Extracting Partial Association Mantel-Haenszel Test based Cause and Effect Relationships using Decision Tree

Dr. D. Mabuni
Assistant Professor, Dept. of Computer Science and Technology, Dravidian University, Kuppam, Chittoor (District), Andhra Pradesh, India

Abstract—Decision trees are very useful tools for both data classification and regression in many real time situations in data mining, machine learning, big data analytics including many distributed data applications. Usage of a single data analysis technique is common in many machine learning techniques. Many recent trends are being becoming popular towards the usage of hybrid techniques in data analytics. As a result, the standard benchmarking decision tree classifiers are combined with many other statistical techniques in order to find cause and effect relationships present in the given datasets. Causal relationships are computed between a predictor (input) variable and the outcome (output) variable. One of the most popular statistical based data analysis techniques called Partial Association Mantel-Haenszel Test combined with bench mark decision tree classifiers in order to elucidate cause and effect relationships in the datasets. These hybrid techniques are applied on the simple and hypothetical dataset for finding cause and effect relationships in the dataset. Experiments are conducted on the selected dataset, customers. Experimental results have revealed that the identified cause and effect relationships present in the dataset are real and well-matched with many of the real time situational scenarios.

Index Terms—cause and effect; causal relationships; hybrid; Partial Association Mantel-Haenszel Test; machine learning; data mining.

I. INTRODUCTION

In modern days effective and intelligent way of data management of very large datasets is crucial for all organizations. In this regard, steps already have been taken for proper management of data in multitude ways. Main goal of machine learning is accurate prediction of unknown data based on learning from known data. Machine learning, data mining, statistical and big data analytics techniques and methods are rapidly and continuously being applied for knowledge extraction from the very large datasets. Large number of statistical techniques are available for proper management of data. When these statistical techniques are used combinedly with appropriate data analysis techniques, very useful results are being generated in many real applications. Usage of such hybrid data analysis techniques are common nowadays.

Finding causal relationships is inevitable in areas. Exploration of causal relationships in the very large datasets necessitates for the invention of more and more, scalable, interoperable, fast, dynamic, correct and convenient machine learning algorithms. Latest data analysis trends are increasing rapidly for finding causality relations. So far, all existing causality finding methods assume that there exists pre-assumed knowledge on the dataset. The main requirement of the many state-of-the-art data analysis tasks is that the methods that are used for finding the causal relationships in the data must have the intelligent capability of finding cause and effect relationships in the data without any pre assumed knowledge in the data. Such latest methods are very useful in medical diagnosis, research, sociality data analysis, and so on. Data analysis research trends have shown that ensemble methods and hybrid methods are increasing in use because such models are identified as better potentially suitable methods in causality exploration in finding accurate and reasonably good causality relationships in many domains including medical diagnosis, defense, physics, research, marketing, science, military, retail, and so on.

II. LITERATURE SURVEY

In the paper [1], authors have constructed causal decision tree using decision tree induction and partial association statistical test. Causal decision tree is constructed without pre-assumption that there exist

causal relationships in the data. The models used in [2], [3], [4] are the fundamental basic building blocks for finding the causal relationships between input and output variables. [5], [6] causal Bayesian networks are very popular in finding causal relationships. [7] interpretability and transparency are the powerful striking features of decision tree and in many cases random forest is the very useful and potentially applicable model for finding causality relations.

[8] causality theory is closely associated with causal sufficiency and faithfulness; the main requirement of causal sufficiency is that all common causes of observed variables must measurable. Recently, usage of local causal variables is increasing frequently in many domains. There are two types of local causal discovery methods. Algorithms belongs to the first category are used for learning a complete causal Bayesian network. Algorithms belongs to the second category are mainly comprising of hybrid approaches. [9] For example, the good decision tree data structure is combined with many statistical techniques for obtaining reasonably good causality relationships in the datasets.

[10], [11] Causal decision trees, causal probability trees, and causal explanation trees are used in finding causal relationships in the data by the pre assumption that there exist causalities in the data.[12], [13], [14] When the dataset contains too many attributes, obtaining perfect strata is very difficult and, in such situations, alternative methods must be needed. Propensity scores determination is one such method for finding strata and logistic regression methods are frequently used for finding correct number of strata.

[15] Authors have used ensemble decision trees for finding causal relationships in the data. To find better predictive accuracy of the proposed model authors have used many error finding and error correcting methods. [16] Random forest algorithms are used not only for classification methods but also for many regression problems. [17] Authors have used many machine learning methods for elicitation of causal relations in the data. [18] Authors have proposed a hierarchical, graphical and probabilistic model for finding causal relationships in the data. Authors in [19] have used a special causality analysis technique called unified granger technique for sequential imaging.

