problem.

These parameters are the following:

- The iteration process stop criteria;
- The crossover and mutation type;
- The type of parents selection and population of the next generation formation, if the Genetic Algorithm is chosen as optimization method;
- The parent-solutions and offspring-solution number;
- Constant or variable value standard deviation;
- Positive or negative dynamics of standard deviation for the Evolution Strategy.

It is known that the convergence in probability holds for a genetic algorithm with an elite selection to a new solutions population, and for $(\lambda + \mu)$ – Evolution Strategy, where λ - parent-solution number, and μ – offspring-solution number with iterations number tending to infinity.

Depending on each problem and desired solution accuracy, other evolution algorithms, like differential evolution, can be used. Besides that, modifications of the classic genetic algorithm can be made to improve the result, besides the coefficient changing. At the same time, it is necessary to test these changes on a simple model before the actual usage to avoid mistakes in future.

5 Conclusions

Working with big data is one of many aspects, connected with modern information society, and nowadays a lot of ranking and clustering methods are designed and improved. They are required to evaluate and compute large amount of statistical data, like economical or financial reports and other. But the main flaw of most of them is that these methods can be used only with the same systematic data.

At the same time, for large companies it is necessary to combine, or at least rank the reviews and opinions of many experts, that is impossible using the classic methods. Also, dealing with these opinions manually is impossible because of their large number.

To solve this problem the fuzzy sets are used. They allow to formalize and combine different opinions and to make some conclusions, based on them. The main principles of performing the fuzzy set objects are used, so that the proper results can be obtained, and the main heuristics are described in the article. This allows to normalize the results, so that more complicated algorithm can be implemented.

Besides, even normalizing the opinions will allow making some conclusions and to find weak spots of the company functioning even without the further computing.

Clustering different objects can be mat in different fields of science, but it's rather new for the fuzzy sets, since it also requires normalized sets of data. The target function may differ very much, it can be undifferentiated and discontinuous, thus, the classic methods can't be used. The evolution algorithm performance is presented, that allows to conduct the clustering process in permissible time period.

The evolution algorithm steps are presented, they are not strict and can be changed depending on every current problem, that allows to improve the algorithm every time instead of using the stereotyped one. The large number of coefficients also allows to adapt the algorithm for each task. This requires more preparation time and some initial analyzes, but allows to get a better result, using less time.

As for the future investigations, the possibility of genetic algorithm improvement must be considered. For example, the usage of evolution strategy, or implementing the penalty method to improve the clusterization results.

References

1. M. Ester, H. P. Kriegel, J. Sander, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Second International Conference on Knowledge Discovery and Data Mining, pp. 226-231, 1996.
2. E. Nakamura and N. Kehtarnavaz, Optimization of fuzzy membership function parameters," 1995 IEEE International Conf on Syst., Man and Cybern., Intelligent Systems for the 21st Century', vol. 1, 22-25 Oct. 1995, pp. 1-6.
3. S. Himavathi, B. Umamaheswari, New membership functions for effective design and implementation of fuzzy systems, IEEE Trans. Syst., Man and Cybern., vol. 31, no. 6, pp. 717 - 723, Nov. 2001.
4. A. Kaufman, M. Gupta, Introduction to fuzzy arithmetic: theory and applications. NY: Van Nostrand Reinhold, 1991.
5. Hnatyenko H.M. Algorithms for determining the membership function by analyzing the values frequency, Proceedings of the III-th international school-seminar "Theory of decision-making", Uzhorod, 2006. - P. 32-34.
6. Hnatyenko H.M., Snityuk V.E. Expert Decision Technology: Monograph, K: LLC "Maclaut", 2008. – 444p.
7. Patrick Ngatchou, Anahita Zarei and M.A. El-Sharkawi, Pareto Multi Objective Optimization, Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems (ISAP 2005), pp. 84-91, IEEE Press, Washington, DC, USA, 6-10 November, 2005.
8. V. Y. Snytyuk ; O. O. Suprun: Evolutionary techniques for complex objects clustering, 2017 IEEE 4th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)
9. Hai-hui Wang ; Wen-jie Zhao, Data clustering based on approach of genetic algorithm, 2008 Chinese Control and Decision Conference, 2-4, July 2008.