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

Probabilistic Causal Decision Trees

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

Abstract: Causal relationships are popular and common in modern data analytics with machine learning techniques. Graphical causal data models are one special attractive and dominant use in many day to day applications. Probabilistic graphical causal data models are also becoming common and their usage is inevitable in numerous artificial intelligence and machine learning applications. Causal decision trees are scalable, fast, automatic and budding data analytics models. Probabilistic causal decision trees are more attractive, convenient and very probable causal data models in medical diagnosis, research, business etc. In this paper two probabilistic causal decision trees are constructed. The first is constructed by taking one hypothetical dataset for easy understanding purpose and the second one is constructed by taking another Adult UCI machine learning dataset. Probabilistic causal decision tree created in this paper for adult dataset is exactly same with causal decision tree created earlier for adult dataset with other standard frameworks. These results show that the proposed probabilistic causal decision tree creation framework is correct, efficient and very useful for taking effective decisions in many real-life applications.

Index terms: **probabilistic causal decision trees, causal decision trees, graphical causal models, probability.**

I. INTRODUCTION

The main goal of probabilistic causation is to find relationships between cause and effect by using probability tools. Examining the cause and effect relationships between predictor and target variables through probability is called probabilistic causation. In general, one variable may or may not cause another variable. In fact, many variables may cause a single variable. Causality can be either deterministic or probabilistic. A deterministically causes B means that if A occurs B should also occur. Probabilistic causality is defined as: X probabilistically causes Y means that if X occurs then Y may occur or may not occur. X causes Y if and only if X increases the probability of Y in all possible situations. Probabilistic reasoning is one of the important forms of knowledge representation. Probability and logic combination are used in many domains to handle data uncertainty. Causality is categorized as absolute causality, conditional causality, and contributory causality. As of today, deep learning-based causality models are increasing in usage for effective decision making. Machine learning models including deep learning models could capture causal as well as probabilistic causal relationships. Uncovering cause and effect relationships in data would be very useful to provide good solutions in many areas such as medical field, physics, research, artificial intelligence, and machine learning. Efficient, smarter, intelligent, and robotic systems can be constructed through better understanding of cause and effect relationships.

In linear causality happening of one thing makes happening of another. Effects that become causes are called domino causalities. Causality detection algorithms are needed in automation of artificial intelligence related tasks.

The main research outcome in probabilistic causation is that causes raise the probabilities of their effects. Mathematization of combination of causal and probability theories play an important role in finding solutions to many problems.Causation is a quantitative measure. In general, the causal effect strength of one predictor variable is different from another predictor variable acting on the same target variable. Determination of causal strength or graded causation is needed in medical diagnosis, military, missile design, and physics etc. To find causal strength many mathematical frameworks were already developed. Formulas are available for finding positive causation, negative causation and no causation. Causal models combined with probabilities are very important when there is a need to make interventions. Increase in the dosage quantity of a specific tablet causes to increase or decrease the pain of patient is an example for cause and effect relationship. An increase in quality in teaching field definitely increases the pass percentage of the students in the class. Probabilistic causal models are specifically useful to analyse, understand and then explain various data generation processes. Deep learning methods combined with probabilistic causality becoming popular.

II. RELATED WORK

[1] Proposed an aided diagnostic method for probabilistic counter examples using possible notations and notations of the causality and responsibility. [2] Employed probabilistic causality analysis techniques and thoroughly studied and then evaluated performance of higher-

level classes and lower level classes separately. They found significant different results against the performance details of upper and lower level classes. Finally, they came to the conclusion that lower level performance causality is not created any improved performance in higher level.[3] A Multivariate probabilistic causal analysis technique is used to create cyber security system with many causal variable-based alarms. Then the system is analysed by applying various probabilistic causal events on the selected datasets. The obtained results are effective and accurate in terms of consistency.[4] Developed a graphical model algorithm with matrices of events called dynamic uncertain causality graph with probabilistic reasoning approach. This algorithm is applied on generator system of a nuclear power plant for finding and diagnosing faults. Finally, causal logic between inference results and observation are displayed in the form of graphs.[5] Dynamic and uncertain causality graph is proposed for finding fault diagnosis of very large industrial systems.

