applied sciences

MDPI

Article

Nonlinear Random Forest Classification, a Copula-Based Approach

Radko Mesiar [1,2] and Ayyub Sheikhi [3,*]

[1] Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Slovak University of Technology in Bratislava, Radlinskeho 11, 810 05 Bratislava, Slovakia; radko.mesiar@stuba.sk
[2] Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, 30. Dubna 22, 701 03 Ostrava, Czech Republic
[3] Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman 7616913439, Iran
* Correspondence: sheikhy.a@uk.ac.ir

Abstract: In this work, we use a copula-based approach to select the most important features for a random forest classification. Based on associated copulas between these features, we carry out this feature selection. We then embed the selected features to a random forest algorithm to classify a label-valued outcome. Our algorithm enables us to select the most relevant features when the features are not necessarily connected by a linear function; also, we can stop the classification when we reach the desired level of accuracy. We apply this method on a simulation study as well as a real dataset of COVID-19 and for a diabetes dataset.

Keywords: random forest; copula; mutual information; classification; COVID-19

Citation: Mesiar, R.; Sheikhi, A. Nonlinear Random Forest Classification, a Copula-Based Approach. *Appl. Sci.* **2021**, *11*, 7140. https://doi.org/10.3390/app11157140

Academic Editors: Stefano Silvestri and Francesco Gargiulo

Received: 2 June 2021
Accepted: 29 July 2021
Published: 2 August 2021

1. Introduction

Dimension reduction is a major area of interest within the field of data mining and knowledge discovery, especially in high-dimensional analysis. Recently, the issue of machine learning has received considerable attention; hence, a number of researchers have sought to perform more accurate dimension reductions in this issue [1,2]. While dimension reduction tries to reduce the dimension of data by selecting some functions of the original dataset, feature selection is one of its special cases, which selects the most important features among all of them. There are many areas of statistics and machine learning that benefit from feature selection techniques. From the statistics point of view, Han and Liu et al. (2013) [3] and Basabi (2008) [4] have applied feature selection for multivariate time series. Debashis et al. (2008) [5] have investigated feature selection and regression in high-dimensional problems.

It is known that selecting the most important and relevant features is the main aim in decision tree/random forest algorithms. Although there are many classification approaches proposed in the literature, they rarely deal with the possible existence of nonlinear relations between attributes. On the other hand, note that mutual information-based filter methods have gained popularity due to their ability to capture the non-linear association between dependent and independent variables in a machine learning setting. Mutual information based on a copula function will be a good choice to carry out a feature selection in which the results are stable against noises and outliers [6,7]. So, one of the major aims of this work is using feature selection in a classification context based on a copula function, especially in random forest classification.

Random forests are commonly used machine learning algorithm, which are a combination of various independent decision trees that are trained independently on a random subset of data and use averaging to improve the predictive accuracy and control over-/under-fitting [8–11]. In this work, in order to extract the most important features in random forest, we use associated copula of features. In this regard, the connection copula between the exploratory variables as well as the associated copula of exploratory attributes and

the class labeled attribute are considered. The rest of the paper is organized as follows: we review preliminaries and introduce our method in the next section; we illustrate our algorithm considering simulated data as well as two real datasets in Section 3; finally, Section 4 is devoted to some concluding remarks.

2. Preliminaries and Related Works

The application of feature selection in machine learning and data mining techniques has been extensively considered in the literature. Kabir et al. (2020) [12] used a neural network to carry out a feature selection, while Zheng et al. (2020) [11] used a feature selection approach in a deep neural network. Li et al. (2017) [13] reviewed the feature selection techniques in data mining; see also the book of Lin and Motoda (2012) [14]. For more information, we refer to Hastie et al. (2009) [3], Chao et al. (2019) [15] and Sheikhpour et al. (2017) [16].

Peng et al. (2019) [17] and Yao et al. (2020) [18] have discussed random forest-based feature selection. It is known that the dependence structure between features plays an important role in dimension reduction. Huag et al. (2009) [19] carried out a dimension reduction based on extreme dependence between attributes. Paul et al. (2017) [5] used feature selection for outcome prediction in medical sciences. Zhang and Zhou (2010) [20] investigated multi-label dimensionality reduction features by maximizing the dependence between the original feature description and the associated class labels; see also Zhong et al. (2018) [21]. Shin and Park (2011) [22] analyzed a correlation-based dimension reduction.

In this work, we use dependence structures between variables to find the best feature selection and construct an agglomerative information gain of random forest. We apply our algorithm to classify influenza and COVID-19 patients. Iwendi et al. (2020) [23] carried out a COVID-19 patient health prediction using a random forest algorithm. Li et al. (2020) [13] applied machine learning methods to generate a computational classification model for discriminating between COVID-19 patients and influenza patients only based on clinical variables. See also Wu et al. (2020) [24], Ceylan (2020) [25] and Li et al. (2020) [13] and references therein for more information. Azar et al. (2014) [26] applied a random forest classifier for lymph diseases. See also Subasi et al. (2017) [27] for chronic kidney disease diagnosis using random forest; Açıcı et al. (2017) [28] for a random forest method to detect Parkinson disease; Jabbar et al. (2016) [29] for a prediction of heart disease using random forest. Additionally, a review work of Remeseiro et al. (2019) [30] may be helpful regarding this subject.

