

基于集成学习的城市空间停车难度预测

谭文安, 刘新乐

(上海第二工业大学 计算机与信息工程学院, 上海 201209)

摘要: 停车位预测技术是解决城市停车难问题的一种可行方案。针对神经网络等预测模型难以应对诸如路边占道停车等复杂情况, 提出了一个基于支持向量机和决策树集成的模型训练方法, 不再着重预测停车位的个数, 而是预测某一位置的停车难度。在每轮训练过程中拟合一个支持向量机模型, 同时收集预测出错的样本, 最后在误分类样本集合上训练决策树模型来提高整个模型的预测准确性。采用该方法训练了一个城市空间停车难度预测模型, 并利用该模型预测了近一周时间的停车难度。实验结果显示, 该方法的预测效果优于单独使用支持向量机、决策树和全连接神经网络模型, 可以较好地捕捉到停车难度随时间变化的基本情况。

关键词: 停车预测; 集成学习; 支持向量机; 决策树

中图分类号: TP391

文献标志码: A

Urban Space Parking Difficulties Prediction Based on Ensemble Learning

TAN Wen'an, LIU Xinle

(School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China)

Abstract: Parking space prediction technology is a feasible solution to solve the problem of urban parking difficulties. In view of the difficulty of predictive models such as neural networks in dealing with complex situations like roadside parking, an integrated model training method based on support vector machine and decision tree is proposed, which no longer focuses on predicting the number of parking spaces, but on predicting the parking difficulty of a certain location. In each training cycle, a support vector machine model is fitted, and the prediction error samples are collected. Finally, the decision tree model is trained on the set of misclassification samples to improve the prediction accuracy of the whole model. This method is used to train a prediction model of parking difficulty in urban space, and the model is used to predict parking difficulty in nearly a week. The experimental results show that the prediction effect of this method is better than that of using support vector machine, decision tree and fully connected neural network model alone, and can capture the basic situation of parking difficulty changing with time.

Keywords: parking prediction; ensemble learning; support vector machine; decision tree

0 引言

随着我国机动车保有量逐年递增, 城市停车场建设已经难以满足机动车的增长速度, 造成“一位难

求”“绕树三匝无枝可依”等停车难问题。许多驾驶人员的出行时间浪费在交通堵塞或寻找停车位上, 这不但进一步加剧拥堵, 也造成出行效率低下与环境污染等问题。

收稿日期: 2019-02-19

通信作者: 谭文安 (1965-), 男, 湖北公安人, 教授, 博士, 主要研究方向为软件工程环境, 服务计算, 智能服务系统等。

E-mail: watan@sspu.edu.cn

基金项目: 国家自然科学基金 (61672022, 61272036), 上海第二工业大学研究生项目基金 (EGD17YJ0034), 上海第二工业大学重点学科 (XXKZD1604) 资助

针对这一日益严峻的城市问题,国内外许多学者进行了深入的研究与讨论。所提出的解决方案主要分为:基于传感器与通信网络的城市停车诱导信息系统建设和基于统计理论和机器学习等数据驱动方法的预测模型研究。

停车诱导系统主要是在停车场的各出入口通过泊位信息采集设备实时检测进出车辆,采集停车场车位变化数据,然后通过统一平台如电子告示牌等将车位实时变化信息告知驾驶员。比如 Hong 等^[1]提出了一个针对停车位管理服务的实时监测系统。但是,该方法主要针对具有固定车位信息的城市停车场,而对于诸如路边占道停车等开放空间的泊车情况却缺乏支持。而后者由于地点分散、路况复杂、灵活度大、通信能力差、多采用人工收费管理等原因,难以集中管理,不能很好地集成到现有的诱导系统中。

