用于航班延误预测的集成式增量学习算法

王 丹[1], 王 萌[1], 王晓曦[2], 杨 萍[1]
(1. 北京工业大学信息学部, 北京 100124; 2. 国家电网管理学院, 北京 102200)

摘 要: 为持续高效地学习不断产生的航班运行信息,提高航班延误预测模型学习新到达数据的效率,采用集成学习思想,提出了一种基于分类与回归树(classification and regression tree,CART)的增量学习算法. 首先,将CART算法与Learn++算法结合实现了增量分类与回归树(incremental classification and regression tree,I-CART)算法;然后,进一步分析了基分类器间的区别和与精确度的关系,使用选择性集成算法来提高I-CART算法预测速率;最后,将该算法应用到航班延误预测中,增量地学习航班动态运行信息. 实验结果表明,该算法有效地提高了模型预测效果.

关键词: 航班延误; 分类与回归树(CART)算法; 增量学习; 集成学习; 选择性集成; 机器学习
中图分类号: U 461; TP 308 **文献标志码**: A **文章编号**: 0254 - 0037(2020)11 - 1239 - 07
doi: 10.11936/bjutxb2019030009

Ensemble of Incremental Learning Algorithm for Flight Delay Prediction

WANG Dan[1], WANG Meng[1], WANG Xiaoxi[2], YANG Ping[1]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. State Grid Management Institute, Beijing 102200, China)

Abstract: To continuously and efficiently learn the constantly generated flight information and improve the efficiency of flight delay prediction model to learn new arrival data, an incremental learning algorithm based on classification and regression tree (CART) was proposed by using ensemble learning ideas. First, incremental classification and regression tree (I-CART) incremental learning algorithms were implemented by combining CART algorithm with Learn++ algorithm. Then, based on the relationship between the difference and accuracy of basic classifiers, and the prediction rate of I-CART algorithm was improved by using the selective ensemble algorithm proposed in this paper. Finally, the incremental learning algorithm was applied to flight delay prediction. Results show that the incremental learning algorithm of flight dynamic information effectively improves the prediction performance of the model.

Key words: flight delay; classification and regression tree (CART) algorithm; incremental learning; ensemble learning; selective ensemble; machine learning

当今大数据时代,每时每刻都有新数据产生,其中往往包含着对后续事件预测有着重要影响的信息. 在机器学习领域,已经实现了多种用于预测的模型和算法. 因为能够持续、高效地学习新数据中包含的信息,并提高模型预测的准确率,所以增量学习技术受到了广泛关注[1]. 例如,在航班预测领域,

收稿日期: 2019-03-15
基金项目: 国家自然科学基金资助项目(61672505)
作者简介: 王 丹(1969—), 女, 教授, 主要从事计算机软件分析、分布式计算方面的研究, E-mail: neuwd@sina.com.cn

有专家提出“对于航班预测而言,获取最新的信息并加入到预测中,会为预测准确度带来更大的提升[2]”. 一般地,越是新到达的航班的运行信息对预测结果的影响也会越大,特别是那些刚刚获得的前航班的延误情况,起飞机场、目的机场的延误情况等. 由此可见,为航班预测模型提供具有增量学习能力的算法十分符合航班预测的实际需求.

在航班预测模型构建中,文献[3-5]均是利用贝叶斯网络来建立航班延误预测模型;文献[6]采用C4.5决策树构建航班预测模型,并与目前常用的贝叶斯分类方法进行对比,证明C4.5决策树算法能进一步提高预测精度;文献[7]利用随机森林算法预测空中交通延误情况. 决策树算法目前已得到广泛应用[8],其中,分类与回归树(classification and regression tree, CART)算法[9]是由Breiman学者提出的,它是一种基于Gini系数的二分法决策树分类算法. 由于它具有“稳定性-可塑性灾难”的特性,为学习新到达数据,只能抛弃现有决策树,利用历史数据集构造新的决策树,所以不具备增量学习的能力. ID5算法、ID5R算法等是具有增量学习功能的决策树算法,它们常常是对原始决策树算法中的节点中存储属性信息加以改造,再利用新获得的数据不断调整原有决策树,进而达到增量学习的目标[10]. 在增量学习过程中,这类算法需要不断地借助存储的信息调整决策树的结构,因此,需要额外的存储空间,若某些参数设置不当会引发过拟合,影响分类效果[11].

