【电子与信息科学 / Electronics and Information Science】

混合属性数据集分布一致性度量的新方法

何玉林^{1,2}, 金 一³, 戴德鑫¹, 黄柏皓¹, 黄家杰¹

1) 深圳大学计算机与软件学院, 广东深圳 518060; 2) 深圳大学大数据系统计算技术国家工程实验室, 广东深圳 518060; 3) 中国刑事警察学院刑事科学技术学院, 辽宁沈阳 110854

摘 要: 数据分布一致性的度量是大数据随机样本划分生成过程中的一个关键问题, 如何针对混合属性的数据集进行合理有效的分布一致性度量是目前随机样本划分技术研究的重点. 提出一种新的基于深度编码和最大平均差异的混合属性数据集分布一致性度量方法, 不直接对两个不同的原始数据集进行分布一致性的度量, 而是首先对混合属性中的离散属性进行独热编码, 得到独热编码数据集; 之后对独热编码数据集进行自编码处理, 得到深度编码数据集; 最后基于最大均值差异指标对两个不同的深度编码数据集进行分布一致性的度量. 在 Adult、Australian、CRX 和 German 标准混合属性数据集上对所提方法进行性能实验验证, 结果表明, 与基于离散属性独热编码的最大平均差异方法和基于连续属性二进制化的相似性度量方法相比, 新方法能够更加准确地对混合属性数据进行分布一致性的度量.

关键词: 人工智能; 随机样本划分; 分布一致性; 最大均值差异; 混合属性数据; 独热编码; 深度编码
中图分类号: TP311 **文献标志码:** A **doi:** 10.3724/SP.J.1249.2021.02170

A new method for measuring the distribution consistency of mixed-attribute data sets

HE Yulin^{1,2}, JIN Yi³, DAI Dexin¹, HUANG Baihao¹, and HUANG Jiajie¹

1) College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen 518060, Guangdong Province, P. R. China

2) National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,
Shenzhen 518060, Guangdong Province, P. R. China

3) College of Forensic Science and Technology, Criminal Investigation Police University of China,
Shenyang 110854, Liaoning Province, P. R. China

Abstract: The measurement of data distribution consistency is a key problem in the process of generating random sample partition (RSP) of big data. How to measure the distribution consistency of mixed-attribute data sets reasonably and effectively is the focus of current research on RSP technology. This paper proposes a new method to measure the distribution consistency of mixed-attribute data sets based on deep encoding and maximum mean discrepancy (DE-MMD). Firstly, we conduct the one-hot encoding to transform the original data set with discrete attributes into the one-hot encoding data set. Then, we construct and train an autoencoder with single hidden layer based on the

Received: 2020-04-20; **Accepted:** 2020-08-18

Foundation: Open Foundation of Key Laboratory of Impression Evidence Examination and Identification Technology, Ministry of Public Security of China (HJKF201901); Basic Research Foundation of Strengthening Police with Science and Technology of the Ministry of Public Security of China (2017GABJC09); Scientific Research Foundation of Shenzhen University for Newly-introduced Teachers (2018060); National Training Program of Innovation and Entrepreneurship for Undergraduates (201910590017)

Corresponding author: Assistant professor HE Yulin. E-mail: yulinhe@szu.edu.cn

Citation: HE Yulin, JIN Yi, DAI Dexin, et al. A new method for measuring the distribution consistency of mixed-attribute data sets [J]. Journal of Shenzhen University Science and Engineering, 2021, 38(2): 170-179. (in Chinese)

one-hot encoding data and thus we get the corresponding deep encoding data set by representing the original data set with hidden layer output. Finally, we measure the distribution consistency of mixed-attribute data sets based on the corresponding deep encoding data sets by using the maximum mean discrepancy index. On 4 benchmark mixed-attribute data sets, which are Adult、Australian、CRX and German, we compare the measure performances of DE-MMD method with those of the one-hot encoding-based MMD (OE-MMD) method and the binarization-based similarity measure (BSM) method. The experimental results show that the proposed method can measure the distribution consistency of mixed-attribute data sets more accurately and more effectively than OE-MMD and BSM methods.