III. PROBLEM DEFINIION

Finding cause and effect relationships are very useful in many fields especially in medical diagnosis, agriculture, scientific, military, and many others. One way of finding causal relationships is by using only potential algorithm. Second way of finding causality relations is causal Bayesian network. Third way of finding causality relations is by using hybrid approaches. Application of hybrid data analysis techniques are common in many areas. In this research a statistical technique, Partial Association Mantel-Haenszel (PAMH) test, is combined with popular machine learning data analysis technique called decision tree classifier induction. In the present paper this hybrid causal data analysis technique is applied for finding cause and effect relationships in the given dataset.

Finding causal relationships in the dataset is very challenging task, in particular it is very tough task in finding causality relationships by using hybrid data analysis techniques. Even many of the local causal analysis tasks have NP-Hard time complexities. Particularly, the time complexity of causal Bayesian network (CBN) is known to be NP-Hard but it is very popular causality finding model in many real time situations including its probability associated special models also. All existing causality models are based on pre assumption that there exists causality in the dataset.

IV. EXPERIMENTS

Age	Job	Income	Education	Experience	Happiness	Count
0	0	0	0	0	0	2
0	0	0	0	1	0	3
0	0	0	1	0	1	2
0	0	0	1	1	0	4
0	0	1	0	0	1	100
0	0	1	0	1	1	90
0	0	1	1	0	1	120
0	0	1	1	1	1	130
0	1	0	0	0	1	60
0	1	0	0	1	1	70
0	1	0	1	0	1	50
0	1	0	1	1	1	40
0	1	1	0	0	1	100
0	1	1	0	1	1	150
0	1	1	1	0	1	110
0	1	1	1	1	1	120
1	0	0	0	0	0	4
1	0	0	0	1	0	5
1	0	0	1	0	0	2
1	0	0	1	1	1	2

1	0	1	0	0	1	60
1	0	1	0	1	1	50
1	0	1	1	0	1	70
1	0	1	1	1	1	80
1	1	0	0	0	1	60
1	1	0	0	1	1	50
1	1	0	1	0	1	40
1	1	0	1	1	1	80
1	1	1	0	0	1	120
1	1	1	0	1	1	130
1	1	1	1	0	1	140
1	1	1	1	1	1	150

Table-1 Customers

A hypothetical Customers dataset is shown in Table-1. The attributes age, job, income, education and experience are input (predictive) attributes and the target (output or outcome) attribute happiness is the output class label attribute. The count attribute tells that the frequency count value of each customer entity shown in the Table-1. Possible and coded values of each attribute are shown in the Table-2. For easy understanding and experimental purpose training dataset (Customers) is completely constructed only with necessary and suitable coded values of the attributes.

Attribute name	**Possible values**	**Coded values**
Age >= 30	High	1
Age < 30	Low	0
Job	Govt	1
Job	Private	0
Income	High	1
Income	Low	0
Education	High	1
Education	Low	0
Experience	High	1
Experience	Low	0
Happiness	Happy	1
Happiness	Sad	0

Table-2 Attributes of the training dataset and their possible values

Algorithm Causal_Decision_Tree
Input - Customers training dataset, D
X is the set of input attributes and Y is the output attribute
Output – Causal Decision Tree
Begin
1. Root = null
2. Height = 0
3. Create_Causal_Tree(Root, X, D, height, null)
4. Prune_Causal-Tree(Root)
5. Return Root

End

Create_Causal_Tree(N, X, D, height, e)
1. if the attribute set, X, is empty or tree height is maximum then
2. add two leaf nodes to the node N and then label leaf nodes with correct label
3. return
4. end if
5. find correlated and threshold satisfied attributes w.r.t Y
6. for each correlated attribute, X_i,
7. compute PAMH(X_i,Y)
8. end for
9. find the attribute, A, with the highest PAMH value
10.if PAMH value is less than the specified threshold then
11. add two leaf nodes to the node N and then label leaf nodes with correct label
12. return
13. end if
14. if e = null then
15. assign A as the root node of tree
16. else
17. add node A as the leaf node of node N and label it correctly
18. end if
19. remove A from attribute set X
20. for each w = 0 or w = 1 do
21. call Create_Causal_Tree(W, X, ReducedD, height, w)
22. end for

Training dataset contains five predictor attributes and one outcome attribute. Causality is computed between only one input attribute with the output attribute at a time by keeping the values of all other attributes as constants. For example, Age, Job, Education, and Experience attributes are kept constant and the causality relationship between income and happiness is computed. For income attribute sixteen partitions are possible but two partitions are computed and shown below for easy understanding purpose. Causality is computed for each partition separately and then aggregate causality score is computed finally by using partial association test and then this aggregate score is used in selecting the best split attribute during causal decision tree creation. Similar procedure is applied for computing causality score of each of the remaining attributes.

