文章编号：2095-2163(2020)07-0155-03 中图分类号：TP311 文献标志码：A

基于 CART 决策树算法的犯罪类型决策研究

史 楠[1]，韩丽娜[1,2]，李振兴[1]
（1 西安石油大学 计算机学院，西安 710065；2 陕西学前师范学院，西安 710100）

摘 要：大数据时代，犯罪案件的数据多样而繁杂，选取关键因素进行准确判决显得越来越重要。本文以美国犯罪记录集作为训练数据，以 CART 算法为工具，构建了基于决策树算法的犯罪类型决策模型，并得出决定犯罪类型最主要的因素是作案工具的结论。同时，此模型具有一定的普适性，可应用到我国的犯罪类型决策中。

关键词：犯罪类型；CART 算法；大数据

Research on crime type decision based on CART decision tree algorithm

SHI Nan[1], HAN Lina[1,2], LI Zhenxing[1]
(1 School of computer science, Xi′an Shiyou University, Xi′an 710065, China;
2 Shannxi Xueqian Normal University, Xi′an, 710100, China)

【Abstract】 In the era of big data, the data of criminal cases are diverse and complicated, so it is important to select key factors for accurate judgment. How to make decision on crime types based on big data is a topic of great interest to researchers. This article studies the America criminal record set as the training data, then CART algorithm is used to construct the criminal type auxiliary decision-making model based on decision tree algorithm, and concluded that the most important factor is the tool type, which is determinant for crime type. At the same time, the universal applicability of the model shows that it can be applied to our country′s criminal type auxiliary decision-making.

【Key words】 Crime prediction; CART algorithm; Big data

0 引 言

现如今，犯罪方式多样且犯罪数据繁杂，案件决策会因此发生误判，所以需要大数据的辅助[1]。针对这类问题，梳理了相关学者的研究。发现犯罪类型决策相关的研究成果较少。决策树（decision tree）算法是一种基于树结构来进行决策的算法，典型决策树算法有 C4.5 和 CART 算法，由于 C4.5 决策树算法采用信息增益比来选择特征，含有大量耗时的对数和多叉树运算，使得运算强度和效率极低，而 CART 算法采用基尼系数代替熵模型，将多叉树改为二叉树，能够极大的提高运算效率[2]。因此，本文采用 CART 决策树算法进行犯罪类型决策研究。

1 CART 决策树算法及其相关概念

CART 算法（classification and regression tree）是一种分类与回归算法，是在给定输入随机变量 X 的条件下，输出随机变量 Y 的条件概率分布的学习方法，决策树递归的二分每一个特征，最终得到决策树[3]。

对于分类，CART 采用 $Gini$ 系数最小化准则来进行特征选择，基尼指数 $Gini$ 可表示为公式（1）：

$$Gini(p) = 2p(1-p). \quad (1)$$

其中，p 为第一个样本输出概率。

CART 算法生成步骤如下[4]：

（1）设结点的训练数据集为 S，对每一个特征 A，对其可能取的每一组值，将 S 分割成 S_l 和 S_r 两部分，计算在特征 A 的条件下，集合 S 的 $Gini$ 指数的定义为公式（2）：

$$Gini(S,A) = \frac{S_1}{S}Gini(S_1) + \frac{S_2}{S}Gini(S_2). \quad (2)$$

（2）在所有可能的划分中，选择 $Gini$ 系数最小的组合作为最优特征与最优切分点，从现结点生成两个子结点，将训练数据集依特征分配到两个子结点中去；

（3）对两个子结点递归地调用步骤（1）~（2），直至没有更多特征；

（4）生成 CART 决策树；

（5）模型调优，通过交叉验证和网格搜索，对模

作者简介：史 楠（1995-），男，硕士研究生，主要研究方向：图像处理、数据挖掘；韩丽娜（1976-），女，博士，教授，硕士生导师，CCF 会员，主要研究方向：数据挖掘、图像处理；李振兴（1994-），男，硕士研究生，主要研究方向：图像处理、数据挖掘。

通讯作者：史楠 Email：1131239464@qq.com

收稿日期：2020-03-30

型的稳定性及超参数进行调优。

2 应用 CART 决策树决策犯罪类型

2.1 数据准备与预处理

本文研究中的数据是美国国家犯罪记录(本文研究数据信息来源于 Kaggle 数据科学竞赛公开数据集 Homicide Reports),其中包含 Record ID、Crime Type、Victim Race、Relationship、Weapon、Crime Type 等 24 个字段特征,共638 454条样本。