近年来伴随着机器学习与深度学习的浪潮,基于数据驱动的停车位预测方法快速发展,并经常作为停车诱导系统信息管理模块来进一步提高对于车位信息实时变化的处理能力。文献[2-4]中研究了基于 BP 神经网络的城市停车需求预测模型。文献[5-6]中研究了基于粒子群优化算法与小波神经网络方法结合的停车场有效停车位预测问题。前者主要利用粒子群算法优化小波神经网络的参数,后者主要通过选择小波函数重构预测时间序列,然后采用小波神经网络进行预测。龙东华等^[7]提出了基于主成分分析的 BP 神经网络停车需求预测模型,利用主成分分析法研究了影响停车需求的各个特征,以此简化了神经网络的输入。孙敏等^[8]研究了基于长、短时记忆网络的停车位预测问题,可以在预测时间为中、长周期内保持数据变化特征。Vlahogiannia 等^[9]提出了一个传感器网络与神经网络相结合的实时预测系统。这些基于神经网络模型的预测算法能够充分利用深度模型对于非线性问题的强拟合能力来处理复杂参数,限于神经网络自身的特点,其也存在着诸如训练开销大,可解释性差等诸多问题。针对深度学习算法的诸多问题,也有学者研究利用传统机器学习的方法进行停车预测, Fan 等^[10]研究了支持向量回归(support vector regression, SVR)分类方法预测停车位的问题;郑喆等^[11]研究了结合深度信念网络模型和 SVR 分类方法构建预测模型。同诱导系统一样,当前这些机器学习方法也主要针

对停车场车位的预测问题,难以处理路边占道停车等复杂情况。复杂的停车方式很难精确实时掌握目的地的车位信息;同时我们注意到大部分出行用户更关注的是有没有停车位可供使用,而非精确的停车位数量,因为停车位供需信息是实时变化的,在用户前往途中往往就会发生变化。

本文从模糊预测的角度来预测目的地的停车难度,而非预测精确的车位数量。用户根据给出的停车难度来规划出行方式。将某一位置的停车难度划分为“简单”与“困难”两个类别,建立一个基于集成学习的监督分类算法,对任意位置的停车难度情况进行预测,包括停车场与路边停车的情况,因此实用性较强。另一方面,在训练模型时采用了非神经网络的机器学习方法,与深度模型相比,其训练效率更高,模型的可解释性更好。

1 算法介绍

根据前文的描述,希望构建一个基于集成学习方法的算法进行停车难度预测。集成学习是指训练一组弱学习器,然后通过某种集成策略将各个子学习器进行结合,以此获得比单一学习器更加优越的泛化性能^[12]。使用支持向量机(support vector machine, SVM)^[13]与决策树(decision tree, DT)集成的方式训练该监督分类模型。首先,这两类算法的运行机制很好理解,且对训练数据中的噪声具有弹性。其次, SVM 引入核方法后可以是非线性学习,能够较好地处理高维数据的情况,便于进一步引入新的特征后保持算法的适应性。同时, DT 模型的结果可解释性好,很容易直观理解每个特定特征,也更容易验证模型的行为是否合理。将二者集成可以获得更稳健的模型与更理想的预测效果。

1.1 基于 SVM 的停车难度二分类模型

将停车难度映射为“简单”与“困难”,以此作为样本数据对应的类别空间 (x, y) , x 是由停车预测相关特征组成的训练样本, y 是样本对应的标签(停车是否困难,即 0 或 1),组成一个二元值。因此,给定样本数据 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, 其中 $x \in R^n$, $y \in \{0, 1\}$, m 为样本个数, n 为训练变量特征维数,任务转变为训练 SVM 二分类模型。考虑到实际问题中由噪声等因素引起的线性不可分情况,同时避免过拟合,通常可引入软间隔 SVM 与正

则化的概念, 于是需要解决以下优化问题:

$$\begin{aligned} \min_{\omega, b, \xi} & \frac{1}{2} \|\omega\|_2^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} & y_i(\omega^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \\ & i = 1, 2, \dots, m \end{aligned} \quad (1)$$

式中: ω 为分类超平面的系数; b 为偏置项; ξ 为松弛变量; C 为惩罚参数, 用于 ξ 与间隔最大化之间取得平衡。式 (1) 对应的拉格朗日函数为:

$$\begin{aligned} L(\omega, b, \xi, \alpha, \beta) = & \frac{1}{2} \|\omega\|_2^2 + C \sum_{i=1}^m \xi_i + \\ & \sum_{i=1}^m \alpha_i [(\omega^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i \end{aligned} \quad (2)$$

