·人工智能与模式识别·　　文章编号：1000-3428(2022)05-0098-06　　文献标志码：A　　中图分类号：TP391

面向停电分类预测的因子分解机模型

冉　懿[1]，王润年[1]，潘红伟[1]，俞海猛[2]，袁培森[3]

（1.国网新疆电力有限公司 营销服务中心，乌鲁木齐 830000；2.国电南瑞南京控制系统有限公司，南京 211106；
3.南京农业大学 人工智能学院，南京 210095）

摘　要：可靠的电力供应对于工业生产和居民日常生活至关重要，通过对电力数据平台中的停电数据进行分析和挖掘，可以更好地了解配电网停电的潜在规律。分类预测是数据挖掘和分析中的常见技术，停电分类预测可以为企事业单位的停电规划安排提供决策参考。针对停电分类预测问题，提出一种基于因子分解机（FM）的停电数据分类预测模型。利用决策树算法计算停电数据中不同特征的基尼系数以得出重要性得分，从中筛选与停电预测关联度较大的非稀疏特征。根据不同地区的地理位置关系构建不同地区间的空间位置矩阵，并通过矩阵分解的方式构造不同地区在空间上的地理位置关联特征。为防止FM模型出现过拟合问题，在模型中加入L2-范数正则化。在此基础上，利用随机梯度下降的方法训练FM模型，通过训练完成的FM模型对停电数据进行分类预测。在真实停电数据集上的实验结果表明，该模型在训练数据集和测试数据集上的F1值和准确率分别高达0.90和0.89，优于DNN、SVM、XGBoost等模型。

关键词：停电分类预测；决策树；矩阵分解；因子分解机；随机梯度下降方法

开放科学（资源服务）标志码（OSID）：

中文引用格式：冉懿，王润年，潘红伟，等. 面向停电分类预测的因子分解机模型[J]. 计算机工程，2022，48(5)：98-103，111.

英文引用格式：RAN Y，WANG R N，PAN H W，et al. Factorization machine model for power outage classification prediction[J]. Computer Engineering，2022，48(5)：98-103，111.

Factorization Machine Model for Power Outage Classification Prediction

RAN Yi[1]，WANG Runnian[1]，PAN Hongwei[1]，YU Haimeng[2]，YUAN Peisen[3]

（1.Marketing Service Center，State Grid Xinjiang Electric Power Co.，Ltd.，Urumqi 830000，China；
2.NARI-TECH Nanjing Control Systems Co.，Ltd.，Nanjing 211106，China；
3.College of Artificial Intelligence，Nanjing Agricultural University，Nanjing 210095，China）

【Abstract】Reliable power supply is important for industrial production and residential daily life. By analyzing and mining the outage data in power data platforms，we can better understand the potential law of network outage distributions. Classification prediction is a common technology in data mining and analysis. Outage classification prediction can provide a decision-making reference for outage planning as well as the arrangement of enterprises and institutions. Concerning blackout classification and prediction，a blackout data classification and prediction model based on the Factorization Machine（FM） model is proposed. The Gini coefficients of different features in outage data are calculated using a decision-tree algorithm to obtain the importance score，and the non-sparse features demonstrating a high correlation with outage prediction are selected. According to the geographical location relationship of different regions，the spatial location matrix between different regions is constructed，and the spatial geographic location correlating features of different regions are constructed using matrix decomposition. To prevent overfitting in the FM model，L2-norm regularization is added to the model. On this basis，the FM model is trained using random gradient descent，and the outage data are classified and predicted by the trained FM model. The experimental results on the real outage dataset show that the F1-score and accuracy of the model on the training and test datasets are as high as 0.90 and 0.89，respectively，which is better than other models，such as DNN，SVM，and XGBoost.

【Key words】power outage classification prediction；decision tree；matrix decomposition；Factorization Machine（FM）；Stochastic Gradient Descent（SGD） method

DOI：10. 19678/j. issn. 1000-3428. 0061529

基金项目：国家自然科学基金（61502236，61806097）。

作者简介：冉　懿（1988—），男，工程师，主研方向为电能计量、电力数据分析、数据挖掘；王润年、潘红伟、俞海猛，工程师；袁培森（通信作者），副教授、博士。

收稿日期：2021-04-30　　**修回日期：**2021-06-06　　**E-mail：**peiseny@163.com

0　概述

可靠的电力供应对于现代社会的发展至关重要，社会生产中的很多方面都依赖于电力建设，因此，电力设施被认为是现代社会中较为关键的基础设施之一[1-2]。一旦停电，会严重影响人们的日常生活[3-4]以及其他关键基础设施系统的运行，从而造成巨大的经济损失。

