Operations Research and Fuzziology 运筹与模糊学, 2023, 13(1), 74-87
Published Online February 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.131009

Hans汉斯

基于属性加权的独依赖条件概率编码方法

梁祖鹏[1]，李秋德[2*]，胡思贵[2]

[1]贵州大学数学与统计学院，贵州 贵阳
[2]贵州医科大学生物与工程学院，贵州 贵阳

收稿日期：2022年11月7日；录用日期：2023年1月28日；发布日期：2023年2月6日

摘 要

包含分类属性和数值属性的混合数据广泛存在于真实世界采集的数据或实验数据，在挖掘或分析这类数据前，通常需要将它们处理(转换/嵌入/表示/编码)为高质量的数值数据。条件概率编码方法(以属性条件独立假设为前提)在大多数情况下能取得不错的性能，但当它面对具有强属性关联的数据集时，性能并不理想。受独依赖值差度量的启发，将放宽属性条件独立的构想应用于条件概率编码方法。此外，还利用属性加权法来优化编码后的数据质量。融合上述这些方法，我们为混合数据的分类编码提出了一个属性加权的独依赖条件概率编码方法。实验结果表明，我们的编码方法可以显著性提高数据转换的质量，从而增强后续数据分析算法的性能。

关键词

混合数据分类，条件概率编码，独依赖值差度量，属性加权

One Dependence Conditional Probability Encoding Method Based on Attribute Weighting

Zupeng Liang[1], Qiude Li[2*], Sigui Hu[2]

[1]School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
[2]School of Biology and Engineering, Guizhou Medical University, Guiyang Guizhou

Received: Nov. 7th, 2022; accepted: Jan. 28th, 2023; published: Feb. 6th, 2023

Abstract

Mixed data containing categorical and numerical attributes are widely available in real-world or

*通讯作者。

文章引用: 梁祖鹏, 李秋德, 胡思贵. 基于属性加权的独依赖条件概率编码方法[J]. 运筹与模糊学, 2023, 13(1): 74-87. DOI: 10.12677/orf.2023.131009

experimental data sets. Before mining or analyzing such data, it is typically necessary to process (transform/embed/represent) them into high-quality numerical data. Conditional probability transformation method (which is premised on the attribute conditional independence assumption) can provide acceptable performance in the majority of cases, but it is not satisfactory for data sets with strong attribute association. Inspired by the one dependence value difference metric method, the concept of relaxing the attributes conditional independence is applied to the conditional probability transformation method. In addition, an attribute weighting method is designed to optimize the quality of data encoding. Combining these methods, we propose an Attribute Weighted One Dependence Conditional Probability Encoding method for categorical encoding on mixed data. Extensive experimental results demonstrate that our method can significantly boost the quality of data encoding, hence enhancing the performance of subsequent data analysis algorithms.

Keywords

Mixed Data Classification, Conditional Probability Encoding, One Dependence Value Difference Metric, Attribute Weighting

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在实际工作中收集的数据，由于收集方式和数据来源的不同，这些数据通常包含数值型、分类型、多媒体等多种类型，这给挖掘出有价值的信息时带来了挑战[1]。为解决数据类型多样化问题，通过数据预处理，将原始数据编码为分类型、数值型或混合型数据(包括数值型和分类型)，然后再应用于机器学习算法进一步开展数据挖掘任务[2]。就我们所知，很多高性能的机器学习算法只能处理单一的数据类型(即分类型或数值型)，例如，神经网络和回归等算法只能输入数值型数据，决策树和朴素贝叶斯等算法只能应用于分类型数据。但是这些机器学习算法中只有少数可以直接处理分类数据，并且数值数据比分类数据有更好的代数特性。因此，为了将混合数据广泛应用于支持数值输入的机器学习算法，在混合数据中把分类数据转化为数字数据是一种常见的方式[3]。

然而，如果编码过程损失过多的信息，将得到低质量的数值数据，这会严重影响后续学习算法的性能和可靠性[4]，因此，寻找一种高质量的编码方法至关重要，如何将混合型数据编码为高质量的单一数据类型吸引了研究者的广泛关注[5]。Kasif 等人[6]提出了能够处理分类型数据的 MBR 转换(或称为条件概率转换，CPT)，它在分类数据的距离度量(如值差度量[7])中取得了不错的性能。此外，我们针对 CPT 的不足，试图寻找对应的缓解策略。CPT 是以条件独立性假设为前提，当数据的属性之间存在强相关关系时，该方法估计的条件概率是不准确的。为此，学者们提出了一系列的改进方法。Jiang 等人[8]提出了隐藏朴素贝叶斯模型，它通过为每一个属性构造隐藏父属性改进条件概率公式，从而放松属性之间的条件独立性假设；Li 等人[8]通过微调条件概率的方法提出了一个微调条件概率转换；Li 等人[9]通过引入 TAN-tree 放宽条件独立性假设以改进条件概率，提出独依赖值差度量(ODVDM)。

此外，在实证研究中我们发现：转换后的数据集中，属性加权对数据挖掘起着重要作用，为此，需要设计一种适合于数据的属性加权方法。Zhang 等人[10]提出属性和实例加权的朴素贝叶斯(AIWNB)，它结合实例加权和属性加权，对于属性权重定义为互相关性和平均互关联度的差值，利用标准 sigmod 函数

处理属性权重并加权[11]。Wang 等人[12]提出双重加权的平均独依赖估计(DWAODE)，他们用属性加权和模型加权对平均独依赖估计器(AODE)进行优化，注意到以不同超父属性结点的超父独依赖估计器(SPODE)性能不同，对 AODE 中的每一个 SPODE 赋予一个权重，并且对 SPODE 中的每一个属性加权，两者结合后的模型可以提升准确性。Zhang 等[5]提出多传递距离嵌入学习(MTDLE)，这也是一种混合数据分类属性转换方法，它通过多传递距离学习得到任意两个样本之间的距离矩阵，再通过必连与勿连约束得到对应样本距离矩阵的权重并加权，作为最后的转换结果。以上运用属性加权方法都可以提升模型的性能，说明属性加权是有效的。因此，我们通过借鉴 ODVDM 和属性加权的构想，提出了一个属性加权的独依赖条件概率编码方法(AWODCPE)，改进 CPT 以提高它的分类数据编码质量。

