

基于概率矩阵分解的不完整数据集特征选择方法

范林歌, 武欣嵘, 童 玮, 曾维军

(中国人民解放军陆军工程大学 通信工程学院, 南京 210007)

摘要: 在机器学习理论与应用中, 特征选择是降低高维数据特征维度的常用方法之一。传统的特征选择方法多数基于完整数据集, 对实际应用中普遍存在缺失数据的情形研究较少。针对不完整数据中含有未被观察信息和存在异常值的特点, 提出一种基于概率矩阵分解技术的鲁棒特征选择方法。使用基于分簇的概率矩阵分解模型对数据集中的缺失值进行近似估计, 以有效测量相邻簇之间数据的相似性, 缩小问题规模, 同时降低填充误差。依据缺失数据值存在少量异常值的情形, 利用基于 $\ell_{2,1}$ 损失函数的方法进行特征选择, 在此基础上给出不完整数据集的特征选择方法流程, 并对其收敛性进行理论分析。该方法利用不完整数据集中的所有信息, 有效应对不完整数据集中异常值带来的影响。实验结果表明, 相比传统特征选择方法, 该方法在合成数据集上选择更少的无关特征, 可降低异常值带来的影响, 在真实数据集上获得了较高的分类准确率, 能够选择出更为准确的特征。

关键词: 矩阵分解; 缺失值填补; 鲁棒特征选择; 不完整数据; $\ell_{2,1}$ 范数

开放科学(资源服务)标志码(OSID):



中文引用格式: 范林歌, 武欣嵘, 童玮, 等. 基于概率矩阵分解的不完整数据集特征选择方法[J]. 计算机工程, 2022, 48(6): 57-64.

英文引用格式: FAN L G, WU X R, TONG W, et al. Feature selection method for incomplete data sets based on probability matrix decomposition[J]. Computer Engineering, 2022, 48(6): 57-64.

Feature Selection Method for Incomplete Data Sets Based on Probability Matrix Decomposition

FAN Linge, WU Xinrong, TONG Wei, ZENG Weijun

(College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China)

[Abstract] In machine learning theory and application, feature selection is one of the common methods of reducing the feature dimension of high-dimensional data. Traditional feature selection methods are mostly based on complete data sets, and a few studies have been conducted on missing data in practical applications. In this study, a robust feature selection method is proposed based on Probability Matrix Decomposition (PMF) for incomplete data containing unobserved information and outliers. First, a probabilistic matrix decomposition model, based on clustering, is used to approximate the missing values in the data set. The model can effectively measure data similarity between adjacent clusters, reduce the scale of the problem, and reduce the imputation error. Secondly, the feature selection method, based on loss function, is used in the case involving missing data values with a few outliers. Finally, the flow of feature selection method for incomplete data sets is constructed, and its convergence is theoretically analyzed. The method proposed in this study utilizes all the information in incomplete data sets and effectively deals with the influence of outliers in incomplete data sets. Experimental results show that when compared with traditional feature selection methods, the proposed method can select fewer irrelevant features in the synthetic data set and reduce the influence of outliers. Conversely, on real data sets, the proposed method realizes higher classification accuracy and selects more accurate features.

[Key words] matrix decomposition; missing value filling; Robust Feature Selection (RFS); incomplete data; $\ell_{2,1}$ norm

DOI: 10.19678/j.issn.1000-3428.0061524

0 概述

高维数据在实际应用中普遍存在, 这些数据通

常被用来构建高质量的机器学习模型, 随后进行数据挖掘分析, 但高维数据具有计算时间长、存储成本高等问题^[1-2]。为了解决这些问题, 研究人员提出特

基金项目: 国家自然科学基金(61802425)。

作者简介: 范林歌(1997—), 女, 硕士研究生, 主研方向为机器学习、数据质量; 武欣嵘、童 玮, 副教授、硕士; 曾维军(通信作者), 讲师、博士。

收稿日期: 2021-04-30 **修回日期:** 2021-07-08 **E-mail:** zwj3103@126.com

征选择、主成分分析等降维方法^[3-4]。特征选择是数据挖掘过程中的重要步骤,特别是在许多生物信息学任务中,有效的鲁棒特征选择方法可以提取有意义的特征并且消除有噪声的特征。

现有的特征选择方法大多以数据是完整的或几乎完整的为前提。但在许多实际应用中,例如生物信息学^[5]和遥感网络普遍存在数据缺失,现有的特征选择方法大部分不能直接用于包含缺失值的数据集。

对不完整数据集进行特征选择,最直接的方法是对不完整数据集进行一定处理使其保持形式上的完整,然后再做特征选择。处理缺失数据最简单易行的方法是完整样本分析(Complete Case Analysis, CCA),即删除包含缺失值的样本或特征^[6]。该方法虽然简单易行,但在缺失样本比例较高和样本总量较小时有很大的局限性,进而影响后续机器学习的准确性。对缺失位置进行填充是另一种较为直接的缺失数据处理方法,应用普遍的是均值填充。均值填充是指以特征观测值的均值作为缺失值的估计值^[7],但该方法会使特征不确定性或方差减小。还有一种基于统计学的缺失填补方法是期望最大化(EM)填补法^[8],该方法利用现有数据的边缘分布对缺失数据进行极大似然估计,从而得到相应的填补值。ZHANG等^[9]提出了k近邻填充,该方法基于近邻样本的特征值相近的假设,以近邻样本的均值作为缺失值的估计值。之后一些改进的k近邻填补方法被提出,其引入灰色关联分析和互信息来更进一步地提高分类精度和填补效果^[10-13]。近邻填充的关键在于准确地度量近邻关系,这也带来了一些不足:不同的特征在相似度度量中的重要性是不同的,而k近邻填充中的距离计算将所有特征同等对待;当数据特征数目较多即在高维特征空间中时,样本分布趋于均匀,此时距离并不能反映样本相似性。

