International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

A Novel Automatic Way of Finding Direct and Linear Sequential Inherent Relationships in the Data

D. Mabuni

Abstract: ***Finding linear sequential relationships (LSRs) in the data and applying them for obtaining fruitful results is an essential task in many modern day to day useful real and simulation based applications. In many previous research applications many research people usually assumed that there exists certain relationships in the data and then they have tried to bring forth some useful results after processing the selected datasets using one more data mining, machine leaning, and big data techniques. Take it for granted assumptions on the data may not be true in all the cases and in all the applications. The purpose of the present study is to bring out some automatic, smart, simple, scalable, fruitful and useful data analysis techniques after analyzing the datasets in the hand and at the same time assumptions on the data are not considered just like the general fashion of take it for granted option. The proposed model is particularly useful and applicable for finding the drug to disease relationships.***

Keywords: Automatic Finding, Data, Direct, Linear Sequential Relationships.

I. INTRODUCTION

Data processing is compulsory in many applications. Now a days intelligent way of automatically finding inherent relationships present in the data is an essential task for effective management of many real time activities including human related or object related processes. At the same time these new techniques must be smart, scalable, sufficient, accurate, and simple enough. Sometimes the data relationships may be very simple or sometimes the data relationships may be very complex. In both the cases there is a strong and compulsory need to find automatic, modern, and intelligent data analysis techniques. Many types of data analysis and management techniques are being evolved continuously from all over the ends of the world. But there is a gap between already developed techniques and techniques which are currently developing for data analysis and management. To fulfill the gap between these two plans, modern research people are trying to find direct and automatic procedures for finding inherent relationships present in the data. Many modern research people are trying to find direct and automatic data management models by applying combination of two or more data analysis techniques including data mining, machine learning, big data, and statistical data analysis techniques. This new data analysis approach is called hybrid data analysis approach. Some of the hybrid data analysis techniques are very useful, common and very sound in finding inherent relationships present in the data and the models generated in that way are popularly using by many people all over the world particularly in the fields of medical, defense, research, simulation and so on. [1] S. Athey and G. Imbens have used regression trees to find the relationships of attributes in each subgroup of a set of sub groups of a big dataset by considering pre assumptions on the dataset. C. Boutilier et. al. [2] developed a special tree called conditional probability tree for explaining the attribute relationships conditionally. The probabilities are assigned in each path of the tree and then the effect of input variables on the output variable is computed in each path. All the computed probabilities are stored in the table. But only the problem is that the table size increases rapidly with exponential time complexity and as a result this model is not applicable for all applications.

A. Applications of the Proposed Model

The proposed data management model is very useful for prescribing or selecting a correct drug for a particular disease. Once all the relationships between the drugs and diseases are known in advance then it is a simple matter of selecting or prescribing a correct drug or a sequence of drugs to the selected disease containing a specific patient. Sometimes a particular drug consumed by the patient directly affects the intensity of the disease or sometimes when the patient consumes a sequence of prescribed drugs for his/her disease then only the patient will get the favorable or unfavorable results. Especially the proposed model will give all the direct and linear sequential relationships between the drugs and diseases.

II. LITERATURE SURVEY

Numerous data analysis techniques are incessantly becoming available for use. In spite of this fabulous list of available techniques many researchers are trying to find smart and automated intelligent data analysis techniques. Exploration of direct or linear sequential relationships between input and output variables in the data is an indispensable task in many applications.

Revised Manuscript Received on April 25, 2020.
* Correspondence Author
D. Mabuni*, Department of Computer Science, Dravidian University, Kuppam, India. Email: mabuni.d@gmail.com

Many research people are continuously applying their efforts for knowing inherent relationships in the data. Many algorithms including hybrid algorithms have been developed for determining linear sequential data relationships.

Z. Jin. Et. al. [3] have pointed out that some special methods are particularly useful for finding causal relationships between the attributes when usual data mining methods are combined with some statistical techniques. Ioannis Kavakiotis et. al. [4] said that in bio technology and health care fields there is a need to find sequential relationships between the attributes.

P. Komarek [5] pointed out that the logistic regression is one way for finding the relationships between the attributes in the data but the time complexity of this method is O(nk) where $2 < k < 3$. So, its applicability is limited to particular applications. Ariel Linden, Dr.P.H. and Paul R. Yarnold [6] have used propensity scores for finding relationships between the data attributes and they have used these propensity scores for comparing maximum accuracy of classification trees. B. K. Lee, J. Lessler, and E. A. Stuart [7] proposed an efficient algorithm for finding relationships between the attributes in the data and they said that the efficiency of the proposed algorithm can be improved by estimating the propensity scores. J. Li et. al. [8] have applied association rule mining techniques for elicitation of causal relationships between the input attributes and output attribute.

