IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

An Efficient First Order Logical Casual Decision Tree in High Dimensional Dataset

S. Preethi[1*] and C. Rathika[2]

[1*]Dept. of Computer Science, Sri Ramakrishna college of Arts and Science for women, Coimbatore, India
[2]Dept. of Computer Science ,Sri Ramakrishna college of Arts and Science for women ,Coimbatore, India

**Corresponding Author: preethisenthil1969@gmail.com*

Available online at: www.ijcseonline.org

Received: 31/Jan//2018, Revised: 08/Feb2018, Accepted: 20/Feb/2018, Published: 28/Feb/2018

***Abstract*:** Uncovering causal interactions in data is a most important objective of data analytics. Causal relationships are usually exposed with intended research, e.g. randomised controlled examinations, which however are costly or insufficient to be performed in several cases. In this research paper aims to present a new Casual Decision tree structure of first-order logical casual decision tree called FOL-CDT structure. The proposed method follows a well-recognized pruning approach in causal deduction framework and makes use of a standard arithmetical test to create the causal relationship connecting a analyst variable and the result variable. At the similar instance, by taking the advantages of standard decision trees, a FOL-CDT presents a compact graphical illustration of the causal relationships with pruning method, and building of a FOL-CDT is quick as a effect of the divide and conquer strategy in use, making FOL-CDTs realistic for representing and resulting causal signals in large data sets.

Keywords: Data Mining, First order Logical, Decision Tree, Pruning, Classification.

I. INTRODUCTION

Data mining is the process of finding previously un known patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.

Detecting causal relationships in data is an important data analytics task as causal relationships [1] can provide better insights into data, as well as actionable knowledge for correct decision making and timely intervening in processes at risk.

Causal relationships can provide better insights into data, as well as actionable knowledge for correct decision making and timely intervening in processes at risk, therefore detecting causal relationships in data is an important data analytics task. Randomized controlled trials (RCTs) are considered as the gold standard for causal inference in many areas such as medicine and social science [2]. However, it is often impossible to conduct RCTs due to cost or ethical concerns. Causal relationships can also be found by observational studies, such as cohort studies and case control studies [3]. An observational study takes a causal hypothesis and tests it using samples selected from historical data or collected passively over the period of time when observing the subjects of interest. Therefore observational studies need domain experts' knowledge and interactions in data selection or collection, and the process is normally time consuming.

Currently there is a lack of scalable and automated methods for causal relationship exploration in data. These methods should be able to find causal signals in data without requiring domain knowledge or any hypothesis established beforehand. The methods must also be efficient to deal with the increasing amount of big data.

Classification methods are fast and have the potential to become practical substitutes for finding causal signals in data since the discovery of causal relationships [4] is a type of supervised learning when the target or outcome variable is fixed. Decision trees [5] are a good example of classification methods, and they have been widely used in many areas, including social and medical data analyses.

However, classification methods are not designed with causal discovery in mind and a classification method may find false causal signals in data and miss true causal signals [6]. For example, Figure 1 shows a decision tree built from a hypothesized data set of the recovery of a disease. Based on the decision tree, we may conclude that the use of Tinder (a match making mobile app) helps cure the disease. However, it is misleading since the majority of people using Tinder are young whereas most people not using Tinder are old [7]. Young people will recover from the disease anyway and old people have a lower chance of recovery. This misleading

decision tree is caused by an unfair comparison between the two different groups of people. It may be a good classification tree to predict the likelihood of recovery, but it does not imply the causes of recovery and its nodes do not have any causal interpretation.

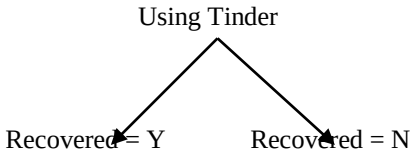

Figure 1. A simple decision tree

In the above figure 1 represents the relationship between using Tinder and the recovery of a disease, the effect of other variables such as age, gender, and health condition of patients (who may or may not use Tinder) should be considered. The objective is not simply to maximize the difference of the conditional probabilities of recovered and not recovered conditioning on the use of Tinder when a classifier is being required [8].

In this paper, we design a first order logical causal decision tree (FOL-CDT) where nodes have causal interpretations. As presented in the following sections, our method follows a well-established causal inference framework, the potential outcome model, and it makes use of a classic statistical test. The proposed FOL-CDT is practical for uncovering causal signals in high dimensional data.