在不填充直接进行特征研究方面,LOU等^[14]提出基于类间隔的不完整数据特征选择(Feature Selection in Incomplete Data, SID)方法,定义了基于类间隔的目标函数,对其优化使每个样本在其相关子空间中的类间隔最大化,以权重范数比为系数降低缺失样本对于类间隔计算的贡献,该方法不再将某个样本的近邻样本当作是固定的,而是去计算所有样本是某个样本近邻样本的概率。SID特征选择最终学习出一个特征权重向量 w , w 中取值较大的分量对应重要特征,取值较小的分量对应次要特征,取值为零的分量对应无关特征甚至噪音特征。与基于常用预处理填充方法(如均值填充、EM填充、k近邻填充)的特征选择相比,SID能够筛除更多无关特征,以该算法选择的特征建立的分类模型分类准确率更高,但其没有考虑异常值的影响。

本文针对不完整数据中含有未被观察信息和存在异常值的特点,提出一种基于概率矩阵分解

(Probability Matrix Decomposition, PMF)技术的鲁棒特征选择方法,使用指示矩阵并利用不完整数据集集中的所有信息进行缺失值近似估计。在特征选择算法上基于鲁棒特征选择(Robust Feature Selection, RFS)方法,在损失函数和正则化项中联合使用 $\ell_{2,1}$ 范数^[15-16],将 $\ell_{2,1}$ 范数作为损失函数可减轻异常值影响,提升特征选择的鲁棒性。

1 基于矩阵分解的缺失值填充方法

考虑不完整数据集 $X = \{(x_n, y_n) | n=1, 2, \dots, N\} \in \mathbb{R}^M$,其中 N 为样本数, M 为特征数, x_n 为第 n 个给定实例, y_n 为其对应的标签值, $y_n=1, 2, \dots, T$, $x_n(j)$ 表示第 n 个样本的第 j 个特征,其中 $j=1, 2, \dots, M$ 。

由于直接使用不完整数据集中所有实例来预估缺失值计算量较大,本文首先使用原始标签将原始数据集分簇,将相似度较高的样本分为一簇,假设第 i 簇数据集为 X^i ,每一簇数据集实例选择如式(1)所示:

$$X^i = \{x_n | y_n = i\}, n=1, 2, \dots, N \quad (1)$$

其包含 L 个实例, X^i 具体表示如式(2)所示:

$$X^i = \begin{bmatrix} x_1^i(1) & \dots & x_1^i(j) \\ \vdots & & \vdots \\ x_q^i(1) & \dots & x_q^i(j) \end{bmatrix} \quad (2)$$

其中: $q=1, 2, \dots, L$, $x_q^i(j)$ 表示第 i 簇数据集中第 q 个实例的第 j 个特征。

本文基于这样一个假设:同一簇内两个样本相似度比不同簇的两个样本高,以达到在减少计算量的基础上不降低填充准确性的目的,之后分别对每簇中的缺失值进行估计。

在计算过程中利用指示矩阵对未观察到的信息进行过滤,从而达到利用完整和不完整样本中所有有效信息的目的。对于给定的不完整数据集 X ,定义指示矩阵 I ,其中 $I_n(j)$ 反映第 n 个实例的第 j 个特征的缺失情况,元素取值如式(3)所示,当 $x_n(j)$ 可观测时对应位置取1,当 $x_n(j)$ 缺失时对应位置取0。

$$I_n(j) = \begin{cases} 1, & \text{当 } x_n(j) \text{ 可观测时} \\ 0, & \text{当 } x_n(j) \text{ 缺失时} \end{cases} \quad (3)$$

则第 i 簇数据集 X^i 对应的指示矩阵为 I^i 。

1.1 矩阵分解

寻找第 i 簇数据集 X^i 的一个近似矩阵 \hat{X}^i ^[17],通过求解近似矩阵来代替原始数据簇中缺失的值, \hat{X}^i 定义如式(4)所示:

$$\hat{X}^i = \begin{bmatrix} \hat{x}_1^i(1) & \dots & \hat{x}_1^i(j) \\ \vdots & & \vdots \\ \hat{x}_q^i(1) & \dots & \hat{x}_q^i(j) \end{bmatrix} \quad (4)$$

将近似矩阵 \hat{X}^i 分解为 U 和 V 矩阵,使得:

$$\hat{\mathbf{X}}^i = \mathbf{U} \cdot \mathbf{V}^T = \begin{bmatrix} \hat{x}_1^i(1) & \cdots & \hat{x}_1^i(j) \\ \vdots & & \vdots \\ \hat{x}_q^i(1) & \cdots & \hat{x}_q^i(j) \end{bmatrix} \quad (5)$$

其中: $\mathbf{U} \in \mathbb{R}^{L \times K}$; $\mathbf{V} \in \mathbb{R}^{M \times K}$, 矩阵 \mathbf{U} 和 \mathbf{V} 分别表示数据集实例和特征特定的潜在特征向量。比如在一个用户对多部电影评分的数据集中, $x_q(j)$ 表示第 q 个用户对第 j 部电影的评分, 此时 \mathbf{U} 和 \mathbf{V} 分别描述用户的特征(比如年龄段等)和电影的特征(比如年份、中英文等)。

均方根误差(RMSE)可以用来衡量真实值与预估值之前的差距, 本文通过计算目标矩阵 \mathbf{X}^i 与近似矩阵 $\hat{\mathbf{X}}^i$ 的 RMSE 来评估模型性能, \mathbf{X}^i 与 $\hat{\mathbf{X}}^i$ 的 RMSE 定义如式(6)所示:

$$R_{\text{RMSE}} = \sum_{q=1}^L \sum_{j=1}^M I_q^i(j) (x_q^i(j) - \hat{x}_q^i(j))^2 \quad (6)$$