式中, 拉格朗日乘子 α, β 满足: $\alpha_i \geq 0, \beta_i \geq 0$ 。式 (2) 分别对 ω, b, ξ_i 求偏导, 并令偏导为零将结果代回原式 (1) 消去 ω, b, ξ , 得到式 (2) 对应的对偶问题:

$$\begin{aligned} \max_{\alpha} & \sum_{i=1}^m \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ & i, j = 1, 2, \dots, m \\ \text{s.t.} & \sum_{i=1}^m \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C; \quad i = 1, 2, \dots, m \end{aligned} \quad (3)$$

考虑到实际问题中样本在低维空间不可分的情况, 式 (2) 中原有 3 个松弛变量, 经过求导代换为单变量的优化问题。同时注意到式 (3) 中训练样本 x 仅出现在 $x_i^T x_j$ 一项之中, 可以考虑映射 $\varphi(x)$ 将样本数据映射到高维甚至无限维空间, 如果 $\varphi(x_i)^T \varphi(x_j)$ 仍满足内积特性, 则可以实现非线性 SVM 学习。此时上述优化问题可以转化为:

$$\begin{aligned} \max_{\alpha} & \sum_{i=1}^m \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ & i, j = 1, 2, \dots, m \end{aligned} \quad (4)$$

式中: $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, 称为核函数。常用的核有线性核、多项式核、高斯核等。引入核是为了避免讨论具体的映射形式。

1.2 基于 DT 的分类模型

DT 分类是一种具有树形结构的机器学习模型。根节点对应全体训练样本集合, 训练时从根节点开始选择某一特征对样本进行测试, 并根据结果将样

本分配到子节点上, 然后递归对子节点重复此“测试 - 分配”过程, 直至满足以下停止条件之一: ① 当前节点样本同属一个类别; ② 当前属性集为空; ③ 当前节点为空节点。最终算法学得一组“if-then”规则集合, 并用它来进行预测。

DT 学习的关键在于迭代过程中如何选择最优的属性来对当前节点进行划分, 即如何定量地度量每个属性对样本集合的划分效果。常用的算法如 ID3 算法选择以信息增益为准则, C4.5 算法是在 ID3 的基础上进一步采用信息增益率, 而 CART 则选择计算基尼 (Gini) 指数。其中使用属性 A 对集合 S 进行划分, 对应的信息增益可表示为:

$$\begin{aligned} \text{Gain}(S, A) = \\ \text{Entropy}(S) - \sum_{i=1}^J \frac{|S_i|}{|S|} \text{Entropy}(S_i) \end{aligned} \quad (5)$$

式中: J 表示由属性 A 对集合 S 进行划分可能得到的子集数目, 假定当前集合中第 k ($k = 1, 2, \dots, K$) 类样本所占比例为 p_k , 则 $\text{Entropy}(S) = -\sum_{k=1}^K p_k \log_2 p_k$ 表示集合 S 的信息熵, 其取值越小表示 S 的纯度越高。因此在划分时选择信息增益最大的属性, 以此获得纯度更高的子集。为了防止由于信息增益偏好取值数目多对结果产生不利的影响, 考虑计算信息增益率 $\frac{\text{Gain}(S, A)}{\text{SplitInfo}(S, A)}$ 。其中:

$$\text{SplitInfo}(S, A) = - \sum_{j=1}^J \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \quad (6)$$

式中: $\text{SplitInfo}(S, A)$ 称为属性 A 的“固有值”, 通常取值数目越多“固有值”越大; $|S_j|$ 表示分支 j 的样本数, 从高于平均信息增益率的属性中选择信息增益率最大的特征对节点进行划分。

Gini 指数可以表示为:

$$\text{Gini}(S) = \sum_{j=1}^J \frac{|S_j|}{|S|} \text{Gini}(S_j) \quad (7)$$

$\text{Gini}(S_j) = 1 - \sum_{k=1}^K p_k^2$ 表示集合 S_j 的不确定性, 本文选择 Gini 指数最小的属性来划分节点。