II. RELATED WORK

In [9] authors discussed a Causal effects are defined as comparisons of potential outcomes under different treatments on a common set of units. Observed values of the potential outcomes are revealed by the assignment mechanism—a probabilistic model for the treatment each unit receives as a function of covariates and potential outcomes. Fisher made tremendous contributions to causal inference through his work on the design of randomized experiments, but the potential outcomes perspective applies to other complex experiments and nonrandomized studies as well.

In [10] authors provided a brief overview to four major types of causal models for health-sciences research: Graphical models (causal diagrams), potential-outcome (counterfactual) models, sufficient-component cause models, and structural-equations models. The paper focuses on the logical connections among the different types of models and on the different strengths of each approach. Graphical models can illustrate qualitative population assumptions and sources of bias not easily seen with other approaches; sufficient-component cause models can illustrate specific hypotheses about mechanisms of action; and potential-outcome and structural-equations models provide a basis for quantitative analysis of effects.

In [11] authors proposed a statistical methodology is presented for analyzing retrospective study data, including chi-square measures of statistical significance of the observed association between the disease and the factor under study, and measures for interpreting the association in terms of an increased relative risk of disease. An extension of the chi-square test to the situation where data are sub-classified by factors controlled in the analysis is given. A summary relative risk formula, R, is presented and discussed in connection with the problem of weighting [12] the individual subcategory relative risks according to their importance or their precision.

In [13] authors provided a new complexity results for algorithms that learn discrete-variable Bayesian networks from data. Their results apply whenever the learning algorithm uses a scoring criterion that favors the simplest structure for which the model is able to represent the generative distribution exactly. The results therefore hold whenever the learning algorithm uses a consistent scoring criterion and is applied to a sufficiently large dataset. The authors showed that identifying high-scoring structures is NP-hard, even when any combination of one or more of the following hold: the generative distribution is perfect with respect to some DAG containing hidden variables; we are given an independence oracle; we are given an inference oracle; we are given an information oracle; we restrict potential solutions to structures in which each node has at most k parents, for all k>=3.

In [14] authors considered a variable selection in high-dimensional linear models where the number of covariates greatly exceeds the sample size. We introduce the new concept of partial faithfulness and use it to infer associations between the covariates and the response. Under partial faithfulness, we develop a simplified version of the PC algorithm, which is computationally feasible even with thousands of covariates and provides consistent variable selection under conditions on the random design matrix that are of a different nature than coherence conditions for penalty-based approaches like the lasso.

In [15] authors proposed an approach to mine causal rules in large databases of binary variables. The corresponding method expands the scope of causality discovery to causal relationships with multiple cause variables, and authors utilized partial association tests to exclude non-causal associations, to ensure the high reliability of discovered causal rules. Furthermore an efficient algorithm is designed for the tests in large databases. The authors assessed the method with a set of real-world diagnostic data. The results show that our method can effectively discover interesting causal rules in large databases.

In [16] authors studied a decision making in environments where the reward is only partially observed, but can be modeled as a function of an action and an observed context. This setting, known as contextual bandits, encompasses a wide variety of applications including health-care policy and Internet advertising. A central task is evaluation of a new policy given historic data consisting of contexts, actions and received rewards. The key challenge is that the past data typically does not faithfully represent proportions of actions taken by a new policy.

III. RESEARCH METHODOLOGY

In this paper, we have proposed a new system FOL-CDT for detection and extraction causal relationships in high dimensional data using Decision Tree classifiers. The aim of this system is to detect the relevant and pure events and extracted portion of casual relationships effectively and efficiently. Our method follows a well established First order logical causal inference framework with pruning and makes use of a classic statistical test. The method is practical for finding causal signals in large data sets. The proposed architecture diagram is described in figure 2.

Data Preprocessing

An important step in the data mining process is data preprocessing [17]. Henceforth, we remember a number of data preprocessing methods that have been used jointly with ensemble approach. A re-sampling methods that learn the consequence of altering class allocation to compact with imbalanced data-sets, where it has been empirically established that the submission of a preprocessing step in order to stability the class division is usually a positive solution.

Noise and the amount of data are reduced and extra information is added during preprocessing to the data. The data is transformed into the occurrence domain, where the noise is reduced. Data cleaning schedules effort to clean data by satisfying in smoothing noisy data, missing values, recognizing or removing anomalies, and resolving unpredictability.

