Statistics and Application 统计学与应用, 2021, 10(1), 70-76
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101007

Hans汉斯

基于C-Vine Copula理论的监督学习分类器的优化

王　蕾[1]，杨　光[1]，付志慧[2]

[1]沈阳师范大学，数学与系统科学学院，辽宁　沈阳
[2]闽南师范大学，数学与统计学院，福建　漳州
Email: 435979862@qq.com, yg19640202@aliyun.com, fuzhihui2001@163.com

收稿日期：2021年1月11日；录用日期：2021年2月15日；发布日期：2021年2月22日

摘　要

由于朴素贝叶斯分类器对特征变量作了独立性假设，忽略了相关性，导致在某些特征相关的情况下分类效果很差。为了提高分类效果，本文对有缺失的数据集利用C-Vine Copula理论进行填补从而得到完整的数据集，并结合Copula函数研究特征变量之间的相关性优化问题，用C-Vine Copula分类器对完整数据集做分类。结果表明，基于C-Vine Copula理论的监督学习分类器具备良好的分类性能。

关键词

缺失数据，C-Vine Copula，监督学习分类器，贝叶斯决策

Optimization for C-Vine Copula-Based Supervised Learning Classification

Lei Wang[1], Guang Yang[1], Zhihui Fu[2]

[1]Collage of Mathematics and Systems Science, Shenyang Normal University, Shenyang Liaoning
[2]Collage of Mathematics and Statistics, Minnan Normal University, Zhangzhou Fujian
Email: 435979862@qq.com, yg19640202@aliyun.com, fuzhihui2001@163.com

Received: Jan. 11th, 2021; accepted: Feb. 15th, 2021; published: Feb. 22nd, 2021

Abstract

Because of the feature independence assumption, the correlation between variables is ignored, causing that the Naive Bayes works poorly in classification for some cases when the features are

文章引用: 王蕾, 杨光, 付志慧. 基于 C-Vine Copula 理论的监督学习分类器的优化[J]. 统计学与应用, 2021, 10(1): 70-76.
DOI: 10.12677/sa.2021.101007

correlated. In this paper, for improving the classification effect, the missing datasets are filled by using C-Vine Copula theory. As a result, the complete datasets are got after imputation. By combining the copula function and investigating on the correlation between features, C-vine copula classifier is used to classify complete datasets. The obtained results show that the supervised learning classifier based on the C-Vine Copula theory has better performance.

Keywords

Missing Data, C-Vine Copula, Supervised Learning Classification, Bayesian Decision Theory

1. 引言

随着信息技术与网络技术的高速发展，各个行业每天都会产生大量数据。然而现实世界中的数据集通常有质量问题，如存在一些错误数据、缺失数据、不确定数据。在数据质量问题中，缺失数据现象尤其常见。

朴素贝叶斯分类是机器学习和数据挖掘中最流行的学习算法之一，主要用于给定特征变量的分类问题。朴素贝叶斯分类器对分类对象的特征变量作了条件独立性假设，该假设忽略了特征变量之间的相关性，从而忽略了实际数据中的相关性[1]。

Sklar (1959) [2]提出的 Copula 理论指出，一个多元联合分布可以分解为 k 个边缘分布和一个 Copula 函数，这个 Copula 函数描述了变量间的相关性。Nelsen (1999) [3]较为系统地介绍了 Copula 函数的定义和构建方法，使得 Copula 理论成为构造多元变量联合分布及描述随机变量间相依结构的重要工具。Joe (1996) [4]首次提出了 Pair Copula 的理念，Bedford 和 Cooke (2001) [5]基于 Joe 的研究提出了 Pair Copula 构建(PCC)方法，利用图论中藤(Vine)描述结构，将高维 Copula 函数仿照树藤结构的分解形式分解为一系列二元成对 Copula 函数，称为 Vine Copula 模型。Aas (2009) [6]等进一步深入研究 Vine Copula 模型，详细论述了 Vine Copula 模型的参数估计和数值模拟的方法。

