DOI: https://doi.org/10.33103/uot.ijccce.19.2.1

Multistage Tree Model for Crime Dataset in Iraq

[1]Reem Razzaq Abdul Hussein, [2]Dr.Muayad Sadik Croock, [3]Dr Salih Mahdi Al-Qaraawi
[1]*Informatics Institute of Postgraduate studies, Baghdad, Iraq*
[2,3]*Computer Engineering Department, University of Technology, Baghdad, Iraq*
reem@uoitc.edu.iq, 120102@uotechnology.edu.iq, drsalihalqaraawi@gmail.com

Abstract— *This research deals with the using of correlation measurement that leads to describing the degree of relationship between variables, quantities or qualities. Therefore, we implement a simple correlation coefficient and conditional correlation to introduce a regular vine copula, which gives different tree structures. Two methods to select tree structures are introduced. The first one adopts the Partial Correlation Constant (PCC) with constant, while the second method depends on the estimation of summation pathway. The proposed method makes modification on Dißmann's algorithm to increase the dependency on each level of the tree using rank correlation measurement. Both methods are adopted to construct the best model with more than three dimensions based on the available label crime dataset in Iraq. The selected model is used for selecting the suitable tree model and generating a decision with the low dimensionality of variables.*

Index Terms— *rank correlation, regular vine, decision tree conditional probability, Dißmann's algorithm.*

I. INTRODUCTION

The graphical structure commonly used for constructing a compact model becomes more popular to establish dependent distribution, such as a decision tree that deals with multidimensions [1]. It also introduces the model of present ordinal attribute and the uncertainty analysis that is used for combining two groups of variables. In this paper, the adopted method is used to construct a tree structure with minimum information obtained from the Correlation Coefficient and Partial Correlation Coefficient PCC, which depend basically on bay'es theorem and correlations [1][2].

The objective of this research is generating observable tree structure model of vine copula depending on the fitting vine. The correlation coefficients in this work are used to determine the degree of strength between two variables such as (crime type and Gender) in a single value. These values are used in selection features and to determine the shape of the tree. The Correlation Coefficient is usually given the symbol r and it lies in the interval [-1, +1], as seen in Figure (1)[3], A correlation coefficient (R) may be (Zero, positive, negative), when A correlation coefficient is close to +1 this denotes that the relationship is positive between A and B, with increases of A of the variables being associated with increases in the other variable B, while A correlation coefficient close to -1 denotes that there is a negative relationship between A and B, with an increase in A which is associated with a decrease in that of B variable. When the correlation coefficient (R) is equal to zero, this indicates that there is no relationship. The correlation can give the strength and the direct correlation of variables, the Pearson correlation is used when the variables are normally distributed, while the Spearman's correlation is more robust than the Pearson correlation coefficient [4].

The data used for this type of correlation are ranked data, which may be given for attribute data, or any measured data, where the calculation depends on a given rank to order observation using the formula:

$$r = 1 - 6\sum D^2 / N(N^2-1) \quad (1)$$

Received 13 Sep 2018; Accepted 4 Dec 2018
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

The weak correlations are eliminated to reduce the unrelated information. The rank correlation coefficient, which represents the measurement of the dependence of the copula between two variables (CrimeType and Hot_Zone), is the joint distribution of (F(crime type), F(Hot_Zone)) which may be the type of uniform or normal copula with correlation (p) [1].

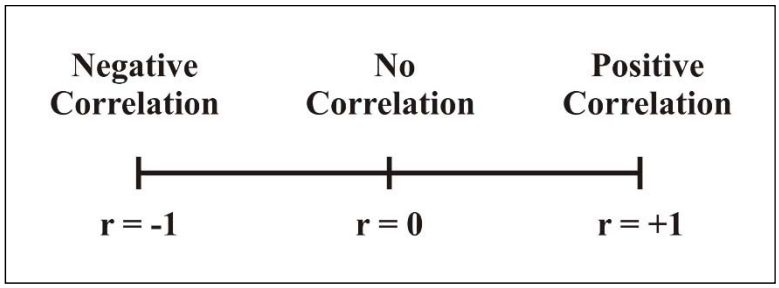

FIG. 1: SCALE OF RANK CORRELATION[4]

In this paper, the sections are organized as follows; In Section II we review the related works on correlation coefficient and vine copula parameters and formulate the problems. Section III explains the vine copula. Section IV explains the strategy of tree selection algorithms, the two methods presented, with the type of correlation need in each algorithm. Section VI presents the result and discussion about the proposed algorithm. Section VII gives the conclusion.