Key words: artificial intelligence; random sample partition; distribution consistency; maximum mean discrepancy; mixed-attribute data; one-hot encoding; deep encoding

分治^[1]是对大数据进行处理和分析的有效途径之一, 当前流行的大数据计算框架, 如 MapReduce^[2]和 Spark^[3]均是基于分治思想开展对大数据存储和计算的. 最近, 一种新的以分治策略为基础的大数据管理模型——随机样本划分 (random sample partition, RSP)^[4-5]被提出并用于大数据的处理和分析中. 该模型通过将大数据划分成与大数据保持概率分布一致性的 RSP 数据块实现对大数据的学习和挖掘, 其中大数据 RSP 数据块的生成和判定是模型的核心.

RSP 数据块^[6]是指在给定显著性水平下具有相同概率分布的数据块, 包括连续属性、离散属性和混合属性数据块. 用于度量不同样本之间相似性的指标很多, 如欧氏距离、马氏距离、余弦相似度、相关系数和 Jaccard 相似系数等. 然而, 用于度量不同数据集分布相似性或一致性的指标却很少, 最经典的就是用于度量连续属性数据集分布一致性的最大平均差异 (maximum mean discrepancy, MMD) 方法^[7]. MMD 在再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS) 中构建度量标准来检验不同数据集是否来自同一概率分布. 李洪奇等^[8]给出了一种度量离散属性数据集分布相似性方法, 本研究将其记为基于二值化的相似性度量 (binarization-based similarity measure, BSM) 方法. 该方法首先将原始数据集转换成二进制数据集; 之后计算二进制数据集的单项目集和双项目集的特征频率, 并将他们组合, 得到原数据集的特征向量; 最后通过计算不同数据集特征向量之间的距离来度量分布相似性. MMD 方法适用于判定连续属性数据集之间的分布一致性, BSM 方法适用于判定离散属性数据集之间的分布相似性. 经检索, 迄今尚未发现对混合属性数据集进行分布一致性判定的相关研究工作的报道.

在当今大数据时代背景下, 数据表现形式的多

样性^[9-11]是大数据的一个显著特征. 如何将 RSP 模型应用到混合属性大数据的处理与分析中是 RSP 模型研究的重点. 对于混合属性大数据 RSP 数据块的生成而言, 其中最关键的就是数据集分布一致性的度量. 度量混合属性数据集分布一致性的难点主要体现在两个方面: ① 通过对连续属性进行离散化处理, 将混合属性数据集转化为离散属性数据集, 这会导致原始数据集信息量的丢失, 最显著的后果就是伴随着连续属性离散化, 数据之间的序关系会丢失; ② 通过对离散属性进行独热编码处理, 将混合属性数据集转化为连续属性数据集, 本质上并没有起到离散属性连续化的目的, 而仅是将原有的离散属性用更多 0 和 1 的二值的离散属性进行了表示.

本研究提出一种基于深度编码 (deep encoding, DE) 和 MMD 的混合属性数据集分布一致性度量方法, 简称 DE-MMD 方法, 基于该方法实现了对混合属性大数据 RSP 数据块的判定. 该方法的核心是通过自编码神经网络得到混合属性数据集的深度编码表示形式, 再利用混合属性数据集的深度编码表示进行分布一致性的度量. 在 4 个标准混合属性数据集上对 DE-MMD 方法进行的性能测试结果表明, 与基于离散属性独热编码的 MMD 方法和连续属性二进制的 BSM 方法^[8]相比, 新方法能够更加准确地对混合属性数据集进行分布一致性的度量.

1 离散属性的独热编码

混合属性数据集 D 和独热编码数据集结构 $\hat{D}$ 如表 1 和表 2. 设 $D = \{x_n | x_n = (a_{n1}, a_{n2}, \dots, a_{nP}; b_{n1}, b_{n2}, \dots, b_{nQ}), n = 1, 2, \dots, N\}$, 其中, a_{np} 为样本 x_n 对应的第 p 个连续属性的取值, $a_{np} \in \mathbf{R}, p = 1, 2, \dots, P; b_{nq} \in \{b_1^{(q)}, b_2^{(q)}, \dots, b_{M_q}^{(q)}\}$ 为样本 x_n 对应的第 q 个离散属性的取值. 离散属性 B_q 对应的 M_q 个取值分别为 $b_1^{(q)}, b_2^{(q)}, \dots, b_{M_q}^{(q)}, q = 1, 2, \dots, Q; P, Q$ 和