1.3 基于集成学习的预测模型

本研究在 Bagging 框架下将 SVM 与 DT 集成, 每轮迭代中从初始训练集中采样生成一个训练样本集, 然后在该采样集上训练一个 SVM 模型, 接着关

注该 SVM 所误分类的样本,并将这些样本记录在另一个集合中,称之为失误集。当失误集的样本数量累加到达某个阈值,比如当与采样集大小相同时,就用该样本集合训练一个 DT 模型。经过 T 轮训练,最终将得到 T 个 SVM 模型与 P 个 DT 模型。使用该模型进行分类预测时,基于简单的投票法集成 T 个 SVM 模型和 P 个 DT 模型的结果,当出现两个类别收到同样票数的情形时,则再次使用投票法取 P 个 DT 模型的结果作为输出(若出现同样的情况则随机选择一个类别输出)。算法的主要框架如下:

输入 训练集;
组件学习器算法: SVM, DT;
训练轮数: T 。

过程 令误分类集合 E 为空集。
Set $P = 1$;
For $t = 1, 2, \dots, T$ do
 $h_t = \text{SVM}(D, D_t)$, D_t 为采样子集, h_t 表示第 t 轮训练获得的子学习器;
 将误分类样本加入集合 E 中;
 If $|E| \geq |D_t|$:
 $f_p = \text{DT}(E)$;
 $P = P + 1$;
 令误分类样本集合 E 为空集;
 End if
End for

输出 $H(x) = \arg \max_{y \in Y} \sum_{t=1}^T \Pi(h_t(x) = y)$
 $F(x) = \arg \max_{y \in Y} \sum_{p=1}^P \Pi(f_p(x) = y)$

最终输出包括 $H(x)$ 表示的 SVM 集成分类器和 $F(x)$ 表示的 DT 集成分类器。

2 实验与结果分析

停车位的供需变化与居民的出行变化情况息息相关,因此收集了上海市区某地的停车相关统计数据,选择与居民的出行模式相关联的特征来训练模型。主要包括:

(1) 时空特征。如:一天当中的第几个小时,星期几,是否节假日;周边一定范围内停车场的数目;目的地的空间属性(商业区,居民区等);交通情况(拥堵/畅通)。

(2) 行为特征。通过收集的用户轨迹数据来标记目的地位置的停车难度,主要评判的规则是监测

用户在目的地周围是否出现“巡游”“兜圈子”找停车位的情况,一旦出现将其标记为“困难”,但是这样也可能误判一些情况,比如在私人车位停车时不会发生上述“巡游”“兜圈子”现象,但显然不应以此就认为该位置的车位充足。另一方面,对乘坐出租车或公共交通抵达的用户,其行为轨迹看起来像是停车位充裕,其实不然,这些误报都有可能误导机器学习系统。

(3) 组合特征。为了剔除误导因素的影响,需要考虑更强的特征,如用户预计到达的时间与实际到达时间的差异,如果两次时间点之间的差异很大则强烈暗示用户在寻找停车位时花费了较长时间,出现了停车难问题。另一方面,当一个位置出现停车难现象之后,车主往往只能选择将车最终停靠在目标位置周边的区域,因此停车位置距目标位置的距离也可以作为一个特征。

利用 Scikit-learn 提供的 SVM 和 DT 工具,以及 TensorFlow 和 Keras 提供的神经网络构建工具来进行模型训练。影响 SVM 模型性能的主要因素是核函数,因此选择不同的核函数训练多个 SVM-DT 模型,主要包括:

线性核

$$k(x_i, x_j) = x_i^T x_j \quad (8)$$

多项式核

$$k(x_i, x_j) = (x_i^T x_j)^d, \quad d \geq 1 \quad (9)$$

高斯核

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad \sigma > 0 \quad (10)$$

Sigmoid 核

$$k(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta), \quad \beta > 0; \theta < 0 \quad (11)$$

DT 模型主要是为了纠正 SVM 的误分类情况,因此训练时可适当放宽参数约束,选择默认设置及选择 Gini 指数与信息增益分别与 SVM 组合,结果如图 1 和 2 所示。可以直观地看到基于 Gini 指数的 DT 偏好特征 X_6 ,该特征对应实际停车位置与目标位置的实际距离,即泊车的空间分布特征。而基于信息增益的 DT 则更倾向于特征 X_7 ,该特征对应周边停车场个数,即目的地的空间属性特征。

