Data Analytics and Mining in Healthcare with Emphasis on Causal Relationship Mining

Sreeraman Y, S. Lakshmana Pandian

Abstract*: High volumes and varieties of data is piling every day from healthcare and related fields. This big data sources if managed and analysed properly will provide vital knowledge. Data mining and data analytics have been playing an important role in extracting useful information from healthcare and related data sources. The knowledge extracted from these data sources guiding patients and healthcare personnel towards improved health conditions. Analytical techniques from statistics, functionalities from data mining and machine learning already proved their capability with significant contributions to healthcare industry. The dominant functionality of data mining is classification which has been in use in mining healthcare data. Though classification is a good learning technique it may not provide a causation model which will be a reliable model for better decision making particularly in the medical field. The present models for causality have limitations in terms of scalability and reliability. The present study is targeted to study causal models for causal relationship mining. This study tried to conclude with some proposals for causal relationship discovery which are efficient, reliable and scalable. The proposed model is going to make use of some qualities of decision trees along with statistical tests and analytics. It is proposed to build the learning models on healthcare big data sources.

***Keywords* : Decision Tree, Causal Models, Causal Relationship Mining, Classification.**

I. INTRODUCTION

Data analytics is the discipline which examines big data to derive conclusions with standard analytical techniques. Data mining has to do with the discovery of vital, valid, novel, and ultimately understandable knowledge from data. This knowledge extraction processes are clearly common in other disciplines like statistics, machine learning, or data analytics. Data mining is different when the size of the data is considered. Today the research community is taking the challenge of data with high volume, multitude varieties and high speed accumulation of data records. Data analytics can be called as data analysis where the process undergone inspection, cleansing, transforming and modelling of data with the main goal of discovering vital knowledge from data. This knowledge is a guide to arrive conclusions at various contexts of data. One of the rich sources of data for analysis and analytics is the data from healthcare field. This area is an upcoming area with advanced features that support the users to make good decisions to improve the service quality. Healthcare analytics can offer potential outcomes that support to upgrade the quality of service at several stages of the system. Using the obtained knowledge treatment costs can be controlled, preventable diseases and deaths can be avoided, and ultimately the quality of life can be improved. Today, huge amounts of data from heterogeneous sources are being generated from health care industry. The sources cover: electronic patient records, diagnostic reports, hospital operations, medical equipment, patient management systems, billing records, and so forth. The near future will depend on wearable sensors, mobile devices, government health records, insurance records and social network sources to extract information for decisions in clinical trials. This key resource with large amount of data can boost the service delivery, quality of research in healthcare, success rate in medical trials .The analysis of this data supply good support to doctors in providing better care for their patients. By analyzing the medical records data, one can discover homogeneous groups of patients. The analysis provides the information about common prescriptions given by doctors to patients with a given disease. It is also possible to identify priorities among the treatment options. A data scientist can also find the relationship between drugs and medical conditions. Classification and clustering are the prime techniques of data mining that can provide rich knowledge through data grouping/classification. Decision tree is a model to learn the behaviour of data under consideration. A model is prepared through a training data set and the same guides the decision making process. The model is evaluated by processing the test data. Though decision trees are powerful tools for data classification, they cannot identify a cause for an outcome. Medical data needs causal inference. The underlying cause for a particular medical condition effects the treatment options. Therefore, the concept of causal decision trees is a key to the present work, where the goal is to infer the causes of a particular event can be extracted.

II. CAUSAL RELATIONSHIPS

Causal relationships among variables provide information regarding the complex associations among the components of a system. Causal relationships are powerful than correlations among variables.

Manuscript published on November 30, 2019.
* Correspondence Author

Sreeraman Y *, Department of CSE, Pondicherry Engineering College, Puducherry, India. Email: sramany@gmail.com
S. Lakshmana Pandian, Department of CSE, Pondicherry Engineering College, Puducherry, India. Email: lpandian72@pec.edu

The causal relationships reflect the cause and effect rather than a simple relationship and therefore they are key elements in prediction and reasoning.

This information avoids erroneous decisions or policy making. In healthcare informatics, causal relationships help the community by providing possible causes of diseases and assists in better diagnosis and discovery of cures for diseases.

One of the most important problems to be dealt with in data mining ,particularly to apply upon the data where the causes influence a lot is Causal inference in which causes are extracted from the outcome. Designed experimental support is needed to make absolute statements about cause and effect. The set up of these experiments is costly, and sometimes infeasible. Naturally, people recognize causal relationships in their life journey. One may infer the cause of an event based on observation. Hard work leads to good results. Healthy intake of food causes better health. Sometimes the same instance may be a cause and an effect as well. Causal relationships help policymakers, practitioners, and scientists by providing them the estimations about cause and effect. Among the sets of possible cause and effect relation pairs most of the candidates were neither feasible nor desirable and a few only are credible. The recent past research in computer science has been attracted by causal discovery methods from observational data. Currently, Bayesian network techniques contributing the core of the methodologies for causal discovery in computer science. Structural equation models (SEMs) are another vital means following the way.

The multitude nature of Big Data increases the chance of higher amount of information. The large quantities of data sources facilitate a higher sample and effect [17]. The heterogeneity in formats of medical data with mostly unstructured nature is a challenge to data analysis [16].The attempts to analyse such data sources brought significant discoveries [3] [29].