考虑到集成学习方法具有高泛化性、分类效果好和增量过程相对简单的特点,本文采用集成学习思想实现了一种增量分类与回归树(incremental classification and regression tree, I-CART)算法. 该算法利用集成学习思想使得CART算法克服了“稳定性-可塑性灾难”[12],实现了增量学习,具备了持续学习不断到达的新数据的能力且提高了学习新信息的效率,并采用选择性集成方法对其存在的集成分类器规模庞大的问题进行了改进. 该方法充分考虑了分类器的差异性与精确度对最终集成分类器的分类效果的影响[13],依据选择性集成理论来构建分类效果更优的基分类器子集. 同时,集成分类器规模显著减少,存储压力得以缓解,分类器预测速率得到提高.

1 I-CART 算法概述

1.1 Learn + + 算法

采用集成学习思想的Learn + +算法[12]具有适用性广、参数设置简单等优势,它将已经学习的历史数据作为基分类器加以保存且在学习新到达的数据时无须访问历史数据,当访问历史数据时也遗忘旧知识,以此方式通过不断学习新到达的数据相应生成多个基分类器,使集成分类器得到扩充.

1.2 I-CART 算法

本文将CART决策树作为基分类器,将CART算法与Learn + +算法相结合,采用集成学习思想设计并实现了I-CART增量学习算法,如算法1所示.

算法1 I-CART算法

输入:将原始数据集Data分为 K 份,标记为 $D(k)$, $k=1,2,\cdots,K$);从 $D(k)$ 中随机选取数据集 S, $S=[(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)]$, $i=1,2,\cdots,m$;基础学习算法为CART算法,子数据集迭代次数 T_k;

输出:集成分类器 C_{final}

---

for $k=1,2,\cdots,K$ do

初始化样本权重:

$w_1(i)=D_1(i)=1/m$, $i=1,2,\cdots,m$, $w_1(i)$ 表示第 i 个样本的权重, m 表示当前训练集样本个数, D 表示样本分布;

if $k>1$ 根据步骤5、6更新样本权重 $w_1(i)$ endif

for $t=1,2,\cdots,T_k$ do

1. 计算 D_t, $D_t(i)=w_t(i)=w_t(i)\Big/\sum_{i=1}^{m}w_t(i)$, 以此保证 D_t 是一个分布;
2. 根据样本分布 D_t,从数据集 S_t 中构建训练集 R_t 和测试集 E_t;
3. 调用CART算法,用训练集 R_t 训练得到决策树 c_t;
4. 计算决策树 c_t 在 $S_t=R_t+E_t$ 上的错误率 ε_t:

   $\varepsilon_t=\sum_{i:c_t(x_i)\neq y_i}D_t(i)$

   if $\varepsilon_t>1/2$ 删除该决策树, $t=t-1$, goto 步骤2 else 正规化错误率: $\beta_t=\varepsilon_t/(1-\varepsilon_t)$ endif
5. 将所有决策树 c_t 集成决策树 C_{count}:

   $C_{\text{count}}=\arg\max_{y\in Y}\sum_{n:c_n(x)=y}\lg(1/\beta_t)$

   计算 C_{count} 在训练集 R_t 的错误率 E_t:

   $E_t=\sum_{i:C_{\text{count}}(x_i)\neq y_i}D_t(i)$

   if $E_t>1/2$,删除 C_{count}, $t=t-1$, goto 步骤2
6. 标准化错误比值 $B_t=E_t/(1-E_t)$,更新每个样本的权重 $w_{t+1}(i)$

$$w_{t+1}(i)=w_t(i)\times\begin{cases}B_t, & C_{\text{count}}(x_i)=y_i\\ 1, & \text{其他}\end{cases}$$

endfor

endfor

使用多数加权集成所有决策树 c_t 得到最终集成分类器 C_{final}

$$C_{\text{final}}=\arg\max_{y\in Y}\sum_{k=1}^{K}\sum_{t:C_t(x)=y}\lg\frac{1}{\beta_t}$$

I-CART 算法由外层循环和内层循环组成，需要输入 K 个子数据集和迭代次数 T_k. 子数据集的遍历在外层循环完成，生成具有差异性的基分类器在内层循环完成，此过程类似于 Adaboost 算法[14].