针对有缺失的数据集，本文设计并提出基于 C-Vine Copula 理论的贝叶斯分类器，进一步优化贝叶斯分类器。首先，确定需要归因的特征变量，通过 C-Vine Copula 理论将数据集的联合概率分布分解为一系列二元 Copula 函数与边缘概率密度函数乘积的形式；然后，将归因特征作为目标特征，对其他特征变量根据特征样本间的相关性，选择适当的二元 Copula 函数，计算条件分布函数和相应的逆函数，从而提出 C-Vine Copula Imputation 算法(CVI)，用该算法得到的预测值替换缺失值得到新的数据集，对新的数据集构建贝叶斯分类器中的条件概率密度函数；最后，将优化后的分类器应用到实际分类问题中，对模型进行分析验证。与样本丢弃法(NADel)、均值归因法(Mean)、预测平均匹配插补法(PMM)、贝叶斯线性回归归因法(Norm)、分类回归树归因法(Cart)和在观测值中随机抽样归因法(Sample)比较，本文提出的分类器具备良好的性能，能够为有缺失的数据集的分类提供新的实现途径。

2. C-vine Copula 贝叶斯分类器

2.1. Copula 函数

根据 Sklar 定理可知，一组 n 维随机向量，其联合概率分布可以分解为 n 个一元的边缘分布函数与一

个 n 维 Copula 函数的乘积。设 $X=\left(X_1,\cdots,X_n\right)$ 为一组 n 维随机变量，其联合概率分布 $F\left(x_1,\cdots,x_n\right)$ 与边缘分布 $F_1,\cdots,F_n$ 的关系可以表示为：

$$F\left(x_1,\cdots,x_n\right)=C\left(F_1\left(x_1\right),\cdots,F_n\left(x_n\right)\right),$$

如果各边缘分布函数都是连续的，则 Copula 函数是唯一的。若 F_k 的逆函数 F_k^{-1} 存在，则

$$C\left(u_1,\cdots,u_n\right)=F\left(F_1^{-1}\left(u_1\right),\cdots,F_n^{-1}\left(u_n\right)\right),$$

其中 $u_k=F_k\left(x_k\right)\in\left(0,1\right)$，$k=1,2,\cdots,p$。当 Copula 函数可微时，可以得到 F 的联合概率分布

$$f\left(x_1,\cdots,x_n\right)=\prod_{k-1}^{p}f_k\left(x_k\right)\cdot c\left(F_1\left(x_1\right),\cdots,F_n\left(x_n\right)\right),$$

$f_k\left(x_k\right)$ 是 F_k 的边缘密度，c 是 C 的 Copula 密度。

$$f\left(x_1,\cdots,x_n\right)=f_n\left(x_n\right)\cdot f\left(x_{n-1}\middle|x_n\right)\cdot f\left(x_{n-2}\middle|x_{n-1},x_n\right)\cdots f\left(x_1\middle|x_2,\cdots,x_n\right) \tag{1}$$

根据 Aas [6]等的论述，可以再次分解条件概率密度函数式(1)，可得

$$f\left(x|v\right)=c_{x,v_j|v_{-j}}\left(F\left(x|v_{-j}\right),F\left(v_k|v_{-j}\right)\right)\cdot f\left(x|v_{-j}\right),$$

其中，条件分布的表达式为

$$F\left(x|v\right)=\frac{\partial C_{x,v_j|v_{-j}}\left(F\left(x|v_{-j}\right),F\left(v_j|v_{-j}\right)\right)}{\partial F\left(v_j|v_{-j}\right)},$$

v 是随机变量 x 去掉 x_i 后的 $n-1$ 维向量；v_{-j} 是 v 中去掉 v_j 后的向量；v_j 为 v 中任意一个成分；$c_{x,v_j|v_{-j}}$ 为相应条件下的二元 Copula 条件概率密度函数。

当 $v_{-k}=\Phi$，x 和 v 都是均匀分布时，式(6)可简化得到

$$h\left(x,v\right)=F\left(x|v\right)=\frac{\partial C_{x,v}\left(x,v\right)}{\partial v}, \tag{2}$$