Real world data are generally affects these things namely

- **Incomplete:** missing attribute values, missing certain features of interest;
- **Inconsistent:** containing discrepancies and inconsistencies suitable to data combination, where a given feature can have unusual names in special databases. Duplicates may also exist;
- Incomplete data is a unsystematic error or variance in a considered variable. Incomplete data could be suitable to defective data collection instruments, data entry problems, data diffusion problems, technology restriction.

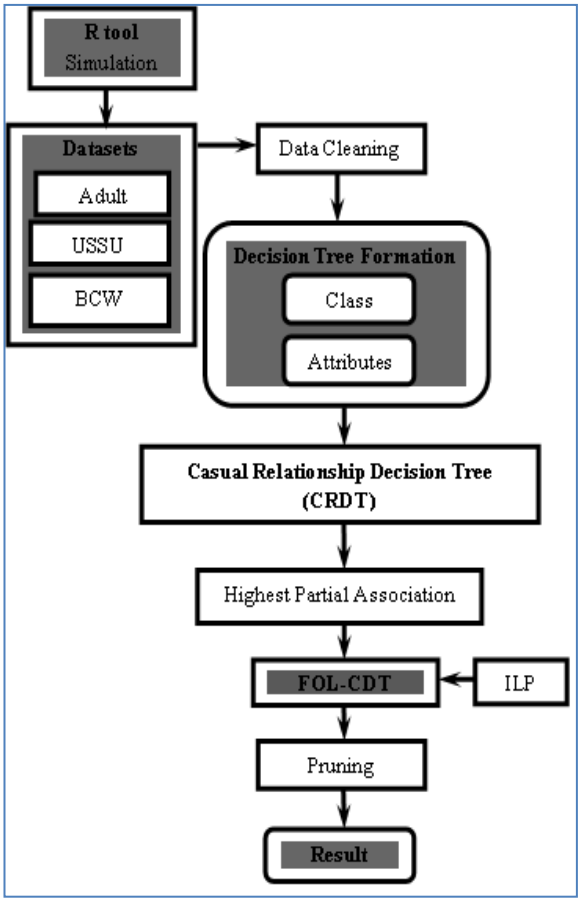

Figure 2. Proposed Architecture

In figure 2, real-world datasets (Adult, USSU and BCW) are taken as inputs in R tool simulation for FOL-CDT process. The first process is called data cleaning or data preprocessing method performs reduced noise, junk and missing values present in the dataset. After that preprocessing process, decision tree formation process is split the dataset into number of classes and number of attributes. For that attributes and classes, initial decision tree is plotted. In third process, casual relationship decision tree (CRDT) is form depends on initial decision tree. In CRDT weighted highest partial association to be predicted in weak learner. Finally FOL-CDT performs the inductive logic programming (ILP) for the casual decision making in terms of the background knowledge (Pruning strategy).

Decision Trees Formation

Decision tree learning is a general method used in data mining. The majority of the profitable packages present

complex tree learning algorithms, but they are extremely much expensive. Decision tree algorithms produce tree-structured classification rules, which are written in a structure of combinations and disjunctions of attribute values (or feature values). These classification policies are created through 1)Choosing the best dividing feature based on a positive measure, 2) separating input data depending on the finest splitting feature values, then 3) recursively duplicate this process until certain ending measure are met. The selected best splitting attribute changes not only the present division of input data, but also the successive best splitting features as it changes the sample allocation of the resultant division. Thus, the best splitting feature collection is possibly the the majority important step in decision tree structure, and dissimilar names are given for decision trees that use dissimilar splitting criteria, for example, C4.5 and ID3 for Shannon entropy-based splitting measures such as and Information Gain ratio and CART for the Gini impurity measure.

Different splitting criteria use their own impurity measures, which are used to calculate "achievable" impurity reduction after a possible split. Consider a nominal feature *X* and target class *Y*.

Causal Relationship Decision Tree (CRDT)

The CRDT consists in training special classifiers with bootstrapped models of the unique training data-set. That is, a original data-set is shaped to train each classifier by arbitrarily representation (with substitute) samples from the new data-set (usually, preserving the original data-set size).

The CRDT deals with class imbalance limitations due to its ease and excellent generalization ability. The hybridization data preprocessing methods is typically simpler than their combination in boosting. A CRDT technique does not involve re-computing some type of weights; therefore, neither is essential to adjust the weight update method nor to modify calculations in the algorithm. In these techniques, the key factor is the approach to collect every bootstrap replica (Algorithm 1), that is, how the class imbalance difficulty is compact to obtain a valuable classifier, without overlooking the importance of the diversity.

