ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/223935113

The mass appraisal of the real estate by computational intelligence

Article *in* Applied Soft Computing · January 2011

DOI: 10.1016/j.asoc.2009.12.003 · Source: DBLP

CITATIONS
30

READS
318

2 authors, including:

Antanas Verikas
Halmstad University

142 PUBLICATIONS **1,475** CITATIONS

SEE PROFILE

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Antanas Verikas
Retrieved on: 23 July 2016

Applied Soft Computing 11 (2011) 443–448

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

The mass appraisal of the real estate by computational intelligence

Vilius Kontrimas[a,*], Antanas Verikas[b,1]

[a] Department of Information Systems, Department of Applied Electronics, Kaunas University of Technology, Studentu g. 50, Kaunas, Lithuania
[b] Department of Applied Electronics, Kaunas University of Technology, Studentu g. 50, Kaunas, Lithuania

ARTICLE INFO

Article history:
Received 10 September 2007
Received in revised form 1 December 2008
Accepted 6 December 2009
Available online 16 December 2009

Keywords:
Ordinary least squares regression
Support vector regression
Multilayer perceptron
Committee
Self-organizing map
Mass appraisal of real estate

ABSTRACT

Mass appraisal is the systematic appraisal of groups of properties as of a given date using standardized procedures and statistical testing. Mass appraisal is commonly used to compute real estate tax. There are three traditional real estate valuation methods: the sales comparison approach, income approach, and the cost approach. Mass appraisal models are commonly based on the sales comparison approach. The ordinary least squares (OLS) linear regression is the classical method used to build models in this approach. The method is compared with computational intelligence approaches – support vector machine (SVM) regression, multilayer perceptron (MLP), and a committee of predictors in this paper. All the three predictors are used to build a weighted data-depended committee. A self-organizing map (SOM) generating clusters of value zones is used to obtain the data-dependent aggregation weights. The experimental investigations performed using data cordially provided by the Register center of Lithuania have shown very promising results. The performance of the computational intelligence-based techniques was considerably higher than that obtained using the official real estate models of the Register center. The performance of the committee using the weights based on zones obtained from the SOM was also higher than of that exploiting the real estate value zones provided by the Register center.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

There are mass and individual appraisals of real estate. The individual appraisal is such an appraisal, when value of the exact object is determined according to all its individual characteristics. Mass appraisal is the systematic appraisal of groups of properties as of a given date using standardized procedures and statistical testing [40]. This valuation method is applied to property objects with many similarities. Mass appraisal of real estate is commonly applied to compute real estate tax.

The purpose of mass valuation is to estimate the market value. It must be distinguished from the market price and other, non-market values [17]. According to the Lithuanian Republic normative documents, the market value is estimated as the money amount, for which a property can be exchanged on the valuation date between a willing buyer and a willing seller in arm's-length transaction after proper marketing, wherein the parties act knowledgeably, without compulsion and impact of other transactions and interests [27,28]. In the international valuation standards 2005 (IVS), issued by the International Valuation Standards Committee (IVSC), the market value is defined as the estimated amount of money for which a property should exchange on the date of valuation between a willing buyer and a willing seller in arm's-length transaction after proper marketing wherein the parties acted knowledgeably, prudently and without compulsion [17]. The market price is formed when curves of supply and demand intersect, it is influenced by many objective and subjective factors. The market price equals to the market value very rarely, because the market of real estate is not an ideal market. The market price of real estate reflects many subjective factors, so a real estate assessor must find the most objective, suitable for all value.

There are three traditional real estate valuation methods: the sales comparison approach, income approach, and the cost approach [30,40]. According to the sales comparison approach, the value is determined by comparing the object with the other objects sold in the market. The value is adjusted according to differences, as real estate objects have the differences. A difference up to 30–35% between characteristics of various objects is acceptable [26]. This method is very suitable for clear land. Reflection of the market price, quick and simple computations is the main advantages of this approach. The income approach is based on the premise that the value is the present worth of future; the value is determined by discounting cache flows generated by the object. The approach is very suitable for objects generating incomes, for example, buildings with leased offices or flats, objects used for services or production. This approach is quite simple too and estimates the economic ben-

[*] Corresponding author.
E-mail address: kontrimv@takas.lt (V. Kontrimas).

[1] Also: Intelligent Systems Laboratory, Halmstad University, Box 823, S-301 18 Halmstad, Sweden.

1568-4946/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2009.12.003

efit from the object. In the cost approach case, the value of the object is determined by construction costs minus depreciation. This approach can be applied only to buildings, and it is very suitable for schools, objects of engineering infrastructure and similar, which do not generate incomes and there are only a few objects to compare with.