为了更好地分析和管理电网系统所产生的电力数据，电网企业搭建了智慧电力大数据平台。通过对平台中的停电数据进行分析和挖掘，可以更好地了解电网停电的潜在规律[5-6]。通过分析历史停电数据，根据分析所得的规律进行停电分类预测，能够为依赖于电力设施的其他公司、公共事业单位等的停电规划安排提供决策参考[7]。从短期看，电网停电的分类预测可以帮助企事业单位提前做好准备，平衡人力、材料成本以及加快电力恢复速度；从长期看，可以根据停电数据分析得到电力系统中需要加强的板块，还能够通过数据分析来设置合适的备用电源数目，以提高本地电网系统的供电可靠性[8]。

目前，国内外学者针对电力停电分类预测问题进行了大量研究，并且取得了一定成果。XIE等[9]通过分析雷电天气的停电数据，提出一种基于通用回归神经网络（GRNN）的方法，以预测雷电天气下的停电情况。该方法对历史雷电天气的停电数据进行特征提取，并将提取的特征作为GRNN的输入以训练模型，然后根据训练的模型来预测停电情况。ZHAI等[10]设计一种新模型，其通过使用公开可用的数据来准确估计各个建筑物级别的停电情况。该模型使用脆弱性函数模拟在危险负荷下单个建筑物级别的停电情况，能够提供更多的本地化、建筑物级别的估计，以评估由于自然灾害而造成的停电。侯慧等[11]采用随机森林方法，结合气象、地理数据来预测不同区域的停电情况，并通过对停电网格进行重要性评估来提高预测的准确性。

本文提出一种基于因子分解机（Factorization Machine，FM）的方法，以对停电数据进行分类预测。为了提取有效的特征来降低数据处理的复杂度，通过决策树算法对停电数据进行特征选择。为了获取更多的有效特征，基于不同地区的空间位置，利用矩阵分解构造空间位置特征。通过梯度下降方法训练因子分解机模型，为了防止模型过拟合，在优化目标中加入L2正则化。在此基础上，利用所训练的模型进行停电数据分类预测，从而为电网公司的用电决策提供参考。

1　特征选择

在特征选择时，需要避免选择太多或太少的特征：如果选择的特征太少，则特征数据中蕴含的信息内容可能会很少；如果选择的特征太多，则可能会存在一些不相关的特征，从而提高任务学习的难度。本文通过决策树[12]进行特征选择，具体地，计算得到不同特征在决策树上所作的贡献，通过贡献度的大小来选择特征。同时，采用基尼系数（Gini）来衡量不同特征的重要性。Gini最早用于经济学领域，主要用于衡量收入分配的公平性。在决策树的构建过程中，Gini通常用来测量数据的纯度或不确定性。

假设样本数据有C个特征$\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_C$，特征$\boldsymbol{X}_j$的基尼系数$\mathrm{VIM}_j^{\mathrm{Gini}}$表示决策树中第$j$个特征节点分裂不纯度的平均变化量。Gini的计算方式如式（1）所示：

$$G_{\mathrm{Gini}}=1-\sum_{k=1}^{|K|}p_k^2 \tag{1}$$

其中：K表示类别总数；p_k表示第k个类别所占的比例。

特征$\boldsymbol{X}_i$的某个取值x将样本数据分成2个部分D_1、D_2，则特征$\boldsymbol{X}_j$的基尼系数$\mathrm{VIM}_j^{\mathrm{Gini}}$的计算方式如式（2）所示：

$$\mathrm{VIM}_j^{\mathrm{Gini}}=\frac{|D_1|}{D}\mathrm{Gini}(D_1)+\frac{|D_2|}{D}\mathrm{Gini}(D_2) \tag{2}$$

其中：$\mathrm{Gini}(D_1)$表示D_1的Gini；$\mathrm{Gini}(D_2)$表示D_2的Gini。

利用决策树算法计算部分特征的重要性分数，对计算出的不同特征的重要性分数进行排序，从中选择对停电分类预测较为重要的特征用于模型训练。

2　停电预测方法

2.1　特征构造

为了利用更多的有效特征，本文在已有停电数据的基础上，根据不同地区的地理位置关系构造新的位置关联特征。本文认为相邻地区之间的停电情况具有相关性，如果2个区域相邻或有重叠的地理区域，则某地区停电，相邻的区域也有很大的可能性停电。