根据研究内容的重点倾向和时效性,仅保留部分特征,666 条样本,数据预处理结果见表 1。

表 1 数据预处理结果

Tab. 1 Results after data preprocessing

	Victim_Race	Perpetrator Race	Relationship	Weapon	Crime_Type
0	White	White	Unknown	Fire	Manslaughter
1	Black	Native	Stranger	Blunt	Manslaughter
2	White	Asian/Pacific	Acquaintance	Knife	Manslaughter
…	…	…	…	…	…
665	White	White	Stepfather	Blunt	Murder

基于对以上数据的处理,本文研究将犯罪类型聚焦在谋杀和误杀之间,通过罪犯种族、受害人种族、作案武器及两者的关系 4 种特征来决策犯罪类型。各字段特征的变量值及其出现次数见表 2。

表 2 各字段特征的变量值及其出现次数

Tab. 2 Variable values of the characteristics of each field and the number of occurrences

Victim_Race	Perpetrator Race	Relationship	Weapon	Crime_Type
Asian/Pacific(14)	Asian/Pacific(14)	Acquaintance(75)	Blunt Object(71)	Manslaughter(317)
Black(177)	Black(200)	Boyfriend(9)	Drowning(7)	Murder(349)
Native(16)	Native(10)	Boy/Girlfriend(1)	Drugs(28)	
Unknown(69)	Unknown(10)	Brother(7)	Fall(1)	
White(390)	White(432)	Daughter(35)	Fire(6)	
		…	…	
		In-Law(4)	Suffocat(13)	
		Wife(20)		

生成决策树之前,将每个特征的取值由类别映射为数值,映射结果见表 3。

表 3 特征映射表

Tab. 3 Feature map

	Victim Race	Perpetrator Race	Relationship	Weapon
0	4	4	21	4
1	1	2	20	0
2	4	0	0	8
…	…	…	…	…
665	4	4	18	0

将数据集拆分成训练集和测试集,拆分结果为 499 条训练数据和 167 条测试数据。

2.2 CART 算法构造决策树

CART 的划分思想:假设特征 A 有 n 个离散值,每次将其中一个特征分为一类,其它 $n-1$ 个特征分为另一类。依照这个标准遍历所有的分类情况,计算每种分类下的基尼指数,最后选择值最小的一个作为最终的特征划分。

对于 Victim Race 特征,第一种分类情况的 $Gini$ 系数计算过程:

其中,S_l 表示 Victim Race 取值为{0}的分组,S_r 表示 Victim Race 取值为{1,2,3,4}的分组,由于决策树的构造依据训练数据集,故 Victim Race 的统计数据见表 4。

表 4 Victim Race 统计数据表

Tab. 4 Victim Race Statistics Table

Crime Type\Victim Race	S_l	S_r
Murder	9	246
Manslaughter	4	240

由公式(1),左右子树的基尼系数如下:

$$Gini_{(S_l)} = 2 * \frac{9}{13} * (1 - \frac{9}{13}) = 0.426\ 035$$

$$Gini_{(S_r)} = 2 * \frac{246}{486} * (1 - \frac{246}{486}) = 0.499\ 923$$

由公式(2),此种划分的基尼系数如下:

$$Gini(S, V) = \frac{13}{499} Gini(S_l) + \frac{486}{499} Gini(S_r) = 0.497\ 998$$

同理,遍历所有特征的不同划分,计算第一次决策的基尼系数见表 5。

分析对比表 5 数据,取 $Gini$ 系数最小的分组作为划分结果,即以特征列 Weapon 为最优特征列;以 13.5 为最优切分点,第一次决策结果如图 1 所示。

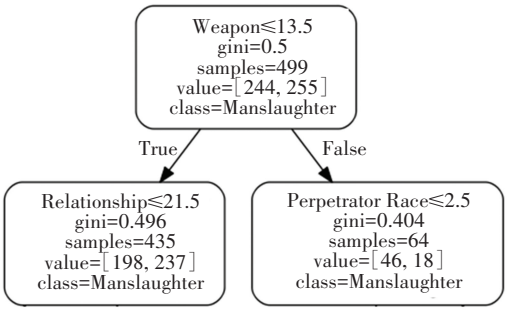

图 1　第一次决策结果

Fig. 1　Results of the first decision

接下来，分别对左子树和右子树进行决策，划分到没有特征值为止。同时，为了防止决策树模型过拟合，提高模型准确率，采用网格搜索寻找最优超参数。为了提高模型的稳定性，进行交叉验证，最终生成 CART 决策树模型（前四层），如图 2 所示。

表 5　第一次决策基尼系数表

Tab. 5　Gini coefficient table for the first decision

Victim_Race	Perpetrator Race	Relationship	Weapon
0:0.497 998	0:0.493 748	0:0.498 142	0:0.497 111
1:0.499 060	1:0.499 750	1:0.499 167	1:0.496 870
2:0.499 167	2:0.498 867	2:0.498 796	2:0.491 424
3:0.499 684	3:0.495 710	3:0.499 507	3:0.491 424
4:0.497 541	4:0.494 171	4:0.499 655	4:0.499 755
		…	…
		22:0.485 697	14:0.484 221