由式(5)得:

$$\hat{x}_q^i(j) = \mathbf{U}_q \mathbf{V}_j^T \quad (7)$$

代入式(6)得:

$$R_{\text{RMSE}} = \sum_{q=1}^L \sum_{j=1}^M I_q^i(j) (x_q^i(j) - \mathbf{U}_q \mathbf{V}_j^T)^2 \quad (8)$$

RMSE 值越小则表示填补效果越好。本文通过求解一组 \mathbf{U} 和 \mathbf{V} 使相似矩阵 $\hat{\mathbf{X}}^i$ 与目标矩阵 \mathbf{X}^i 的 RMSE 最小, 至此该优化问题可以表示为:

$$\min \sum_{q=1}^L \sum_{j=1}^M I_q^i(j) (x_q^i(j) - \mathbf{U}_q \mathbf{V}_j^T)^2 \quad (9)$$

至此, 将上述问题转化为一个无约束优化问题, 本文采用梯度下降法^[18]迭代计算 \mathbf{U} 和 \mathbf{V} , 首先固定 \mathbf{V} , 对 \mathbf{U} 求导, 如式(10)所示:

$$\frac{\partial R}{\partial \mathbf{U}_q} = \sum_{j=1}^M I_q^i(j) (-2x_q^i(j) \mathbf{V}_j^T + 2\mathbf{U}_q \mathbf{V}_j^T) \quad (10)$$

更新 \mathbf{U}_q , 如式(11)所示:

$$\mathbf{U}_q^{(t+1)} = \mathbf{U}_q^{(t)} + \alpha \left[\sum_{j=1}^M I_q^i(j) (-2x_q^i(j) \mathbf{V}_j^T + 2\mathbf{U}_q \mathbf{V}_j^T) \right] \quad (11)$$

其中: α 为更新速率, 表示迭代的步长。

固定 \mathbf{U} , 对 \mathbf{V} 求导得:

$$\frac{\partial R}{\partial \mathbf{V}_j} = \sum_{q=1}^L I_q^i(j) (-2x_q^i(j) \mathbf{U}_q + 2\mathbf{U}_q \mathbf{V}_j^T) \quad (12)$$

更新 \mathbf{V}_j , 如式(13)所示:

$$\mathbf{V}_j^{(t+1)} = \mathbf{V}_j^{(t)} + \alpha \left[\sum_{q=1}^L I_q^i(j) (-2x_q^i(j) \mathbf{U}_q + 2\mathbf{U}_q \mathbf{V}_j^T) \right] \quad (13)$$

重复式(11)与式(13), 迭代优化 \mathbf{U} 和 \mathbf{V} , 直到 $\text{RMSE} < \zeta$, ζ 为自定义误差。

遍历所有 i , 对每一簇数据进行如上操作, 直到 I_o 等于 0 即数据集无缺失为止。至此, 求出 \mathbf{X}_i 的近似矩阵 $\hat{\mathbf{X}}^i$, 用 $\hat{\mathbf{X}}^i$ 中的数据填充 \mathbf{X}_i 中对应位置的缺失值。 I_o 定义如式(14)所示:

$$I_o = \sum_{n=1}^N \sum_{j=1}^M I_n(j) \quad (14)$$

1.2 概率矩阵分解

为提升算法精度以及降低算法运算复杂度, 本文将上述的矩阵分解扩展为概率矩阵分解。假设数据集第 i 簇数据集 \mathbf{X}^i , 归一化后服从均值为 $\mu=0$, 方差为 σ^2 的高斯分布。因此, 可定义 \mathbf{U} 和 \mathbf{V} 为一个多变量的零均值球面高斯分布, 如式(15)、式(16)所示:

$$p(\mathbf{U} | \sigma_u^2) = \prod_{n=1}^N N(\mathbf{U}_n | 0, \sigma_u^2 \mathbf{I}) \quad (15)$$

$$p(\mathbf{V} | \sigma_v^2) = \prod_{j=1}^M N(\mathbf{V}_j | 0, \sigma_v^2 \mathbf{I}) \quad (16)$$

该先验概率可使式(11)和式(13)最终的收敛值不会远离 0, 从而避免了矩阵 \mathbf{U} 和 \mathbf{V} 出现幅度较大的值。若无该先验知识, 1.1 节中的矩阵概率分解迭代运算次数将增加, 运算复杂度提升。

综合上述两个先验概率, 可将数据集 \mathbf{X} 上的取值条件概率分布定义为:

$$p(\mathbf{X} | \mathbf{U}, \mathbf{V}, \sigma^2) = \prod_{n=1}^N \prod_{j=1}^M \left[N(X_{nj} | \mathbf{U}_n^T \mathbf{V}_j, \sigma^2) \right]^{I_{nj}} \quad (17)$$

其中: $N(X_{nj} | \mathbf{U}_n^T \mathbf{V}_j, \sigma^2)$ 是均值为 μ 的高斯分布的概率密度函数; 方差 σ^2 , I_{nj} 为式(3)定义的指示矩阵中的元素。

基于式(15)、式(16)、式(17)的概率密度函数, 利用经典的后验概率推导可以得到 $p(\mathbf{U}, \mathbf{V} | \mathbf{X}) = p(\mathbf{X} | \mathbf{U}, \mathbf{V}) p(\mathbf{U}) p(\mathbf{V})$, 最大化该后验概率, 就可以通过已有的近似矩阵估计出系统参数 \mathbf{U} 和 \mathbf{V} 。

为了计算方便, 对后验概率取对数, 在观测噪声方差和先验方差保持固定的情况下, 最大化该对数后验分布等价于使用二次正则化项使误差平方和最小。最后, 为了限制数据集中数据的取值范围, 对高斯函数的均值施加 logistic 函数 $g(x) = 1/(1 + \exp(-x))$, 其取值在 (0, 1) 之间。最终的能量函数为:

$$E = \frac{1}{2} \sum_{n=1}^N \sum_{j=1}^M I_{nj} (X_{nj} - \mathbf{U}_n^T \mathbf{V}_j)^2 \quad (18)$$