Jiuyong Li et. al. [9] pointed out that finding relationships between the attributes, particularly causal relationships, is an essential task in many domains involving data analytics and they have developed decision tree model for determining relationships between the input attributes and output attribute. Also experimentally they obtained useful results. Jianmo Ni et. al. [10] Authors have developed context-aware sequential models containing personalized as well as temporal pattern details of data belonging to fitness. U. H. Nielsen et. al. [11] used Bayesian networks for creating causal explanation trees by attaching attribute values to the attributes sequentially but these trees represent global sequential relationships but not local sequential relationships. Values for the attributes are assigned path wise sequentially for analyzing path wise relationships by using known Bayesian network rules. Juan de Ona et. al. [12] have developed a new data model for finding relationships between the data attributes without considering any pre-defined and assumed relationships in the data.

E. A. Stuart [13] have noticed that data must be divided into sub groups to find the inherent relationships between the attributes but the convention is that the optimal number of sub groups must be in the range between 5 to10 only. M. K. P. B¨uehlmann and M. Maathuis [14] have noticed that some algorithms especially operate on complete Bayesian network learning technique for determining local sequential relationships between the attributes of the dataset. Larissa Westerdijk [15] have studied many machine learning algorithms energetically for finding relationships between the attributes during breast cancer medical research for predicting whether tumor is malignant or else benign.

The work presented in this paper is towards the goal of automatic determination of direct or linear sequential inherent relationships in the data. Well designed and developed standard frameworks are required for ease of handling real time tasks and applications conveniently. For obtaining quality frameworks, present trend is running rapidly towards the usage of hybrid techniques. Once the accurate data model is constructed with high quality and quantitative relationships between the input and output variables, then the data model can be applied in various domains for getting sought-after results.

Sometimes the relationships may be either local or global. In data mining, many existing works are executing their tasks with the assumption that certain relationships in the data are known in advance. Day by day usages of machine learning methods are increasing with incredible speed in different domains. Changes are inevitable in almost all the applications. So, there is an urgent need of designing and development of a collection of state-of-the-art and standard data analysis tools not only for elicitation of quality results but also for finding quantitative relationships between the data variables.

III. PROBLEM DEFINITION

Data analysis techniques are mandatory for effective and efficient management of real and original activities. During the time of application of these data analysis techniques generally researchers are assuming that there exists certain relationships in the data and then they will try to obtain desired inherent relationships from the data using various types of data analysis techniques. The main problem in this process is that the assumed assumptions may not be correct in all cases. So, there is a need to find good and accurate techniques which can be applied directly without any pre assumptions on the data. That is, desired methods must be able to apply directly on the data for obtaining useful results without any assumptions or without knowing domain knowledge. The goal of the present study is to find such automatic and direct data analysis techniques for directly extracting inherent relationships present in the data.

IV. PROPOSED TECHNIQUE-1

In this paper a new technique is proposed for direct analysis of data without considering any pre granted assumptions. This new technique analyzes the data directly and then outputs intrinsic relationships present in the data. Now a days many researchers are applying their efforts for elicitation of direct automatic data analysis techniques. Recent research results have shown that hybrid techniques usage in data analysis is increasing rapidly not only in one type of research but in multitude research disciplines. In this paper a new framework consisting of a classification tree and a special statistical metric is proposed and then used directly for constructing a new data model that is useful for both data analysis and data management conveniently. In the first step data classification tree is constructed for the given data. During the data classification tree construction special statistical metric is used for selecting the best attribute in the given dataset and then this best attribute is used for dividing the data into two sub groups. The data classification tree is constructed in top down fashion using greedy method recursively. In the second step, developed data model is used for finding solutions to some problems. During data classification tree construction process the given dataset is analyzed systematically and then processed for finding the best data attribute.

The best data attribute from the given dataset is identified based on the special statistical measure. Form the given dataset, all attributes are considered and then each attribute is compared with output attribute and then processed in order to determine the best attribute at each level of the data classification tree construction. In the next level, only remaining attributes are considered for selecting the best attribute and then sub grouping or dividing the data in hand at that time. The data classification tree is created recursively with depth first search technique. The time complexity for the data classification tree creation or data classification tree usage is O(log n) where n is the number of nodes in the data classification tree.

Without imposing any pre assumptions on the data the statistical metric is used directly to find the best attribute based on the inherent relationships between the attributes of the dataset. That is, this special statistical metric does not assume any pre assumptions on the data. This statistical technique derives a numeric measure which signifies how strongly the input attribute(s) are related with the output attribute. The numeric measure is directly proportional to the relationship strength of the output variable. The relationships from the input attributes to output attribute may be either one-to-one or linear sequence-to-one. One-to-one relationships are simple whereas linear sequence-to-one relationships are complex. Different types of relationship examples for one-to-one relationships are A → Y, B → Y, C → Y, and D → Y. Some of the examples for linear sequence-to-one relationships are: AB → Y, ABC → Y, and ABCD → Y. Here AB, ABC, and ABCD are linear sequential paths from the root node A to the leaf node of the tree.

