Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 3852 – 3859

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

An ensemble learning based model for real estate project classification

Worapat Paireekreng[a,]*, Worawat Choensawat[b]

[a]*Dhurakij Pundit University, Bangkok, Thailand*
[b]*Bangkok University, Bangkok, Thailand*

Abstract

The demand of accommodation has been increasing especially in the metro area. The process of credit assessment for loan is increasingly high. One of the criteria for banks to determine the amount of loan is a grade of the real estate project. This paper addresses the problem of determining the grade of a real estate project and proposes a real estate classification model. The model helps loaners to make a decision for the further stage of the loan. For the model construction, the data were gathered from 407 real estate projects in Thailand which are launched in 2014 and forthcoming projects that will be launched in 2015. The variables of our model involve the project infrastructure and characteristics such as the facilities, number of units, location, parking space, and size of a developer. For all 407 projects, the grades were provided by a bank. We use the supervised learning with ensemble technique and vote algorithm for training and testing against the dataset where the dataset is separated into the training set and the testing set of 307 records and 100 records, respectively. The efficiency of our proposed model is measured by classification accuracy and user satisfaction with the model.

Peer-review under responsibility of AHFE Conference

Keywords: Classification; Ensemble; Real estate

* Corresponding author. Tel.: +66-8-1431-1414; fax: +66-2-954-7300.
E-mail address: worapat.png@dpu.ac.th

Peer-review under responsibility of AHFE Conference
doi:10.1016/j.promfg.2015.07.892

1. Introduction

One of the most important basic needs for human is an accommodation. Almost all people need to acquire the accommodation for their living such as house, apartment, commercial building and condominium. There are many social changes which influence the people's behavior in terms of acquiring accommodation. For example, more young people tend to search for accommodation in larger metro because of the job opportunities and urban life. Some of these young people are the first-time home buyer. On the other hand, many people are seeking for a real estate investment. Despite of the reasons in buying the real estate property, in the decision-making process for property products, information related to the property are needed especially the maximum loan amount of the considering real estate.

Most of the people who need to buy the accommodation might consider the mortgage amount and the monthly payments. However, some information cannot be determined by loaner themselves. Banks which provide the loan service have criteria to consider the amount of loan. In addition, different bank has different criteria in terms of loan consideration. Not only a loaner factor but the real estate developer is also the important criteria. Although the customer has a potential for the loan but some real estate developers may be determined the grade of the real estate project differently. Some projects may be graded low because the performance of the company which reflects to loan amount to customer. For example if a loaner with good credit may get 100% loan for the real estate price of grade "A" and 70% for grade "C". Although, a loaner can apply for a pre-approval of the loan amount, the process shall take 1-2 weeks and require an additional fee.

Hence, determining the grade of the real estate project is important. However, building the classification to predict real estate grade seems to be complicated. Some information are provided by banks but as mentioned above, different bank have a different guidelines for determining the real estate grade. Moreover, there is a variety of characteristics of the real estate project such as the number of units in the project, facilities, or the performance of the real estate developer in the past years.

This paper addresses the problem of determining the grade of a real estate project and proposes a real estate classification model. The model helps people to make a decision for the further stage of loan. The structure of this paper is as follows: Section 2 presents a literature review of real estate classification and techniques, while Section 3 shows the research methodology. Section 4 presents the experimental results for the proposed methodology. Finally, Section 5 draws some conclusions.

2. Related works

2.1. Real estate classification

The real estate industry has been the main factor to drive economic in almost all countries. However, there are several perspectives for an appraisal of real estate. The criteria used in each appraisal and perspective are different. The main approach can be divided into government appraisal and consumer appraisal. Firstly, the government appraisal focuses on the price of real estate based on areas or location including demographic and economic factors such as specific industrial area or commercial area [1]. Secondly, the appraisal is based on consumers' needs. The second appraisal can be viewed as an accommodation appraisal. Consumers need to know the information related to the accommodation that they are buying. However, the mass appraisal model for the real estate commonly based on the sales and other criteria related to commercial [2]. In addition, most of the consumers prefer to know the information related to the grade and ranking of the accommodation which is the primary assessment of the loan. Therefore, the real estate classification seems to be important towards consumers' needs. This is based on available that the consumers can gather information about the individual real estate project. The criteria that consumers use for decision-making is such as infrastructure and facilities.

