Evolutionary Optimization of TSK Fuzzy Model to Assist with Real Estate Appraisals

Tadeusz Lasota¹, Bogdan Trawiński², and Krzysztof Trawiński³

¹ Wrocław University of Environmental and Life Sciences, Faculty of Environmental Engineering and Geodesy, C.K. Norwida 25/27, 50-375 Wrocław, Poland
   tadeusz.lasota@wp.pl

² Wrocław University of Technology, Institute of Applied Informatics, Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, Poland
   bogdan.trawinski@pwr.wroc.pl

³ Wrocław University of Technology, Faculty of Electronics, Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, Poland
   krzysztof.trawinski@wp.pl

Abstract. A Takagi-Sugeno-Kang-type fuzzy model to assist with real estate appraisals was developed and optimized using evolutionary algorithms. Three approaches were compared in the paper. The first one consisted in learning the rule base, the second one in tuning the membership functions having the rule base optimized and the third one in combining both methods in one process. Six fuzzy models comprising from two to seven input variables referring to the attributes of a property being appraised were evaluated. The evolutionary algorithms were based on Pittsburgh approach with the real coded chromosomes of constant length. The experiments were conducted using training and testing sets prepared on the basis of actual 134 sales transactions made in one of Polish cities and located in a residential section.

Key words: fuzzy system, TSK fuzzy model, real estate appraisal, evolutionary learning, evolutionary tuning

1 Introduction

For years fuzzy systems have been gaining acceptance in many fields, among others in control and automation, engineering and manufacturing, pattern recognition, biomedicine, medical diagnosis and forecasting. The fuzzy systems developed are mostly treated as black boxes therefore an analytical theory for fuzzy systems to eliminate the misunderstanding and controversy is needed. Further investigation and optimization of fuzzy models proposed and comparative analysis of different approaches is needed [3], [8].

One of the new areas of fuzzy system application is real estate appraisal market with automated valuation models [2], [9], [10], [15]. The most popular approach to determining the market value of a property is sales comparison approach. Applying this approach it is necessary to have transaction prices of the
properties sold which attributes are similar to the one being appraised. If good comparable transactions are available, then it is possible to obtain reliable estimates. Prior to the evaluation the appraiser must conduct a thorough study of the appraised property using available sources of information such as cadastral systems, transaction registers, performing market analyses, accomplishing on-site inspection. Despite the fact, that he sometimes uses the expertise of the surveyor, the builder, the economist or the mortgage lender, his estimations are usually subjective and are based on his experience and intuition. Automated valuation models (AVMs) are based on statistical models such as multiple regression analysis, soft computing methods and geographic information systems (GIS) [2], [17]. Many intelligent methods based on artificial intelligence have been developed to support appraisers’ works: neural networks [5], [16], case-based reasoning [1], [7], data mining [13] and hybrid approaches [11], [14].

The concept of a fuzzy system to assist real estate appraisals was developed basing on sales comparison method. It was assumed that whole appraisal area, that means the area of a city or a district, is divided into sections of comparable property attributes. For each section a representative property and rule bases should be determined. The architecture of the proposed system is shown in Fig. 1.

![Figure 1](image-url)

**Fig. 1.** Architecture of the fuzzy system for real estate appraisal

The appraiser accesses the system through the internet and chooses an appropriate section and input the values of the attributes of the property being evaluated. Then the system using the parameters of the representative property for the section indicated, calculates the input values to the fuzzy model. The classic fuzzy inference mechanisms, applying a rule base generated for that section, calculates the output. Then on the basis of the parameters of the representative property the final result is determined and as a suggested value of the property is sent to the appraiser.

Mamdani and Takagi-Sugeno-Kang-type (TSK) fuzzy models were developed and initially evaluated [9], [10]. The models were built with the aid of experts and comprised 7 input variables relating to main attributes of a property being appraised, namely area, front, infrastructure, neighbourhood, arrangement, distance and communication. The rule bases of the fuzzy models were generated using evolutionary algorithms. The experiments were conducted using MATLAB.
The complete TSK-type fuzzy model for real estate appraisal was built with the aid of experts. The model comprised 7 input variables and they referred to the difference or proportion of attribute values between a property being appraised and the representative one. The representative properties were determined for one section comprising residential properties, having similar characteristics, by means of calculating average values of attributes of all properties in the set of data used in the experiment. For each input variable five triangular and trapezoidal membership functions were defined (see Fig. 2). Therefore the input of the fuzzy system was defined by the vector of seven following variables:

- **Distance** - it is the difference in the distance from a local centre expressed in meters. The domain of this variable is the interval form -1000 to 1000 meters. Negative values denote that the representative property is located closer to the local centre than the appraised one.
- **Front** - it is the difference in the length of fronts of parcels expressed in meters. The domain of this variable is the interval form -50 to 50 meters. Positive values mean that the examined parcel has longer front than the representative one what is considered as a better result.
- **Area** - it is the ratio of the area of the examined parcel to the area of the representative one. The domain of this variable is the interval form 0 to 10. Values greater than 1 indicate that the examined property has the bigger area.
- **Infrastructure, arrangement neighborhood and communication** - values of these four attributes are appraiser’s judgments of what is the difference in a given attribute between the appraised and the representative parcels. The values are taken from the range 0-200 where 100 means that both parcels are equal in this respect, values greater than 100 - that the examined parcel is better and the ones lower than 100 - the opposite.
Infrastructure - refers to what extent the technical infrastructure of a given property is better than the representative one.

Arrangement - pertains to the assessment of how better a given property was arranged than the representative one.

Neighborhood - concerns the difference in the quality of properties’ neighborhood.

Communication - deals with the evaluation of the means of public transportation available to the residents.

The TSK-type model was a zero-order one where output functions were constants representing the difference of values between appraised and representative properties. The final output of the system was calculated as the weighted average of all rule outputs using the following formula: 

$$ PVD = \sum_{i=1}^{N} \frac{y_i w_i}{\sum_{i=1}^{N} w_i} $$

where 

- $PVD$ denotes a property value difference, 
- $y_i$ - output of $i$-th rule, 
- $w_i$ - firing strength of $i$-th rule computed as the product of the membership functions for each input and 
- $N$ is the number of rules in the model. 

Using the complete model with 7 inputs as the base further models with lower and lower number of input variables were created. As the criterion of eliminating the dimensions the variability coefficient, which is expressed by the standard deviation divided by the mean, was employed. The coefficient was calculated for system input variables using whole set of transaction data (see Table 1). Each successive reduced model comprised input variables with the biggest value of variability coefficient. Thus five reduced models were obtained and together with the complete one were
subjected to evaluation. The resulting models were denoted by means of codes reflecting the input criteria in the following way 1234567, 123456, 12345, 1234, 123, 12.

Table 1. Input variables ranked by variability coefficient

<table>
<thead>
<tr>
<th>No. of variable</th>
<th>Input variable</th>
<th>Std dev.</th>
<th>Mean</th>
<th>Variability coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>area</td>
<td>1479.6650</td>
<td>1031.2933</td>
<td>1.4348</td>
</tr>
<tr>
<td>2</td>
<td>front</td>
<td>16.8018</td>
<td>21.8333</td>
<td>0.7695</td>
</tr>
<tr>
<td>3</td>
<td>arrangement</td>
<td>26.6112</td>
<td>52.3704</td>
<td>0.5081</td>
</tr>
<tr>
<td>4</td>
<td>distance</td>
<td>294.4379</td>
<td>580.5333</td>
<td>0.5072</td>
</tr>
<tr>
<td>5</td>
<td>infrastructure</td>
<td>17.2505</td>
<td>78.0667</td>
<td>0.2210</td>
</tr>
<tr>
<td>6</td>
<td>communication</td>
<td>13.4888</td>
<td>66.3373</td>
<td>0.2033</td>
</tr>
<tr>
<td>7</td>
<td>neighbourhood</td>
<td>10.1453</td>
<td>94.6963</td>
<td>0.1071</td>
</tr>
</tbody>
</table>