首先，初始化每个子数据集的样本权重，设置为 $1/m$，m 表示子数据集中有 m 个样本. 之后进入内层循环，执行步骤 1，计算样本分布与样本权重，然后，针对样本分布 D_t 产生新的训练集. 样本权重决定训练集的构成，因为训练集由样本分布构造而成，而样本分布由样本权重产生. 再执行步骤 3，将训练集输入 CART 决策树算法，训练得到基分类器 c_t，并将其加入集成分类器达到不断学习新数据的目的，充分体现了集成的思想.

为构造具有差异性的基分类器，提高集成分类器的泛化性，需要更新样本权重，得到不同的训练集，更新规则见步骤 6. B_t 为标准化的错误率，取值区间为(0，1)，由集成分类器 $C_{\text{count}}(x_i)$ 得到正确分类的样本的权重与分布概率将会减小，其被选进训练集的概率会降低. 同理，被 $C_{\text{count}}(x_i)$ 错误分类的样本的权重会增加，其被选进训练集的概率会增大. 这一点体现在步骤 5 和步骤 6 中，就是说，当有新类别出现时，现有的分类器会产生错误的分类结果，该分类错误的样本的权重就会增大，受到学习算法的特殊关注. 通过以上分析，由于 I-CART算法不需要访问历史数据，对新数据学习的效率更具有优势.

步骤 5、6 也是使 I-CART 算法能够学习新类别的关键. 算法在迭代中通过增大分类错误样本的权重，增加其再次被选进训练集的概率，使得学习算法能够更多地关注被分类错误的样本. 因此，当有新的类别出现时，一定会被现有的分类器分类错误，成为分类错误的样本，该样本权重就会增大，使得学习算法关注这些具有新类别的样本，实现对新类别的学习. 从算法的执行来看，当有新数据加入时，应当返回到对子数据集的遍历即外层循环，首先需要执行的是每个子数据集的初始化部分，语句“if $k>1$，根据步骤 5、6 更新样本权重和样本分布”会对新的数据样本进行评价，根据对新数据集的表现更新样本分布，增大分类错误样本的权重. 其中，具有新类别的样本一定包含在其中，这些样本被选进训练集中，实现了对新类别的学习，使算法具备了增量学习能力.

2　基于基分类器差异性与精确度关系的选择性集成算法

2.1　算法设计思路

在 I-CART 算法执行过程中，由于不断生成新的基分类器加入到集成分类器中，所以不可避免地会形成规模庞大的集成分类器，导致预测速率下降，影响预测性能. 选择性集成思想，就是不断剔除性能不好的基分类器，或者只挑选性能优异的基分类器来提高分类器性能[15]. 因此，本文借鉴了选择性集成的思想，解决集成分类器规模不断庞大的问题. 那么，如何选择基分类器成为选择性集成算法设计的首要问题.

实验证明：集成分类器中的基分类器分类精度越高、基分类器之间的差异越大时，集成效果越明显，两者缺一不可. 据此，本文首先对分类器间差异性与精确度的关系进行了探究，然后提出了一种基于差异性与精确度关系的选择性集成算法，并将此算法加入到 I-CART 算法中. 在 I-CART 算法外层循环执行完成后 ，即所有的基分类器训练完成后再进行此选择性集成算法，选择出更优的基分类器子集构成最终的集成分类器. 该算法提高了分类器正确率，显著减小了集成分类器的规模和存储压力，集成分类器的预测性能也得到提高.

2.2　基分类器差异性与精确度关系探究

本文利用航班数据集和加州大学欧文机器学习知识库(UC Irvine machine learning repository，UCI)中的 Ionosphere 数据集进行 I-CART 算法关于基分类器间的差异性和精确度的探究实验，并选择 2 个基分类器间的 Kappa 系数 κ 作为差异性的度量指标，2 个基分类器间的平均分类误差率 $\bar{e}$ 作为精确度的度量指标.

Kappa 系数是一种衡量分类精度的指标，将 Kappa 系数作为 2 个分类器的差异性度量准则时，其取值范围为 $\kappa\leqslant 1$. $\kappa=1$ 时表示分类结果完全相同，不具有差异性；$\kappa=0$ 时表示分类结果相同，是偶

然产生的,具有差异性;$\kappa<0$ 表示分类器分类结果具有很大差异性. κ 值愈小,表征 2 个分类器的差异性愈大. Kappa 系数由算法 2 计算得到.