[6] Probabilistic reasoning is a good tool for handling uncertainty with constraints and causal relations. In some applications probabilistic causality frameworks are combined with deep earning algorithms to create efficient and effective decision-making systems. Probabilistic based kernels are constructed for robust-decision making systems.[7] Naïve Bayesian classifier is extended with probabilistic causality framework and causality-based attribute weighting system is proposed.[8] A special probabilistic framework is designed for model reference and adaptive control called causal form of the randomized controller.[9] A state space reconstruction algorithm is created for causal discovery in coupled time series. [10] A new technique is proposed with uncertain causality representation and probabilistic reasoning for online fault diagnoses of large and complex industrial systems. [11] Causal techniques play an important role in risk-based applications. Two types of causal relationships are used in risky decision-making systems. The main goal of the first type of causal relationship is to find the tendency of one event to cause another. The goal of second causal type is to find dominant cause among the many potential causes that affects the target variable. This paper explains how probabilistic causality is applied and tested on different types of risk analysis applications.

[12] Probabilistic causality models are used to represent probability distributions in terms of selected variables to analyse real systems and then make good decisions. Now a days automation of tasks is very important and it requires reasoning to take correct decisions and then execute desired actions. For example, artificial intelligence-based program accepts symptoms of patients and then analyse those symptoms thoroughly in order to find correct disease of the patients.[13] Causal strength is measured by two popular methods – causal effect measure and difference measure. Some important causal strength measures finding methods are discussed and explained clearly.[14] Causal relationships are thoroughly analysed using mathematical and probabilities framework tools and also many causal concepts are explained in detailed.[15] Bayesian nets and causality are one of the fastest growing technology in artificial intelligence with many newly added causality-based algorithms.[16] Causality and responsibility are applied for skyline queries.

[17] probabilistic causality is applied in multiple causation view.[18] Causal independence between sets of factors is explained in more detail.[19] Various types of causalities such as linear causality, domino causality, cyclic causality, spiralling causality, mutual causality, and relational causality are explained neatly.[20] Physical causalities are empirically studied rather than studying logical or conceptual causalities. All test cases that are used are only empirical test cases only. [21] Observed practically and said that time series parameters are very important in many fields including causal inference-based domains. The core assumption of probabilistic theory says that correlation between two variables is the indication of causal connection between variables. [22] Pointed out that health sciences are dependent not only on the physical mechanisms but also on the probabilistic dependences.

III. MATHEMATICAL FORMULATION OF PROBABILISTIC CAUSAL DECISION TREES

3.1 Conditional Probability based decision trees

$$P(gutka = 1/cancer = 1) = \frac{P(gutka = 1 \cap Cancer = 1)}{P(cancer = 1)} = \frac{\{Total\ number\ of\ cancer\ patients\ who\ have\ gutka\ habit\}}{\{Total\ number\ of\ cancer\ patients\}} \dots\dots\dots\dots\dots (1)$$

3.2 Causal Probability based decision trees

$$\begin{aligned} P(gutka = 1/cancer = 1) &= \frac{P(gutka = 1 \cap Cancer = 1)}{P(cancer = 1)} - \frac{P(gutka = 0 \cap Cancer = 1)}{P(cancer = 1)} \\ &= \frac{\{Total\ number\ of\ cancer\ patients\ with\ gutka\ habit\}}{\{Total\ number\ of\ cancer\ patients\}} \\ &- \frac{\{Total\ number\ of\ cancer\ patients\ without\ gutka\ habit\} \dots\dots\dots\dots\dots (2)}{\{Total\ number\ of\ cancer\ patients\}} \end{aligned}$$

Probabilistic causal decision trees are constructed by using both conditional probability and causal probability and it has been observed that both output results are same and correct. Up to moderate data size datasets computation of causal probabilities and subsequently creation of probabilistic causal decision treesis not a problem but for very large datasets with curse of dimensionality it is a challenging task becausehigh computational complexities are incurred. Cause and effect relationships are needed for effective decision making and also for handling data uncertainty. Artificial intelligence algorithms are mainly based on probabilistic causal data models.