Partition – 1 of the income attribute.

{Age, Job, Education, Experience} Happiness

{0,0,0,0}	1	0	
Income = 1	100	0	100
Income = 0	0	2	2
	100	2	102

Partition -2 of the income attribute

{Age, Job, Education, Experience}	Happiness		
{0,0,0,1}	1	0	
Income = 1	90	0	90
Income = 0	0	3	3
	90	3	93

Computations are shown only for the partition – 1 of the income attribute.

$$\frac{n_{11k}n_{22k}-n_{21k}n_{12k}}{n_{..k}} = \frac{100*2-0*0}{102}$$

$$= \frac{200}{102} = 1.96$$

$$\frac{n_{1.k}n_{2.k}-n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)} = \frac{100*2*100*2}{102^2(102-1)}$$

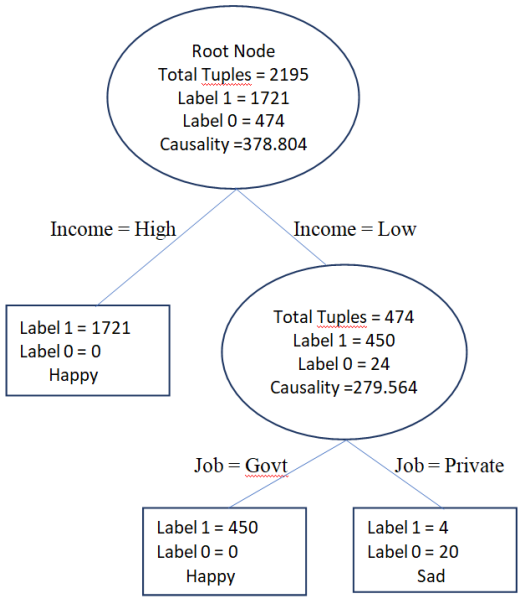
$$= \frac{40000}{1050804} = 0.038066$$

$$PAMH(Income, Hapiness) = \frac{\left(\left|\sum_{k=1}^{r}\frac{n_{11k}n_{22k}-n_{21k}n_{12k}}{n_{..k}}\right|-0.5\right)^2}{\sum_{k=1}^{r}\frac{n_{1.k}n_{2.k}-n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)}} \quad \ldots\ldots (1)$$

PAMH value is computed over for all partitions from k = 1 to k = r. that is, k runs form 1 to r where r represents total number of potential partitions of the selected attribute. Whenever the partition contains all vertical zeros or all horizontal zeros in any column or row then its causality score is 0 hence its contribution is zero to the aggregate causality score.

For each attribute, for all the k partitions, an aggregate PAMH is computed. It gives causal strength between input and output variable. During the causal decision tree induction select the attribute with the highest causality score as the best split attribute and then divide the data into partitions based on the best split attribute values. In general, for k number of attributes,2^{k-1} number of partitions are possible for each attribute separately and then partial association test is applied for all the partitions separately and then aggregately by using the statistical partial association test, PAMH, formula shown in the equation (1).

For ease of computation and understanding purpose only computations of Income attribute with only two separate partitions are shown. Actually, maximum of eight partitions are possible for income attribute. Similar procedure is applied for all other attributes in computing causality score. During tree growing phase the attribute with the highest causality score is used for selecting the best split attribute at each level of the tree induction.



Figure–1 Causal Decision Tree for the Customers Dataset

Causal decision tree shown in Figure-1 perfectly represents cause and effect relationships between input and output attributes. Age, Job, Income, Education and Experience are the input attributes and Happiness is the output (Target) attribute. In many real-life situations high incomes directly correlates more and more happiness. It is also true in the present experimental results and it is very close to the reality also. Job attribute also showing the same relationships with the reality situations. It is true that people doing government jobs are more and more happy than people doing private jobs because private job timings are very high, salaries are very less, salaries are not paying regularly and more over there is no job satisfaction and job security and plenty of other problems are inherently associated with private jobs. This direct and real relationships are clearly shown by the job attribute.

Experimental results conducted and shown in this paper clearly are showing the causality relationships between income and happiness and also between job and happiness. Initially income attribute exhibits very strong causality relationship with the happiness attribute. Therefore, income attribute is selected as the best split attribute at the starting. In the next level, job attribute is selected as the best split attribute because causality value of the job with happiness attribute is very high among the remaining attributes except the income attribute. This causality is computed in terms of statistical measure called PAMH, partial association test. Causality tree is generated based on the causality value but not with the usual splitting rules of the conventional decision tree creation. Therefore, the procedure that is used for creating the causal decision tree is completely different from the procedure used for creating normal decision trees. Normal decision trees are used for high accuracy classification results whereas causal decision trees are used for finding causality relationships among the data values.

