Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 173-186
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Application of Decision Tree Algorithm in Housing Purchase Problems

—A Case Study of Xining City

Siyu Chen, Li Fu*

School of Mathematics and Statistics, Qinghai Nationalities University, Xining, China
Email: *727397895@qq.com

How to cite this paper: Chen, S.Y. and Fu, L. (2024) Application of Decision Tree Algorithm in Housing Purchase Problems. *Journal of Computer and Communications*, **12**, 173-186.
https://doi.org/10.4236/jcc.2024.1211013

Received: October 25, 2024
Accepted: November 24, 2024
Published: November 27, 2024

Abstract

Decision tree is an effective supervised learning method for solving classification and regression problems. This article combines the Pearson correlation coefficient with the CART decision tree, replacing the Gini coefficient with the correlation coefficient to consider the correlation between conditional attributes, prioritizing the selection of conditional attributes with higher correlation coefficients as leaf nodes. The collected data on homebuyers is divided into age groups, including youth, middle-aged, and elderly groups. Both traditional CART decision tree and improved CART decision tree are applied to this problem, and after comparison, it is found that the depth of the CART decision tree in this study is reduced, the number of leaf nodes is decreased, the time complexity is shortened, efficiency is improved, and pruning issues are avoided. Finally, corresponding housing recommendations are given to homebuyers of different ages.

Keywords

Decision Tree, Gini Coefficient, Correlation Coefficient

1. Introduction

Decision tree is a type of supervised learning in machine learning, representing a mapping relationship between sample values and attributes. The decision tree algorithm is easy to understand and implement, robust, readable, and able to handle both categorical and numerical attributes simultaneously. Its applications are widespread, including credit assessment in the financial industry, disease diagnosis in the medical industry, and quality inspection in industrial manufacturing, among many other fields.

The CART decision tree algorithm was first introduced in reference [1], and it is studied based on the Gini index. Literature [2] discusses the ID3 algorithm, which calculates using information gain. Building upon the ID3 algorithm, literature [3] utilizes information gain ratio and upgrades it to the C4.5 algorithm. Taking Lanzhou City, Chengguan District as an example, literature [4] addresses the problem of garage location selection by analyzing data through the mutual information attribute reduction algorithm of fuzzy rough sets, and then generates a decision tree through inductive learning algorithms. This model can be further studied in other problem domains. Literature [5] focuses on the ID3 decision tree, improving it with correlation coefficients to overcome its bias towards multi-values and simplifying the Gini gain formula using Taylor series and McLaurin series. Building on literature [5], literature [6] uses a more relaxed Spearman correlation coefficient instead of the stringent Pearson correlation coefficient, enhancing the algorithm's classification accuracy to a certain extent and optimizing its runtime speed. For village location selection problems, literature [7] considers geographical, transportation, and employment factors as conditional attributes, constructing an evaluation model using expert groups, weight methods, and analytic hierarchy process. There is still room for improvement in the selection of conditional attributes in this model. To reduce algorithm complexity, literature [8] proposes a reduction algorithm based on k-nearest neighbor attribute importance and integrates the correlation coefficient method to eliminate redundant information from reduced attributes. Finally, literature [9] combines Pearson correlation coefficient with decision tree algorithms to minimize noise accumulation in large-scale datasets, further enhancing the algorithm's utility in handling large-scale datasets.

Literature [8] and literature [9] indicate that correlation coefficients can be applied in decision information systems. The former involves the reduction of conditional attributes in rough sets, while the latter focuses on enhancing efficiency in decision tree algorithms. Literature [10] applies the ID3 decision tree in the metal field. In literature [11], both information entropy-based and Gini coefficient-based decision trees are seen as having room for improvement. A new decision tree algorithm is designed by merging the two to establish a fusion metric of information gain and Gini coefficient with a knowledge-based weighted linear combination, highlighting their equivalent utility for functional purposes. Literature [12] applies the C4.5 decision tree in the aerospace field, while literature [13] utilizes the CART decision tree in the domain of bridge earthquake analysis. Additionally, literature [14] applies an improved CART decision tree in the area of oak branch and leaf management, and literature [15] implements the improved CART decision tree in workshop scheduling problems.

The observation of the research conducted by the aforementioned scholars reveals that while studies have been carried out on location selection issues, they have not extended into the field of real estate purchases. Additionally, the methodologies employed do not incorporate the concept of correlation coefficients.

Scholars researching correlation coefficients have only applied them to decision tree problems constructed using information entropy, without integrating them with the Gini coefficient. Therefore, combining correlation coefficients with the Gini coefficient and applying this approach to real estate purchasing issues represents a worthy area of research.

This study employs the CART decision tree, which, compared to information entropy decision trees, can handle both classification and regression problems. In contrast to ID3, C4.5, and C5.0 decision trees, the CART decision tree exhibits higher stability, algorithm efficiency, lower time complexity, and lower computational complexity. The mathematical model established in this study combines the Pearson correlation coefficient with the CART decision tree. While the root node is still selected using the Gini coefficient, other nodes are now chosen using the correlation coefficient. The correlation coefficients between each conditional attribute and the previous node are calculated, with the node with the highest correlation coefficient chosen as the next node. Subsequently, the improved decision tree is applied to the housing purchase issue in Xining City, providing recommendations to homebuyers based on different age groups.

