Big Data Analyzing Techniques in Mathematical House Price Prediction Model

Jiahao Yang*
Nankai University, Tianjin, 450000, China
*Corresponding author's e-mail: jiahao.yang.20@neoma-bs.com

2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA) | 978-1-6654-1606-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/EEBDA53927.2022.9744970

***Abstract*—The Chinese house market has been flourishing in the past three decades as an increasingly bigger population moves into cities. To keep the rise of house price within a proper range and acceptable to the public, the government has adopted various approaches to bring the price under control. After intermittent fluctuations, house purchasing has become a hot topic for both the media and the public. It is of great significance to study the change of house price considering all the related factors. Some scientists prefer machine learning. Machine learning is a subject which involves many subjects such as probability theory, statistics, approximation theory, convex analysis, and algorithm complexity theory. In this paper, we will discuss the details of the machine learning algorithms, three typical models (i.e., AdaBoost, Linear Regression and KNN) and their strengths and weaknesses. In the future, the growing complexity of all the factors that influence house price will cause more and more trouble for scientists pursuing more precise results. Also, based on its own specialty that computer is still not able to think like a man, machine learning cannot achieve 100% right, which often results in some silly outcomes.**

Keywords—Machine learning, House price, Prediction, Algorithms

I. Introduction

As the house price is vital for both the Ill-being of the public and the economic development, many experts in different research fields have explored and predicted it with the machine-learning strategies. The three main approaches which are popular in this field to predict prices are AdaBoost, Linear Regression and KNN. Because the price is susceptible to multiple factors, it is very challenging to obtain the accurate number. Thus, it is important to decide on the key factors which result in the change of the price. Collect the related information, and transform them into data, so that scientists can use suitable models and algorithms to analyze and evaluate the results in an increasingly accurate way. It has become a hot and vital issue to develop more advanced models and algorithms, which scientists have been working on to overcome for decades.

To fully understand machine learning and some classic models, people first need to know exactly what machine learning is. Machine learning is a subject related to the artificial intelligence, whose research mainly concentrates on computer algorithms and models that can automatically evolve through experiences. So, mostly people regard 1950 as the year that machine learning and artificial intelligence was born when Turing invented the first computer. But tracing back to the 17th centuries, human already got some basic skills and tools of machine learning, like Bayes' and Laplace's derivation of least square method, and the Markov chain. Generally, I can divide all the models and algorithms into two main parts, supervised learning, and unsupervised learning. Supervised learning is the largest part of machine learning, which includes most classic models.

TABLE I FOUNDERS AND YEARS OF DIFFERENT MODELS.

Name	Founder	Year
Linear discriminant analysis [1]	Fisher	1936
Bayes classifier	Bayes	1950
Logistic Regression [2]	Cox	1958
KNN [3]	Thomas	1967
CART [4]	Breiman	1984
ID3 [5]	Quinlan	1986
CNN	LeCun	1989
AdaBoost [6]	Freund	1995
Random Forest [7]	Breiman	2009

Prior to 1980, the algorithms are all fragmented and unsystematic, but they made great contributions to the overall development of machine learning, which could not be ignored. For example, KNN was raised in 1967, simple but very useful, and is still being used up to now. Machine learning did not become an independent research direction until 1980. After that, neural network, inheritance learning algorithm and decision tree, lots of algorithms Ire raised and put into massive use. AdaBoost was raised in 1995, and can reach an impressing accuracy level [6]. We can see the history of different models from Table I above.

Predicting house price is a complex and challenging issue. There are lots of factors that may influence the house price, including the location, orientation of rooms, stores, decoration, neighborhoods, schools, traffic conditions, and security issues, etc. It is impossible to take into consideration all the factors concerned in predicting the price. Moreover, rental housing price is vulnerable to the economic condition both in China and worldwide. While this condition is changing all the time, it is challenging research for scientists to get the exact data and predict the price change. Another question is that machine learning models are constraint to math progresses. Therefore, without a better algorithm model with the focus on the key factors which impact the real price, it's difficult to make improvement in the price prediction.

The remaining of this paper will analyze the models in machine learning which are frequently used in predicting house price. Then I pick three typical examples to discuss in detail. The analysis includes the strength and weakness of different models and their usage in some real-life research.

II. Main body

A. Overview

There are 5 main steps of using machine learning models to predict the house price. First scientists should decide on the places to predict and get related data, either through data retrieval or other ways. Second is processing the data.

Because the data from websites are mainly raw ones, including null value, invalid value, so functions are needed to delete or ignore them to insure the normal use of data. Third is the import of models, which is a crucial part for scientists to get more precise results. Forth is an evaluation of the results to check if the model is suitable to predict. The last step is the implementation of the whole program. After all these processes, the program can finally predict house prices with a higher precision.

Different machine learning methods show different prediction performances even if the same feature set and training data are used [8]. Here I list 5 main models that are frequently used by scientists, and their strengths and weaknesses in Table II.

TABLE II THE STRENGTHS AND WEAKNESSES OF DIFFERENT MODELS.