2.2. Classification problems

The problem context of the classification can be interpreted by data mining techniques. The wide range of techniques can be used. Wu et al. [3] have shown that some commonly used algorithms in data mining such as k-means, SVM, Apriori, and PageRank including Naïve Bayes. Classification techniques also have been incorporated into real estate prediction. Other research by Cufoglu et al. [4] and Nurmi and Hessinen [5] has proposed which classifier is the most appropriate for classification problems. The potential of these kinds of classification techniques can be seen from the recommendation system for mobile content usage [6].

The measurement to evaluate the classification performance of the built model is accuracy rate (1). It has been used for an empirical measure. The measurement shows the numbers of correctly classified instances of a different class in the data sets. This is also applied for binary class and multiclass classification problem.

$$ACC = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

The attempt to increase the speed of computational time for model building can be seen from pre-processing stage. The feature selection has been used for the objective of selecting the most relevant attributes of the data set as input variables. The variables impact on the performance in terms of accuracy rate and prediction results of classification model. The pre-processing stage includes elimination of irrelevant attributes in order to increase the computational time of the model building stage. Feature selection has been applied in data mining, machine learning, and pattern recognition [7]. In addition, the input variables are the challenge of this area by selecting the minimum subset of attribute with little or no loss of classification performance [8][9].

Feature selection can be divided into two categories: model-free method and model-based methods. Model-free methods are based on statistical tests, properties of function and available data such as linear regression, whereas, model-based methods, such as neural network develop model, are to find the significant features and minimize the model output error [10].

2.3. Real estate classification model

The real estate appraisal model mostly focused on mass appraisal [1], [2]. The linear regression technique seems to be the most popular for mass appraisal due to marketing forecasting aspect and sales. However, the other classification techniques can be applied for real estate appraisal. Those techniques are such as artificial neural networks, support vector machine or k-nearest neighbor. The techniques to be used in the model building for real estate classification and forecasting should be used carefully because each technique has some advantages and drawbacks which is suitable for specific situation and data set. In addition, the characteristics of the data set may be varied from attributes to attributes. The attribute characteristics also affect the use of classification techniques. It can be seen that the quality of the classification models depends on the amount of data and relationship among the attributes [11].

3. Research methodology

3.1. Experimental design

The data for the experiment were gathered from two sources. The first source of data is from a commercial bank in Thailand. The data were related to real estate project name, facilities, developer profile and ranking by the bank. In addition, the bank rated the real estate project by their criteria which were unable to disclose due to trade secret. The other source of collected data is from a survey. The incomplete data for each project was filled. Field study, interview and secondary source from brochure and real estate project handbook were applied to complete the data. It combined with primary from the bank to consolidate the matrix of real estate project data. Furthermore, the bank rated the grade of the real estate project using 4 categories which are AA, A, B and C grade.

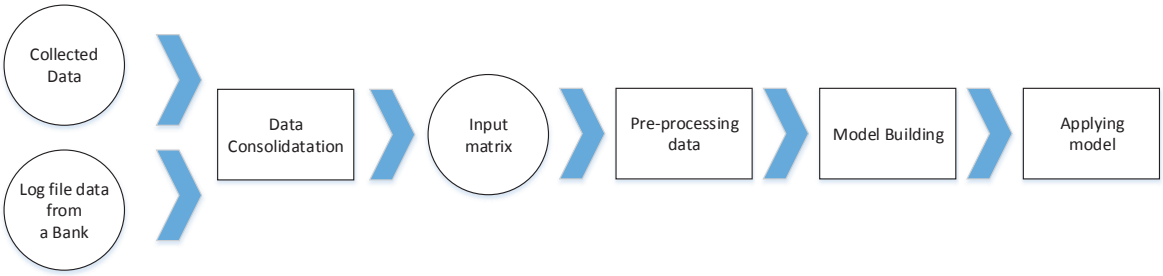

Fig. 1. The proposed framework for ensemble learning based model for real estate project classification.

The sample size was calculated from Taro Yamane's sample size statistics. Then, 407 real estate projects, which are launched in 2014 and will be launched in 2015 in Thailand, were selected. The variables of input matrix involve two main categories. Firstly, developer profile is related to number of year experience in the real estate sector, listed in stock market and size of the developer. Secondly, it concerns project infrastructure and characteristics such as the facilities, swimming pool, fitness center, park or garden, number of units, location, and parking space.

3.2. The proposed ensemble learning based model

The process of the proposed ensemble learning based model for real estate project classification can be seen from Figure 1. The data source combined collected data and logged file data from the bank which the data were related to real estate projects. After that, the data were consolidated and the input matrix was derived. Next, the pre-processing process was run to find missing values and replace those values. The pre-processing process includes the feature analysis for each attribute. After the pre-processing process, the model was built using the classification techniques with appropriate attributes. We also chose the appropriate classifier for real estate project classification. Then, the selected model was applied to the testing data.