2. Basic Knowledge

2.1. Commonly Used Decision Trees

CART Decision Tree [1]: The CART decision tree is a binary tree that can handle both classification and regression problems. It was first proposed by Breiman *et al.* in 1984. This decision tree is widely recognized for its ease of understanding, use, and interpretation, as well as its more accurate predictions and significant algorithmic advantages.

The basic calculation process of the CART decision tree is as follows:

$$Gini(N)=\sum_{d=1}^{D}p_d(1-p_d)=1-\sum_{d=1}^{D}p_d^2 \tag{1}$$

where N represents the sample set, $Gini(N)$ denotes the Gini coefficient of sample set N, which reflects the uncertainty. A higher value of the Gini coefficient indicates a greater uncertainty in the sample set. D represents the decision attribute, d represents the decision class, and p_d represents the probability of a sample belonging to decision class d.

$$Gini(N|A)=\frac{|N_1|}{|N|}Gini(N_1)+\frac{|N_2|}{|N|}Gini(N_2) \tag{2}$$

The above formula represents the Gini coefficient of sample set N under the condition attribute A. Here, N_1 and N_2 respectively denote the two possible partitions of sample set N based on the different values of condition attribute A.

ID3 Decision Tree [2]: The ID3 algorithm was proposed by J. R. Quinlan in 1975 at the University of Sydney as a classification prediction algorithm with information entropy as its core. It calculates the information gain for each conditional attribute, selects the conditional attribute with the highest information gain

as a node, and repeats this process to generate the decision tree.

C4.5 Decision Tree [3]: The C4.5 decision tree is an upgraded version of the ID3 decision tree, and its main advantage lies in the ability to handle both continuous data and missing data. It uses information gain ratio as a measure, with a higher information gain ratio indicating a higher priority for selecting the corresponding conditional attribute. The differences between various types of decision trees are outlined in Table 1 below.

Table 1. Comparison of different types of decision trees.

Type	Principle	Method	Issue	Result	Node selection
ID3	Information entropy	Information gain	Discrete	Maximum	Maximum information gain value
C4.5	Information entropy	Information gain ratio	Continuous	Maximum	Maximum information gain ratio value
C5.0	Information entropy	Information gain ratio	Continuous	Maximum	Maximum information gain ratio value
CART	Gini index	Gini index	Continuous	Minimum	Minimum gini index value

2.2. Pearson Correlation Coefficient

The Pearson correlation coefficient, also known as Pearson's *r*, is a measure of the linear relationship between two variables. It was introduced by Karl Pearson in the 1880s. It is used to quantify the degree of correlation between two variables and its value ranges from −1 to 1. The closer the value is to the endpoints, the stronger the correlation; as it approaches 0, the correlation becomes weaker.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{3}$$

When the sample size n is large, the Pearson correlation coefficient r will become the population correlation coefficient ρ. Compared to Spearman, Kendall, and Hoeffding correlation coefficients, Pearson correlation coefficient is suitable for studying a large number of samples, provides an intuitive understanding of the correlation between variables, and can simplify mathematical models. However, its usage conditions are more stringent.

Note: Conditions for using Pearson correlation coefficient:

The variables are continuous variables.

The variables follow a normal distribution.

There is a linear relationship between the variables.

3. Decision Tree Based on Pearson Correlation Coefficient

Gini Index Literature [3] indicates the heterogeneous independence of information entropy and the Gini coefficient. Literature [7] improved the ID3 decision

tree using the Spearman correlation coefficient, while the Pearson correlation coefficient, compared to the Spearman correlation coefficient, is more stringent in its conditions. Therefore, this paper combines the Pearson correlation coefficient with the CART decision tree.

