

基于 C4.5 决策树的视频车辆车型分类算法

卞建勇¹, 徐建闽²

(1 东莞职业技术学院 电子工程系, 广东 东莞 523808; 2 华南理工大学, 广东 广州 510640)

摘 要: 为了克服 BP 神经网络、支持向量机、K-近邻和贝叶斯等分类器通常存在训练时间长、编程和运算复杂、难以满足实时性等问题, 提出了基于 C4.5 决策树的视频车辆车型分类算法, 介绍了层次化分类器设计方法, 并以二分“小型车”为“小客车”和“小货车”两类的 C4.5 决策树的构造为例, 对视频车辆车型分类 C4.5 决策树的构造过程作了详细介绍, 最后通过实验证明了该算法进行车型分类的有效性。

关键词: C4.5 决策树; 层次化分类器; 车型分类

中图分类号: TP391.4

文献标识码: A

文章编号: 1000-7180(2017)05-0104-06

The Video Vehicle Classification Algorithm Based on C4.5 Decision Trees

BIAN Jian-yong¹, XU Jian-min²

(1 Department of Electronic Engineering, Dongguan Polytechnic, Dongguan 523808, China;

2 South China University of Technology, Guangzhou 510640, China)

Abstract: In order to overcome the BP neural network, support vector machine (SVM), k nearest neighbor and Bayesian classification's shortcomings which are long training time, complicated programming and operation, bad real-time performance. In this paper, the video vehicle classification algorithm based on C4.5 decision trees is proposed, the design method of the hierarchical classifier is introduced, and the construction process of C4.5 decision trees as an example which was used to divide light-duty vehicle into minibus and van is introduced in detail. Finally, the experimental results demonstrate the effectiveness of the video vehicle classification algorithm based on C4.5 decision trees.

Key words: C4.5 decision trees; hierarchical classifier; vehicle classification

1 引言

基于视频处理的车型分类通常先设法获取待分类车辆的分类特征, 再选择合适的分类器进行分类^[1-9], 目前常见的分类器有 BP 神经网络^[1-3]、支持向量机^[4-7]、K-近邻^[8]和贝叶斯^[9]等。BP 神经网络分类器容易出现局部最小和维数灾等问题; 支持向量机运算比较复杂; K-近邻主要用于线性分类; 贝叶斯分类器对样本数量要求较高。常见分类器大都存在训练时间长, 编程和运算复杂, 难以满足实时性等问题。

决策树(decision tree)^[10]是一种类似流程图的

树结构, 其中每一个内部结点表示在一个属性特征上的测试, 每一个分支代表一个测试输出, 每一个叶子结点代表类或者类的分布。学习得到的决策树能够被表示为多个 if-then 的规则, 很容易编程实现, 运算复杂度低, 能够满足实时性的要求, 已经被成功应用到医疗诊断、贷款信用风险的评估等多个领域。因此, 本文对 C4.5 决策树算法进行扩展并将其应用于视频车辆的车型分类。

2 C4.5 决策树的算法过程

C4.5 算法^[11-12]采用自顶向下的贪婪搜索遍历

可能的决策树空间. 该算法的构造过程从“哪一个特征将在树的根结点被测试?”这个问题开始时,分类能力最好的特征将被选作树的根结点,然后为该根结点特征的每个可能值产生一个分支,并将训练样本排列到适当的分支(即样本的属性特征值对应的分支)之下,然后重复整个过程,用每个分支结点关联的训练样本来选取在该结点被测试的最佳特征. 假定 Examples 代表训练样本集,Characters 代表属性特征集,Target_classes 代表要分类的目标类别,则 C4.5 决策树具体算法过程如下所示.

Step1 如果样本集 Examples 都为正,则返回 label=+ 的单结点的树.

Step2 如果样本集 Examples 都为负,则返回 label=- 的单结点的树.

Step3 否则执行以下步骤,直至最后一个特征被测试.

(1)使用信息增益比率找出属性特征集 Characters 中分类样本集 Examples 能力最好的属性特征 A.

(2)将属性特征 A 作为决策树的根结点 Root.

(3)对属性特征 A 的每一个可能值 v_i .

①在根结点 Root 下增加一个新的分支对应测试 v_i ;

②令 Examples $_{v_i}$ 为样本集 Examples 中满足值 v_i 的子集:

- 如果 Examples $_{v_i}$ 都为正,则在新的分支下返回 label=+ 的叶子;
- 如果 Examples $_{v_i}$ 都为负,则在新的分支下返回 label=- 的叶子;
- 如果 Examples $_{v_i}$ 有正有负,则在新的分支下增加一个子树.

Step4 使用规则后修剪(rule post-pruning)方法对生成的树进行修剪;

Step5 返回经过修剪后最终生成的树.

由以上算法过程可知,在 C4.5 决策树的构造过程中,最为关键的两个步骤就是:基于信息增益比率的决策树结点属性特征选择以及决策树的规则后修剪.

3 基于信息增益比率的 C4.5 决策树结点属性特征选择

将 C4.5 将决策树看作是含有分类信息的信息源,假设样本集为 S ,样本类别标号为 $C_i, i = 1, \dots, n$,则该样本集的信息熵(entropy)为

$$Entropy(S) = - \sum_{i=1}^n \left[\frac{num(C_i, S)}{|S|} \right] \times \log_2 \frac{num(C_i, S)}{|S|} \quad (1)$$

式中, $num(C_i, S)$ 表示样本集 S 中属于类 C_i 的样本数目, $|S|$ 表示样本集 S 的样本总数. 熵刻画了样本集的纯度(purity).

