International Journal of
Mathematical Modelling & Computations
Vol. 12, No. 03, Summer 2022, 183- 190
DOI:10.30495/ijm2c.2022.1947200.1239

Dimension Reduction of Big Data and Deleting Noise and Its Efficiency in the Decision Tree Method and Its Use in Covid 19

F. B. Farahabadi[a], K. F. Vajargah[b,*] and R. Farnoosh[c]

[a]*Department of Statistics, Islamic Azad University, Science and Research Branch, Tehran, Iran*,
[b]*Department of Statistics, Islamic Azad University, Tehran North Branch, Tehran, Iran*,
[c]*School of Mathematics, Iran University of Science and Technology, Tehran, 16844, Iran.*

Abstract. In today's world, with the advancement of science and technology, data is generated at high speeds, and with the increase in the size and volume of data, we often face a lot of extensions and redundant data and noise data that make the task of analysis difficult. Therefore, dimension reduction of the data without losing useful information in the data is very important to prepare the data for data mining and can increase the speed and even accuracy of the analysis. In this research, we present a dimensional reduction method using a copula function that reduces the dimensions of the data by identifying the relationships between the data. The copula function provides a good pattern of dependence for comparing multivariate distributions to better identify the relationship between data. In fact, by fitting the appropriate copula function to the data and estimating the copula function parameter, we measure the structural correlation of the variables and eliminate variables that are highly structurally correlated with each other. As a result, in the method presented in this study, using the copula function, we identify noise data and data with many common features and remove them from the original data.

Received: 13 December 2021, Revised: 01 February 2022, Accepted: 10 February 2022.

Keywords: Copula function; Gaussian copula function (normal); Decision tree; C4.5; Covid 19.

Index to information contained in this paper

*Corresponding author. Email: k_fathi@iau-tnb.ac.ir; fathi_kia10@yahoo.com

1. Introduction

In data mining methods, dimensional reduction is very important and necessary as one of the steps of data preprocessing, especially for big data. In classification methods, dimensional reduction even improve the classification method and make the analysis easier and more understandable. The dimensional reduction method that is commonly used in data mining is the principal component analysis (PCA) method, However, since the nature of the data changes in this method, it can not be used to reduce the dimensions for some classification methods, such as the decision tree, so we present a dimensional reduction method based on feature selection.

In this research, we present a feature-based method that uses detailed function and estimation of community parameters, identifies dimensions that are highly structurally correlated with each other and reduces data dimensions by eliminating noise and redundant dimensions and prepares the data for analysis.

In this study, we try to show how this proposed dimensional reduction method improves the efficiency of decision tree classification methods and simplifies the analysis task. To do this, we first apply the dimension reduction method to the data using copula function, and then apply the decision tree constructed with the C4.5 pattern to the original data and the reduced data, and demonstrate the efficiency of the dimension reduction method. In previous studies, best pattern for the decision tree was examined, which creates a better classification of the data [2] and the use of supervised and semi-supervised methods to improve the performance of the decision tree method were proposed [14], In this research, we present a feature selection method based on the measure of correlation between dimensions to identify redundant dimensions.

2. Copula

In brief, a copula is a function that links multivariate distributions to their marginal distributions. In other words, a copula is a multivariate distribution, the marginal distributions of which follow a normal distribution within (0, 1).

A copula is used for various reasons. First, it is a method for measuring the free-scale dependence. Second, it is a starting point for developing joint distributions with known margins. In fact, a considerable number of general studies on copulas analyze the dependence of random variables, for they allow us to distinguish between the dependence of variables and the effects of marginal distributions. This characteristic resembles the bivariate normal distribution where there are no links between its mean vector and its covariance matrix, both of which indicate the distribution simultaneously [15].