表 1 混合属性数据集

Table 1 Mixed-attribute data set

D	连续属性				离散属性			
	A_1	A_2	$\cdots$	A_P	B_1	B_2	$\cdots$	B_Q
x_1	a_{11}	a_{12}	$\cdots$	a_{1P}	b_{11}	b_{12}	$\cdots$	b_{1Q}
x_2	a_{21}	a_{22}	$\cdots$	a_{2P}	b_{21}	b_{22}	$\cdots$	b_{2Q}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_N	a_{N1}	a_{N2}	$\cdots$	a_{NP}	b_{N1}	b_{N2}	$\cdots$	b_{NQ}

表 2 独热编码数据集

Table 2 One-hot encoding data set

$\dot{D}$	连续属性				二值属性								
	A_1	A_2	$\cdots$	A_p	B_1				$\cdots$	B_Q			
					$B_1^{(1)}$	$B_2^{(1)}$	$\cdots$	$B_{M_1}^{(1)}$		$\cdots$	$B_1^{(Q)}$	$B_2^{(Q)}$	$\cdots$
$\dot{x}_1$	a_{11}	a_{12}	$\cdots$	a_{1P}	$b_{11}^{(1)}$	$b_{12}^{(1)}$	$\cdots$	$b_{1M_1}^{(1)}$	$\cdots$	$b_{11}^{(Q)}$	$b_{12}^{(Q)}$	$\cdots$	$b_{1M_Q}^{(Q)}$
$\dot{x}_2$	a_{21}	a_{22}	$\cdots$	a_{2P}	$b_{21}^{(1)}$	$b_{22}^{(1)}$	$\cdots$	$b_{2M_1}^{(1)}$	$\cdots$	$b_{21}^{(Q)}$	$b_{22}^{(Q)}$	$\cdots$	$b_{2M_Q}^{(Q)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\dot{x}_N$	a_{N1}	a_{N2}	$\cdots$	a_{NP}	$b_{N1}^{(1)}$	$b_{N2}^{(1)}$	$\cdots$	$b_{NM_1}^{(1)}$	$\cdots$	$b_{N1}^{(Q)}$	$b_{N2}^{(Q)}$	$\cdots$	$b_{NM_Q}^{(Q)}$

N 分别为数据集 D 含有连续属性、离散属性和样本的个数. 采用独热编码技术^[12-13]对数据集中的离散属性 $B_1, B_2, \cdots, B_Q$ 进行编码, 可得表 2 的独热编码数据集, 编码方式为

$$b_{nm}^{(q)} = \begin{cases} 1, & b_{nq} = b_m^{(q)} \\ 0, & \text{其他} \end{cases} \quad (1)$$

其中, $n=1, 2, \cdots, N; q=1, 2, \cdots, Q; m=1, 2, \cdots, M_q$.

经过对原始数据集中离散属性的独热编码处理, 离散属性 $B_q, q=1, 2, \cdots, Q$, 被扩展成 M_q 个二值属性 $B_1^{(q)}, B_2^{(q)}, \cdots, B_{M_q}^{(q)}$. 其中, $B_m^{(q)}$ 的取值为 0 或 1; $m=1, 2, \cdots, M_q$.

2 基于自编码神经网络的深度编码

由于原始数据集 D 中含有离散属性, 无法对其

进行概率密度函数估计, 进而不能进行不同数据集间概率分布一致性的度量. 将原始数据集转化成独热编码数据集 $\dot{D}$, 尽管离散 $B_q (q=1, 2, \cdots, Q)$ 被表示成了 0 和 1 的二值属性 $B_m^{(q)} (m=1, 2, \cdots, M_q)$, $B_m^{(q)}$ 仍是特殊的离散属性, 即 $b_{1m}^{(q)}, b_{2m}^{(q)}, \cdots, b_{Nm}^{(q)}$ 是由若干个 0 和若干个 1 构成, 无法较好地获得其概率密度函数估计.

