Multi-level hybrid recommender decision support system with verbal output

S V Novikova¹, Y A Tunakova¹, A R Shagidullin² and E S Kremleva¹

¹ Kazan National Research Technical University named after A.N. Tupolev (KNRTU–KAI), Kazan, 420111, Russia
² Institute for Environmental Problems and Subsoil Use of the Academy of Sciences of Tatarstan, Kazan, 420087, Russia

E-mail: SVNovikova@kai.ru

Abstract. The article describes decision-making methods based on intelligent learning algorithms, for the construction of which verbal elements are used. Such algorithms and methods operate in calculations with strictly quantitative data, however, taking into account the human way of perceiving information in verbal form. A person does not directly participate in the process of constructing a model, that is, its structure does not depend on expert or other human opinions, however, qualitative verbal information (for example, elements of regulatory acts, documents, orders, etc.) is embedded in the algorithm in encoded form.

Unlike the classical methods of fuzzy logic and fuzzy inference, the proposed models consist of several interconnected classical and author’s models, using author’s numerical methods. This increases their accuracy and adequacy. The great advantage is the automatic way to build these models, only on the basis of sets of initial quantitative and qualitative data, which will allow you to quickly and efficiently create recommender systems for solving a wide range of decision-making problems.

The system is built in three stages:

First, the fuzzy logical inference Takagi – Sugeno – Kang system (TSK) is automatically built on the sounding of the available quantitative and qualitative data. The advantage of such a system is that it can be retrained (adjust its parameters).

Then, after the system is built according to the available data, at the second stage its accuracy is increased using a fuzzy neural network. The drawback of the constructed system is that it does not give out a verbal answer.

At the third stage, based on the adjusted parameters of the TSK system, a Mamdani type system is constructed using a special algorithm. As a result, it can produce a verbal answer from a recommender system, and at the same time gets all the advantages of a neural network refinement.

Computational experiments are presented.

1. Problem statement

The opinions of expert analysts are used to create models of decision-making in those areas of knowledge where decisions are made on the basis of primarily not quantitative, but qualitative data (jurisprudence, psychology, ecology, etc.). This makes it possible to build flexible human-centered models capable of carrying in their structure information about the experience and qualifications of experts, and during operation to receive verbal responses as recommendations.
However, in practice, situations arise when only quantitative data are at the disposal of the researcher to create recommendatory models, as a rule, these are measurements of the parameters of the modeled system.

Based on the available quantitative data, as well as current verbal definitions and concepts of the system, such as regulations, verbal definitions, job descriptions, etc., it is necessary to be able to build a decision support system that would generate recommendations in natural language.

2. Development of the concept of automatically generated DSS taking into account quality information.

Taking into account the formulation of the problem, the data at the disposal of researchers are real multidimensional tuples of the form:

| System parameter №1 | System parameter №2 | …… | System parameter №n | System output parameter |
|----------------------|----------------------|-----|----------------------|------------------------|
| \( x_1 \)             | \( x_2 \)             | \( x_n \) | \( y \)               |
| Qualitative interpretation | \( A_1 \)  | \( A_2 \) | \( A_n \) | \( B \)             |

Here:

\( y \) – measured or uniquely calculated based on values \( x_i \), DECISIVE quantitative value, on the basis of which qualitative recommendations are given by the DS system

\( x_i \) – measured quantitative parameters of the system, on which the DECISIVE variable \( y \) depends.

The dependence of \( y \) on \( x_i \) can be either explicit, given in the form of a formal description, or implicit.

\( A_i \) – known qualitative expressions of quantitative parameters \( x_i \) (verbal definitions of parameters in technical documentation, regulations, instructions, etc.).

\( B \) – the required qualitative expression for the decisive parameter \( y \). It is this qualitative meaning (in the form of a lexical variable) that will be the response of the DSS with elements of human perception of information.