本文的组织结构如下。第 2 节介绍了一些预备知识，包括混合数据集介绍、条件概率转换方法、TAN-tree 和属性加权方法的主要内容。第 3 节介绍属性加权的独依赖条件概率编码方法的主要内容。第 4 节设计实验来证明我们方法的数据编码性能。最后在第 5 节总结。

2. 预备知识

2.1. 混合数据集

混合数据集是指包含分类属性和数值属性的结构化数据集，如表 1 (“Automobile”数据集所示)。其中，“engine type”、“aspiration”和“fuel-type”是分类属性，“width”、“horsepower”是数值属性，“class”为分类标签。

Table 1. Description of mixed datasets
表 1. 混合数据集介绍

ID	Width	Engine-type	Aspiration	Horsepower	Fuel-type	Class
1	64.0	ohc	std	70	diesel	1
2	64.2	ohc	std	68	gas	3
3	63.6	ohcv	turbo	62	diesel	2
4	65.4	L	std	88	gas	0
5	67.9	dohc	turbo	200	gas	1
6	66.5	ohc	std	152	gas	0

2.2. 条件概率转换方法

令 $X=\{x_i\}_{i=1}^{N}$ 是一个混合数据集， $M\left(=M_c+M_n\right)$ 是所有属性 $A=\left\{A_j^c\right\}_{j=1}^{M_c}\cup\left\{A_j^n\right\}_{j=M_c+1}^{M}$ 的总数，这里 $\left\{A_j^c\right\}_{j=1}^{M_c}$ 和 $\left\{A_j^n\right\}_{j=M_c+1}^{M}$ 分别是 M_c 个分类属性和 M_n 个数值属性。数值属性可表示为：

$$A_j^n\to\left[x_{1,j}^{(n)},\cdots,x_{N,j}^{(n)}\right]^{\mathrm{T}},\ j=M_c+1,\cdots,M \tag{1}$$

这里 $x_{i,j}^{(n)}$ 是第 i 个实例第 j 个数值属性的属性值，其中 $i=1,\cdots,N;\ j=1,\cdots,M_c$。同理分类属性可表示为

$$A_j^c\to\left[x_{1,j}^{(c)},\cdots,x_{N,j}^{(c)}\right]^{\mathrm{T}},\ j=1,2,\cdots,M_c \tag{2}$$

这里 $x_{i,j}^{(c)}$ 是第 i 个实例第 j 个分类属性的属性值。CPT 将分类属性值 $x_{i,j}^{(c)}$ 转换为数值行向量 $a_j\left(x_i^{(c)}\right)$：

$$x_{i,j}^{(c)}\to a_j\left(x_i^{(c)}\right)_{1\times l}=\left[P\left(c_1\mid x_{i,j}^{(c)}\right),P\left(c_2\mid x_{i,j}^{(c)}\right),\cdots,P\left(c_l\mid x_{i,j}^{(c)}\right)\right]_{1\times l} \tag{3}$$

其中，$P\left(c_l \mid x_{i,j}^{(c)}\right)=\dfrac{\sum_{t=1}^{N}\delta\left(x_{t,j}^{(c)},x_{i,j}^{(c)}\right)\delta\left(c_i,c_t\right)+1}{\sum_{t=1}^{N}\delta\left(x_{t,j}^{(c)},x_{i,j}^{(c)}\right)+l}$，这里 l 是标签个数，$P\left(c_l \mid x_{i,j}^{(c)}\right)$ 是给定标签 c_l 属性值 $x_{i,j}^{(c)}$ 的概率。由公式(3)可知，CPT 是以属性条件独立性假设为前提，然而在所有的属性之间可能存在依赖关系，例如表 1 中“engine type”和“aspiration”或“fuel-type”之间存在联系，如果忽略这种依赖关系可能会导致估计的概率不准确[9]。

2.3. TAN-tree

独依赖值差度量[9]实行与贝叶斯分类器[13]相同的策略学习属性之间的网络结构，生成一棵 TAN-tree 以获得不同属性的依赖关系，具体步骤如下：

1) 使用下式计算任意两个分类属性之间的条件互信息

$$I\left(x_{i,j}^{(c)},x_{i,k}^{(c)} \mid c\right)=\sum_{x_{i,j}^{(c)},x_{i,k}^{(c)},c\in y}P\left(x_{i,j}^{(c)},x_{i,k}^{(c)} \mid c\right)\log\frac{P\left(x_{i,j}^{(c)},x_{i,k}^{(c)} \mid c\right)}{P\left(x_{i,j}^{(c)} \mid c\right)P\left(x_{i,k}^{(c)} \mid c\right)}; \tag{4}$$

2) 以属性为结点构建完全图，将图中任意两个结点之间无向边的权重设为 $I\left(x_{i,j}^{(c)},x_{i,k}^{(c)} \mid c\right)$；

3) 构建完全图的最大带权生成树，挑选根变量，将边设置为有向；

4) 加入标签结点 c，增加从 c 到每个属性的有向边。

2.4. 属性加权方法

属性加权[14]为 0 到 1 之间的每个属性分配连续的权重。在 AIWNB [10]中，通过以下公式计算权重，首先计算属性与标签之间和属性与属性之间的条件互信息：

$$I\left(A_j;C\right)=\sum_{x_{i,j}^{(c)}\in A_j}\sum_{c\in C}P\left(x_{i,j}^{(c)},c\right)\log\frac{P\left(x_{i,j}^{(c)},c\right)}{P\left(x_{i,j}^{(c)}\right)P(c)} \tag{5}$$

$$I\left(A_j;A_k\right)=\sum_{x_{i,j}^{(c)}\in A_j}\sum_{x_{i,k}^{(c)}\in A_k}P\left(x_{i,j}^{(c)},x_{i,k}^{(c)}\right)\log\frac{P\left(x_{i,j}^{(c)},x_{i,k}^{(c)}\right)}{P\left(x_{i,j}^{(c)}\right)P\left(x_{i,k}^{(c)}\right)} \tag{6}$$

再将公式(5)和公式(6)的条件互信息归一化：

$$NI\left(A_j;C\right)=\frac{I\left(A_j;C\right)}{\frac{1}{m}\sum_{j=1}^{m}I\left(A_j;C\right)} \tag{7}$$

$$NI\left(A_j;C\right)=\frac{I\left(A_j;A_k\right)}{\frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{k=1\wedge k\neq j}^{m}I\left(A_j;A_k\right)} \tag{8}$$