Mass appraisal models are commonly based on the sales comparison approach. The linear regression is the most popular technique used to build mass appraisal valuation models. However, other techniques such as neural networks, support vector machines (SVM), and committees of models can also be used.

Most studies on this topic compare the performance of linear regression and neural network-based models. One of the first well known studies was performed by Do and Grudnitski [37]. The authors used a one hidden layer perceptron trained by backpropagation. A very small 6.9% mean absolute error was achieved using sales data of individual houses in San Diego at 1991. The mean absolute error achieved with the linear regression on the same data was equal to 11.26%. Similar results were achieved in other studies by Borst [38], Tay and Ho [11], and Evans et al. [1]. However, Worzala et al. tested the previous studies with similar data and their results were not so promising [12]. For example, while trying to replicate the study of Do and Grudnitsky, for the neural networks they achieved only 10% and 13.1% mean absolute errors using different software packages, while the mean absolute error of linear regression was equal to 11.1%. The conclusion, therefore, was that neural networks must be used very carefully for the real estate valuation. There were many more studies by the other authors: for example, Amabile and Rosato [39], Rossini [32,33], Nguyen and Cripps [31], Ge et al. [21], Wilson et al. [16]. Their results show slight advantage of neural networks against the linear regression. It is obvious that both techniques may show advantage against each other depending on the quality and amount of the data, dependencies between the variables.

The purpose of this study is to explore the usefulness of the most prominent computational intelligence techniques for mass appraisal. The linear regression which is the most popular technique in various studies, a multilayer perceptron (MLP), a support vector machine, and a committee of the models are the techniques used in the investigation. The number of unacceptable valuations is the main parameter of usefulness of a mass appraisal technique. Valuation differing from real value of an object more than certain percentage, 20% in Lithuania [28], is called an unacceptable valuation. There are two main reasons of using the linear regression instead of some higher order regression. Firstly, the linear regression is still used in most mass appraisal systems. Thus, old mass appraisal models can be included into a committee and, therefore, a much easier way of moving from legacy appraisal systems to ones proposed in this work is created. Secondly, Worzala has demonstrated that the linear regression is sometimes more accurate than non-linear computational intelligence techniques [12]. Our results also confirmed this fact. SVM and committees of models have shown excellent performance in the recently studied task of detecting fictious real estate transactions [46–48]. We present a way of obtaining data-dependent weights for aggregating the models into a committee. The weights are based on real estate value zones obtained from the SOM. We demonstrate that committees based on value zones generated by SOM are more accurate than those exploiting the real estate value zones provided by the Register center. The finding means that the time consuming and expensive step of establishing the real estate value zones by an expert can be avoided.

The remainder of the paper is organized as follows. In the next section, a brief description of the techniques used to solve the task is given. The results of the experimental tests are presented in Section 3. The conclusions of the study are given in Section 4.

2. Methods

2.1. Linear regression

The standard regression model is given by:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of dependent variable values with n being the number of observations, $\mathbf{X}$ is a $n \times m$ matrix containing values of independent variables, $\boldsymbol{\beta}$ is a $m \times 1$ vector of regression coefficients, $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of true errors with the standard deviation σ, and m is the number of independent variables. The meaning of $\mathbf{X}$ and $\mathbf{y}$ stated here is kept through entire paper. The estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ is obtained as a solution to:

$$\min_{b} Q_{OLS}(\boldsymbol{b}) \tag{2}$$

where $Q_{OLS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$ in the ordinary least squares (OLS) case, where e_i is the estimation of the true error, and $\hat{y}_i$ is an estimate of y_i.

2.2. Support vector machine for regression

The support vector method can be applied to the case of regression, maintaining all the main features that characterize the maximal margin algorithm developed for classification: a non-linear function is learned by a linear learning machine in a kernel-induced feature space, while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space [20].

The standard form of the support vector regression provided by Vapnik [50] is:

$$\min_{w,b,\xi,\xi^*} = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\xi_i^*,$$

subject to:

$$\begin{gathered} y_i - \boldsymbol{w}^T\phi(\boldsymbol{x}_i) - b \le \varepsilon + \xi_i; \\ -y_i + \boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b \le \varepsilon + \xi_i^*; \;\; \xi_i, \xi_i^* \ge 0, \;\; i = 1, \ldots, n \end{gathered} \tag{3}$$

where $\mathbf{w}$ stands for the weights vector, $\phi(\mathbf{x}_i)$ is an image of $\mathbf{x}$ in the feature space, y_i is the target, n is the number of learning data points, ξ_i, ξ_i^* are the slack variables, which measure the cost of the errors on the training points, b is the threshold, the regularization constant C controls the trade-off between the norm and the losses, and ε is the threshold for ignoring the errors.

