computer sciences and mathematics forum

MDPI

Proceeding Paper

Research on a Streamlined Causal Tree Algorithm Based on Factor Space Theory †

Kaile Lin [1,2,*], Fanhui Zeng [1,2,*], Xiaotong Liu [1,2], Ying Wang [1,2] and Kaijie Zhang [1,2]

[1] College of Science, Liaoning Technical University, Fuxin 123000, China; liu_xiaotong2299@163.com (X.L.); 15309877632@163.com (Y.W.); 19824855218@163.com (K.Z.)
[2] Institute of Intelligent Engineering and Math, Liaoning Technical University, Fuxin 123000, China
* Correspondence: 15525501190@163.com (K.L.); zfh3351@sina.com (F.Z.); Tel.: +86-155-2550-1190 (K.L.); +86-139-4183-3880 (F.Z.)
† Presented at the 2023 Summit of the International Society for the Study of Information (IS4SI 2023), Beijing, China, 14–16 August 2023.

Abstract: Decision rule extraction is an important tool for artificial intelligence and data mining, but decision rule redundancy reduces the generalization ability of causal trees. In order to better reduce the size of causal trees and improve the classification accuracy, based on factor space theory and aiming at the elimination of noise and special samples in the dataset using the extension decision degree criteria, the conditional factor corresponding to the optimal extension decision degree is used as the branch node of the tree, and the abnormal state object is removed from the conditional factor, recurring to obtain the streamlined causal tree algorithm. Comparison with other classification algorithms shows that the streamlined causal tree algorithm produces the smallest causal tree size, the least redundant rules, and the best classification accuracy.

Keywords: factor space; extension decision degree; streamlined causal tree algorithm; α threshold; C4.5

Citation: Lin, K.; Zeng, F.; Liu, X.; Wang, Y.; Zhang, K. Research on a Streamlined Causal Tree Algorithm Based on Factor Space Theory. *Comput. Sci. Math. Forum* **2023**, *8*, 72. https://doi.org/10.3390/cmsf2023008072

Academic Editors: Zhongzhi Shi and Wolfgang Hofkirchner

Published: 16 August 2023

1. Introduction

In 1982, Wang Peizhuang [1] proposed the idea of factor space from the origin of object cognition and based on it, established the mathematical theory of knowledge representation—factor space theory—which is the earliest basic theory of artificial intelligence in international intelligence mathematics. In 2014, Wang Peizhuang [2] et al. Proceeded with the rapid extraction of causal rules based on the logical nature of reasoning and proposed the factor analysis method, which is one of the core algorithms in factor space and provides important tools for artificial intelligence and data mining. Bao Yanke [3] et al. proposed a subtraction and rotation calculation to improve the utilization of factor analysis methods in the training set sample information. Liu Haitao [4] et al. provided a reasoning model for the factor analysis method, which solved the problem of object recognition caused by incomplete training set samples and improved the accuracy of the factor analysis method. Wang Huadong [5] adopts a column-by-column advancement method when selecting factors for superposition division to improve the accuracy and running speed of the factor analysis method.

However, current literature studies have not significantly reduced the size of the causal tree in the factor analysis method. The main method to reduce the size of the causal tree is pruning [6]. Current literature research shows that pruning can reduce the size of causal trees to a certain extent, but the resulting causal trees are not streamlined. The size of the causal tree reflects the generalization ability of the tree to a certain extent. The more complex the rules extracted from the dataset, the larger the size of the tree. Rule redundancy will lead to overfitting and weaken the generalization ability. It is particularly important to minimize the size of the causal tree without affecting classification accuracy. Therefore,

this paper proposes a streamlined causal tree algorithm where, by using a self-defined threshold, the noise samples in the training set are filtered out and the optimized causal tree is trained in the same step, thereby greatly reducing the size of the causal tree and improving its classification performance. In addition, the deletion of the determining region is a key factor in reducing the computational complexity of the algorithm and achieving fast convergence. The streamlined causal tree algorithm can find a larger determining region, enabling the algorithm to converge faster under the optimal threshold.

2. Basic Knowledge

Factor is a key to describing everything and can be understood as a generalized gene. From a mathematical perspective, factors are a special mapping that maps objects onto their phases. The basic theories related to factor space [7] are as follows:

Factors influence each other, restrict each other, and cause and affect each other. In the factor analysis method, the factor g that is concerned is called the result factor, and those factors $\{f_1, f_2, \cdots f_n\}$ that have an influence on it are called conditional factors.