假设共有n个区域$p_1,p_2,\cdots,p_n$，现构造一个关联矩阵$\boldsymbol{A}^{n\times n}$，如果2个区域$p_i$、$p_j$相邻或有重叠的地理区域，那么对应的矩阵元素值$a_{ij}$为1，其他非对角线元素值均为0。由不同地区构造出来的关联矩阵$\boldsymbol{A}^{n\times n}$的形式如下：

$$\boldsymbol{A}=\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{3}$$

其中：a_{ii}的值为1；a_{ij}表示地区p_i和p_j之间的停电关联值，2个区域相邻或有重叠的地理区域，$a_{ij}=1$，否则$a_{ij}=0$。

显然，矩阵$\boldsymbol{A}$的值不能直接作为停电特征，需要对$\boldsymbol{A}$进行矩阵分解，本文采用LU分解的方式来完成。LU分解[13]是将矩阵变成下三角矩阵与上三角矩阵的乘积，其形式如下：

$$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U} \tag{4}$$

其中：$\boldsymbol{L}$为下三角矩阵，即当$i>j$时，$l_{ij}=0$；$\boldsymbol{U}$为上三

角矩阵，即当$i<j$时，$u_{ij}=0$。

对于停电样本数据，每一行代表一个区域的停电情况，一共有n个区域$p_1,p_2,\cdots,p_n$。式(4)将矩阵$\boldsymbol{A}$分解成$n\times n$矩阵$\boldsymbol{L}$与$n\times n$矩阵$\boldsymbol{U}$的乘积，假如样本数据的第i行是区域p_i的停电情况，那么该行对应的空间位置构造特征就是矩阵$\boldsymbol{L}$的第i行数据。通过矩阵分解的方式构造不同区域的空间位置特征，可以为模型训练提供更多的有效特征，从而提升模型的分类准确性。

2.2 因子分解机

FM将支持向量机的优点与因子分解模型相结合[14]，其可与任何实值特征向量一起使用，是一种通用预测模型。FM的机制是使用分解参数对变量之间的交互进行建模，即使在具有稀疏性的问题中也可以估计交互，对于停电数据中存在的大量稀疏特征，如月份、年份等，FM具有适用性。此外，由于FM的模型方程可以在其因子数量k和特征数量n方面都降为线性复杂度，因此FM的计算效率很高，这也意味着FM模型的预测时间是线性的，且减少了训练阶段要学习的参数量[15-16]。

假设$\boldsymbol{X}\in\mathbb{R}^n$表示维度为$n$的特征向量，$\langle\cdot,\cdot\rangle$表示2个大小为$k\in N_0^+$的向量的点积，向量的点积计算如下：

$$\left\langle \boldsymbol{v}_i,\boldsymbol{v}_j\right\rangle=\sum_{f=1}^{k}v_{if}v_{jf} \tag{5}$$

FM能够通过使用因子分解模型来对不同特征之间的交互进行建模，尤其是FM模型通过分解交互特征以估计交互，从而打破交互特征之间的独立性。FM模型考虑到不同特征之间的关联关系，引入交叉项，通过对特征两两相乘来找到一些组合特征。对于二阶交叉，FM模型方程如下：

$$\hat{y}=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left\langle \boldsymbol{v}_i,\boldsymbol{v}_j\right\rangle x_ix_j \tag{6}$$

其中：$\hat{y}$表示预测值；$\boldsymbol{x}\in\mathbb{R}^n$表示模型方程的输入向量；$x_i$表示向量$\boldsymbol{x}$的第$i$个元素；$w_0\in\mathbb{R}$表示全局偏差；$\boldsymbol{w}\in\mathbb{R}^n$表示输入向量$\boldsymbol{x}$的权重向量；$\boldsymbol{V}\in\mathbb{R}^{n\times k}$是潜在的特征矩阵，用来表示第$i$个变量和第$j$个变量之间的交叉项；$\boldsymbol{v}_i$表示$x_i$的特征向量；$\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle$用于建模$x_i$和$x_j$的相互交叉。直接采用式(6)计算模型方程的时间复杂度为$O(kn^2)$，本文对其进行进一步优化，优化过程如下：