II. RELATED WORKS

In terms of vine copula, past researchers have depended on the correlation between variables, the Bayesian is combined copula parameters, to produce a flexible model that deals with high dimensionality of the real dataset [4], the process of eliciting a conditional correlation of ordinal parameter uses two methods, the first method is fast and deals with binary approach, however the problem of this method is that the feasibility of correlation matrices is limited, while the second deals with the mean mapping method, which overcomes the problem, however its disadvantage is that it takes longer time [5] .

The researcher presented a survey on Pair Copula Construction, which had been used in multivariate, data that found intricate patterns of dependence model using bivariate copula in financial applications, the model's tests of goodness [6]. A simple vine couples structure was used, to produce a simple tree structure strategy which is adapted to select tree model based on Diβmann's algorithm which had been developed by depending on positive Kendall measurement [7].

III. VINE COPULA

Some of the concepts used in this paper are illustrated to construct the idea of vine copula [1][[2][7]. In addition, a high dimension N for scale information is introduced. Tree of T={V, E} is a noncyclic graph, where V is denoted as vertexes and can be V={crime, gender, Age,…..,v_N} for o}N dimension. Moreover, E is denoted as a set of edges of Tree and can be E={e_1,e_2,….., the e_N} that links between vertex. They are correlated with multiple inputs of variables probability distribution, such as {CimeTypes, Gender, Time,…} [1][7]. On the other hand, D is correlated with R to produce the binary _variate tree B_i while the conditional probability D is within the range of i=[1,n], where{i,j ∈E} are the copula C_{ij} of D that can be an item of B_i(i,j). Figure (2) shows an example of conditional probability that deals with three nodes, the parent type as a murder, and thieves are the child nodes. The Bayes term is defined as:

$$P(node1|node2)=p(node2\cap node1)/p(node1) \quad (2)$$

Received 13 Sep 2018; Accepted 4 Dec 2018
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

Where, is the probability that both events are node1 and node2. The parent node is node 1, while nodes 2, and 3 represent the child nodes. In our work, we depend on Regler vine[8], with a Spearman correlation coefficient, partial correlation and higher order, as shown in the next section.

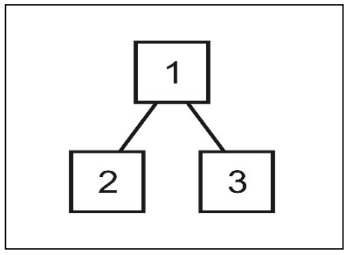

FIG. 2: PARTIAL CONDITIONAL [8]

IV. THE STRATEGY OF TREE SELECTION ALGORITHMS

The sequence structure, which is used for selecting the suitable model, is employed to construct a vine in many strategies. The data of these strategies need to be ranked by computing a correlation matrix. The random ordinal variable can be represented as a matrix (n*n), i=1,…d, which is the number of rows, and j= 1..d which is the number of columns.

$$\boldsymbol{C} = \begin{pmatrix} \mathrm{Cor}(A_1, A_1) & \mathrm{Cor}(A_1, A_2) & \dots & \mathrm{Cor}(A_1, A_d) \\ \mathrm{Cor}(A_2, A_1) & \mathrm{Cor}(A_2, A_2) & \dots & \mathrm{Cor}(A_2, A_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cor}(A_d, A_1) & \mathrm{Cor}(A_d, A_2) & \dots & \mathrm{Cor}(A_d, A_d) \end{pmatrix}.$$