The causal analysis table takes the object as the row and the conditional and result factors as the columns, as shown in Table 1. The i-th row and j-th column elements in Table 1 represent the state of the i-th object under the j-th factor.

Table 1. Causal analysis table.

U	f_1	f_2	$F \to g$ $\cdots$	f_n	g
u_1					
u_2			$f_j(u_i)$		$g(u_i)$
$\vdots$					
u_m					

Each row of the causal analysis table is the coordinate of an object in the factor space. A finite number of objects constitute a domain $U = \{u_1, u_2, \cdots, u_m\}$. The conditional factors are $F = \{f_1, f_2, \cdots, f_n\}$. The state space of conditional factors is $I(f_j) = \{a_{j1}, a_{j2}, \cdots, a_{jk}\} (j = 1, 2, \cdots, n)$. The result factor is g. The state space of the result factor is $I(g) = \{g_1, g_2, \cdots, g_s\}$.

Definition 1. *Given a conditional factor f_j and the state a_t taken by that factor, remember $[a_t] = \{u_i | f_j(u_i) = a_t\}$, if all objects in $[a_t]$ have the same result, (there is a state or level where the result factor g exists such that $[g_l] = \{u_i | g(u_i) = g_l\} \supseteq [a_t]$), it is said to be a determining class $[a_t]$ of factor f_j. The union of all determining classes of factor f_j is called the determining region for the result factor. The ratio of the number of rows h in the determining region of the factor f_j to the number of rows in the table (i.e., the number of all objects) m is called its determining degree on the result factor g, denoted as $d(f_j) = h/m$.*

Definition 2. *If the class $[a_t]$ of conditional factors f_j is a determining class and all objects in the class $[a_t]$ have a unique and definite result, then it is called "if f_j is a_t, then the result g is g_l". This sentence is a reasoning sentence determined by conditional factors f_j, denoted as $f_j = a_t \to g = g_l$.*

3. The Streamlined Causal Tree Algorithm

The factor analysis method in the factor space can quickly and concisely analyze the causal relationships contained in the dataset, establish causal rules, and obtain a causal tree. However, when using the factor analysis method to train causal trees, when there are too many conditional factors in the dataset or when there are many states of conditional factors, the trained causal tree rules are redundant, and the prediction effect is poor. Since the calculation principle of determining degree is too absolute, noisy object data and special object data generated due to input errors, measurement equipment failures, and other reasons in the dataset will have a significant negative impact on the training of the factor analysis method. This means that it cannot cope with noisy data, has poor robustness, and the decision effectiveness of factors cannot be fully utilized, thus limiting the application

of this algorithm. Even if pre- and post-pruning are used for the trained causal tree, the negative impact is inevitable.

In order to solve the interference of noisy data, improve the robustness of causal tree algorithms, reduce the size of causal trees, and improve the accuracy of classification prediction, a streamlined causal tree algorithm is proposed.

3.1. Algorithm Principle

The purpose of factor analysis in factor space is to transform a table into a set of inference sentences (decision rules). Since the determining class is contained by the result class, an inference sentence is formed from the determining class to the result class containing it, and finally a rule causal tree is obtained from the conditional factor to the result factor.

3.1.1. Theoretical Knowledge

Definition 3. *(extended determining class) Given a conditional factor f_j and a state $[a_t]$ taken by that factor, remember $[a_t] = \{u_i | f_j(u_i) = a_t, u_i \in U\}$. All objects in $[a_t]$ have $[g_l] = \{u_i | g(u_i) = g_l,\ u_i \in [a_t]\} \subseteq [a_t]$, $l = 1, 2, \cdots, s$ for all states $\{g_1, g_2, \cdots, g_s\}$ of the result factor g. Given a α threshold ($\alpha \in (0.5, 1]$), if $\frac{|[g_l]|}{|[a_t]|} > \alpha$, it is said $[a_t]$ is an extended determining class of factor f_j. The union of all extended determining classes of factor f_j is called the extended determining region of the result factor g.*

Definition 4. *(extended determining degree) The ratio of the number of objects q in the extended determining region of the factor f_j to the number of all objects m is called the extended determining of the result factor, denoted as $d(f_j) = q/m$.*

3.1.2. Algorithm Principle

The extended determining degree criterion, which adopts the extended determining degree with the essence of reasoning set logic as the criterion, selects the optimal conditional factors and achieves fast convergence of the algorithm by expanding the determining region.