It follows from this that the task of creating a recommendatory model with verbal output should be solved in stages:

1. Determination of the decisive parameter \( y \) in the form of a quantitative value.

2. Analysis of the obtained result and its interpretation in the form of a qualitative assessment.

Since we need to calculate \( y \) taking into account the qualitative representation \( A_i \) values of \( x_i \), in this case it is impossible to apply such calculation methods as building a regression model of the form

\[
\begin{align*}
\text{defuzzification operation} & = \text{calculation of a weighted average with weights proportional to the degree of operation of each rule, where the right-hand sides of the rules are } \\
\end{align*}
\]

where \( M \) is the number of tuples in the set, \( X^k = (x_1^k, x_2^k, \ldots, x_3^k) \) - specific parameter values \( (x_1^k, x_2^k, \ldots, x_3^k) \) for the \( k \)-th dataset, \( y^k \) - specific value of the output quantity \( y \) for the \( k \)-th dataset

Since it is necessary to form a model that, on the one hand, is able to process the quantitative values of the system parameters, and on the other hand, to interpret them qualitatively, the preferred way to solve the problem is fuzzy logic methods. In such methods, we can set the input parameters as words and then represent them as numbers – values of some transform function.

The calculation of the parameter \( y \), as a quantitative estimate, is most conveniently carried out on the basis of the Takagi-Sugeno-Kanga (TSK) fuzzy inference system [1]. In the Takagi-Sugeno system, the defuzzification operation is the calculation of a weighted average with weights proportional to the degree of operation of each rule, where the right-hand sides of the rules are
specified as nonrandom linear functions of the parameters $x_i$. In order to define the right-hand sides of the rules as linear combinations of parameters $x_i$, it is advisable to apply the method of linear multivariate regression.

For the subs OR operation equent refinement of the regression coefficients, as well as the parameters of the membership functions of the left sides of the rules, it is necessary to use the possibility of representing the TSK system in the form of a fuzzy neural network with its subsequent training.

The qualitative interpretation of the parameter $y$ as a linguistic variable is most conveniently carried out on the basis of the Mamdani fuzzy inference system, in which the right-hand sides are set in the form of a condition for the output variable to belong to one of the fuzzy values, the composition is a logical OR operation, and defuzzification consists in calculating the center of gravity of the figure under the graph of the final membership function [2].

3. Model for calculating the final parameter in the number form based on TSK-system.

The TSK model should be described as $m$ fuzzy rules:

$$\text{If } (x^*_1 \in A_1) \text{ AND } (x^*_2 \in A_2) \text{ AND } ... \text{ (} x^*_n \in A_n) \text{ THEN } y = b_{j0} + b_{j1}x_1 + b_{j2}x_2 + ... + b_{jn}x_n$$

$j=1,2,...,m; \ b_{ji}$ some numbers.

Here:

$x^*_i$ - quantitative value of the $i$-th parameter of the system

$A_i$ - fuzzy sets that verbally correspond to qualitative gradations of quantitative parameters $x_i$

Based on the form of fuzzy rules of the Takagi-Sugeno model, we can say that they switch control between several linear laws. Since the input variables are qualitative and are described by the membership functions, several rules can be fulfilled simultaneously, and, accordingly, several linear control laws will act simultaneously, but with different strengths. The degree of membership of the input vector $X=(x^*_1,x^*_2,...,x^*_n)$ to values $d_j = b_{j0} + \sum_{i=1}^{n} b_{ji}x_i$ is calculated as follows:

$$\mu(d_j(X^*)) = \bigwedge_{i=1}^{n} \mu^j_{A_i}(x_i^*) \bigg| \ j=1,m \quad (2)$$

Here $\bigwedge$ - operation from the t-norm, i.e. from many implementations of logical AND operations, $\mu^j_{A_i}(x_i)$ - membership function of the set $A_i$.

