Stability Performance of Copula based Feature Selection Algorithm

Parisa Tavakoli[1], Mahdi Emadi[2*], Morteza Mohammadi[3]

[1,2]Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

[3]Department of Statistics, University of Zabol, Zabol, Iran

Abstract

Feature selection is a key step in many machine-learning tasks. The copula Based Feature Selection (CBFS) method is one of the best available algorithms for this purpose. This method maximizes the relevance of selected features while minimizing redundant information. In this paper, the stability of this feature selection method in the presence of noisy data is investigated by analyzing a real data set.

Keywords: Feature selection, Copula, Mutual information, Classification accuracy.
Mathematics Subject Classification (2020): 62B10, 62H20, 62H05.

1 Introduction

Advancement of technology has caused the production of more data, both in terms of volume and dimensions, which this increase affects the machine learning model Dash and *et al.* (2019). Therefore, it is necessary to implement new algorithms to understand this volume of data. One of the popular mutual information-based approaches is the Minimal-Redundancy-Maximal-Relevance criterion (MRMR) Peng and *et al.* (2005), which considers feature relevance concerning class labels and ensures that redundant features are not present in the final feature subset. The MRMR algorithm evaluate bivariate feature dependencies. Moreover, the current algorithm is susceptible to transformations, such as scaling. To address this problem Lall and *et al.* (2021) proposed a feature selection algorithm based on copula, which on several real and synthetic datasets, the proposed algorithm performed competitively in maximizing

*Corresponding author, emadi@um.ac.ir

classification accuracy. in this article, we present *Copula Based Feature Selection* (CBFS), a filter based forward sequential search technique for feature selection, which In a real data set, the proposed algorithm performed in maximizing classification accuracy. CBFS works by minimizing the copula mutual information among selected features and maximizing the same between a candidate feature and class, And that this method does not depend on the data set. We have shown that the CBFS algorithm performs more stably than the MRMR algorithm in selecting features from noisy data. Section 2 presents the main concepts. Section 3 presents Copula Based Feature Selection Algorithm and minimal-redundancy-maximal-relevance Algorithm and Stability Performance. In Section 4, we have simulated the stability of the CBFS method in feature selection from noisy data on a real data set.

2 Concepts

In this section, we show some general concepts about copulas and mutual information.

2.1 Copula Theory

Copula produces a multivariate probability distribution from multiple uniform marginal distributions. Copula (Nelsen , 2007) is also extensively used in high dimensional data applications to obtain joint distributions from a random vector, easily by estimating their marginal functions.

Let $X = (X_1, X_2)$, denote a 2-dimensional random vector with distribution function

$$F_X(X) = P(X_1 \leq x_1, X_2 \leq x_2), \qquad X = (x_1, x_2) \in \mathbb{R}^2$$

and marginal distribution functions $F_i(x) = P(X_i \leq x)$ for $x \in \mathbb{R}$ and $i = 1, 2$. If not stated otherwise, we will always assume that the marginal distribution functions F_i are continuous.

Following Sklar (1959) theorem, there exists a unique *copula* $C : [0,1]^2 \to [0,1]$ such that

$$F_X(X) = C(F_1(x_1), F_2(x_2)) \qquad \text{for all } X \in \mathbb{R}^2$$

The copula C is the joint distribution function of the random variables $U_i = F_i(X_i)$, where $i = 1, 2$. It also admits the representation

$$C(u) = F_X\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right) \qquad \text{for all } u \in [0,1]^2$$

Where F_i^{-1} denotes the quantile function of X_i , $i = 1, 2$.

2.2 Information theory

Here, we discuss some basic parts of information theory based on entropy and Mutual Information. The entropy is defined as a measure of uncertainty and average information in a random variable. The entropy of discrete random variable, $X = (x_1, x_2, \ldots, x_n)$is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

where, $p(x_i)$ is the probability mass function.