算法 2 2 个基分类器间的 Kappa 系数

输入:h 表示当前基分类器,$h_i(x_i)$ 表示 h_i 对样本 x_i 的分类,y_i 表示样本 x_i 对应的真实分类

输出:Kappa 系数 κ

1. $p_o=\dfrac{\sum_{a=1}^{L}C_{a,a}}{m}$,$a$ 代表分类,L 代表类别的数量,$C_{a,b}$表示 $h(x)=a,y=b$,m 代表样本个数

2. $p_e=\sum_{a=1}^{L}\left(\sum_{b=1}^{L}\frac{C_{a,b}}{m}\cdot\sum_{b=1}^{L}\frac{C_{b,a}}{m}\right)$

3. $\kappa=\dfrac{p_o-p_e}{1-p_e}$

图 1、2 分别为在航班数据集和 Ionosphere 数据集上基分类器的 κ 与分类误差率 e 关系图.

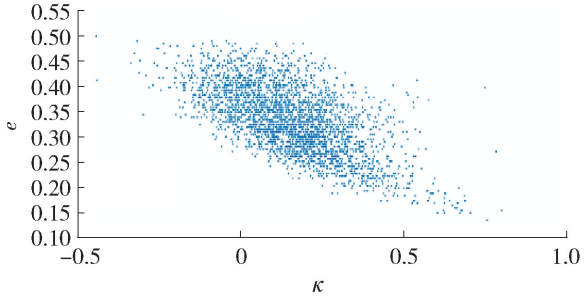

图 1 κ 与 e 在航班数据集上的关系

Fig. 1 Relation on the flight dataset between κ and e

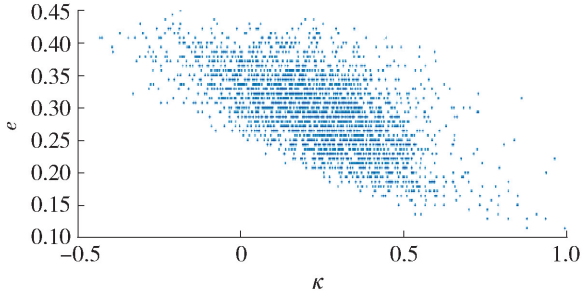

图 2 κ 与 e 在 Ionosphere 数据集上的关系

Fig. 2 Relation on the Ionosphere dataset between κ and e

从图 1 和图 2 可以看出,κ 减小时,基分类器的 e 变大,分类精确度变低. 根据选择性集成原理,应当尽可能集成那些差异性大且精确度高的基分类器. 然而,通过实验发现基分类器的差异性与精确度成反比,当选择差异性大的基分类器时,其精确度极低;当选择精确度高的基分类器时,它们彼此又不具备差异性.

2.3 选择性集成算法

本文根据基分类器两两间产生的 κ,结合 $\bar{e}$ 值,绘制得到图 3. 为兼顾基分类器间差异性与精确度,本文通过阈值对基分类器差异性加以控制,即从低于 Kappa 阈值的基分类器中选择具有最高精确度的基分类器,对它们进行集成,如图 3 所示,从最小的 $\bar{e}$ 值开始选择.

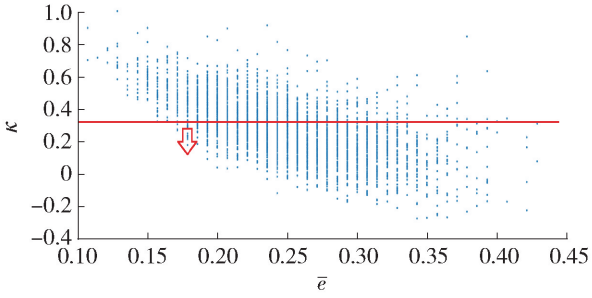

图 3 选择性集成算法示意图

Fig. 3 Schematic diagram of selective ensemble algorithm

算法中 threshold 表示差异性阈值,N 表示 I-CART算法中得到的基分类器的个数,S_N 表示选择基分类器的个数,$S_N>0\&\&S_N<N$,S_N 一般设置为 $\frac{1}{3}N\sim\frac{1}{2}N$,当 S_N 太接近 N 时,集成分类器的规模变化效果不明显,太小无法得到集成的目的. h_c 表示选择后的基分类器. h_i、h_j 表示不同的基分类器. 具体执行步骤描述如下.

1)不重复地计算所有基分类器两两间的 κ 与 $\bar{e}$,保存于 Kappa 数组和 Error 数组中.

