Research on Optimization of Machine Learning Algorithm Based on Feature Selection

Lei Yang, Yungui Chen and Longqing Zhang
Guangdong University of Science and Technology, Dongguan, China

Keywords: Feature Selection, Machine Learning, Prediction.

Abstract: Feature selection, as a method of machine learning algorithm optimization, aims to improve the performance and efficiency of the algorithm by selecting important features. This research selects 3 A-share stocks of China's Shanghai Stock Exchange as the research object and proposes a prediction model based on feature selection decision tree optimization algorithm, which is used to predict the closing price of the stock. Our research results show that the optimized Models can improve the performance of predictions.

1 INTRODUCTION

In machine learning, feature selection is a commonly used data preprocessing technique, which can reduce the data dimension, reduce the influence of noise and redundant data, and improve the generalization ability and interpretability of the model at the same time (Zhou H F-Alsahaf A). Research on machine learning algorithm optimization based on feature selection is an important research direction, which aims to improve the performance and accuracy of machine learning algorithms, so as to better solve practical problems (Sharma, 2022).

Feature selection refers to the selection of the most useful features in machine learning algorithms, thereby reducing the complexity of the model and improving the generalization ability of the model. There are many methods of feature selection, including filtering, wrapping, and embedding. The filtering method refers to evaluating the importance of features by calculating the correlation between features and labels, and then sorting according to the importance index to select features with higher rankings 0 (Christo,2022). The advantage of the filtering method is fast calculation, but the disadvantage is that it is easily affected by noisy data. The wrapping method refers to searching the feature space through a search algorithm to find the optimal feature subset. The advantage of the wrapping method is that it can find the optimal feature subset, but the disadvantage is that it is computationally intensive and time-consuming (Chen, 2021). The embedding method refers to limiting the complexity of the model through methods such as regularization during the training process of the machine learning algorithm, thereby realizing feature selection. The advantage of the embedding method is that it can optimize the accuracy and generalization ability of the model at the same time, but the disadvantage is that it is computationally intensive and time-consuming 0 (Yun, 2021).

This research uses the filtering method for feature selection, uses spearman correlation to analyze the characteristics of the data, and extracts more relevant data to become another data set separately, using the same machine learning algorithm, that is, decision tree algorithms to learn and predict on two datasets.

2 DATASET AND METHOD

The data used in this research comes from the data of China's Shanghai Stock Exchange Index in the first half of 2022. The historical transaction data of three China's Shanghai Stock Exchange A-shares were selected as the research objects, namely Gd Power Development Co., Ltd., Hundsun Technologies Inc. and Ningbo Joyson Electronic Corp. The stock codes are 600795, 600570 and 600699 respectively.

This research analyzes and processes the data of the opening date, opening price, highest price, lowest price, closing price, and trading volume of these three stocks, and uses the machine learning algorithm of decision tree to predict the stock price

Yang, L., Chen, Y. and Zhang, L.
Research on Optimization of Machine Learning Algorithm Based on Feature Selection.
DOI: 10.5220/0012275300003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 100-103
ISBN: 978-989-758-677-4

trend. The original data set of each stock has 117 samples, each sample contains 8 features, the first 102 samples are used as the training set, and the remaining 15 samples are used as the test set.

We set the maximum depth of the decision tree to 100 layers, use the inductive binary tree, set the minimum number of samples on the leaf nodes to 3, and no longer split when the number of subsets is less than 3, the regression prediction model is used in the prediction, and the prediction Visualize the analysis results after outputting the results(Luo M,2021). Meanwhile, we also performed performance analysis of our predictive model using different evaluation metrics.

Divide the input space into M regions R_1, R_2, ... R_M and generate the decision tree:

$$f(x) = \sum_{m=1}^{M} \hat{c}_m I(x \in R_m) \quad (1)$$

The tree can be created as shown in Algorithm 1.

Algorithm 1: Tree creation.