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left\langle \boldsymbol{v}_i,\boldsymbol{v}_j\right\rangle x_ix_j=\\&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\langle \boldsymbol{v}_i,\boldsymbol{v}_j\right\rangle x_ix_j-\frac{1}{2}\sum_{i=1}^{n}\left\langle \boldsymbol{v}_i,\boldsymbol{v}_i\right\rangle x_ix_i=\\&\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{f=1}^{k}v_{if}v_{jf}x_ix_j-\sum_{i=1}^{n}\sum_{f=1}^{k}v_{if}v_{if}x_ix_i\right)=\\&\frac{1}{2}\sum_{f=1}^{k}\left[\left(\sum_{i=1}^{n}v_{if}x_i\right)^2-\sum_{i=1}^{n}v_{if}^2x_i^2\right]\end{aligned} \tag{7}$$

将式(7)中的结果代入式(6)，得到预测值的表达式为：

$$\hat{y}=w_0+\sum_{i=1}^{n}w_ix_i+\frac{1}{2}\sum_{f=1}^{k}\left[\left(\sum_{i=1}^{n}v_{if}x_i\right)^2-\sum_{i=1}^{n}v_{if}^2x_i^2\right] \tag{8}$$

通过数学变换，FM计算方程的时间复杂度降为$O(kn)$，这也说明FM模型的计算成本相对于潜在特征的维数和特征是线性的。

停电分类预测属于二分类问题，本文采用logitloss函数作为FM模型的损失函数，其表达式如下：

$$\text{loss}\left(\hat{y},y\right)=-\ln\frac{1}{1+\text{e}^{-\hat{y}y}} \tag{9}$$

由式(9)可以看出，模型的预测值和真实值越接近，则损失函数的值越小。

FM模型的目标是总损失函数最小化，总损失函数计算如下：

$$\text{Loss}_{总}=\sum_{i=1}^{N}\text{loss}\left(\hat{y}\left(x_i\right),y\left(x_i\right)\right) \tag{10}$$

由于本文损失函数选择的是logitloss函数，因此总损失函数被更新为：

$$\text{Loss}_{总}=\sum_{i=1}^{N}\left(-\ln\frac{1}{1+\text{e}^{-\hat{y}\left(x_i\right)y\left(x_i\right)}}\right) \tag{11}$$

针对FM模型的优化，目标如下：

$$\min_{\varTheta}\sum_{i=1}^{N}\text{loss}\left(\hat{y}(x_i),y(x_i)\right) \tag{12}$$

为防止FM模型过拟合，本文引入一种L2-范数正则化优化技术[17]，L2范数正则化的基本思想是在原始成本函数中添加一个额外项，称为正则项，其中仅包含L2范数误差项，并带有一个用于控制正则化相对量的超参数。可以将这种技术视为在原始L2范数误差项和L2范数正则项之间的一种折衷方法，其能增强FM模型的泛化能力。加入L2-范数正则项后的优化目标更新为：

$$\min_{\varTheta}\left(\sum_{i=1}^{N}\text{loss}\left(\hat{y}\left(x_i\right),y\left(x_i\right)\right)+\sum_{\theta\in\varPhi}\lambda_\theta\theta^2\right) \tag{13}$$

其中：λ为L2正则化的系数，λ的取值影响FM模型的泛化能力。

本文通过随机梯度下降[18]的方法训练FM模型。随机梯度下降是不断地沿着目标函数梯度的反方向去寻找损失函数值最小的参数。求损失函数$y(\boldsymbol{x})$关于θ的偏导，如下：

$$\begin{aligned}\frac{\partial\text{loss}\left(\hat{y},y\right)}{\partial\theta}&=-\frac{1}{\sigma\left(\hat{y}y\right)}\sigma\left(\hat{y}y\right)\left[1-\sigma\left(\hat{y}y\right)\right]y\frac{\partial\hat{y}}{\partial\theta}=\\&\left[\sigma\left(\hat{y}y\right)-1\right]y\frac{\partial\hat{y}}{\partial\theta}\end{aligned} \tag{14}$$

其中：$\sigma(\cdot)$表示sigmoid函数。

利用随机梯度下降的方法训练FM模型的算法描述如算法1所示。

算法1 FM模型训练算法

输入 训练数据集S，正则化参数λ，梯度下降参数学习率η，初始化参数σ

输出　模型参数 $\Theta=\left(w_0, w, V\right)$

1.$w_0, w, V \leftarrow \text{Initialization}(0, 0, \sigma)$

$j \leftarrow 1$

2.WHILE！ stopping criterion DO

3.FOR $(x, y) \in S$ DO

4.$w_0 \leftarrow w_0 - 2\eta\lambda_0 w_0 - \eta \frac{\partial}{\partial w_0} \text{loss}\left(\hat{y}(x|\Theta), y\right)$