其中 $h\left(\cdot\right)$ 称为 h-函数，是用于生成伪观测值的函数，之后将用它来拟合 C-Vine 结构的模型[7]。

2.2. C-Vine Copula Imputation 算法

在构建多维变量的联合概率分布函数时，可以用多种二元 Copula 结构描述变量之间的相关性，其中 C-vine Copula 和 D-vine Copula 是最典型的两种[8]。本文采用 C-vine Copula 结构对随机变量的复杂相关性进行建模。图 1 为一个 4 维 C-Vine 的分解实例。

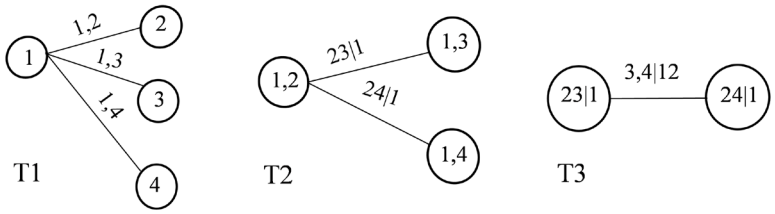

Figure 1. A 4-dimensional C-vine example
图 1. 4 维 C-vine 结构

逆采样方法是较为常见的生成随机数的方法，令 F 是 $\Re$ 上连续递增的分布函数，它的逆函数 F^{-1} 定义为：

$$F^{-1}:(0,1)\to\Re,\quad F^{-1}(x):=\inf\{y\in\Re:F(y)\geq x\}。$$

如果$U\sim U[0,1]$是在$[0,1]$上的均匀分布随机变量，则$F^{-1}(U)$有分布函数F。如果X有分布函数F，则$F(X)$是$[0,1]$上的均匀分布。

Bedford 和 Cooke [5] [9]以及 Kurowicka 和 Cooke 等[10]对 Vine 抽样算法都有讨论。Aas 等[6]给出 C-Vine 和 D-Vine 的一般性算法。对于$X=(X_1,\cdots,X_{n-1},X_n)$这组$n$维随机变量，$X_n$上有缺失数据。首先，生成与缺失数据数量相等的随机数v_n；再根据条件分布函数和 h-函数可以得到

$$F(x_n|x_1,\cdots,x_{n-1})=\frac{\partial_{n-1}C_{n,n-1|1:n-2}\left(F(x_n|x_1,\cdots,x_{n-2}),F(x_{n-1}|x_1,\cdots,x_{n-2})\right)}{\partial F(x_{n-1}|x_1,\cdots,x_{n-2})}。$$

所以，缺失数据X_n可以表示为

$$X_n=F^{-1}_{n|1:n-1}(V_n|X_1,\cdots,X_{n-1})。$$

2.3. C-Vine Copula 分类器

在n维 C-Vine 结构中，有$n-1$颗树$T_j\ (j=1,\cdots,n-1)$，每棵树由节点和边组成。每条边对应一个二元 Copula，树T_j的边是树T_{j+1}的节点。整个分解由n个边缘分布和$n(n-1)/2$个二元 Copula 的乘积得到的。由此可知，C-vine Copula 结构的联合概率分布函数可以分解为

$$f(x_1,\cdots,x_n)=\prod_{k=1}^{n}f(x_k)\times\prod_{i=1}^{n-1}\prod_{j=1}^{n-i}c_{i,i+j;1:(i-1)}\left(F(x_i|x_1,\cdots,x_{i-1}),F(x_{i+j}|x_1,\cdots,x_{i-1})\right),\tag{3}$$

其中，$F(\cdot|\cdot)$是条件分布函数，可由式(2)得到；$c_{i,i+j;1:(i-1)}$为二元 Copula 条件概率密度。

根据贝叶斯准则，向量$X=(X_1,\cdots,X_n)$是类别$e\in E$的概率为

$$\Pr(e|x)\propto\Pr(x|e)\cdot\Pr(e),\tag{4}$$

其中E是类向量，$E=(e_1,\cdots,e_l)'$，l是类别总数。结合联合概率分布函数式(3)，式(4)写作

$$\Pr(e|x)\propto\underbrace{\prod_{k=1}^{n}f_k(x_k|e)\prod_{i=1}^{n-1}\prod_{j=1}^{n-i}c_{i,i+j;1:(i-1)}\left(F(x_i|x_1,\cdots,x_{i-1}),F(x_{i+j}|x_1,\cdots,x_{i-1})|e\right)}_{f(x|e)}\times\Pr(e)。$$

