Design and Development of OCPDT by computing Path Probability by multiplying Probability differences of all Branch wise internal nodes

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

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

ABSTRACT: The most powerful and widely used tool for classification and prediction is the Decision Tree. A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label. The strengths of Decision Trees are: Decision trees are able to generate understandable rules. Decision trees perform classification without requiring much computation time. But it also suffers some limitations such as: The training of decision trees can be computationally expensive. A decision tree's growth requires extensive computational work. Each candidate splitting field at each node must first be sorted in order to determine which split is best. Some algorithms employ combinations of fields, so it is necessary to look for the best combining weights. Due to the necessity of creating and comparing numerous candidate sub-trees, pruning algorithms can also be costly. 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 Path Probability by multiplying Probability differences of all Branch wise internal nodes.

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

1. INTRODUCTION:

A popular category of probabilistic graphical models are Bayesian networks. They are made up of a structure and parameters. The structure, which expresses conditional dependencies and independencies among random variables connected to nodes, is a directed acyclic graph (DAG). Each node's conditional probability distributions make up the parameters. A compact, adaptable, and understandable representation of a joint probability distribution is a Bayesian network. Given that directed acyclic graphs allow for the representation of Causal relationships between variables, it is also a useful tool in knowledge discovery. A Bayesian network is typically trained using data. Probabilistic models can be used to quantify probabilities and define relationships between variables. For instance, fully conditional models might need a massive amount of data to account for every scenario, and probabilities might be practically impossible to calculate. Although it is a significant simplification, simplifying assumptions like the conditional independence of all random variables can be useful, as in the case of Naive Bayes.

The known conditional dependence with directed edges in a graph model is explicitly captured by Bayesian networks, a Probabilistic Graphical model. The conditional independencies in the model are defined by all missing connections. As a result, Bayesian Networks offer a practical tool to visualise the probabilistic model for a domain, examine all of the connections between the random variables, and infer Causal probabilities for scenarios based on the evidence at hand.

In this paper, the author is computing the best Causal Probability decision trees. Classification methods are commonly used to build decision trees using graphical model of Bayesian network, but they have limitations in terms of decision tree height and time complexities. Causal Decision trees were introduced to overcome the aforementioned limitations. Further enhancing a tree's performance is pruning. It entails cutting off branches that rely on minor features. We lessen the tree's complexity in order to boost its predictive ability by lowering overfitting. Either the leaves or the roots can be pruned first. The most fundamental pruning technique starts with the leaves and eliminates every node belonging to the most popular class in each leaf; if this change does not impair accuracy, it is kept. Additionally called Reduced Error Pruning. However, these techniques fall short of locating a single Cause. A straight forward approach to discovering a combined cause is to include both individual and combined variables in the Causal discovery, the author proposed a metric for evaluating the Finest Causal Probability Decision Trees by Path Probability multiplying Probability differences of all Branch wise internal nodes.

2. Related work:

Jiuyong Li, et al.'s[1] investigation focused on how classification techniques fall short in their attempts to establish Causal links between input and output variables by failing to take other variables into account. The authors made the case that classification methods are not meant to be used for determining Causal relationships, and they proposed a new scalable, automated Causal decision

tree framework model based on a unique statistic-based Causal relationship framework for determining real Causal relationships in sizable data sets. Big data applications can also make use of the new method that is being proposed.
Sajida.S, et al[2] 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.

Sajida.S, et al[3] has proposed various metrics for the identifying of Optimum Causal Probability Decision Trees. The author discovered combined Causal relationships in this paper by employing propose a metric to construct Optimal Causal Probability Decision Trees with improved correlation and Causal inference. As being present in many real-world applications, Peter Spirtes [3] addressed and highlighted this. The author also investigated all of these problems with determining Causal relationships when using algorithms for graphical Causal modelling. A number of Causal modelling-related issues were raised by the author, including how to match Causal models and search algorithms to Causal problems, model selection, prior knowledge, ways to improve the effectiveness and efficiency of search algorithms, ways to characterise search algorithms, and ways to add and remove simplifying algorithms. The author also investigated different Causal models, examined potential problems, and talked about the actual problems with Causal inference.