3.3 Probabilistic Causal Decision Tree for Patient Dataset

Attribute	Yes	No
Age > 30	1	0
Smoking	1	0
Drinking	1	0
Gutka	1	0
Cancer	1	0

Table-1 Patient Dataset Attribute Values.

Age	Smoking	Drinking	Gutka	Cancer	Count
0	0	0	0	0	4
0	0	0	1	1	6
0	0	1	0	0	5
0	0	1	1	1	30
0	1	0	0	0	10
0	1	0	1	1	30
0	1	1	0	1	30
0	1	1	1	1	30
1	0	0	0	0	4
1	0	0	1	1	30
1	0	1	0	1	20
1	0	1	1	1	40
1	1	0	0	1	10
1	1	0	1	1	40
1	1	1	0	1	30
1	1	1	1	1	60

Table-2 Patient Dataset for Probabilistic Causal Decision Trees

Probabilistic causal decision tree is created by taking hypothetical patient dataset. Patient dataset consists of four predictor attributes and one target attribute. Count attribute represents frequencies of each row. All attributes are binary attributes whose details are shown in the Table-1. Note that probabilistic causal decision tree is constructed for the given patient training dataset using correlations and conditional causal probabilities of predictor attributes against target attribute. If patients' age > 30 is true then the value of the attribute is Yes (1); otherwise value of the attribute is false (0). If the patient has the habit of smoking then smoking attribute value is Yes (1); otherwise smoking attribute value is zero. Same is applicable for drinking and gutkapredictor attributes.

Causal probability relationships between predictor attribute and target attribute is represented by using simple conditional probability.Causal probability of age attribute given that the randomly selected patient has cancer disease is represented and then computed using the standard formula as

$$P(age = 1/cancer = 1) = \frac{P(age = 1 \cap Cancer = 1)}{P(cancer = 1)} = \frac{\{Total\ number\ of\ cancer\ patients\ whose\ age\ >\ 30\}}{\{Total\ number\ of\ cancer\ patients\}}$$

In a similar manner probabilities of other predictor attributes are computed. For example,

$$P(smoking = 1/cancer = 1) = \frac{P(smoking = 1 \cap Cancer = 1)}{P(cancer = 1)} = \frac{\{Total\ number\ of\ cancer\ patients\ who\ have\ smoking\ habit\}}{\{Total\ number\ of\ cancer\ patients\}}$$

Initially correlation is computed for each input predictor attribute with the output attribute and also correlation threshold is specified. Then causal probabilities between predictor and target attribute are computed to only predictor attributes which satisfy correlation threshold. Finally, only the attribute whose causal probability is highest is selected for node split during probabilistic causal decision tree creation. In the beginning all tuples are put in the root node. In the presently selected patient dataset total number of tuples are 380. Computed correlations between input and output attributes are shown below:

Age	Smoking	Drinking	Gutka	Correlation threshold
0.232260	0.105095	0.228422	0.390163	0.390163

Correlations values of predictor attributes

Age	Smoking	Drinking	Gutka
0	0	0	0.747899

Causality values of predictor attributes

Gutka predictor attribute has the highest correlation threshold value = 0.390163 and selected correlation threshold value is 0.390163. So, Only Gutkapredictor attribute satisfies the correlation threshold. So, causal probability is computed to only for Gutka attribute and by default causal probabilities values of all the remaining attributes is zero. In general, causal probabilities are computed for all predictor attributes whose correlation value is greater than the selected correlation threshold.Causal probability of cancer patients due to Gutka consumption = P(Gutka=1/Cancer=1) = 267/357 = 0.747899

Age	Smoking	Drinking	Cancer	Count
0	0	0	1	6
0	0	1	1	30
0	1	0	1	30
0	1	1	1	30
1	0	0	1	30
1	0	1	1	40
1	1	0	1	40
1	1	1	1	60

Table-3 dataset with Gutka = 1 (left sub group)

Initially Gutka attribute is taken as the root node attribute. All the tuples of the training dataset are stored in the Gutka root node attribute because Gutka attribute is selected as the root node attribute. All tuples in the root node are divided into left sub group and right sub group based on the values of Gutka = 1 and Gutka = 0. Gutka = 1 forms the left sub group and Gutka = 0 forms the right sub group.