In the present paper the hybrid data model is applied for data analysis. This hybrid data model is combination of one statistical data analysis model and the another tree data analysis model. Initially, entire data is stored in the tree root node. Statistical data model is used for measuring the inherent relationship between input and output attributes. Based on the result of the statistical measure tree nodes are divided into sub trees. Tree is created level by level. At each level, best data dividing or sub grouping attribute is selected based on the currently computed statistical measure value by considering currently available data. At each level the attribute whose statistical relation measure is the highest is selected as a dividing or splitting attribute. Same procedure is repeated at each level. Tree growing is stopped after reaching a specified height of the tree or the currently computed statistical measure falls below a certain threshold level or all the data attributes are exhausted.

To reduce the computational effort, before applying the actual main statistical measure another sub statistical measure is used and in many cases this sub statistical measure is a correlation measure between input attributes and output attribute. Statistical measure is applied for each attribute at each level of the tree creation process for finding the best data attribute at that level and time. The best data attribute is one whose statistical measure is the highest among all the currently computed statistical measures of attributes. Number of data attributes to be considered for tree growing decreases as the number of levels of the tree increases. Tree data model is selected because its time complexity is O(log n) for all tree operations irrespective of the number of branches of the tree data model.

TABLE-1 Sample Dataset

B.Tech (A)	M.Tech (B)	Ph.D. (C)	Experience (D)	GATE (E)	NET (F)	Selected for job (Y)
0	0	0	0	0	0	0
0	0	0	0	0	1	0
0	0	0	0	1	0	0
0	0	0	0	1	1	1
0	0	0	1	0	0	0
0	0	0	1	0	1	0
0	0	0	1	1	0	0
0	0	0	1	1	1	1
0	0	1	0	0	0	1
0	0	1	0	0	1	1
0	0	1	0	1	0	1
0	0	1	0	1	1	1
0	0	1	1	0	0	1
0	0	1	1	0	1	1
0	0	1	1	1	0	1
0	0	1	1	1	1	1
0	1	0	0	0	0	0
0	1	0	0	0	1	0
0	1	0	0	1	0	0
0	1	0	0	1	1	0
0	1	0	1	0	0	0
0	1	0	1	0	1	0
0	1	0	1	1	0	1
0	1	0	1	1	1	1
0	1	1	0	0	0	1
0	1	1	0	0	1	1
0	1	1	0	1	0	1
0	1	1	0	1	1	1
0	1	1	1	0	0	1
0	1	1	1	0	1	1
0	1	1	1	1	0	1
0	1	1	1	1	1	1
1	0	0	0	0	0	0
1	0	0	0	0	1	0
1	0	0	0	1	0	0
1	0	0	0	1	1	1
1	0	0	1	0	0	0
1	0	0	1	0	1	0
1	0	0	1	1	0	0
1	0	0	1	1	1	1
1	0	1	0	0	0	1
1	0	1	0	0	1	1
1	0	1	0	1	0	1
1	0	1	0	1	1	1
1	0	1	1	0	0	1
1	0	1	1	0	1	1
1	0	1	1	1	0	1
1	0	1	1	1	1	1
1	1	0	0	0	0	0
1	1	0	0	0	1	0
1	1	0	0	1	0	0
1	1	0	0	1	1	1

1	1	0	1	0	0	0
1	1	0	1	0	1	0
1	1	0	1	1	0	0
1	1	0	1	1	1	1
1	1	1	0	0	0	1
1	1	1	0	0	1	1
1	1	1	0	1	0	1
1	1	1	0	1	1	1
1	1	1	1	0	0	1
1	1	1	1	0	1	1
1	1	1	1	1	0	1
1	1	1	1	1	1	1
1	1	0	1	1	0	1
0	1	1	1	0	1	1
0	1	1	1	1	0	1
0	1	1	1	1	1	1
1	0	0	0	0	0	0
1	0	0	0	0	1	0
1	0	0	0	1	0	0
1	0	0	0	1	1	1
1	0	0	1	0	0	0
1	0	0	1	0	1	0
1	0	0	1	1	0	0

A. Dataset Description

TABLE-1 contains a sample dataset of recruitment details of employees of a particular software company, SOFTTECH. At the beginning itself company has framed two clear rules for job selection. The first rule is that Ph.D qualification is given first priority in selecting employees. The second rule is that job will be given only for those students having both GATE and NET qualifications. For easy understanding purpose only a small set of attributes are taken into consideration but in real life situations the number of attributes in the dataset may be very large. For handling such lengthy high dimensional datasets definitely scalable, accurate and intelligent data analysis techniques must be needed.