Sun et al. (2020) [31] have implemented a mutual information-based feature selection.

Assume that $F_{X_1,X_2,\ldots,X_d}$ is the joint multivariate distribution function of the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ and F_{X_i}, $i = 1, 2, \ldots, d$, are the related marginal distribution functions. A grounded d-increasing uniformly marginal function $C : [0,1]^d \to [0,1]$ is called a copula of $\boldsymbol{X}$ whenever it couples the multivariate distribution function $F_{X_1,X_2,\ldots,X_d}$ to its marginals $F_{X_i}, i = 1, 2, \ldots, d$, i.e.,

$$F_{X_1,X_2,\ldots,X_d}(x_1, x_2, \ldots, x_d) = C_X\big(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_d}(x_d)\big). \tag{1}$$

Note that if $\boldsymbol{X}$ is a continuous random vector, then the copula C_X is unique. For more details concerning copulas, their families and association measures, we recommend Nelsen (2006) [32] and Durante and Sempi (2016) [33]. Merits of copulas and dependence measures in dimension reduction have been discussed in the literature. See, for instance, Snehalika et al. (2020) [34] and Chang et al. (2016) [35] for copula-based feature selection; Ozdemir et al. (2017) [36] and Salinas-Gutiérrez et al. (2010) [37] for classification algorithms using copulas; Marta et al. (2017) [38] and Lascio et al. (2018) [39] for copula-based clustering approaches; Houari et al. (2106) [40] and Kluppelberg and Kuhn (2009) [41] for copula functions used in dimension reduction.

A well-known measure of uncertainty in a probability distribution is its average Hartley information measure called (Shannon) entropy. For a discrete random variable X with values $x_1, x_2, \ldots, x_n$ and mass density function $p(.)$, its entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \tag{2}$$

and for a continuous random variable X, its (differential) entropy is given by:

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x) dx, \tag{3}$$

where $\mathcal{X}$ is the support of X. Similarly, for a (continuous) multivariate random vector $\boldsymbol{X}$ of dimension k with the multivariate density $p(\boldsymbol{X})$, the entropy is defined as:

$$H(\boldsymbol{X}) = -\oint_{\mathcal{X}} p(\boldsymbol{X}) \log p(\boldsymbol{X}) d\boldsymbol{X}, \tag{4}$$

where $\oint$ is an k-integral on $\boldsymbol{\mathcal{X}}$. For two random variables X and Y with joint distribution $p(x, y)$, the conventional information gain (IG) or mutual information (MI) is defined as:

$$I(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x) p(y)} dx dy, \tag{5}$$

which is used to measure the amount of information shared by X and Y together, with convention $\frac{0}{0} = 1$. Moreover, one may generalize this concept to a continuous random vector $\boldsymbol{X} = (X_1, X_1, \ldots, X_k)$ as:

$$I(\boldsymbol{X}) = \oint_{\mathcal{X}} p(\boldsymbol{X}) \log \frac{p(\boldsymbol{X})}{\prod_{i=1}^{k} p(x_i)} d\boldsymbol{X}, \tag{6}$$

Ma and Sun (2011) [42] defined the concept of "copula entropy". Based on their definition, for a multivariate random vector $\boldsymbol{X}$, which is associated with copula density $c(\boldsymbol{u})$, its copula entropy is:

$$h_c(\boldsymbol{X}) = -\oint_{u} c(\boldsymbol{u}) \log c(\boldsymbol{X}) d\boldsymbol{u}.$$

Additionally, they have pointed out that the mutual information is a copula entropy. Indeed we have the following lemmas

Lemma 1. *Ref. [32] For a multivariate random vector $\boldsymbol{X}$ with the multivariate density $p(\boldsymbol{X})$ and copula density $c_X(\boldsymbol{u})$,*

$$I(\boldsymbol{X}) = -h_c(\boldsymbol{X}).$$

Finally, the conditional mutual information is useful to express the mutual information of two random vectors conditioned by a third random vector. If we have a k-dimensional random vector $\boldsymbol{X}$, m-dimensional random vector $\boldsymbol{Y}$ and n-dimensional random vector $\boldsymbol{Z}$, such that $\boldsymbol{X} \sim p_X(\boldsymbol{x})$, $\boldsymbol{Y} \sim p_Y(\boldsymbol{y})$, $\boldsymbol{Z} \sim p_Z(\boldsymbol{z})$, $(\boldsymbol{X}, \boldsymbol{Z}) \sim p_{X,Z}(\boldsymbol{x}, \boldsymbol{z})$, $(\boldsymbol{Y}, \boldsymbol{Z}) \sim p_{Y,Z}(\boldsymbol{y}, \boldsymbol{z})$ and $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim p_{X,Y,Z}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, then mutual information of $\boldsymbol{X}$,$\boldsymbol{Y}$ given $\boldsymbol{Z}$ which is referred to as "conditional information gain" or "conditional mutual information" of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is obtained as:

$$I(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z}) = \oint_{\mathcal{Z}} \oint_{\mathcal{Y}} \oint_{\mathcal{X}} p_{X,Y,Z}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \log \frac{p_Z(\boldsymbol{z}) p_{X,Y,Z}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{p_{X,Z}(\boldsymbol{x}, \boldsymbol{z}) p_{\boldsymbol{Y},\boldsymbol{Z}}(, y, z)} d\boldsymbol{x} d\boldsymbol{y} d\boldsymbol{z}. \tag{7}$$