5.FOR $i=1$ to $p \wedge x_i \neq 0$ DO

6.$w_i \leftarrow w_i - 2\eta\lambda_w^{\pi(i)} w_i - \eta \frac{\partial}{\partial w_i} \text{loss}\left(\hat{y}(x|\Theta), y\right)$

7.FOR $f=1$ to k DO

8.$v_i \leftarrow w_{if} - 2\eta\lambda_v^{f,\pi(i)} v_{if} - \eta \frac{\partial}{\partial v_{if}} \text{loss}\left(\hat{y}(x|\Theta), y\right)$

9.END FOR

10.END FOR

11.END WHILE

12.RETURN C, γ

算法1通过随机梯度下降算法训练FM模型：首先，对FM模型中的参数进行初始化操作；接着，利用随机梯度下降算法根据设置的学习率对FM中的参数进行更新，每次更新后计算给定条件是否满足，如果满足，则停止迭代，如果不满足，则继续迭代更新直到满足更新条件；最后，返回停止迭代后的模型参数值。

3　实验分析

3.1　实验数据集与评估标准

本文采用的实验数据集一共包括23 768条数据，记录了14个地区的停电情况，数据包括天气、人口、停电时间等非稀疏特征，以及年份、月份、地区等稀疏特征。

本文采用基于混淆矩阵[19]的评估度量。混淆矩阵如表1所示。

表1　混淆矩阵

Table 1　Confusion matrix

真实值	预测值	
	Predicted 1	Predicted 0
True 1	T_{TP}	F_{FN}
True 0	F_{FP}	T_{TN}

T_{TP}是真阳性，表示被分类器正确分类的正类数据；T_{TN}是真阴性，表示被分类器正确分类的负类数据；F_{FP}是假阳性，表示被分类器错误地标记成正类数据而实际是负类的样本数据；F_{FN}是假阴性，表示被分类器错误地标记为负类而实际为正类的样本数据。根据混淆矩阵计算分类器的准确率（Accuracy）和F1值（F1-score），以衡量模型的性能[20-21]。准确率和F1值的计算公式如下：

$$A_{\text{Accuracy}} = \frac{T_{TP} + T_{TN}}{T_{TP} + F_{FN} + F_{FP} + T_{TN}} \tag{15}$$

$$\text{F1} = 2 \times \frac{P \times R}{P + R} \tag{16}$$

其中：P表示精确率；R表示召回率。P和R的计算方式如下：

$$P = \frac{T_{TP}}{T_{TP} + F_{FP}} \tag{17}$$

$$R = \frac{T_{TP}}{T_{TP} + F_{FN}} \tag{18}$$

根据分类器混淆矩阵的值，即可计算最终的分类器分类准确率和F1值。

3.2　结果分析

本文利用决策树算法计算城市停电数据非稀疏特征的重要性，根据计算出的不同特征的重要性和设定的阈值，筛选出对停电数据分类预测相对重要的特征以训练FM模型，从而提高FM模型的分类预测性能。通过决策树算法计算的非稀疏特征的重要性得分结果如图1所示，将重要性分数的阈值设置为0.1，本文选取重要性分数排在前4位的非稀疏特征进行模型训练。

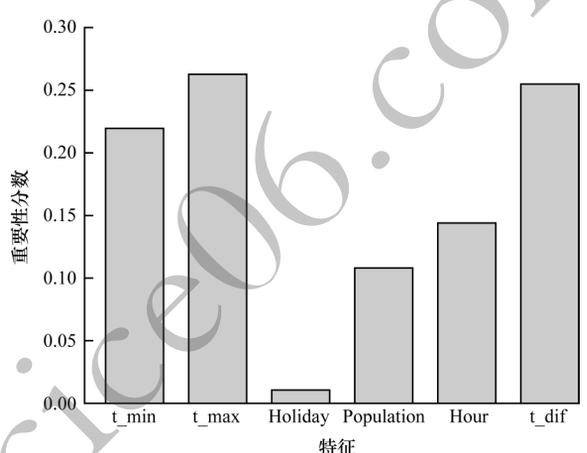

图1　非稀疏特征的重要性分数

Fig.1　Importance score of non-sparse features

为了进一步说明特征提取的必要性以及根据决策树算法提取特征的有效性，本文分别利用上述选取的特征以及全部的非稀疏特征进行实验，比较2种情况下的F1值、准确率以及运行时间，实验结果如表2所示。从表2可以看出：本文特征提取方法能够有效提升FM模型的性能，在F1和准确率2项指标上均有提升；利用全部非稀疏特征训练模型的时间大于特征提取后的模型训练时间，这说明经过特征提取后再进行模型训练，能够节约训练时间，节省CPU资源。