For any two discrete random variables X and Y, mutual information is then given by

$$I(X,Y) = H(X) - H(X \mid Y)$$

2.3 Relation of Copula with Mutual Information

The mutual information between two random variables X and Y can be described in terms of the copula function as

$$I(X,Y) = -H(C(u,v))$$

where, u and v are the individual marginal distributions, respectively, $P(x)$ and $P(y)$. So, It can be seen that the Mutual Informations of the random variables are similar as negative entropy of their corresponding Copula distributions.

3 Feature Selection Algorithm

Let, a dataset be D with N dimensions. The total feature space is $F = \{f_1, f_2, \ldots, f_N\}$ We want to select a sub set of features with n dimensions, where $n \ll N$. The feature subset will be $S = \{f_1, f_2, \ldots, f_n\}$. We aim to select a feature subset S, which have same or better classifier accuracy than the feature set F. Any noise in the dataset may change the selected feature subset. We develop a copula-based feature selection (CBFS) method, which is much more stable than the MRMR method.

3.1 MRMR Algorithm

Max Relevance is to search features satisfying, with the mean value of all mutual information values between individual feature x_i and class c:

$$\max D(S,c), \quad D = \frac{1}{|S|} \sum_{f_i \in S} I(f_i; c).$$

It is likely that features selected according to Max Relevance could have rich redundancy, i.e., the dependency among these features could be large. When two features highly depend on each other, the respective class discriminative power would not change much if one of them were removed. Therefore, the following minimal redundancy (Min Redundancy) condition can be added to select mutually exclusive features:

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j).$$

The criterion combining the above two constraints is called "minimal-redundancy-maximal-relevance" (mRMR). We define the operator $\Phi(D,R)$ to combine D and R and consider the following simplest form to optimize D and R simultaneously:

$$\max \Phi(D,R), \quad \Phi = D - R.$$

In practice, incremental search methods can be used to find then ear optimal features defined by $\Phi(\cdot)$. Suppose we already have S_{n-1}, the feature set with $n-1$ features. The task is to select the nth feature from the set $\{F - S_{n-1}\}$.This is done by selecting the feature that maximizes $\Phi(\cdot)$. The respective incremental algorithm optimizes the following condition:

$$\max_{f_j \in F - S_{n-1}} \Big[I(f_j; c) = \frac{1}{n-1} \sum_{f_i \in S_{n-1}} I(f_j; f_i) \Big].$$

One of the disadvantages of the MRMR method is that it depends on the data set and any noise in the data set may change the selected feature subset. Moreover, the MRMR algorithm are susceptible to transformations, such as scaling. We develop a copula-based feature selection (CBFS) method to optimize the relevancy and redundancy, which is much more stable than the MRMR method. We minimize the copula mutual information between f_i and f_s (to reduce the1 redundancy between them) and maximizing the copula mutual information be tween class label Y and f_i. So, we are indeed using the first order incremental search to select one feature in each step.

3.2 CBFS Algorithm

Algorithm 3.1. *The algorithm is presented as follows:*

- *Step 1. In this algorithm takes the data matrix M with the class label y, and the number of features that are selected as K input parameters.*
- *Step 2. The first most relevant feature is selected by maximizing the copula based mutual information between class label Y and all features in F.*
- *Step 3. Multivariate mutual information values based on copula between each unselected feature (f_i) with all selected features $(f_1, f_2, \ldots, f_s)$ are stored in D.*
- *Step 4. Copula based multivariate mutual information values between each unselected feature (f_i) with class label Y are stored in R.*
- *Step 5. Subtraction of values in D from values in R are kept in E.*
- *Step 6. A feature is selected with maximum value, obtained from E, and merged in selected feature list S iteratively.*
- *Step 7. The above process (step 3 to step 6) is repeated k times.*
- *Step 8. Thus, an optimal feature subset is obtained in S.*
- *Step 9. S is returned as output.*

3.3 Stability Performance

Most of the existing mutual information based feature selection methods are dataset dependent. One of the striking advantages of the presented method is that it overcomes the downside of the primitive mutual information based approach due to its *scale invariant* property (Nelsen , 2007). the proposed method has the potential to select features from noisy data. the copula has the power to preserve the same dependency measure in transformed noisy fea tures.