2.1 *Main features of a copula*

Assume that $C : I^2 \to I$ has the following features:

1) For every $u, v \in [0, 1]$, we will have:

$$C(u, 0) = C(0, v) = 0, C(u, 1) = u, C(1, v) = v.$$

2) For every $0 \leq v_1 < v_2 \leq 1, 0 \leq u_1 < u_2 \leq 1$, we will have:

$$C(U_2, v_2) + C(U_1, v_1) - C(U_1, v_2) - C(U_2, v_1) \geq 0.$$

Such function like C implied in the two above conditions is called copula function ([4]).

2.2 *Sklar's theory*

Assume that H is a joint probability distribution function with marginal distributions of F and G. Then C is a copula if the following equation is true for every $x, y \in \mathcal{R}$,

$$H(x, y) = C(F(x), G(y)).$$

If F and G are continuous, then the copula C is unique; otherwise, C is defined as unique on $\text{Rang}(F) \times \text{Rang}(G)$.

Conversely, if C is a copula with marginal univariate distributions F and G, then H is a function with margins F and G.

According to the Sklar's theory, if F and G have normal distributions, then:

$$H(x, y) = C(x, y).$$

It represents a copula of bivariate distribution with a normal marginal distribution within (0, 1). In other words, a copula is a bivariate distribution function with normal marginal distributions within (0, 1).

Assume that c, g, f, and h are density functions of distributions C, G, F, and H, respectively. Based on the Sklar's theory, the following equation is true:

$$h(x, y) = c(F(x), G(y)).f(x).g(y),$$

where $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ [1].

The important application of a copula is to present an appropriate method for generating distributions of correlated random multivariate variables and offer a solution to the problem of density estimation conversion. To show the problem of reversible transforms of m-dimensional random continuous variables $X_1, \ldots, X_m$ based on their distribution function into m normal independent variables $U_1 = F_1(X_1), \ldots, F_m(X_m)$, it should be assumed that $f(x_1, \ldots, x_m)$ and $c(u_1, \ldots, u_m)$ are the density probability function of $x_1, \ldots, x_m$and the join density function of $U_1, \ldots, U_m$, respectively. Since the estimation the density probability function $f(x_1, \ldots, x_m)$ can be a nonparametric form (i.e. an unknown distribution), the density probability function $c(u_1, \ldots, u_m)$ is estimated for $U_1, \ldots, U_m$ instead of $x_1, \ldots, x_m$ in this case to simplify the density estimation problem. It is then simulated to obtain the random samples $x_1, \ldots, x_m$ through the inverse transform $X_i = F^{-1}(U_i)$.

The scalar field theory indicates that there is a unique m-dimensional copula in $[0, 1]^m$ with standard normal marginal distributions $U_1, \ldots, U_m$, whereas Nelson stated that every distribution function F with margins $F_1, \ldots, F_m$ could be written as follows [11]:

$$\forall (X_1, \ldots, X_m) \in \mathbb{R}^m, \qquad F(X_1, \ldots, X_m) = C(F_1(X_1), \ldots, F_m(X_m)).$$

To evaluate a copula selected with an estimated parameter and avoid defining any hypotheses on $F_i(X_i)$, the empirical distribution function of a marginal distribution $F_i(X_i)$ can be employed to transform m samples of X into m samples of U [1, 3, 5, 11].

2.3 *Gaussian copula*

The difference between a Gaussian copula and a joint normal distribution is that the Gaussian copula allows us to have different types of a distribution function for a joint distribution. However, according to the probability theory, the multivariate normal distribution is the generalization of a one-dimensional normal distribution.

The standard multivariate Gaussian copula is defined as below:

$$c(\Phi(X_1),\ldots,\Phi(X_m)) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}X^T(\Sigma^{-1}-I)X\right),$$

where $\Phi(x_i)$ is the standard distribution of $f_i(x_i)$, whereas $X_i \sim N(0,1)$ and Σ are the correlation matrices. As a result, $c(u_1,\ldots,u_m)$ is called the Gaussian copula, and the joint density is obtained from the following equation:

$$c(u_1,\ldots,u_m) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\xi^T(\Sigma^{-1}-I)\xi\right],$$

where $u_i = \Phi(x_i)$ and $\xi = (\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_m))^T$ [8, 9].