The number of the real estate project participated in the experiment was 407. We combined data from two different sources in the data collection stage. The next stage concern pre-processing which the feature selection was applied. The feature selection process was incorporated in reducing the number of attribute for the model building stage. The feature selection process includes finding the weight of attributes that affects the model building. The linear regression technique was used in the feature selection. After that, the collected data implemented data normalization using a range transformation method. The data normalization is to ensure suitability for each comparison technique. The normalized dataset from the previous stage was separated into a training dataset and a testing dataset. 307 randomly selected records were used as the training data. The remaining 100 records were used as the testing data. The training process in the model building stage used 10-fold cross-validation to train the data. Each classification techniques applied in this research was Decision Tree, Naïve Bayes, Neural Networks and Support Vector Machine. The results from the classification used in the experiment were used in the next process to enhance the performance of grade prediction of the real estate project. Vote algorithm was used in the ensemble technique for classification model building. Next, the model was applied to the testing data in order to find the results of unseen data. Then, the performance of this stage was analyzed. Finally, the output of the predicted real estate project grade was derived. Figure 2 demonstrates the proposed methodology for model building and applying the ensemble learning based model for real estate project classification. Table 1 shows the information related to data set and number of instance in each real estate grade.

Table 1. Data set information.

Grade	Instances
AA	107
A	100
B	100
C	100
Total	407

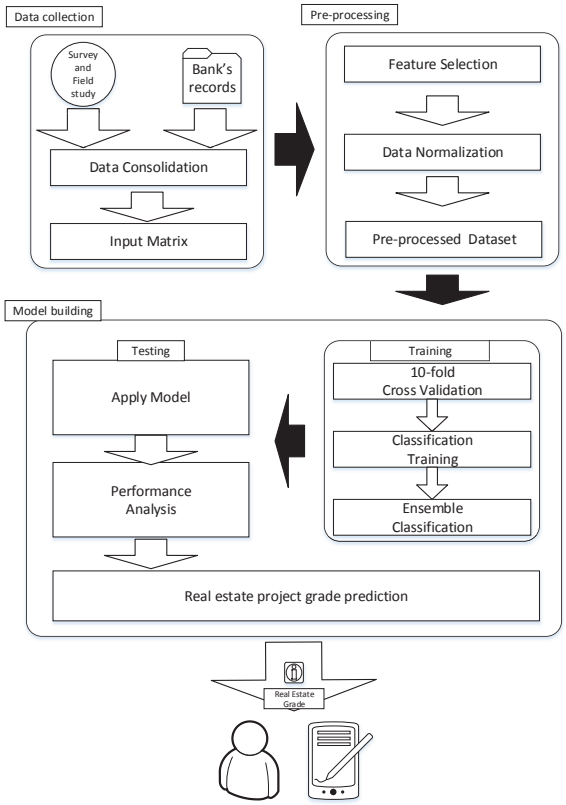

Fig. 2. Explanation of the proposed methodology for model building and applying the ensemble learning based model for real estate project classification.

4. Research methodology

4.1. The feature selection of real estate data

For the experimental stage, the input matrix was prepared in the data collection stage ready for feature selection. The feature selection in the experiment was performed by an evolutionary approach with neural networks. The weights of the attributes are calculated using a Genetic Algorithm. The result can be seen from Table 2. The attributes were reduced from 14 attributes to 11 attributes excluding a label attribute. Therefore, a subset of new input variables were derived and would supply to the next stage for model building. The attributes in the data set were selected based on the feature selection results shown above.

Table 2. Coefficient of attributes on feature selection.

Attribute	Coefficient
Swimming Pool (Salt)	1.000
Swimming Pool (Chlorine)	0.966
Fitness	0.888
Experience	0.784
Wireless	0.706
Sauna	0.570
Park	0.396
Unit	0.394
Listed in Stock Market	0.267
Security	0.170
CCTV	0.120

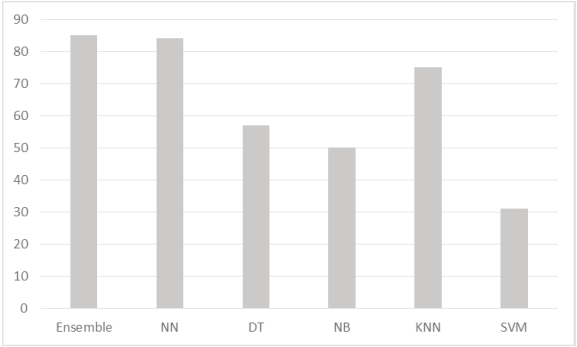

Fig. 3. A comparison of performance of each classification technique and ensemble technique.