最后由属性权重定义为互相关性和平均互关联度的差值(即公式(9))，并利用 sigmoid 函数将权重映射到[0, 1]中(即公式(10))：

$$\Delta w_j^{att}=\underbrace{NI\left(A_j;C\right)}_{\text{互相关性}}-\underbrace{\frac{1}{m-1}\sum_{j=1}^{m}\sum_{k=1\wedge k\neq j}^{m}NI\left(A_j;A_k\right)}_{\text{平均互关联度}} \tag{9}$$

$$w_j^{att} = \frac{1}{1+\mathrm{e}^{-\Delta w_j^{att}}} \tag{10}$$

AIWNB [10]中的实验证明属性加权确实取得了不错的效果，每个属性的权重被纳入到公式(11)所示的朴素贝叶斯分类公式中：

$$c(x_i) = \arg\max_{c\in C} P(c)\prod_{j=1}^{m} P\left(x_{i,j}^{(c)} \mid c\right)^{w_j^{att}} \tag{11}$$

上式中，$P(c)$是标签为c时的先验概率，$P\left(x_{i,j}^{(c)} \mid c\right)$是标签为$c$时分类属性值为$x_{i,j}^{(c)}$的条件概率。除了上述的加权方式，在 MTDLE 中，通过判别嵌入的方法为不同共现行为得到的距离矩阵加权，具体用公式(12)为距离矩阵加权：

$$D = \sum_{j=1}^{n} w_j D_j \tag{12}$$

上式中，D_j是由不同共现行为得到的距离矩阵，D是加权后的距离矩阵。而我们的方法借鉴 MTDLE 的加权方式为转换后的数据加权。因此，我们结合独依赖值差度量放松属性依赖性的方法和属性加权对条件概率编码进行改进，提出属性加权的独依赖条件概率编码方法(AWODCPE)。

3. 属性加权的独依赖条件概率编码方法

3.1. 算法框架

AWODCPE 具体的编码过程分为两个部分：

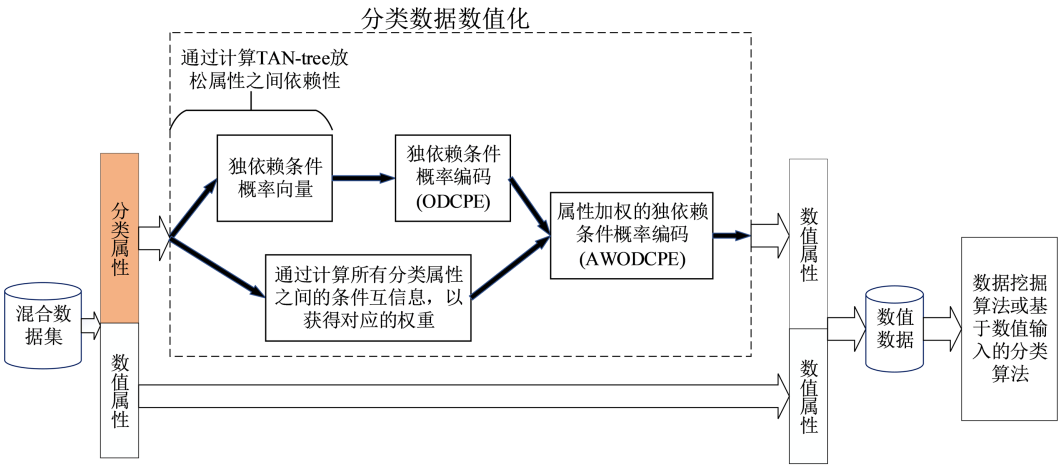

Figure 1. AWODCPE framework
图 1. AWODCPE 框架

根据图 1 第一部分提出了一个独依赖条件概率编码方法，将混合属性数据中的分类数据转换为数值数据。首先，计算每个数据集所有分类属性之间的 TAN-tree 得到分类属性之间的依赖性，改进 CPT 的公式，将一个分类属性值转换成一条条件概率向量，并替换原分类属性A_j^c中对应的分类属性值得到数值属性A_j'，以此提高混合数据的转换质量。

第二部分是基于属性之间的条件互信息的属性加权方法。由于公式(6)和公式(8)中没有考虑标签对属性与属性之间条件互信息的影响，所有我们对这两个公式进行改进，求出每个数据集的所有分类属性的权重之后并加权，表示为：

$$w_1A_1', w_2A_2', \cdots, w_{M_c}A_{M_c}' \tag{13}$$

3.2. 独依赖条件概率编码

由前面所述，CPT 具有现实中不成立的条件独立性假设，因此，我们使用独依赖值差度量相同的方法进行改进。经过计算 TAN-tree 的四个步骤，我们能够获得每个数据集中分类属性依赖关系，但仅保留了最相关属性之间的依赖性，并没有包括所有的分类属性信息。我们从 TAN-tree 中得到除根结点之外所有分类属性的父属性之后，再将分类属性值 $x_{i,j}^{(c)}$ 转换成数值向量 $a_j\left(x_i^{(c)}\right)$：

$$\begin{aligned} x_{i,j}^{(c)} &\to a_j\left(x_i^{(c)}\right)_{1\times\left(l\cdot n_{jp}\right)} \\ &= \left(\left[P\left(c_1 \mid x_{i,j}^{(c)}, x_{1,jp}^{(c)}\right), \cdots, P\left(c_l \mid x_{i,j}^{(c)}, x_{1,jp}^{(c)}\right)\right], \cdots, \left[P\left(c_1 \mid x_{i,j}^{(c)}, x_{n_{jp},jp}^{(c)}\right), \cdots, P\left(c_l \mid x_{i,j}^{(c)}, x_{n_{jp},jp}^{(c)}\right)\right]\right)_{1\times\left(l\cdot n_{jp}\right)} \end{aligned} \tag{14}$$

上式中 n_{jp} 为父属性 A_{jp} 的属性值个数，则第 j 个分类属性值可转换成

$$A_j' \to \left[a_j\left(x_1^{(c)}\right), a_j\left(x_2^{(c)}\right), \cdots, a_j\left(x_N^{(c)}\right)\right]_{N\times\left(l\cdot n_{jp}\right)}^{\mathrm{T}} \tag{15}$$