2.4 *Copula estimation*

There are several methods for estimating a copula:

(1) Maximum Likelihood Estimation (MLE): This method is often considered difficult to use, for there are many parameters to estimate.
(2) Pseudo-MLE: There are two types of pseudo-MLE, i.e. parametric pseudo-MLE and semi-parametric pseudo-MLE. They are used more often than MLE. In pseudo-MLE, the margins are estimated through the cumulative distribution function, and the copula is then estimated through MLE.

2.5 *Maximum likelihood estimation*

Consider $Y = (Y_1,\ldots,Y_m)$ a random diagram. Assume that $F_{Y_1}(\cdot|\theta_1),\ldots,F_{Y_m}(\cdot|\theta_m)$ is a parametric model for marginal distribution functions and that $c_Y(\cdot|\theta_C)$ is a parametric model for copula Y. The following equation is true:

$$f_Y(y) = f_Y(y_1,\ldots,y_m) = c_Y(F_{Y_1}(y_1),\ldots,F_{Y_m}(y_m))\prod_{j=1}^{m} f_{Y_j}(y_j).$$

Assume that an instance of IID is $Y_{1:n} = (Y_1,\ldots,Y_n)$. The likelihood logarithm is then obtained

$$\log L(\theta_1,\ldots,\theta_m,\theta_C) = \log\prod_{i=1}^{n} f_Y(y_i)$$

$$= \sum_{i=1}^{n}(\log[c_Y(F_{Y_1}(y_{i,1}|\theta_1),\ldots F_{Y_m}(y_{i,m}|\theta_m)|\theta_C)]$$

$$+\log(f_{Y_1}(y_{i,1}|\theta_1)) + \cdots + \log(f_{Y_m}(y_{i,m}|\theta_m))).$$

ML estimators $\widehat{\theta_1}, \ldots, \widehat{\theta_2}, \widehat{\theta_C}$ are obtained from the maximization of the above equation based on $\theta_1, \ldots, \theta_m, \theta_C$.

This method has a few setbacks:

(1) There are too many parameters to estimate, especially for large values of m. As a result, optimization can be difficult.
(2) If any of the univariate parametric distributions $F_{Y_i}(\cdot|\theta_i)$ are defined incorrectly, bias can emerge in univariate distributions and the copula [4].

2.6 *Pseudo-MLE*

Pseudo-MLE helps solve the above mentioned MLE problems. This method has two steps:

(1) The marginal distribution functions are first estimated to define $\widehat{F}_{Y_j}$, for $j = 1, \ldots, m$. For this purpose, the following two methods can be adopted:
   - The empirical distribution function is defined as below for $y_{1,i}, \ldots, y_{n,j}$:

$$\widehat{F}_{Y_i}(y) = \frac{\sum_{i=1}^{n} I_{\{y_{i,j} \leq y\}}}{n+1}.$$

   - A parametric model is developed with $\widehat{\theta_j}$ obtained from the univariate conventional MLE.

(2) The parameters of copula θ_C are obtained by maximizing the following expression:

$$\sum_{i=1}^{n} \log[c_Y(\widehat{F}_{Y_1}(y_{i,1}), \ldots, \widehat{F}_{Y_m}(y_{i,m})|\theta_C].$$

It should be noted that the above expression is obtained directly from the likelihood logarithm only by using marginal distributions in Step 1 and using the parameters of θ_C that were not estimated [4].

3. New method

This study proposes a novel method for dimensionality reduction of multidimensional data. This method uses the copula theory to estimate an unlimited multivariate copula distribution in specific types of marginal distributions of random variables showing data dimensions. A copula-based model presents a complete and unscaled description of dependence. Estimating the copula parameters can facilitate the use of this model to compare the dependence of random variables. This dependence is then employed to identify the additional values and noisy data in order to cleanse the original data.