Company rules for student job selection are summarized below:

1) If a student has Ph.D qualification then he/she will be directly selected for a job. (or)
2) If a student has both GATE and NET qualification he/she will be selected for a job

There are six input attributes (B.Tech, M.Tech, Ph.D, Experience, Gate_Qualified, and Net_Quaified) and one output attribute (selected for job). Present goal is to find the direct or sequential relationships in the data. For simplicity purpose all the input attributes are renamed to A, B, C, D, E, and F respectively and output variable is renamed to Y.

TABLE-2 Renamed attributes

Serial No.	Original attribute name	Renamed attribute
1	B.Tech.	A
2	M.Tech.	B
3	Ph.D.	C
4	Experience	D
5	GATE_QUALIFIED	E
6	NET_QUALIFIED	F
7	Selected for job	Y

V. ALGOITHM

Algorithm Linear_Sequential_Relation( )

1.create a tree root node, T
2.read data into D
3.find the set of attributes of the dataset
4.find the size of the data, n
5. initialize tree height h = 0
6.call Linear_Sequential_Tree_Creation(T,D,A,n,h)

Algorithm Linear_Sequential_Tree_Creation (T, D,A,n,h)

Input:

T is the root node of the tree model
D is the initial dataset
A is the set of attributes of the dataset
n is the size of the dataset
h is the height of the tree model

Output:

Tree data model - representing the relationships between the input and output attributes in the dataset.

1. if T is null or n = 0 or h = maximum height then
2. Create leaf node for the data D
3. return
4. end-if
5. Find the sub statistical measure to minimize the number of potential attributes to be used in finding statistical measure
6. select potential attributes
7. find statistical score for all the potential attributes in the data in D
8.if all the statistical scores less than threshold then
9. Create leaf node for the data D
10. return
11.end-if
12.divide data in D into two sub groups D1 and D2
13. h = h + 1
14.call Linear_Sequential_Tree_Creation(T,D1,n1,h)
15.call Linear_Sequential_Tree_Creation(T,D2,n2,h)
16.print tree before pruning
17.prune the tree
18.print tree after pruning

A. Algorithm Explanation

Lines-1, 2, 3, and 4 are for stopping the tree growing process. There are three conditions for stopping the tree growing process. When the node pointer T is null or when all the tuples are exhausted during tree construction process or when the specified maximum height of the tree is reached. In each of these cases a node is created and it is converted into a leaf node. The tree growing process is stopped in that particular branch only.

Lines-5 and 6 are for finding the potential attributes of the tree by finding the correlations between input and output attributes. After calculating correlations only a sub set of attributes are selected for further processing. This sub set of attributes is called potential set of attributes.

Line-7 computes statistical measures for all the potential attributes.

Lines-8, 9, 10, and 11 are for stopping the tree creation process when all the statistical measures computed are less than a pre specified threshold value. In that case current node is converted into leaf node and only in that specific branch the tree growing process is stopped.

Line-12 when there is no possibility for stopping the tree growing process, the current node data must be divided into two sub branches, namely left branch and right branch and data must be distributed accordingly.

Line-13 increases the height of the tree, h, by one level.

Line-14 allows the left branch to grow recursively

Line-15 allows the right branch to grow recursively

The output data model must be simple, elegant and minimum height. Tree pruning is generally used to reduce the tree height. There are two pruning types, pre pruning and post pruning. In post pruning tree growing is allowed to its maximum height and then after the completion of the tree growing process post pruning is applied in the bottom up process starting from the leaves. When two siblings (children) of the parent node have the same class label, those two siblings are removed from the tree and the corresponding parent is converted into leaf node. Within each leaf node class labels are computed for each class separately and the majority class name is computed and leaf class is fixed with this majority class label. Same process is repeated for each branch of the tree.

A small example dataset is given to demonstrate the computational details of statistical measure. Only for simplicity purpose three input attributes and one output attribute is taken into consideration. Fifth column represents corresponding frequency occurrence counts of each data tuple.