至此, 可以使用梯度下降方法, 求解 \mathbf{U}_i 、 \mathbf{V}_j 中的每一个元素。

2 鲁棒特征选择算法

2.1 鲁棒特征选择

为增强算法面对异常值时的鲁棒性, 本文采用基于 $\ell_{2.1}$ 范数的损失函数来去除异常值, 而不使用对异常值敏感的 ℓ_2 范数损失函数。以最小二乘回归为例, 使用 $\ell_{2.1}$ 鲁棒损失函数后目标函数如式(19)所示:

$$\min_{\mathbf{W}} \frac{1}{\gamma} \|\mathbf{X}^T \mathbf{W} - \mathbf{Y}\|_{2.1} + \|\mathbf{W}\|_{2.1} \quad (19)$$

其中: γ 为正则化项参数; $\mathbf{W} \in \mathbb{R}^{M \times C}$ 。

将上述目标函数变为带约束的形式, 如式(20)所示:

$$\min_{W,E} \|E\|_{2,1} + \|W\|_{2,1} \text{ s.t. } X^T W + \gamma E = Y \quad (20)$$

将上述的问题改写为:

$$\min_{W,E} \left\| \begin{bmatrix} W \\ E \end{bmatrix} \right\|_{2,1} \text{ s.t. } \begin{bmatrix} X^T & \gamma H \end{bmatrix} \begin{bmatrix} W \\ E \end{bmatrix} = Y \quad (21)$$

使 $A = [X^T \ \gamma H]$, $B = \begin{bmatrix} W \\ E \end{bmatrix}$, 其中 $H \in \mathbb{R}^{N \times N}$ 为单位

矩阵, 则目标函数进一步改写成式(22):

$$\min_B \|B\|_{2,1} \text{ s.t. } AB = Y \quad (22)$$

RFS 使用一个非常简单的方法来求解式(22), 同时这种方法也很容易应用到其他一般的 $\ell_{2,1}$ 范数最小化问题中。

该算法主要基于拉格朗日方法, 构造拉格朗日函数如下:

$$L(B) = \|B\|_{2,1} - \text{Tr}(A^T(AB - Y)) \quad (23)$$

求上式对 B 的偏导并令其为 0, 得到式(24):

$$\frac{\partial L(B)}{\partial B} = 2DB - A^T A = 0 \quad (24)$$

其中: D 是对角矩阵。

第 n 个元素为:

$$d_n(n) = \frac{1}{2\|b_n\|_2} \quad (25)$$

在式(24)两边同乘 AD^{-1} , 并有 $AB = Y$, 得式(26):

$$2AB - AD^{-1}A^T A = 0 \Rightarrow 2Y - AD^{-1}A^T A = 0 \Rightarrow A = 2(AD^{-1}A^T)^{-1}Y \quad (26)$$

将式(26)代入式(24), 得:

$$B = D^{-1}A^T(AD^{-1}A^T)^{-1}Y \quad (27)$$

由于式(22)中的问题是一个凸问题, 当且仅当满足式(27)时, B 是该问题的全局最优解。其中, D 依赖于 B , 因此也是一个未知变量。依然使用迭代算法来获得满足式(27)的解 B : 在每次迭代中, 用当前的 D 计算 B , 然后根据当前计算的 B 更新 D 。不断重复迭代过程, 直到算法收敛。具体描述如算法 1 所示。

算法 1 基于概率矩阵分解的不完整数据集特征选择算法

输入 不完整数据集 $X \in \mathbb{R}^{N \times M}$, $Y \in \mathbb{R}^{N \times C}$

输出 $B \in \mathbb{R}^{M \times C}$

初始化 U 和 V , 初始化 D 为单位阵

1. 根据式(1)将原始数据集分簇, 读取第 i 簇数据。

重复

2. 根据式(11)更新 U 。

3. 根据式(13)更新 V 。

直到 $\text{RMSE} < \zeta$

4. 遍历所有 i , 直到 $I_0 = 0$ 。

重复

5. 计算 $B = D^{-1}A^T(AD^{-1}A^T)^{-1}Y$ 。

6 计算对角阵 D , 其元素取值为 $d_n(n) = \frac{1}{2\|b_n\|_2}$ 。

直到收敛

2.2 算法分析

本文提出的算法在每次迭代计算中都单调地减少式(22)中问题的目标。

引入以下引理对其进行证明:

引理 1 对于任意非零向量 $b, b_i \in \mathbb{R}^c$, 以下不等式成立:

$$\|b\|_2 - \frac{\|b\|_2^2}{2\|b_i\|_2} \leq \|b_i\|_2 - \frac{\|b_i\|_2^2}{2\|b_i\|_2} \quad (28)$$

证明 从 $(\sqrt{v} - \sqrt{v_i})^2 \geq 0$ 开始:

算法的收敛性总结为以下定理:

$$(\sqrt{v} - \sqrt{v_i})^2 \geq 0 \Rightarrow v - 2\sqrt{vv_i} + v_i \geq 0 \Rightarrow$$

$$\sqrt{v} - \frac{v}{2\sqrt{v_i}} \leq \frac{\sqrt{v_i}}{2} \Rightarrow \sqrt{v} - \frac{v}{2\sqrt{v_i}} \leq \sqrt{v_i} - \frac{v_i}{2\sqrt{v_i}} \quad (29)$$