使用常用的分类任务评价指标来度量模型在测试样本集合上的效果。对于二分类问题,用 T_P 表示

真实情况为正例而预测结果为正例的样本数, T_N 表示真实情况为反例而预测结果也为反例的样本数, F_N 表示真实情况为正例而预测结果为反例的样本数, F_P 表示真实情况为反例而预测结果却为正例的样本数。通用指标包括准确率、精度和召回率。

定义 A_{cc} 表示分类正确的样本数占样本综述的比例, 计算式如下:

$$A_{\text{cc}} = \frac{T_{\text{P}} + T_{\text{N}}}{T_{\text{P}} + F_{\text{N}} + F_{\text{P}} + T_{\text{N}}} \quad (12)$$

定义精度表示预测正确的正例占所有预测为正的比值, 计算式如下:

$$P = \frac{T_P}{T_P + F_P} \quad (13)$$

定义召回率表示预测正确的正类与实际为正类的比值, 公式如下:

$$R = \frac{T_P}{T_P + F_N} \quad (14)$$

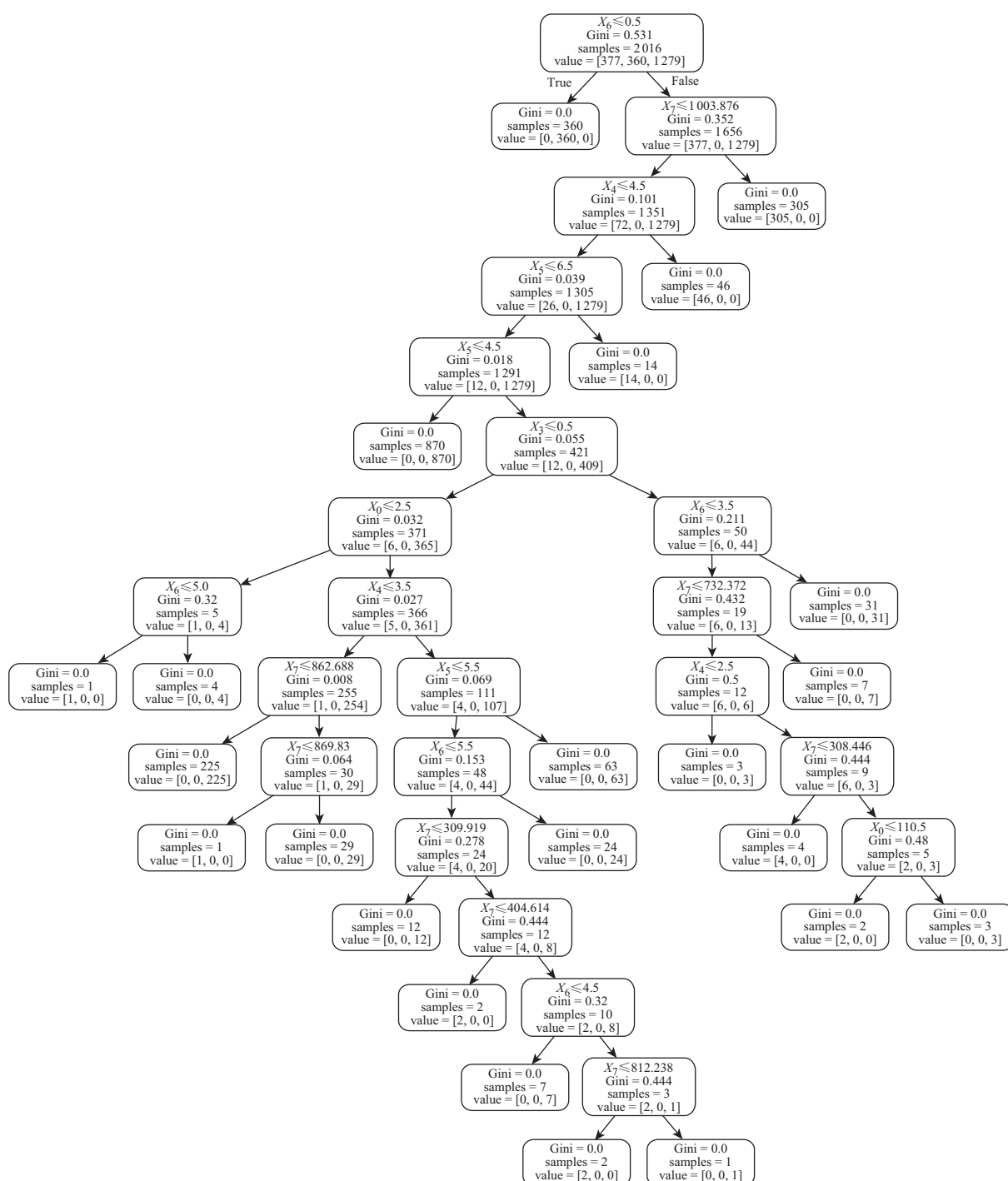


图 1 基于 Gini 指数的停车难度分类 DT 模型

Fig. 1 DT model of parking difficulty classification based on Gini coefficient

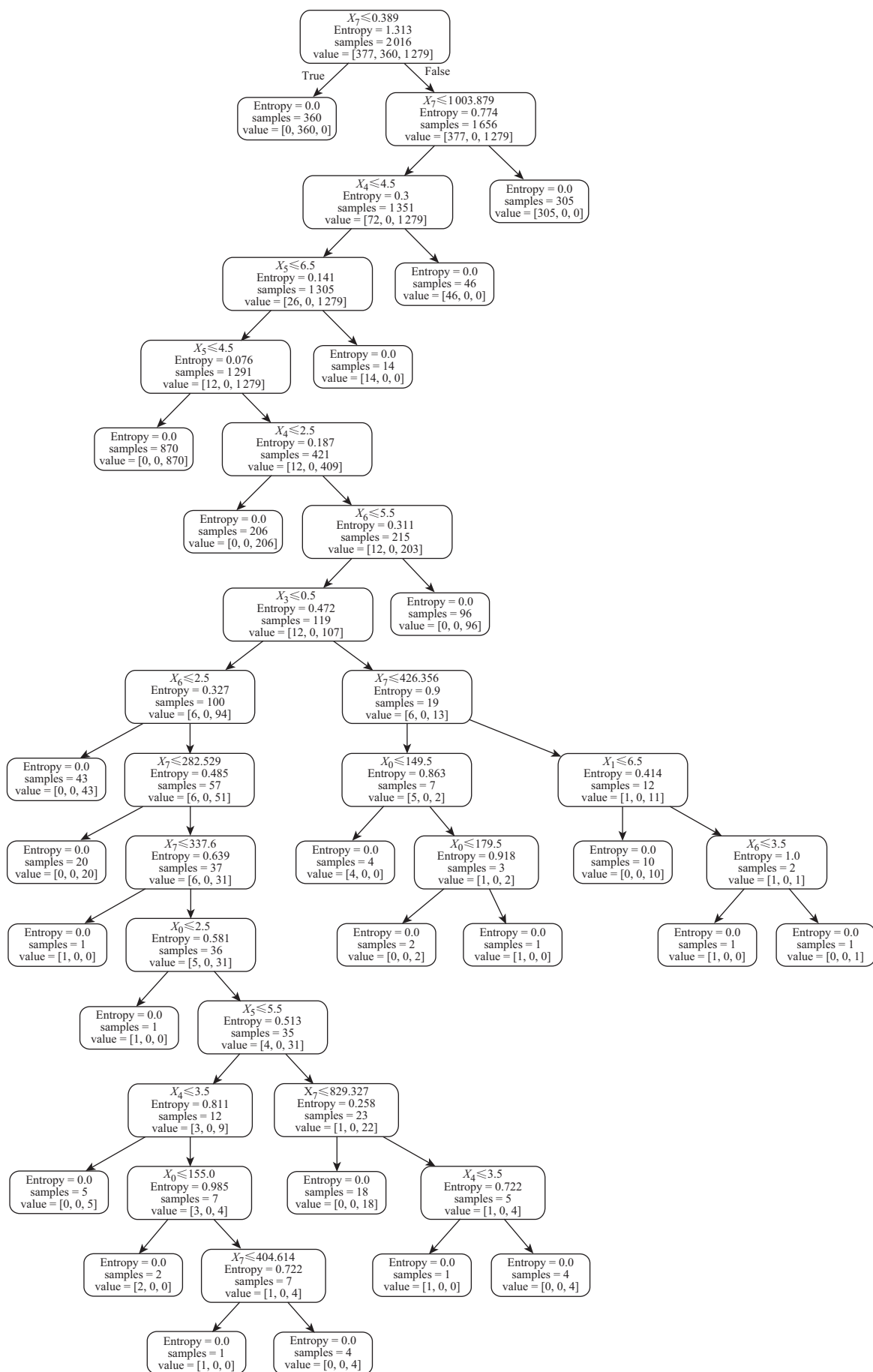