因此, 本研究继续对独 $\dot{D}$ 进行深度编码处理^[14]. 首先, 构建一个单隐含层的全连接前馈神经网络, 隐含层节点数为 L ; 其次, 网络的输入和输出都为式(2)的矩阵 $\dot{D}$, 并称该网络为自编码神经网络; 最后, 采用误差反传算法完成对自编码神经网络的训练, 得到如式(3)的隐含层输出矩阵 H , 即原始数据集 D 对应的深度编码数据集.

$$\dot{D} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_N \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1P} & b_{11}^{(1)} & \cdots & b_{1M_1}^{(1)} & \cdots & b_{11}^{(Q)} & \cdots & b_{1M_Q}^{(Q)} \\ a_{21} & \cdots & a_{2P} & b_{21}^{(1)} & \cdots & b_{2M_1}^{(1)} & \cdots & b_{21}^{(Q)} & \cdots & b_{2M_Q}^{(Q)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NP} & b_{N1}^{(1)} & \cdots & b_{NM_1}^{(1)} & \cdots & b_{N1}^{(Q)} & \cdots & b_{NM_Q}^{(Q)} \end{bmatrix} \quad (2)$$

$$H = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1L} \\ h_{21} & h_{22} & \cdots & h_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{NL} \end{bmatrix} \quad (3)$$

自编码神经网络的隐含层使用 sigmoid 激活函数, 因此可得 $h_{nl} \in (0, 1), n=1, 2, \cdots, N; l=1, 2, \cdots, L$. 至此, 完成了从混合属性数据集 D 向连续属性数据集 H 的转换.

3 基于 MMD 的分布一致性度量

MMD 方法^[7]通过在再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS) 构建统计量, 检验不同数据集是否来自同一概率分布.

设有一个 D 维连续属性数据集 $S = \{s_i | s_i = (s_{i1}, s_{i2}, \dots, s_{iD}) \in \mathbf{R}^D, i = 1, 2, \dots, M\}$ 和 $T = \{t_j | t_j = (t_{j1}, t_{j2}, \dots, t_{jD}) \in \mathbf{R}^D, j = 1, 2, \dots, N\}$, 采用 MMD 方法构建的检验量为

$$\text{MMD}(S, T) = \left[\frac{1}{M^2} \sum_{i=1}^M \sum_{j=1}^M K(s_i, s_j) + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N K(t_i, t_j) - \frac{2}{MN} \sum_{i=1}^M \sum_{j=1}^N K(s_i, t_j) \right]^{\frac{1}{2}} \quad (4)$$

当

$$\text{MMD}(S, T) < \sqrt{\frac{K(M+N)}{MN}} + \varepsilon \quad (5)$$

时, 数据集 S 和 T 具有一致的概率分布. 其中, ε 为一致性阈值, $\varepsilon > 0$; K 为核函数 $K(u, v)$ 的上界,

$$K(u, v) = \exp\left(-\frac{\|u - v\|_2^2}{2\sigma^2}\right) \quad (6)$$

其中, $u = (u_1, u_2, \dots, u_D)$ 和 $v = (v_1, v_2, \dots, v_D)$ 为两个 D 维连续属性向量, σ^2 为核宽度.

利用自编码神经网络获得两个混合属性数据集 D_1 和 D_2 对应的深度编码数据集 H_1 和 H_2 之后, 基于 MMD 方法的 $\text{MMD}(H_1, H_2)$ 判断 D_1 和 D_2 是否具有一致的分布.

4 实验验证与结果分析

基于表 3 的 Adult、Australian、CRX 和 German 标准 KEEL^[15] 混合属性数据集, 对所提的基于深度编码和最大平均差异的混合属性数据集分布一致性度量方法 DE-MMD 的可行性和有效性进行验证.

表 3 四个标准 KEEL 数据集

Table 3 Four KEEL benchmark data sets

数据集	样本数量	连续属性数量	离散属性数量
Adult	45 222	6	8
Australian	690	8	6
CRX	653	6	9
German	1 000	7	13

4.1 DE-MMD 方法的可行性

4.1.1 自编码神经网络对混合属性数据集深度编码稳定性的影响

采用代入熵 (re-substitution entropy, RE)^[16] 衡量混合属性数据集所转换的连续属性数据集所包含的信息量. 对于给定的含有 N 个样本和 D 个连续属性的数据集

$$\mathbf{Z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1D} \\ z_{21} & z_{22} & \cdots & z_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{ND} \end{bmatrix} \quad (7)$$

代入熵为

$$\text{RE}(\mathbf{Z}) = -\frac{1}{N} \sum_{n=1}^N \ln[\tilde{f}(z_n)] \quad (8)$$