In Sugeno’s fuzzy inference, the product is most often used as a t-norm. As a result, we obtain a fuzzy set $\tilde{y}$ corresponding to the input vector $X^*$:

$$\tilde{y} = \frac{\mu(d_1(X^*))}{d_1} + \frac{\mu(d_2(X^*))}{d_2} + ... + \frac{\mu(d_m(X^*))}{d_m} \quad (3)$$

Note that the above fuzzy set is an ordinary first-order fuzzy set. It is given on a set of clear numbers. The resulting value of the output $y$ is determined as a superposition of linear dependencies performed at a given point $X^*$ of the $n$-dimensional factor space. To do this, defuzzify a fuzzy set $\tilde{y}$ by finding a weighted average:
or weighted amount:

\[ y = \sum_{j=1}^{m} \mu(d_j(x^*)) \cdot d_j. \tag{5} \]

3.1. Defining the parameters of the left parts of the rules

To set the parameters of the left parts of the inference rules of the form (1), it is necessary to define:

1. The form of membership functions \( \mu_j(x^*_i) \) to each fuzzy set \( A_i \).
2. Parameters of the selected membership function.

There are two approaches to solving these problems - expert and automatic.

With an expert approach, subject matter experts independently, from non-formalized considerations, determine the type and parameters of functions \( \mu \) [3].

Based on the need to create an automatically generated system, we will consider an automatic procedure. Generation requires an adequate set of quantitative inputs / outputs provided by a trusted source.

The procedure for automatically setting the left-hand sides of the TSK system inference rules based on the available tuples of numerical data:

1. Each element \( x_i \) of the input vector \( X=(x_1, x_2, \ldots x_n) \) is ordered and ranked according to different sample quantiles. The number of quantiles is allocated in accordance with the number of qualitative gradations of the parameter \( x_i \). For example, if the parameter \( x_i \) is described by four qualitative gradations: “low parameter”, “medium parameter”, “high parameter”, “above average parameter”, then the ordered set of all possible values of this parameter should be divided into 4 quantiles:

- 0,25\% quantile \( \Rightarrow \) “low parameter”;
- 0,5\% quantile \( \Rightarrow \) “medium parameter”;
- 0,75\% quantile \( \Rightarrow \) “high parameter”;
- >0,75\% quantile \( \Rightarrow \) “above average parameter”;

2. Each value of the qualitative estimate of the parameters of the vector \( X \) is associated with the corresponding quantile.

3. For each of the linguistic variables \( A \), a membership function of the Gaussian type is determined.

\[ \mu_A(x_i) = \frac{1}{1 + \left( \frac{x_i - c_i}{\sigma_i} \right)^2} \]
The parameter $c$ of the Gaussian (its center) will be taken as the center of the corresponding quantile, and the spread of the Gaussian ($\sigma$) is calculated as the 25% outer offset from the quantile boundaries divided by three (according to the 3-sigma rule).

3.2. Defining the parameters of the right parts of the rules.

The right-hand sides of the rules of the form (1) are linear combinations of elements of the input vector $X$. Therefore, to determine the coefficients $b_{ji}$, it is advisable to use the method of constructing linear multiple regression models described, for example, in [4].

Algorithm for finding the coefficients of the right-hand sides of the inference rules of the TSK system based on the available tuples of numerical data:

1. Each element $x_i$ of the input vector $X=(x_1, x_2, \ldots, x_n)$ is ordered and ranked according to different sample quantiles. The number of quantiles is allocated in accordance with the number of qualitative gradations of the parameter $x_i$.
2. Each value of the qualitative estimate of the parameters of the vector $X$ is associated with the corresponding quantile.
3. Inside each quantile, the procedure for constructing a linear regression model using the method of the least squares.

The coefficients of the regression dependences found as a result are taken as the coefficients of the right-hand sides of the inference rules of the TSK system.

As a result of the sequential application of the two developed procedures, the left and right sides of the TSK rules will be formed. However, for their automatic matching (composition of each part into a single rule), an additional procedure must be developed.