Fig. 2 DT model of parking difficulty classification based on information gain

对每一个组合分别计算上述3个指标, 模型评价指标如表1~3所示。可以直观地看到线性核模型即可取得较好的预测准确率。另一方面, 为了对比分析, 利用相同的训练数据训练了一个DT分类预测模型以及一个两层全连接神经网络模型(FC), 然后在相同测试集上分别计算模型正确率、精度和召回率, 结果见表4。可以看出, SVM和DT集成的模型在分类性能方面优于单一的SVM或DT模型; 另一方面SVM-DT模型的性能与FC相比, 具有更小的计算开销和更少的超参数设置, 这在训练时具有更强的实用意义。

表1 普通SVM模型评价指标

Tab. 1 General SVM model evaluation indexes

	线性核	多项式核	高斯核	Sigmoid核
正确率	0.918	0.926	0.922	0.762
精度	0.917	0.901	0.916	0.771
召回率	0.904	0.911	0.902	0.860

表2 基于信息增益的SVM-DT模型评价指标

Tab. 2 SVM-DT model evaluation indexes based on information gain

	线性核	多项式核	高斯核	Sigmoid核
正确率	0.988	0.956	0.972	0.783
精度	0.987	0.951	0.976	0.776
召回率	0.984	0.971	0.992	0.868

表3 基于Gini指数的SVM-DT模型评价指标