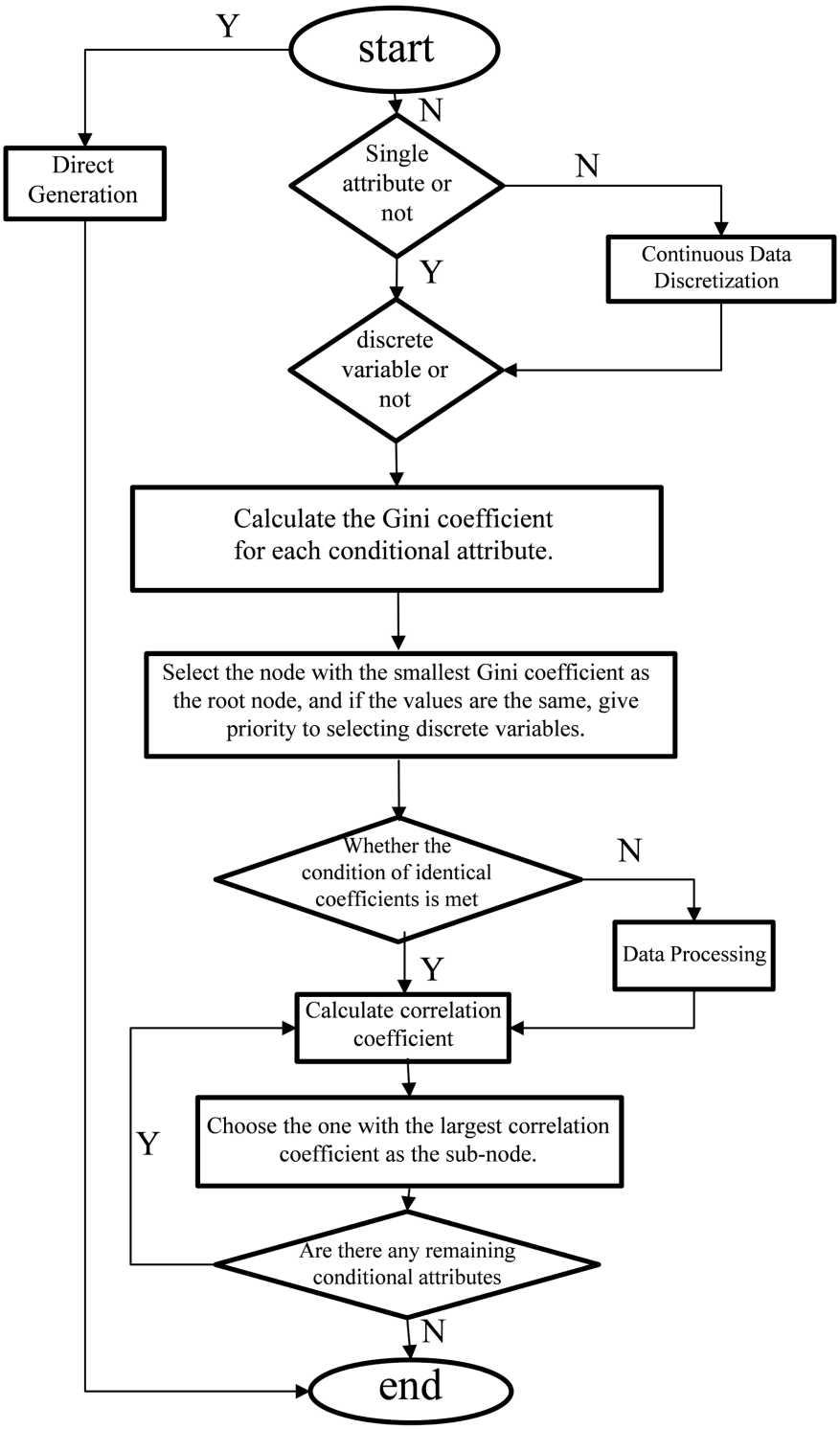

Figure 1. Correlation coefficient decision tree flowchart.

Below is an introduction to the algorithm principle and process: initially, the root node is determined by the Gini index, and the conditional attribute with the lowest Gini index is selected as the root node. Once the root node is determined, subsequent nodes are no longer sorted using the Gini index but are sorted based on the Pearson correlation coefficient. However, data processing is required to meet the necessary conditions. The advantage of this approach is that it does not excessively increase the time complexity, but using correlation coefficient ranking in the subsequent steps can lead to decision results that are more realistic and align better with actual expectations. The complete flowchart of the decision tree using correlation coefficients is shown in Figure 1. The algorithm steps after integration are provided below:

Step 1: Discretize continuous data, calculate the Gini index for each conditional attribute, and select the conditional attribute with the minimum value as the root node.

Step 2: Process the data for the remaining conditional attributes to meet the conditions of the Pearson correlation coefficient.

Step 3: Calculate the correlation coefficient between the remaining conditional attributes and the previous node, selecting the largest value as the next node.

Step 4: Repeat Step 3 until there are no more remaining conditional attributes.

In other words, the selection of the root node is still defined by the classic CART decision tree, as this method is well-established. The difference from traditional methods lies in the selection of subsequent nodes, where the original approach of minimizing the Gini coefficient has been changed to maximizing the correlation coefficient. This adjustment can further strengthen the relationships between the various attribute conditions.

The comparison between the traditional CART decision tree and the correlated-CART decision tree is shown in Figure 2, and the feasibility verification can be seen in the subsequent case studies.

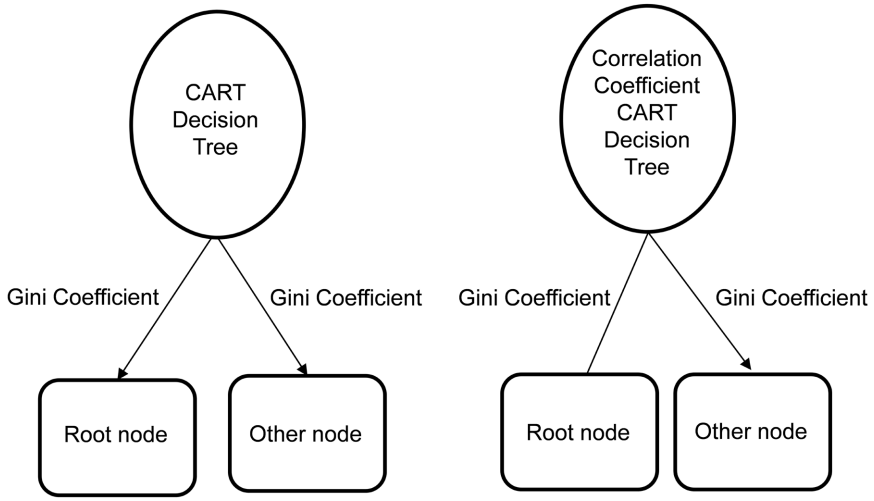

Figure 2. Comparison between CART decision tree and CART decision tree with correlation coefficient.

