Research and application of data mining algorithm

Cong Huang¹ ² *, Yao Yang¹ ², Huajun Wang¹ ², Xueyu Zhang¹ ², Jinquan Zhao¹ ², and Jun Wan¹ ²

¹Department of Geophysical Institute, Chengdu University of Technology, 610059, China
²Department of Key Laboratory of Earth Exploration and Information Technology of Ministry of Education, Chengdu University of Technology, 610059

*Cong Huang: 717662413@qq.com

Abstract. With the breakthrough development of a series of technologies such as machine learning and internet of things, data plays an increasingly important role in everyone's daily affairs and work. With the explosive exponential growth of data volume, data types are also quietly changing, from the traditional single, structured data to today's diverse, semi-structured data. In the face of emerging new massive data, data mining technology has gradually become the focus of attention. Data mining was also known as knowledge discovery (KDD) at that time. It was mainly defined as a pattern hidden in massive data, which must be understood by people and bring potential benefits. In this paper, we mainly studies the basic principle and algorithm knowledge of data mining, and applies ridge regression and random forest algorithm model in real estate price forecasting. Finally, through the stacking thinking of ensemble learning in ensemble learning, we propose a fusion method of ridge regression and random forest model, and obtain a more accurate and stable prediction model.

1. Introduction
With the rapid development of artificial intelligence and database technology, almost everyone’s life and work are producing a lot of data all the time. In the face of the increasing amount of data, the traditional data analysis technology is more difficult to effectively discover the information and rules behind the data. People have to face the phenomenon of "rich data but poor knowledge and information". Therefore, in order to solve this problem, data mining technology gradually developed and matured, and well meet people’s expectations and requirements. Data mining technology has been widely used in various fields, especially in the financial industry, telecommunications industry and retail industry. Data mining has formed a mature theoretical knowledge and technical system. The most classic example is Wal Mart’s "beer and diaper". Through data mining technology, it is found that there is a correlation between the sales of beer and diapers. When people buy diapers, they also tend to buy some beer. Therefore, through the discovery of beer and diaper potential correlation, so as to guide the management personnel to place goods reasonably, put the related items together as far as possible, and ultimately improve the sales of goods. Therefore, data mining technology can guide us to make correct decisions and improve the efficiency of the Department.

2. Related work
In the paper, we mainly introduces the basic principle of data mining algorithm and applies it to the prediction of commercial housing price. The reasons behind the high and low house prices are very
complex, which are affected by many factors, such as market supply and demand, the introduction of relevant policies and regulations, housing characteristics, GDP, real estate investment, total population, etc., which makes the study of housing prices more difficult. There is a complex nonlinear relationship between the change trend of house price and the influencing factors, while the traditional research methods, such as linear model, often have large errors in dealing with the nonlinear relationship. In the paper, a method of combining ridge regression and random forest algorithm to build house price forecasting model is proposed, which combines the stable regression coefficient of ridge regression and the strong data capacity of random forest, which can solve the nonlinear problem and improve the prediction accuracy.

3. The method

3.1. Basic principles of data mining
Data mining refers to the extraction of information and knowledge hidden in it, which is unknown in advance, but it has potential application value from a large number of incomplete, noisy, fuzzy and random application data. Data mining is an application-oriented multidisciplinary and interdisciplinary research field, which combines theories and technologies in many different fields, such as machine learning, database and mathematical statistics. For data mining, data management technology is provided by database, while data analysis technology is provided by machine learning and mathematical statistics. At present, the commonly used data mining algorithms mainly include rule induction, decision tree, genetic algorithm, neural network and fuzzy technology. Each mining technology has its own characteristics and application scope. In the application of forecasting the price of commercial housing, the price of real estate is mainly formed by the conditions of scarcity, usefulness and demand, and is affected by many factors. The fusion ridge regression and random forest method proposed in this paper can fit the real estate price data well.