3. Copula-Based Random Forest

The connection between mutual information and copula function has been investigated in the literature. We can also represent the conditional mutual information via the copula function through the following proposition:

Proposition 1. *If the random vector* $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ *is associated with copula* $C_{X,Y,Z}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ *, then Equation (7) is*

$$I(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = h_c(\boldsymbol{X}, \boldsymbol{Z}) + h_c(\boldsymbol{Y}, \boldsymbol{Z}) - h_c(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) - h_c(\boldsymbol{Z}) \tag{8}$$

Proof of Proposition 1. By an appropriate equivalent modification of the argument of the log function in the integrand of (7), we readily obtain:

$$\begin{aligned} I(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) &= \oint_{\mathcal{Z}} \oint_{\mathcal{Y}} \oint_{\mathcal{X}} p_{X,Y,Z}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \log \frac{p_Z(\boldsymbol{z}) p_{X,Y,Z}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{p_{X,Z}(\boldsymbol{x}, \boldsymbol{z}) p_{Y,Z}(, \boldsymbol{y}, \boldsymbol{z})} d\boldsymbol{x} d\boldsymbol{y} d\boldsymbol{z} \\ &= -h(\boldsymbol{X}, \boldsymbol{Z}) - h(\boldsymbol{Y}, \boldsymbol{Z}) + h(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) + h(\boldsymbol{Z}) \\ &= h_c(\boldsymbol{X}, \boldsymbol{Z}) + h_c(\boldsymbol{Y}, \boldsymbol{Z}) - h_c(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) - h_c(\boldsymbol{Z}). \end{aligned}$$

The last equality comes from one of the results of Ma and Sun (2011), which proves that for $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$,

$$h(\boldsymbol{X}) = \sum_{i=1}^{n} h(\boldsymbol{X}_i) + h_c(\boldsymbol{X}).$$

In order to use the mutual information in the decision trees, assume we have a dataset $\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the i-th input or observation and $y_i \in \mathcal{Y}$ is the corresponding outcome variable. In a machine learning approach, the major goal is constructing (or finding) a classification map $f : \mathcal{X} \to \mathcal{Y}$ which takes the features $\boldsymbol{x} \in \mathcal{X}$ of a data point as its input and outputs a predicted label. The special case of the outcome variable is a class label $y_i \in \{-1, 1\}$; i.e., it has two possible values, such as: negative/positive, pathogenic/benign, patient/normal, etc. The general objective function that must be maximized is:

$$J_{CMI}(\boldsymbol{X}_k) = I(\boldsymbol{X}_k, Y) - \beta \sum_{i=1}^{n} I(\boldsymbol{X}_i, \boldsymbol{X}_k) - \gamma \sum_{i=1}^{n} I(\boldsymbol{X}_i, \boldsymbol{X}_j | \boldsymbol{Y}) \tag{9}$$

where $I(\boldsymbol{X}_k, Y)$ measures the relation between term $\boldsymbol{X}_k$ and target variable Y, $I(\boldsymbol{X}_i, \boldsymbol{X}_k)$ quantifies the redundancy between $\boldsymbol{X}_i$ and $\boldsymbol{X}_k$; while $I(\boldsymbol{X}_i, \boldsymbol{X}_j | Y)$ measures the complementarity between terms $\boldsymbol{X}_i$ and $\boldsymbol{X}$.□

Similar to Proposition 1, one may state the Equation (9) based on copula as:

$$\begin{aligned} J_{CMI}(\boldsymbol{X}_k) &= h_c(\boldsymbol{X}_k, Y) + \beta \sum_{i=1}^{n} h_c(\boldsymbol{X}_i, \boldsymbol{X}_k) \\ &- \gamma \sum_{i=1}^{n} [h_c(\boldsymbol{X}_i, Y) + h_c(\boldsymbol{X}_k, Y) - h_c(\boldsymbol{X}_i, \boldsymbol{X}_k, Y) - h_c(Y)] \\ &= (1 - \gamma(n+1)) h_c(\boldsymbol{X}_k, Y) + n\gamma h(Y) \\ &+ \beta \sum_{i=1}^{n} h_c(\boldsymbol{X}_i, \boldsymbol{X}_k) - \gamma \sum_{k \neq i=1}^{n} h_c(\boldsymbol{X}_i, Y) + \gamma \sum_{i=1}^{n} h_c(\boldsymbol{X}_i, \boldsymbol{X}_k, Y), \end{aligned}$$

where the first term of the last part of equality refers to the relevancy of the new feature $\boldsymbol{X}_k$. Peng et al. (2019) [17] introduced the "Minimum Redundancy Maximum Relevance (mRMR)" criterion to set the value of β to be the reverse of the number of selected features

and $\gamma = 0$. We generalize their results by simplifying $J_{CMI}(\boldsymbol{X}_k)$. In particular, we have the following criterion, which we have to maximize:

$$J_{MRMR}(\boldsymbol{X}_k) = I(\boldsymbol{X}_k, Y) - \frac{1}{|\mathcal{S}|}\sum_{i=1}^{n} I(\boldsymbol{X}_i, \boldsymbol{X}_k) = -h_c(\boldsymbol{X}_k, Y) + \frac{1}{|\mathcal{S}|}\sum_{i=1}^{n} h_c(\boldsymbol{X}_i, \boldsymbol{X}_k) \tag{10}$$

