Exploration of Bagging Ensembles Comprising Genetic Fuzzy Models to Assist with Real Estate Appraisals

Tadeusz Lasota[1], Zbigniew Telec[2], Bogdan Trawiński[2], and Krzysztof Trawiński[3]

[1] Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management
Ul. Norwida 25/27, 50-375 Wroclaw, Poland
tadeusz.lasota@wp.pl
[2] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl
[3] European Centre for Soft Computing, Edificio Científico-Tecnológico, 3ª Planta,
C. Gonzalo Gutiérrez Quirós S/N, 33600 Mieres, Asturias, Spain
krzysztof.trawinski@softcomputing.es

Abstract. The study reported was devoted to investigate to what extent bagging approach could lead to the improvement of the accuracy machine learning regression models. Four algorithms implemented in the KEEL tool, including two evolutionary fuzzy systems, decision trees for regression, and neural network, were used in the experiments. The results showed that some bagging ensembles ensured higher prediction accuracy than single models.

Keywords: bagging ensembles, genetic fuzzy systems, regression models, real estate appraisal, KEEL.

1 Introduction

Ensemble learning systems combine the output of machine learning algorithms, called "weak learners", in order to get smaller prediction errors (in regression) or lower error rates (in classification). The individual estimator must provide different patterns of generalization, thus in the training process diversity is employed. Otherwise, the ensemble would be composed of the same predictors and would provide as good accuracy as the single one. It has been proved that the ensemble performs better when each individual machine learning system is accurate and makes errors on different examples [2], [10]. Although, there are many taxonomies for that, there is one recognized group the so-called data resampling, which generates different training sets to obtain unique regressor or classificator. To this group we may include bagging [3], boosting [18], and stacking [21].

The most popular method - bagging, which stands for bootstrap aggregating, is one of the most intuitive and simplest ensemble algorithms providing a good performance. Diversity of regressors is obtained by using bootstrapped replicas of the training data. That is, different training data subsets are randomly drawn with replacement from the original training set. So obtained training data subsets, called also bags, are used then to train different regression models. Finally, individual regressors are combined

E. Corchado and H. Yin (Eds.): IDEAL 2009, LNCS 5788, pp. 554–561, 2009.

through an algebraic expression, such as minimum, maximum, sum, mean, product, median, etc.[16].

Although the ensemble learning systems are more common in classification, they are also applied to regression problems. Bagging has been employed among others to regression trees [4], Gaussian process [7], SVM [22], and neural networks [9]. You can find also some works on applying ensemble fuzzy systems to solve classification [5], [12] and prediction problems [11], but they are not too numerous.

In our previous works [13], [14] we tested different machine learning algorithms, among others genetic fuzzy systems trying to select the most appropriate ones to build data driven models for real estate appraisal using MATLAB and KEEL. Since ensemble learning can be used to improve the performance of a model and reduce the likelihood of an unfortunate selection of a poor one, in this paper we focused our efforts into investigating to what extent the bagging approach could lead to the improvement of the accuracy machine learning regression models devoted to assist with real estate appraisals.

The concept of a data driven models for premises valuation, presented in the paper, was developed based on the sales comparison method. The architecture of the proposed system is shown in Fig. 1. The appraiser accesses the system through the internet and input the values of the attributes of the premises being evaluated into the system, which calculates the output using a given model. The final result, that is a suggested value of the property, is sent back to the appraiser.

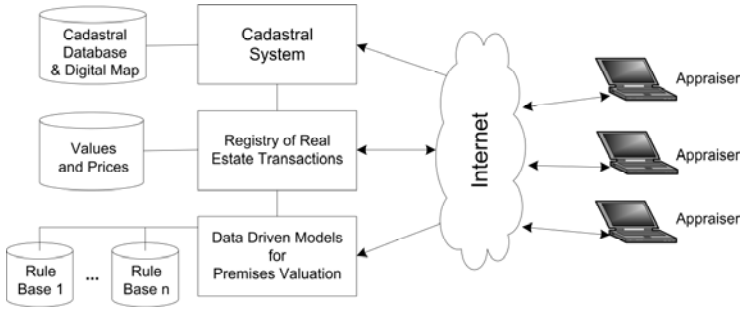

Fig. 1. Information systems to assist with real estate appraisals

Actual data used to generate and learn appraisal models came from the cadastral system and the registry of real estate transactions referring to residential premises sold in one of big Polish cities at market prices within two years 2001 and 2002. They constituted original data set of 1098 cases of sales/purchase transactions. Four attributes were pointed out as price drivers: usable area of premises, floor on which premises were located, year of building construction, number of storeys in the building, in turn, price of premises was the output variable.