3.2. Ridge regression
Ridge regression, also known as ridge regression, is a common regression method in regression analysis. It is mainly used to process multiple linear data and is essentially an improved least square method. However, compared with the least square method, ridge regression can obtain better regression coefficients, thus improving the accuracy of prediction. The least square method is an unbiased estimation method in regression analysis. It is mainly used when the matrix X is full rank:

\[ \mathbf{X}\theta = \mathbf{Y} \quad (1) \]

When the least square method is used, the loss function is the sum of squares of residuals:

\[ \| \mathbf{X}\theta - \mathbf{Y} \|^2 \quad (2) \]

regression coefficient \( \theta \):

\[ \theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad (3) \]

When the column of matrix X is not in the rank, ridge regression method can be used. The loss function is to add a regularization term on the basis of the sum of squares of residuals:

\[ \| \mathbf{X}\theta - \mathbf{Y} \|^2 + \| \lambda \mathbf{I}\theta \|^2 \quad (4) \]

regression coefficient \( \theta(\lambda) \):

\[ \theta(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \quad (5) \]

among \( \mathbf{I} \) is the identity matrix.

As a biased estimation regression method, ridge regression is the first method used to deal with the number of samples with more features than samples. Due to the addition of the regularization term \( \lambda \mathbf{I}\theta \) is added to the loss function, the regression coefficient \( \theta(\lambda) \) changes with the value of \( \lambda \), forming a constantly changing track on the coordinate axis, which is called ridge trace. Among \( \lambda \), the value is determined by the ridge trace map \( \theta(\lambda) \) when the value is no longer changed, the abscissa is taken as \( \lambda \) Value. Due to the addition of regularization term, ridge regression loses its unbiased feature, but also makes the regression coefficient more stable and the calculation results more accurate.
3.3. Random forest

Random forest is an integrated learning model based on the combination of multiple decision trees, which has two functions of classification and regression prediction. Figure 1 shows the principle of random forest model, in which the decision tree is a sub classifier in the random forest model. When there is sample data input, the decision tree will select the features of the input data and split it to form multiple nodes until the data reaches the classification expectation. The result of random forest classification is determined by the mode of multiple decision trees, that is, the output of random forest tends to the result of classification. Therefore, compared with decision tree, random forest has better classification performance. For the regression prediction of random forest, bootstrap resampling technology is generally used to collect a fixed number of sample data from training data, which is a random sampling process with return. Each decision tree is combined together and the numerical variable is taken as a regression prediction variable to form a multivariate nonlinear random forest regression prediction model.

Through bootstrap sampling method, the randomness of samples can be achieved, thus the anti over fitting ability of the model is improved, and the prediction results obtained are more accurate.

4. Experiment and analysis

The real estate data studied in this paper comes from Boston house price data on the competition website kaggle. Figure 2 shows part of the house price data. The data is divided into two parts. One part is the training data train, including 79 house properties such as lotarea, street and overallqual. The data contains 42 types of data and 37 types of numerical data. Some of the data are shown as follows.

![Figure 2: House price data.](image)

Figure 3 shows the attribute description of some data, mainly describing the meaning of each property of the house and its value range. Some of the data are shown as follows.

![Figure 3: Attribute description of some data.](image)
4.1. Data preprocessing
In order to improve the performance of the model, it is necessary to simplify the time and space processing of the general model before it is cleaned up. In this study, a series of processing processes, such as statistical analysis, missing value processing, correlation analysis, data format conversion and data standardization, are carried out to establish a better prediction model. The treatment process is as follows:

1) Statistical analysis: the missing value statistics and the average value, median, variance and other characteristic statistics of numerical data are mainly carried out.

2) Missing value processing: the missing values in the data are replaced by the average value, and the attributes with more missing values are directly deleted.

3) Correlation analysis: Figure 4 shows that Pearson correlation analysis and thermal mapping are used to determine the degree of correlation between other attributes and the sales price attribute, as well as the correlation between each attribute. The degree of correlation was as follows.

Figure 3 data attribute description.