4. Case Study in This Experiment

Xining City's five districts and two counties are taken as examples, and relevant data is collected through online sources. Data collection methods include but are not limited to acquiring data from online maps such as Gaode Maps, Baidu Maps, Tencent Maps, Google Maps, Baidu Baike, Xining Statistical Yearbook, Qinghai Statistical Yearbook, etc. Different data sources have varying update frequencies; for example, yearbooks are updated annually, while maps will be updated quarterly. As a result, the data collected from different sources may exhibit discrepancies.

Due to the timeliness and lag of data, the study selected the most recent data for the experiment. Surveys were conducted with multiple real estate agencies in Xining City to gather information on the top 10 factors that residents are most concerned about when purchasing a home. The collected customer information consists of local residents of Xining City, aiming to improve the reliability and authenticity of housing selection in Xining City. The information for each district and county in Xining City is presented in Table 2.

Table 2. Information of districts and counties in Xining City.

Attribute / Region	Hospital a_1	School a_2	Park a_3	Market a_4	House price a_5	Shopping mall a_6	Transportation a_7	Parking a_8	Express a_9	Greening a_{10}
Urban Center	55	10	16	9	7672	25	1	89	92	40.5
East City	56	21	9	34	7850	18	4	148	90	35
West City	49	28	15	14	11,453	22	1	138	83	42.8
North City	28	22	11	45	8076	4	2	74	91	30
Huangzhong District	17	4	3	13	7243	6	1	24	27	40
Huangyuan County	10	8	3	5	6858	1	6	45	26	37.5
Datong Hui and Tu Autonomous County	12	13	4	10	7600	8	5	85	47	45.38

4.1. Data Acquisition and Experimental Validation

To verify the effectiveness of the improved decision tree model, the experiment selected the top 5 real estate agencies in terms of comprehensive strength in Xining City as the data source. A total of 1000 sets of data were collected, ranging from near too far, for the study. The conditional attributes include 10 aspects such as hospitals, schools, parks, markets, housing prices, malls, transportation, parking, express delivery, and greenery. After obtaining the data, preprocessing was done on the data in the sample set. Due to errors by employees or unforeseen circumstances resulting in some data anomalies, these data points were removed to ensure the accuracy of the experiment. There was a total of 68 sets of abnormal data and 932 sets of valid data.

Due to the varying types of information recorded in the database, there are both clear and fuzzy data types. The fuzzy data primarily consists of linguistic values, while the clear data includes both discrete and continuous types. Therefore, this paper processes the original data to obtain data that can be directly applied to the model. A screenshot of the processed data from some homebuyers is shown in Figure 3.

	A	B	C	D	E	F	G	H	I	J	K
1	*a1*	*a2*	*a3*	*a4*	*a5*	*a6*	*a7*	*a8*	*a9*	*a10*	*d*
2	34	15	8	16	9437	6	6	82	28	34	Y
3	45	26	3	10	10894	12	6	108	58	37	Y
4	16	16	7	40	9825	4	4	119	69	45	Y
5	20	4	6	23	9196	6	5	107	45	30	Y
6	38	7	6	43	10588	5	3	122	54	45	Y
7	35	12	4	31	9914	1	2	134	38	37	Y
8	12	19	7	37	10703	16	4	57	85	37	Y
9	20	18	9	24	7911	10	1	143	38	33	Y
10	53	9	12	21	10465	8	4	141	92	34	Y
11	24	4	16	15	10550	3	6	136	34	35	Y
12	32	7	16	43	9748	15	5	51	82	35	Y
13	15	6	8	18	11173	21	6	32	80	35	Y
14	53	4	14	30	7284	16	3	66	51	38	Y
15	49	6	5	20	10714	21	3	134	57	38	Y
16	39	16	6	45	9602	17	1	89	49	45	Y
17	17	16	15	45	7845	14	5	141	84	31	Y
18	55	8	9	28	11429	17	1	145	61	32	Y
19	14	6	15	26	11356	25	4	99	82	41	Y
20	50	21	10	41	8357	7	6	131	76	35	Y
21	53	10	13	13	9291	12	6	114	35	41	Y
22	49	19	5	23	8207	20	3	30	34	41	Y
23	26	4	5	21	7855	9	2	45	29	30	Y
24	24	25	12	38	8846	20	1	68	29	32	Y
25	44	24	10	40	9848	2	4	90	79	38	Y
26	49	18	3	45	9730	2	4	132	33	44	Y
27	31	13	14	43	11284	15	1	67	84	37	Y
28	18	25	14	33	9437	2	4	102	34	40	Y
29	24	12	14	11	10983	14	4	77	74	42	Y
30	49	27	15	8	8827	3	6	46	71	44	Y

Sheet1

Figure 3. Screenshot of data from some homebuyers.