Age	Smoking	Drinking	Cancer	Count
0	0	0	0	4
0	0	1	0	5
0	1	0	0	10
0	1	1	1	30
1	0	0	0	4
1	0	1	1	20
1	1	0	1	10
1	1	1	1	30

Table-4 dataset with Gutka = 0 (right sub group)

In Table-3 all tuples belong to the same class label. So, it is converted into leaf node further causal probabilities are not computed. In the next level,correlations and causal probabilities are computed for all the remaining predictor attributes against cancer target attribute as shown in Table-4.

Age	Smoking	Drinking	Correlation threshold
0.400338	0.303718	0.626240	0.607044

Correlations values of predictor attributes

Age	Smoking	Drinking
0	0	0.888888

Causality values of predictor attributes

Out of three predictor attributes, only drinking attribute satisfies correlation threshold. So, causal probability is computed to only drinking attribute and causal probabilities of other attributes are set to zero. Hence, drinking attribute is taken as the node split attribute for the right partition of data shown in Table-4. Left sub group is created for drinking =1 and right sub group is create for drinking = 0.

Age	Smoking	Cancer	Count
0	0	0	5
0	1	1	30
1	0	1	20
1	1	1	30

Table-5 left partition for drinking = 1

Age	Smoking	Cancer	Count
0	0	0	4
0	1	0	10
1	0	0	4
1	1	1	10

Table-6 right partition for drinking = 0

Similar procedure is applied for finding split attributes in the next respective higher levels of probability causal decision tree construction. The final probabilistic causal decision tree is shown in Figure-2 with post pruned version after its creation.