其中, $\tilde{f}(z)$ 为利用 Parzen 窗口法估计数据集 $\mathbf{Z}$ 的概率密度函数, $z = (z_1, z_2, \dots, z_D)$ 为 D 维自变量, 即

$$\tilde{f}(z) = \tilde{f}(z_1, z_2, \dots, z_D) = \frac{1}{N} \sum_{n=1}^N \prod_{d=1}^D \frac{1}{\sqrt{2\pi}h_d} \exp\left[-\frac{1}{2}\left(\frac{z_d - z_{nd}}{h_d}\right)^2\right] \quad (9)$$

这里, h_d 是窗口宽度参数, 且有

$$h_d = \frac{1}{\sqrt{N}}, \quad d = 1, 2, \dots, D \quad (10)$$

对每个数据集随机抽取 200 个样本, 然后利用自编码神经网络 (权重最多更新 2 000 次, 终止阈值为 1×10^{-4}) 进行属性转换, 再计算转换后数据集的代入熵, 重复该过程 500 次, 并统计 500 个代入熵的均值 m 和标准差 s . 图 1 给出了针对 4 个混合属性数据集的代入熵计算结果. 由图 1 可见, 自编码神经网络能够稳定地对混合属性数据集进行深度编码, 因为对于每一个数据集, 500 次转换的代入熵 RE 值基本上位于区间 $[m - 3s, m + 3s]$ 内 (Adult 数据集有 498 次、Australian 数据集有 500 次、CRX 数据集有 499 次、German 数据集有 499 次). 该实验证实了利用自编码神经网络对混合属性数据集进行深度编码是可行的, 它基本能够保证转换后所得连续属性数据集所含信息量的稳定.

4.1.2 DE-MMD 方法对非 RSP 和 RSP 数据块分布一致性的判定

当利用 DE-MMD 方法判定不同数据集 D_1 和 D_2 的概率分布一致性时, 从式 (5) 可发现阈值 ε 会对判定结果产生影响. 由 GRETTON 等^[7] 的推导可知, 当 $M = N$ 时,

$$\varepsilon = 2 \sqrt{K \ln \alpha^{-1}/N} \quad (11)$$

图 1 自编码神经网络对混合属性数据集深度编码稳定性的影响

Fig. 1 The impact of autocoder neural network on transformation of mixed-attribute data set

其中, $K = 1$ 为式(6)所示的核函数上界; α 为显著性水平. 图 2 给出了 $N = \{100, 200, \dots, 10\,000\}$ 和 $\alpha = \{0.001, 0.002, \dots, 0.100\}$ 时, 对应的 ε 取值情况.

由图 2 可见, 较小的样本数量和显著性水平对

图 2 阈值 ε 与样本个数和显著性水平 α 之间的关系

Fig. 2 The relationship among threshold, size of data set, and significance level

应较大的阈值, 而较大的样本数量和显著性水平往往对应较小的阈值. 在 MMD 方法的分布一致性判别式 (式(5)) 中, 发现阈值实际上仅与数据集的规模相关, 而与数据集的具体性质无关. 本研究给出一种基于“数据驱动”的方式来确定阈值.

首先, 将式(5)修改为

$$\text{MMD}(S, T) < \varepsilon' \quad (12)$$

其中, ε' 为修正的一致性阈值, 且 $\varepsilon' > 0$. 确定阈值 ε' 的具体流程为:

1) 确定数据集 S 的 RSP 数据块 S_{i1}, S_{i2} 和数据集 T 的 RSP 数据块 T_{i1}, T_{i2} . 其中,

$$|S_{i1}| = |S_{i2}| = \frac{1}{2} |S| \quad (13)$$

$$|T_{i1}| = |T_{i2}| = \frac{1}{2} |T| \quad (14)$$

RSP 数据块的生成方法见文献[6, 17], 在此不再赘述.

2) 分别计算 $\text{MMD}(S_{i1}, S_{i2})$ 和 $\text{MMD}(T_{i1}, T_{i2})$ 的值.

