JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IMPLEMENTAION OF OPTIMAL CAUSAL PROBABILITY DECISION TREES IN DIFFERENT SCENARIOS TO DISCOVER CAUSAL RELATIONSHIPS

[1]S.Sajida, [2]Dr. K.Vijaya Lakshmi,[3]Prof.M.Padmavathamma

[1]Research Scholar, ,[2]Assistant Professor ,[3] Professor

[1,2,3] Department of Computer Science,

[1]Sri Venkateswara University, Tirupati, AP, India.

Abstract: *In* computer science causality is a fundamental notion and it plays a vital role in exploration and prediction of Decision making control. Many real time application scenarios, it is very useful to discover causal relationships between multiple causal variables and single predicted outcome variable. Causal decision trees (CDT) construction is similar to the classification decision tree construction but the former uses different static based framework for finding causal relationships between variables. Sometimes two or more causal decision trees will be created for the given same dataset with pre specified tree height. Present study proposes new techniques to find the best causal probability decision tree from among the many possible causal decision trees. Extensive experiments are conducted and the results show that proposed techniques are reasonably good in finding optimal causal probability decision trees**.**

IndexTerms **-** *CDT, node size, node probability, node type, node causality, aggregated score*

I. Introduction

The methods used to construct normal decision trees and causal decision tree differ. It is necessary to design experimental support before making definitive statements about cause and effect. In order to construct causal decision trees different measures are used. Generally people seeks causal relationships in their life, By observing an event, one might be able to infer its cause. For example, Hard work leads to good results. Better health results from eating healthy foods. Sometimes the same instance may be a cause and an effect as well. Causal relationships benefit policymakers, practitioners, and scientists by providing them the cause and effect pair assessments. Among the sets of probable cause and effect relation pairs most of the options were neither practicable nor desirable and a few alone are plausible. The use of causal discovery techniques using observational data has drawn attention in computer science research during the last three decades. At the moment, Bayesian network techniques make up the majority of the strategies used in computer science to find causal relationships.

Statistical methods were mostly used to study causal relationships. Bayesian theorem provided the foundation to analyse and forecast a cause from the observable sets of effects under examination. This method can produce probabilistic forecasts. Probabilistic reasoning alone is insufficient to rely on in situations requiring vital decisions. Generally speaking, it might not be possible to identify a trustworthy real causal model for the supplied inputs.It can be very challenging to determine how well causal model algorithms work at times. A causal model is frequently broken down into two pieces, the first of which is referred to as a statistical model and the second as a causal graph that specifies the relationships between the variables.

As a result, Causal Decision trees must be better interpretable and shorter in Height in order to find optimum existing causal relationships in the data sets. A reasonable level of performance has been achieved by the **optimal causal probability decision trees** proposed in this research through the experimental verification of the proposed techniques. Using the proposed techniques, the results from the present research were experimentally verified and showed that optimal causal probability decision trees had good outcomes. This research verified that optimal causal probability decision trees have been able to produce reasonable results using the proposed techniques.

II.Related Work:

Yeying Zhu, etal [3] studied a casual inference problem with a continuous treatment variable based on propensity scores. The authors defined Propensity score as the conditional density of the treatment level given covariates. Propensity scores were used to estimate the weights of inverse probability. A boosting algorithm was suggested to estimate the mean function of the treatment given covariates.In this studied a casual inference problem with a continuous treatment variable based on propensity scores. The authors defined Propensity score as the conditional density of the treatment level given covariates. Propensity scores were used to estimate the weights of inverse probability.