III. HEALTHCARE AS A BIG SOURCE FOR DATA ANALYTICS

Healthcare is a huge sector providing support to large scale employment and contributing high revenue. The sector covers a large range of medical sub sectors that include hospitals, medical devices, clinical trials, diagnosis, telemedicine, medical tourism, health insurance, and medical equipment. This sector in India is growing at a brisk pace.

The nature of health data is evolving fast. The analytical techniques are emerged to deal with complex data with large volumes, high velocity and varieties.

A number of the models have been investigated and successfully deployed for predicting the data nature in many of the clinical practices [7]. Prediction is more important in health diagnosis for suggesting right treatment.

Supervised machine learning techniques are successfully being applied for clinical prediction tasks and those successful techniques are categorized into three fundamental classes: the first class includes linear-regression, logistic-regression, Bayesian-models and many other statistical models, second class contains some of the machine learning techniques and some of the data mining techniques and the third class contains survival models with different features of well-defined aims in order to predict survival related outcomes. The actual need is to look forward to the time estimation for an even to be occurred. These models are gaining popularity and used in the situation of medical data analysis in terms of predicting the patient’s survival time. Various test techniques are available for the examination of these models.

Analytical techniques using big data paradigm against the sources of patient health record information can give valuable outputs for patient’s healthcare. The population growth, the mortality rate, the modern health care techniques are all moving around the data. This big data source can provide much about the health conditions of individuals and can predict the milestones of a person. If all the data regarding health information is analyzed properly one can get valid outcomes. The need of big data process is in inevitable.

The size and heterogeneity of the data being collected reached the stage where the sampling techniques cannot be employed successfully [18]. Most of the information about patients is encoded in the form of clinical notes. These notes are typically stored in an unstructured data format and are the backbone of much of healthcare data. These records manage information with respect to prescriptions from the doctors, suggestions from the specialists. The record generation and management make use of modern information gathering techniques. The data piled through these data sources is a challenge to analysts due to the complexity involved in pre-processing such medical records. The main bottleneck is the nature of data formats most of which are not structured. Many techniques from Natural language processing (NLP) and entity extraction are providing better means to deal with the data. [20].

IV. MEDICAL DATA ANALYTICS

Now a day’s very large sizes of medical data sets are available but the knowledge available from such data sets is minimal. The sizes of medical datasets are growing exponentially and these data sets contain valuable information that is useful for effective decision making during medical diagnosis process of patients. Reliable diagnosis result when intelligent data analytics algorithms are used in knowledge processing. Discovering useful knowledge from medical data sets and other related medical fields is a very difficult task. Latest trend is to automate all the techniques useful for knowledge discovery particularly from medical data sets and other medical related areas. Medical Data analytics means applying data analytics principles, techniques, methods, ideas and plans on the medical data sets particularly in the medical area belonging to the electronic health records. The goal of medical data analytics is how to automate all the existing medical diagnosis methods

The steps in the medical data analytics include:

1. Collect medical data from hospital raw database
2. Pre process all the targeted data sets
3. Perform the feature selection

4. Apply the data analytics technique to the selected data for obtaining useful pattern and result.

Various types of tools and techniques available in the data analytics include:

A. Classification

Classification is a supervised learning model that is generally builds against labelled data. The model built after classification suggests a label for unlabeled data from the same context from which the model is built. Classification is the most useful and popular data analytical technique used in classifying data in real applications such as medicine, research, agriculture, railway, banking sector, healthcare and so on. Classification simplifies the task of causal inference mechanism by providing the useful groups for decision making. When a large data source is classified into meaningful groups then it makes the further study simplified. Classification is a data analytical techniques used for analyzing very large datasets. Classification model categorizes the data.

Classification proceeds in two steps. In the first step classifies or model is constructed using the values of the attributes given in the training data set. In the second step classifier is used for testing the test tuples. Classification is a supervised learning technique where class labels are provided in the training data sets classifier predicts or maps the class label of the test tuple whenever classification model is used to predict the unknown result then it is called prediction. Basically classification tree is a prediction technique. Classifier will be able to classify or predict both ordered and unordered values of attributes. In data mining various types of classification techniques are available. Some of the most important and accurate classification methods are

i. Support vector methods (SVM)
ii. Decision trees
iii. Bayesian classification
iv. Fuzzy classification
v. Artificial neural network (ANNs)
vi. Roughest classification and so on.

B. Clustering

Clustering is an unsupervised learning model that groups data into clusters based on the features present in the data. Sometimes clustering is used as a preprocessing technique before data classification. Data are clustered using similarity measures. Many real applications need data clustering. Ranges of algorithms are available for data clustering. The main applications of clustering are medicine, research, biology, zoology, and so on. Clustering process makes the object identification easier. Clustering eases the role of the case analyst by providing meaningful groups of data. Now the analysis among the groups of data is easier rather than the whole individual data records. In medical data analytics the task of a problem specialist is simple with the clustered data as they have the summary gist. It saves the diagnosis cost by reducing the number of tests to be carried out on the patients.

Clustering is an unsupervised learning technique where objects are grouped based on the features of objects and class labels are not given in the training data set. Some of the popular clustering schemes include:

i. Partitioning methods
ii. Hierarchical Clustering Methods
iii. Density based clustering methods
iv. Grid based clustering methods
v. Model based clustering methods
vi. Constraint based clustering methods and so on.

C. Association rule mining

Associations among the data attributes provide better understanding of the behaviour of the data under study. Such associations guide the further analysis. In the context of causal inference association analysis provides useful insight to proceed further in the way of getting causal inferences. The purpose of association rule mining is to find relationships, frequently occurring patterns, correlations, causality relationships between or among the attributes of the data sets such as relations in the relational databases. It is a rule based technique useful for finding interested relationships between the attributes of the data sets obtained from the relational databases.