3.3 Determination of correspondences between the left and right sides of the TSK system rules.

For the reasonable formation of rules (dependence of the left and right sides), the following procedure has been developed:

The procedure for generating a complete base of TSK system rules based on the available tuples of numerical data:

1. For each linguistic variable $A_i$ on the left side of the rules, the “middle range” is determined - this is the range covered by the category “middle parameter”. Accordingly, expanding the regression coefficients of the equation describing the category “average parameter” for each variable, we obtain the weights of the individual parameters:

$$V(x_i) = v_i$$

2. Normalize the weights:

$$v_i^N = \frac{v_i}{\sum_{i=1}^{n} v_i}$$

The sum of the normalized weights is 1.

3. Introduce a coding scale for fuzzy categories of qualitative definitions $x_i$. The number of scale gradations corresponds to the number of categories. For example, if the parameter $x_i$ is described by four quality gradations: “parameter low”, “parameter medium”, “parameter high”, “parameter above average”, then a 4-digit scale will be entered. Each category is assigned a CODE - a score from 1 to $K$, where $K$ is the capacity of the scale (acceptance score). For each left side of the rules generated in step 3.1. We replace the linguistic expression of the parameter state with the corresponding score, and
calculate the weighted average value of the final score based on the normalized values of \( V_i^N \) found at the previous stage.

4. The resulting value of the final score shall be rounded to the nearest integer. Each calculated score is replaced by the corresponding linguistic expression.

According to the developed procedure, the complete rule base of the TSK fuzzy inference system is formed.

3.4. Improving the accuracy of the TSK model based on machine learning.

It is possible increase model accuracy by adjusting system parameters: that is, the parameters of the membership functions of the left sides of the rules (IF blocks) and the coefficients of the linear formulas of the right sides (TO blocks). Fuzzy neural networks can be effectively used to adjust these parameters. In the learning process, such neural networks change the values of the above parameters of the Takagi-Sugeno system.

In this case, an approach is possible when a fuzzy neural network replaces the Takagi-Sugeno system, becoming its trained clone. In the process of training the network, the difference between the result of modeling a fuzzy system and the experimental data available to researchers is minimized. Minimization is carried out by correcting the parameters of the membership functions of the left sides of the rules and parameters of the linear functions of the right sides. After training such neural fuzzy network, the accuracy and, consequently, the adequacy of the original fuzzy TSK system increases.

3.5. Verbal interpretation of the quantitative response of the TSK system

To form a model capable of producing qualitative (verbal) conclusions regarding the input numerical parameters, it is necessary to use the Numerical method for obtaining a verbal recommendation based on the quantitative response of a fuzzy algorithm [2], [5]:

1) Set the number of input variables of the Mamdani system equal to the number of inputs of the Takagi-Sugeno system.
2) All fuzzy terms and their membership functions of the Takagi-Sugeno system are transferred to the Mamdani system without changes.
3) Set the number of fuzzy terms in the output variable of the Mamdani system equal to the number of fuzzy terms in the output variable Takagi-Sugeno. For each fuzzy term of the output variable, set the membership function in the form of a Gaussian.
4) The values of the centers of the Gaussians \( c^j \) are determined by the formula
   \[
   c^j = \sum_{i=1}^{N} t_i^j a_i^j + t_0^j,
   \]
   where \( a_i^j \) - the centers of the membership function of the input variables, defined as Gaussians, \( t_i^j \) - linear coefficients of the right-hand sides of the rules of the original TSK system:
   \[
   y_j = \sum_{i=1}^{N} t_i^j x_i + t_0^j
   \]
5) The values of the scatter of the Gaussians \( \sigma_i^j \) are determined from the condition of uniform coverage of the domain of definition.
6) All Takagi-Sugeno inference rules become the inference rules of the Mamdani system.
7) To obtain a verbal response, subject the response of the Mamdani system to reverse fuzzification.