Tab. 3 SVM-DT model evaluation indexes based on Gini coefficient

	线性核	多项式核	高斯核	Sigmoid核
正确率	0.992	0.966	0.988	0.794
精度	0.994	0.971	0.986	0.857
召回率	0.991	0.981	0.992	0.893

表4 不同模型的分类性能

Tab. 4 Classification performance of different models

模型	正确率	精度	召回率
SVM	0.918	0.917	0.904
DT	0.958	0.955	0.946
SVM-DT (Gini 指数)	0.992	0.994	0.991
SVM-DT (Information gain)	0.988	0.987	0.984
FC	0.987	0.984	0.984

利用训练好的SVM-DT预测模型, 预测了1周的停车难度情况。训练时间从1天的零点开始, 间隔时间为5 min, 第1个5 min计为第1个时间片, 以此类推, 统计出周一至周日的各个时间片出现停车难现象的频率, 将真实数据与预测结果的统计情况进行比较, 结果如图3、4所示。真实数据和预测结果均显示停车难的现象集中出现在时间片70~110及200~250, 这两处时间片基本对应上午6:00~9:00和下午5:00~8:00, 也正是城市早晚高峰的时段, 由于上下班等因素导致机动车活跃数目激增, 路况变得更加复杂, 由此导致停车难现象出现的概率增大, 通过预测模型可以较好地捕捉到这一特点。