D. Ensemble learning

Ensemble learning partitions large complex data into manageable groups. The desired analysis/learning model is primarily applied on the individual groups. Finally the individual results are combined to get the aggregated result. Ensemble learning is a new learning technique available in machine learning where a set of two or more separate base learners are combined into a single complex learning model in order to increase the accuracy of the proposed learning model. When earlier classification results were observed thoroughly, previous results have shown that in many cases the ensemble classification results were more accurate than individual classification results. As a result of this, ensemble learning techniques are being applied in many of the diversified applications.

E. Regression

Regression is a statistical technique used for finding the relationships between two or more variables. It finds how a dependent variable changes when there is a change in the corresponding independent variable. Usually the output of the classification result is yes or no answer but the output of the regression is a number. Regression results may be either positive or negative. Regression tree is used for finding the regression result similar to the classification tree, which finds classification results.

F. Text mining

Finding useful and quality results from text data is called text data mining. Text analytics play an important role in modern text data analytics and its need is gradually increasing in the processing fields of languages. Its main task is transforming unstructured text data into structured and more useful format so that processed text is directly available as input to the many of the data mining algorithms directly.

Text data mining has profound effect on different applications such as e-commerce, natural-language-processing, parser construction, information retrieval and so on.

G. Stream data mining

Stream data mining is a method of extracting useful knowledge from continuously moving large amounts data. Stream data consists of ordered sequences of data records.

Mining dynamically moving continuous large data is called stream data mining. Stream data mining applies advanced data mining techniques on continuously moving data instead of static data. Nowadays stream data mining is inevitable in many applications because stream data usage is compulsory in different applications. In many stream data applications the task is to find class or value of an unseen instance in the stream data. Incremental data learning techniques are generally used in stream data mining. Various statistical measures are usually applied during stream data mining for managing data drifting features.

H. Hybrid classification

A hybrid classification is a classification approach which makes use of other classification models to build a final modal. The other models contribute their intelligence in the process of building the classification model.

I. Graph data mining

Graph data mining is a way of finding patterns or sub-graphs in the given graph and sometimes it is used for classification or clustering. Graph data mining applications are abundant in daily lives of people. Graph data mining is mandatory in many researches, scientific and medical related applications.

J. Incremental data mining

Incremental data mining means applying data mining techniques on the fly many times on the incrementally adding data without starting from the scratch data. Approximately the same computation power is needed is for incremental data mining.

There are three types of data analytics models. The first one is called predictive model and the second one is called descriptive model. Predictive model generally used supervised learning techniques for prediction whereas descriptive model used unsupervised techniques for discovering knowledge from the large amount of data. The third model named prescriptive model make use of the first two models to prescribe solutions for the current problem undertaken.

Medical data analytics means applying data analytics techniques on electronic health record for getting data relationships, patterns, hidden knowledge, and many other useful result inherently present in the data medical data analytics is a specially identified, multi-domain and multi-dimensional inter disciplinary research area where people from diversified fields interact, discuss, analyze, identify and elucidate desired information .The people that participate in the interaction are doctors, software engineers, healthcare taking persons and other medical service experts.

V. CHALLENGES AND OPPORTUNITIES IN DEVELOPING CAUSAL MODELS IN MEDICAL DATA ANALYTICS

The availability of medical data is huge today in terms of electronic health records, medical diagnosis data repositories from hospitals, government and other sources. Diseases are evolving and attaching with high speeds. The personnel of medical field are taking the challenge with modern treatment techniques merged with the commitment and intelligence of the hospitals and diagnosis units. Though the technology and intelligence is growing fast the proper utilization of large amounts of medical data is an ever demanding issue. Data mining and analytics can provide a lot to fight against the medical challenges. To do so smarter techniques of data mining and data analytics are needed. Causal inference is one of the useful tools that can be applied on medical data to mine some interesting causes of medical treatment outcomes. The present causal inference techniques are lagging in to cope with the higher data volumes in meeting the scalability issues. There is a need to model better means to cope with the large volumes of medical data that can assist in treatment decisions.

Discovery of causal relationship is a type of supervised learning with a label is fixed for a target /outcome. In such cases classification methods are capable of finding the signals of causality. Causal inference needs continuous reassessment whereas decision tree analysis is more of an inference down a probability model. At a fundamental level of thinking and application causal inference and decision analysis are connected .At advanced level of application both went on separately unless the some connection criteria is defined. The two approaches allow manipulation and led to potential outcomes. In causal inference, the “causal effect is the difference among what would happen under the n number of specified treatments. In the case of decision analysis the concern was on what would happen at one of the decisions taken. The research in these two areas may or may not closer. As classification methods are not intended for causal discovery in mind there is a possibility that classification methods may find false causal signals in data and the true causal signals may be missed. As classification method ignore the effects of other variables on the class label or outcome variable at the time of relationship examining between a variable and the class label, it may leads to false discoveries of causal relationships. To get the true causalities it is needed to work up on the effects of all the variables in the subject of classification study. Therefore it is a fundamental challenge of present research to study the casual inference approaches and develop scalable and compatible tools to infer the causalities in current data.