INPUT: Training set T;
Feature set F
OUTPUT: A tree
Procedure : Function TC(T, F)
0: Initialize a node;
1: **if** The tests in T all have the same sort W **then**
2: Stamp node as a W-type leaf node; **return**
3: **end if**
4: **if** $F = \emptyset$ **OR** The test in T has the same value on F **then**
5: Stamp node as a leaf node;
6: The type is singled out in T; **return**
7: **end if**
8: Select the ideal parcel feature $a,*$ from F;
9: **for** Each value $a^{t},*$ in $a,*$ **do**
10: Create a branch for node;
11: Let T_i indicate a subset of tests in T that have a value of $a^{t},*$ on $a,*$;
12: **if** T_i is vacant **then**
13: Stamp branch nodes as leaf nodes;
14: The type is singled out in T; **return**
15: **else**
16: TC(T_i, $F\backslash\{a,*\}$) as a branch node;
17: **end if**
18: **end for**

To optimize predictive models for machine learning, we apply feature selection to the original dataset. Feature selection algorithms include filtering, wrapping, and embedded three types. Filtering feature selection algorithms are independent of any specific machine learning model and select features by performing statistical tests or mathematical transformations on them 0. Wrapped feature selection algorithms treat feature selection as a search problem, taking a subset of features as input, and evaluating the performance of each subset using a given machine learning algorithm 0. The embedded feature selection algorithm embeds feature selection into the training process of machine learning algorithms and selects the best features by optimizing the loss function of the model.

For the selection of feature evaluation indicators, feature evaluation indicators are mainly used to evaluate the importance and contribution of features. Commonly used indicators include information gain, chi-square test, mutual information, and correlation coefficient 0.

In order to better predict the stock price, we use Spearman correlation to analyze the characteristics of the data, conduct correlation analysis on the 8 features in the data set, and finally select the strongest 3 features plus the opening date, A total of 4 features were used to construct a new data set, and the same decision tree algorithm and parameters were used to perform machine learning on the new data set again. Our research results show that the optimized model can improve prediction performance.

3 EXPERIMENT AND ANALYSIS

A decision tree algorithm is an algorithm used in machine learning that uses a tree model to build a decision process from input to output. It is a type of classification and regression tree (CART) algorithm and can be used for classification and regression tasks 0. The advantage of the decision tree algorithm is that it can represent the decision process in a readable form, and it is also fast to train. As shown in Figure 1, it is the decision tree model of the stock price prediction of the stock Gd Power, which is visualized with the Pythagorean tree. The decision tree models for the other 2 stocks are similar.

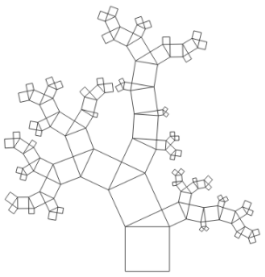

Figure 1: The decision tree model with Gd Power.

In this research, the historical transaction data of 3 stocks are divided into training sets and test sets. There are 351 pieces of data in total. The total amount of data for a single stock is 117. In order to better compare the effect of feature selection on the prediction model, the training set and test set adopt a fixed sampling mode. There are 102 samples in the training set and 15 samples in the test set. The prediction results are shown in Table 1.

Table 1: The predicted results and real values of the 3 stocks.

SN	Gd Power real	Gd Power pred	Gd Power devi	Hundsun real	Hundsun pred	Hundsun devin	Joyson real	Joyson pred	Joyson devi
1	3.8	3.78	0.47%	42.9	42.52	0.90%	13.45	13.31	1.04%
2	3.73	3.73	0.00%	42.02	41.67	0.84%	13.7	13.77	0.51%
3	3.8	3.78	0.47%	42.18	42.52	0.79%	13.94	13.77	1.22%
4	3.72	3.76	0.99%	43.86	44.07	0.49%	13.57	13.59	0.18%
5	3.69	3.66	0.88%	41.76	41.67	0.23%	13.48	13.59	0.85%
6	3.78	3.78	0.05%	42.78	42.52	0.62%	13.63	13.59	0.26%
7	3.77	3.76	0.35%	42.79	42.56	0.54%	13.96	14.05	0.66%
8	3.96	3.92	1.01%	45.2	44.88	0.70%	13.66	13.59	0.48%
9	3.89	3.92	0.77%	43.17	43.25	0.17%	13.63	13.59	0.26%
10	3.78	3.76	0.62%	45.01	44.88	0.28%	14.65	14.43	1.48%
11	3.74	3.73	0.27%	45.25	45.17	0.18%	15.56	15.41	0.96%
12	3.72	3.73	0.27%	45.69	45.77	0.18%	16.24	16.19	0.29%
13	3.77	3.78	0.32%	45.7	45.17	1.17%	16.28	16.19	0.54%
14	3.92	3.92	0.00%	44.27	44.88	1.39%	15.98	16.19	1.33%
15	3.91	3.92	0.26%	43.54	43.25	0.68%	15.71	15.84	0.83%

It can be seen from the above table that most of the deviations between the predicted values of the 3 stocks and the real values are less than 1%. The curves of the predicted values and the real values of the 3 stocks are shown in Figure 2.