The dual form is:

$$\min_{\alpha,\alpha^*} = \frac{1}{2}(\alpha - \alpha^*)^T(\alpha - \alpha^*)\kappa(x_i, x_j) + \varepsilon\sum_{i=1}^{l}(\alpha + \alpha^*) + \sum_{i=1}^{l} y_i(\alpha - \alpha^*),$$

subject to:

$$\sum_{i=1}^{l}(\alpha - \alpha^*) = 0; \quad \alpha_i \ge 0; \;\; \alpha_i^* \le C; \;\; i = 1, \ldots, n \tag{4}$$

where $\kappa(x_i, x_j)$ is the kernel (we used the polynomial kernel in this study) and the parameters α_i^* and α_i are found during the optimization process. Threshold b is computed as follows [20]:

$$b = y_i - \sum_{j=1}^{n} \alpha_i^* \kappa(\boldsymbol{x}_j, \boldsymbol{x}_i) - \varepsilon, \quad \text{for } i \text{ such that } 0 < \alpha_i^* < C. \tag{5}$$

The regression function is:

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\kappa(x_i, x) + b. \tag{6}$$

2.3. Multilayer perceptron

A feed-forward MLP with one hidden layer was employed in the study. Cross-validation data set based early stopping is a usual way to control overfitting. Due to the small data sets available, we avoided using cross-validation data sets. To avoid overfitting, we used Bayesian regularization [7,13], which is implemented by minimizing the following objective function:

$$E = \beta E_D + \alpha E_W = \frac{\beta}{2} \sum_{i=1}^{n} \sum_{j=1}^{Q} \{y_k(\boldsymbol{x}_i; \boldsymbol{w}) - t_j^i\}^2 + \frac{\alpha}{2} \sum_{s=1}^{n_W} (w_s)^2, \tag{7}$$

where $\mathbf{x}_i$ is the input data point, n is the number of the observations, Q is the number of outputs in the network, $\mathbf{w}$ is the weights vector, n_W is the number of weights, and α, β are hyper parameters. The second term of the objective function performs regularization. We used the Levenberg–Marquardt algorithm to optimize the objective function. As in the case of linear regression and SVM, MLP takes independent regression variables x_i as input and predicts the value y.

2.4. Feature selection

It is well known that not all features are useful for solving the task at hand. In the case of limited data sets, some features may even deteriorate the prediction accuracy. Therefore, we applied the feature selection in this study.

There are many feature selection techniques ranging from the sequential forward selection or backward elimination [15,25], to the genetic [41] or tabu search [14]. Usually feature selection methods are categorized as being based on filter or wrapper approaches. Filter methods estimate some measure of saliency for all the features and rank the features on the basis of that measure. The problem is how to find a cut-off point of the ranked list. By contrast, wrapper methods usually evaluate all possible feature combinations by using some learner. Therefore, the approach is computationally prohibitive for large sets of features. In this work we use the filter approach for the feature selection and the predictor output sensitivity to assess the feature saliency. The saliency can be measured as [22,24]:

$$\Psi_i = \frac{1}{QP} \sum_{j=1}^{Q} \sum_{i=1}^{P} \left| \frac{\partial y_{jp}}{\partial x_{ip}} \right|, \tag{8}$$

where y is the output of the predictor, Q is the number of outputs, P is the number of training samples, x_{ip} is the ith feature of the input vector (observation) $\mathbf{x}_p$. In order to find the cut-off point, we compare the saliency of the candidate feature with the saliency score of the noise feature. This technique is employed to select features for the MLP and SVM. All redundant features are eliminated based on the paired t-test comparing the saliency of the candidate and noise features. In the case of OLS regression, variables were selected by testing their t values with the cut-off t values. This approach was adopted because of its use in almost all studies with linear regression [23].

2.5. Self-organizing map

Kohonen's self-organizing feature map (SOM) is such a neural network, wherein the observations are classified so that those, which share related characteristics are located in the same zone of the map (topologic ordering of the data). So it learns the pattern of the input data, it is output commonly is only the index of a winning neuron. The same or neighbour neurons must be activated by the similar input data vectors. A winning neuron, called the best matching unit, is determined by the Euclidian distance between the input vector and the weight vector of the neuron (node). The SOM is trained according to the following rule [45]:

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \alpha(t)\nu(i^*, i, t)[\boldsymbol{x} - \boldsymbol{w}_i(t)], \tag{9}$$

where $\mathbf{w}_i(t)$ is the weight vector of the ith SOM node at the time step t, $\alpha(t)$ is the decaying learning rate, $\nu(i^*,i,t)$ is the decaying Gaussian neighborhood function, and i^* stands for the index of the winning node.