3) 重复步骤 1) 和步骤 2) I 次, 并令

$$\varepsilon' = \frac{1}{2} \left[\frac{\sum_{i=1}^I \text{MMD}(S_{i1}, S_{i2})}{I} \right] + \left[\frac{\sum_{i=1}^I \text{MMD}(T_{i1}, T_{i2})}{I} \right] \quad (15)$$

针对表3中的每一个数据集分别抽取2个非RSP数据块和2个RSP数据块,使用DE-MMD方法分别度量非RSP数据块和RSP数据块之间的分布一致性.其中,两个数据块的规模均为200,自编码神经网络的权重最大更新次数为 5×10^4 ,权重更新终止阈值为 1×10^{-6} ,一致性阈值计算公式执行 $I = 10$ 次,核宽度 $\sigma^2 = 9$.

图3至图6分别显示了在4个混合属性数据集上DE-MMD方法对非RSP和RSP数据块分布一致性的判定情况.

图3 在Adult数据集上DE-MMD方法对非RSP和RSP数据块分布一致性的判定 ($\epsilon' = 0.0131$)

Fig. 3 The distribution consistency determined by DE-MMD method on non-RSP and RSP data blocks of Adult data set ($\epsilon' = 0.0131$)

由图3至图6可见,对于非RSP数据块,其选定的离散属性和连续属性在分布上差异非常明显;而对于RSP数据块,其选定的离散属性和连续属性在分布上基本保持了一致.对于非RSP数据块,MMD值不满足式(12),即不同数据块的MMD值大

图4 在Australian数据集上DE-MMD方法对非RSP和RSP数据块分布一致性的判定 ($\epsilon' = 0.1098$)

Fig. 4 The distribution consistency with determined by DE-MMD method on non-RSP and RSP data blocks of Australian data set ($\epsilon' = 0.1098$)

于阈值;而对于RSP数据块,MMD值满足了式(12),即不同数据块的MMD值小于阈值.实验结果表明,DE-MMD方法能够量化不同混合属性数据集的概率分布一致情况并做出判定.

4.2 DE-MMD方法的有效性

4.2.1 基于离散属性独热编码的OE-MMD方法和离散属性二进制化的BSM方法

OE-MMD方法是基于表2的独热编码数据集,使用MMD方法对不同混合属性数据集的分布一致性进行检验.对于OE-MMD方法,同样使用式(12)判定数据块的分布一致性,阈值按照4.1.2节设计方法确定.

基于连续属性二进制化的BSM方法^[8]首先对连续属性进行离散化处理,将混合属性数据集转换成离散属性数据集,为简便起见,在本实验对连续

图 5 在 CRX 数据集上 DE-MMD 方法对非 RSP 和 RSP 数据块分布一致性的判定 ($\epsilon' = 0.0397$)

Fig. 5 The distribution consistency with determined by DE-MMD method on non-RSP and RSP data blocks of CRX data set ($\epsilon' = 0.0397$)

值属性采用二值离散化;之后对离散属性数据集进行独热编码,得到独热编码数据集;然后计算独热编码数据集对应的单项目集和双项目集的特征频率向量;最后对单项目集和双项目集的特征频率向量进行归一化处理,得到混合属性数据集对应的特征向量。

BSM 方法通过计算特征向量之间的距离判断混合属性数据集之间的相似性:距离越小,数据集相似性越高;距离越大,数据集相似性越低。假设现有两个混合属性数据集对应的特征向量 $w_1 = (w_{1,1}, w_{1,2}, \dots, w_{1,18})$ 和 $w_2 = (w_{2,1}, w_{2,2}, \dots, w_{2,18})$,当

$$\sqrt{\frac{1}{18} \sum_{d=1}^{18} (w_{1,d} - w_{2,d})^2} < \delta \quad (16)$$

时,表明两个混合属性数据集具有相似的概率分

图 6 在 German 数据集上 DE-MMD 方法对非 RSP 和 RSP 数据块分布一致性的判定 ($\epsilon' = 0.0381$)

Fig. 6 The distribution consistency with determined by DE-MMD method on non-RSP and RSP data blocks of German data set ($\epsilon' = 0.0381$)