3 Evolutionary Methods of Optimizing Fuzzy Models

The models were implemented employing the MATLAB Fuzzy Logic Toolbox and the evolutionary algorithms were programmed to optimize them, using the functions included in the MATLAB Genetic Algorithm and Direct Search Toolbox. The structure of an evolutionary algorithm was the same as the structure of a classic genetic algorithm [6], [12], but the algorithm applied differed in the way of chromosome coding and crossover and mutation operations. Three methods of optimizing the models using evolutionary algorithms were compared. The first one consisted in learning the rule base, the second one in tuning the membership functions having the rule base optimized and the third one in combining both methods in one process. The optimization process of learning the rule base is depicted in Fig. 3a and was denoted as the SR process. In this case it was assumed that the membership functions were determined earlier by the expert and were unchangeable. The second method consisted in tuning the parameters of membership functions using the rule base obtained as the product of the previous SR process described above. The optimization process of genetic tuning the membership functions of the fuzzy model is presented in Fig. 3b and was denoted as the SF process. Finally the third optimization process combined both previous approaches and was denoted as the SM process. Both the rule base and membership functions were genetically tuned during one process as it was shown in Fig. 3c.

Training and testing sets. The set of data used in the process of generating rules comprised 134 sales transactions made in one of Polish cities and located in a residential sections what assured comparable attributes of properties. The data were taken from the governmental registry of real estate sales transactions. The
Fig. 3. Evolutionary optimizing a) SR process, b) SF process, c) SM process

attributes of the properties embraced by those transactions were determined by an expert, who had visited and studied personally all of them. The set of data was bisected into training and testing sets by clustering the property descriptions including their prices using the k-means method and then by splitting randomly each cluster into two parts. Finally the training data set counted 64 properties and the testing one 70 properties.

Fig. 4. Examples of coding chromosomes a) – f) rules in 12, 123, 1234, 12345, 123456, 1234567 models respectively, g) membership functions
Coding chromosomes. The rule base was real-coded using the Pittsburgh method, where one chromosome comprised whole rule base. The constant length of the chromosome composed of \( N \) rules was assumed. Each \( i \)-th rule was represented by \( M+1 \) genes: \( g_1^i, g_2^i, \ldots, g_{M+1}^i \), where \( M \) was the number of input variables and first \( M \) genes contained natural numbers from 1 to 5 corresponding to linguistic values of seven input variables, e.g. MLT (much less than), LT (less than), EQ (equal), GT (greater than), MGT (much greater than) respectively. Zero value on the position of a given input meant that this attribute did not occur in the rule. The \( M+1 \)-th gene represented the output and contained natural numbers which were drawn at random from the set of successive numbers from 1 to 81 representing the difference of values between appraised and representative properties expressed in Polish currency (PLN). The difference could range from -200 to 200 PLN with the step of 5 PLN, what established altogether 81 values. In the SF process chromosomes contained only the representation of membership functions. For the purpose of tuning each membership function was transformed into trapezoidal one. In a chromosome each trapezoid was represented by four genes containing real numbers which during initialization were drawn at random from the interval \( < d_i - \Delta, d_i + \Delta > \), where \( d_i \) is one of four parameters of a trapezoid and \( \Delta \) was established as 10 percent of \( d_i \) (see Fig. 5). Similar approach is described in [4]. Thus each input variable comprising 5 linguistic values was represented by 20 genes. In turn in the SM process the chromosome was constructed by attaching the representation of membership functions to the one reflecting the rule base. The examples of coding chromosomes in respective TSK fuzzy models are presented in Fig 4. The lengths of chromosomes used in SR, SF and SM processes are given in Table 2.

Initialization. At the beginning the set of all possible rules was confined to that comprising only those rules which could be activated by the set of training data and in the case of the 1234567 model it counted above 40 thousand rules. Let us call this set as activated rules. Then each chromosome was composed of rules taken randomly out of the activated rules (see Fig 6) and so the initial population of randomly generated chromosomes was obtained. In the case SF and SM processes the chromosomes or chromosome fragments representing
Table 2. Lengths of chromosomes comprising 50 rules

<table>
<thead>
<tr>
<th>Model</th>
<th>SR model</th>
<th>SF model</th>
<th>SM model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234567</td>
<td>400</td>
<td>140</td>
<td>540</td>
</tr>
<tr>
<td>123456</td>
<td>350</td>
<td>120</td>
<td>470</td>
</tr>
<tr>
<td>12345</td>
<td>300</td>
<td>100</td>
<td>400</td>
</tr>
<tr>
<td>1234</td>
<td>250</td>
<td>80</td>
<td>330</td>
</tr>
<tr>
<td>123</td>
<td>200</td>
<td>60</td>
<td>260</td>
</tr>
<tr>
<td>12</td>
<td>150</td>
<td>40</td>
<td>190</td>
</tr>
</tbody>
</table>

membership functions the values of genes were determined randomly following the principles shown in Fig. 3.