A key role for causal inference in public health is the evaluation of the health outcomes after different interventions. These comparisons in general need randomized experiments. Unfortunately, such randomized experiments are often unethical, impractical, or simply too lengthy for timely decision making. Therefore, causal inferences from public health data are usually derived from observational studies.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

There is a need to develop models that help to identify the causes from effects using observational data.

VI. RELATED WORK

A causal effect of a treatment plays a significant role in clinical or human population studies, where a causal effect of a treatment variable is the difference between the measure of individuals in the population with the outcome variable of interest with treatment, and the proportion of the same individuals with the outcome of interest without treatment. The identification of the causal effect lies in understanding the bias in the Assignment Mechanism (AM). The Assignment Mechanism structure is used to assign treatment to individuals. The assignment mechanism plays a vital role in finding cause and effect relationships between treatment and outcome variables. Therefore the assignment mechanism should be understood as clearly as possible. When the AM is understood clearly with the details such as a randomization clinical trial the rest of the process for a clinical trial is interesting. Randomization eliminates confusing and, therefore is a best choice in clinical trials to find causality. Randomization is one of the favorite study designs of clinicians and population scientists desiring to formulate causal inference. In randomized treatments, the balance between the covariates is needed with respect to treated and untreated groups. With this type of randomization, the difference between the probabilities P(Y=y|T=t) and P(Y=y|T=c), provides causal inference about treatment T on outcome Y.

In their tutorial of modern causal inference, A Yazdani and E Boerwinkle [1] gave a broad discussion on the concept of the subpopulation causal effect as a path toward improved decision medicine. It is very interesting to study the relationships among prediction, causation and association. Knowledge about causes of an outcome improves prediction. This prediction is easier when variables associated with the response variable are estimated. The authors presentedwith a classical example with body mass index glucose, triglycerides, HDL-cholesterol and total cholesterol as variables of interest. A cause and effect relationship can be claimed from A to B, when the probability of B given A, exceeds some given threshold. Causality and probability are directly related to each other. Graphical models were applied to visualize the AM and SEM as a part of estimating the causal effect. A constraint-based restricted algorithm named Peter and Clark (PC), was used to identify the causal structure. The causal graph data structure across the whole sample set revealed that body mass index influences TRG levels both directly and via HDL levels. It was found that the total causal effect of body mass index on TRG is significant which comprises direct effect as well as indirect effect through the mediators TC and HDL. Hypertension is a major risk factor in humans that will lead to more life threatening problems. Systolic Blood Pressure Intervention Trial (SPRINT) trial is a mean to lower the systolic blood pressure, have been in discussions in recent past. Kipp W. Johnson et al. [13] proposed a method of causal inference called parametric g formula aimed to assess the benefits of blood pressure treatments at four levels/targets. The method was applied to blood pressure measurements obtained from the electronic health records of about 200,000 patients. The records obtained from electronic sources were used to get the optimal values regarding the treatment given. These records maintain the information like visit dates, patient demographics, medication prescription orders, disease and procedure billing codes, and most importantly, blood pressure measurements. G methods are applied on blood pressure data to estimate the effect of different inputs on an outcome. Extended g methods are called parametric G formulas and these formulas are applied to model the effect of different BP treatment targets on major adverse cardiovascular outcomes (MACE).Parametric g formula uses three sub-tasks for analyzing patient records. During the first stage of parametric modeling Bayesian logistic regression is used to model effect sizes and conditional probabilities for all person-times for all covariates and outcomes. The probabilities computed in the first step are used as inputs to perform Monte Carlo simulation for ten thousand individuals for each treatment target. In the next step models based on Cox proportional hazards are fitted to evaluate the relative efficacy to the results from each of the different simulated treatment policies and estimate their efficacy. Among the 10,000 patients simulated for each BP target, 40,000 experienced a total of 14,501 major adverse cardiovascular events. The number of MACE was highest in the 150 mmHg target group, followed by 140 mmHg group followed by 130 mmHg group, followed finally by the 120 mmHg target group (3367 events).

The effect of the treatments on four different SBP targets is evaluated. The acyclic directed graph model is used to get the inferences. This formula can determine the chances for an outcome with respect to covariates and exposures. It is not possible to get probabilities from continuous variables like blood pressure and heart rate. The role of the proposed formula to model probabilities instead of direct calculations is significant here. These methods are able to handle continuous covariates. It was found that the association between normal blood pressure target and the decreased incidence of adverse heart events. Causal inference methods applied in this context to may provide increased benefit. Discovering data relationships among different variable in observational data is very important. Association relationships are most popular and useful in many real time applications. For example in the medical field, the association relationships between diet habits and a particular disease may be used to infer potential causes of the disease. Market basket data analysis and the associated relationships between the items may help to the increased sales of supermarkets. Note that association relationships between the data do not necessarily mean causality. When we purchase two products together it does not mean that buying one product is the cause of buying the other product. It is not possible to find causality relations based on the manipulations forcibly in controlled manner. A causal relationship is an indication of influence of one variable on another variable rather than mere associations. Hence the application of the relationships provides vital outcomes in many contexts of real life.