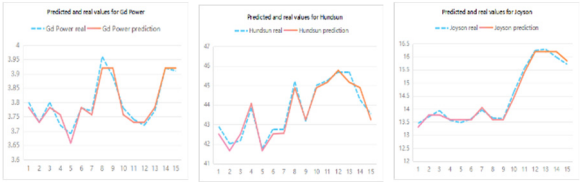

Figure 2: The curves of the predicted value and the real value of the 3 stocks.

In order to optimize our prediction model, for the original data set, we use Spearman correlation to analyze the characteristics of the data and perform correlation analysis on the 8 features in the data set. The correlation analysis of the three stocks is shown in Table 2.

Table 2: The correlation analysis of the 3 stocks' features.

SN	Gd Power			Hundsun			Joyson		
1	0.989	close	high	0.994	close	high	0.995	close	low
2	0.986	close	low	0.99	close	low	0.994	close	high
3	0.971	close	open	0.983	close	open	0.987	close	open
4	0.818	amount	close	-0.785	close	date	-0.747	close	date
5	0.727	close	volume	-0.342	close	volume	0.437	amount	close
6	0.554	close	date	0.029	amount	close	0.018	close	volume

As can be seen from Table 2, the three features of the three stocks, the opening price, the highest price of the day, and the lowest price of the day, have the strongest correlation with the closing price. Therefore, we choose these three features, plus the opening time, to form a the new data set uses the same decision tree algorithm and parameters to perform machine learning on the new data set again, and the prediction results are shown in Table 3.

Table 3: The predicted results and real values of the new dataset.

SN	Gd Power real	Gd Power pred	Gd Power devi	Hundsun real	Hundsun pred	Hundsun devi	Joyson real	Joyson pred	Joyson devi
1	3.8	3.79	0.18%	42.9	42.43	1.10%	13.45	13.46	0.04%
2	3.73	3.73	0.13%	42.02	41.89	0.31%	13.7	13.82	0.88%
3	3.8	3.79	0.18%	42.18	42.43	0.58%	13.94	13.82	0.86%
4	3.72	3.73	0.13%	43.86	44.07	0.49%	13.57	13.60	0.22%
5	3.69	3.69	0.00%	41.76	41.89	0.31%	13.48	13.56	0.56%
6	3.78	3.79	0.35%	42.78	42.79	0.01%	13.63	13.56	0.55%
7	3.77	3.78	0.13%	42.79	42.79	0.01%	13.96	14.05	0.64%
8	3.96	3.94	0.51%	45.2	45.07	0.28%	13.66	13.67	0.04%
9	3.89	3.90	0.26%	43.17	43.36	0.43%	13.63	13.60	0.22%
10	3.78	3.78	0.13%	45.01	45.07	0.14%	14.65	14.43	1.48%
11	3.74	3.73	0.27%	45.25	45.17	0.18%	15.56	15.46	0.64%
12	3.72	3.73	0.27%	45.69	45.54	0.34%	16.24	16.11	0.80%
13	3.77	3.77	0.13%	45.7	45.17	1.17%	16.28	16.28	0.03%
14	3.92	3.94	0.51%	44.27	44.60	0.75%	15.98	16.11	0.81%
15	3.91	3.90	0.26%	43.54	43.36	0.42%	15.71	15.84	0.83%

MSA (Mean Squared Absolute Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R2 (R Squared) are commonly used regression error indicators or scoring indicators. The value of R2 is between 0 and 1, the closer to 1, the better the fitting effect of the model. The performance of the predictive scoring indicators of the two machine learning models on the two datasets is shown in Table 4.

Table 4: The performance of the decision tree model.