将式(29)中的 v 和 v_i 分别带入 $\|b\|_2^2$ 和 $\|b_i\|_2^2$, 可以得到式(28)。

定理 1 算法每次迭代都会单调地降低式(22)中问题的目标, 并收敛于问题的全局最优。

证明 可以验证式(27)是以下问题的解:

$$\min_B \text{Tr}(B^T DB) \text{ s.t. } AB = Y \quad (30)$$

因此在 t 次迭代中:

$$B_{t+1} = \arg \min_{B \text{ s.t. } AB=Y} \text{Tr} B^T D_t B \quad (31)$$

这表明:

$$\text{Tr}(B_{t+1}^T D_t B_{t+1}) \leq \text{Tr}(B_t^T D_t B_t) \quad (32)$$

即:

$$\sum_{i=1}^m \frac{\|b_{t+1}^i\|_2^2}{2\|b_t^i\|_2} \leq \sum_{i=1}^m \frac{\|b_t^i\|_2^2}{2\|b_t^i\|_2} \quad (33)$$

其中: 向量 b_t^i 和 b_{t+1}^i 分别表示 B_t 和 B_{t+1} 的第 i 行。

根据引理 1, 对每个 i 有:

$$\|b_{t+1}^i\|_2 - \frac{\|b_{t+1}^i\|_2^2}{2\|b_t^i\|_2} \leq \|b_t^i\|_2 - \frac{\|b_t^i\|_2^2}{2\|b_t^i\|_2} \quad (34)$$

因此, 以下不等式成立:

$$\sum_{i=1}^m \left(\|b_{t+1}^i\|_2 - \frac{\|b_{t+1}^i\|_2^2}{2\|b_t^i\|_2} \right) \leq \sum_{i=1}^m \left(\|b_t^i\|_2 - \frac{\|b_t^i\|_2^2}{2\|b_t^i\|_2} \right) \quad (35)$$

结合式(33)和式(35)可以得到:

$$\sum_{i=1}^m \|b_{t+1}^i\|_2 \leq \sum_{i=1}^m \|b_t^i\|_2 \quad (36)$$

即:

$$\|B_{t+1}\|_{2,1} \leq \|B_t\|_{2,1} \quad (37)$$

因此, 本文提出的算法在每次迭代中将单调地降低式(22)中问题的目标。在收敛性上, B_t 和 D_t 满足式(27), 又因为式(22)是一个凸问题, 所以满足等式(27)表示 B 是式(22)的全局最优解, 因此, 算法将收敛于问题式(22)的全局最优。

3 实验与结果分析

为了检验所提算法的有效性,本节分别在合成数据集和真实数据集上进行实验,在不完整数据集中比较了本文提出的PMF、SID算法、传统的先填充再进行特征选择算法,其中传统方法中的填充算法分别为概述中介绍的均值填充(mean)、KNN填充、EM填充,特征选择算法为本文使用的鲁棒特征选择算法RFS和两种基于边缘的特征选择算法:Simba^[19]和Relief^[20],分别将以上几种算法交叉组合形成9套完整的针对不完整数据集进行特征选择的流程,以评估本文所提算法在高维数据上的分类性能。

本节使用分类精度(ACC)作为评估指标。ACC表示样本分类正确的百分比,即:

$$A_{ACC} = \frac{N_c}{N} \quad (38)$$

其中: N 是样本总数; N_c 为正确被分类的样本个数。

3.1 实验设置

本实验环境为MATLAB2019a,为了观察缺失率对算法的影响,本节人工随机地向数据集注入不同比例的缺失值。缺失率范围为5%~65%,以10%为间隔递增,本文缺失率定义为缺失值占所有值的百分比。在此仅考虑缺失机制为完全随机缺失的情况。为了体现特征选择算法的效果,具体地,本文会向特征较少的数据集中添加一些无关特征,实验中向数据集中添加的无关特征并不被注入缺失。

如前所述, K 为矩阵 U 和 V 的维度,当其取不同值时均方根误差的收敛速度如图1所示,这里分别取 $K=1, 5, 10, 15, 20$,可以看到:当 $K=1, 5$ 时收敛速度较慢并且无法收敛到0,当 $K=10, 15, 20$ 时均方根误差收敛较快,并且可以收敛到0,可见, K 的取值不光影响均方根误差收敛速度,还会影响最终收敛结果。为了不增加计算量并且防止过拟合,本文所有实验均设置 K 值取10。

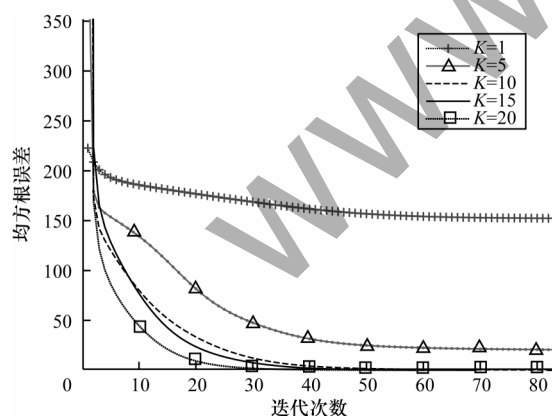


图1 不同 K 值时均方根误差随迭代次数的变化

Fig.1 Variation of root mean square error with iteration times at different K values

不止 K 的取值,学习速率 α 的取值不同也会影响均方根误差的收敛速度,如图2所示, α 分别取0.1、0.01、

0.001、0.000 1,由图2可见, α 取0.01时收敛速度最快, α 取0.000 1时收敛速度最慢,可见步长越小,损失函数到达底部的时间越长,步长越大,损失函数收敛越快,但步长并不能无限大,经过实验发现当 α 取0.1时,目标函数不会收敛,所以在该合成数据集上 α 取0.01。经过在不同数据集上的实验也发现,同一个步长并不适用于所有数据集,所以要通过多次实验发现最适合本数据集的步长,本文中所有数据集步长设置均经过多次实验,设其为最恰当数值。