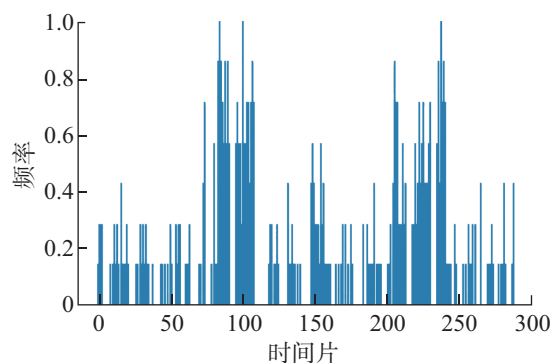


图3 1周时间实际停车情况统计柱状图

Fig. 3 Histogram of actual parking statistics during one week

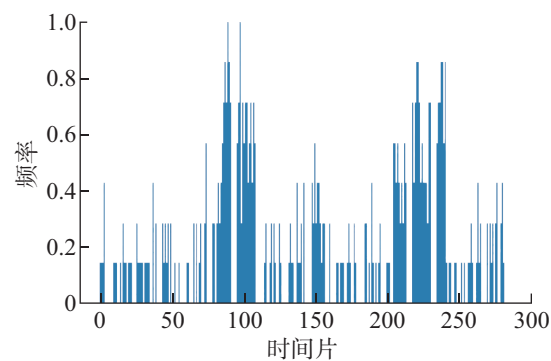


图4 1周时间停车情况预测柱状图

Fig. 4 Forecasting histogram of parking during one week

3 结 论

基于SVM与DT构建了一个集成学习的监督分类算法, 该方法可以较准确地预测任意位置的停车难度, 包括停车场与路边停车的情况, 实用性较强。训练模型时采用了非神经网络的机器学习方法, 相比深度模型来说训练效率更高, 模型的可解释性较好。该方法还可用于预测实时停车难度, 具体做法是在训练数据中加入待预测点历史停车难度信

息, 利用待预测时间点前一段时间的停车数据构造时序特征及利用时间滑动窗口从序列数据建立角度训练模型。

进一步的研究可以考虑结合众包集体智慧的方式收集用户反馈停车信息, 并将其作为训练数据, 比如利用移动终端追踪用户轨迹信息等方式, 以期模型的预测结果能够作为辅助用户出行的决策支持, 帮助用户合理规划出行方式, 进而缓解停车难问题, 营造智慧化的和谐城市。

参考文献:

- [1] HONG T P, SOH A C, JAAFAR H, et al. Real-time monitoring system for parking space management services [C]//Proceedings of 2013 IEEE Conference on Systems, Process and Control. Kuala Lumpur, Malaysia: IEEE, 2013: 149-153.
- [2] 高健夫, 王培松. 基于神经网络的城市停车需求预测模型 [J]. 交通科技与经济, 2015, 17(5): 5-9.
- [3] 高广银, 丁勇, 姜枫, 等. 基于 BP 神经网络的停车诱导泊位预测 [J]. 计算机系统应用, 2017, 26(1): 236-239.
- [4] 赵戊辰, 张玉茹. BP 神经网络用于停车场空余泊位的预测研究 [J]. 哈尔滨商业大学学报 (自然科学版), 2015, 31(1): 44-46.
- [5] 张军, 王书华, 杨正瓴. 基于粒子群优化的小波神经网络的停车位数量短时预测 [J]. 计算机应用与软件, 2015, 32(11): 66-68.
- [6] 季彦婕, 陈晓实, 王伟, 等. 基于小波变换和粒子群小波神经网络组合模型的有效停车泊位短时预测 [J]. 吉林大学学报 (工学版), 2016, 46(2): 399-405.
- [7] 龙东华, 邵毅明, 向红艳. 基于神经网络的停车需求预测模型及应用 [J]. 交通信息与安全, 2010, 28(5): 6-9.
- [8] 孙敏, 彭磊, 李慧云. 基于长短期记忆神经网络的可用停车位预测 [J]. 集成技术, 2018, 7(6): 39-48.
- [9] VLAHOIANNIA E I, KEPAPTSOGLLOU K, TSET-SOSA V, et al. A real-time parking prediction system for smart cities [J]. Journal of Intelligent Transportation Systems: Technology, 2015, 20(2): 192-204.
- [10] FAN J K, HU Q, TANG Z Z. Predicting vacant parking space availability: An SVR method with fruit fly optimization [J]. IET Intelligent Transport Systems, 2018, 12(10): 1414-1420.
- [11] 郑喆, 韩印. 基于 DBN-SVR 模型的短时有效停车泊位预测 [J]. 物流工程与管理, 2017, 39(6): 83-85.
- [12] ZHOU Z H. Ensemble methods: Foundations and algorithms [M]. London: Chapman and Hall, 2012.
- [13] BURGESS C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998(2): 121-167.