From this formula, it can be seen that mRMR tries to select features that have high correlation to the target variable while they are mutually far away from each other, and in this case, the couple of functions plays an important role in the connection between the input and class level variables.

Since decision trees are prone to overfitting and do not globally find an optimal solution, their generalization, random forests, are suggested to overcome these disadvantages. Our algorithm considers the dependence between attributes to provide the best feature selection set and embeds these selected features to a random forest procedure. In this approach, we first use the dependence between attributes to choose the max-dependent as well as the max-relevant features to the class label and eliminate the max-redundant features. From the point of view of these three criteria, our approach is equivalent to the method presented by Peng et al. (2019) [17].

The confusion matrix is a metric that is often used to measure the performance of a classification algorithm. It is also called a contingency table and in a binary classification it is a 2×2 table, as shown in Figure 1.

$$Sensitivity = \frac{TP}{TP + FN}. \tag{11}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{12}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Figure 1. Confusion matrix.

We use our copula-based random forest to find the most relevant features and carry out a classification task. For this, using the copula function which connects input variables with each other as well as with the class variable, we find the most important variables by maximizing of these three criterions and then, based on their priorities, we embed them to a random forest approach to classify the class label feature. We continue our selection to find the most important feature until we reach the desired level of criteria. Traditional criteria can define some values of accuracy/sensitivity/specificity. Inspired by Snehalika et al. (2020) [34], Algorithm 1 presents a pseudo code of this method. Without loss of generality, we consider that the criterion is the accuracy. The algorithm for the sensitivity and specificity is the same.

Algorithm 1. Algorithm of copula-based random forest classification.

Result Data: data set $\mathcal{D} = (\boldsymbol{X}, Y)$, thereshold value δ.
Result: Selected feature set $\mathcal{S}$, Classification results.
Initialization: $\mathcal{S} = \mathbf{0}$, *accuracy* = 0, $\boldsymbol{F}$ = all features,
while accuracy $\leq \delta$ do
$\boldsymbol{x}_R = \underbrace{argmax}_{\boldsymbol{X}_k \in \boldsymbol{F} \setminus \mathcal{S}} \left[-h_c(\boldsymbol{X}_k, Y) + \frac{1}{|\mathcal{S}|} \sum_{i=1}^{n} h_c(\boldsymbol{X}_i, \boldsymbol{X}_k) \right]$;
$\mathcal{S} = \mathcal{S} \bigcup \boldsymbol{x}_R$;
$\boldsymbol{F} = \boldsymbol{F} \setminus \mathcal{S}$;
Perform a random Forrest classification;
the accuracy of random forest classification using 13;
accuracy = Acc_{new} + *accuracy*;
end

4. Numerical Results

A simulated dataset as well as real data analysis is presented to illustrate our method.

4.1. Simulation Study

In order to carry out a simulation study, we generated data from normal distribution with copula dependence. Our considered copulas were Gaussian, *t* and Gumbel copulas; see, e.g., Nelsen (2006) [32]. Using the copula library, we first generated n = 10,000 random samples $x_1, \ldots, x_{10}$ from a 10-variate Gaussian copula where all off-diagonal elements of their correlation matrix equal to ρ = 0.85 and their marginals follow the standard normal distribution. In a similar fashion, again, we generated another 10 variates x_{11}, $\ldots$, x_{20} independent from the first 10 variables. Then, for simulating from *t*-copula, we generated 10-variates $x_{21}, \ldots, x_{30}$ from t-copula with all correlation values equal to ρ = 0.85, df = 19 [43] and their marginals follow the standard normal distribution. Finally, a bivariate Gumbel copula with θ = 5 [44] and normal marginals were generated and inserted into x_{31}, x_{32}. A schematic heatmap plot of these 32 features is shown in Figure 2. Using a linear combination, we added values of these feathers and made the outcome variable. In order to obtain a class-valued variable, we recoded the negative values of the outcome variable to "0" and other values to "1".

Using Algorithm 1, we started with n = 2 features. The most important features to classify y were x_{26} and x_{32} with sensitivity = 0.869, specificity = 0.867 and accuracy = 0.875. Continuing the selection of the most relevant features has led us to x_{26}, x_{32} and x_{31} as the first three relevant features. In order to obtain unbiased results, we performed a 10-fold cross validation, and in each fold, we left out 1000 cases as a test group and the remainder for the train set. Averages of sensitivity, specificity and accuracy were calculated to assess the algorithm. Table 1 shows the most relevant and least redundant features with their evaluation scales sensitivity, specificity and accuracy. This table helps us to assess our algorithm by monitoring its running time as well as its comparison with other algorithms. Since, after selecting the features, we use the traditional random forest approach, it is reasonable that we compare our results with the results of the traditional random forest approach. Comparing the last two rows of Table 1, we deduce that the results of our algorithm and the traditional random forest algorithm are the same.