Figure 4 correlation degree of data attributes.
(4) Data format conversion: for classified data in the dataset, each discrete value is regarded as a state by using one hot feature extraction method. If a feature has n different values, we abstract the feature into n different states. Figure 5 shows the classification data conversion results.

(5) Data standardization: in order to eliminate the dimensional difference between data and abnormal data distribution, the Z-score normalization method is used to transform the data into the same range, so that each attribute data conforms to the normal distribution.

4.2. Model training and fusion
In this paper, the random forest and ridge regression algorithms are used to construct the prediction model, and the performance index of the classifier is evaluated by cross validation method. Then, the stacking thinking in ensemble learning is used to fuse the two algorithm models to construct a combined model. Generally speaking, the combination model constructed by stacking method is better than any single model. The effect of the model is better. Finally, the test data is used to predict the real estate price, and the accuracy rate is calculated by comparing with the real data.

4.2.1. Model training
Random forest and ridge regression models are trained by cross validation. A stable and reliable prediction model can be obtained by cross validation method. Figures 6 and 7 show the parameters and error training results of ridge regression model and random forest model.

4.2.2. Model fusion
Through stacking thinking in ensemble learning, ridge regression and random forest algorithm models are fused to construct a more accurate and stable prediction model. Figure 8 shows the partial results and error rate of the fusion model prediction.
It can be seen from figure 8 that the prediction accuracy is good, but there is a small part of data, and there is a big difference between the predicted price and the real price. Through the statistical analysis of the experimental results, it can be concluded that: the average value of the real price is 178839, the average value of the predicted price is 175833, the average prediction accuracy is 91.8%, and there are 14 forecast data whose prediction accuracy is more than 99.5%.

5. Conclusion
In the paper, we first introduce the research background of data mining technology, point out that data mining technology can guide us to make correct decisions, and then analyze the basic principles and characteristics of ridge regression and random forest algorithm, and apply them to real estate price forecasting. We propose a method of combining ridge regression and random forest algorithm model, which can be used in real estate price forecasting a more accurate and stable prediction model. Through the analysis of the experimental results: the average value of the real price of commercial housing is 178839, the average value of the predicted price is 175833, the average prediction accuracy is 9.2%, and there are 14 forecast data with the prediction accuracy above 99.5%. Because the method proposed in this paper combines the stable regression coefficient of ridge regression and the strong data capacity of random forest, it has obvious accuracy advantage for real estate price prediction.

References
[1] Jitender S D, Vijay V R, Hayri S. (2015) Data Mining: Research Trends, Challenges, and Applications. The Department of Computer Science University, Nebraska.
[2] Sigurður Olafsson, Li X N, Wu S N. (2006) Operations research and data mining. Iowa State University, Iowa.
[3] Tang Q Y. (2012) A Study On Data Mining Algorithm Gragh - based. Baoding Vocational and Technical College, Baoding.
[4] Chen, M.W. (2016) Research on price prediction of commercial housing based on RS-SVM. East China Jiaotong University, Nanchang.
[5] Cai, L.L. (2012) Model selection and parallelization of random forests. Harbin Institute of technology, Harbin.
[6] Du, X.W. (2011) Research on sales forecasting based on data mining. Shandong Institute of light industry, Jinan.
[7] Guan, D.W. (2005) Data preprocessing in data mining. Jilin University, Jilin.
[8] Gao, D.L. (2009) Research on the application of data mining classification algorithm in student information. Qingdao University, Qingdao.
[9] Guo, Z.Y. (2015) prediction and analysis based on reduced random forest. Tianjin University, Tianjin.
[10] Lu, D.B. (2008) Research and application of data mining algorithm based on decision tree. Wuhan University of technology, Wuhan.

[11] Li, A. (2003) Research on data mining algorithm. Dalian Maritime University, Dalian.

[12] Wang, X.X. (2015) Real estate price prediction based on data mining. J. Journal of Inner Mongolia Normal University., 20: 27-34.