2)threshold 阈值为所有 Kappa 系数的平均值. 通过设置平均值而不是某个固定阈值使其能更好地适应不同的情况.

3)按升序排序 Error 数组,产生基分类器序号并保存到数组 Index,确保精确率最高的基分类器得到优先选择.

4)从第一个 κ 值小于 threshold 的序号位置开始选择,节省了选择时间设计,不需要再依次比较.

5)从 begin 的位置开始选择基分类器,当 $\kappa<$ threshold 时,将对应的基分类器加入到集成分类器中,直到基分类器的个数达到 S_N,最终将选出的 S_N 个基分类器构成集成分类器. 具体实现过程如算法 3 所示.

算法 3　选择性集成算法

输入：指定选择基分类器的个数 S_N，开始选择位置 begin；

输出：选择出最优基分类器

1. 计算 h_i 与 h_j 之间的 κ，统计 $\bar{e}$

   for $i=1,2,\cdots,N$

   $e_i=h_i$ 在测试集上的误差率

   for $j=i+1,2,\cdots,N$

   $e_j=h_j$ 在测试集上的误差率

   $\mathrm{Error}(m)=(e_i+e_j)/2, m\in[1,\cdots,N^2]$

   按照算法 2 计算 h_i 与 h_j 之间的 κ

   $\mathrm{Kappa}[m]=\kappa, m\in[1,\cdots,N^2]$

   endfor

   endfor

2. 设置阈值 threshold = avg(Kappa)；

3. 排序(升序) Kappa 数组，并记录基分类器序号 Index，SORT( ) 表示排序函数

   Index = SORT(Error)

4. 设置开始选择性集成的位置

   begin = INDEX (NO. 1 → Kappa (Index (i)) < threshold)

5. 选择 S_N 个基分类器 $h_c, c=1,2,\cdots,S_N$

   h_c 表示选择出的基分类器，h_{Index} 表示序号为 Index 的基分类器

   while $c\leqslant S_N$ do

   if Kappa(Index(j)) < threshod

   $h_c=\mathrm{hIndex}, c=c+1$

3　实验设计与结果分析

3.1　实验数据集

本文利用美国交通运输统计局(Bureau of Transportation Statistics)网站的航班准点数据集(airline on-time performance data，AOTP)[16]，它包含了美国 1987 年至今的航班数据，并选择纽约 2018 年 2 月份与 5 月份产生的航班数据，共 105 159 条. 其中，删除了 2 505 个具有缺失值的样本后，将剩余的 102 654 个样本作为本文实验数据集. 对航班数据特征进行分析和数据特征选择后，本文选择了 13 个特征，其中，第 1 ~ 12 个特征为标识航班信息的重要属性，第 13 个为是否延误标志，准点到达目的地机场，其值为 0，晚点到达15 min 以上的，其值为 1. 具体特征描述如表 1 所示.

表 1　航班数据特征描述

Table 1　Description of flight data characteristics

特征字段	特征名称	特征描述
1	Quarter	季度
2	Month	月份
3	Day of Month	日期
4	Day of Week	星期
5	Origin Airport ID	起飞机场编号
6	Origin City Market ID	起飞城市编号
7	Dest Airport ID	目的机场编号
8	Dest City Market ID	目的城市编号
9	CRS_DEP_TIME	系统起飞时间
10	DET_TIME	实际起飞时间
11	CRS_ARR_TIME	系统到达时间
12	Distance	飞行距离
13	ARR_DEL15	到达延误标识

3.2　实验设计与结果分析

实验 1 的目标是将 I-CART 算法和 CART 算法的分类误差率进行统计，以验证 I-CART 集成式增量学习算法的分类性能. 实验 2 和实验 3 分别从增量次数与增量数量验证 I-CART 集成式增量学习算法能够高效地学习新到达的数据并具有良好的预测性能. 实验 4 用来验证分类器规模是否减少，预测效率是否能够提高.

3.2.1　实验 1

I-CART 算法参数见表 2，设置 I-CART 算法参数，K 表示子数据集个数，T_k 表示迭代次数，根据样本数量将 K 设置为 5、10、20，保证每个子数据集中样本个数相同. 表 3 是算法在不同样本数量时的分类误差率，性能提升率为分类误差率的差值，公式为

$$\text{性能提升率} = \text{CART 算法分类误差率} - \text{I-CART 算法分类误差率} \tag{1}$$

由表 3 可知，I-CART 算法与 CART 算法相比，在相同的训练数据量下 I-CART 算法的误差率普遍低于 CART 算法. 当样本数量为 10 万时，能达到 0.565% 的性能提升，证明 I-CART 算法当面临海量数据训练时分类性能比单分类器 CART 算法更优.