Stock	MSE before	MSE after	RMSE befor	RMSE after	MAE before	MAE after	R2 before	R2 after	R2 impr	Avg devi befor	Avg devi after	Avg devi impr
Gd Power	0.011	0.006	0.021	0.01	0.017	0.009	0.93	0.983	5.70%	0.45%	0.23%	48.97%
Hundsun	0.097	0.058	0.311	0.242	0.267	0.191	0.945	0.966	2.22%	0.61%	0.44%	28.74%
Joyson	0.015	0.011	0.122	0.103	0.106	0.084	0.987	0.991	0.41%	0.73%	0.57%	21.12%

As shown in Table 4, the performance of the prediction model after the feature selection process has been improved in both R2 and average deviation, especially the average deviation index, which has increased by more than 20%. Our experimental results show that after feature selection processing, the performance of machine learning model prediction can indeed be improved.

4 CONCLUSION

Based on the feature selection algorithm, this research improves the prediction model of the decision tree algorithm and uses this model to predict the closing price of the stock. The research shows that the performance of the prediction model has been improved. The optimization research of machine learning algorithms based on feature selection can also help Solve other types of practical problems, such as text classification, image recognition, bioinformatics, financial risk control, etc. For stock price prediction, there are many optimization methods based on feature selection, which we will further study in future topics research.

ACKNOWLEDGMENTS

This research was funded by the following programs: Research Capability Enhancement Project of Guangdong University of Science and Technology: Application Research of Artificial Intelligence Technology Based on Kunpeng Computing Platform; Natural Science Project of Guangdong University of Science and Technology(GKY-2022KYZDK-12, GKY-2022K YZDK-9); Innovation and School Strengthening Project of Guangdong University of Science and Technology(GKY-2022CQTD-2, GKY-2022CQTD-4, CQ2020062); 2022 Guangdong Province Undergraduate University Quality Engineering Construction Project - Exploration of Integrated Teaching Reform of "Curriculum Chain, Practice Chain, and Competition Chain"; Education Research and Reform Project of Online Open Course Alliance of Universities in Guangdong–Hong Kong–Macao Greater Bay Area(WGKM2023166); Quality Engineering Project of Guangdong University of Science and Technology(GKZLGC2022018, GKZLGC2022271).

REFERENCES

Zhou H F, Zhang J W, Zhou Y Q, et al. A feature selection algorithm of decision tree based on feature weight[J]. *Expert Systems with Applications*, 2021, 164: 113842.

Qian H, Wang B, Yuan M, et al. Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree[J]. *Expert Systems with Applications*, 2022, 190: 116202.

Alsahaf A, Petkov N, Shenoy V, et al. A framework for feature selection through boosting[J]. *Expert Systems with Applications*, 2022, 187: 115895.

Sharma A, Mishra P K. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis[J]. *International Journal of Information Technology*, 2022: 1-12.

Christo V R E, Nehemiah H K, Brighty J, et al. Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest[J]. *IETE Journal of Research*, 2022, 68(4): 2508-2521.

Chen W, Jiang M, Zhang W G, et al. A novel graph convolutional feature based convolutional neural network for stock trend prediction[J]. *Information Sciences*, 2021, 556: 67-94.

Yun K K, Yoon S W, Won D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process[J]. *Expert Systems with Applications*, 2021, 186: 115716.

Luo M, Wang Y, Xie Y, et al. Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass[J]. *Forests*, 2021, 12(2): 216.

Bhandari H N, Rimal B, Pokhrel N R, et al. Predicting stock market index using LSTM[J]. *Machine Learning with Applications*, 2022, 9: 100320.

Kumbure M M, Lohrmann C, Luukka P, et al. Machine learning techniques and data for stock market forecasting: A literature review[J]. *Expert Systems with Applications*, 2022, 197: 116659.

Buchaiah S, Shakya P. Bearing fault diagnosis and prognosis using data fusion based feature extraction and feature selection[J]. *Measurement*, 2022, 188: 110506.

Qiu Y, Song Z, Chen Z. Short-term stock trends prediction based on sentiment analysis and machine learning[J]. *Soft Computing*, 2022, 26(5): 2209-2224.

Kurani A, Doshi P, Vakharia A, et al. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting[J]. *Annals of Data Science*, 2023, 10(1): 183-208.