布。其中, δ 为相似性阈值,且 $\delta > 0$ 。

4.2.2 DE-MMD 方法与 OE-MMD 方法和 BSM 方法的比较

针对表 3 的每一个标准混合属性数据集,对其 RSP 数据块分别使用 OE-MMD、BSM 和 DE-MMD 方法进行概率分布一致性的判定。其中,自编码神经网络的权重最大更新次数为 5×10^4 ,权重更新终止阈值为 1×10^{-6} , $I = 10$,核宽度 $\sigma^2 = 9$ 。针对每个数据集,选取 3 种不同规模(200、300 和 400 个)的 RSP 数据块。对每种规模的数据块分别进行 100 次的分布一致性判定,并统计判定的准确率、阈值参数的选取,对比结果如表 4—表 7 和图 7 所示。

RSP 数据块具有一致的概率分布^[5-6],从对比结果可以发现,DE-MMD 方法在 4 个混合分布数据集上均获得了优于 OE-MMD 和 BSM 方法的一致性

判别准确率, 证明本研究提出的 DE-MMD 方法是有效的. 在此, 本研究尝试对 DE-MMD 方法取得优势的原因进行讨论: ① 与基于离散属性独热编码的 OE-MMD 方法相比, DE-MMD 方法并没有直接使用 0 和 1 二值化的离散属性进行分布一致性度量, 因为直接使用 0 和 1 的二值化离散属性在计算不同数据集的 MMD 值时容易增加式(4)失效的概率. 举一个最极端情况的例子: 假设有两组数据(1, 0)和(0, 1)以及(2, 2)和(1, 1), 经计算可发现这两组

数据对应的 MMD 值相同, 但是相比数据 (1, 0) 和 (0, 1) 之间的分布一致性, (2, 2)和(1, 1)应该具有更大的分布一致性. ② 与基于连续属性二进制化的 BSM 方法相比, DE-MMD 方法是通过确定原始数据集的一种深度编码形式来计算数据集之间的分布一致性, 这种深度编码通过输入和输出完全相同的神经网络将数据转化过程中的信息丢失率 达到最小化, 从而保证了基于深度编码的分布一致性能够反映原始数据的分布一致性.

表 4 OE-MMD、BSM 和 DE-MMD 在 Australian 数据集对应的 RSP 数据块上的对比

Table 4 The comparative results of OE-MMD, BSM, and DE-MMD on RSP data blocks of Australian data set

方法	200 个 RSP 数据块			300 个 RSP 数据块			400 个 RSP 数据块		
	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围
OE-MMD	0.91	0.067 3	[0.046 5, 0.080 7]	0.95	0.055 8	[0.050 5, 0.064 5]	0.97	0.043 9	[0.036 2, 0.047 3]
BSM	0.88	0.080 0	[0.080 0, 0.080 0]	0.84	0.070 0	[0.070 0, 0.070 0]	0.96	0.060 0	[0.060 0, 0.060 0]
DE-MMD	1.00	0.114 7	[0.098 9, 0.131 7]	1.00	0.094 2	[0.084 9, 0.101 9]	1.00	0.086 3	[0.074 1, 0.100 2]

表 5 OE-MMD、BSM 和 DE-MMD 在 Adult 数据集对应的 RSP 数据块上的对比

Table 5 The comparative results of OE-MMD, BSM, and DE-MMD on RSP data blocks of Adult data set

方法	200 个 RSP 数据块			300 个 RSP 数据块			400 个 RSP 数据块		
	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围
OE-MMD	0.94	0.082 5	[0.071 1, 0.107 3]	0.95	0.079 2	[0.066 8, 0.097 3]	0.97	0.058 4	[0.052 0, 0.063 7]
BSM	0.85	0.050 0	[0.050 0, 0.050 0]	0.93	0.050 0	[0.050 0, 0.050 0]	0.92	0.050 0	[0.050 0, 0.050 0]
DE-MMD	1.00	0.015 9	[0.014 8, 0.017 7]	1.00	0.013 7	[0.012 4, 0.016 0]	1.00	0.011 7	[0.010 7, 0.012 5]

表 6 OE-MMD、BSM 和 DE-MMD 在 CRX 数据集对应的 RSP 数据块上的对比

Table 6 The comparative results of OE-MMD, BSM, and DE-MMD on RSP data blocks of CRX data set