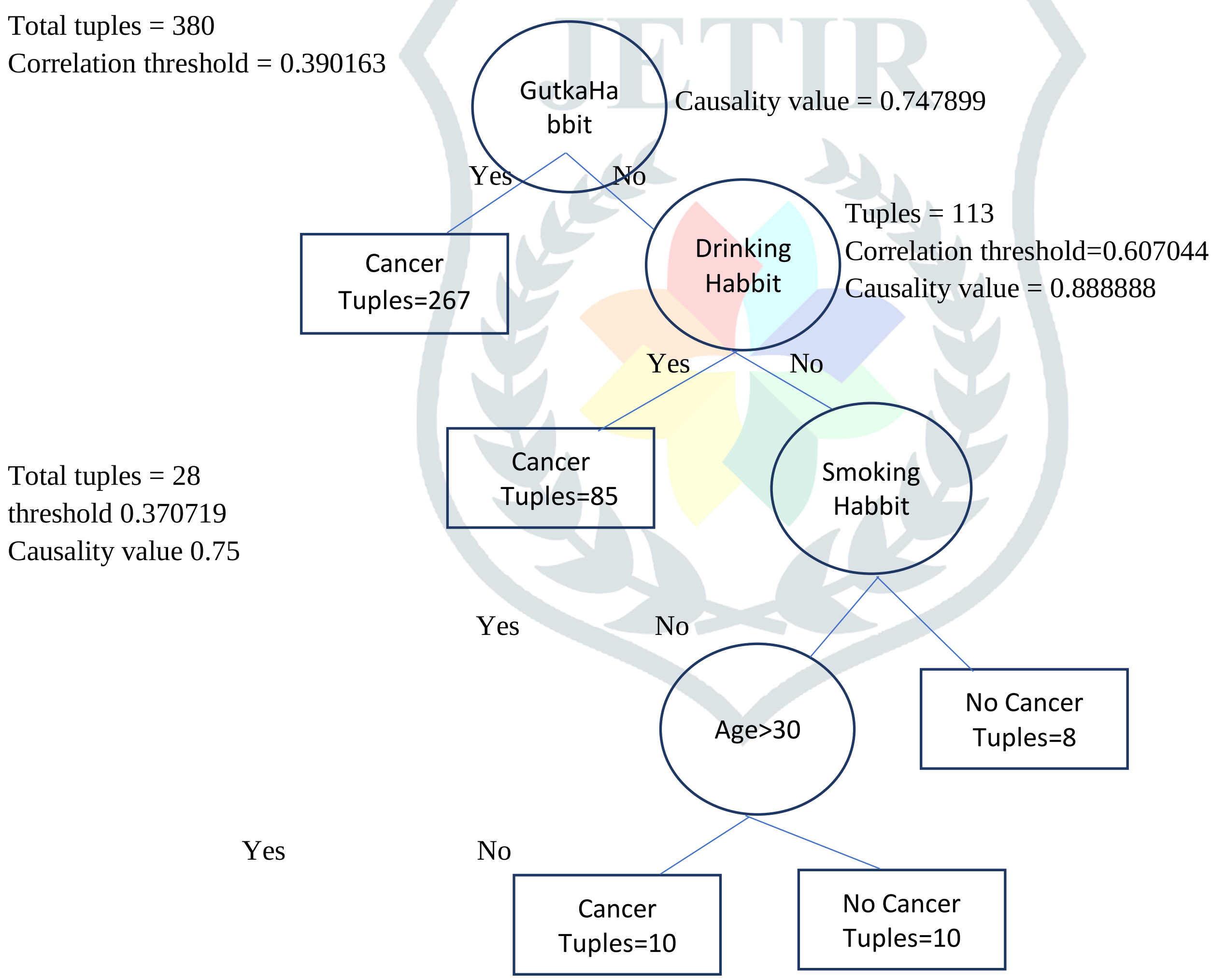

Figure-1 Probabilistic Causal Decision Tree for Patient Dataset.

3.4 Causal Decision Tree for Adult UCI machine learning dataset

The Adult data set			
Attributes	Yes	No	Comment
Age < 30	14,515	34,327	Young
Age > 60	3,606	45,236	Old
Private	33,906	14,936	private company employee
Self-emp	5,557	43,285	self-employment
Gov	6,549	42,293	government employer
Education-num > 12	12,110	36,732	bachelor or higher
Education-num < 9	6408	42,434	education years
Prof	23,874	24,968	professional occupation
White	41,762	7,080	Race
Male	32,650	16,192	Gender
Hours > 50	5,435	43,407	weekly working hours
Hours < 30	6,151	42,691	weekly working hours
US	43,832	5,010	Nationality
> 50k	11,687	37,155	annual income, outcome

Table-7 Adult UCI machine learning dataset attributes details.

Adult dataset consists of 13 predictor attributes and 1 target attribute. Initially correlations of all predictor attributes with respect to target attribute are computed and correlation threshold is selected. Causal probabilities are computed to only those predictor attributes whose correlation is greater than the selected correlation threshold. Then the one predictor attributes whose causal probability is the highest is selected as the best split attribute of the probabilistic causal decision tree and then tuples in the current node are partitioned into left partition and right partition based on the values of the split attribute. Similar procedure is applied at each level of the probabilistic causal decision tree. Final probabilistic causal decision tree is shown in Figure-2.

Attributes	Yes	No
Age < 30	1	0
Age > 60	1	0
Private	1	0
Self-emp	1	0
Gov	1	0
Education-num > 12	1	0
Education-num < 9	1	0
Prof	1	0
White	1	0
Male	1	0
Hours > 50	1	0
Hours < 30	1	0
US	1	0
> 50k	1	0

Table-8 Attribute binary values of the Adult dataset

Probabilistic Causal Decision Tree Construction for Adult dataset

Initially correlations and probabilistic causal values are computed and shown in the following:
correlation threshold value is = 0.2837

Attributes	Correlation	Causal probability
Age < 30	0.2837	0.06466
Age > 60	0.0023	0
Private	0.1178	0
Self-emp	0.0979	0
Gov	0.0559	0
Education-num > 12	0.2383	0

Education-num < 9	0.0789	0
Prof	0.1829	0
White	0.0823	0
Male	0.2161	0
Hours > 50	0.1389	0
Hours < 30	0.1479	0
US	0.0391	0

