Design and Development of Optimal Causal Probability Decision Tree by ComputingPath probability of Internal Causality nodes

[1]S.Sajida, [2]Dr. K.Vijaya Lakshmi

[1]Research Scholar, ,[2]Assistant Professor
[1,2] Department of Computer Science
[1,2]Sri Venkateswara University, Tirupati, AP, India.

ABSTRACT: Data becomes the driving force of the modern world, almost everyone has come across terms like data science, machine learning, artificial intelligence, and data mining. A tree has many real-world analogies, and it turns out that it has influenced a broad area of machine learning, including classification and regression. A decision tree can be used in decision analysis to visually and explicitly represent to make decisions. Though it is a common tool in data mining for developing a strategy to achieve a specific goal, it is also widely used in machine learning, which will be the primary focus in this research paper. Since the trees are generated with a cause and effect relationship, the decision tree's consequence is a Causal probability decision tree. The author proposed a metric for evaluating the Finest Causal Probability Decision Trees by Computing Path Probability of Internal Causality Nodes.

Keywords: Decision Tree, Optimal Probability, Correlation, Causal inference internal node Causality,path probability,path scores.

1. INTRODUCTION:

In the real world, we are surrounded by humans who have the ability to learn from their experiences, as well as computers or machines that work on our instructions. Machine learning algorithms establish a mathematical model that assist in making predictions or decisions without being explicitly programmed using sample historical data known as training data. Machine learning aims to combine computer science and statistics to develop predictive models and strengthens or implements algorithms that learn from historical data. At a broad level, it can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

Supervised learning is a type of machine learning in which machines are trained using well-labeled training data and then predict the output based on that data. Labeled data indicates that some input data has already been tagged with the appropriate output. In supervised learning, the training data provided to the machines acts as a supervisor, teaching the machines how to correctly predict the output. It employs the same concept that a student would learn under the supervision of a teacherA variety of business applications, including the following, can be built and advanced using supervised learning models:Supervised learning algorithms can be used to locate, isolate, and categorise objects in videos or images, making them useful when applied to various computer vision techniques and imagery analysis.

Predictive analytics: Creating predictive analytics systems to provide deep insights into various business data points is a common use case for supervised learning models. This enables enterprises to forecast specific outcomes based on a given output variable, assisting business leaders in justifying decisions or pivoting for the benefit of the organisation..

Supervised learning is the process of providing correct input and output data to a machine learning model. A supervised learning algorithm's goal is to find a mapping function that maps the input variable (x) to the output variable (y). Supervised learning can be grouped further in two categories of algorithms:

- Classification
- Regression

Unsupervised learning is a type of machine learning in which the machine learns from data without the need for external supervision. Unsupervised models can be trained on unlabeled datasets that have not been classified. They can be classified into two types:

- Clustering
- Association

In reinforcement learning, an agent interacts with its environment by performing actions and learning through feedback. Like Q-Learning algorithm.

Classification:

The Classification algorithm is a Supervised Learning technique that uses training data to identify the category of new observations. A programme in Classification learns from a given dataset or observations and then classifies new observations into one of several classes or groups. For example, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, and so on. Classes can also be referred to as targets/labels or categories.

The Classification algorithm is a Supervised Learning technique that uses training data to categorise new observations. In classification, a programme learns how to categorise new observations into various classes or groups by using the dataset or observations provided. For example, 0 or 1, red or blue, yes or no, spam or not spam, and so on. Classes can be described using targets, labels, or categories. Because it is a supervised learning technique with input and output data, the Classification algorithm employs labelled input data. In the classification process, a discrete output function (y) is assigned to an input variable (x).

In this paper, the author is computing the best Causal Probability decision trees. Classification methods are commonly used to build decision trees, but they have limitations in terms of decision tree height and time complexities. Causal Decision trees were introduced to overcome the aforementioned limitations. Pruning can improve a tree's performance even more. It entails removing branches that make use of low-importance features. We reduce the complexity of the tree to, increase its predictive power by reducing overfitting.Pruning can begin at the root or at the leaves. The most basic pruning method begins with leaves and removes each node with the most popular class in that leaf; this change is kept if it does not degrade accuracy. It's also known as Reduced Error Pruning.However, these methods fail to identify a combined causeA straightforward approach to discovering a combined cause is to include both individual and combined variables in the causal discovery, but this scheme is computationally infeasible due to the large number of combined variables.