Two methods need to calculate C matrix at the beginning, the method can be summarized [6] as the following:

- Method 1: This method computes the highest summation of ranked data for underlying tree structure models data. There are three dimensions of crime type, Gender, and Hot Zone represented as x1, x2, and x3, respectively. So, the possible tree structure is T1, T2, and T3, as shown in Figure (3)[7][6].

$$T1=correltion(CrimeType,HotZone) + correlation\ (HotZone, Gender) \quad (2)[7]$$

$$T2=correltion(CrimeType,Gender) + correlation\ (HotZone,CrimeType) \quad (3)[7]$$

$$T3=correltion(CrimeType,Gender) + correlation\ (HotZone, Gender) \quad (4)[7]$$

  Another option to optimize a value that helps to choose the structure of the tree by weighting path is done by applying the formula [6]:

$$D= alpha*.e_1+(alpha-1)\ *e_2 \quad (5)[6]$$

  Where alpha is the weight and e defines the length of the path.

- Second Method: The partial correlation can be computed using the following formula:

$$rC=(crime\ type,Age)=(crime\ type\sim,\ Age\sim) \quad (6)[1]$$

  where r(crime type~, Age~) is denoted as the distribution of (crime type, Age) based on the random variable C. The Spearman rank correlation measurement of *copula* association with two variables is computed using the following formula:

$$r12.3 = \frac{r12-r13\ r23}{\sqrt{(1-r13^2)}\sqrt{(1-r23^2)}} \quad (7)$$

  where r12, r13, and r23 define the partial correlation.

Received 13 Sep 2018; Accepted 4 Dec 2018
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

Partial rank correlation is the correlation that can be used to test the constant conditional correlation of two variables. Therefore, the test of partial rank correlation can be computed based on the standard rank. In this case, the correlations are performed between the three variables as shown in equation (7). For more variables, the number of constants can be increased to two or more. The r_{AB} defines the correlation between A and B. With the ordinal correlation coefficient, the partial rank correlation is constructed with three parameters of x1, x2, and x3 and fixing other variables such as x4, x5,…xN, where N is the number of parameters, as shown in figure 4 [1][4][9].

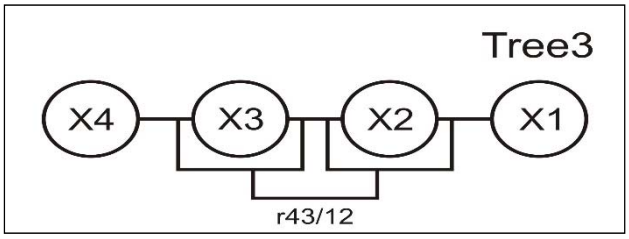

FIG. 3: FOUR-DIMENSIONAL R_VINE INCLUDING CORRESPONDING PAIR-COPULAS [9].

The equation of Four-dimensional R_vine including the corresponding pair-copulas is shown in the following equation:

$$r12.34 = \frac{.(r12\_3 - r14\_3 * r24\_3)}{(1 - r14\_3)2 * (1 - r24\_3)2} \quad (8)[1][9]$$

Where r_{12_3} r_{14_3}, r_{24_3} have identified a partial correlation with a constant example. Therefore, this way depends on maximum patrial or higher correlation for selection of the tree structure. For example, r_{14_3} represents the relationship between Gender and Crime types with constant Hot Zone. The selection of the tree can be evaluated using the Akaike Information Criterion (AIC) estimator of quality for modes. A set of trees structures can be given, AIC estimates the quality of each tree structure to be compared with each of the other models. AIC is calculated for each tree, the lowest value of AIC had been chosen. AIC, is defined as:

$$AIC = -2Log\ L + 2V \quad (8)$$

Where L is the maximum likelihood of selected models, and V is a free parameter that increases the probability, that the selected model produces observed data.
AIC is employed for model selection as well as dealing with the exchange between the quality of the suitability for such tree structure and the simplicity of the tree model. Bayesian Information Criterion (BIC) is another estimation method with the lowest cost is defined as:

$$BIC = 2Log\ L + VLog\ n \quad (9)$$

Where n is the number of observation data. Moreover, the BIC is preferred in most applications [6].