Due to the famous topologic ordering property, SOM is widely used for data visualization and clustering. In this work we use SOM for obtaining data-dependent aggregation weights to build a data-dependent weighted averaging committee of aforementioned predictors. Next section explains the way the aggregation weights are obtained.

2.6. Committee

It is well known that a committee of predictors can improve the prediction accuracy. A variety of schemes have been proposed for combining predictors. The approaches used most often include averaging [29,36], weighted averaging [2,6,10,34,44], the fuzzy integral [3,35,42,43], probabilistic aggregation [18], and aggregation by a neural network or SVM [5,8,42]. The aggregation weights assigned to committee members can be the same in the entire data space or can be different – data dependent – in various regions of the space [2,4,34,49]. The use of data-dependent weights, when properly estimated, provides higher estimation accuracy [4]. See Verikas et al. [4] for a comparative study of different combination schemes.

In this study, three members, namely the OLS regression model, SVM and MLP comprise a committee. Different architectures chosen for different committee members is one of possible way to increase the diversity of the committee members. It is well known that an efficient committee should consist of members that are not only very accurate, but also diverse.

The data-depended weights-based weighted averaging has been used to aggregate decisions obtained from the members. The aggregation weights were specific for each predictor and each value zone (cluster). The value zones used were obtained from the Register center or automatically using SOM. The variables used to train the SOM are those regression variables x_i found being significant by at least two committee members. These variables are: the size of the house, the year of the construction, the type of the heating system, the type of the outside finishing, the value of the lot, and the percentage of the house completion.

The data-dependent aggregation weights are given by the reciprocal of entries of the decision matrix **DSC**:

$$\boldsymbol{DSC}_{r,j} = \frac{1}{N_{obsr}} \sum_{i=1}^{N_{obsr}} \boldsymbol{ACM}_{i,j}, \quad r = 1, \ldots, n_r, \;\; j = 1, \ldots, n_j, \tag{10}$$

where n_r is the number of clusters, n_j is the number of predictors, N_{obsr} is the number of observations in the rth cluster, and the matrix

ACM is given by:

$$\boldsymbol{ACM}_{i,j} = \frac{|y_i - (|y_i - \hat{y}_{i,j}|)|}{\sum_{j=1}^{n_j} |y_i - (|y_i - \hat{y}_{i,j}|)|}, \tag{11}$$

where y_i stands for the target value and $\hat{y}_{i,j}$ is the predicted value of the jth predictor given the ith observation. Thus, the matrix **ACM** expresses the accuracy of each predictor in the vicinity of each observation. The matrix **DSC** expresses the average accuracy of each predictor in each cluster. Thus, the aggregation weights are data-dependent.

2.7. Assessment of models

Care must be taken of OLS regression violations when using linear models. Correlation coefficient, Durbin–Vatson coefficient, and the variance inflation factor are the parameters usually used to assess ordinary least squares linear regression models.

The correlation coefficient r_{xy} indicates the strength and direction of a linear relationship between the variables x and y [19]:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}, \tag{12}$$

where $\bar{x}$ and $\bar{y}$ are means of the variables, s_x and s_y are the standard deviations, and n is the number of observations.

Durbin–Watson (D–W) coefficient is used to investigate autocorrelation [19]. The coefficient is computed as:

$$d = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \tag{13}$$

where e is the residual. The parameter d is compared with the upper and lower limits of autocorrelation. Values of the limits depend on the significance level, the number of observations, and the number of independent variables. These values are usually taken from special tables.

The variance inflation factor (VIF) is used to assess multi-colinearity [19]:

$$VIF = \frac{1}{1 - R_j^2}, \tag{14}$$

where R_j^2 is the determination coefficient from the regression of x_j on the other independent variables. The value greater than 10 is an indication of potential multi-colinearity problems [19].

The quality of non-linear models, such as MLP and SVM, is usually assessed by estimating the generalization performance (the prediction error for unseen data) of the models. The mean absolute percentage difference between the real y and predicted $\hat{y}$ ($MAPD$), the mean absolute error between the real and predicted $\hat{y}$ (MAE), and the number of unacceptable valuations (UV) are the model quality (prediction accuracy) measures used to asses all the models in this work. As it has already been mentioned, UV is equal to the number of predicted $\hat{y}$ values differing from the real y values more than 20%.