3.2. Setting of the α Threshold

If the α threshold is too low, during the training process, the conditions are easy to meet, which will delete too many non-noise objects and special objects, easily leading to underfitting, resulting in a single decision tree rule and loss of decision value.

If the α threshold is too high, during the training process, the conditions are difficult to satisfy, which is not enough to delete noisy objects and special objects and cannot achieve the purpose of optimizing the training set and reducing the size of the causal tree.

Through experiments, it was found that the α threshold range is generally 0.8~0.95.

3.3. Algorithm Steps

Domain $U = \{u_1, u_2, \cdots, u_m\}$, the conditional factor is $F = \{f_1, f_2, \cdots, f_n\}$, and the state space of the conditional factor is $I(f_j) = \left\{a_{j1}, a_{j2}, \cdots, a_{jk}\right\} (j = 1, 2, \cdots, n)$; the result factor is g, and the state space of the result factor is $I(g) = \{g_1, g_2, \cdots, g_s\}$. The steps for a streamlined causal tree algorithm are:

Input: Dataset.

Step 1. Divide the dataset into Train_data and Test_data. Given the initial α threshold.

Step 2. Calculate $q(a_{jt})$ and $q(a_{jt}, g_l)$. Traverse all conditional factors, calculate the number of objects $q(a_{jt}) (j = 1, 2, \cdots, n;\ t = 1, 2, \cdots, k)$ corresponding to all states of the conditional factor f_j. Calculate the number of objects $q(a_{jt}, g_l)$ $(j = 1, 2, \cdots, n;\ t = 1, 2, \cdots, k;\ l = 1, 2, \cdots, s)$ in all states of the conditional factor f_j corresponding to the result factor.

Step 3. Calculate the ratio r of $q(a_{jt}, g_l)$ to $q(a_{jt})$.

Step 4. Determine the extended determining class. Compare $r_{a_{jt,1}}, r_{a_{jt,2}}, \cdots, r_{a_{jt,s}}$ under the same state of the conditional factor f_j with α. If $r_{a_{jt,l}} > \alpha$, then all objects whose

state a_{jt} of the conditional factor f_j corresponds to the result factor state g_l are extended determining classes.

Step 5. Determine the extended determining region. Union of the extended determining classes of each conditional factor to obtain the extended determining region.

Step 6. Calculate the extended determining degree. Calculate the number of objects in the extended determining region of each conditional factor and obtain the extended determining degree $d = \{d_1, d_2, \cdots, d_n\}$. Calculate the maximum extended determining degree $d_{\max} = \max\{d_1, d_2, \cdots, d_n\}$.

Step 7. Update the training set. Suppose that the conditional factor corresponds to $d_{\max}$ is f_j, if there are $q(a_{jt}, g_l)$ extended determining classes in a certain state of the conditional factor f_j, label all objects in the extended determining class as normal objects. At the same time, in this state, there are $Q = q(a_{jt}, g_1) + q(a_{jt}, g_2) + \cdots + q(a_{jt}, g_{l-1}) + q(a_{jt}, g_{l+1}) + \cdots + q(a_{jt}, g_s)$ objects that are not extended determining classes and are marked as abnormal objects, namely noise objects and special objects to be deleted. In the training set Train_data, objects marked abnormal were deleted to obtain a new training set Train_data1.

Step 8. Extraction rules. For Train_data1, it uses $d_{\max}$ corresponding to conditional factors to extract decision rules and divides the dataset to obtain sub-datasets.

Step 9. Building a causal tree. Repeat steps 2 to 8 on the sub-dataset to construct a causal tree under the α threshold. Each node of the causal tree satisfies the condition α, and each branch is carried out on the updated training set under the condition α.

Step 10. Select the optimal α threshold. Given a step size $step = 0.01$, repeat steps 2 to 9. Analyze the relationship between α threshold and the accuracy of causal tree prediction and select the optimal α threshold on the training set.

Output: The causal tree under the optimal α threshold.

3.4. Instance Analysis

Five classification datasets in the UCI database were analyzed using the streamlined causal tree algorithm, the factor analysis method, the ID3 algorithm, and the C4.5 algorithm. Tenfold cross-validation was used to obtain the number of decision rules, accuracy, precision, recall, F1-measure, and running time. The running time of the streamlined causal tree algorithm is the causal tree training time under the optimal threshold. The experimental results are shown in Table 2.