2. Related work:

Jiuyong Li ,et al[1] investigated how classification methods fail to account for other variables when attempting to establish causal relationships between input variables and outcome variables. The authors argued that classification methods are not intended for discovering causal relationships, and they proposed a new scalable, automated causal decision tree framework model based on a special statistic-based causal relationship framework for discovering true causal relationships in large data sets. The proposed new technique can also be used for big data applications.

Sajida.S, et al[4] has proposed various metrics for the identifying of Optimum Causal Probability Decision Trees. The author discovered combined Causal relationships in this paper by employing constraint-based Causal relationship discovery. The author proposes a new method for constructing a Causal Probability Decision Tree.

Jiuyong Li ,et al[6] proposed that, The collection of observational data has increased dramatically in recent decades. As a result, finding causal relationships directly from data is preferable. The Causal Bayesian Network (CBN) theory has made significant progress in the field of discovering causal relationships. CBN applications, on the other hand, are severely limited due to their high computational complexity. In another direction, the author proposed an association rule mining has been demonstrated to be an effective data mining method for discovering relationships.

Swati Hira & P. S. Deshpande [5] proposed that,A mechanical, biological, or social-economic system is made up of independent components. These components influence one another to keep their activity going in order for a system to exist and achieve its goal. When a component is significantly changed or removed, the system's behaviour changes. This motivates us to investigate the reason or cause of a fault and identify the cause parameters in explaining the interactions of system or process components. The causal discovery indicates not only that the indicators are correlated, but also how changing one of the cause variables is expected to cause a change in one of the effect variables.

3.Design and Development of Optimal Causal Probability Decision Tree by computing path probability of internal Causality nodes

The causality decision tree has two variables one is **predicted variable** and another one is **outcome variable**. The predicted variable is chosen by using Partial association test which is classical stastical test. So for each branch the causality of internal nodes are generated. In order to find the best Causal Probability Decision trees the author has proposed one metric.

Avg of

$$\sum_{i=1}^{n} \text{int node causality} * \text{its leafnode size/total tuples}$$

The contributions of this paper are listed as below by proposing a metric:-

In each tree path, complete path probability is multiplied with the sum of causalities of all internal nodes of the path and then finally tree aggregated causality score is computed by averaging all these path scores of the tree. The best Causal Probability Decision Tree is preferred based upon the highest tree score. In the paper Sajida et.al[4] the author has generated the Causal Probability Decision Trees .By considering that trees it has certain internal causality values as follows.

Causality details of internals nodes :

in this paper the author has taken ADULT DATA SET from UCI machinelearning.There are 14 attributes in the ADULT DATASET.These attributes are named as A,B,C,D,E,F,G,H,I,J,K,L,M,Y.Causality values of each node is computed with the help of Stastical test that is Partial Association Mantel Haszel test and constructed a Causal Probability Decision Trees.

A=316.8423269

F=115.926822

J=271.012342

and

H=25.73002343.

On the above nodes,The proposed metric was implemented and below mentioned observations are tabulated:

ranch number	Internals nodes	Causality Value/Values of internal nodes	Sum of Causality values ∑i=1 to 5
branch-1	A	316.8423269	316.8423269

branch-2	{A,F,J,H}	{316.8423269, 115.926822, 271.012342, 25.73002343}	729.5115143
branch-3	{A,F,J,H}	{316.8423269, 115.926822, 271.012342, 25.73002343}	729.5115143
branch-4	{A,F,J}	{316.8423269, 115.926822, 271.012342}	703.7814908
branch-5	{A,F}	{316.8423269, 115.926822}	432.7691489

Table-1 :To find the Causality of node/nodes and its summation

Sizes of leaf nodes 1, 2, 3, 4 and 5 respectively are 13044, 2999, 9736, 5217 and 14226

To calculate Branch scores of each branch:

(∑i=1 to 5) *(Leaf node size)/ Total number of tuples in the dataset.

Branch Wise Score	**Final Causality Value**
Branch-1 score	(∑1 * 13044)/45222 = 91.39116606.
Branch-2 score	(∑2* 2999) /45222 = 48.37921877
Branch-3 score	(∑3* 9736)/45222 = 157.0590443
Branch-4 score	(∑4*5217)/45222 = 81.19119096
Branch-5 score	(∑5*1426)/45222 = 136.1411241

Table-2:To Compute Branch Score Causality values

Similarly, for Tree-2 (which is generated by changing a correlation value), we compute the internal node causality and leaf node size, as well as the branch scores for each branch. Then, by comparing the Causality of internal nodes in both trees, the Optimal Causal Probability Decision Tree with the highest value is chosen.