表2　特征提取前后的实验结果

Table 2　Experimental results before and after feature extraction

指标	特征提取前	特征提取后
F1值	0.89	0.90
准确率	0.89	0.90
运行时间/s	170	155

本文利用筛选出的非稀疏特征和稀疏特征来训练FM模型,训练次数设置为200次。随着训练次数的增加,模型分类预测的准确率和F1值如图2所示。从图2可以看出,随着模型的不断训练,准确率和F1值都在提高,这说明FM模型的分类预测效果越来越好,且当迭代次数较多时,对于停电数据能够取得较好的预测效果。此外,当迭代次数达到一定数值时,指标值变化很小,基本趋于稳定,说明模型的分类性能保持稳定。

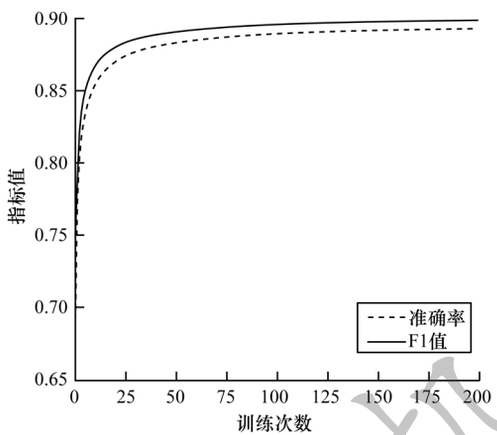

图2 模型的准确率和F1值变化曲线

Fig.2 The accuracy and F1 value curve of the model

本文将FM模型和线性分类(Linear)模型、深度神经网络(DNN)模型、支持向量机(SVM)模型、朴素贝叶斯(Bayes)模型、XGBoost模型[22]进行对比,比较不同模型的准确率和F1值。对于不同的分类预测模型,训练数据集和测试数据集的比例均为3∶1。

训练数据集在不同模型下的准确率和F1值结果如表3所示。从表3可以看出,FM模型的分类效果均优于另外5种模型。F1值为分类模型召回率和精确率的综合指标,也是衡量模型分类预测性能的综合指标,其值越大,说明模型的分类预测效果越好。FM模型在训练数据集下的F1值为0.90,比线性分类模型提高11%,比DNN模型提高34%,比SVM模型提高17%,比朴素贝叶斯模型提高23%,比XGBoost模型提高8.4%。准确率是表示模型将正负样例正确分类的能力,其值越大,说明模型分类正确的比例越大,模型性能越好。FM模型在训练数据集下的准确率为0.89,比线性分类模型提高11%,比DNN模型提高30%,比SVM模型提高17%,比朴素贝叶斯模型提高22%,比XGBoost模型提高8.4%。根据准确率和F1值可以看出,训练数据集在FM模型下效果较好,并且优于其他模型,这是因为在电力数据中存在大量的稀疏数据,而FM针对稀疏性问题也可以估计交互,其借助因子分解技术打破了交叉项参数之间的独立性,使得在稀疏性问题中具有更强的学习能力。

表3 训练集中不同模型的评价指标结果

Table 3 Evaluation index results of different models in the training set

模型	F1值	准确率
FM	0.90	0.89
线性分类	0.81	0.81
深度神经网络	0.67	0.69
支持向量机	0.77	0.77
朴素贝叶斯	0.73	0.74
XGBoost	0.83	0.83

测试数据集在不同模型下的准确率和F1值结果如表4所示。从表4可以看出,FM模型的分类性能评价指标均优于其他模型。FM模型在测试集下的F1值为0.90,F1值越接近1,模型的分类预测效果越好。FM模型的F1值比线性分类模型提高12.5%,比DNN模型提高43%,比SVM模型提高18%,比朴素贝叶斯模型提高23%,比XGBoost模型提高9.8%。测试集下FM模型的准确率为0.89,说明有89%的样本数据被正确分类。FM模型的准确率值比线性分类模型提高11.3%,比DNN模型提高38.5%,比SVM模型提高18.4%,比朴素贝叶斯模型提高21.6%,比XGBoost模型提高8.5%。无论是准确率还是F1值,测试数据集在FM模型下的结果都较好,利用训练集训练出来的模型,用测试集去验证其性能,得到的结果依旧较好,并且都优于另外5种模型,这再次验证了使用FM模型对停电数据集进行分类预测具有有效性。