2.3　模型评估

导入 167 条测试数据检验模型，借助交叉验证和超参数调优，准确率达到了 71.25%。说明该模型具有一定的参考价值，模型科学性的辅助，会大幅度提升犯罪类型决策的正确率。结果表明，犯罪类型决策的最重要的因素是作案工具，作案工具越倾向于预备型（如：枪、刀具），犯罪类型越可能是谋杀。其次，双发关系对决策的影响也很重要，因而在决策中应着重考虑作案工具和双方关系。

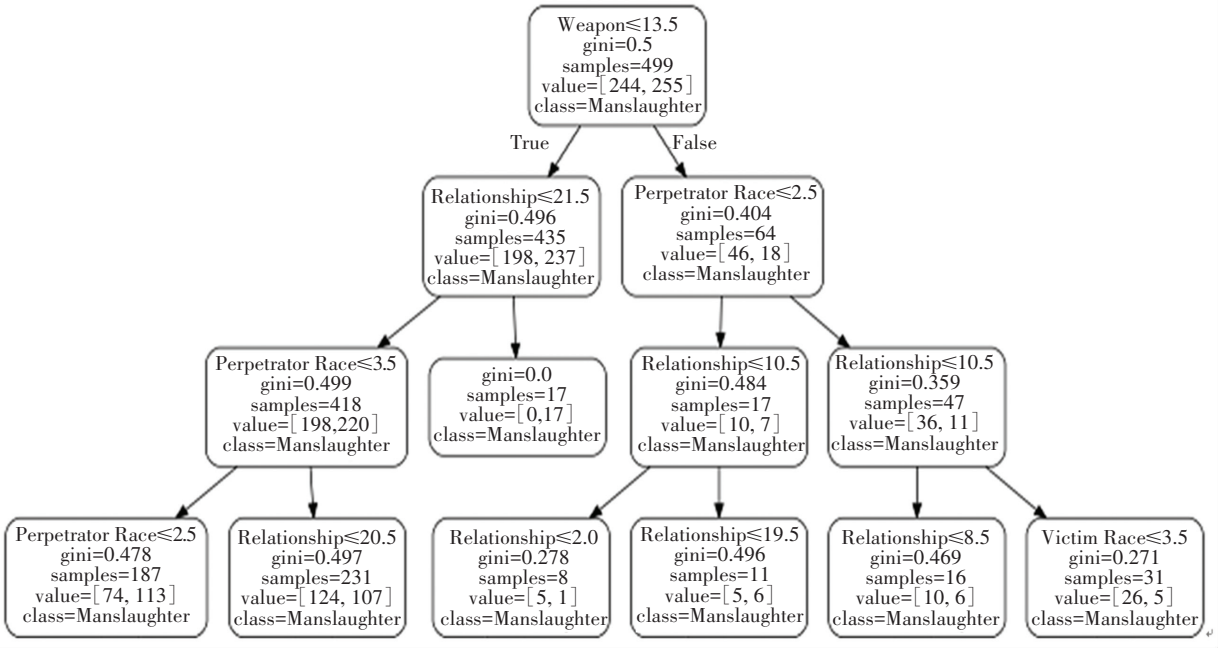

图 2　CART 决策树模型

Fig. 2　CART decision tree model

3　结束语

本文采用了 CART 算法辅助决策犯罪类型，通过对数据的处理和决策树模型的建立，并采用交叉验证和网格搜索优化模型。最终的实验结果表明，基于决策树算法的犯罪类型辅助决策，有一定的科学依据，能够协助执法人员在决策中，做出更科学、公正的选择。在排除国家文化种族差异的情况下，此模型可适用于我国的犯罪类型决策中。由于数据集采集的年份较早，一部分相关的字段特征没能顾及，没能够对更多变量（作案地点、心理性格特征）之间的因果关系进行进一步考察。因此，需要在今后的研究中，涵盖更广的目标范围，应用更多、更全面的数据，来预测犯罪类型。

参考文献

[1] 赵永军. 犯罪决策的心理学研究[D]. 河南大学，2003.

[2] 倪海鸥.决策树算法研究综述[J].宁波广播电视大学学报，2008(3)：113-115

[3] Lin Shuqiong. A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes[J]. Multivariate behavioral research，2019：1-15.

[4] 张亮，宁芊. CART 决策树的两种改进及应用[J]. 计算机工程与设计，2015，36(5)：1209-1213.