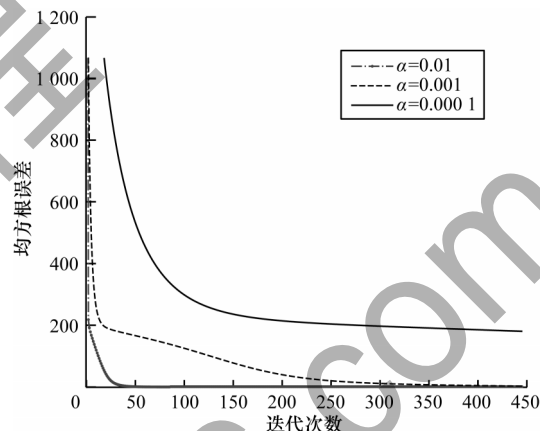


图2 不同 α 值均方根误差随迭代次数变化

Fig.2 Variation of root mean square error with iteration times at different α values

3.2 数据集合成

本节通过设计合成数据集来评价本文算法在存在大量不相关变量的不完整数据中填充缺失值后选择出相关特征的能力。合成数据集包含500个实例和100维特征。其中二维特征是随机产生的0、1序列,将这两维特征做异或运算,产生结果即为合成数据集的标签,剩下的98个特征服从于以0为均值、1为方差的标准正态分布。

本节以所选择的无关特征数目为标准评价特征选择方法。为了简便期间,只比较本文提出的算法与SID算法,以及使用LBFS算法作为特征选择算法的3种传统方法,实验结果如图3所示。

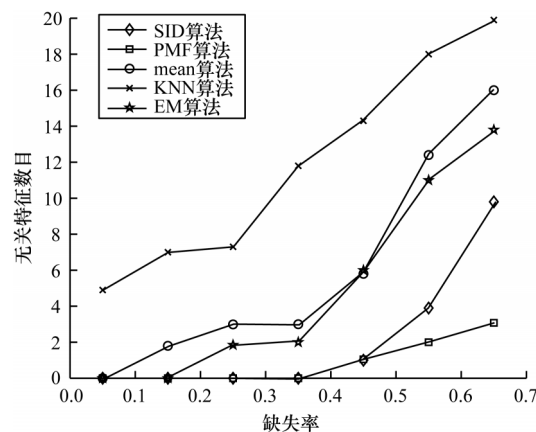


图3 无关特征数随缺失率的变化

Fig.3 Variation of irrelevant feature number with deletion rate

由图3可以看出,在该数据集上,KNN填充的方法选择出的无关特征数目最多,效果最差,本文提出的算法PMF效果最好,在缺失比例较高的情况下依然能选择出较少的无关特征。SID算法在缺失率较低的情况下效果较好,但在缺失率较高时不如PMF。

3.3 真实数据集

本节分别从LIBSVM数据网站、UCI website^[21]、肯特岗生物医学数据集等网站下载DLBCL、Mnist、Splice、Wpbc、USPS、Arcene 6个真实数据集。由于生物医学方面数据集往往包含大量冗余特征,更能体现出特征选择的作用,本节选用的数据集大部分与生物医学有关。DLBCL为弥漫大B细胞淋巴瘤基因的相关数据集,Splice是关于识别DNA序列中两种类型的剪接点的数据集,Wpbc数据集则取自威斯康辛大学医学院的乳腺癌病例数据库,Arcene原始数据来自于美国国家癌症研究中心和东维多利亚医学院,收集了通过SELDI技术采集的癌症病人和健康人的质谱信息,用于癌症预测。由于Splice和Wpbc数据集特征较少,这里人为的向其添加2 000个无关特征,无关特征的特征值均通过从正态分布 $N(0,1)$ 中抽样得到。数据集详细信息如表1所示。

表1 数据集详细信息

数据集	实例数	特征数	类别数
DLBCL	141	661	3
Mnist	5 000	780	10
Splice	1 000	60+2 000	2
Wpbc	198	33+2 000	2
USPS	9 298	256	10
Arcene	200	10 000	2

对于真实数据集,因为无法预知其中特征的重要性,所以使用分类准确率为标准评价特征选择方法。在训练集上先进行特征选择,之后对所选特征进行分类,分类准确率高说明特征选择的效果越好。

使用交叉验证方式^[22]选择最佳参数组合,将原始数据集分为70%的训练集和30%的测试集,其中训练集的类标签是已知的,假设测试集的类标签未知,通过在训练集上训练得到分类器来预测测试实例的类标签,比较预测得到的类标签与真实的类标签就可以得到该分类器的分类精度。本实验使用SVM分类器进行训练。对于基于KNN填充进行的特征选择,本文选取 $k=5$ 。

各算法在DLBCL数据集上的分类效果柱状图如图4所示,其中纵坐标为分类准确率,横坐标依次展示11种不同的算法,不同底纹代表不同的缺失率,这里选取了缺失率为25%、35%和55%时的分类结果,可以看到,本文提出的算法在不同缺失率时准确率都高于其他10个算法,均值填充的3种算法在该数据集上的整体表现略高于KNN填充和EM填充,并且3种不同的特征选择算法效果相差不大,SID算法效果仅次于PMF,

因为其利用了不完整数据集中的全部信息,但该算法没有考虑异常值的影响。

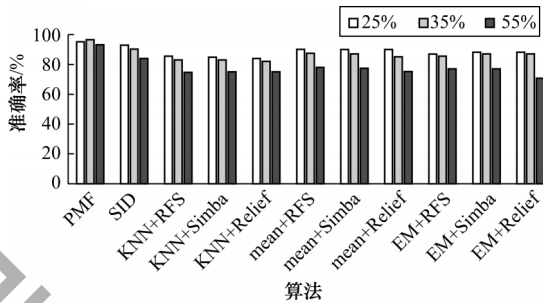


图4 不同缺失率下DLBCL数据集的分类准确率

Fig.4 Classification accuracy of DLBCL dataset under different miss rates

各算法在Mnist数据集上的分类效果如图5所示,可以看到,在缺失率较高的情况下,本文提出的算法依然能达到90%的准确率,在该数据集中用KNN填充的3种算法效果较差,可能是该数据集比较稀疏导致近邻关系度量不够准确,EM填充与SID算法效果相差不大,略高于均值填充。