其中，$P\left(c_l \mid x_{i,j}^{(c)}, x_{i,jp}^{(c)}\right) = \dfrac{\sum_{t=1}^{N}\delta\left(x_{t,j}^{(c)}, x_{i,j}^{(c)}\right)\delta\left(x_{t,jp}^{(c)}, x_{i,jp}^{(c)}\right)\delta\left(c_i, c_t\right)+1}{\sum_{t=1}^{N}\delta\left(x_{t,j}^{(c)}, x_{i,j}^{(c)}\right)\delta\left(x_{t,jp}^{(c)}, x_{i,jp}^{(c)}\right)+l}$，式中 l 是标签个数，$\delta(\cdot)$ 是示性函数(当函数中的两个元素相等时，返回 1，否则返回 0)，$x_{i,jp}^{(c)}$ 是第 j 个属性 A_j 的父属性 A_{jp} 同一行的属性值，$P\left(c_l \mid x_{i,j}^{(c)}, x_{i,jp}^{(c)}\right)$ 是给定标签 c_l 和父属性值 $x_{i,jp}^{(c)}$ 的 $x_{i,j}^{(c)}$ 的概率。通过公式(14)和公式(15)，我们就能将分类属性转换为数值属性。

3.3. 属性加权

因为不同的属性的重要性不同，需要为其分配权重并加权。我们借鉴 AIWNB [10]的方法计算权重，但是他们的公式(6)中没有考虑标签的信息，这可能导致计算权重时出现偏差。在 DWAODE [12]中，他们通过使用属性与标签之间的条件概率计算条件互信息，我们借鉴他们的方法改进 AIWNB 的权重计算公式(6)。对公式改进如下：

$$I\left(A_j; A_k \mid C\right) = \sum_{x_{i,j}^{(c)}\in A_j}\sum_{x_{i,k}^{(c)}\in A_k}\sum_{c\in C} P\left(x_{i,j}^{(c)}, x_{i,k}^{(c)} \mid c\right)\log\frac{P\left(x_{i,j}^{(c)}, x_{i,k}^{(c)} \mid c\right)}{P\left(x_{i,j}^{(c)} \mid c\right)P\left(x_{i,k}^{(c)} \mid c\right)} \tag{16}$$

其中，

$$P\left(x_{i,j}^{(c)} \mid c\right) = \frac{P\left(x_{i,j}^{(c)}, c\right)}{P(c)} \tag{17}$$

$$P\left(x_{i,k}^{(c)} \mid c\right) = \frac{P\left(x_{i,k}^{(c)}, c\right)}{P(c)} \tag{18}$$

$$P\left(x_{i,j}^{(c)}, x_{i,k}^{(c)} \mid c\right) = \frac{P\left(x_{i,j}^{(c)}, x_{i,k}^{(c)}, c\right)}{P(c)} \tag{19}$$

公式(17)(18)(19)中的概率又由以下四个公式计算：

$$P(c)=\frac{\sum_{t=1}^{N}\delta(c_t,c)+1/q}{N+1} \tag{20}$$

$$P\left(x_{i,j}^{(c)},c\right)=\frac{\sum_{t=1}^{N}\delta\left(x_{t,j}^{(c)},x_{i,j}^{(c)}\right)\delta(c_t,c)+1/(n_j\cdot q)}{N+1} \tag{21}$$

$$P\left(x_{i,k}^{(c)},c\right)=\frac{\sum_{t=1}^{N}\delta\left(x_{t,k}^{(c)},x_{i,k}^{(c)}\right)\delta(c_t,c)+1/(n_k\cdot q)}{N+1} \tag{22}$$

$$P\left(x_{i,j}^{(c)},x_{i,k}^{(c)},c\right)=\frac{\sum_{t=1}^{N}\delta\left(x_{t,j}^{(c)},x_{i,j}^{(c)}\right)\delta\left(x_{t,k}^{(c)},x_{i,k}^{(c)}\right)\delta(c_t,c)+1/(n_j\cdot n_k\cdot q)}{N+1} \tag{23}$$

这里 q 是标签数量， n_j 和 n_k 分别是是第 j 个和第 k 个分类属性中属性值的数量，N 是实例个数。由以上四个公式计算出任意两个属性之间给定标签情况下的条件信息值后，我们归一化条件互信息：

$$NI(A_j;C)=\frac{I(A_j;C)}{\frac{1}{m}\sum_{j=1}^{m}I(A_j;C)} \tag{24}$$

$$NI(A_j;A_k\mid C)=\frac{I(A_j;A_k\mid C)}{\frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{k=1\wedge k\neq j}^{m}I(A_j;A_k\mid C)} \tag{25}$$

公式(9)可变为：

$$\Delta w_j=NI(A_j;C)-\frac{1}{m-1}\sum_{j=1}^{m}\sum_{k=1\wedge k\neq j}^{m}NI(A_j;A_k\mid C) \tag{26}$$

然而，通过公式(26)计算的权重可能为负，所以我们通过公式(10)将权重映射在区间[0, 1]中，得到分类属性的 M_c 个权重： $w_1,w_2,\cdots,w_{M_c}$ ，并对编码后的分类属性 $A_1',A_2',\cdots,A_{M_c}'$ 加权，可表示为：

$$w_1A_1',w_2A_2',\cdots,w_{M_c}A_{M_c}' \tag{27}$$

原始的数值数据不做任何处理，并与式(27)的数据拼接获得最后的转换数据：

$$w_1A_1',\cdots,w_{M_c}A_{M_c}',A_{M_c+1}^{n},\cdots,A_{M}^{n} \tag{28}$$

3.4. 算法设计与时间复杂度分析

我们的数据编码算法设计见表 2，对于 Algorithm 1 的时间复杂度如下：步骤 1 和步骤 8 至步骤 12 中计算考虑标签后的任意两个属性之间的条件互信息值花费时间按复杂度为 $O\left(cn_jn_kM_c^2\right)$ (这里 c 是标签数量，M_c 是类别属性数量，n_j 和 n_k 分别是第 j 个分类属性的属性数量和第 k 个分类属性的属性数量)；步骤 2 至步骤 6 计算每个属性对应的概率向量，时间复杂度为 $O\left(cn_jn_{jp}M_c\right)$ (这里 n_{jp} 是第 j 个分类属性的父属性属性值数量)。

4. 实验设计和评估

4.1. 实验设置

4.1.1. 参数设置

在我们的实验中，我们使用多重感知机分类器[15] (Multilayer Perceptron, MLP)评估后续数据转换学习算法的质量，构造一层隐藏层，最大迭代次数(max iter)设置为 1000，激活函数为“relu”函数。

Table 2. Algorithm design
表 2. 算法设计

Algorithm1 属性加权的独依赖条件概率编码方法

Input：分类属性 $A_1^c,\cdots,A_{M_c}^c$，数值属性 $A_{M_c+1}^n,\cdots,A_M^n$ 和标签 c.