The age range of the homebuyers was between 25 and 65 years old, categorized into youth (25 to 35), middle-aged (35 to 55), and elderly (55 to 65) age groups, with the proportions of each age group shown in Figure 4. The traditional CART decision tree and the improved CART decision tree proposed in this study were separately applied to this case.

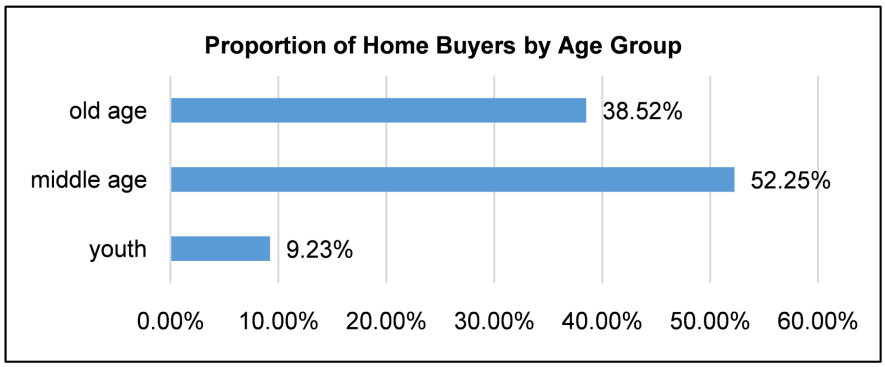

Figure 4. Proportion of home buyers by age group.

4.1.1. CART Decision Tree

The optimal binary split points for the 10 conditional attributes and their respective Gini coefficients were calculated using formulas (1) and (2) as shown in Table 3. For continuous data, the splitting points were discretized based on a binning method. The Gini coefficient values for each interval were calculated in sequence, with the minimum value chosen as the optimal splitting point.

The process for determining the root node is as follows:

Preprocess the collected raw data to convert it into a format suitable for direct calculation.

Using the obtained samples, calculate the Gini coefficient and the split point for the 10 conditional attributes based on formulas (1) and (2).

For continuous data, split the values by taking one less than the maximum value. For example, if the maximum value of a certain attribute is in the hundreds, the split point would be at 10. After dividing, continue to calculate using the Gini coefficient method and select the minimum value as the split point.

Choose the conditional attribute with the smallest Gini coefficient as the root node, using the split point as the division point for the node.

Based on the information in Table 3, it can be observed that the attribute of schools is most important to homebuyers, while the attribute of parks is the least important. The CART decision tree was constructed based on the data in Table 3, and the tree diagram is presented in Figure 5.

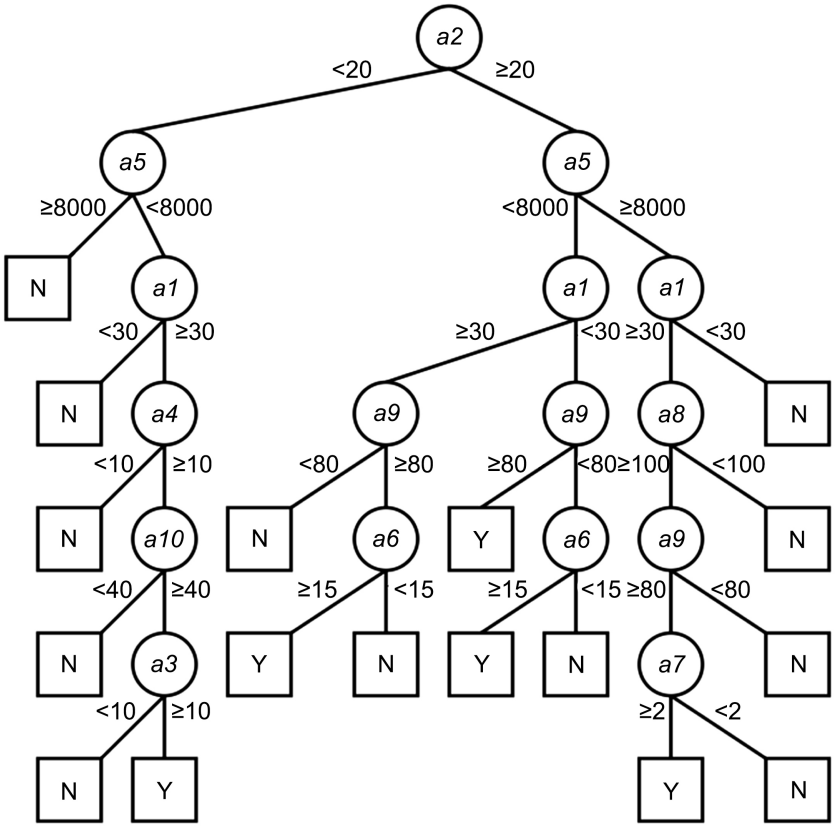

Figure 5. CART decision tree diagram.