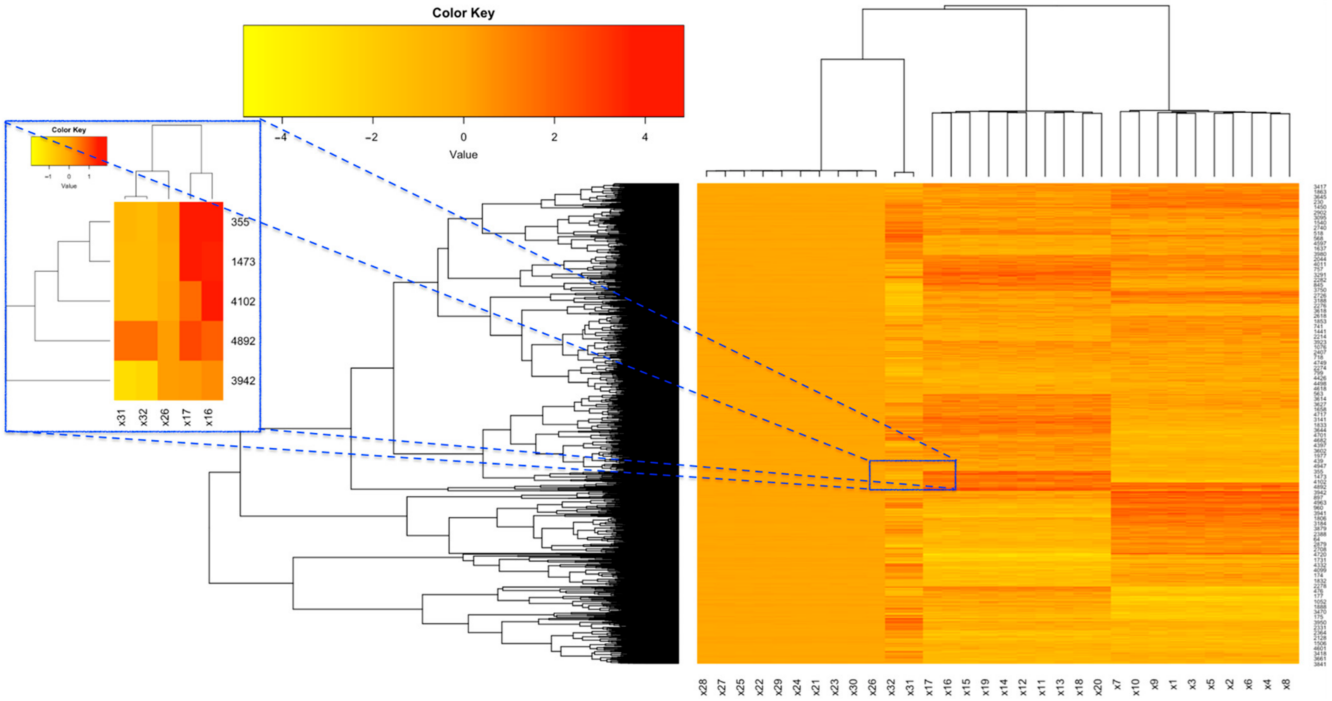

Figure 2. Heatmap plots of simulated data.

Table 1. Selected feature with their measures of assessment in simulate data.

n	Relevant Attributes	Sensitivity	Specificity	Accuracy	Running Time
2	x26, x32	0.869	0.867	0.875	4.51
3	x26, x32, x31	0.879	0.869	0.879	4.91
4	x26 x32 x31 x16	0.880	0.873	0.880	5.12
5	x26, x32, x31, x16, x13	0.881	0.878	0.881	5.55
6	x26, x32, x31, x16, x13, x12	0.888	0.880	0.886	5.83
7	x26, x32, x31, x16, x13, x12, x20	0.891	0.888	0.891	6.33
8	x26, x32, x31, x16, x13, x12, x20, x18	0.893	0.892	0.892	6.55
10	x26, x32, x31, x16, x13, x20, x12, x18, x17, x14	0.898	0.895	0.893	6.92
15	x26, x32, x31, x16, x13, x12, x20, x18, x14, x17 x15, x11, x19, x3, x7	0.908	0.896	0.901	8.34
20	x26, x32, x31, x16, x13, x12, x20, x18, x14, x17 x11, x15, x19, x3, x7, x9, x8, x5, x6, x1	0.918	0.909	0.917	11.30
25	x26, x32, x31, x16, x13, x12, x20, x18, x17, x14, x11, x15, x19, x3, x7, x8, x9, x6, x5, x1 x4, x2, x10, x27, x29	0.929	0.939	0.934	13.61
32	All attributes: x1, x2, . . . ,x32	0.982	0.979	0.981	16.75
32	Traditional random forest	0.982	0.979	0.981	16.75