Finding causal connections in huge databases of observational data is particularly challenging, according to Z. Jin et al. [9]. The fundamental drawback of using Bayesian networks in this field for causal link discovery is that learning them is an NP-complete issue. As a result, several constraint-based algorithms have been created and developed for efficient causal relationship discovery from big data sets

Jiuyong Li ,et al[10] said that causal relationships are generally found with designed experiments such as randomized controlled trails but these are costly to conduct. They also said that causal relationships can also be discovered with well designed observational studies by taking the help of domain expert's knowledge and also pointed out that this is a time consuming process. They observed that more advanced scalable and automated state-of-the-art techniques are needed for finding potential causal relationships between the variables and the outcome variable in the case of large data sets. Authors pointed out that classification methods may appear that they are good for finding causal relationships but in reality the classification methods may find false causal relationships and could miss true causal relationships. They studied that classification methods fail to take accounts of other variables while trying to establish causal relationships between the input variables and the outcome variable. Authors argued that classification methods are not designed for finding causal relationships and they proposed a new scalable, automated causal decision tree framework model based on special statistic based causal relationship framework for finding true causal relationships from the large data sets. The proposed new technique is also applicable for big data applications also.

Large numbers of variables, short sample sizes, and the utilisation of unmeasured causes are issues that Peter Spirtes [12] addressed and highlighted as being present in many real-world applications. When using algorithms for graphical causal modelling, the author also explored all of these issues with determining causal relationships. The author brought up a number of issues related to causal modelling, including how to match causal models and search algorithms to causal problems, model selection and prior knowledge, how to increase the effectiveness and efficiency of search algorithms, how to characterise search algorithms, and how to add and remove simplifying algorithms. Additionally, the author explored various causal models, analysed potential issues, and discussed the real issues with causal inference

Sara Magliacane, et al. [13] discussed about Joint Causal Inference (JCI) by taking multiple data sets to learn both causal structure and outcomes interactively. They observed that JCI offers many advantages when compared with the many existing constraint based causal inferences for finding causal relationships from the pooled data of multiple data sets.

.

In [13]matching approaches are employed under comparable observations for identifying cause effect linkages by comparing treatment units with control units. The use of various matching methods in various applications, including medicine, criminology, science, economics, education, social sciences, public policy analysis, scientific disciplines, statistics, machine learning, data mining, sociology, psychology, research, behavioural sciences, and so on, is reportedly widespread, according to authors. Most often, data linkages between covariates, treatments, and output variables are only loosely modelled by matching approaches. Matching techniques can be used for several transdisciplinary applications. Matching techniques primarily rely on covariate distance measurements between the treatment and control groups.

.III.PROPOSED METHODOLOGY

The causal decision tree, as the name implies, is a special type of tree that represents the cause and effect relationship between input attributes and target attributes. Usually, causal decision trees are special types of trees that represent cause-and-effect relationships between input attributes and target attributes by means of a causal decision tree.Although the accessibility and suitability of present approaches for casual inference seem promising, more sophisticated methods are required to meet the demands of the current data analysis. Data availability in a wide range of formats and in large numbers is emerging as a major concern. The choice of the appropriate collection of qualities (parameters) for the causation process from this data with a vast set of attributes is once again a difficult problem. The issue of estimating causal parameters is not immediately addressed by existing statistical, data mining, and machine learning (ML) approaches to estimation, model selection, and robustness.

The two different causal probability decision trees created for the same dataset by considering different correlation threshold values. The problem now is how to select the best causal probability decision tree from the two different causal probability decision trees with the same height. Both trees are equally eligible and acceptable for the consideration of decision making. But when the causality effect value is quantized, generally, different causal probability decision trees will give different results.

Proposed techniques for causality score quantization are:

1. Average of causality values of all the non-leaf nodes.
2. An average of the sum of the causality and probability products of all internal nodes on all branches.
3. The probability of a path is multiplied by the sum of causalities of all internal nodes in the path.
4. The probability of a path is calculated by multiplying the probability differences between all nodes within the path. Tree score is computed as sum of all path scores.