Output：编码后的数值属性 $A_1',\cdots,A_{M_c}'$ 和对应的权重 $w=\left[w_1,\cdots,w_{M_c}\right]$。

begin

//独依赖条件概率编码

1：计算 TAN-tree 获得属性之间的依赖关系。

2：**for** 每一对分类属性 A_j^c 和它的父属性 A_{jp}^c，j=1 to M_c do

3：　　**for** 每一个可能的标签 c do

4：　　通过公式(14)计算数值概率向量 $a_j\left(x_i^{(c)}\right)$。

5：　　**end for**

6：**end for**

7：由步骤 2 to 6, 我们得到每一个分类属性对应的数值向量并替换，分类属性 $A_1^c,\cdots,A_{M_c}^c$ 可被编码为数值属性 $A_1',\cdots,A_{M_c}'$。

//属性加权

8：**for** 每一对分类属性 A_j^c 和其它属性 A_k^c，j=1 to M_c, k=1 to M_c, $j\neq k$ do

9：　　**for** 每一个可能的标签 c do

10：　通过公式(16)~(23)计算所有属性之间的条件互信息值。

11：　　**end for**

12：**end for**

13：：通过公式(24)~(26)所有分类属性的权重。

14：通过公式(10)将权重映射在区间[0, 1]中。

return 编码后的数值属性 $A_1',\cdots,A_{M_c}'$ 和对应的权重 $w=\left[w_1,\cdots,w_{M_c}\right]$。

本节给出的所有实验都是在 Intel(R) Core(TM) i5-10500 CPU @ 3.10GHz 3.10GHz 处理器上执行的，内存为 8GB，在 Windows 10 Professional x64 上运行。MLP 和 22 个编码方法是从 scikit-learn 的程序包中调用的，scikit-learn 是一个广泛使用的用 Python 实现的机器学习库。实验在 Python 3.10 环境下运行。

4.1.2. 数据集和预处理

我们基于以下原因使用来自加州大学机器学习领域 UCI 数据库的 18 个数据集：1) 这些数据集的样本实例数量差距较大，方便我们比较方法之间的转换效率；2) 它们包括两种属性，即分类属性和数值属性；3) 它们来源广泛，比如医学和商业等。

我们对这些数据集有如下处理：1) 缺失值替换处理，数值属性用平均值替换缺失值，分类属性用出现次数最多的属性值替换缺失值；2) 删除数据集中的重复对象；3) 移除属性值太多或因数据缺失导致只有一个属性值的无用属性(比如“Zoo”数据集中的属性“animal”有 100 个以上的属性值，“Anneal”数

据集中的“exptl”和“ferro”属性只有一个属性值)。我们将经过以上处理后的 18 个数据集信息总结在表中(见表 3)。

Table 3. Description of 18 UCI datasets used in the experiments
表 3. 实验中 18 个 UCI 数据集介绍

ID	数据集	实例个数	分类属性个数	数值属性个数	标签个数
1	adult	45222	8	6	2
2	automobile	159	9	15	6
3	autos	205	10	15	6
4	car	1728	6	0	4
5	census	142521	27	12	2
6	colic	368	15	7	2
7	credit-a	690	9	6	2
8	heart-statlog	270	0	13	2
9	hepatitis	155	13	6	2
10	hypothyroid	3772	21	6	4
11	kr-vs-kp	3196	36	0	2
12	labor	57	8	8	2
13	nursery	12960	8	0	5
14	primary-tumor	339	17	0	21
15	sick	3772	21	6	2
16	splice	3190	60	0	3
17	vowel	990	3	10	11
18	zoo	100	0	16	7

4.1.3. 评估标准

我们使用 20 折交叉验证的 F1 得分[16]对实验结果的性能进行评价，并在每个单元格中标出平均值 ± 标准差，使用置信水平为 0.95 的双尾 t 检验[17]在统计意义上比较在所有数据集上所有方法之间的显著性差异，并且单元格中的○和●的意义分别是我们的方法与对比方法相比在统计学意义上有显著的退化和改善。每个数据集最好的 F1 得分加粗。表底部还总结了在所有数据集上的平均值(average)，最优 F1 得分(# of best)，Win/Tie/Loss (W/T/L)和平均序值(AR)。W/T/L 意义是我们的方法与对比方法相比，赢 W 个数据集，平 T 个数据集，输 L 个数据集。我们还使用箱线图总结 F1-score，每个箱体中的直线是该方法的中位线，并且将我们的方法(AWODCPE)的中位线用红色虚线延伸，并标出数值，直观地确定方法的优缺点。

除了以上评估方法之外，我们为了进一步分析 AWODCPE 在多个数据集上的泛化性能，使用 Friedman 检验及其后续 Nemenyi 检验的统计假设检验[18]。在使用 Friedman 检验时，χ_F^2 对应的分布服从自由度为 $k-1$ 和 $(k-1)(N_d-1)$ 的 F 分布(这里 k 是算法个数，N_d 是数据集个数)。若在 95%显著性水平下，χ_F^2 值大于对应 F 分布临界值，则拒绝原假设，表明不同方法之间有显著性差异。最后，我们使用后验 Nemenyi 检验直观的确定方法之间是否有显著差异。在 p 值小于 0.05 时，Nemenyi 检验的平均序值差别的临界值为：

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N_d}} \tag{29}$$

其中 q_α 为 Tukey 分布的临界值。

4.2. 编码算法性能评估

我们从 scikit-learn 的 Category Encoder 库中选出 4 种编码方法与 CPT 和 AWODCPE 比较数据编码性能，这 4 种编码方法如下：