3. Experimental investigations

3.1. Data

The data set used in the study contains sale transactions of houses in one city, dated from 2005 till 2006. Parameters of the sample are as follows: the sample size is 100, the number of available variables is 12, the average price per square meter is equal to 2368.36 Lt (686.48€), the median of the price is 2340.41 Lt, the standard deviation is 688.40 Lt, the minimum price is 179,000 Lt, the maximum price is 875,000 Lt, the average size of the house is 195.90 m^2, the median of the size is 190 m^2, the standard deviation is 47.38 m^2, the minimum and the maximum sizes are 92 m^2 and 384.68 m^2, respectively. Table 1 summarizes the variables of the sample used in this study.

Table 1
Available parameters.

Available parameters	Notation
Place of the house	x_1
Size of the house	x_2
Year of construction	x_3
Canalization type	x_4
Type of the heating system	x_5
Type of the house construction (material)	x_6
Type of the outside finishing	x_7
Number of floors	x_8
Type of the house (separate or blocked)	x_9
Size of the lot	x_{10}
Value of the lot	x_{11}
Percentage of the house completion	x_{12}

Recent studies show that there are many real estate transactions recorded with a fictive price in Lithuania [46]. The data used was checked by the human expert, who did not find any outliers in this data set.

3.2. Results

All experiments were implemented using MATLAB. Feature selection was performed as described in Section 2.4. The leave-one-out approach has been used in the experiments to assess the quality of the models, due to the small data set. In the leave-one-out technique, the model is trained n (the number of observations) times using all the data points for training, except one, which is used for testing. The data point kept aside for testing is different in each run. The model accuracy is estimated as the average accuracy obtained in these n runs.

3.2.1. Linear regression

Values of the linear regression equation parameters, estimated by the ordinary least squares using all the data, are:

$$\begin{aligned}\hat{y} = {} & -6916.16x_1 - 1643.80x_2 - 79.18x_3 - 220902.75x_4 \\ & -4004.69x_5 + 160681x_6 + 8153.31x_7 - 2217.86x_8 \\ & -14444.76x_9 - 7030.47x_{10} + 0.15x_{11} + 6059.69x_{12},\end{aligned} \tag{15}$$

where $\hat{y}$ is the predicted price of the house.

The Student's t values are: $t_1 = -1.81$, $t_2 = 7.25$, $t_3 = -0.74$, $t_4 = -2.14$, $t_5 = -0.36$, $t_6 = 2.17$, $t_7 = 0.67$, $t_8 = -0.07$, $t_9 = -0.5$, $t_{10} = -2.78$, $t_{11} = 0.69$, and $t_{12} = 10.24$. The cut-off t value is: $t_{1-\lambda/2}(100-12) = \pm 1.987$, $\lambda = 0.95$. Parameters, the t values where of which do not exceed the cut-off value, must be removed. Therefore, the new equation is given by:

$$\begin{aligned}\hat{y} = {} & 1536.64x_2 - 332888.62x_4 - 144819.92x_6 \\ & -6429.9x_{10} + 6174.71x_{12}.\end{aligned} \tag{16}$$

Table 2
The results of mass appraisal of the real estate.

Predictor	$MAPD$ (%)	MAE (Lt)	UV
OLS regression	15.02	65260.24	31
MLP	23.30	90255.57	42
SVR	13.62	59055.77	18
Committee 1	13.61	56450.59	1
Committee 2	13.36	56030.86	1
Register center models	29.67	135592.53	71

The new Student's t values are: $t_2 = 7.48$, $t_4 = -4.58$, $t_6 = 2.66$, $t_{10} = -4.19$, and $t_{12} = 11.38$ and the new cut-off value is $t_{1-\lambda/2}(100-5) = \pm 1.985$. Thus, all the remaining parameters exceed this value.

The matrix of the estimated correlation coefficients $\mathbf{R}$ is:

$$\boldsymbol{R} = \begin{bmatrix} y & x_2 & x_4 & x_6 & x_{10} & x_{12} \\ 1 & 0.22 & 0.13 & 0.21 & 0.07 & 0.56 \\ 0.22 & 1 & -0.06 & 0.14 & 0.30 & -0.33 \\ 0.13 & -0.06 & 1 & 0.007 & -0.07 & 0.32 \\ 0.21 & 0.14 & 0.007 & 1 & 0.07 & -0.04 \\ 0.07 & 0.30 & -0.07 & 0.07 & 1 & 0.20 \\ 0.56 & -0.33 & 0.32 & -0.04 & 0.20 & 1 \end{bmatrix} \qquad (17)$$

As it can be seen, the correlation coefficient values between the independent variables are small. The highest value equal to −0.33 is between the size of the house and the percentage of the house implementation (opposite correlation), the correlation coefficient value equal to 0.32 is between the type of canalization and the percentage of the house implementation, and 0.30 between the size of the house and the size of the parcel.