TABLE-3 Sample dataset for statistical measure computation

Record Id	A	B	C	Y	Total
1	0	0	0	0	4
2	0	0	1	0	8
3	0	1	0	0	4
4	0	1	1	1	16
5	1	0	0	0	2
6	1	0	1	1	10
7	1	1	0	1	5
8	1	1	1	1	20

TABLE-4

{B,C}={0,0}	Y = 1	Y = 0	Row sum
A = 1	0	2	2
A = 0	0	4	4
Column sum	0	6	6

TABLE-5

{B,C}={0,1}	Y = 1	Y = 0	Row sum
A = 1	10	0	10
A = 0	0	8	8
Column sum	10	8	18

TABLE-6

{B,C}={1,0}	Y = 1	Y = 0	Row sum
A = 1	5	0	5
A = 0	0	4	4
Column sum	5	4	9

TABLE-7

{B,C}={1,1}	Y = 1	Y = 0	Row sum
A = 1	20	0	20
A = 0	16	0	16
Column sum	36	0	36

To determine the relationship between attribute A and Y, TABLE-3 is divided into four separate small tables and then A to Y aggregate relationship is computed separately. Similarly the aggregate relationships between B to Y and C to Y are computed separately. Once aggregate influence relationships are determined for all input attributes the best input attribute whose aggregate score is maximum is determined. At each level of the tree growing process the best input attribute is used to split the current node into sub trees. Same process is repeated at each level of the tree growing process. The aggregated score is a statistical measure derived from all of its constituent tables separately for each input attribute. This aggregate score computation will become complex when the number of input attributes increases. For easy understanding purpose the aggregate score computational details of input attribute, A, using TABLE-4, TABLE-5, TABLE-6, and TABLE-7 are given

$$A11 = \frac{product\ of\ diagonal1 - product\ of\ diagonal2}{row1sum + row2sum} \quad (1)$$

$$A12 = \frac{column1sum * column2sum * row1sum * row2sum}{square\ of\ (row1sum + row2sum) * (row1sum + row2sum - 1)} \quad (2)$$

$$LSR = \frac{Square\,(absolute(A11 + A21 + A31 + A41) - 0.5)}{(A12 + A22 + A32 + A42)} \quad (3)$$

A11 means first measure of TABLE-1 and A12 means second measure of TABLE-1

A21 means first measure of TABLE-2 and A22 means second measure of TABLE-2

A31 means first measure of TABLE-3 and A32 means second measure of TABLE-3

A41 means first measure of TABLE-4 and A42 means second measure of TABLE-4

Equation-3, LSR score, represents aggregate relationship measure between selected input attribute and output attribute. Linear Sequential relationships (LSRs) scores computations are shown below:

For the TABLE-4

$$A11 = \frac{0*4-0*2}{6} = \frac{0}{6} = 0 \text{ and}$$

$$A12 = \frac{0 * 6 * 2 * 4}{square(6) * (6-1)} = \frac{0}{6} = 0$$

For the TABLE-5

$$A21 = \frac{10*8-0*0}{18} = \frac{80}{18} = 4.4444 \text{ and}$$

$$A22 = \frac{10 * 8 * 10 * 8}{square(18) * (18-1)} = \frac{6400}{5508} = 1.161946$$

For the TABLE-6

$$A31 = \frac{5*4-0*0}{9} = \frac{20}{9} = 2.2222 \text{ and}$$

$$A32 = \frac{5 * 4 * 5 * 4}{square(9) * (9-1)} = \frac{400}{648} = 0.61284$$

For the TABLE-7

$$A41 = \frac{20*0-16*0}{18} = \frac{0}{18} = 0 \text{ and}$$

$$A42 = \frac{36*0*20*16}{square(36)*(36-1)} = \frac{0}{36*36*35} = 0$$

$$LSR = \frac{Square(absolute(0+4.4444+2.2222+0)-0.5)}{(0+1.161946+0.61284+0)} = 21.42622$$

If (LSR > 3.84) then the input attribute A is considered as a potential attribute. In each level of data tree construction process the same process is repeated for all the remaining input attributes for finding LSR score values. Input attributes whose LSR score > 3.84 are included in the list of potential attributes. Among all the potential attributes the input attribute with the highest LSR score is selected for the current node splitting during data tree construction. The data tree created with LSR score must be more interpretable and small in height. The threshold value 3.84 is a statistical threshold value that is used whether to continue or stop tree growing process in that particular branch. If the LSR score value is less than or equal to 3.84 then a leaf node is created in that path. In the present method only post pruning procedure is used for pruning the constructed tree for better interpretability than the better accuracy of the tree. The proposed model needs better interpretability than the high classification accuracy.