- Ordinal Encoder：顺序编码将分类属性值映射为连续的整数值。
- Target Encoder [19]：由 Micci 等人于 2001 年提出，使用参数函数作为加权因子计算对应的数值标签均值和总标签均值的加权平均值作为最后的数值估计。
- Catboost Encoder [20]：由 Prokhorenkova 等人于 2018 年提出，计算当前所在行的标签平均值与总标签数之比，并使用先验概率进行平滑处理，处理后的值作为最后的数值估计。
- Quatile Encoder [21]：由 Mougan 等人于 2021 年提出，利用对应属性的样本数和正则化参数定义权重，对具有相同标签的样本的 p 分位数和全局 p 分位数进行加权，加权后的值作为最后的数值估计。

AWODCPE 与以上 6 种方法的比较结果被总结在表 4、图 2 和图 3 中，我们将重点做如下总结：

1) 在表 4 中，AWODCPE 的平均的 F1 得分是最高的(86.87%)，与 CPT(84.92%)相比高 1.95%。我们的方法比其它 4 种方法高至少 3%。我们方法的最高成绩数为 12 个，远高于其它比较方法。在统计学意义上，我们方法与 CPT 相比，赢 7 个数据集，平 10 个数据集，输 1 个数据集，明显优于 CPT，并且也显著由于其它 4 个方法。平均排序为 1.72 在所有方法中最小，仅以平均排序来看 CPT 与 AWODCPE 的差距最小。

2) 从图 2 中可以看出，AWODCPE 的箱体上下限差距最小区别于其它 5 种方法，这说明我们的方法有更优秀的稳定性。并且我们的方法比其他方法有更高的箱体位置和中位线，说明拥有更好的转换性能。

3) 在 95%显著性水平下，Friedman 检验的 $\mathcal{X}_F^2(=5.21)$ 大于临界值 $F_{0.95}(5,85)(=2.32)$，说明各方法之间有显著区别。因此，我们计算 CD 值为 1.78，并以表 4 中的平均排序(AR)为基础，在图 3 中直观表现各个编码方法之间的区别。可以看出 AWODCPE 最佳，各方法之间有显著性差异，显著优于 Ordinal、Target、catboost 和 Quatile。

Table 4. F1-score comparisons for our method versus 6 encoding methods
表 4. 我们的方法与 6 个编码方法的 F1 得分比较

Datasets	AWODCPE	CPT	Catboost	Quatile	Target	Ordinal
adult	85.27 ± 0.50	85.10 ± 20.9	85.20 ± 0.24	82.99 ± 0.28●	**85.28 ± 0.22**	84.70 ± 0.34●
automobile	**78.23 ± 14.8**	62.27 ± 2.42●	68.04 ± 3.14●	72.17 ± 3.81	69.42 ± 2.38●	56.73 ± 4.93●
autos	**73.90 ± 15.2**	68.53 ± 18.8	67.82 ± 3.61	71.92 ± 2.08	68.81 ± 3.24	60.96 ± 3.78●
car	**91.37 ± 2.96**	89.82 ± 3.53●	80.05 ± 2.78●	82.99 ± 0.07●	77.87 ± 2.89●	82.85 ± 5.21●
census	**96.27 ± 0.22**	96.08 ± 1.15●	96.11 ± 0.07●	96.12 ± 0.08●	96.10 ± 0.10●	96.16 ± 0.21
colic	**83.72 ± 6.72**	80.94 ± 8.75	80.83 ± 1.59	76.07 ± 1.49●	81.22 ± 1.42	80.64 ± 2.08
credit-a	86.46 ± 4.25	86.29 ± 8.58	86.42 ± 0.61	**87.94 ± 0.81**	86.45 ± 0.47	87.90 ± 1.35
heart-statlog	**85.42 ± 8.05**	81.83 ± 6.35	81.33 ± 1.35●	81.34 ± 1.29●	81.40 ± 1.54●	82.08 ± 1.33

Continued

hepatitis	**86.87 ± 9.62**	**85.11 ± 2.26**	**87.08 ± 2.83**	89.14 ± 1.27	**87.01 ± 2.58**	**85.36 ± 1.62**
hypothyroid	**97.61 ± 1.12**	96.32 ± 0.52●	95.33 ± 0.36●	95.25 ± 0.05●	95.27 ± 0.40●	95.98 ± 0.29●
kr-vs-kp	95.18 ± 1.54	96.32 ± 0.53○	94.56 ± 0.18	96.04 ± 0.05○	94.60 ± 0.19	**99.38 ± 0.14**○
labor	**95.67 ± 14.9**	93.98 ± 2.90	92.64 ± 2.54	93.94 ± 1.68	93.43 ± 2.13	90.29 ± 3.49
nursery	**92.63 ± 1.28**	91.94 ± 1.06	89.72 ± 0.93●	78.43 ± 0.27●	90.22 ± 0.80●	90.13 ± 4.56●
primary-tumor	48.94 ± 17.9	**54.30 ± 2.37**	44.96 ± 2.07	44.62 ± 1.95	44.91 ± 1.24	47.80 ± 2.54
sick	**96.83 ± 0.92**	96.29 ± 0.97●	96.37 ± 0.25●	95.76 ± 0.11●	96.24 ± 0.14●	96.38 ± 0.15●
splice	**96.35 ± 1.24**	95.66 ± 0.17●	92.57 ± 0.35●	93.98 ± 0.38●	92.75 ± 0.37●	82.04 ± 1.58●
vowel	**75.48 ± 6.85**	70.46 ± 4.25●	70.01 ± 4.04●	62.47 ± 2.81●	69.82 ± 3.52●	67.72 ± 4.64●
zoo	97.40 ± 6.36	97.29 ± 1.71	98.12 ± 1.58	**99.91 ± 0.40**	97.76 ± 1.56	98.46 ± 2.02
average	**86.87**	84.92	83.73	83.41	83.81	82.53
# of best	**12**	1	0	3	1	1
W/T/L	-	7/10/1	9/9/0	10/7/1	9/9/0	9/8/1
AR	1.72	3.44	4.22	3.83	3.94	3.83