表4 测试集中不同模型的评价指标结果

Table 4 Evaluation index results of different models in the test set

模型	F1值	准确率
FM	0.90	0.89
线性分类	0.80	0.80
深度神经网络	0.63	0.65
支持向量机	0.76	0.76
朴素贝叶斯	0.73	0.74
XGBoost	0.82	0.82

4 结束语

针对停电分类预测问题,本文提出一种基于随机梯度优化的FM分类预测模型。利用决策树算法计算停电数据中不同特征的重要性得分,根据设定的阈值筛选出与停电预测关联度较大的特征。利用不同地区的空间位置建立位置矩阵,通过矩阵分解构造空间特征。基于选择的特征对应的样本数据使用随机梯度下降方法训练FM模型,并加入L2正则化以防止模型过拟合,最终利用训练好的FM模型对停电数据进行分类预测。实验结果验证了FM模型

在停电数据分类预测任务中的有效性。下一步将对不同季节的停电数据进行分类,利用自编码器模型对停电数据完成特征提取并对多维数据作降维处理,使用降维后的数据特征训练FM模型,从而提高模型的分类预测精度。

参考文献

[1] 于群,屈玉清,石良. 基于相对值法和Hurst指数的电网停电事故自相关性分析[J]. 电力系统自动化,2018,42(1):55-60,124.
YU Q,QU Y Q,SHI L. Self-correlation analysis of power grid blackouts based on relative value method and Hurst exponent[J]. Automation of Electric Power Systems,2018,42(1):55-60,124. (in Chinese)

[2] 罗鸿轩,肖勇,金鑫. 基于电力客户分群特征的停电敏感度预测算法研究[J]. 西南师范大学学报(自然科学版),2020,45(10):106-112.
LUO H X,XIAO Y,JIN X. On prediction algorithm of blackout sensitivity based on characteristics of power customer clustering[J]. Journal of Southwest China Normal University (Natural Science Edition),2020,45(10):106-112. (in Chinese)

[3] 郑旭,丁坚勇,尚超,等. 计及多影响因素的电网停电损失估算方法[J]. 武汉大学学报(工学版),2016,49(1):83-87.
ZHENG X,DING J Y,SHANG C,et al. An assessment method of grid outage cost considering multifactorial influences[J]. Engineering Journal of Wuhan University,2016,49(1):83-87. (in Chinese)

[4] 盛银波,仲立军,张利庭,等. 基于停电明细数据的配电网可靠性监测与研究[J]. 浙江电力,2017,36(12):70-74.
SHENG Y B,ZHONG L J,ZHANG L T,et al. Reliability monitoring and research of distribution networks based on detailed outage data[J]. Zhejiang Electric Power,2017,36(12):70-74. (in Chinese)

[5] 孙毅,毛烨华,李泽坤,等. 面向电力大数据的用户负荷特性和可调节潜力综合聚类方法[J]. 中国电机工程学报,2021,41(18):6259-6271.
SUN Y,MAO Y H,LI Z K,et al. A comprehensive clustering method of user load characteristics and adjustable potential based on power big data[J]. Proceedings of the CSEE,2021,41(18):6259-6271. (in Chinese)

[6] 陈江兴,梁良,付俊峰,等. 基于大数据的智能电网数据调度与快速分发方法研究[J]. 电测与仪表,2020,57(6):88-93.
CHEN J X,LIANG L,FU J F,et al. Research on smart grid data scheduling and fast distribution method based on big data[J]. Electrical Measurement & Instrumentation,2020,57(6):88-93. (in Chinese)

[7] 刘天浩,朱元振,孙润稼,等. 极端自然灾害下电力信息物理系统韧性增强策略[J]. 电力系统自动化,2021,45(3):40-48.
LIU T H,ZHU Y Z,SUN R J,et al. Resilience-enhanced strategy for cyber-physical power system under extreme natural disasters[J]. Automation of Electric Power Systems,2021,45(3):40-48. (in Chinese)

[8] 张俊潇,唐俊熙,曹华珍,等. 配电终端全寿命周期成本模型与智能优化求解[J]. 电测与仪表,2020,57(20):81-89.
ZHANG J X,TANG J X,CAO H Z,et al. Life cycle cost model and intelligent optimization of distribution automation terminal unit[J]. Electrical Measurement & Instrumentation,2020,57(20):81-89. (in Chinese)

[9] XIE Y Y,LI C J,LÜ Y J,et al. Predicting lightning outages of transmission lines using generalized regression neural network[J]. Applied Soft Computing,2019,78:438-446.