Theorem 3.2. *Consider two random variables X and Y with joint copula function C and the functions α and β, let*

1. *the functions α and β both to be strictly increasing according to random variables X and Y respectively Copula based mutual information $I_c(\alpha(X), \beta(Y))$ is same as $I_c(X, Y)$.*
2. *the functions α is strictly increasing and β is strictly decreasing according to random variables X and Y respectively. Copula based mutual information $I_c(\alpha(X), \beta(Y))$ is same as $I_c(X, Y)$.*
3. *the functions α is strictly decreasing and β is strictly increasing according to random variables X and Y respectively. Copula based mutual information $I_c(\alpha(X), \beta(Y))$ is same as $I_c(X, Y)$.*
4. *the functions α and β both to be strictly decreasing according to random variables X and Y respectively. Copula based mutual information $I_c(\alpha(X), \beta(Y))$is same as $I_c(X, Y)$.*

So according to the theorem 3.2, opula based mutual information of random vari ables $\alpha(X)$ and $\beta(Y)$ are here same as that of the random variables X, Y, where α and β are increasing or decreasing function of X and Y. Since the proposed CBFS method is strictly based on this copula based mutual informa tion, the selected feature set with this method also remains the same under such transformations.

4 Real Dataset

Musk's dataset contains 476 samples and 168 features. Musk is a continuous type of dataset that is processed by discretization process. Gaussian noise is added to this dataset, which uses a random normal distribution to generate normal noise and adds it to the input matrix. Then

the accuracy and stability percentage is tested with different number of selected features (10, 20, 30, 40).
The results of checking the accuracy percentage of the CBFS algorithm and the rival MRMR algorithm are shown in Table 1. The accuracy percentage of the rival MRMR method between the features of the real data set and noise is lower than the CBFS method. This shows the stability of the CBFS compared to the MRMR. The performance of MRMR against noisy data is as follows: As the standard deviation of noisy data increases, the percentage of accuracy decreases and as the number of selected features increases, the percentage of accuracy increases.

5 Discussion and Resuls

In this paper, we have presented CBFS, a multivariate copula based approach for feature selection. The main characteristics of the CBFS method is that the copula based objective function for feature selection has a scale invariance feature, hence CBFS can be stable in noisy datasets. Therefore, CBFS performance is stable against noisy data. The high accuracy percentage between the features in the real dataset and the noise proves this fact.

Table 1: The percentage stability of selected features from the real and noise data set with different standard deviation using CBFS and MRMR methods

Number of selected feature	Gaussian noise parameters	Stability percentage	
		CBFS	MRMR
10	$\mu = 0, \sigma = 1$	1	0.40
	$\mu = 0, \sigma = 2$	1	0.30
	$\mu = 0, \sigma = 5$	1	0.20
	$\mu = 1, \sigma = 10$	1	0.10
20	$\mu = 0, \sigma = 1$	1	0.55
	$\mu = 0, \sigma = 2$	1	0.45
	$\mu = 0, \sigma = 5$	1	0.40
	$\mu = 1, \sigma = 10$	1	0.35
30	$\mu = 0, \sigma = 1$	1	0.70
	$\mu = 0, \sigma = 2$	1	0.66
	$\mu = 0, \sigma = 5$	1	0.63
	$\mu = 1, \sigma = 10$	1	0.56
40	$\mu = 0, \sigma = 1$	1	0.82
	$\mu = 0, \sigma = 2$	1	0.80
	$\mu = 0, \sigma = 5$	1	0.77
	$\mu = 1, \sigma = 10$	1	0.75

References

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, **6(1)**, 1-25.

Peng, H., Long, F., and Ding, C. (2005).Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy *IEEE Transactions on pattern analysis and machine intelligence*, **27(8)**, 1226-1238.

Lall, S., Sinha, D., Ghosh, A., Sengupta, D., and Bandyopadhyay, S. (2021).Stable feature selection using copula based mutual information *Pattern Recognition*, **112**, 107697.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer science and business media.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP* **8(3)**, 229-231.