表 2 I-CART 算法参数

Table 2 Parameters of I-CART algorithm

样本数量/万	K	T_k
2.5	5	3
5.0	10	3
10.0	20	3

表 3 实验 1 分类误差率统计

Table 3 Statistics of classification error rate in experiment 1

样本数量/万	误差率/%		性能提升率/%
	CART 算法	I-CART 算法	
2.5	12.198	12.116	0.082
5.0	10.494	9.857	0.637
10.0	10.039	9.474	0.565

3.2.2 实验 2

在该实验中,将 5 000 作为固定增量,即每次增加 5 000 个新样本;增量次数分别设置为第 5、10、20 次,即当新数据在第 5 次、第 10 次、第 20 次到来时,学习新数据更新时需要的时间. 表 4 为实验 2 在不同增量次数下的时间统计表,耗时比计算方式为

$$\text{耗时比} = 100 \times \frac{\text{I-CART 算法执行时间}}{\text{CART 算法执行时间}} \quad (2)$$

表 4 实验 2 算法执行时间统计

Table 4 Statistics of algorithms execution time in experiment 2

增量次数	执行时间/s		耗时比/%
	CART 算法	I-CART 算法	
5	2.484 0	1.232 8	49.63
10	4.531 8	2.315 2	51.09
20	8.604 0	3.999 4	46.48

由表 4 可知,针对不同的增量次数, I-CART 算法的执行时间分别是 CART 算法的 49.63%、51.09%、46.48%,大大减少. 随着增量次数的增加,虽然 2 个算法的时间均有增多,但总体来看, I-CART 算法学习新数据的时间还是大大缩短.

3.2.3 实验 3

此实验从增量数量的角度验证 I-CART 算法能提高新数据的学习效率. 在该实验中,增量次数固定为 10 次,新数据每次以 2 000、5 000、10 000 个到达. 表 5 为时间统计表,耗时比计算方式见式(2).

表 5 实验 3 算法执行时间统计

Table 5 Statistics of algorithms execution time in experiment 3

增量数量	执行时间/s		耗时比/%
	CART 算法	I-CART 算法	
2 000	2.256 8	1.217 5	53.95
5 000	4.531 8	2.184 7	48.21
10 000	8.604 0	4.265 1	49.57

由表 5 可以看出, I-CART 算法执行时间分别是 CART 算法的 53.95%、48.21%、49.57%,约是 CART 算法所需时间的 49%,表明该算法大大缩短了学习新数据所需时间.

3.2.4 实验 4

为证明本文提出算法的有效性,将选择性集成算法加入 I-CART 算法进行选择性集成. 设置 I-CART算法子数据集个数 $K=30$,迭代次数 $T_k=3$,选择后基分类器的个数设置为 $S_{N_1}=20$、$S_{N_2}=30$、$S_{N_3}=45$,如表 6 所示. 表 7 为最终集成分类器中对应个数的分类误差率,表 8 为对应的预测时间.

表 6 实验 4 参数设置

Table 6 Parameters settings in experiment 4

数据集	K	T_k	训练样本数量	测试样本数量	基分类器数量
航班数据集	30	3	97 522	5 132	90

表 7 实验 4 分类误差率统计

Table 7 Statistics of classification error rate in experiment 4 %

数据集	I-CART	S_{N_1}	S_{N_2}	S_{N_3}
航班数据集	12.64	13.225	10.012	10.980

表 8 实验 4 预测时间统计

Table 8 Statistics of prediction time in experiment 4 s

数据集	I-CART	S_{N_1}	S_{N_2}	S_{N_3}
航班数据集	0.756 2	0.158 8	0.240 6	0.385 3

I-CART 算法表示,当未执行选择性集成算法时,生成的集成分类器中包含 90 个基分类器;S_{N_1}、S_{N_2}、S_{N_3}表示当执行选择性集成算法时,原分类器规模减小为原规模的 2/9、1/3、1/2. 可以看出,选择性集成算法在参数 S_{N_2}时产生最好的分类效果,集成分

类器规模减小为原来的 1/3，节省了约 67% 的存储空间，预测速率提高约 68%.