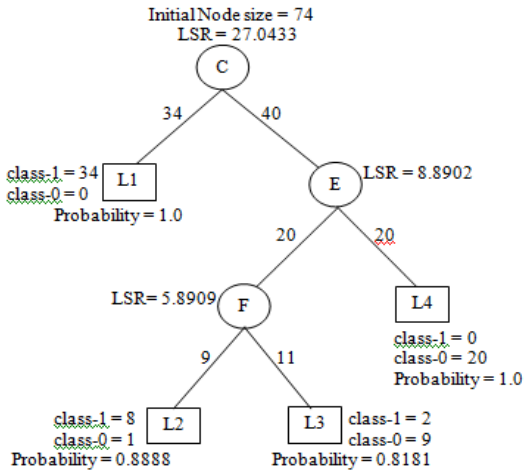

FIGURE-1 A data model of Linear sequential relationship strengths (LSRs)

VI. EXPERIMENTAL RESULTS

Experiments are conducted by taking one hypothetical recruiting dataset of a software company. The results show that which attributes influence more in getting a job in that company. The output variable is selected for job. All these influences are numerically computed and the inherent relationships are thoroughly analyzed. Some relationships are directly related to output attribute whereas other relationships are linear sequentially related to the output attribute.

L1 Influence score from Ph.D. to selected for job (C → Y) and path is (C → Y)

$$27.0433 * \frac{34}{74} * \frac{34}{34} = 12.4253$$

L2 Influence score from Ph.D. to selecting for job (CEF → Y) and path is (CEF → Y)

$$27.0433 * \frac{40}{74} * \frac{1}{2} * 8.8902 * \frac{20}{40} * \frac{1}{4} * 5.8909 * \frac{9}{20} * \frac{1}{8} * \frac{8}{9} = 2.392585$$

L3 Influence score from Ph.D. to selecting for job (CEF → Y) and path is (CEF → Y)

$$27.0433 * \frac{40}{74} * \frac{1}{2} * 8.8902 * \frac{20}{40} * \frac{1}{4} * 5.8909 * \frac{9}{20} * \frac{1}{8} * \frac{2}{11} = 0.489352$$

L4 Influence score from Ph.D. to selecting for job (CE → Y) and path is (CE → Y)

$$27.0433 * \frac{40}{74} * \frac{1}{2} * 8.8902 * \frac{20}{40} * \frac{1}{4} * \frac{0}{20} = 0$$

L1 Influence score from Ph.D. to not selecting for job (C → Y) and path is (C → Y)

$$27.0433 * \frac{34}{74} * \frac{0}{0} = 0$$

L2 Influence score from Ph.D. to not selecting for job (CEF → Y) and path is (CEF → Y)

$$27.0433 * \frac{40}{74} * \frac{1}{2} * 8.8902 * \frac{20}{40} * \frac{1}{4} * 5.8909 * \frac{9}{20} * \frac{1}{8} * \frac{1}{9} = 0.299048$$

L3 Influence score from Ph.D. to not selecting for job (CEF → Y) and path is (CEF → Y)

$$27.0433 * \frac{40}{74} * \frac{1}{2} * 8.8902 * \frac{20}{40} * \frac{1}{4} * 5.8909 * \frac{9}{20} * \frac{1}{8} * \frac{9}{11} = 2.202082$$

L4 Influence score from Ph.D. to not selecting for job (CE → Y) and path is (CE → Y)

$$27.0433 * \frac{40}{74} * \frac{1}{2} * 8.8902 * \frac{20}{40} * \frac{1}{4} * \frac{20}{20} = 8.122309$$

TABLE-8 Influence scores from Ph.D to selected for job

Leaf	Selected for job	Not selected for job
L1	12.4253	0
L2	2.392385	0.299048
L3	0.489352	2.202082
L4	0	8.122309

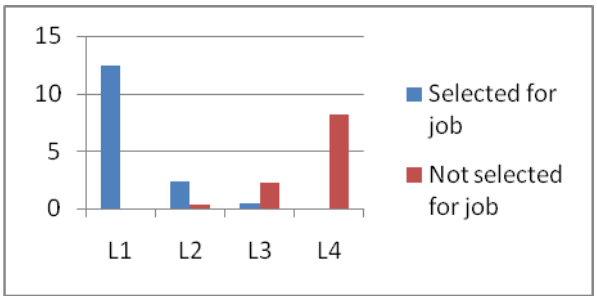

FIGURE-2 Influence scores from Ph.D to selected for job

From the graph it is clear that there is a very high probability of getting a job for Ph.D. holders. Blue color bar graph in the leaf, L1, indicates this property. Bar graph at leaf, L4, indicates probability for not getting a job who is not a Ph.D holder. Blue color graph at the leaf, L2, indicates selection of students who have GATE and NET qualifications for jobs. That is without Ph.D qualification students should have both GATE and NET qualification for getting a job in the company.