4.2. An ensemble learning based model for real estate project classification

Many classification techniques have been used to deal with classification problems and predictions. In this experiment the techniques to be used for real estate grade classification are Artificial Neural Network (ANN), Decision Trees (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). However, the individual classifier may not perform well with the data that has been reduced with the feature selection. Therefore, to increase the performance of the individual classifier, the ensemble learning based model was applied in the experiment. The vote algorithm was implemented for ensemble technique which combines three best-performed classifiers in the experiment. Table 3 presents an accuracy rate of the real estate classification in each technique. The results showed that the ensemble technique performed better than individual classifier.

Table 3. An accuracy rate of the real estate classification in each technique.

Technique	Ensemble	NN	DT	NB	KNN	SVM
Accuracy	85%	84%	57%	50%	75%	31%

4.3. The prototype of the real estate classification application

After the model for real estate classification has been built, the prototype and mobile application has been developed for consumers. The mobile app used Android platform to demonstrate. The app allows user input the criteria related to new accommodation and evaluate the grade of the real estate project. The information to be displayed on the application includes the grade of the real estate project and the loan amount related to the grade of the project. It also showed primary loan payment in each period which supports user for decision-making about loan and real estate buying. The screenshots of prototype application of real estate classification can be seen from Figure 4.

5. Discussion and conclusions

One of the most important needs for human is accommodation. Therefore, seeking for the accommodation seems to be an important activity for consumers. However, there is a variety of real estate project. The grade of the real estate project reflects consumers towards their decision making on a loan and buying. The real estate project grade may be rated differently based on criteria related to the property characteristics such as the experience of a builder or facilities in the project.

Fig. 4. Screenshots for mobile application of real estate classification.

This research proposed the real estate classification model. The data from the real world was collected from the survey and bank's data. The pre-processing stage performed to select the relevant information in the data set to build the real estate classification model. The feature selection in the experiment was performed by an evolutionary approach with neural networks. The weights of the attributes are calculated using a Genetic Algorithm. Next, the reduced attributes were used for the model building. The ensemble learning based classification model was implemented using vote algorithm. The performance in terms of accuracy rate to the unseen data is higher than the well-known classification techniques which were compared in the experiment.

Future works in this area should investigate in the case of feature selection. The high-performance algorithms could be incorporated in the feature selection stage for choosing the most relevant attributes. This can help to reduce the computational time for model building but provide the comparable classification performance. In addition, the data set of real estate have its unique characteristics therefore, the appropriate approach needed to address the problem of real estate classification in terms of the project grade. The novel techniques in classification may be investigated to build the classification model for real estate classification.

References

[1] R. Zhang, Q. Du, J. Geng, B. Liu, and Y. Huang, An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. Habitat International. 46 (2015) 196-205.

[2] V. Kontrimas and A. Verikas, The mass appraisal of the real estate by computational intelligence. Applied Soft Computing, 11 (2011) 443-448.

[3] X. Wu, et al., Top 10 Algorithms in Data Mining. Knowledge and Information Systems 14 (2008) 1-37.

[4] A. Cufoglu, et al. A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling, in: Proceedings of 2009 World Congress on Computer Science and Information Engineering, 2009.

[5] P. Nurmi, et al., A Comparative Analysis of Personalization Techniques for a Mobile Application, in: Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), 2007.

[6] W. Paireekreng, K. W. Wong, and C. C. Fung, A Framework for Integrated Mobile Content Recommendation. International Journal of Electronic Commerce Studies, 4(2013) 185-202.
[7] P. Mitra, C. A. Murthy and S. K. Pal, Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 3(2002) 301–312.
[8] H. Almuallim and T. G. Dietterich, Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69 1-2 (1994) 279–305.
[9] D. Koller and M. Sahami, Toward optimal feature selection, in: Proceedings of the Thirteenth International Conference on Machine Learning, 1996.
[10] J. M. C. S. Susana M. Vieira, Thomas A. Runkler, Two cooperative ant colonies for feature selection using fuzzy models. Expert Systems with Applications, 37(2010) 2714-2723.
[11] Q, Do, D. Grudnitski, A neural network approach to residential property appraisal. Real Estate Appraiser 38-45 (1992).