The value of the Durbin–Watson coefficient is 1.54 and the bounds at the 95% significance level are $d_{L,0.25} = 1.53$, $d_{U,0.25} = 1.69$. Thus, there is no negative autocorrelation between observations and there is no evidence that there is positive autocorrelation. This can be explained by a short period of sale transactions. The variance inflation factors are: $VIF_1 = 1.36$, $VIF_2 = 1.00$, $VIF_3 = 1.00$, $VIF_4 = 1.28$, and $VIF_5 = 1.46$, so there is no multi-collinearity.

The mean absolute percentage difference (*MAPD*) between the real and predicted $\hat{y}$ values is 15.02%, the mean absolute error (*MAE*) between the real and predicted $\hat{y}$ values is 65260.24 Lt. The number of unacceptable valuations (*UV*) is 31, which is the most important parameter. The test results are summarized in Table 2.

3.2.2. Multilayer perceptron

The Matlab neural networks toolbox has been used to simulate both MLP and SOM. To train MLP, the *trainbr* function implementing Levenberg–Marquardt optimization with Bayesian regularization was employed. MLP with 13 input nodes, one hidden layer with log-sigmoid transfer functions and one linear node in the output layer was used at this step. Targets and inputs were auto-scaled to fall into the range [−3;3]. There was no need for robust scaling, because there were no outliers. The applied feature selection procedure has shown that the saliency of only 6 variables exceeded the saliency of the noise variable: place of the house, size of house, type of the heating system, level of the outside finishing, value of the parcel, and percentage of the house completion. As it can be seen, there are only two variables, x_2 and x_{12} common with those found in the linear regression case. The number of hidden nodes has been found experimentally. The best results were obtained with 7 nodes in the hidden layer. The mean absolute percentage difference (*MAPD*) between the real and predicted $\hat{y}$ values was 23.30%, the mean absolute error (MAE) was 90255.57 Lt, and the number of unacceptable valuations (*UV*) was 42. Thus, by all the parameters, the MLP performed worse than the linear regression.

3.2.3. Support vector machine

The SVM has been implemented using the software written for MATLAB and available at http://www.csie.ntu.edu.tw/~cjlin/libsvm [9]. The data auto-scaling was done in the same way as in the MLP case. Almost the same variables were significant, except x_3 (year of construction), which proved to be significant. The best results were achieved with the cubic polynomial kernel with gamma value $\gamma = 0.6$, and the value of $C = 1000$. The mean absolute percentage difference (*MAPD*) between the real and predicted $\hat{y}$ values was 13.62%, the mean absolute error (*MAE*) was 59055.77 Lt, and the number of unacceptable valuations (*UV*) was 18. As it can be seen, the results obtained from the SVM are clearly better than those obtained from the linear regression and MLP. The results indicate that non-linear modeling is required. MLP performs non-linear modeling. However, MLP training probably terminates in a local minimum, while in the case of SVM a global minimum is reached.

3.2.4. Committee

As expected, the results obtained from the committee were better than those provided by the separate predictors. Two types of committees were studied: *Committee 1* using the value zones provided by experts of the Register center and *Committee 2* exploiting the value zones obtained automatically from SOM. The following results were obtained from the *Committee 1*: the mean absolute percentage difference (*MAPD*) between the real and predicted $\hat{y}$ values is 13.61%, the mean absolute error (*MAE*) is 56450.59 Lt, and the number of unacceptable valuations (*UV*) is only 1. We remind that *UV* is the most important parameter in the mass appraisal of the real estate.

No information from the Register center has been used when obtaining the value zones (clusters) by the SOM. The results obtained from *Committee 2*, exploiting the SOM created zones, are slightly better than the results obtained from *Committee 1* based on the Register center's value zones. The mean absolute percentage difference (*MAPD*) between the real and predicted $\hat{y}$ values is 13.36%, the mean absolute error (*MAE*) is 56030.86 Lt, and the number of unacceptable valuations (*UV*) is 1 (see Table 2). The numbers of used value zones was the same in both cases. However, the assignment of data to the zones made by the Register center and SOM was quite different. Many of members of the Register center clusters were split to 3–4 different clusters by SOM. For example, 30% of members of the first Register center value zone were assigned to one cluster by SOM, 40% to another cluster, 30% to the third cluster. Observe that all the houses were sold in the eyre of one city, therefore the place of the real estate is not as important as it is in the heterogenous areas.

3.2.5. Register center models

The same data was used to test the performance of the official real estate valuation models, prepared by the Register center. All these models can be obtained via Internet at http://www.registrucentras.lt/masvert/. The results obtained from the models are as follows. The mean absolute percentage difference (*MAPD*) between the real and predicted $\hat{y}$ values is 29.67%, the mean absolute error (*MAE*) is 135592.53 Lt, and the number of unacceptable valuations (*UV*) is 71. All the test results are summarized in Table 2. The superiority of the computational intelligence-based techniques should be obvious.