Name	Strength	Weakness
AdaBoost	1. Makes good use of weak classifier to cascade 2. Has high accuracy 3. Fully considers the weight of each classifier	1. When the data set is not balanced, it leads to the decline of classification accuracy 2. Training is time-consuming, and the best segment point of the current classifier is re-selected each time
Logistic	1. Low computation cost 2. Easy to understand and implement	Low classification accuracy
SVM	1. Can solve machine learning problems with small samples 2. No local minimum problems 3. Can handle high dimensional data sets very Ill	1. Sensitive to missing data 2. The interpretation power of higher dimensional mapping of kernel function is not strong
Decision Tree	1. Easy to understand and explain 2. Runs fast when testing the data 3. Extends well to large database and the size is independent of the database size	1. Difficult to process missing data 2. Easy to ignore the correlation of attributes in the dataset
KNN	1. An online technique where new data can be added directly to a data set without retraining 2. The theory is simple and easy to implement	1.For the data set with large sample size, the calculation is larger 2. When the sample is unbalanced, the prediction deviation is relatively large

B. AdaBoost

AdaBoost, abbreviation of Adaptive Boosting, is an iterative algorithm which has the core idea to train different classifiers (weak classifiers) for the same training set, and then aggregate these weak classifiers to ultimately form a stronger classifier (strong classifier).

From the essays I can find that AdaBoost performs well in some fields. In one case, scientists compared the effects of different models to achieve model optimization. Except for the low independent of standard Re of 59.09% [9], most of the results for AdaBoost standard were greater than 75.0%. In addition, concerning different strategies of AdaBoost, they all got similar marks and have slight differences. Take SMOTE for example, the training set had marks roughly higher than 80%. The testing set also presented approaching 60% Re.

AdaBoost has some weaknesses in the meanwhile. In the case predicting drug properties, the RSFBoost method performs better than that of AdaBoost [10]. Because of the special characteristic of the AdaBoost, it is more vulnerable to the high diversity of the data, which will decrease the accuracy of the results.

All in all, AdaBoost works well when the dataset is balanced and not so big. The reason behind this its perfect performance is that, although it uses weak classifier, the aggregation of many weak classifiers will show great power in doing calculation and modifying. AdaBoost performs better than other models in many cases, its advantage lies in the framework itself, which combines the classifier with weak performance to form a stronger one. Research also shows that with more weak classifiers, AdaBoost performs better. When I try to train AdaBoost, I found that with different number of weak classifiers, the accuracy of the result is different. If the number of classifiers is less than 20, the rate of mistake is higher than 25%. But after I increase the number to 200, the rate decreases to less than 5%. The high accuracy of AdaBoost makes it powerful and a tool in hand in predicting house price.

C. Linear Regression

Linear Regression is a statistical analysis method that uses regression analysis to determine the interdependent quantitative relationship between several variables. It is also widely used in predicting house price because it is easier to understand and runs fast when testing the data. Linear regression is widely used in machine learning fields. It is the first type of regression analysis that has been strictly studied and widely used in practical applications. This is because linear models depend on their unknown parameters, which makes them easier to fit than nonlinear models that depend on their unknown parameters, and the statistical properties of the resulting estimates are easier to determine.

In a chemical study using near-infrared (NIR) spectroscopy to characterize the biomass of algal, scientists use single and multiple linear regression [11]. The results show that by using the multiple regression model, of the 36 predicted values, only 6 of them show lower accuracy with more than 15% relative difference. Other results are acceptable and perfect with the rate less than 10%. In the article the author expressed his expectation towards the future of machine learning prediction. He believed the improvement of accuracy will make the detection of outliers easier.

In another research on predicting the house price in Boston, the writer uses Linear Regression in the functions. By using Rescaling and Standardization, the writer adopted feature scaling method in order to standardize the range of data features and speed up convergence. As for the result, both Rescaling and Standardization have the RMSE (i.e., Root Mean Square Error) standard lower than 20. The Rescaling shows better performance with the standard lower than 14.

Here one thing needs to be paid attention to. Mostly when coping with results that are vulnerable to several variables, scientists use multiple linear regression (MLR) to fully consider all the factors. And in the real life, cases are that things are usually influenced by several factors. Thus, multiple linear regression is more practical in the real life than single linear regression.

D. KNN

This part will analyze the KNN model in detail. K-Nearest Neighbor (KNN) classification algorithm is one of the simplest methods in data mining classification. K nearest neighbors means that every sample can be represented by its closest K neighbors. The nearest neighbor algorithm is a method to classify every record in the data set [16]. The advantage of KNN is that the idea is simple and easy to understand or to implement, without the need of estimating parameters [12]. However, KNN also have aspects to improve. When the sample is imbalanced, such as the sample size of one class is very large while the sample size of other classes is very small, it may lead to the sample with large volume class become the majority of K neighbors when a new sample is imported [13]. What is more, because KNN needs to calculate the distance between different samples, it has a large calculation quantity.