Table 3. Gini coefficient table for various condition attributes.

Condition attribute	*Bipoint*	*Gini*
a_1	30	0.2
a_2	20	0.1
a_3	10	0.5
a_4	10	0.25
a_5	8000	0.12
a_6	15	0.35
a_7	2	0.45
a_8	100	0.3
a_9	80	0.34
a_{10}	40	0.42

4.1.2. Decision Tree Based on Correlation Coefficients

In this study, correlation coefficients were used to enhance the decision tree. The root node still selects the conditional attribute with the smallest Gini index, but the leaf nodes are selected based on the conditional attribute with the highest correlation coefficient to construct the tree. The improved CART decision tree was applied to the youth, middle-aged, and elderly groups. Initially, the root nodes for different age groups were calculated using formulas (1) and (2), and then the Pearson correlation coefficients with the previous node were calculated using formula (3). The conditional attribute with the highest absolute correlation coefficient value was chosen as the next node to build the decision tree. The decision trees for each age group are presented in Figure 6, Figure 7, and Figure 8.

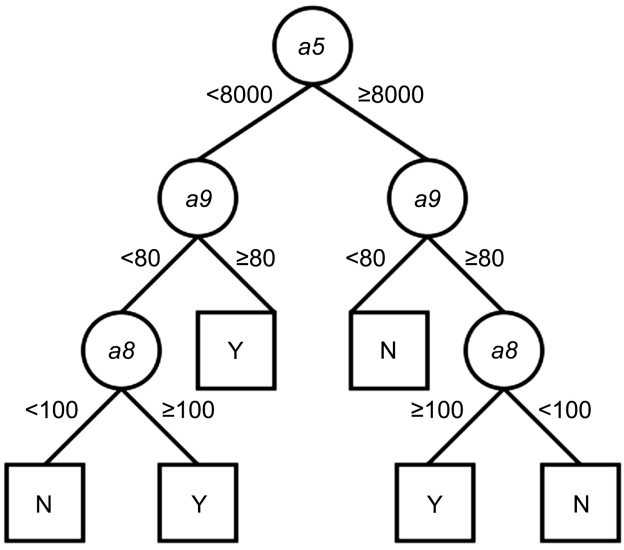

Figure 6. Decision tree for young adults.

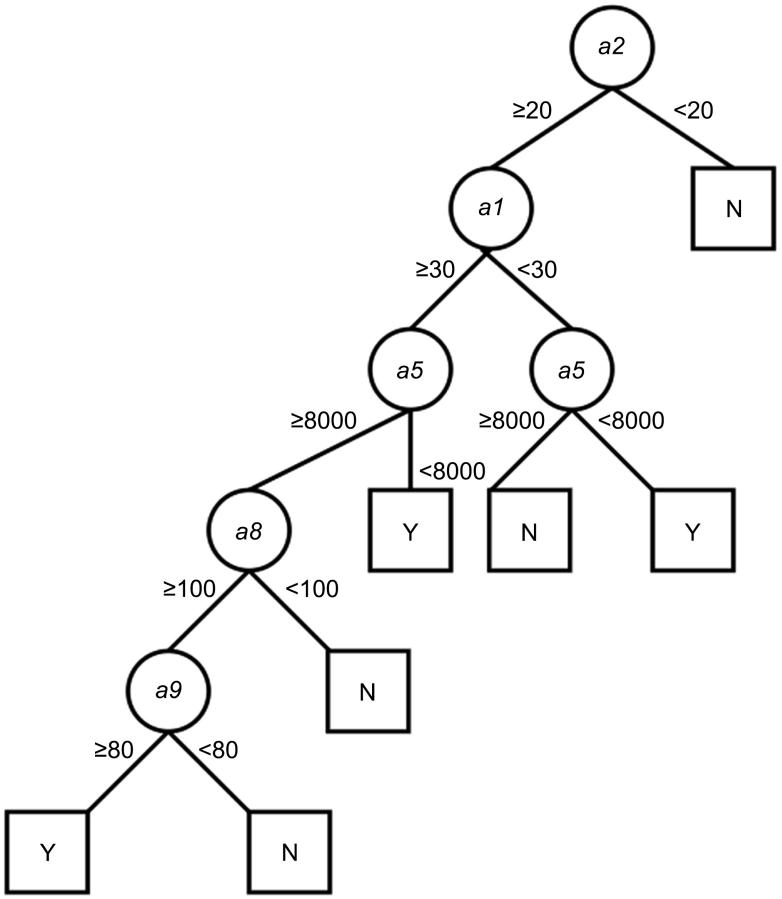

Figure 7. Decision tree for middle-aged adults.