4. Computational experiments
To check the performance of the developed model, test calculations of the risk parameter for the state of the environment were carried out $P_{\text{total}}$ (environmental risk in total) based on the risk values for its individual components:

- $P_{\text{air}}$ (environmental risk by air), $R_{\text{snow}}$ (environmental risk by snow), $P_{\text{soil}}$ (environmental risk by soil cover), $R_{\text{bioenv}}$. (environmental risk for biological environment of the population)

For practical implementation, testing and use of the developed fuzzy logical system, the MatLab mathematical package with the Fuzzy Logic Toolbox subsystem was used [1].

The learning error of the TSC system after training with a neural network was $7.5 \times 10^{-6}$ (less than 0.1%). That is, the accuracy of the calculations has increased by more than 30 thousand times compared to the original model without adjusting the parameters. The error of the corrected fuzzy inference system on the test set averaged $6.43 \times 10^{-4}$ (0.14%) (Figure 1):

![Figure 1. Comparative values of hazard levels according to the fuzzy inference system ($P_{\text{total}}$) and reference values ($P_{\text{real}}$) after increasing the accuracy.](image)

The results of the work of the designed Takagi-Sugeno fuzzy inference system are well demonstrated by the following examples:

Example №1:
Initial data for calculations:
1) “Air condition” $P_{\text{air}} = 0.1$ (to a greater extent corresponds to the linguistic concept “satisfactory”)
2) “Snow condition” $P_{\text{snow}} = 0.8$ (to a greater extent corresponds to the linguistic concept of “high risk”)
3) “Soil condition” $P_{\text{soil}} = 0.9$ (to a greater extent corresponds to the linguistic concept of “high risk”)
4) “State of the biological environment” $P_{\text{bioenv}} = 0.9$ (to a greater extent corresponds to the linguistic concept of “high risk”)

Calculation result: “State of the environment” $P_{\text{total}} = 0.472$ - to a greater extent corresponds to the linguistic concept of “intense”.
Example №2:
Initial data for calculations:
1) “Air condition” $P_{air} = 0.9$ (to a greater extent corresponds to the linguistic concept “high risk”)
2) “Condition of snow cover” $P_{snow} = 0$, (to a greater extent corresponds to the linguistic concept “average”)
3) “Soil condition” $P_{soil} = 0.5$, (to a greater extent corresponds to the linguistic concept “increased risk”)
4) “State of the biological environment” $P_{bioenv} = 0.4$ (to a greater extent corresponds to the linguistic concept “average”)

Calculation result: “State of the environment” $P_{total} = 0.712$ - to a greater extent corresponds to the linguistic concept of “crisis”.

5. Conclusion
The developed multilevel hybrid recommender model with verbal output can be effectively used for the automatic creation of intelligent DS systems that take into account human perception of information.

Funding
The research was carried out with the financial support of the Ministry of education and science of the Russian Federation in the framework of fulfilling obligations under Agreement number 075-03-2020-051/3 dated 09.06.2020 (topic number fzu-2020-0021).

References
[1] Sivanandam S N, Sumathi S and Deepa S N 2007 *Introduction to Fuzzy Logic using MATLAB.* (Springer-Verlag Berlin Heidelberg) 425
[2] Wang L X 1992 Fuzzy systems are universal approximators Proc. of the IEEE Int. Conf. on Fuzzy Systems. San Diego 1163-69.
[3] Tunakova Y A , Novikova S V and Shagidullin A R 2017 Calculation of the probability integral indicator of the level of air pollution *Environmental Engineering* 7
[4] Judd C, Judd M, Carey G and Carey R 2017 *Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond* (Routledge) 378
[5] Kremleva E S and Novikova S V 2014 The method of obtaining a linguistic answer from a logical system with a quantitative output *Actual issues of modernization of science Conf. Proc.* 24-26 (in Russian)