*Fitness function* was calculated as a mean relative error between values of properties included in the training set and the values of corresponding properties determined by the fuzzy system using a rule base produced by a subsequent generation of the evolutionary algorithm.

*Reproduction.* MATLAB allowed to set parameters of reproduction to determine how the genetic algorithm would create the next generation. So we set following parameters of reproduction: elite count was set to 2, crossover fraction was set to 0.8, and therefore the mutation fraction was close to 0.2.

*Crossover.* Uniform crossover operation was employed, where the pattern of the position of rules to be exchanged was determined randomly for each pair of parents separately with the probability of 0.5. According to this pattern whole rules were exchanged between the parents of a given pair instead of individual genes. In Fig. 7 it is shown that $i$-1-th and $i$+1-th rules are exchanged between parent chromosomes. In the case of a chromosome representing membership functions the operation is performed similarly, i.e. randomly determined four genes corresponding to a given trapezoid are exchanged between parents. In the SM process chromosome fragments representing rule base and membership functions are processed separately.
Evolutionary of Optimization of TSK Fuzzy Model

Fig. 7. Crossover operation a) parents, b) crossover pattern, c) children

Mutation. The mutation operation consisted in altering rules randomly selected with the probability of 0.05 in each chromosome, remained in the mutation fraction. The selected rule was replaced in a chromosome by a new one taken randomly from the set of activated rules (see Fig. 8).

Fig. 8. Mutation operation a) mutation pattern, b) parent, c) child

4 Results of the Experiment

The main goal of the investigation was to compare three fuzzy model optimization processes (SR, SF and SM) described in section 3 in respect of the convergence of the evolutionary algorithm and the best fitness achieved using training data and mean relative error with testing set of data.

All the experiments were conducted with the same number of rules in one chromosome equal to 50 and the same number of membership function parameters equal to 20 per one input variable and the size of population counting 500
Fig. 9. Comparison of SR, SF and SM processes for training and testing sets
Evolutionary of Optimization of TSK Fuzzy Model

chromosomes. The experiments were repeated for different number of generations equal to 100, 200 and 500.

The graphs for best fitness achieved using training data and mean relative error with testing set of data for different number of generations are presented in Fig. 9. Some observations can be made when analysing Fig. 9. For all models the SR process assures worse results than the other two processes so that it may be concluded it is worth to tune the membership functions of the model proposed by experts. Taking into account both best fitness and mean relative errors the 12345 model optimized by SM process provides the best results. For 500 generations the values are equal to about 0.039 and 0.155 respectively. The 12, 123 and 1234 models revealed unsatisfactory accuracy, so the reduction of input variables should be performed carefully.

5 Conclusions

Three approaches to evolutionary optimization of a TSK-type fuzzy system for assisting the property appraisers’ work were compared in the paper. The first one consisted in learning the rule base (SR process), the second one in tuning the membership functions having the rule base optimized (SF process) and the third one in combining both methods in one process (SM process). Six fuzzy systems with the decreasing number of input variables from 7 to 2 were also tested. The experiments were conducted using training and testing sets prepared on the basis of actual 134 sales transactions of residential properties.

The investigation revealed that the second and third methods provide better results compared to the model containing the rule base generated without the optimization of membership functions. So beside learning the rule base it is worth to tune membership functions during the optimization process of a fuzzy model proposed by experts. The reduced models with the least number of input variables revealed unsatisfactory accuracy, so the reduction of input variables in order to simplify the models and therefore to decrease time needed to optimize them should be performed carefully.

The experiments allowed us to determine the model providing the best output taking into account both best fitness with training set and mean relative errors with testing set. The values of the measure obtained for the training set were low, especially for 500 generations and the results for the testing set were fairly acceptable. Thus further experiments are planned with different parameters of fuzzy model and evolutionary algorithm such as different number of rules and population size. Moreover, evaluation of the forecasting capabilities of the models will be carried out in a more robust way, for example using the 10-fold cross validation method. This approach is more efficient for smaller data sets than a single data partition.

It is also planned to implement the model and to develop an internet system which can be exploited by the appraisers. The main design problem is how to refresh efficiently the rule base of the fuzzy model when new data of sales transactions will be available.
References