A simple method to conquer the class imbalance problem in every casual relationship is to obtain into description the classes of the samples when they are randomly drawn from the novel data-set. Therefore, instead of performing a random sampling of the entire data-set, an oversampling method can be approved out prior to training each classifier. This process can be developed in two ways. Oversampling consists in rising the amount of minority class samples by their duplication, all common class samples can be incorporated in the new bootstrap, but a different option is to resample them trying to amplify the diversity. Note that in all samples will possibly take part in at least one bag, but every bootstrapped replica will contain many more instances than the original data-set.

Algorithm 1: Causal Relationship Decision Tree (CRDT)
Input: S: Training set; T: Number of iterations; n: Casual relationship size; I: Weak learner
Output: CRDT classifier: $CRDT(x) = \sum_{t=1}^{T} CRDT_t(x)$ sign where $CRDT_t \in$ [-1, 1] are the induced classifiers
for t = 1 to T do
Wt ←find attribute *W* with the highest partial association (n, S)
$CRDT_t$ ←I(Wt)
end for

First-Order Logical casual Decision Trees in Lift-Boosting approach

The Decision Tree (DT) approach is more powerful for data mining classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and logical tree. In DT pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications.

A first-order logical casual decision tree (FOL-CDT) is a binary decision tree in which (1) the nodes of the tree contain a conjunction of literals, and (2) different nodes may share variables, under the following restriction: a variable that is introduced in a node (which means that it does not occur in higher nodes) must not occur in the right branch of that node.

It is shown that first order logical decision trees have an advantage with respect to expressiveness over the kind of logic programs that are normally induced by most inductive logic programming (ILP) systems. We call the latter flat logic programs, because the target predicate is defined immediately in terms of the background knowledge, without any intermediate predicates being defined (in the latter case we would call them layered). Given a set of predicates, the set of all first order logical decision trees is a strict superset of the set of flat logic programs. In order to enable an ILP system to produce any theory that is represented by a first order decision tree, the ILP system should either have the possibility to invent auxiliary predicates (so that layered logic programs can be induced), use a format for its theories that allows the use of both universal and existential quantification, or induce Prolog programs with cuts (so-called first order decision lists).

The FOL-CDT technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

IV. SIMULATION RESULTS

The simulation studies work has been evaluated The Adult data set was retrieved [18] . And it is an extraction of 1994 USA census database. It is a well known classification data set used in predicting whether a person earns over 50 K or not in a year. We recoded the data set to make the causes for high/low income more clearly and easily understandable. The objective is to find the causal factors of high (or low) income.

Table 1: A Comparison of Classification Accuracy of FOL-CDT, CDTs and C4.5 Trees

Dataset	C4.5	CDT	FOL-CDT
Adult	80.80%	80.64%	81.65%
USSU	81.17%	81.05%	82.73%
BCW (Original)	94.71%	91.7%	95.21%
Car Evaluation	92.36%	93.98%	95.07%
Mushroom	100.00%	89.56%	92.21%
Sick	98.00%	94.25%	95.51%

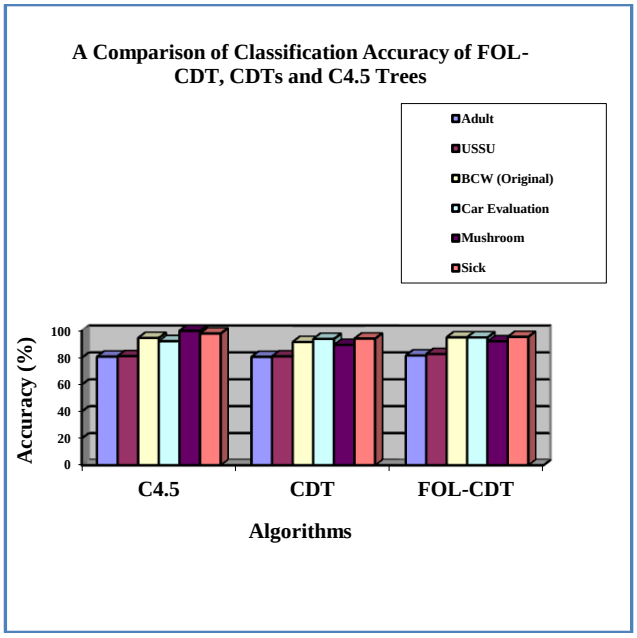

Figure 3. A Comparison of Classification Accuracy of FOL – CDT, CDTs and C4.5 Trees