L2 Influence score from GATE to selected for job (EF → Y) and path is (EF → Y)

$$8.8902 * \frac{20}{40} * \frac{1}{2} * 5.8909 * \frac{9}{20} * \frac{8}{9} * \frac{1}{4} = 1.30928198$$

L3 Influence score from GATE to selecting for job (EF → Y) and path is (EF → Y)

$$8.8902 * \frac{20}{40} * \frac{1}{2} * 5.8909 * \frac{11}{20} * \frac{2}{11} * \frac{1}{4} = 0.327320495$$

L4 Influence score from GATE to selecting for job (E → Y) and path is (E → Y)

$$8.8902 * \frac{20}{40} * \frac{1}{2} * \frac{0}{20} = 0$$

L2 Influence score from GATE to not selecting for job (EF → Y) and path is (EF → Y)

$$8.8902 * \frac{20}{40} * \frac{1}{2} * 5.8909 * \frac{9}{20} * \frac{1}{9} * \frac{1}{4} = 0.1636600247$$

L3 Influence score from GATE to not selecting for job (EF → Y) and path is (EF → Y)

$$8.8902 * \frac{20}{40} * \frac{1}{2} * 5.8909 * \frac{11}{20} * \frac{2}{11} * \frac{1}{4} = 1.472942227$$

L4 Influence score from GATE to not selecting for job (E → Y) and path is (E → Y)

$$8.8902 * \frac{20}{40} * \frac{1}{2} * \frac{0}{20} = 0$$

TABLE-9 Influence scores from GATE (E) to selected for job (Y)

Leaf	Selecting for job	not selecting for job
L2	1.30928198	0.163660247
L3	0.327320495	1.472942227
L4	0	2.22255

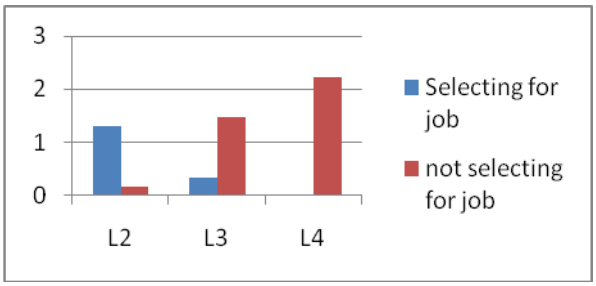

FIGURE-3 Influence scores from GA TE (E) to selected for job (Y)

The graph reveals that without Ph.D. qualification getting a job is very difficult. Only the other possibility for getting a job is either the student must have Ph.D qualification or must processes both GATE and NET qualifications. Otherwise there is no possibility for getting a job.

Other measures used for job selection are given below.

VII. PROPOSED TECHNIQUE-2

Leaf-1 score for class-1 = 74 * 27.0433 * 34 = 68040.94

Leaf-1 score for class-0 = 0

Leaf-2 score for class-1 = (74 * 27.0433 + 40 * 8.8902 + 20 * 5.8909) * 8 = 19797.04

Leaf-2 score for class-0 = (74 * 27.0433 + 40 * 8.8902 + 20 * 5.8909) * 1 = 2474.63

Leaf-3 score for class-1 = (74 * 27.0433 + 40 * 8.8902 + 20 * 5.8909) * 2 = 4949.26

Leaf-3 score for class-0 = (74 * 27.0433 + 40 * 8.8902 + 20 * 5.8909) * 9 = 22271.7

Leaf-4 score for class-1 = (74 * 27.0433 + 40 * 8.8902) * 0 = 0

Leaf-4 score for class-0 = (74 * 27.0433 + 40 * 8.8902) * 20 = 47136.4

Average class-1 score for the entire tree = 68040.94 + 19797.04 + 4949.26 + 0 = 92787.24

Average class-0 score for the entire tree = 0 + 2474.63 + 22271.7 + 47136.4 = 71882.7

Total class score = 92787.24 + 71882.7 = 164669.94

Normalized class-1 score = 92787.24/164669.94 = 0.56347406

Normalized class-0 score = 71882.7/164669.94 = 0.436525938

C input direct variable influence score with output variable for class-1 = 68040.9428/92787.24 = 0.733300606

CEF linear sequential relationship with output variable for class-1 = 0.213359516 + 0.053339879 = 0.266699395

CE linear sequential relationship with output variable for class-1 = 0

C input variable strength with output variable for class-0 = 0

CEF linear sequential relationship with output variable for class-0 = 0.034426023 + 0.309834209 = 0.344260232

CE linear sequential relationship with output variable for class-0 = 0.65574

All the input and output attribute relationships are determined branch wise for the tree shown in FIGURE-1. After observing branch wise influence scores some relationships are identified as direct and some other relationships are identified as linear sequential relationships. Normally the influence scores of direct relationships are higher than sequential relationship scores. From the above calculations, the probability for getting a job is more than the probability of not getting a job. The attribute Ph.D. directly influences the **selected for job** attribute because its score is very high.

Leaf nodes one (L1) and two (L2) must be strengthen for increasing the job getting relationship.