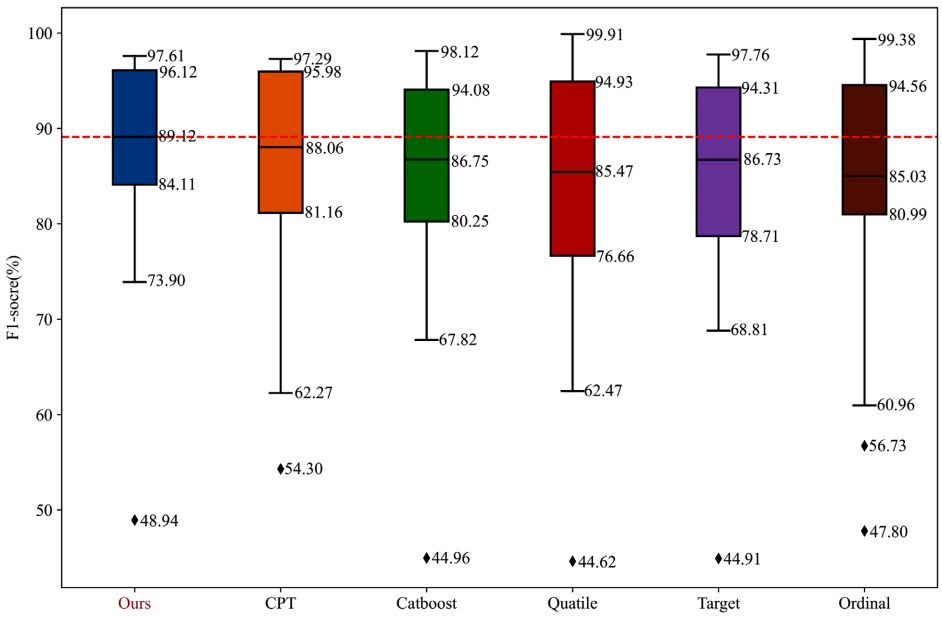

Figure 2. Comparison of AWODCPE with its competitors in terms of F1-score
图 2. 我们的方法与竞争者在 F1 得分方面的比较

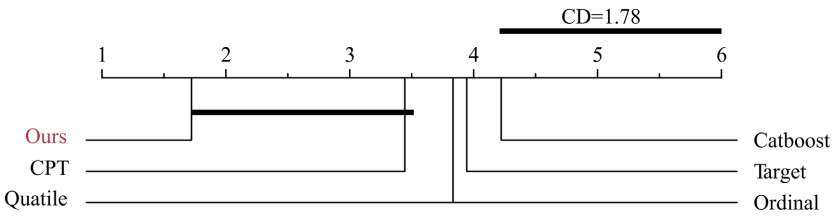

Figure 3. F1-score comparison of our method versus its competitors per Nemenyi test
图 3. 根据 Nemenyi 检验，我们的方法与竞争者的 F1 得分比较

4.3. 时间开销评估

我们将本节实验中，将转换后的数据集输入多重感知机分类器，记录分类器在这些数据集上的时间花费，并结果总结在了表 5 中。同样地，每个数据集最好的时间花费加粗。从表中可以看出，我们的方法由于维度扩张造成在“census”数据集上的时间花费远高于其他方法，导致拉高了平均值。在其他数据集上，比其他方法的花间花费稍高。

进一步地，我们在 95%显著性水平下，计算 Friedman 检验的 $\chi_F^2(=11.37)$ 大于临界值 $F_{0.95}(5,85)(=2.32)$，说明各方法之间有显著区别。因此，我们计算 CD 值为 1.78。同样地，根据表 5 中的平均排序(AR)画出图 4。从图中可以看出，我们方法的时间开销不算理想，但在转换性能和时间成本的权衡之下，是可以接受的。

Table 5. F1-score comparisons for our method versus 6 encoding methods
表 5. 我们的方法与 6 个编码方法的 F1 得分比较

Datasets	AWODCPE	CPT	Catboost	Quatile	Target	Ordinal
adult	10.978	3.5989	5.2116	**2.0965**	3.6111	4.5885
automobile	0.8736	0.3795	**0.3417**	0.3421	0.3477	0.3461
autos	1.3798	0.4071	0.3644	0.3698	0.3708	**0.3641**
car	3.0485	2.0577	0.8963	**0.0772**	1.2467	1.8829
census	231.48	18.202	20.083	17.860	11.682	**10.938**
colic	0.7452	**0.2475**	0.3864	0.3262	0.4279	0.5496
credit-a	0.8782	**0.3069**	0.5276	0.4421	0.3652	0.5132
heart-statlog	0.3254	0.3142	**0.2035**	0.2204	0.2164	0.2144
hepatitis	0.2083	0.3009	0.2264	**0.1831**	0.2396	0.3002
hypothyroid	32.703	2.6624	**1.1992**	1.9817	4.4219	3.7632
kr-vs-kp	3.2243	**1.4947**	1.6723	1.7346	1.7036	1.5515
labor	0.2748	**0.2225**	0.2713	0.2489	0.2812	0.2493
nursery	23.716	12.947	10.445	**2.1918**	15.463	3.7767
primary-tumor	3.8233	1.3706	0.3341	0.3722	**0.1895**	0.7441
sick	3.7469	**0.2184**	1.6378	1.4081	1.8878	2.3352
splice	29.048	1.9616	**0.9277**	1.0736	1.1389	3.0414
vowel	3.0248	1.5958	1.3601	**1.3582**	1.3637	1.4001
zoo	0.3169	0.3067	**0.2995**	0.3093	0.3162	0.3752
average	19.434	2.6997	2.5771	1.8109	2.5152	2.0519
# of best	0	5	5	5	1	2
AR	5.67	3.33	2.61	2.28	3.56	3.56

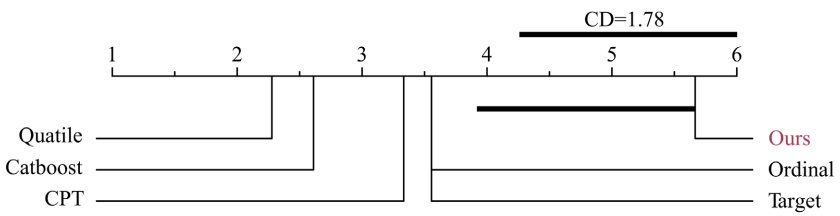

Figure 4. Time Cost comparison of our method versus its competitors per Nemenyi test
图 4. 根据 Nemenyi 检验，我们的方法与竞争者的时间花费比较