4. Conclusions

Ordinary (OLS regression) and computational intelligence-based techniques (MLP, SVM, and a committee comprised of all the three types of models) have been evaluated in the mass appraisal of real estate task. Many previous authors reported diverse results when comparing the MLP and OLS regression based approaches. Our results also indicated the superiority of OLS regression over the MLP. However, the SVM clearly outperformed both the OLS regression and MLP based models. The results indicate that non-linear modeling is required. SVM as being capable of non-linear modeling and finding the global minimum of the cost function suits very well for the task.

The proposed committee of models has shown an excellent performance and clearly outperformed the separate predictors. The number of unacceptable valuations, which is the main parameter in the mass appraisal tasks, was only 1. It means that only 1% of valuations do not satisfy the accuracy limits for the mass appraisal. A new approach to obtaining data-dependent aggregation weights to build a weighted averaging committee of models has been presented in this work. Two ways of obtaining the weights were explored, namely the weights based on value zones obtained from experts of the Register center and the weights based on clusters (zones) provided by SOM without using any information from the experts. Clusters obtained from the Register center and SOM were quite different. This can be explained by the homogeneity of the whole area. Thus, the impact of house place was not so important in comparison to its main characteristics. The committee results obtained using SOM clusters were slightly better than those based on the Register center value zones. Therefore, SOM can be used for effective clustering of the real estate data.

The much better performance obtained from the computational intelligence-based techniques than from the official models shows the great practical impact of this study.

References

[1] A. Evans, H. James, A. Collins, Artificial neural networks: an application to residential valuation in the UK, Journal of Property Valuation & Investment 11 (1992) 195–203.

[2] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, MIT Press, 1995, pp. 231–238.

[3] A. Verikas, A. Lipnickas, Fusing neural networks through space partitioning and fuzzy integration, Neural Processing Letters 16 (2002) 53–65.

[4] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, A. Gelzinis, Soft combination of neural classifiers: a comparative study, Pattern Recognition Letters 20 (1999) 429–444.

[5] A. Verikas, A. Lipnickas, K. Malmqvist, Selecting neural networks for a committee decision, International Journal of Neural Systems 12 (2002) 351–361.

[6] A. Verikas, M. Signahl, K. Malmqvist, M. Bacauskiene, Fuzzy committee of experts for segmentation of colour images, in: Proceedings of 5th European Congress on Intelligent Techniques and Soft Computing, vol. 3, Aachen, Germany, 1997, pp. 1902–1906.

[7] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.

[8] C.L. Liu, Classifier combination based on confidence transformation, Pattern Recognition 38 (2005) 11–28.

[9] Ch.Ch. Chang, Ch.J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] C.J. Merz, M.J. Pazzani, Combining neural network regression estimates with regularised linear weights, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, MIT Press, 1997, pp. 564–570.

[11] D.P.H. Tay, D.K.K. Ho, Artificial intelligence and the mass appraisal of residential apartments, Journal of Property Valuation & Investment 10 (1991) 525–540.

[12] E. Worzala, M. Lenk, A. Silva, An exploration of neural networks and its application to real estate valuation, Journal of Real Estate Research 10 (2) (1995).

[13] F.D. Foresee, M.T. Hagan, Gauss-Newton approximation to Bayesian learning, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 1997, pp. 1930–1935.

[14] H. Zhang, G. Sun, Feature selection using tabu search method, Pattern Recognition 35 (2002) 701–711.

[15] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.

[16] I. Wilson, S. Paris, A. Ware, D. Jenkins, Residential price forecasting at national and regional levels, in: FIG XXII International Congress in Washington, 2002.

[17] International Valuation Standards Seventh edition, Appraisal Inst, 2005.

[18] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans Pattern Analysis and Machine Intelligence 20 (1998) 226–239.

[19] J. Neter, M. Kutner, W. Wasserman, Christopher Nachtsheim, Applied Linear Statistical Models, McGraw-Hill, Irwin, 1996.

[20] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK, 2004.

[21] J.X. Ge, G. Runeson, K.C. Lam, Forecasting Hong Kong housing prices: an artificial neural network approach, in: Methodologies in Housing Research Conference, 2003.

[22] J.M. Steppe, K.W. Bauer, Improved feature screening in feedforward neural networks, Neurocomputing 13 (1996) 47–58.

[23] J.O. Rawlings, G.S. Pantula, D.A. Dickey, Applied Regression Analysis: A Research Tool, Springer-Verlag, New York, 1998.