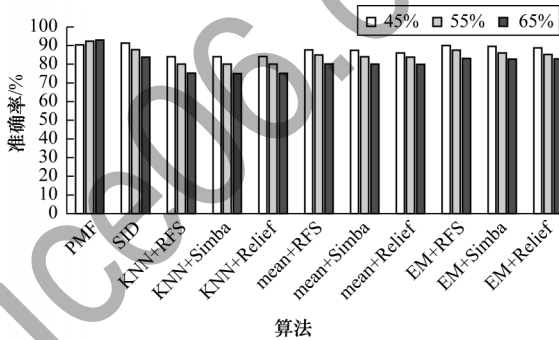


图5 不同缺失率下Mnist数据集的分类准确率

Fig.5 Classification accuracy of Mnist dataset under different miss rates

各算法在Splice数据集上的分类效果柱状图如图6所示,可以看到,基于KNN填充的算法效果较差,在缺失率为25%时只可以达到70%左右的准确率,PMF在相同缺失率时可以达到80%以上的准确率,可见PMF在该数据集上效果有了较大的提升。基于均值填充的算法和SID算法效果较为接近。

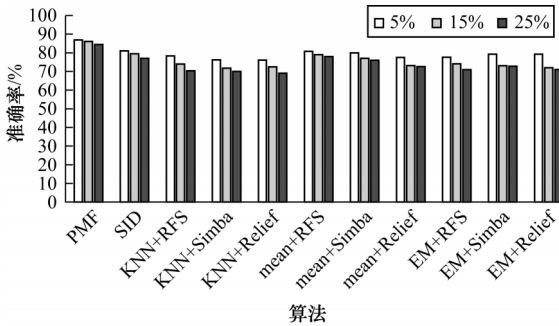


图6 不同缺失率下Splice数据集的分类准确率

Fig.6 Classification accuracy of Splice dataset under different miss rates

各算法在 Wpbc 数据集上的分类效果柱状图如图 7 所示,可以看到,在该数据集上各个算法的分类准确率普遍较低。在该数据集上基于 KNN 填充的算法效果较差,均值填充在 25% 缺失率时的效果较好,在 KNN 填充和均值填充上 RFS 特征选择算法的效果比 Simba 和 Relief 略高。相比其他算法,本文提出的算法在该数据集上效果依然是最好的。

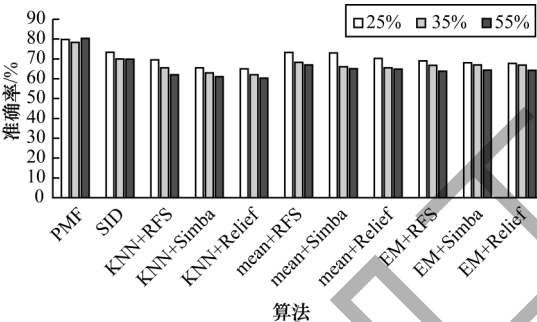


图 7 不同缺失率下 Wpbc 数据集的分类准确率
Fig.7 Classification accuracy of Wpbc dataset under different miss rates

各算法在 USPS 数据集上的分类效果柱状图如图 8 所示,可以看到,本文提出的算法在该数据集上效果较好。在缺失率为 5% 时,所有算法的分类准确率都高达 92% 左右,但是随着缺失率增大到 55%,其他算法的分类效果有了明显下降,此时本文提出的算法优势更加明显。相比之前的基于填充的算法, SID 算法在该数据集上并不具有明显的优势。

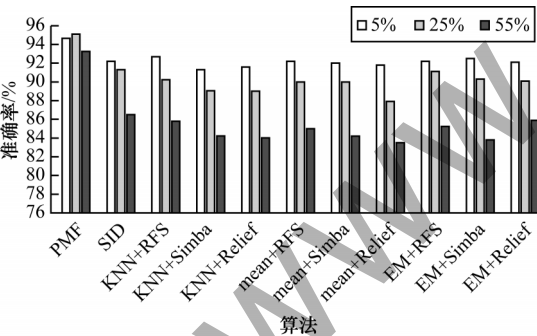


图 8 不同缺失率下 USPS 数据集的分类准确率
Fig.8 Classification accuracy of USPS dataset under different miss rates

各算法在 Arcene 数据集上的分类效果如图 9 所示,可以看到,本文提出的算法在该数据集上的优势非常明显,其他 10 种算法在所有缺失比例时的准确率基本都在 60% 左右,然而 PMF 算法可以达到 90% 左右的准确率,可见本文提出的算法非常适用于该数据集,这应该与数据集本身有关。

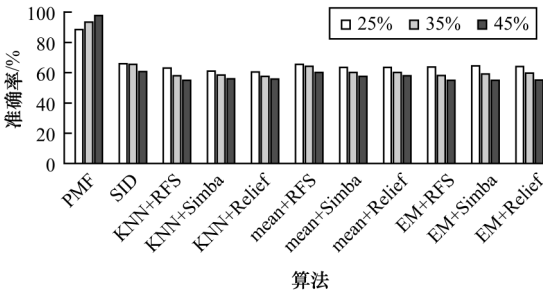


图 9 不同缺失率下 Arcene 数据集的分类准确率
Fig.9 Classification accuracy of Arcene dataset under different miss rates

3.4 异常值检测

本节使用 3.2 节中的合成数据集进行异常值检测,为了方便展示,人工随机地向数据集注入不同比例的异常值,这里使用固定值来模拟异常值。分别在异常值比率为 1%、2%、3% 时进行实验求无关特征,异常值比率定义为注入异常值的数量占整个数据集的百分比,结果如表 2 所示,本文固定缺失值比率为 5%。从表 2 的数据可以看出,在异常值比例较低时,本算法能筛选出所有的重要特征,随着异常值比率增加,筛选出的无关特征数目有所增加,但仍然保持较低的数量,可见本文算法可以在一定程度上应对不完整数据集中异常值带来的影响。