5. 结论

为了将混合数据应用于数值输入的机器学习算法，需要将分类属性数据转换为数值属性数据，而如何高质量地完成转换，这对于学者们是一大挑战。对于这个问题，本文提出了一种属性加权的独依赖条件概率编码算法(AWODCPE)。它通过建立 TAN-tree 达到放松条件独立性假设的目的，通过放松后的依赖关系求出条件概率向量，再通过所有属性之间的条件互信息值求出每一个分类属性对应的权重并加权，得到最终的转换数据。在 18 个数据集上的实验说明了 AWODCPE 的转换性能比其它 5 个竞争者更优秀。

虽然 AWODCPE 从总体上已经取得了很好的效果，然而在如下的部分还可以做进一步改进。例如，在一个数据集中有些属性并不依赖于其他属性，然而为它赋予一个父属性后，可能会降低数据转换质量，所以我们可以寻找最先进的放松条件独立性假设的方法进行改进。或者是寻找一套求每个属性值权重的更有效的方法，从而降低转换过程的时间成本，使其具有更高的转换效率，这些都是我们将来的工作方向。

基金项目

本研究由国家自然科学基金(No. 62166009)、贵州省科技厅自然科学基金 ZK [2021]333, ZK [2022]350 资助；贵州省卫健委科学基金(No. gzwkj2023-258)、贵州医科大学博士研究启动基金(No. 2020-051)资助。

参考文献

[1] Ramírez-Gallego, S., Krawczyk, B., García, S., Wozniak, M. and Herrera, F. (2017) A Survey on Data Preprocessing for Data Stream Mining: Current Status and Future Directions. *Neurocomputing*, **239**, 39-57. https://doi.org/10.1016/j.neucom.2017.01.078

[2] García, S., Luengo, J. and Herrera, F. (2015) Data Preprocessing in Data Mining. Intelligent Systems Reference Library. https://doi.org/10.1007/978-3-319-10247-4

[3] Li, Q., Xiong, Q., Ji, S., Yu, Y., Wu, C. and Yi, H. (2021) A Method for Mixed Data Classification Base on RBF-ELM Network. *Neurocomputing*, **431**, 7-22. https://doi.org/10.1016/j.neucom.2020.12.032

[4] 李秋德. 混合属性数据的处理及其分类算法研究[D]: [博士学位论文]. 重庆: 重庆大学, 2020. https://doi.org/10.27670/d.cnki.gcqdu.2020.000081

[5] Zhang, K., Wang, Q., Chen, Z., Marsic, I., Kumar, V., Jiang, G. and Zhang, J. (2015) From Categorical to Numerical: Multiple Transitive Distance Learning and Embedding. *Proceedings of the* 2015 *SIAM International Conference on Data Mining* (*SDM*), Vancouver, 30 April-2 May 2015, 46-54. https://doi.org/10.1137/1.9781611974010.6

[6] Kasif, S., Salzberg, S., Waltz, D. L., Rachlin, J. and Aha, D. W. (1998) A Probabilistic Framework for Memory-Based Reasoning. *Artificial Intelligence*, **104**, 287-311. https://doi.org/10.1016/S0004-3702(98)00046-0

[7] Stanfill, C. and Waltz, D. L. (1986) Toward Memory-Based Reasoning. *Communications of the ACM*, **29**, 1213-1228. https://doi.org/10.1145/7902.7906

[8] Jiang, L., Zhang, H. and Cai, Z. (2009) A Novel Bayes Model: Hidden Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1361-1371. https://doi.org/10.1109/TKDE.2008.234

[9] Li, C. and Li, H. (2011) One Dependence Value Difference Metric. *Knowledge-Based Systems*, **24**, 589-594. https://doi.org/10.1016/j.knosys.2011.01.005

[10] Zhang, H., Jiang, L. and Yu, L. (2021) Attribute and Instance Weighted Naive Bayes. *Pattern Recognition*, **111**, Article ID: 107674. https://doi.org/10.1016/j.patcog.2020.107674

[11] Jiang, L., Zhang, L., Li, C. and Wu, J. (2019) A Correlation-Based Feature Weighting Filter for Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, **31**, 201-213. https://doi.org/10.1109/TKDE.2018.2836440

[12] Wang, L., Xie, Y., Pang, M. and Wei, J. (2022) Alleviating the Attribute Conditional Independence and I.I.D. Assumptions of Averaged One-Dependence Estimator by Double Weighting. *Knowledge-Based Systems*, **250**, Article ID: 109078. https://doi.org/10.1016/j.knosys.2022.109078

[13] Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian Network Classifiers. *Machine Learning*, **29**, 131-163. https://doi.org/10.1023/A:1007465528199

[14] Lee, C.-H. (2018) An Information-Theoretic Filter Approach for Value Weighted Classification Learning in Naive

Bayes. *Data & Knowledge Engineering*, **113**, 116-128. https://doi.org/10.1016/j.datak.2017.11.002

[15] Popescu, M.-C., Balas, V., Perescu-Popescu, L. and Mastorakis, N. (2009) Multilayer Perceptron and Neural Networks. *WSEAS Transactions on Circuits and Systems*, **8**, 579-588.

[16] Yang, C. C. (2010) Search Engines Information Retrieval in Practice. *The Journal of the Association for Information Science and Technology*, **61**, 430. https://doi.org/10.1002/asi.21194

[17] Nadeau, C. and Bengio, Y. (2003) Inference for the Generalization Error. *Machine Learning*, **52**, 239-281. https://doi.org/10.1023/A:1024068626366

[18] Demsar, J. (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, **7**, 1-30.

[19] Micci-Barreca, D. (2001) A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. *ACM SIGKDD Explorations Newsletter*, **3**, 27-32. https://doi.org/10.1145/507533.507538

[20] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) CatBoost: Unbiased Boosting with Categorical Features. *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*, Montréal, 3-8 December 2018, 6639-6649.

[21] Mougan, C., Masip, D., Nin, J. and Pujol, O. (2021) Quantile Encoder: Tackling High Cardinality Categorical Features in Regression Problems. 18*th International Conference*, *MDAI* 2021, Umeå, 27-30 September 2021, 168-180. https://doi.org/10.1007/978-3-030-85529-1_14