Many researchers have been conducting research in the areas of graphical causal modelling to find causal relationships using Bayesian networks or probabilistic based graphical model. Bayesian networks are modeled using a Directed Acyclic Graph (DAG) representing relationships among the nodes where each node represents a variable of interest. Bayesian network learning is a very good method for causal relationships discovery but the computational costs for learning Bayesian network is very high and the proposed graph model can handle datasets with limited dimensions only. Some constraint based Bayesian networks are more efficient for discovering causal relationships on observational data using local casual structures. Association rule mining provides association relationships. Casual relationships always imply associations but the converse may not be true always. Analyzing and then finding casual relationships is very difficult when the number of variables in the data set is very high because all the variables are taken into consideration. There is possibility to find casual relationships through normal association rule mining because association rule mining is efficient method for finding association relationships in large datasets. In association rule mining there is a problem of generating spurious association rules. Integrating cohort studies with association with rule mining provides the way to automatically generate causal relationship from the very large data sets. CBN cannot find causal associations involving sets of variables as a cause. The Causal Bayesian network embedded with integrated methods allows discovering causes consisting of combined variables. Mainly observed point is that sometimes each individual variable does not provide any cause separately even though it is associated with response variable but the combination of individual variables produces a cause effect. An association is necessary for a causal relationship but associations may not indicate casual relationships. Therefore, there is need to conduct cohort study to find true causal relationships from the identified association rules.

In the case of medical and social research activities whenever randomized controlled trails are practically impossible, the alternate way is to apply a cohort study to find risk factors. Cohort studies are observational which are divided into two types –prospective and retrospective. An association rule A->B represented or indicates a causal relationship between the two variables A and B and then it is called causal association rule, if its odds ratio of the fair data set value is significantly greater than 1. If the odds ratio of a specific association rule on its fair dataset is significantly greater than 1, then a change of the response variable is resulted from the change of the exposure variables. If a variable is associated with the response variable then that variable is said to be relevant variable. Otherwise the variable is irrelevant. There is a need to control relevant variables only but not irrelevant variables.

In [21] authors carried out a study on the models leading to the techniques of finding causal associations. The study also concentrated on the applicability of various models along with the suitable data contexts. The study moved around variety of models named graphical models, potential outcome models, sufficient component cause models and structural equations models and so on. The study brought out the fundamental links among the models with respect to logic, process and applicability along with their SWOT analysis. A graph can be interpreted as a causal model provided the edge is mapped to causal relationships. These models can present the causal signals among the population. Potential outcome models provide outcomes which are probabilistic in nature because of the uncertainty of the parameters. The study concluded that causal models constructed on graphical means can be linked with other models of causation.

Jiuyong Li et al [12] proposed an algorithm for mining causal associations. The algorithm integrated the association rule mining process with conditions or tests for causality. In the process of causal linkage the property of anti-monotonic is utilized. This property states the applicability of frequentness from superset to its subsets which states that a superset of an infrequent pattern is infrequent. This statement is used to reduce the effort through pruning.

The process of detecting causal rules contains three important steps- determining the controlled variables, creating fair dataset, and selecting the required controlled variables. Irrelevant variables are those variables which are not associated with the response variable. Suppose for a given variable x, its association with the response or target variable Y can be determined by the odds ratio of x->y. The odds ratio value is automatically adaptive to the size of the dataset. The set of samples selected from the given dataset must be modified so that the resulted dataset is fair. Odds ratio is used to test whether the selected association rule is a causal association rule or not. If the odds ratio of x->y is significantly higher than 1, then the conclusion drawn is that x is a cause of Y. odds ratio must be taken from the fair dataset only. During experiments authors have used many public datasets for testing the effectiveness and efficiency of the proposed causal rule discovery method. Selected specific variables are assigned binary values either 0 or 1. Two medical datasets hypothyroid and sick are used during experimentation and both of these datasets are related to thyroid disease. For scalability purpose authors also used the adult dataset related to census Income. Another additional medical dataset called the Harvard Lung cancer dataset is used for testing the scalability of the dataset. The selected additional medical data set contains many discrete attributes. Harvard Lung cancer dataset is a microarray of data samples. The selected original dataset contains a set of eleven thousand six hundred and fifty seven genes records. Out of the all these genes top-89 genes records are obtained by a special processing technique continuous values of genes expressions are descretized as “up” and “down”.

In all the experiments involving medical datasets class attributes of the original datasets are set as response variables. The default minimum support value is 0.05 but for Adult dataset it is 0.01 and the minimum local support for the Harvard Lung cancer dataset is 0.35 because it is a small dataset. Top-89 genes are obtained by information gain ratio implementation of Weka. Causal association rules obtained are significantly smaller than other types of rules such as association rules in number.

That is cardinality of the causal association rule set is very much less than the cardinality of the association rule set.

The size of causal rule set is very small and this small set of causal rules is not sufficient for classification because only some of the data records are covered by a causal association rules. Causal association rules are small in number, powerful, popular, useful, and reliable data relationships because each of the causal rules is tested by the cohort study in data. Causal rules are tested by the cohort study in data. Causal rules can be interpreted very easily and these rules contain one or two variables only.

That is majority of the causal rules are short.

Graphical causal modeling methods are most widely used methods for causal discovery in the data. The two important properties of graphical causal modelling are causal sufficiency and causal faith fullness bias is most important in selecting the fair data set. For reducing the effect of bias authors have run the experiments multiple times the some fair data set and finally selected consistent rules in multiple causal rule sets as final rules. Authors were noticed that final causal rules are quite stable because the variance is small. In many datasets the change of causal rules between different runs is very small. More than 80% of the causal rules are stable and consistent in not more than three runs. For finding efficiency of the causal association rules with the corresponding response variable authors have experimented by taking many different fair data sets with multiple runs.