方法	200 个 RSP 数据块			300 个 RSP 数据块			400 个 RSP 数据块		
	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围
OE-MMD	0.95	0.087 8	[0.077 8, 0.098 3]	0.97	0.070 2	[0.062 2, 0.079 2]	0.98	0.057 6	[0.053 0, 0.059 7]
BSM	0.87	0.080 0	[0.080 0, 0.080 0]	0.95	0.070 0	[0.070 0, 0.070 0]	0.97	0.060 0	[0.060 0, 0.060 0]
DE-MMD	1.00	0.043 6	[0.030 7, 0.073 7]	1.00	0.029 6	[0.028 1, 0.032 4]	1.00	0.025 3	[0.014 7, 0.043 3]

表 7 OE-MMD、BSM 和 DE-MMD 方法在 German 数据集对应的 RSP 数据块上的对比

Table 7 The comparative results of OE-MMD, BSM, and DE-MMD on RSP data blocks of German data set

方法	200 个 RSP 数据块			300 个 RSP 数据块			400 个 RSP 数据块		
	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围	准确率	阈值均值	阈值范围
OE-MMD	0.96	0.106 7	[0.103 4, 0.110 1]	0.97	0.081 7	[0.079 8, 0.086 6]	0.99	0.078 8	[0.067 9, 0.079 7]
BSM	0.91	0.080 0	[0.080 0, 0.080 0]	0.91	0.070 0	[0.070 0, 0.070 0]	0.96	0.060 0	[0.060 0, 0.060 0]
DE-MMD	1.00	0.036 6	[0.033 7, 0.039 1]	1.00	0.025 6	[0.023 4, 0.027 9]	1.00	0.024 3	[0.022 8, 0.026 1]

图 7 OE-MMD、BSM 和 DE-MMD 在 4 个 KEEL 数据集上的对比结果

Fig. 7 The comparison of results of OE-MMD, BSM, and DE-MMD on 4 KEEL data sets

结 语

提出一种新的基于深度编码和最大平均差异的混合属性数据集分布一致性度量方法 DE-MMD, 它能够对混合属性数据集的分布一致性进行有效度量, 其表现优于基于离散属性独热编码的 OE-MMD 方法和基于连续属性二进制的 BSM 方法. 不同

于 OE-MMD 方法和现有的 BSM 方法, DE-MMD 方法并未直接使用 0 和 1 的二值离散属性表示连续属性也没有对连续属性进行离散化处理, 因此不仅没有增加分布一致性判定过程中的不确定性, 且最大限度地保留了原始数据集属性转化中的信息量. 通过对自编码神经网络的使用, DE-MMD 将原始数据集转换成可靠的深度编码表示形式, 这对增强数据集分布一致性判定方法的稳定性起着至关重要的作用.

下一步研究计划结合不确定性理论深入分析 DE-MMD 方法的优势, 尝试采用 JS 散度替换 MMD 进行分布一致性的度量, 并考虑将 DE-MMD 方法应用到混合属性大数据的随机样本划分生成方法中, 同时考虑使用极限学习机(extreme learning machine, ELM)^[18-19] 替换本研究的自编码神经网络以提升 DE-MMD 方法的判别速率.

致谢: 衷心感谢张晓亮博士对本文数学公式的推导及检查.

基金项目: 公安部重点实验室开放基金资助项目 (HJKF 201901); 公安部科技强警基础工作专项资助项目 (2017GABJC09); 深圳大学新引进教师科研启动资助项目 (2018060); 大学生创新创业训练计划资助项目 (201910590017)

作者简介: 何玉林 (1982—), 深圳大学副研究员、博士. 研究方向: 大数据系统计算技术与应用. E-mail: yulinhe@szu.edu.cn

引 文: 何玉林, 金 一, 戴德鑫, 等. 混合属性数据分布一致性度量的新方法[J]. 深圳大学学报理工版, 2021, 38(2): 170-179.

参考文献 / References:

- [1] 何 清, 李 宁, 罗文娟, 等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4): 327-336.
HE Qing, LI Ning, LUO Wenjuan, et al. A survey of machine learning algorithms for big data [J]. Pattern Recognition and Artificial Intelligence, 2014, 27 (4): 327-336. (in Chinese)
- [2] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.
- [3] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: cluster computing with working sets [C] // Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. Berkeley, USA: USENIX Association, 2010: 10.
- [4] 魏丞昊, 黄哲学, 何玉林. 基于统计感知的大数据系统计算框架[J]. 深圳大学学报理工版, 2018, 35