[10] ZHAI C W,CHEN T Y J,WHITE A G,et al. Power outage prediction for natural hazards using synthetic power distribution systems[J]. Reliability Engineering & System Safety,2021,208:107348.

[11] 侯慧,耿浩,肖祥,等. 台风灾害下用户停电区域预测及评估[J]. 电网技术,2019,43(6):1948-1954.
HOU H,GENG H,XIAO X,et al. Research on prediction and evaluation of user power outage area under typhoon disaster[J]. Power System Technology,2019,43(6):1948-1954. (in Chinese)

[12] 丁飞鸿,刘鹏,卢暾,等. 基于遗传优化决策树的建筑能耗短期预测模型[J]. 计算机工程,2019,45(6):280-289,296.
DING F H,LIU P,LU T,et al. Short-term prediction model of building energy consumption based on genetically optimized decision tree[J]. Computer Engineering,2019,45(6):280-289,296. (in Chinese)

[13] 程凯,田瑾,马瑞琳. 基于GPU的高效稀疏矩阵存储格式研究[J]. 计算机工程,2018,44(8):54-60.
CHENG K,TIAN J,MA R L. Study on efficient storage format of sparse matrix based on GPU[J]. Computer Engineering,2018,44(8):54-60. (in Chinese)

[14] RENDLE S. Factorization machines[C]//Proceedings of the 10th IEEE International Conference on Data Mining. Washington D. C.,USA:IEEE Press,2010:995-1000.

[15] 赵衍衍,张良富,张静,等. 因子分解机模型研究综述[J]. 软件学报,2019,30(3):799-821.
ZHAO K K,ZHANG L F,ZHANG J,et al. Survey on factorization machines model[J]. Journal of Software,2019,30(3):799-821. (in Chinese)

[16] 于龙飞. 基于深度因子分解机的分类算法研究[D]. 北京:北京邮电大学,2020.
YU L F. Research on deep factorization machine based classification algorithm[D]. Beijing:Beijing University of Posts and Telecommunications,2020. (in Chinese)

[17] 袁广林,薛模根. L2范数正则化鲁棒编码视觉跟踪[J]. 电子与信息学报,2014,36(8):1838-1843.
YUAN G L,XUE M G. Robust coding via L2-norm regularization for visual tracking[J]. Journal of Electronics & Information Technology,2014,36(8):1838-1843. (in Chinese)

[18] 史加荣,王丹,尚凡华,等. 随机梯度下降算法研究进展[J]. 自动化学报,2021,47(9):2103-2119.
SHI J R,WANG D,SHANG F H,et al. Research advances on stochastic gradient descent algorithms[J]. Acta Automatica Sinica,2021,47(9):2103-2119. (in Chinese)

[19] 徐健锋,苗夺谦,张远健. 基于混淆矩阵的多目标优化三支决策模型[J]. 模式识别与人工智能,2017,30(9):859-864.
XU J F,MIAO D Q,ZHANG Y J. Three-way decisions model for multi-object optimization based on confusion matrix[J]. Pattern Recognition and Artificial Intelligence,2017,30(9):859-864. (in Chinese)

(下转第111页)

（上接第103页）

[20] 郭慧，刘忠宝，柳欣. 基于云模型与决策树的入侵检测方法[J]. 计算机工程，2019，45(4)：142-147.
GUO H，LIU Z B，LIU X. Intrusion detection method based on cloud model and decision tree[J]. Computer Engineering，2019，45(4)：142-147.（in Chinese）

[21] 张国芳，刘通宇，温丽丽，等. 基于变分自编码器的日线损率异常检测研究[J]. 华东师范大学学报(自然科学版)，2020(5)：146-155.
ZHANG G F，LIU T Y，WEN L L，et al. Research on abnormal detection of daily loss rate based on a variational auto-encoder[J]. Journal of East China Normal University (Natural Science)，2020(5)：146-155.（in Chinese）

[22] 严道波，杨勇，邱丹，等. 基于天气因素和XGBoost算法的配电线路故障停电预测[J]. 电力与能源，2019，40(2)：168-171.
YAN D B，YANG Y，QIU D，et al. Failure prediction of distribution line based on weather factors and XGBoost algorithm[J]. Power & Energy，2019，40(2)：168-171.（in Chinese）

编辑　吴云芳