Causality value is computed for each attribute initially. Income attribute has the highest causality value. At the root node income attribute is the split attribute. In the next level causality is computed for all attributes except income. Number of attributes gets reduced as the number of tree levels increases from top to bottom.

1. The causality strength between income and happiness is very high and, in the experiment, also same relationship is reflected and its strength is computed first because its causality is the highest among the given attributes.
2. Next the causality relationship between job and happiness is computed as the next highest value. It is also very closely resembling the any real time situations.
3. Age, education, and experience do not cause any causal relation with happiness or sometimes there may exist very small quantity of causal relationship.

CONCLUSIONS

Decision tree classifiers are very famous in machine learning and their construction follows with Greedy and top down approach using predefined node data split attribute measuring rules. But the procedure for creating causal decision trees is completely different from the usual decision trees. Nowadays state-of-the-art and hybrid approaches for finding causal data relationships are popular. A special hybrid approach is used in this paper for finding causality relationships. The best and more popular decision tree data structure is combined with the best PAMH, partial association test, statistical measure for finding causality relationships. Also, experimentally verified that causal relationships are identified correctly if they present in the data. In the future there is a possibility to apply different hybrid techniques for finding causality relationships. Even many potentially possible ensemble methods are available for finding causality relationships in the datasets.

REFERENCES

[1] Jiuyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu, “Causal Decision Trees”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 2, FEBRUARY 2017

[2] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[3] D. B. Rubin, “Causal inference using potential outcomes: Design, modeling, decision,” J. Amer. Statistical Assoc., vol. 100, no. 469, pp. 322–331, 2005.

[4] G. W. Imbens and D. B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2015.

[5] P. Spirtes, “Introduction to causal inference,” J. Mach. Learn. Res., vol. 11, pp. 1643–1662, 2010.

[6] R. E. Neapolitan, Learning Bayesian Networks. Englewood Cliffs, NJ, USA: Prentice Hall, 2003.

[7] KajaBalzereit , Alexander Maier , Bjorn Barig ¨ , Tino Hutschenreuther and Oliver Niggemann, “Data-driven Identification of Causal Dependencies in Cyber-Physical Production Systems”, In Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019), pages 592-601 ISBN: 978-989-758-350-6.

[8] J. Li, L. Liu, and T. D. Le, Practical Approaches to Causal Relationship Exploration. Berlin, Germany: Springer, 2015.

[9] 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,” J. Mach. Learn. Res., vol. 11, pp. 171–234, 2010

[10] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in Bayesian networks," in Proc. 12th Conf. Uncertainty Artif. Intell., 1996, pp. 115–123.
[11] U. H. Nielsen, J. philippe Pellet, and A. Elisseeff, "Explanation trees for causal Bayesian networks," in Proc. Uncertainty Artif. Intell., 2008, pp. 427–434.
[12] R. P. Rosenbaum and B. D. Rubin, "The central role of the propensity score in observational studies for causal effects," Biometrika, vol. 70, no. 1, pp. 41–55, 1983.
[13] D. B. Rubin, "Estimating causal effects from large data sets using propensity scores," Ann. Internal Med., vol. 127, no. 8, pp. 757–763, 1997.
[14] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," Statistical Sci., vol. 25, no. 1, pp. 1–21, 2010.
[15] Neelam Younas, Amjad Ali, Hafsa Hina, Muhammad Hamraz, Zardad Khan and Saeed Aldahmani,
"Optimal Causal Decision Trees Ensemble for Improved Prediction and Causal Inference", Received January 11, 2022, accepted January 19, 2022, date of publication January 26, 2022, date of current version February 4, 2022. Digital Object Identifier 10.1109/ACCESS.2022.3146406
[16] S. Wager and S. Athey, ''Estimation and inference of heterogeneous treatment effects using random forests,'' J. Amer. Stat. Assoc., vol. 113, no. 523, pp. 1228–1242, Jul. 2018.
[17] S. Singla, S. Wallace, S. Triantafillou, and K. Batmanghelich, ''Using causal analysis for conceptual deep learning explanation,'' in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. USA: Inst. of Mathematical Statistics, 2021, pp. 519–528
[18] M. Pastorino, A. Montaldo, L. Fronda, I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, ''Multisensor and multiresolution remote sensing image classification through a causal hierarchical Markov framework and decision tree ensembles,'' Remote Sens., vol. 13, no. 5, p. 849, Feb. 2021.
[19] Z. Hu, F. Li, X. Wang, and Q. Lin, ''Description length guided unified Granger causality analysis,'' IEEE Access, vol. 9, pp. 13704–13716, 2021.