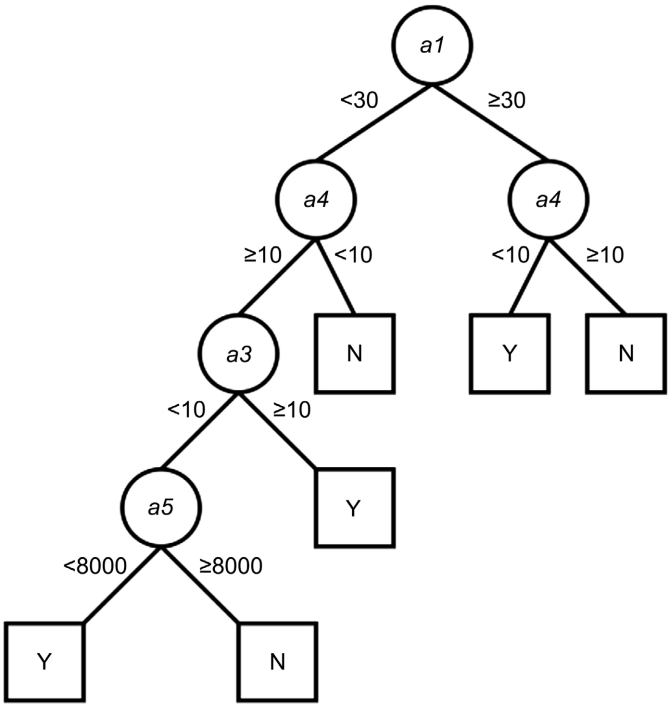

Figure 8. Decision tree of older group.

The process for determining the leaf nodes is as follows:

After determining the root node, calculate the correlation coefficients of the remaining conditional attributes with the root node using formula (3), and select the highest value as the leaf node for that level.

Repeat the above process, computing the correlation coefficients of the remaining conditional attributes with the previous node, and take the maximum value as the leaf node for the current level.

By comparing the traditional CART decision tree with the correlation coefficient-based CART decision tree, it was found that for homebuyers of different ages, the results of root node selection varied, leading to changes in subsequent leaf node selection. By altering the selection of leaf nodes, some conditional attributes were excluded from the decision tree construction, reducing the number of leaf nodes, decreasing the height of the tree, shortening the time complexity, avoiding pruning issues, and improving the efficiency of the model.

4.2. Extraction of Decision Rules and Recommendations

Extraction of rules was performed on the 932 sets of valid data in the sample set, and the decision rules were described using the *if/then* principle. The extracted decision rules are presented in Table 4.

Table 4. Decision rules.

	If	*Then*
R_1	$a_3 \geq 10$	Purchase
R_2	$a_6 \geq 15$	Purchase
R_3	$a_7 \geq 2$	Purchase

Among the 932 sets of samples, there were a total of 765 homebuyers and 167 non-homebuyers, as shown in the proportion in Figure 9. The proportion of homebuyers in each district can be seen in Figure 10. To ensure the effectiveness of the recommendations, a selection was made from the 765 sets of samples of homebuyers, and recommendations for homebuying were provided to individuals of different age groups based on the decision rules.

Based on the decision trees for the three age groups, it can be observed that the youth group prioritizes attribute a_5 as the main factor for homebuying, possibly due to being unmarried, without children, and having limited savings, and therefore it is advised to buy a home in the city center district. The middle-aged group prioritizes attribute a_2, possibly due to having some savings and children needing education, and therefore it is recommended to buy a home in the western district of the city. The elderly group prioritizes attribute a_1, potentially due to not having children in school, having some savings, and having more health issues, and therefore it is suggested to buy a home in the eastern district of the city.

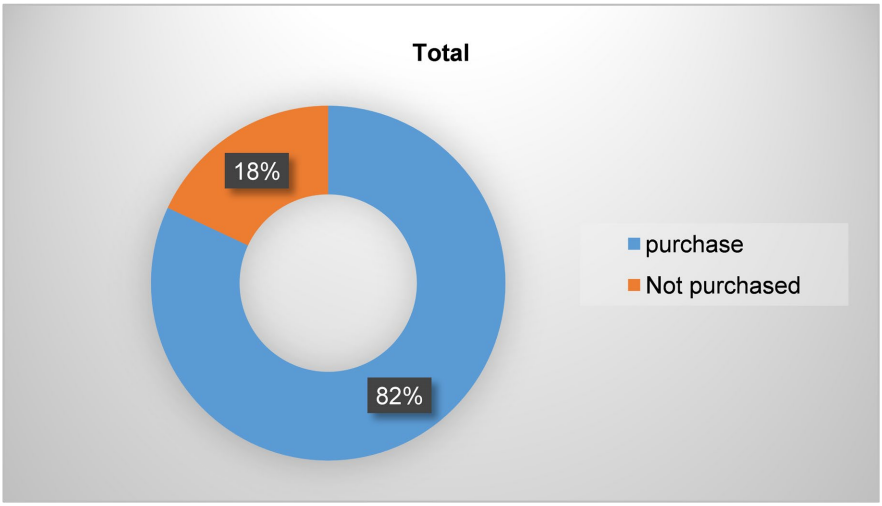

Figure 9. Home purchase status.