4　结论

1）基于集成学习思想的 I-CART 增量学习算法，使 CART 决策树算法具备了增量学习的能力，在处理不断新增的数据时能够增量地学习，显著提高了学习新数据的效率且拥有较好的分类性能.

2）综合考虑分类器间的差异性与精确度关系的选择性集成算法可确保分类性能提升，并显著减小集成分类器规模，提高预测速率.

3）为达到最好的分类效果，如何设定基分类器选择个数，通常需要多次尝试，是算法进一步改进的方向.

参考文献：

[1] LI Z, HOIEM D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2935-2947.

[2] 胡皓月. 航班延误预测的大数据方法研究[D]. 南京：南京航空航天大学, 2017.
HU H Y. Research on prediction of flights delay based on big data methods [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2017. (in Chinese)

[3] QIANYA L, LEI W, RONG F, et al. An analysis method for flight delays based on Bayesian network [C] // The 27th Chinese Control and Decision Conference. Piscataway: IEEE Society, 2015: 2561-2565.

[4] LIU Y J, MA S. Flight delay and delay propagation analysis based on bayesian network [C] // 2008 International Symposium on Knowledge Acquisition and Modeling. Piscataway: IEEE Society, 2008: 318-322.

[5] SHAO Q, LUO X, WU K, et al. The analysis of flight delay in airport based on Bayesian networks[J]. Science Technology and Engineering, 2012, 12(30): 8120-8124.

[6] 程华，李艳梅，罗谦，等. 基于 C4.5 决策树方法的到港航班延误预测问题研究[J]. 系统工程理论与实践, 2014, 34(增刊 1): 239-247.
CHEN H, LI Y M, LUO Q, et al. Study on delay prediction of arrival flight based on decision tree method of C4.5[J]. System Engineering Theory and Practice, 2014, 34(Suppl 1): 239-247. (in Chinese)

[7] REBOLLO J J, BALAKRISHNAN H. Characterization and prediction of air traffic delays[J]. Transportation Research Part C: Emerging Technologies, 2014, 44: 231-241.

[8] 刘俊龙. 高维数据下决策树的快速构造[D]. 北京：中国科学技术大学, 2017.
LIU J L. Fast construction of decision tree in high dimensional data[D]. Beijing: University of Science and Technology of China, 2017. (in Chinese)

[9] BREIMAN L I, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression tree [J]. Encyclopedia of Ecology, 2015, 57(3): 582-588.

[10] 孙静. 面向情报大数据的决策树增量学习算法研究[D]. 哈尔滨：哈尔滨工程大学, 2017.
SUN J. Research on incrementla learning algorithm of decision tree for intlligence large data [D]. Harbin: Harbin Engineering University. 2017. (in Chinese)

[11] 李华峰. 一种基于增量的集成学习方法的研究[D]. 哈尔滨：哈尔滨工程大学, 2016.
LI H F. Reseach on ensemble learning method based on incremental learning [D]. Harbin: Harbin Engineering University, 2016. (in Chinese)

[12] POLIKAR R, BYORICK J, KRAUSE S, et al. Learn++: a classifier independent incremental learning algorithm for supervised neural networks[C] // Proceedings of the 2002 International Joint Conference on Neural Networks. Piscataway: IEEE, 2002: 1742-1747.

[13] 张亮，李智星，王进. 基于动态权重的 AdaBoost 算法研究[J]. 计算机应用研究, 2017(11): 38-41.
ZHANG L, LI Z X, WANG J. Research on dynamic weights based AdaBoost [J]. Application Research of Computers, 2017(11): 38-41. (in Chinese)

[14] RANDHAWA K, LOO C K, SEERA M, et al. Credit card fraud detection using AdaBoost and majority voting [J]. IEEE Access, 2018, 6: 14277-14284.

[15] 倪志伟，张琛，倪丽萍. 基于萤火虫群优化算法的选择性集成雾霾天气预测方法[J]. 模式识别与人工智能, 2016, 29(2): 143-153.
NI Z W, ZHANG S, NI L P. Haze forecast method of selective ensemble based on glowworm swarm optimization algorithm [J]. Pattern Recognition and Artificial Intelligence, 2016, 29(2): 143-153. (in Chinese)

[16] BELCASTREO L, MAROZZO F, TALIA D, et al. Using scalable data mining for predicting flight delays[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2016, 8(1): 1-20.

（责任编辑　梁　洁）