[24] K.L. Priddy, S.K. Rogers, D.W. Ruck, G.L. Tarr, M. Kabrisky, Bayesian selection of important features for feedforward neural networks, Neurocomputing 5 (1993) 91–103.

[25] K.Z. Mao, Orthogonal forward selection and backward elimination algorithms for feature subset selection, IEEE Transactions on Systems, Man, & Cybernetics Part B: Cybernetics 34 (2004) 629–634.

[26] Lithuania: Resolution of the Lithuanian government on the confirmation of the real estate valuation rules. Valstybes Zinios 117 (2005) 2005-09-29 (in Lithuanian).

[27] Lithuania: The law amending the 8 Subsection of the law on valuation basics of property and bussinnes No. IX-1428. Valstybes Zinios 38 (2003) 2003-04-03 (in Lithuanian).

[28] Lithuania: The law on valuation basics of property and bussinness No. VIII-1202. Valstybes Zinios 52 (1999) 1999-05-25 (in Lithuanian).

[29] M. Taniguchi, V. Tresp, Averaging regularized estimators, Neural Computation 9 (1997) 1163–1178.

[30] M.R. Linne, S.M. Kane, G. Dell, A Guide to Appraisal Valuation Modeling, 2000.

[31] N. Nguyen, A. Cripps, Predicting housing value: a comparison of multiple regression analysis and artificial neural networks, Journal of Real Estate Research 22 (3) (2001).

[32] P. Rossini, Improving the results of artificial neural network models for residential valuation, in: Proceedings of Fourth annual Pacific Rim Real Estate Society Conference, 1998.

[33] P. Rossini, P. Kershaw, Using neural networks to estimate constant quality house price indices, in: Proceedings of Fifth annual Pacific Rim Real Estate Society Conference, 1999.

[34] S. Sollich, A. Krogh, Learning with ensembles: how over-fitting can be useful, in: D.S. Touretzky, M.C. Mozer, H.E. Hasselmo (Eds.), Advances in Neural Information Processing Systems, vol. 8, MIT Press, 1996, pp. 190–197.

[35] P.D. Gader, M.A. Mohamed, J.M. Keller, Fusion of handwritten word classifiers, Pattern Recognition Letters 17 (1996) 577–584.

[36] P.W. Munro, B. Parmanto, Competition among networks improves committee performance, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, MIT Press, 1997, pp. 592–598.

[37] Q. Do, D. Grudnitski, A neural network approach to residential property appraisal, Real Estate Appraiser (1992) 38–45.

[38] R.A. Borst, Artificial Neural Networks: The Next Modeling/Calibration Technology for the Assessment Community, Artificial Neural Networks 10 (1992) 69–94.

[39] R. Amabile, P. Rosato, The use of neural networks in the spatial analysis of property values, in: Proceedings of the Sixth Joint Conference on Agriculture, Food, and the Environment, Minneapolis, Minnesota, August 31–September 2, 1998.

[40] R.J. Gloudemans, Mass appraisal of real property, 1999.

[41] S. Yu, S.G. Backer, P. Scheunders, Genetic feature selection combined with composite fuzzy nearest neighbor classiers for hyperspectral satellite imagery, Pattern Recognition Letters 23 (2002) 183–190.

[42] S.P. Kim, J.C. Sanchez, D. Erdogmus, Y.N. Rao, J. Wessberg, J.C. Principe, M. Nicolelis, Divide-and-conquer approach for brain machine interfaces: nonlinear mixture of competitive linear models, Neural Networks 16 (2003) 865–871.

[43] S.B. Cho, J.H. Kim, Combining multiple neural networks by fuzzy integral for robust classification, IEEE Transactions on Systems Man and Cybernetics 25 (1995) 380–384.

[44] T. Heskes, Balancing between bagging and bumping, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, MIT Press, 1997, pp. 466–472.

[45] T. Kohonen, Self-organizing Maps, 3rd ed., Springer-Verlag, Berlin, 2001.

[46] V. Kontrimas, A. Verikas, Neural networks based screening of real estate transactions, Neural Network World 16 (1) (2007) 17–30.

[47] V. Kontrimas, A. Verikas, Tracking of doubtful real estate transactions by outlier detection methods: a comparative study, Information Technology and Control 35 (2) (2006) 94–105.

[48] V. Kontrimas, An application of hybrid computational methods to real estate markets, in: Proceedings of Information technologies 2005, Technologija, Kaunas, 2005, pp. 141–144.

[49] V. Tresp, M. Taniguchi, Combining estimators using non-constant weighting functions, in: G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, MIT Press, 1995.

[50] V. Vapnik, The Nature of Statistical Learning, Springer, New York, 1995.