When talk about using KNN in the prediction of house price, we need to attention first that house price is a question about regression but not classification. Although KNN can be used to do the prediction, it may not be easy to get the same accuracy level as AdaBoost or Linear Regression. Usually, scientists need to apply restrictions on the form of data.

The following is an example of house price prediction using KNN and several other models. The author used a figure to depict the comparison of algorithms based on the performance measure of the model to predict the most suitable and least error prone output [14]. Among the 19 different models, KNN ranks the 15th, lower than most of the boost and linear regression, among which Catboost ranks the 1st.

Despite the disadvantage of classification over regression of KNN, we can add limitation and standardization to enhance the accuracy. Standardizing the source data is a good way, because KNN is a cluster based on European distance, if the influence of different features varies greatly due to the difference in value, data standardization is needed. I tried to value the result of prediction using MAE standard, the difference of which is that one set used standardization while the other didn't. The MAE of the one standardized is 0.39, 0.75 less than the other, which means after standardizing the data, the accuracy improved.

E. Conclusion of the models

At the end of this chapter, I will give a comprehensive conclusion of the three main models. The summary is important because when dealing with classification and regression problems, although all models can solve the problem, different models have different performance and sometimes even a slightly different can cause the opposite result. So fully understand the feature of the models is important.

All in all, of the three models above mentioned in detail, AdaBoost can reach a high accuracy with more classifiers, multiple linear regression is good at dealing with several factors than single linear regression, KNN needs restrictions to ensure the accuracy. If we want to predict house price, after preprocessing the data, it's better to use the means of cross validation to determine which is the best model for a certain circumstance.

III. Conclusion

In conclusion, by retrieving data from the website and using the statistic learning strategy of machine learning, scientists can easily obtain the various data in a proper form and perform prediction at any given time. As for the accuracy, we can see from above that basically the models can satisfy the level of scientific research including house price prediction. What is more, the result is supportive enough to meet the basic requirements of the ordinary people.

There indeed exists many difficulties when analyzing different models and their characters. A limited knowledge of mine causes great trouble in reading the essays and understanding how and why the scientists use them. Especially when faced with multiple technical terms, I need to find what they truly mean in the articles, otherwise I will misunderstand the meaning of the author and get the wrong conclusion.

In the future, house price problems will be more and more complex, susceptible to much more factors. To achieve high accuracy, scientists need to know exactly how the factors interact with each other and influence the house price. They also need to pick the proper model to predict. But on the other hand, scientists will produce more powerful and convenient models which can be used in considering all the factors and predicting the prices more precisely. As one author says, 'The model allows a person to easily find a price for their house or for their-to-be house without needing for them to consult a real estate broker or any third party for the same price calculation and saving a lot of money that any third party would take for providing their service to the customer. Many people look forward to the future of machine learning in house price prediction, because the market is large. The demand will motivate scientists to develop more and more advanced models which will make our life more and more convenient.

References

[1] R. A. Fisher, The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 7(2), 179–188 (1936)

[2] DR. Cox, The regression analysis of binary sequences (with discussion). J Roy Stat Soc B. 20 (2), 215–242 (1958)

[3] Thomas M Cover, Peter E Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory (1967)

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees, Wadsworth (1984)

[5] J. R. Quinlan, Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106 (1986)

[6] Y. F. CrossRef and R. Schapire, Boosting a weak learning algorithm by majority. Information and Computation, 121: 256–285 (1995)

[7] Breiman, Leo. Random Forests. Machine Learning 45 (1), 5-32 (2001)

[8] An Chen, Xu Zhang, Letian Chen, Sai Yao, and Zhen Zhou. A Machine Learning Model on Simple Features for CO2 Reduction Electrocatalysts. The Journal of Physical Chemistry, 124 (41): 22471-22478 (2020)

[9] Lijun Dou, Xiaoling Li, Lichao Zhang, Huaikun Xiang, and Lei Xu. iGlu_AdaBoost: Identification of Lysine Glutarylation Using the AdaBoost Classifier. Journal of Proteome Research, 20 (1): 191-201. (2021)

[10] Tomasz Arodź, David A. Yuen, and Arkadiusz Z. Dudek. Ensemble of Linear Models for Predicting Drug Properties. Journal of Chemical Information and Modeling, 46 (1): 416-423 (2006)

[11] L. M. L. Laurens and E. J. Wolfrum. High-Throughput Quantitative Biochemical Characterization of Algal Biomass by NIR Spectroscopy; Multiple Linear Regression and Multivariate Linear Regression Analysis. Journal of Agricultural and Food Chemistry, 61 (50): 12307-12314 (2013)

[12] Xiangyang Zeng. Intelligent underwater target recognition. National Defense Industry Press, 2016.03: 107 (2016)

[13] Guang Cheng, Aiping Zhou, and Hua Wu. Internet Big data mining and classification. Southeast University Press, 2015.12: 18 (2015)

[14] G. K. Kumar, D. M. Rani, N. Koppula and S. Ashraf. Prediction of House Price Using Machine Learning Algorithms. 2021 5th ICOEI, 1268-1271 (2021)