V. THE PROPOSED METHOD

In our work, we proposed to use the previous two methods, which were mentioned in the last section, as follows:

- At first, input the vertices and compute the Spearman correlation coefficient of crime dataset.
- Apply to Select the highest correlation of features (vertices).
- Compute PCC of spearman and maximum of PCC of the tree.
- Compute the average estimation of total paths, then choose the maximum path.
- The use of evaluation methods, namely AIC and BIC, that help to choose the best model of the tree structure, as shown in the flowchart of Figure (4).

Received 13 Sep 2018; Accepted 4 Dec 2018
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

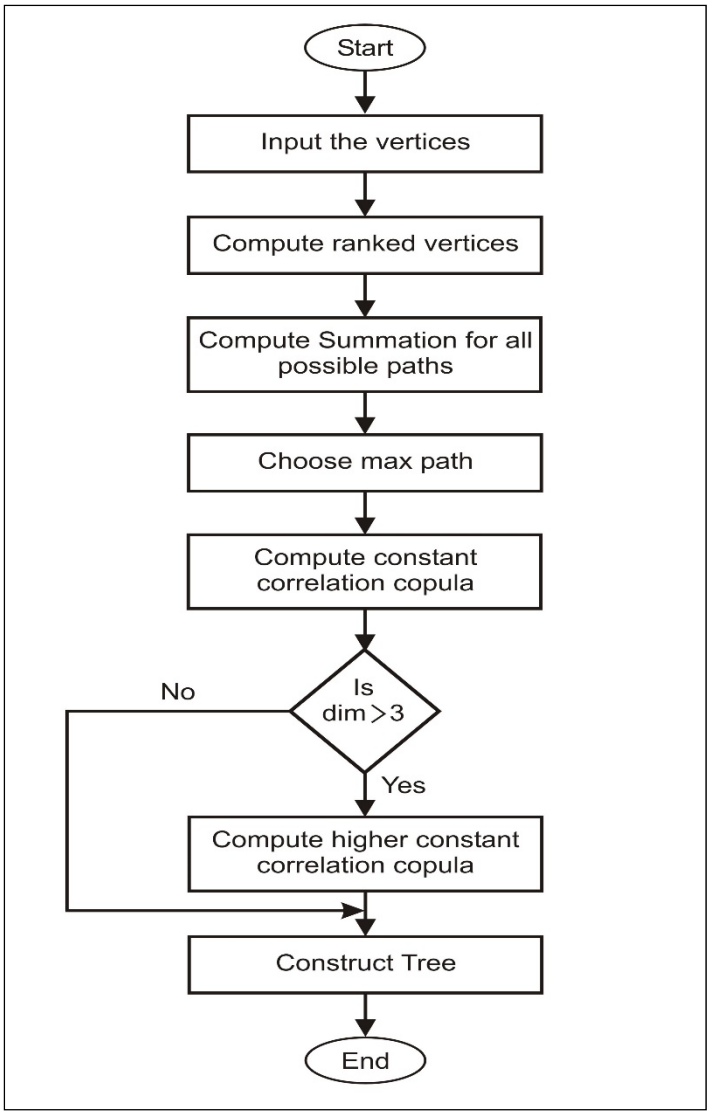

FIG. 4: THE TREE GENERATION FLOW CHART

In this paper, we compare our work with that of Daniel K and Claudia in 2017 [7], which had been basically depended on [7]. However, there are many differences which are shown as follows:

1. The dataset is different.
2. The type of correlation coefficient used in this work is Spearman.
3. For the patrial correlation, also the Spearman correlation type has been used.
4. The evaluation used in our work is based on two types; AIC and BIC.
5. The final results can be employed to generate an emission matrix (probability) which helps in making a decision.
6. The height dimension problem is solved by eliminating weak attributes and applying the regular Vine Copula.