Additionally, from the running time point of view, as seen from the last column of the table, for a small number of attributes, the running time (based on seconds) is negligible, and by increasing the number of attributes, the running time increases significantly. From the pros and cons point of view of the proposed approach, as understood from this table, there is a design-of-experiment approach that physicians may encounter. They can regulate the number of desired attributes to carry out a reasonable random forest classification based on the percentage of accuracy, specificity and sensitivity. Evidently, as seen from the last column of Table 1, after selecting attributes using copula, such a classification algorithm will run fast for a small number of attributes; one may think this is an operation research problem. Specifically, the sample size, the number of attributes and the complexity

of relationship between attributes play important roles in such a classification procedure. So, from the point of view of the practical implications, these results enable researchers to specify the number of attributes based on the desired levels of sensitivity, specificity and accuracy, and if the relationship between attributes is not complicated, one can choose a greater number of attributes and achieve more accuracy, and vice versa.

4.2. COVID-19 Dataset

Li et al. (2020) [13], in a meta-analysis, merged 151 datasets of COVID-19 including patient symptoms and routine test results. Nineteen clinical variables were included as explanatory inputs. The variables included age, sex, serum levels of neutrophil (continuous and ordinal), serum levels of leukocytes (continuous and ordinal), serum levels of lymphocytes (continuous and ordinal), results of CT scans, results of chest X-rays, reported symptoms (diarrhea, fever, coughing, sore throat, nausea and fatigue), body temperature, and underlying risk factors (renal diseases and diabetes) [13]. By applying machine learning methods, they reanalyzed these data and investigated correlation between explanatory variables and generated a computational classification model for discriminating between COVID-19 patients and influenza patients based on clinical variables alone.

In a COVID-19 patient case, an agglomerative approach test may help diagnosis of illness. We used our copula-based feature selection to identify the most effective attributes to make a discrimination between COVID-19 patients and influenza patients. We started with two attributes. The most relevant attributes were "age" and "fatigue". We then applied these two attributes to separate the COVID-19 and influenza patients and obtained evaluation values sensitivity, specificity and accuracy, respectively as 0.755, 0.864 and 0.836. Seeking the three most important classification attributes lead us to "age", "fatigue" and "nausea/vomiting" with sensitivity equaling 0.840, specificity equaling 0.886 and accuracy equaling 0.873. Table 2 summarizes the 10 most important features with their classification evaluation's scores. As understood from this table, there is a design-of-experiment approach that a physician may encounter. In fact, the required percentage of information determines the number and types of tests of patients. For example, if there is a required 85% accuracy of classification only, then it is enough to know "age", "fatigue" and "nausea/vomiting" of patients, while for 91.4% accuracy, we need to test the 15 most important attributes.

Table 2. Selected features with their measures of assessment in COVID-19 dataset.

n	Names of Attributes	Sensitivity	Specificity	Accuracy
2	Age, Fatigue	0.755	0.864	0.836
3	Age, Fatigue, Nausea/Vomiting	0.840	0.886	0.873
4	Age, Fatigue, Nausea/Vomiting, Diarrhea	0.826	0.875	0.860
5	Age, Fatigue, Nausea/Vomiting, Diarrhea, Sore Throat	0.783	0.891	0.860
10	Age, Fatigue, Nausea/Vomiting, Diarrhea, Sore Throat, X-ray Results, Shortness of Breath, Neutrophil, Serum Levels of White Blood Cell, Risk Factors	0.735	0.922	0.865
15	Age, Fatigue, Nausea/Vomiting, Diarrhea, Sore Throat, X-ray Results, Shortness of Breath, Neutrophil, Serum Levels of White Blood Cell, Risk Factors, Temperature, Coughing, Lymphocytes, Neutrophil Categorical, Sex	0.873	0.929	0.914

4.3. Diabetes 130-US Hospitals Dataset

In this subsection, we assess our approach in a big data analysis. We apply our algorithm to classify the Diabetes 130-US hospitals dataset [45]. This dataset represents 10 years (1999–2008) of clinical care at 130 US hospitals and integrated delivery networks. It is comprised of 101,721 observations of 50 features representing patient and hospital

outcomes. The data contains such attributes as "race", "gender", "age", "admission type", "time in hospital" and another 45 attributes. Detailed descriptions of all the attributes are provided in Strack et al. (2014).

We used the "diabetesMed" variable (0 and 1) as our response/target class variable and applied the other attributes to classify patients into two groups: no medical prescription needed and medical prescription needed. Similar to the previous subsection, the results are summarized in Table 3.

Table 3. Selected features with their measures of assessment in Diabetes 130-Us hospital dataset.

n	Names of Attributes	Sensitivity	Specificity	Accuracy
2	num_medications, num_procedures	0.742	0.756	0.708
3	num_medications, num_procedures, A1Cresult	0.766	0.771	0.780
5	num_medications, num_procedures, A1Cresult, epaglinide, max_glu_serum	0.837	0.806	0.801
10	num_medications, number_diagnoses, age, A1Cresult, repaglinide, max_glu_serum, weight, glimepiride, rosiglitazone, pioglitazone	0.921	0.948	0.871
20	num_medications, number_diagnoses, age, A1Cresult, repaglinide, max_glu_serum, weight, glimepiride, rosiglitazone, pioglitazone, glyburide, number_emergency, glipizide, number_outpatient, race, metformin, diag_2, readmitted, repaglinide, diag_3	0.981	0.977	0.972
50	All attributes	0.986	0.981	0.978

5. Conclusions

A copula-based algorithm has been employed in a random forest classification. In this regard, the most important features were extracted based on their associated copulas. The simulation study as well as real data analysis have shown that the proposed couple-based algorithm may be helpful when the explanatory variables are connected nonlinearly and when we are going to extract the most important features instead of all features.