This method consists of two steps:

Step 1: In this step, pseudo-MLE (explained in the previous subsection) is adopted to link the univariate marginal distributions to their joint multivariate distribution function. After that, the copula parameters are estimated to place the dimensions with strong correlation in a smaller set. If ρ of a copula is greater than 0.7 for two random continuous variables X_1 and X_2,

these variables are strongly correlated; thus, they are placed in the subset of interest.

Step 2: In the second step, the dimensions of this subset are analyzed to delete the dimensions that are the linear combinations of the subset dimensions. Finally, the greatest value of ρ is selected from the subset, and the rest of the dimensions are deleted, for the other dimensions behave like the selected dimension and can be used as additional dimensions.
Step1:

$$X_i, X_j$$

if $\rho \geq 0.7$ then $X_i, X_j \in S_i$
otherwise $X_i, X_j \in$Reduced variables set.

Step2: if $\forall X_i \in S_i$, X_i is Linear dependent then $X_i \in$ Reduced variables set.

Finally, we select the variable that had the highest correlation with the other variables in set S_i and place it in the reduced set with the other variables.

4. Investigation and comparison of the decision tree constructed using the dimensional method based on the copula function

In this section, we examine the decision tree classification method after reducing the size of the data. First we reduce the dimensions of the data using the method presented based on copula function, and then we compare the decision tree method for the reduced data with the decision tree for the original data [7, 12].

To compare the classification methods, we use the accuracy criterion and make the training and testing set in a ratio of 80 to 20 and we make the decision tree by C4.5 criterion [6, 10, 13].

Dataset 1: Covid-19

A series of data on Covid-19 related to mortality and morbidity indices and the number of samples taken from individuals relative to the population of countries on different continents with 10 different variables is available in https://www.kaggle.com/.

The variables are as follows:

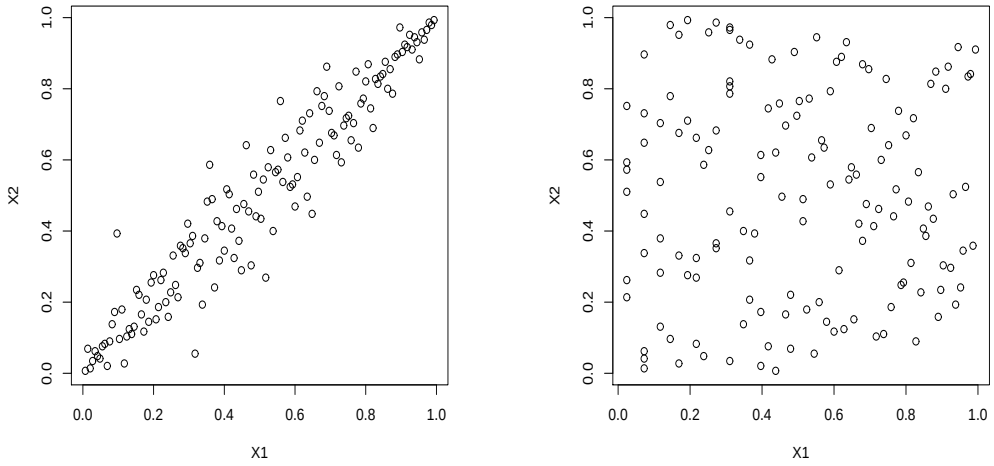
{continent, total_confirmed, total_deaths, total_recovered, active_cases, serious_or_critical,
total_cases_per_1m_population,
total_deaths_per_1m_population, total_tests, total_tests_per_1m_population}

Now we use the dimensional reduction method:

(1) First, by fitting the Gaussian copula function to the data set, we obtain the set of highly depended variables as follows:
{rotal_confirmed, total_deaths, total_recovered, active_cases, total_tests},
{total_cases_per_1m_population, total_deaths_per_1m_population}.