表 2 不同比例异常值与无关特征数目	
Table 2 Different proportion outliers and irrelevant features number	
异常值比率/%	无关特征的数目
0	0.000
1	0.000
2	0.075
3	0.175

综上所述,基于 KNN 填充算法在 Mnist 和 Splice 数据集上效果较差,虽然该算法利用样本近邻关系来进行填充,在一定程度上更多地利用了不完整数据集的信息,但是近邻关系度量的准确性也会对填充效果造成影响。基于均值填充算法与 EM 填充算法效果略高于 KNN 填充算法、SID 算法效果仅次于本文算法,因为该算法使用指示矩阵利用了不完整数据集中的全部信息,但是它没有考虑数据集中异常值的影响。PMF 算法效果最好,因为其不仅充分利用了不完整数据集中的所有有效信息,还使用 $\ell_{2,1}$ 范数作为损失函数,增强了其应对异常值的鲁棒性,所以与其他算法相比,在所有数据集上分类准确率都有所提升。

4 结束语

高维数据集中通常包含缺失值和离群值,对其进行降维是数据挖掘的重要步骤之一。本文针对不完整数据中含有未被观察信息和存在异常值的特点,提出一种基于概率矩阵分解技术的鲁棒特征选择方法。该方法引入指示矩阵利用数据集中全部信息,对不完整数据集进行近似估计,在考虑异常值情形下,利用 $\ell_{2,1}$ 范数对数据点中异常值具有鲁棒性的特点,将其作为回归损失函数,实现鲁棒特征选择。实验结果表明,该方法能够充分利用不完整数据集中的所有信息,避免繁琐的距离运算,可有效应对不完整数据集中异常值带来的影响。下一步考虑将概率矩阵分解填充拓宽到半监督或无监督的特征选择流程上,在数据集标签有缺失甚至无标签的情况下,提升特征选择的效果。

参考文献

- [1] ZHU X F, LI X L, ZHANG S C, et al. Robust joint graph sparse coding for unsupervised spectral feature selection [J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(6): 1263-1275.
- [2] ZHANG Z, LIU L, SHEN F M, et al. Binary multi-view clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1774-1782.
- [3] ZHU X F, ZHANG S C, ZHU Y H, et al. Self-weighted multi-view fuzzy clustering [J]. ACM Transactions on Knowledge Discovery from Data, 2020, 14(4): 1-17.
- [4] HU R Y, ZHU X F, ZHU Y H, et al. Robust SVM with adaptive graph learning [J]. World Wide Web, 2020, 23(3): 1945-1968.
- [5] TSAGRIS M, PAPADOVASILAKIS Z, LAKIOTAKI K, et al. The γ -OMP algorithm for feature selection with application to gene expression data [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 99(1): 1-10.
- [6] CISMONTI F, FIALHO A S, VIEIRA S M, et al. Missing data in medical databases: impute, delete or classify? [J]. Artificial Intelligence in Medicine, 2013, 58(1): 63-72.
- [7] GARCIA C, LEITE D, ŠKRJANC I. Incremental missing-data imputation for evolving fuzzy granular prediction [J]. IEEE Transactions on Fuzzy Systems, 2020, 28(10): 2348-2362.
- [8] SIMONE R. An accelerated EM algorithm for mixture models with uncertainty for rating data [J]. Computational Statistics, 2021, 36(1): 691-714.
- [9] ZHANG S C, LI X L, ZONG M, et al. Efficient kNN classification with different numbers of nearest neighbors [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1774-1785.
- [10] PAN R L, YANG T S, CAO J H, et al. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information [J]. Applied Intelligence, 2015, 43(3): 614-632.
- [11] SEFIDIAN A M, DANESHPOUR N. Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model [J]. Expert Systems with Applications, 2019, 115: 68-94.
- [12] HUANG C C, LEE H M. A grey-based nearest neighbor approach for missing attribute value prediction [J]. Applied Intelligence, 2004, 20(3): 239-252.
- [13] TIAN J, YU B, YU D, et al. Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering [J]. Applied Intelligence, 2014, 40(2): 376-388.
- [14] LOUS, OBRADOVIC Z. Margin-based feature selection in incomplete data [C]//Proceedings of AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2012: 125-136.
- [15] NIE F, HUANG H, XIAO C, et al. Efficient and robust feature selection via joint $\ell_{2,1}$ -norms minimization [C]//Proceedings of International Conference on Neural Information Processing Systems. [S. l.]: Curran Associates Inc., 2010: 389-397.
- [16] YU H Y, GAO L R, LIAO W Z, et al. Global spatial and local spectral similarity-based manifold learning group sparse representation for hyperspectral imagery classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(5): 3043-3056.
- [17] ZHANG C K, WANG C. Probabilistic matrix factorization recommendation of self-attention mechanism convolutional neural networks with item auxiliary information [J]. IEEE Access, 2020, 8: 208311-208321.
- [18] MA C, LI Y X, CHI Y J. Beyond procrustes: balancing-free gradient descent for asymmetric low-rank matrix sensing [J]. IEEE Transactions on Signal Processing, 2021, 69(1): 867-877.
- [19] GILAD-BACHRACH R, NAVOT A, TISHBY N. Margin based feature selection-theory and algorithms [C]//Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM Press, 2004: 452-466.
- [20] KIRA K, RENDELL L A. Feature selection problem: traditional methods and a new algorithm [C]//Proceedings of the 20th IEEE National Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 1992: 129-134.
- [21] AMINI M R, USUNIER N, GOUTTE C. Uci-dataset-url [EB/OL]. [2021-03-20]. <https://dblp.org/rec/journals/prl/Brito>.
- [22] LIU Y, JIANG Z S, XIANG J W. An adaptive cross-validation thresholding de-noising algorithm for fault diagnosis of rolling element bearings under variable and transients conditions [J]. IEEE Access, 2020, 8: 67501-67518.