A data model example shown in FIGURE-1 is considered for easy understanding. Initial root node size is 74 tuples and linear sequence relation strength (LSR) with output variable, **selected for job,** is shown on the top of each node. This model consists of three non-leaf nodes (C, E, F) and four leaf nodes (L1, L2, L3, and L4). Different linear sequence paths are C, CEF, and CE. From the above generated data model two things are clear that the probability of getting a job for the Ph.D. student is 1.0 and if the student is not a Ph.D. student then he/she must qualify both GATE and NET exams for getting a job. The generated model directly reflects the rules imposed in the job selection process. That means, there is a correlation between the dataset considered and the model obtained after processing the input dataset by applying hybrid data mining techniques.

VIII. CONCLUSIONS

It is quiet common that in many cases some assumptions are assumed on the data before data processing. These assumed assumptions may not be valid or correct in all the data processing situations. Previous research trends were normally following this principle. In this paper the goal of the present study is to find directly applicable data processing techniques without considering any pre data relationship assumptions on the data. Now the availability of such directly applicable techniques is awfully limited and very unusual in usage. In the future, in many domains, there is a possibility to find and then apply data processing techniques directly without considering any assumptions on the data. The proposed model can be applied for finding drug to disease relationships in the future.

If the doctor knows the effect of a particular drug on a particular disease in advance then he will be able to provide correct and exact treatment to the patient without any side effects or other problems.

REFERENCES

1. S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," in Proc. Natl Academy Sci., vol. 113, no. 27, pp. 7353–7360, 2016.
2. C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in Bayesian networks," in Proc. 12th Conf. Uncertainty Artif. Intell., 1996, pp. 115–123.
3. Z. Jin, J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang. Discovery of causal rules using partial association. In Data Mining (ICDM), 2012 IEEE 12th Int. Conf. on, pages 309–318.
4. **IOANNIS KAVAKIOTIS, OLGATSAVE, ATHANASIOS SALIFOGLOU, NICOS MAGLAVERAS, IOANNISVLAHAVAS, AND IOANNA CHOUVARDA, "MACHINE LEARNING AND DATA MINING METHODS IN DIABETES RESEARCH", COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL, VOLUME 15, 2017, PAGES 104-116**
5. P. Komarek, "Logistic regression for data mining and high-dimensional classification," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2004.
6. Ariel Linden, Dr.P.H. and Paul R. Yarnold, "Some Machine Learning Algorithms Find Relationships Between Variables When None Exist -- CTA Doesn't", Optimal Data Analysis Copyright 2019 by Optimal Data Analysis, LLC, Vol. 8 (March 21, 2019), 64-67 2155-0182/10/$3.00
7. 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.
8. J. Li, T. Le, L. Liu, J. Liu, Z. Jin, and B. Sun. Mining causal association rules. In Data Mining Workshops (ICDMW), 2013 IEEE 13th Int. Conf. on, pages 114–123, 2013.
9. Jiuyong Li, Saisai Ma, Thuc Duy Le, Lin Liu and Jixue Liu, "Causal Decision Trees", School of Information Technology and Mathematical Sciences, University of South Australia, Australia Mawson Lakes, SA 5095, Australia, arXiv:1508.03812v1 [cs.AI] 16 Aug 2015
10. Jianmo Ni, Larrymuhlstein, and Julian McAuley, "Modeling Heart Rate and Activity Data for Personalized Fitness Recommendation", WWW '19, May 13–17, 2019, San Francisco, CA, USA © 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License. ACM ISBN 978-1-4503-6674- 8/19/05. https://doi.org/10.1145/3308558.3313643
11. U. H. Nielsen, J. philippe Pellet, and A. Elisseeff, "Explanation trees for causal Bayesian networks," in Proc. Uncertainty Artif. Intell., 2008, pp. 427–434.
12. Juan de Ona, Rocio de, and Francisco J.Calvo, "A classification tree approach to identify key factors of transit service of quality", TRYSE Research group. Department of Civil Engineering, University of Granada, SPAIN
13. E. A. Stuart, "Matching methods for causal inference: A review and a look forward," Statistical Sci., vol. 25, no. 1, pp. 1–21, 2010.
14. M. K. P. B¨uehlmann and M. Maathuis. Variable selection for highdimensional linear models: partially faithful distributions and the PCsimple algorithm. Biometrika, 97:261–278, 2010.
15. Larissa Westerdijk, "Predicting malignant tumor cells in breasts", RESEARCH PAPER Master Business Analytics, VRIJE UNIVERSITEIT AMSTERDAM

AUTHORS PROFILE

D. Mabuni, completed M.Sc. (Computer Science), MCA and M.Phil. (Computer Science). Currently working as Assistant Professor in the Department of Computer Science at Dravidian University, Kuppam, Andhra Pradesh, India. My interested research areas are Data Mining, Databases, and User Interfaces.