(2) We now check that this set of variables are linearly independent:
According to the software output, these variables are linearly independent. As a result, the reduced dimensions are as follows:
{serious_or_critical, total_cases_per_1m_population, total_tests, total_tests_per_1m_population}.

Figure 1 shows that for reduced data the correlation of the variables decreases, ie the redundant dimensions that were highly interdependent were removed.



Full data Reduced data

Figure 1. Empirical distribution function for full data and reduced data.

Table 1. Accuracy for both cases.

	Full data	Reduced data
accuracy	0.5862	0.6207

Now we use the decision tree method using the C4.5 criterion for all data and reduced data and accuracy parameter for both cases is according to Table 1.

5. Conclusion

As shown in Figure 1, we found that the new dimension reduction method eliminated highly correlated data as noise and redundant data and reduced the data size, which improved the speed of analysis. According to Table 1, I saw that in the decision tree method, the accuracy parameter is also greater than the full data for the reduced data.

As a result, the copula function-based dimensional reduction method, which reduces data dimensions by identifying and eliminating redundant variables, is an effective and efficient method that reduces data dimensions well and increases the efficiency of the decision tree method. This method can also be used for other classification methods in data mining.

References

[1] F. Badakhshan Farahabadi, K. F. Vajargah and R. Farnoosh, Dimension reduction big data using recognition of data features based on copula function and principal component analysis, Advances in Mathematical Physics, **2021** (2021), Article ID 9967368, doi:10.1155/2021/9967368.
[2] B. Charbuty and A. Abdulazeez, Classification based on decision tree algorithm for machine learning, Journal of Applied Science and Technology Trends, **2 (1)** (2021) 20–28.
[3] F. Durante, J. Fernandez-Sanchez and C. Sempi, A topological proof of sklars theorem, Applied Mathematics Letters, **26 (9)** (2013) 945–948.
[4] M. Haugh, An introduction to copulas, IEOR E4602: quantitative risk management, Lecture notes, Columbia University, (2016).

[5] R. Houari, A. Bounceur, M.-T. Kechadi, A.-K. Tari and R. Euler, Dimensionality reduction in data mining: A copula approach, Expert Systems with Applications, **64** (2016) 247–260.
[6] A. Gajewicz et al., Decision tree models to classify nanomaterials according to the DF4nanogrouping scheme, Nanotoxicology, **12 (1)** (2018) 1–17.
[7] F. Gorunescu, Data Mining: Concepts, Models and Techniques, Springer Science & Business Media, (2011).
[8] D. Lopez-Paz, J. M. Hernández-Lobato and G. Zoubin, Gaussian process vine copulas for multivariate dependence, in International Conference on Machine Learning, PMLR, (2013) 10–18.
[9] D. MacKenzie and T. Spears, The formula that killed wall street: The Gaussian copula and modelling practices in investment banking, Social Studies of Science, **44 (3)** (2014) 393–417.
[10] C. E. Metz, Basic principles of roc analysis, in Seminars in Nuclear Medicine, **8** (1978) 283–298, Elsevier.
[11] R. B. Nelsen, An Introduction to Copulas, Springer Science & Business Media, (2007).
[12] K. Nigam, J. Lafferty and A. McCallum, Using maximum entropy for text classification, in IJCAI-99 Workshop on Machine Learning for Information Filtering, **1 (1)** (1999) 61–67.
[13] H. H. Patel and P. Prajapati, Study and analysis of decision tree based classification algorithms,International Journal of Computer Sciences and Engineering, **6 (10)** (2018) 74–78.
[14] J. Tanha, M. van Someren and H. Afsarmanesh, Semi-supervised self-training for decision tree classifiers, International Journal of Machine Learning and Cybernetics, **8 (1)** (2017) 355–370.
[15] E. W. Weisstein et al., Mathworld–a wolfram web resource, (2004).