Finnian Lattimore The Gradient Institute, Paris David Rohde Criteo AI Lab, Paris[4] In this paper, the authors has showed how a probabilistic graphical model can represent the assumptions encoded by Causal graphical models (PGM). The main benefit of doing this is conceptual: it enables Bayesian practitioners to represent and analyse the modelling presumptions necessary for Causal inference in a context that is familiar to them. If the do-calculus is unable to identify Causal queries, there may still be practical advantages. In these circumstances, it is fundamentally impossible to predict how an intervention will turn out, even when given without any additional assumptions, infinite pre-interventional data. Modelling these issues in a We take advantage of a large body of existing research on fusing assumptions with data to derive finite sample estimates for distributions of interest using standard Bayesian inference settings. Despite the unless we add assumptions about the prior, posterior distribution will always remain sensitive to the prior. They might still be able to get useful bounds if the relationships between the variables take a functional form. The drawback of explicitly modelling Causal questions as a single PGM is that it is more time-consuming and complex. Computationally costly (unless we use the do-calculus machinery to identify appropriate re-parameterisation

3. Design and Development of OCPDT by computing Path Probability by multiplying Probability differences of all Branch wise internal nodes:
The predicted variable and the outcome variable are the two variables in the Causality decision tree. Using the partial association test, a traditional statistical test, the predicted variable is chosen. Therefore, the Causality of internal nodes is generated for each branch. The author has suggested one metric to help determine which Causal Probability Decision trees are the best. The predicted variable and the outcome variable are the two variables in the Causality decision tree. Using the partial association test, a traditional statistical test, the predicted variable is chosen. Therefore, the Causality of internal nodes is generated for each branch. The author has suggested one metric to help determine which Causal Probability Decision trees are the best.
Avg of $\sum_{i=1}^{n}$ path probability $*$ leafsize

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

The left branch and right branch path probability difference for each internal node of each tree is computed, and the resulting probabilities and leaf node size are then multiplied to obtain the path score. All of these path scores are combined to form the tree score. The best Causal Probability Decision Tree is preferred based upon the highest tree score. In the paper Sajida et.al[3] the author has generated the Causal Probability Decision Trees .By considering that trees it has certain internal Causality values as follows.

Details of internal nodes' cause and effect:

The author of this paper used the UCI machine learning ADULT DATA SET. The ADULT DATASET contains 14 attributes. They go by the names A, B, C, D, E, F, G, H, I, J, K, L, M, and Y. With the aid of the Statical test, also known as the Partial Association Mantel Haszel test, the Causality values of each node are calculated and a Causal Probability Decision Tree is built.

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:

Branch number	Internals nodes	Causality Value/Values of internal nodes
branch-1	A	316.8423269

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

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-1 size * path- probability)

Branch Wise Score	**Final Causality Value**
Leaf-1 size * path-1 probability	13044 * 0.423112644 = 5519.081328
Leaf-2 size * path-2 probability	2999 * 0.010854177 = 32.55167682
Leaf-3 size * path-3 probability	9736 * 0.010854177 = 105.6762673
Leaf-4 size * path-4 probability	5217 * 0.020517729 = 107.0409922
Leaf-5 size * path-5 probability	14226 * 0.048993651 = 696.9836791

Table-2:To Compute Branch wise Score Causality values

In a similar manner, we calculate the internal node Causality and leaf node size, as well as the branch scores for each branch, for Tree-2 (which is produced by altering a correlation value). Then, the Optimal Causal Probability Decision Tree with the highest value is selected by comparing the Causality of internal nodes in both trees.

Branch	**Final Causality Value** Tree-1	**Final Causality Value** Tree-2
Leaf-1 size * path-1 probability	5519.081328	5519.081328
Leaf-2 size * path-2 probability	32.55167682	10.19104724
Leaf-3 size * path-3 probability	105.6762673	228.8199777
Leaf-4 size * path-4 probability	107.0409922	107.041
Leaf-5 size * path-5 probability	696.9836791	696.9837
Tree average score	1292.2668	**1312.423407**

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

The Optimal Causal Probability Decision Tree will be determined by comparing the average Causality scores of the two trees. The Tree-2 is chosen as the best Causal Probability Decision Tree among the generated trees because its average tree score is the highest. There may, in general, be a variety of causal probability decision trees for the same dataset and predetermined tree height. In these situations, the suggested techniques identify the optimal causal probability decision tree for the supplied dataset.

CONCLUSION:

Decision tree learners create Bayesian networks graphical model trees but it has a major limitation of time complexity. To overcome the limitation Causal effect decision trees were created when certain classes are dominant. Therefore, it is advised to balance the data set before fitting it to the decision tree. It is possible for decision-tree learners to create excessively complex trees that do not

adapt well to new data. Overfitting is the term for this. The discovery of Causal relations shows not only the correlation between the indicators but also how altering one cause variable is predicted to alter another. Building Causal Probability Decision Trees is therefore necessary to obtain scalable and understandable data. The various Causal Proposed Decision Trees are created, and using the suggested metric, the best Causal Probability Decision Tree is selected.

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] 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).

[3] Sajida S,K.VijayaLakshmi," Design and Development of Optimal Causal Probability Decision Tree by ComputingPath probability of Internal Causality nodes" www.ijsdr.org International Journal of Scientific Development and Research (IJSDR) 316.

[4] Finnian Lattimore The Gradient Institute, Paris David Rohde Criteo AI Lab, Paris Causal inference with Bayes rule.

[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.