VI. RESULTS AND DISCUSSION

In this research, we illustrate the design and implementation of the vine structure model. The four dimensions of the vertices are (hot zone, gender, crime type, and social status) as listed in Table 1. These vertices are represented as x1,x2, x3, and x4, respectively.

TABLE 1: DATA SET ATTRIBUTES

Received 13 Sep 2018; Accepted 4 Dec 2018

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

Seq	Names	Classes
1	Gender	{Female,Male}
2	Crime Type	{Murder, Thief, Multiple Crimes, non_criminal
3	Hot Zone	{Hot Zone, Safe Zone}
4	Social status	{Single,Married,Widowed }

The adopted data set is the crime in Iraq. In addition, the implementation is done using the Matlab programming language, as follows; We create a linear correlation matrix named as R:

```
R =

    1.0000    0.2683    0.5723    0.9368
    0.2683    1.0000    0.3384    0.2495
    0.5723    0.3384    1.0000    0.5479
    0.9368    0.2495    0.5479    1.0000
```

In matrix R, the diagonal values are usually equal to 1. The ranges of correlation values can be ranked as:

- Rank_correlation(Gender, Hot Zone)=0.9368 is a very strong correlation .
- Rank correlation(Gender, Crime Type)=0.5723 is a strong correlation .
- Rank correlation(Social Status, Gender)=0.2683 is a very weak correlation.
- Rank correlation(CrimeType,Social Status) =0.338 is a very weak correlation.
- Rank correlation(CrimeType , Hot Zone)=0.5479 is a strong correlation.

So, we notice that the degree of correlation of social status is very weak in comparison with others. Thus, this attribute is eliminated.

We test the partial correlation or conditional copula of the considered three attributes (Gender, CrimeTyper, and HotZone) to generate three different trees, each one gave different values of PC and average estimation which have been computed in a recursive way. The PC results are shown in Table (2) and compared with Table (3) in average estimation.

The PC is the value of conditional (partial) correlation that depends on heuristic approach structures. There are three structures that can be generated in the following ways:

- Partial Correlation (PC) (Gender, CrimeTyper, and HotZone are constant) = 0.2016.
- Partial Correlation (Gender, HotZone and CrimeType are constant) = 0.9085.
- Partial Correlation (CrimeType, HotZone, and Gender are constant) =0.0410 .

In the first way, T2 has been selected as it has the highest value of PC, as shown in Table (2) result, colored with gray and with a value of 0.9085. T2 means the selection of the structure tree (2), where the edges are (2,1) and (2,3).In the second way, the estimation test is adopted as shown in the Table (3). The highest average of conditional correlational occurs with the fitted vine of T1, which equals 0.7546, colored in gray.

The regular vine based on weighted conditional correlation test is used to determine the maximum spanning tree. The alpha factor can affect the selection of the tree. The best value is 0.7813 when the value of alpha= 0.6, that is related to tree structure 2, as shown in Table (4).

Received 13 Sep 2018; Accepted 4 Dec 2018

 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

TABLE 2: CONDITIONAL COPULA STRATEGY

No	Tree	Conditional_copula	PC
T1	2->1->3	C23_1	0.0410
T2	1->2->3	C13_2	0.9085
T3	1->-3->2	C12_3	0.2016

TABLE 3: RESULTS OF AVERAGE ESTIMATION

No	Tree	Sum of R	Ave _Estimation
T1	2->1->3	r12+r13	0.7546
T2	1->2->3	r12+r23	0.5601
T3	1->-3->2	r13+r32	0.7424

TABLE 4: WEIGHTED PATH

No	Weighted conditional correlation Test
T1	0.5577
T2	0.7813
T3	0.7181

The AIC and BIC values of vine copula that generate the tree have been evaluated using Diβmann's algorithm. T2 has been selected with AIC and BIC values of -997.9555 and -1.0007, respectively, as seen in Table (5).