V. CONCLSUION

In this paper, presents a new Casual Decision tree framework of first-order logical casual decision tree called FOL-CDT that consistently results in higher Accuracy values over the class imbalanced data sets. The proposed work modifies traditional pruning rules in FOL_CDT algorithm to directly reflect an evaluation metric based on conditions. The First order logical relationship tree is a good generalization for unobserved instance, only if the instances are described in terms of features that are correlated with the target concept. The Decision is provide a clear indication of which fields are most important for prediction or classification.

REFERENCES

[1] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," Statistical Sci., Vol. 25, No. 1, pp. 1–21, 2010

[2] N. Cartwright, "What are randomised controlled trials good for?" Philosophical Studies, vol. 147, no. 1, pp. 59–70, 2009.

[3] P. R. Rosenbaum, Design of Observational Studies. Berlin, Germany: Springer, 2010.

[4] R. P. Rosenbaum and B. D. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," J. Amer. Statistical Assoc., Vol. 79, No. 387, pp. 516–524, 1984.

[5] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," J. Nat. Cancer Inst., Vol. 22, No. 4, pp. 719–748, 1959.

[6] 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.

[7] J. Foster, J. Taylor, and S. Ruberg, "Subgroup identification from randomized clinical trial data," Statistics Med., Vol. 30, No. 24, pp. 2867–2880, 2011

[8] 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.

[9] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decision," J. Amer. Statistical Assoc., Vol. 100, No. 469, pp. 322–331, 2005.

[10] S. Greenland and B. Brumback, "An overview of relations among causal modelling methods," Int. J. Epidemiology, Vol. 31, pp. 1030–1037, 2002

[11] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," J. Nat. Cancer Inst., Vol. 22, No. 4, pp. 719–748, 1959.

[12] B. K. Lee, J. Lessler, and E. A. Stuart, "Improving propensity score weighting using machine learning," Statistics Med., Vol. 29, No. 3, pp. 337–46, 2010.

[13] D. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," J. Mach. Learn. Res., Vol. 5, pp. 1287–1330, 2004.

[14] M. K. P. B€uehlmann and M. Maathuis, "Variable selection for highdimensional linear models: Partially faithful distributions and the PC-simple algorithm," Biometrika, Vol. 97, pp. 261–278, 2010.

[15] Z. Jin, 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.

[16] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," in Proc.Natl Academy Sci., Vol.113, No. 27, pp. 7353–7360, 2016.

[17] P.T.Kavitha, Dr.T.Sasipraba , Knowledge Driven HealthCare Decision Support System using Distributed Data Mining, Indian

Journal of Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Jun-Jul 2012.

[18] K. Bache and M. Lichman, " Evolutionary Learning of concepts" , Journal of Computers and Communications,Vol.2 No. 8, June 27, 2014.

Author's profile

Ms. S.Preethi, received B.Sc degree from Sri Ramakrishna college of Arts and Science for Women in 2014. She has received M.Sc degree in 2016 from Sri Ramakrishna college of Arts and Science for Women and M.Phil degree from Sri Ramakrishna college of Arts and Science for Women She has published 1 in National Conference and 1 in International Conference. Her interested area is Data Mining, Cloud computing, Distributed computing.

Mrs. C. Rathika is currently working as an Assistant Professor in the Department of Computer Science at Sri Ramakrishna College of Arts and Science for Women. She is currently pursuing her Doctorate degree in computer science at Sri Ramakrishna College of Arts and Science for Women. She received her Master of Philosophy degree from Bharathidhasan University in 2007.She received her Master's degree in Applied Sciences Computer Technology from Maharaja Engineering College and Bachelor's degree in Computer Science From Vidhyasagar College of Arts and Science. She has 14 years of experience in teaching. Her research areas include Data mining and Big Data analytics. She has guided 6 M.Phil scholars, around 40 UG projects and 20 PG projects. She acted as an Adjunct Faculty for a period of one month in Texila American University, South America to develop the curriculum for the course Diploma in Information Technology. She acted as Technical Trainer for E-Mudhra Technologies, Bangalore. She published papers in 6 papers in International Journals and presented papers in international and national Conferences.