Table 2. Comparison of experimental results.

Datasets	Indexes	Factor Analysis Method	Streamlined Causal Tree Algorithm	ID3	C4.5-PEP
Lymphography	Number of decision rules	45	24	50	28
	Accuracy	0.7614	0.8514	0.7438	0.7567
	Precision	0.7822	0.8781	0.8106	0.8074
	Recall	0.7614	0.8514	0.7438	0.7567
	F1	0.7619	0.853	0.7549	0.7627
	Time/ms	55	40	57	74
Dermatology	Number of decision rules	98	26	125	31
	Accuracy	0.7593	0.9167	0.7011	0.9134
	Precision	0.8238	0.9525	0.8073	0.9487
	Recall	0.7593	0.9167	0.7011	0.9134
	F1	0.7776	0.929	0.7359	0.9224
	Time/ms	201	133	216	358

Table 2. *Cont.*

Datasets	Indexes	Factor Analysis Method	Streamlined Causal Tree Algorithm	ID3	C4.5-PEP
Cancer	Number of decision rules	86	34	85	75
	Accuracy	0.9312	0.9634	0.9224	0.9313
	Precision	0.9394	0.9608	0.9326	0.9389
	Recall	0.8622	0.9297	0.8413	0.8637
	F1	0.895	0.9441	0.8838	0.8983
	Time/ms	90	50	108	132
Australian	Number of decision rules	263	38	222	160
	Accuracy	0.7348	0.8768	0.7739	0.8116
	Precision	0.7795	0.922	0.8045	0.8506
	Recall	0.7236	0.854	0.7834	0.8077
	F1	0.7479	0.884	0.7917	0.8259
	Time/ms	209	90	203	274
Tic-tac-toe	Number of decision rules	271	76	190	122
	Accuracy	0.7828	0.8487	0.8476	0.7975
	Precision	0.8367	0.8481	0.8939	0.8378
	Recall	0.8277	0.9395	0.8687	0.8588
	F1	0.8306	0.8903	0.8805	0.847
	Time/ms	240	120	193	257

3.5. Conclusions

The causal tree trained by the streamlined causal tree algorithm has the smallest size, the best classification accuracy and F1-measure, and the least redundant rules. Therefore, the streamlined causal tree algorithm can not only reduce rule redundancy and significantly reduce the size of the causal tree but also improve the classification performance of the causal tree to a certain extent, expanding the theory and application of factor space.

Author Contributions: There are five authors in this paper. K.L. provided the algorithm and software coding for verification analysis and writing of the paper; F.Z. provided the guidance for writing and preparing the first draft; X.L., K.Z. and Y.W. reviewed and edited the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Liaoning Provincial Department of Education Project, grant number LJ2019JL019, and the Key Research Projects of Basic Scientific Research Projects in Higher Education Institutions of the Liaoning Provincial Department of Education, grant number LJKZZ20220047.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used in this article is all from the UCI dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, P.Z.; Sugeno, M. Background structure of factor field and Fuzzy Set. *Fuzzy Math.* **1982**, *2*, 45–54.
2. Wang, P.Z.; Guo, S.C.; Bao, Y.K.; Liu, H.T. Factor Analysis Method in Factor Space. *J. Liaoning Tech. Univ. (Nat. Sci.)* **2014**, *33*, 865–870.
3. Bao, Y.K.; Ru, H.Y.; Jin, S.J. A new algorithm of knowledge mining in factor space. *J. Liaoning Tech. Univ. (Nat. Sci.)* **2014**, *33*, 1141–1144.
4. Liu, H.T.; Guo, S.C. Reasoning model of factor analysis method. *J. Liaoning Tech. Univ. (Nat. Sci.)* **2015**, *34*, 124–128.
5. Wang, H.D.; Wang, P.Z.; Guo, S.C. Improved factor analysis on factor spaces. *J. Liaoning Tech. University. Nat. Sci.* **2015**, *34*, 539–544.

6. Fan, S.B.; Zhang, Z.J.; Huang, J. Association Rule Classification Method for Decision Tree Pruning Strengthening. *Comput. Eng. Appl.* **2023**, *59*, 87–94.
7. Wang, P.Z.; Liu, H.T. *Factor Space and Artificial Intelligence*, 1st ed.; Beijing University of Posts and Telecommunications Press: Beijing, China, 2021; pp. 89–92.