Z. Jin ,et al.[12] carried out the concept of obtaining causation through association based on causal relationships. Bayesian networks are predominantly used in this area for discovering causal relationships but only its main disadvantage is that Bayesian network learning is a NP-complete problem, as a result of this many constraint based algorithms have been designed and developed for effective discovery of causal relationships from large data sets. All these new methods are based on Bayesian learning either directly or indirectly and uses single cause variable in causal relationships exploration. Authors have proposed a new approach for finding causal relationships form the very large data sets without predefining any thresholds. They said that causal relationship is more powerful than associated relationship. Finding complete or local causal inferences using causal graphical models need very high computational cost and to overcome this problem constraint based algorithms were proposed.

Nearest neighbour matching method is commonly used for finding causal treatments and counterfactuals effects from the given observational data. After forming pairs of nearest neighbour covariates average treatment effects are estimated from the selected set of matched pairs. For high dimensionality this estimation will give more and more biased results as the dimensionality increases. To address this problem, Sheng Li et al.[23]proposed a novel estimator technique that first projects the selected data into a number of random linear sub-spaces, and then it computes median treatment effects using nearest neighbour matches. Authors empirically computed the mean square error of the proposed estimator using semi-synthetic data, and demonstrate the method on real-world digital marketing campaign data. The results show marked improvement over baseline methods.

Vikas Ramachandra [28]used separate machine learning techniques called deep learning techniques for estimating individual and average treatment effects. In order to reduce the dimensionality auto encoders are used by keeping the local neighborhood structure as it is. Most of the times it was observed that deep learning based technique was produced better results when compared with the k-nearest neighbor matching. Authors applied a modified model based on deep neural network learning for matching propensity scores. The proposed model is a generalization of a well-known classification technique named the logistic regression. This model is used to estimate propensity scores. The proposed model performed better than logistic regression at propensity score matching. Authors concluded that the proposed approach can also be applied in other domains that involve high-dimensional observational data, such as text analysis and public health.

Epidemiologists always wish to estimate quantities that are easy to transmit to the results of realistic public health contexts. Causal inference models provide the ways to answer these questions. Alexander P. Keilaet al. [2] tried the applicability of g formula for building Bayesian approach. This integrated model proved its efficiency in estimation causal effects when the sample sizes are small and in the case of sparse data. The model was tested against the data with respect to tobacco smokers and studied the effect of this smoking on body mass index. The authors concluded that the proposed approach can provide easily interpretable knowledge even the relationships are complex.

VII. RECENT TRENDS IN CAUSAL RELATIONSHIP DISCOVERY

Causal relationships are very important in many fields such as medicine, research, science, social and biomedical including statistics. Causal relationships finding methods really find actual or real causal relationships present in the data by finding the relationships between input and output variables and putting other extra variables as constant. Cohort based observational studies and case based controlled studies are two important techniques for finding causal relationships among the data elements. Randomly selected and then experimented with controlled trails are treated as gold standard during causal relationships determinations in many applications such as research, medicine, science, and social and so on Randomly selected and then experimented with controlled trails are treated as gold standard during causal relationships determinations in many applications such as research, medicine, science, and social and so on .The potential outcome model contains well established rules and tools and it is the most popular model used for finding the causal relationships between predictor and target variable .Mantel-Haenszel test is a well-known statistically based association test and it is particularly useful for elicitation of causal relationships between input and target variables .Some authors also applied statistical related association rule mining techniques for finding causal relationships.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Causal explanation trees are special data structure that are particularly useful for finding cause and effect relationships by assigning a sequence of values to the selected attributes of the given dataset .Sometimes data are portioned into sub groups using regression tree models and then causal effects are computed on each small group .It was observed that on an average five to ten sub-groups are moderately good for grouping the large dataset into sub-groups before finding causal relationships .Average causal effect is the good measure for finding intensity of the actual causal effect between the selected variables .

Stratification is a sub-grouping technique where each record is placed in a separate sub-group so that each record in the sub-group has the same score .This technique is pointed out by few researchers. Logistic regression have been in use for finding propensity scores of sub-groups of larger datasets but the time complexity, O(n3), of finding propensity scores is very high .From the last two decades several authors has been contributing to the field of causal discovery from large data source [5][6][8][9][10][11][14] [15][19][22][27].

VIII. REVIEW SUMMARY

Causal inference is playing a vital role in data analysis of big data sources particularly in medical and healthcare domains. Extraction of causes from effects is a challenging task. This needs practical data but it is not possible to do the whole set of experiments as they need large amounts of time and efforts. Observational data provide the alternative to this problem. Association rule mining, correlations and the related statistical tests are providing means for finding causal indications. But correlations may not mean causation. Some of the causal inference finding procedures includes Structural Equation Models, Bayesian Belief Networks, Association rule mining, Graphical inference models, Nearest-neighbour matching, deep neural networks, and decision trees. These models have been used by many researchers in past and these methods have limitations with respect to scalability and reliability. Causality has connection with probability and other statistical tests. The search for causation with added statistical tests and mining procedures can provide desirable outcomes. There is a large scope and need with respect to medical big data sources to mine causal relationships in data that can provide better means for decision making in medication and treatments.

IX. THE PROPOSED FRAMEWORK

Keeping the entire context in view a framework is proposed to discover causal relationships from large datasets with an additional goal of scalability.