根据最大后验决策规则(MAP)，可以得到

$$\text{classify}(x)=\{e:\arg\max f(x|e)\cdot\Pr(e)\}。$$

本文采用核函数方法估计边缘概率密度函数。核函数方法是一种非参数方法，$x_1,\cdots,x_n$是独立同分布的n个样本点，设概率密度函数为f，则测试样本x的概率密度估计值为

$$\hat{f}_n(x)=\frac{1}{n}\sum_{i=1}^{n}K_n(x-x_i)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right),$$

其中，$K(\cdot)$是高斯核函数；$h>0$称为带宽[11] [12]。

3. 实证分析

为验证本文提出方法的有效性和可行性，用 UCI 数据库中的数据集对比不同填补缺失数据方法下的分类结果。

3.1. 实验设计

首先，对数据集某一列 X_n 做完全随机缺失，缺失率分别为 5%、10%、20%、50%和 70%，将数据集分为完整数据(complete)和缺失数据(missing)。然后，根据 CVI 得出缺失数据 u_n，并将 u_n 与 70%的完整数据(complete)放在一起作为新的训练集，剩下的 30%完整数据作为测试集，依据新的训练集用 C-vine Copula 分类器对测试集进行分类。接下来，分别用样本丢弃法(NADel)、均值归因法(Mean)、预测平均匹配插补法(PMM)、贝叶斯线性回归归因法(Norm)、分类回归树归因法(Cart)及从观测值中随机抽样归因法(Sample)填补缺失数据，再在 C-Vine Copula 分类器上对测试集分类。最后，重复以上步骤 100 次，得到不同填补算法的平均分类准确性。各个填补算法的分类结果如下。

3.2. 结果分析

对 UCI 中的 2 个数据集分别在缺失率为 5%、10%、20%、50%和 70%的情况下，用 CVI、NADel、Mean、PMM、Norm、Cart 和 Sample 做填补处理。分类准确率的测试结果如表 1 和表 2 所示。由表 1、表 2 可知，相比其他填补算法，CVI 算法在大多数情况下均能得到更高的分类准确率。

Table 1. Classification accuracy for Iris data set
表 1. Iris 数据集填补算法分类性能表

缺失率	CVI	NADel	Mean	PMM	Norm	Cart	Sample
5%	0.95644	0.81200	0.94800	0.95778	0.78556	0.95578	0.95667
10%	0.95778	0.81244	0.94778	0.95533	0.72733	0.95556	0.95489
20%	0.95489	0.86022	0.95022	0.95156	0.56889	0.95422	0.95067
50%	0.94733	0.86044	0.95289	0.94867	0.46089	0.94844	0.94667
70%	0.93711	0.85356	0.94667	0.93733	0.41422	0.94356	0.93267
5%	0.95644	0.81200	0.94800	0.95778	0.78556	0.95578	0.95667
10%	0.95778	0.81244	0.94778	0.95533	0.72733	0.95556	0.95489

Table 2. Classification accuracy for Breast Cancer data set
表 2. Breast Cancer 数据集填补算法分类性能表

缺失率	CVI	NADel	Mean	PMM	Norm	Cart	Sample
5%	0.94443	0.92057	0.94352	0.92567	0.76876	0.91319	0.94067
10%	0.94529	0.93729	0.94429	0.90848	0.71419	0.91381	0.93948
20%	0.94686	0.93729	0.94638	0.92533	0.69410	0.92014	0.94100
50%	0.95129	0.90781	0.94343	0.92614	0.65114	0.87686	0.92571
70%	0.95252	0.91243	0.94867	0.89952	0.65752	0.87848	0.90086
5%	0.94443	0.92057	0.94352	0.92567	0.76876	0.91319	0.94067
10%	0.94529	0.93729	0.94429	0.90848	0.71419	0.91381	0.93948