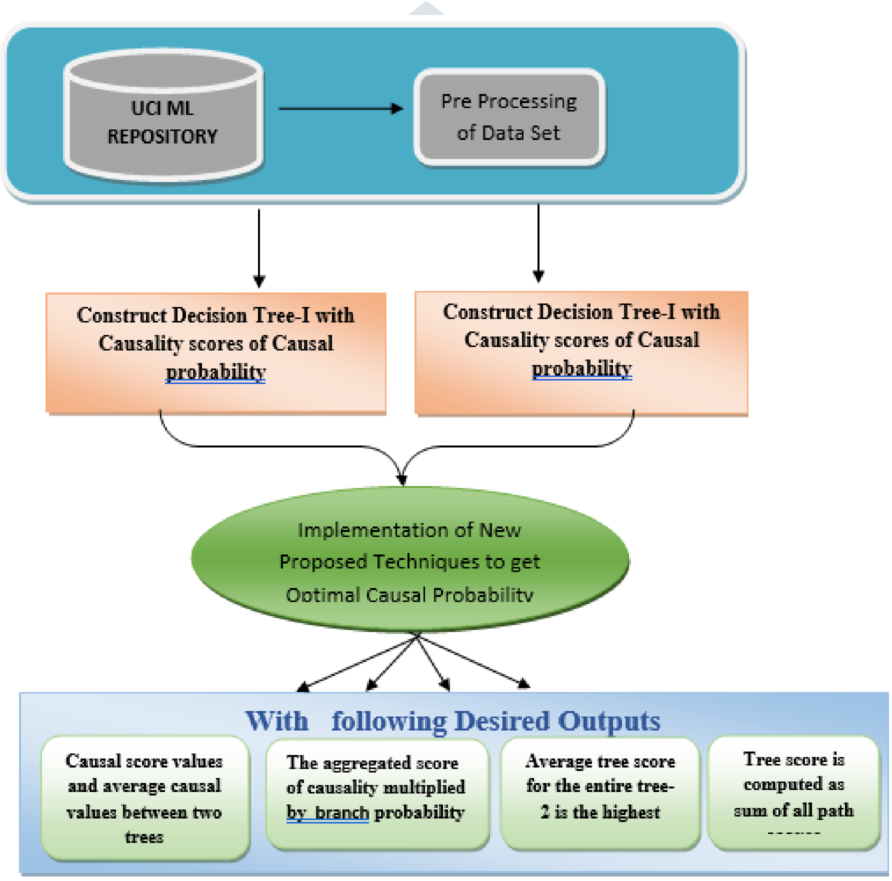
UCI ML REPOSITORY
Pre Processing of Data Set
Construct Decision Tree-I with Causality scores of Causal probability
Construct Decision Tree-I with Causality scores of Causal probability
Implementation of New Proposed Techniques to get Optimal Causal Probability
With following Desired Outputs
Causal score values and average causal values between two trees
The aggregated score of causality multiplied by branch probability
Average tree score for the entire tree-2 is the highest
Tree score is computed as sum of all path ------

Figure: Contextual Architectural Model of Proposed Work

Various techniques are proposed for deriving causality scores of causal probability decision trees. Each technique is directly related to the actual causal probability decision tree parameters such as node size, node probability, node type, node causality, class labels information of tuples and so on. The purpose of all these techniques is to represent causal probability decision tree quantitatively. This quantitative representation is useful to compare two causal probability decision trees in terms of desired parameters.

Machine learning UCI repository INCOME CENSUS and the data set “ADULT DATASET” is used as dataset for this research. The description of the Dataset is given below.

DATASET DESCRIPTION

Name of the dataset “ADULT DATASET” consisting of 45222 tuples, thirteen predictive attributes and one target attribute. For simplicity purpose thirteen predictive attributes are taken as A, B, C, D, E, F, G, H, I, J, K, L, M and target attribute is taken as Y.