The main components of the proposed work include:

A. Data pre-processing:

Fundamental analytical techniques such as tests like correlation, chi-square test, PCA will be conducted to reduce the size of big data considered for data analytics and mining. The result of the tests like correlation, chi-square test, PCA guide us to keep strong attention on subset of attributes or dimensions rather than considering all the given set of dimensions/attributes. Correlation found the association among the variables. Through association test we can eliminate the concentration on variables/attributes which are less associated with decision variable .We can also reduce the size of dimension by elevating similar independent variables with redundant influence on dependent variable .This can be done through association tests. Chi-square and PCA provide the information about association among the variables. Further PCA picks up the principle dimensions /attributes that causes the changes in dependent variable mostly. This dissection guides us to reduce the size of dimensionality.

B. Causal Tree generation:

A general decision tree tries to classify the given data context into labelled sub groups so that one can classify the data based on the label information model. The proposed model concentrates on preparation of a decision tree model with embedded intelligence at each phase of data division or place of node formation. The intelligence is aided by a set of statistical tools and techniques. That is while constructing the tree decisions about generation of a node, termination of a path through a leaf, labelling of a node/leaf will be guided by the outcomes of the statistical tests. At each node of the generation process the causality between the parent node and the child being generated will be found through statistical models and the outcomes will be considered for tree production. The degree of causal relationships only will allow the tree to grow further. The aim is to construct procedures that can infer causes from effects with desired scalability.

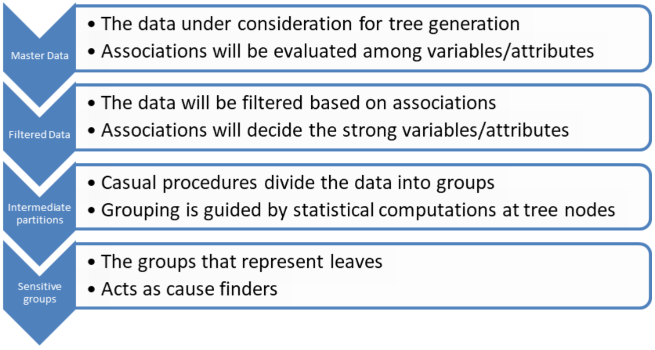
Master Data
• The data under consideration for tree generation
• Associations will be evaluated among variables/attributes
Filtered Data
• The data will be filtered based on associations
• Associations will decide the strong variables/attributes
Intermediate partitions
• Casual procedures divide the data into groups
• Grouping is guided by statistical computations at tree nodes
Sensitive groups
• The groups that represent leaves
• Acts as cause finders

Fig. 1. Partition flow of data

C. Data

Medical records are potential sources of data with inherent causation. The preparation of the causal models will make use of medical data to mine causes for medical situations. Mining such causes helps in better decision making in medical context. The pictorial representation of the proposed process is given below:

As shown in Fig. 2, the input to the proposed model is a dataset with labeled information. The model tries to mine the causes to the effects ended by the label. To do this the model follows a preprocessing process followed by tree generation. In preprocessing stage the model reduces the computational burden by sizing the dimension through association tests. Dimensions/attributes having low correlations with the dependent variable will be removed here.

The tree generation process generates a casual tree. In this phase at every node the data is divided in to sublevels through causal association tests.

This division continues until no significant causation is observed. The tree with causal nodes can be interpreted as a model that can find causes to the labelled effects at the leaf level of the tree. By navigating the causal tree from bottom level to the top one can list the significant causes to a picked effect at a leaf.

X. CONCLUSIONS

Big data analytics is very much useful to speed up decision making process in all big data contexts related to a good number of real life case studies and real life problem situations. The data from health care sector is one among the useful big data sources. Healthcare and related fields are the potential sources for generating massive datasets. There is a need for the proper management of such data efficiently for effective decision making. Doctors have to take cost effective, accurate, and fast and constraint based decisions with limited availability of details and resources. Inferring causes from effects is a vital practice for predicting and prescription in problem solving. For the last two decades the research community investigated various causal models. The existing models of causal inference have limitations with respect to volume and varieties of data. To cope with these limitations an interesting proposal is presented in this paper. The proposed approach will certainly add some knowledge to pursue further research in the context of causal inference models.