图 2 描绘了数据集 Breast Cancer 在 5 种不同数据缺失率的情况下，用 7 种填补算法得到的分类性能比较图。由折线图可知，CVI 算法在 7 种填补算法中具有最好的分类性能。

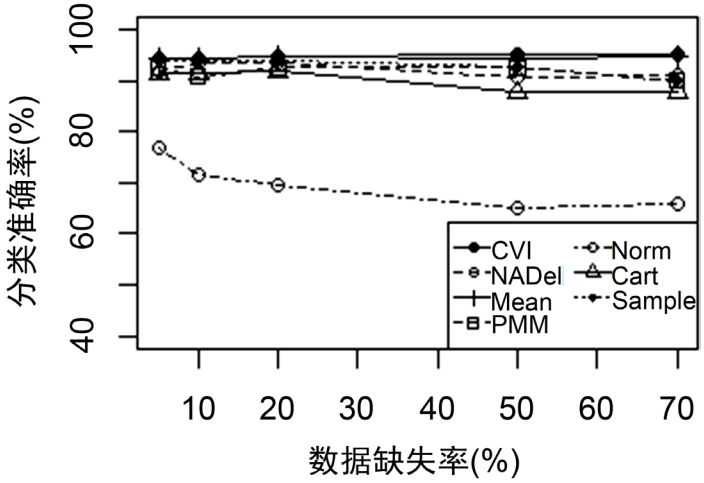

Figure 2. Classification performance of Breast Cancer dataset
图 2. Breast Cancer 的分类性能图

4. 结论

本文针对数据集中存在缺失数据的情况，结合 C-vine Copula 函数，利用条件分布函数填补缺失数据，再将变量的联合概率分布分解成二元 Copula 函数与边缘概率密度函数乘积的形式，分别对二元 Copula 函数和边缘概率密度函数进行优化估计，将特征变量之间的复杂相关性构建在条件概率密度函数中。与其他算法相比，提高了贝叶斯分类器处理具有复杂相关性特征变量数据的分类性能，在实际应用中得到了较好的分类结果。

基金项目

辽宁省教育厅自然科学基金项目(LJC201914)；辽宁省自然科学基金(2019MS285)。

参考文献

[1] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 48-53.

[2] Sklar, A. (1959) Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Universite de Paris*, **33**, 229-231.

[3] Nelsen, R.B. (1999) An Introduction to Copulas. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-3076-0

[4] Joe, H. (1997) Multivariate Models and Dependence Concepts. Chapman and Hall, London. https://doi.org/10.1201/9780367803896

[5] Bedford, T. and Cooke, R. (2001) Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. *Annals of Mathematics and Artificial Intelligence*, **32**, 245-268. https://doi.org/10.1023/A:1016725902970

[6] Aas, K., Czado, C., Frigessi, A., *et al*. (2009) Pair-Copula Constructions of Multiple Dependence. *Insurance Mathematics and Economics*, **44**, 182-198. https://doi.org/10.1016/j.insmatheco.2007.02.001

[7] Czado, C., Schepsmeier, U. and Min, A. (2011) Maximum Likelihood Estimation of Mixed C-Vines with Application to Exchange Rates. *Statistical Modelling*, **12**, 229-255. https://doi.org/10.1177/1471082X1101200302

[8] Brechmann, E.C., Schepsmeier, U., Grün, B., *et al*. (2013) Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. *J of Statistical Software*, **52**, 1-27. https://doi.org/10.18637/jss.v052.i03

[9] Bedford, T. and Cooke, R. (2002) Vines a New Graphical Model for Dependent Random Variables. *The Annals of Statistics*, **30**, 1031-1068. https://doi.org/10.1214/aos/1031689016

[10] Kurowicka, D. and Cooke, R. (2006) Uncertainty Analysis with High Dimensional Dependence Modeling. John Wiley

& Sons, Manhattan. https://doi.org/10.1002/0470863072

[11] 韦艳华, 张世英. Copula 理论及其在金融分析上的应用[M]. 北京: 清华大学出版社, 2008: 1-40.

[12] Chen, Y.H. (2014) A Copula-Based Supervised Learning Classification for Continuous and Discrete Data. *Journal of Data Science*, No. 13, 769-790.