Causal probability is computed to only for age < 30 attribute because whose correlation is greater than or equal to the correlation threshold. In case if more than one attribute satisfies correlation threshold then causal probability is computed for all those attributes and then highest causal probability node is selected as the split node attribute and then tuples are partitioned according to the values of the split attribute.

$$P(age < 30 = 1/income > 50k = 1) = \frac{P(age < 30 = 1 \cap \ income > 50k = 1)}{P(cancer = 1)}$$
$$= \frac{\{Total\ number\ of\ persons\ with\ age < 30 = 1\ and\ their\ income > 50k \geq 1\}}{\{Total\ number\ of\ persons\ whose\ income > 50k = 1\}}$$
$$P(private = 1/income > 50k = 1) = \frac{P(private = 1 \cap \ income > 50k = 1)}{P(cancer = 1)}$$
$$= \frac{\{Total\ number\ of\ persons\ with\ private = 1\ and\ their\ income > 50k \geq 1\}}{\{Total\ number\ of\ persons\ whose\ income > 50k = 1\}}$$

So, age < 30 is selected as a root node and split attribute of the probability causal decision tree and then all tuples are stored in the root node. Now tuples are divided into two partitions, left partition with age < 30 is true and right partition with age < 30 is false. Similar procedure is applied in the next levels of the probability causal decision tree creation. The final probability causal decision tree for the adult dataset is in Figure-2.

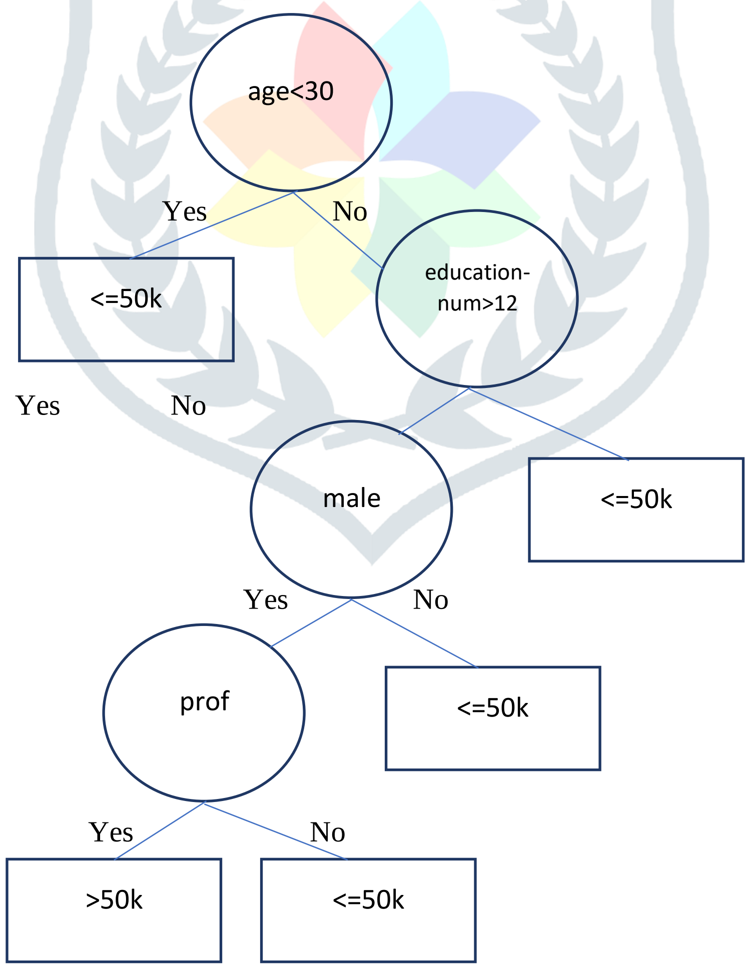

Figure-2 Probabilistic Causal Decision Tree for the Adult UCI machine learning dataset

Conclusion

Probabilistic causal decision trees are constructed for two training datasets. Both conditional probability and probabilistic causality are used during decision tree creation procedure and final output results are compared. Both methods have produced the same results. In the feature, comparisons will be performed by using many possible mathematical causality formulae and new formulas will be traced out for the same.