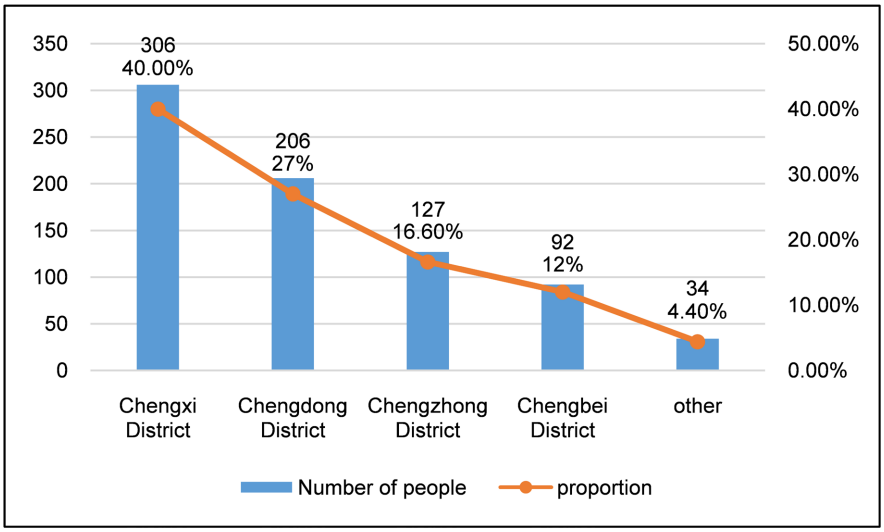

Figure 10. Distribution of population proportion in each city district.

5. Conclusions

In conclusion, this study optimized the algorithm to a certain extent by combining the Pearson correlation coefficient with the CART decision tree. A comparison showed a reduction in the depth of the decision tree, a decrease in the number of leaf nodes, a shorter time complexity, increased efficiency, and the avoidance of pruning issues. The improved decision tree algorithm was then applied to the housing purchase issue in Xining City, providing consumer recommendations through experimental validation.

Through the enhanced CART decision tree algorithm application in addressing housing issues in Xining City, the aim was to provide recommendations to consumers based on existing data, with the hope of assisting individuals with future homebuying needs. In future research, the complexity can be further reduced, and accuracy can be improved from the perspective of attribute reduction.

Funding

Innovation Project Number: Qinghai Minzu University 2023 Graduate Innovation Project (No. 07M2023008).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Breiman, L., Friendman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression trees. Wadsworth.

[2] Quinlan, J.R. (1986) Induction of Decision Tree. *Machine Learning*, **1**, 81-106. https://doi.org/10.1007/BF00116251

[3] Quinlan, J.R. (1993) C 4.5: Program for Machine Learning. Morgan Kaufmann Publishers, 21-31.

[4] Tang, M.A., Wang, X.M., Cao, J., *et al.* (2015) GIS Analysis and Entropy-Based Attribute Reduction for Parking Garage Site Selection Decision Making. *Systems Engineering-Theory & Practice*, **35**, 175-182.

[5] Dong, Y.H. and Liu, L. (2015) Optimization Algorithm of Decision Tree Based on Correlation Coefficient. *Computer Engineering and Science*, **37**, 1783-1793.

[6] Wu, S.B., Chen, Z.G. and Huang, R. (2016) ID3 Optimization Algorithm Based on Correlation Coefficient. *Computer Engineering and Science*, **38**, 2342-2347.

[7] Tang, J.Y., Yang, Z.Q. and Lu, J.X. (2019) Evaluation of Geospatial Suitability for Village Relocation Site Selection in Coal Mining Subsidence Areas in the Loess Plateau. *Journal of Xi'an University of Science and Technology*, **39**, 334-340.

[8] Lin, Z.X., Liu, Z.R. and Ji, J. (2020) Attribute Reduction Based on k-Nearest Neighbor Attribute Importance and Correlation Coefficient. *Computer Engineering and Design*, **41**, 2488-2494.

[9] Yan, Q. (2021) Research on Differential Privacy Decision Tree Method Based on Pearson Correlation Coefficient. Master's Thesis, Guangxi Normal University.

[10] Sun, C.H., Wang, J., Yang, F., *et al.* (2021) Study and Implementation of ID3 Decision Tree in Predicting Aluminum Output in Electrolytic Cells. *Light Metals*, No. 8, 59-62.

[11] Xie, X., Zhang, X.Y. and Yang, J.L. (2022) Decision Tree Algorithm Integrating Information Gain and Gini Index. *Computer Engineering and Applications*, **58**, 139-144.

[12] Wu, T., Wang, Z.H., Chen, Q., *et al.* (2023) Simulation of Aircraft Wing Icing Risk Monitoring Based on C4.5 Decision Tree. *Computer Simulation*, **40**, 44-48.

[13] Ma, D.H., Luo, L., Wang, W., *et al.* (2023) Extraction of Bridge Damage State Decision Association Rules Based on DRSA and CART. *Journal of Beijing University of Technology*, **49**, 1167-1179.

[14] Pan, Z.S., Ma, K.S., Long, Y., *et al.* (2024) Study on Oak Branch and Leaf Point Cloud Classification with Improved Classification and Regression Tree Model. *Journal of Nanjing Forestry University* (*Natural Sciences Edition*), **48**, 123-131.

[15] Wang, Y.H., Zhao, Y.J. and Liu, W.X. (2024) Application of Data Mining Algorithm in Job Workshop Scheduling Problem. *Computer Integrated Manufacturing Systems*, **30**, 520-536.