Actual attribute name	Renamed attribute
age < 30	A
age > 60	B
private	C
self-emp	D
gov	E
education-num >12	F
education-num < 9	G
prof	H
white	I
male	J
hours > 50	K
hours < 30	L
US	M
>50K	Y

New Proposed Causal_Tree_Creation Algorithm :

```
Step 1. Read dataset from the UCI repository
Step 2. Create the root node and store all the attributes, dataset tuples, dataset size etc.
Step 3. Call the Create_Causal_Probability_Tree() algorithm
Step 4: Split the current node data into left subset and right subset using splitting attribute
Step 5: Recursively call Create_Causal_Probability_Tree(left node address, $n_1$, h)
Step 6: Recursively call Create_Causal_Probability_Tree(right node address, $n_2$, h)
```

New Proposed Create_Causal_Probability_Tree(Root, n, h) Algorithm:

Root is the root node of the decision tree
n is the number of tuples in the dataset
h is the desired height of the causal probability tree

```
begin
    Step1.   If the root node contains all the tuples with the same class label then
    Step 2.  Create leaf then return
    Step 3.  else
    Step 4. If all attributes are exhausted or maximum height is reached
             then
                 Step 4.1  Create leaf node
    Step 5   return
    Step 6.  else
                 Step 6.1 Find the correlation between input attributes and target attribute
                 Step6.2 Find the best attribute among the correlation threshold satisfied attributes by using statistical measure
    Step 7.  h = h + 1
end
```

Proposed Technique-1:

In the causal probability decision tree, all **non-leaf nodes** are summed to compute causality values. An average is then calculated based on the causality values of all the non-leaf nodes. After that, it computes the average causality value of all the non-leaf nodes.In order to determine the best causal probability decision tree for a given height, the causal probability decision tree with the highest average score value is selected as the best causal probability decision tree.

Proposed Technique-2:

Each tree branch consists of a set of sequential internal nodes. In each path, the causality of **each internal node** is multiplied with corresponding path probability and these results are summed and treated it as branch score. In the present method five branches are there, correspondingly five leaf scores are computed and then finally tree score is computed by averaging all these five branch wise scores. The best causal probability decision tree is one whose tree score is the highest tree score.

Proposed Technique-3:

A tree aggregate causality score is computed by **averaging all the path scores of the tree and multiplying each complete path** probability by the sum of causalities of all internal nodes. Based on the highest tree score the best causal probability decision tree is selected.

Proposed Technique-4:

For each **internal node of each tree path probability** difference of left branch and right branch is computed and then the resulted outcome probabilities and leaf node size are multiplied to get the path score. Tree score is the aggregation of all these path scores.

IV. CONCLUSION

Causal probability decision tree is an important tool in many applications including medical field. For the same dataset with the pre specified tree height sometimes, generally, there may be many possible causal probability decision trees. In such cases, the proposed techniques find the best causal probability decision tree for the given dataset. The best causal probability decision tree is called optimal causal probability decision tree. In the future there is a scope to investigate many other efficient and effective techniques for finding optimal quantitative causal probability decision trees.

REFERENCES

1. Alfieri's C. F., et al. [C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D.Koutsoukos, "Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation", Journal of Machine Learning. Res., vol. 11, pp. 171–234, 2010.
2. Birch M.W." The Detection of Partial Association", Journal of Royal Statistical Society Aeries B (Methodological), Vol. 26. No. 2 (1964), pp. 313-324.
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.
4. 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)
5. Christopher D. Ittner, "Strengthening causal inferences in positivist field studies", Accounting, Organizations and Society 39 (2014) 545–549.
6. 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.
7. 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]
8. 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.]
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, 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
11. Jiuyong Li, Saisai Ma, Thuc Duy Le, Lin Liu and Jixue Liu "Causal Decision Trees," arXiv: 1508.03812v1 [cs.AI] 16 Aug 2015]
12. *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]
13. Spirtes Peter. "Introduction to Causal Inference, " Journal of Machine Learning Research 11 (2010) 1643-1662
14. 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]
15. Magliacane Sara, Tom Claassen, Joris M. Mooij, "Joint Causal Inference from Observational and Experimental Datasets", Journal of Machine Learning Research, March 2017.
16. S. L. Morgan and D. J. Harding, "Matching estimators of causal effects: Prospects and pitfalls in theory and practice," Sociological Methods Res., vol. 35, pp. 3–60, 2006.
17. Sander Greenland and Babette Brumback, "An overview of relations among causal modelling methods",*International Journal of Epidemiology*, Volume 31, Issue 5, 1 October 2002, Pages 1030–1037.