The idea of this paper may be extended in some manners. One may use this idea in a multi-class random forest classification. Additionally, a random forest regression considering the connecting copula of features will be useful. Moreover, the associated copula of features in order classification tasks such as the support vector machine, discriminant analysis and naive Bayes classification will be of interest. Many extensions of random forest have been investigated by several authors, for example, boosted random forest, deep dynamic random forest, ensemble learning methods random forest, etc. Each extension of the random forest classification may be combined with our approach to obtain better results. We are going to extend these results in a longitudinal dataset in which the outcome variables are connected using some copulas.

Author Contributions: Conceptualization, R.M. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: APVV-18-0052 and VEGA 1/0006/19.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 (accessed on 2 June 2021).

Acknowledgments: The work of the first author was supported by the project APVV-18-0052 and VEGA 1/0006/19. Additionally, the authors wish to thank the anonymous reviewers for the comments and suggestions that have led to a significant improvement of the original manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Han, M.; Liu, X. Feature selection techniques with class separability for multivariate time series. *Neurocomputing* **2013**, *110*, 29–34. [CrossRef]
2. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
3. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin, Germany, 2009.
4. Chakraborty, B. Feature selection for multivariate time series. In Proceedings of the IASC 2008 4th World Conference of IASC on Computational Statistics and Data Analysis, Yokohama, Japan, 5–8 December 2008; pp. 227–233.
5. Paul, D.; Su, R.; Romain, M.; Sébastien, V.; Pierre, V.; Isabelle, G. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Comput. Med. Imaging Graph.* **2017**, *60*, 42–49. [CrossRef]
6. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef]
7. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 1–45. [CrossRef]
8. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
9. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Ensemble Machine Learning*; Springer: Berlin, Germany, 2012; pp. 157–175.
10. Lall, S.; Sinha, D.; Ghosh, A.; Sengupta, D.; Bandyopadhyay, S. Stable feature selection using copula-based mutual information. *Pattern Recognit.* **2021**, *112*, 107697. [CrossRef]
11. Chen, Z.; Pang, M.; Zhao, Z.; Li, S.; Miao, R.; Zhang, Y.; Feng, X.; Feng, X.; Zhang, Y.; Duan, M.; et al. Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* **2020**, *36*, 1542–1552. [CrossRef]
12. Kabir, M.M.; Islam, M.M.; Murase, K. A new wrapper feature selection approach using neural network. *Neurocomputing* **2010**, *73*, 3273–3283. [CrossRef]
13. Li, W.T.; Ma, J.; Shende, N.; Castaneda, G.; Chakladar, J.; Tsai, J.C.; Apostol, L.; Honda, C.O.; Xu, J.; Wong, L.M.; et al. Using machine learning of clinical data to diagnose COVID-19. *medRxiv* **2020**, *20*, 247.
14. Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 454.
15. Chao, G.; Luo, Y.; Ding, W. Recent advances in supervised dimension reduction: A survey. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 341–358. [CrossRef]
16. Sheikhpour, R.; Sarram, M.A.; Gharaghani, S.; Chahooki, M.A.Z. A survey on semi-supervised feature selection methods. *Pattern Recognit.* **2017**, *64*, 141–158. [CrossRef]
17. Peng, X.; Li, J.; Wang, G.; Wu, Y.; Li, L.; Li, Z.; Bhatti, A.A.; Zhou, C.; Hepburn, D.M.; Reid, A.J.; et al. Random forest based optimal feature selection for partial discharge pattern recognition in hv cables. *IEEE Trans. Power Deliv.* **2019**, *34*, 1715–1724. [CrossRef]
18. Yao, R.; Li, J.; Hui, M.; Bai, L.; Wu, Q. Feature selection based on random forest for partial discharges characteristic set. *IEEE Access* **2020**, *8*, 159151–159161. [CrossRef]
19. Haug, S.; Klüppelberg, C.; Kuhn, G. Copula structure analysis based on extreme dependence. *Stat. Interface* **2015**, *8*, 93–107.
20. Zhang, Y.; Zhou, Z.H. Multilabel dimensionality reduction via dependence maximization. *ACM Trans. Knowl. Discov. Data* **2010**, *4*, 1–21. [CrossRef]
21. Zhong, Y.; Xu, C.; Du, B.; Zhang, L. Independent feature and label components for multi-label classification. In *2018 IEEE International Conference on Data Mining (ICDM)*; IEEE: Piscataway, NJ, USA, 2018; pp. 827–836.
22. Shin, Y.J.; Park, C.H. Analysis of correlation based dimension reduction methods. *Int. J. Appl. Math. Comput. Sci.* **2011**, *21*, 549–558. [CrossRef]
23. Iwendi, C.; Bashir, A.K.; Peshkar, A.; Sujatha, R.; Chatterjee, J.M.; Pasupuleti, S. COVID-19 patient health prediction using boosted random forest algorithm. *Front. Public Health.* **2020**, *8*, 357. [CrossRef] [PubMed]
24. Wu, J.; Zhang, P.; Zhang, L.; Meng, W.; Li, J.; Tong, C.; Li, Y.; Cai, J.; Yang, Z.; Zhu, J.; et al. Rapid and accurate identification of covid-19 infection through machine learning based on clinical available blood test results. *medRxiv* **2020**. [CrossRef]
25. Ceylan, Z. Estimation of COVI-19 prevalence in Italy, Spain, and France. *Sci. Total Environ.* **2020**, *729*, 138817. [CrossRef]
26. Azar, A.T.; Elshazly, H.I.; Hassanien, A.E.; Elkorany, A.M. A random forest classifier for lymph diseases. *Comput. Methods Programs Biomed.* **2014**, *113*, 465–473. [CrossRef]
27. Subasi, A.; Alickovic, E.; Kevric, J. Diagnosis of chronic kidney disease by using random forest. In *CMBEBIH 2017*; Springer: Berlin, Germany, 2017; pp. 589–594.
28. Açıcı, K.; Erdaş, Ç.B.; Aşuroğlu, T.; Toprak, M.K.; Erdem, H.; Oğul, H. A random forest method to detect parkinsons disease via gait analysis. In *International Conference on Engineering Applications of Neural Networks*; Springer: Berlin, Germany, 2017; pp. 609–619.