TABLE 5: AIC AND BIC

No	AIC	BIC
T1	-1.7166	-1.7192
T2	-997.9555	-1.0007
T3	1.0016	-1.0042

While based on the above tests, the tree T2 has been chosen as it obtains the highest values. The tree T2 is represented as a matrix:

T=

0.9091	0.0182	0	0.0727
0.8654	0.0096	0.0096	0.1154
0.8943	0.0325	0	0.0732
0.1818	0.0070	0.0280	0.7832

Then, we apply Bayes's theorem to produce a decision tree with their probability p. The results are divided into two parts: criminal and noncriminal:

a) *The probability of criminal*:

P (Thief /Hot Zone,male)= 0.9091
P(Thief /Hot Zone, Female)=0
P(Thief /safeZone,male)= 0.0182
P(Thief /Safe Zone,Female)= 0.0727
P(murder/Hot Zone,male)= 0.9091
P(murder/Hot Zone,Female)=0
P(murder/Safe Zone,male)= 0.0182

Received 13 Sep 2018; Accepted 4 Dec 2018
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

P(murder/Safe Zone,Female)= 0.0727
P(Multiple Crimes /Hot Zone,male)= 0.8943
P(Multiple Crimes /Hot Zone,Female)= 0
P(Multiple Crimes /Safe Zone,male)= 0.0325
P(Multiple Crimes /Safe Zone,Female)= 0

b) *The Probability of non criminal*:
P(Non_criminal /Hot Zone,male)= 0.1818
P(Non_criminal /Hot Zone,Female)= 0.0280
P(Non_criminal /Safe Zone,male)= 0.0070
P(Non_criminal /Safe Zone,Female)= 0.7832

c) The dimension is eliminated by:
   1. Selecting the highest correlate of attributes
   2. By extracting one parameter from two parameters, for example, Probability of (Multiple Crimes /Hot Zone, male)= 0.8943, the Hot Zone and male give a single value Multiple Crimes, thus we reduce three parameters into two dimensions (i.e. location and Gender) combined as one parameter.

VII. CONCLUSION

This research proposed an algorithm for selecting a suitable tree structure model of vine copula, the algorithm is the combination of the Partial Correlation with Constant (PCC) and estimation of summation pathway. The aim of this work is to generate an observable tree structure model of vine copula depending on the fitting vine that helped to make the decision using maximum summation paths of ranked data. The proposed algorithm was used to get a fit model, especially with a high number of dimensions, which was represented as a matrix that shows each criminal theft,..other types with the probability of crime happened in such locations. The obtained results showed that the proposed algorithm performed in an accurate way. We extended with the method using a test model (AIC and BIC) which was applied on the crime dataset.

REFERENCES

[1]: Bedford, T., and Cooke, R. M. (2002), "Vines: A new graphical model for dependent random variables," Annals of Statistics, 30, 1031–1068

[2]: Joe, H., Cooke, R. M., and Kurowicka, D, "Regular vines: generation algorithm and number of equivalence classes," Dependence Modeling: Vine Copula Handbook, 219–231.2011.

[3]: Mukaka, Mavuto M. "A guide to appropriate use of correlation coefficient in medical research." *Malawi Medical Journal* 24.3 (2012): 69-71.

[4]: Elidan, Gal. "Copula bayesian networks." *Advances in neural information processing systems*. 2010

[5]: Kaiser, Sebastian, Dominik Träger, and Friedrich Leisch. "Generating correlated ordinal random values." (2011).

[6] Aas, K. (2016), "Pair-Copula Constructions for Financial Applications: A Review," Econometrics, 4.

[7]: D Kraus, "Growing simplified vine copula trees: improving Dißmann's algorithm", arXiv.org,March 16, 2017.

[8]: Pereira, G., Veiga, ´A., Erhardt, T., and Czado, C. (2016), "Spatial R-vine copula for streamflow scenario simulation," Power Systems Computation Conference (PSCC), 2016, IEEE, pp. 1–7.

[9]: Geidosch, Marco, and Matthias Fischer. "Application of vine copulas to credit portfolio risk modeling." *Journal of Risk and Financial Management* 9.2 (2016): 4.

Received 13 Sep 2018; Accepted 4 Dec 2018

 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352