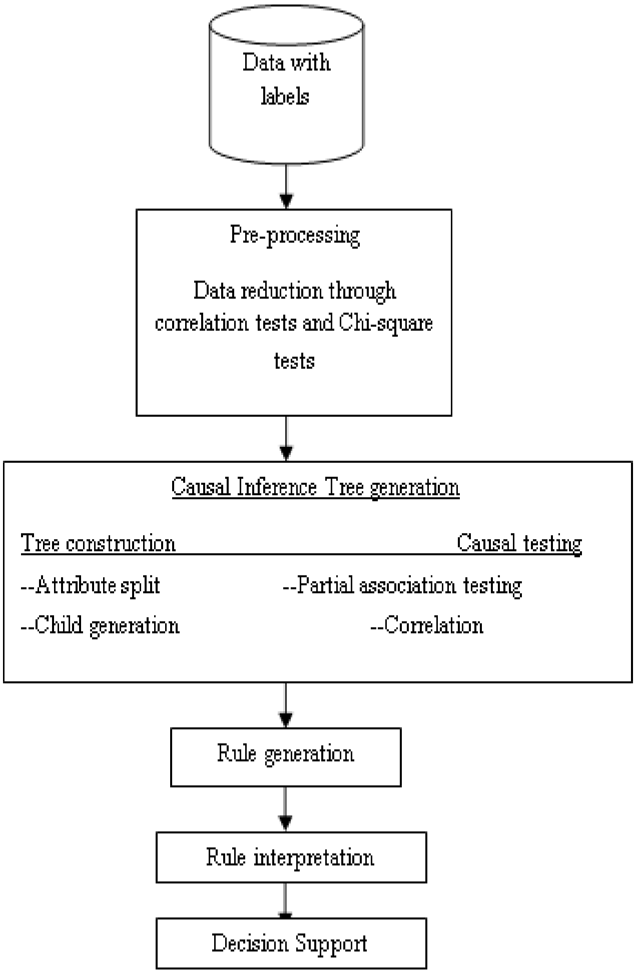

Fig. 2. Proposed Methodology

REFERENCES

1. A Yazdani and E Boerwinkle,Causal Inference in the Age of Decision Medicine, J Data Mining Genomics Proteomics. 2015 January; 6(1): doi:10.4172/2153-0602.1000163.
2. Alexander P. Keila , Eric J. Daza , Stephanie M. Engel , Jessie P. Buckley , and Jessie K. Edwards, A Bayesian approach to the g-formula, December 16, 2015.
3. Baker, E.W., Relational Model Bases: A Technical Approach to Real-time Business Intelligence and Decision Making. Communications of the Association for Information Systems, 2013.33(1): p.23.
4. Birch M.W.” The Detection of Partial Association”, Journal of Royal Statistical Society Aeries B (Methodological), Vol. 26. No. 2 (1964), pp. 313-324.
5. 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)
6. Christopher D. Ittner, “Strengthening causal inferences in positivist field studies”, Accounting, Organizations and Society 39 (2014) 545–549.
7. Constantin F. Aliferis, Alexander Statnikov, IoannisTsamardinos, 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.
8. 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]
9. 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.
10. 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.
11. Jiuyong Li, Saisai Ma, ThucDuy Le, Lin Liu and Jixue Liu “Causal Decision Trees,” arXiv: 1508.03812v1 [cs.AI] 16 Aug 2015.
12. Jiuyong Li, ThucDuy 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
13. Kipp W. Johnson, Benjamin S. Glicksberg, Rachel A. Hodos, KhaderShameer, and Joel T. Dudley, Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula, Pac SympBiocomput. 2018; 23: 180–191.
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. May, M., Life Science Technologies: Big biological impacts from big data. Science, 2014.344(6189): p. 1298-1300.39. OECD, Strengthening Health Information Infrastructure for Health Care Quality Governance.2013: OECD Publishing. 180.
17. Mayer-Schonberger, V. and K. Cukier, Big data: A revolution that will transform how we live, work, and think. 2013: Houghton Mifflin Harcourt.
18. OECD, Strengthening Health Information Infrastructure for Health Care Quality Governance. 2013: OECD Publishing.
19. 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.
20. S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, Extracting information from textual documents in the electronic health record: A review of recent research. Yearbook of Medical Informatics, pages 128–144, 2008.
21. 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.
22. 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.

23. Sheng Li, Nikos Vlassis, Jaya Kawale,and Yun Fu, Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16).
24. Spirtes P., C. C. Glymour, and R. Scheines, Causation, Predication, and Search, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.
25. Spirtes Peter. “Introduction to Causal Inference, ” Journal of Machine Learning Research 11 (2010) 1643-1662.
26. Stephen L. Morgan, David J. Harding, “Matching Estimators of Causal Effects”, Sociological Methods& Research Volume 35 Number 1August 2006 3-60 _ 2006 Sage Publications
27. Swati Hira and P. S. Deshpande ,“Mining precise cause and effect rules in large time series data of socio economic indicators,” SpringerPlus (2016) 5:1625.
28. Vikas Ramachandra, Deep Learning for Causal Inference, Stanford University Graduate School of Business,655 Knight Way, Stanford, CA 94305.
29. White, R.W., et al., Web-scale pharmaco vigilance: listening to signals from the crowd. Journal of the American Medical Informatics Association, 2013: p. amiajnl-2012-001482/
30. 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.
31. Zhou JIN, Rujing WANG, He HUANG, Yimin HU, “Mining Top-k High Persistent Causal Rules Through Non-graphic Methodology” Journal of Computational Information Systems 10: 9 (2014) 3955-3963.

AUTHORS PROFILE

Sreeraman Y received his B.Tech. degree in Computer Science and Information Technology from the Jawaharlal Nehru Technological University(JNTU) , Hyderabad, India, in 2002 and M.E. degree in Computer Science and Engineering from the Sathyabama University, Chennai, India, in 2007.Currently he is pursuing his Ph.D in Pondicherry Engineering College in the Department of Computer Science and Engineering, Pondicherry University (A Central University),Pondicherry, India. He is a Life Member of the Indian Society for Technical Education (ISTE) and Computer Society of India (CSI). He has published papers in various National and International journals. His current research interests include Causal Discovery, Machine Learning and Data Analytics.

S. Lakshmana Pandian received his Bachelor of Engineering in Electrical and Electronics Engineering, from Government College of Engineering, Tirunelveli in 1993, and Master of Engineering in 1998 in Computer Science and Engineering, from Government College of Engineering, Tirunelveli and Ph.D. degree in the Department of CSE from Anna University, Chennai in 2011. He is working as Associate Professors in the department of CSE in Pondicherry Engineering College, India. He has more than 20 publications in International Journals. He has presented more than 20 papers in International conferences. He has guided both UG and PG candidate’s projects. His areas of interest include Language Technology (Natural language processing, Compilers, Automata theory and computations), Embedded Systems and Data Analytics.