References

[1] Hichem Debbi; Mustapha Bourahla, Causal analysis of probabilistic counterexamples
Published in: 2013 Eleventh ACM/IEEE International Conference on Formal Methods and Models for Codesign (MEMOCODE 2013)
[2] Sunny Raj; Sumit Kumar Jha, Predicting Success in Undergraduate Parallel Programming via Probabilistic Causality Analysis, Published in: 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)
[3]Xiaolu Chen, Jing Wang, Jinglin Zhou,"Process Monitoring Based on Multivariate Causality Analysis and Probability Inference", special section on data-driven monitoring, fault diagnosis and control of cyber-physical system, Received December 12, 2017, accepted January 13, 2018, date of publication January 23, 2018, date of current version March 9, 2018
[4] Qin Zhang, Chunling Dong, Yan Cui, and Zhihui Yang, "Dynamic Uncertain Causality Graph for Knowledge Representation and Probabilistic Reasoning: Statistics Base, Matrix, and Application", IEEE Transactions on neural networks and learning systems, vol. 25, no. 4, April 2014
[5] Qin Zhang, Dynamic Uncertain Causality Graph for Knowledge Representation and Probabilistic Reasoning: Directed Cyclic Graph and Joint Probability Distribution
Published in: IEEE Transactions on Neural Networks and Learning Systems (Volume: 26, Issue: 7, July 2015)
[6] Nimish Shah; Laura I. GalindezOlascoaga; WannesMeert; Marian Verhelst, "Acceleration of probabilistic reasoning through custom processor architecture", Published in: 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Date Added to IEEE *Xplore*: 15 June 2020
[7] Ming Li, Kefeng Li, "Causality-Based Attribute Weighting via Information Flow and Genetic Algorithm for Naive Bayes Classifier", IEEE Access, Received September 15, 2019, accepted October 4, 2019, date of publication October 15, 2019.
[8] RandaHerzallah, "A Fully Probabilistic Design for Tracking Control for Stochastic Systems With Input Delay", Published in: IEEE Transactions on Automatic Control (Volume: 66, Issue: 9, September 2021)
[9]Guanchao Feng; J. Gerald Quirk; Petar M. Djurić, "Detecting Causality using Deep Gaussian Processes", Published in: 2019 53rd Asilomar Conference on Signals, Systems, and Computers
[10] Zhenxu Zhou; Qin Zhang, "Model Event/Fault Trees With Dynamic Uncertain Causality Graph for Better Probabilistic Safety Assessment
", Published in: IEEE Transactions on Reliability (Volume: 66, Issue: 1, March 2017)
[11]L. James Valverde A, "Probabilistic Causality and Its Applications to Risk Analysis", Part of the Advances in Risk Analysis book series (AIRA,volume 9)
[12] Probabilistic Graphical Models N. Srihari Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA
[13] Foundations for a Probabilistic Theory of Causal Strength, Jan Sprenger, February 3, 2016
[14] Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press, Cambridge.
[15] Williamson, J. (2005). Bayesian nets and causality: philosophical and computational foundations. Oxford University Press, Oxford
[16]Yunjun Gao, Qing Liu, Gang Ch, "Finding Causality and Responsibility for Probabilistic Models", Published in: 2017 IEEE 33rd International Conference on Data Engineering (ICDE)
[17]Paul Humphreys, Probabilistic Causality and Multiple Causation, Published online by Cambridge University Press: 21 March 2022
[18] Galavotti, M.C., Suppes, P., and Costantini, D., (eds.), 2004, Stochastic Causality, Stanford, CSLI Publications.
[19] Tina Grotzer, Learning Causality in a Complex World: Understandings of Consequence,
Rowman & Littlefield Education (2012)
[20] Dowe, P. (2000). Physical causation. Cambridge University Press, Cambridge.
[21] Reiss, J. (2007). Time series, nonsense correlations and the principle of the common cause. In Causality and probability in the sciences, pages 179–196. College Publications, London
[22] Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. International Studies in the Philosophy of Science, 21(2):157–170.