29. Jabbar, M.A.; Deekshatulu, B.L.; Chandra, P. Prediction of heart disease using random forest and feature subset selection. In *Innovations in Bio-Inspired Computing and Applications*; Springer: Berlin, Germany, 2016; pp. 187–196.
30. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [CrossRef]
31. Sun, L.; Yin, T.; Ding, W.; Qian, Y.; Xu, J. Multilabel feature selection using ml-relieff and neighborhood mutual information for multilabel neighborhood decision systems. *Inf. Sci.* **2020**, *537*, 401–424. [CrossRef]
32. Nelsen, R.B. *An Introduction to Copulas*; Springer Science & Business Media: Berlin, Germany, 2006.
33. Durante, F.; Sempi, C. *Principles of Copula Theory*; CRC Press: Boca Raton, FL, USA, 2015.
34. Snehalika, L.; Debajyoti, S.; Abhik, G.H.; Debarka, S.; Sanghamitra, B. Feature selection using copula-based mutual information. *Pattern Recognit.* **2021**, *112*, 107697.
35. Chang, Y.; Li, Y.; Ding, A.; Dy, J. A robust-equitable copula dependence measure for feature selection. In Proceedings of the Artificial Intelligence and Statistics, Cadiz, Spain, 9–11 May 2016; pp. 84–92.
36. Ozdemir, O.; Allen, T.G.; Choi, S.; Wimalajeewa, T.; Varshney, P.K. Copula-based classifier fusion under statistical dependence. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2740–2748. [CrossRef]
37. Salinas-Gutiérrez, R.; Hernández-Aguirre, A.; Rivera-Meraz, M.J.; Villa-Diharce, E.R. Using gaussian copulas in supervised probabilistic classification. In *Soft Computing for Intelligent Control and Mobile Robotics*; Springer: Berlin, Germany, 2010; pp. 355–372.
38. Martal, D.F.L.; Durante, F.; Pappada, R. Copula—Based clustering methods. In *Copulas and Dependence Models with Applications*; Springer: Cham, Switzerland, 2017; pp. 49–67.
39. Di Lascio, F.M.L. Coclust: An R package for copula-based cluster analysis. *Recent Appl. Data Clust.* **2018**, *93*, 74865.
40. Houari, R.; Bounceur, A.; Kechadi, M.T.; Tari, A.K.; Euler, R. Dimensionality reduction in data mining: A copula approach. *Expert Syst. Appl.* **2016**, *64*, 247–260. [CrossRef]
41. Klüppelberg, C.; Kuhn, G. Copula structure analysis. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 737–753. [CrossRef]
42. Ma, J.; Sun, Z. Mutual information is copula entropy. *Tsinghua Sci. Technol.* **2011**, *16*, 51–54. [CrossRef]
43. Demarta, S.; McNeil, A.J. The t copula and related copulas. *Int. Stat. Rev.* **2005**, *73*, 111–129. [CrossRef]
44. Wang, L.; Guo, X.; Zeng, J.; Hong, Y. Using gumbel copula and empirical marginal distribution in estimation of distribution algorithm. In *Third International Workshop on Advanced Computational Intelligence*; IEEE: Piscataway, NJ, USA, 2010; pp. 583–587.
45. Strack, B.; DeShazo, J.P.; Gennings, C.; Olmo, J.L.; Ventura, S.; Cios, K.J.; Clore, J.N. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Res. Int.* **2014**, *2014*, 781670. [CrossRef]