Avg of

$$\sum_{i=1}^{n} \text{int node causality} * \text{its leafnode size/total tuples}$$

Branch	Internal node causality value Tree-1	Internal node causality value Tree-2
Branch-1 score	91.39116606	91.39116606
Branch-2 score	48.37921877	8.987584475
Branch-3 score	157.0590443	201.7985818
Branch-4 score	81.19119096	81.19119096
Branch-5 score	81.19119096	136.1411241
Tree average score	128.5404	**129.8774**

Table -1 Comparison of Tree average scores of two CPDT's

By comparing the two trees average causality scores, the highest average scored trees will be the Optimum Causal Probability Decision Tree. Average tree score for the entire Tree-2 is the highest and it is selected as the best Causal Probability Decision Tree between the generated trees.

CONCLUSION:

If some classes dominate, decision tree learners produce biased trees. It is therefore recommended that the data set be balanced before fitting with the decision tree. Decision-tree learners can produce overly complex trees that do not generalise well to new data. This is known as over fitting. The discovery of causal relations indicates not only that the indicators are correlated, but also how changing a cause variable is expected to cause a change in an effect variable. So there is a necessity of building Causal Probability Decision Trees in order to get scalable and interpretable data .The various Causal Proposed Decision Trees are generated ,and hence the best Causal Probability Decision trees is chosen with the proposed metric.

References

1. Jiuyong Li, Saisai Ma, Thuc Duy Le, Lin Liu and Jixue Liu "Causal Decision Trees," arXiv: 1508.03812v1 [cs.AI] 16 Aug 2015]
2. Jiuyong Li, et al [Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma", From Observational Studies to Causal Rule Mining". ACM Trans. Intell. Syst. Technol.2015 3. Yeying Zhu, Donna L. Coffman and Debashis Ghosh, "A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments", J. Causal Infer. 2015; 3(1): 25–40.
3. Donald B. Rubin [Donald B. Rubin, "Estimating Causal Effects from Large Data Sets Using Propensity Scores" 15 October 1997 | Volume 127 Issue 8 Part 2 | Pages 757-763, Annals of Internal Medicine, American College of Physicians]
4. Sajida S,K.VijayaLakshmi,M.Padmavathamma," "Implementaion of optimal Causal probability Decision Trees In different scenarios to discover Causal relationships",journal of emerging technologies and innovative research(2022).

5. Swati Hira & P. S. Deshpande Mining precise cause and effect rules in large time series data of socio-economic IndicatorsHira and Deshpande SpringerPlus (2016) 5:1625 DOI 10.1186/s40064-016- 3292-0

6. Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bing-Yu Sun:Mining Causal Association Rules. ICDM Workshops 2013: 114-123
7. Frey L., D. Fisher, I. Tsamardinos, C. Aliferis, and A. Statnikov, "Identifying Markov blankets with Decision Tree induction," in Proc. 3rd IEEE Int. Conf. Data Mining, Nov. 2003, pp. 59–66.]
8. Bollen K.A., Pearl J. "Eight Myths About Causality and Structural Equation Models. In: Morgan S." (eds), Handbook of Causal Analysis for Social Research. Springer, Dordrecht (2013)
9. Jin Z., J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang, "Discovery of Causal rules using partial association," in Proc. IEEE 12th Int. Conf. Data Mining, Dec. 2012, pp. 309–318]
10. *Jiuyong Li, Lin Liu, Thuc* Duy *Le. Jiuyong Li • Lin Liu • Thuc* Duy *Le, "Practical Approaches to Causal Relationship Exploration" Jiuyong Li* School of Information Technology and Mathematical Sciences University.2015]
11. Spirtes Peter. "Introduction to Causal Inference, " Journal of Machine Learning Research 11 (2010) 1643-1662
12. Li. j, Liu. L, Le. T., "Practical approaches to Causal relationship exploration"2015. X. 80 p. 55 illu., softcover, ISBN:978-3-319-14432-0, http://www.springer.com/978-3-319-14432-0]
13. Magliacane Sara, Tom Claassen, Joris M. Mooij, "Joint Causal Inference from Observational and Experimental Datasets", Journal of Machine Learning Research, March 2017.
14. Christopher D. Ittner, "Strengthening Causal inferences in positivist field studies", Accounting, Organizations and Society 39 (2014) 545–549.
15. Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, Xenofon D. Koutsoukos, "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification", Journal of Machine Learning Research 11 (2010) 171-234.