有了熵作为衡量训练样本集纯度的标准,则属性特征分类训练样本能力的度量标准——信息增益(information gain)就是使用属性特征分割样本集而导致的期望熵的降低. 假设属性特征为 A , 样本集为 S , 则信息增益 $Gain(S, A)$ 为

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

式中, $Values(A)$ 是属性特征 A 所有可能值的集合, S_v 是样本集 S 中属性特征 A 的值为 v 的子集(即 $S_v = \{s \in S \mid A(s) = v\}$). 式(2)中的第一项是样本集 S 的熵,第二项是用属性特征 A 分类样本集 S 后熵的期望值,第二项描述的期望熵是样本集 S 中属性特征 A 的每个子集的熵的加权和,其加权值为属于子集 S_v 的样本数占原样本集 S 样本数的比例 $|S_v| / |S|$. $Gain(S, A)$ 描述的就是给定属性特征 A 的值而得到的关于目标函数值的信息.

C4.5 算法引入了分裂信息(split information)以及信息增益比率的概念. C4.5 使用分裂信息来衡量属性特征分裂数据的广度和均匀性. 信息增益中对熵的使用只考虑了样本集 S 关于学习到的树要预测的目标类别属性值的熵,分裂信息实际上描述的就是样本集 S 关于属性特征 A 的各值的熵. 分裂信息可表示为

$$SplitInformation(S, A) = - \sum_{i=1}^c \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (3)$$

式中, S_1 到 S_c 是 c 个值的属性特征 A 分割样本集 S 而形成的 c 个样本子集.

通过信息熵、信息增益以及分裂信息可以获得信息增益比率:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)} \quad (4)$$

信息增益比率表示了由分支产生的有用信息的比率,因此,信息增益比率越大,分支包含的有用信息就越多.

C4.5 算法就是使用信息增益比率作为在增长树的每一步中选取最佳属性特征的度量标准,即使用信息增益比率来评价可能的分支点,获得最大信息增益比率的点被选中作为新的结点,然后算法再

在这个新的结点上生成新的子树.

4 决策树的规则后修剪

C4.5 决策树的构造过程中另一个重要的关键步骤就是使用规则后修剪(rule post-pruning)方法对初步生成的可能由于噪声或巧合已经过度拟合的决策树进行修剪,以避免过度拟合的风险. 假定从决策树根结点到叶子结点路径上的每一个属性特征的测试称为一个规则的先行词,即前件(preconditions),叶子结点的分类称为规则的结论. 则规则后修剪主要包括以下步骤:

- (1)将生成的可能过度拟合的决策树转化为等价的规则集合,方法是为从决策树根结点到叶子结点的每一条路径创建一条规则;
- (2)通过删除任何可能导致估计精度提高的先行词来修剪(泛化)每一条规则;
- (3)按照修剪过的规则的分类精度对其进行排序,并按照排序的顺序应用这些规则来分类后来的实例.

在 C4.5 决策树的具体构造过程中,一般将样本集的三分之二作为前期决策树生成的训练样本集合,另外三分之一作为规则后修剪过程中的验证样本集合.

5 视频车辆车型分类 C4.5 决策树的构造

本文采用层次化设计方法将 C4.5 决策树算法扩展到多类车型分类识别,即:按照二分类的处理方法,构造 C4.5 决策树,先将车辆二分成“大型及拖挂车”和“小型及摩托车”两类,然后再构造 C4.5 决策树将“大型及拖挂车”二分成“拖挂车”和“大型车”两类,接着构造 C4.5 决策树将“小型及摩托车”二分成“摩托车”和“小型车”两类,最后构造 C4.5 决策树将“大型车”二分成“大客车”和“大货车”两类,构造 C4.5 决策树将“小型车”二分成“小客车”和“小货车”两类. 整个分类过程共需要构造 5 个不同的 C4.5 决策树,并将车辆分类为摩托、小客、小货、大客、大货及拖挂六类.

以下以二分“小型车”为“小客车”和“小货车”两类的 C4.5 决策树的构造为例,对视频车辆车型分类 C4.5 决策树的构造作详细介绍.

5.1 基于信息增益比率的车型分类决策树结点属性特征选择

对图 1 和图 4 所示采用图像分割算法所获得的小货和小客真实车辆图像进行水平和垂直方向灰度投影,分别获得图 2 和图 5 所示的小货和小客水平

方向灰度投影直方图及图 3 和图 6 所示的小货和小客垂直方向灰度投影直方图.

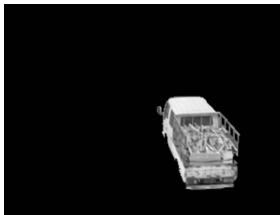


图 1 小货原始灰度图



图 2 小货水平灰度投影图



图 3 小货垂直灰度投影图



图 4 小客原始灰度图



图 5 小客水平灰度投影图



图 6 小客垂直灰度投影图

通过计算可以从投影直方图获得以下八类几何特征:

- (1)垂直方向灰度投影直方图曲线的宽度,简称车宽;
- (2)水平方向灰度投影直方图曲线的宽度,简称车长;
- (3)水平方向灰度投影直方图曲线的宽度与垂直方向灰度投影直方图曲线的宽度的比值,简称长宽比;
- (4)水平方向灰度投影直方图曲线中斜率与车道标准线斜率比较接近的曲线段的宽度,简称顶部长度;
- (5)水平方向灰度投影直方图曲线中斜率与车道标准线斜率比较接近的曲线段的宽度与水平方向灰度投影直方图曲线宽度的比值,简称顶长比;
- (6)以水平方向灰度投影直方图曲线中斜率与车道标准线斜率比较接近的曲线段的中垂线为界,界前与界后曲线宽度的比值,简称前后比;
- (7)水平方向灰度投影直方图曲线中斜率与车道标准线斜率比较接近的曲线段在水平方向灰度投

影直方图曲线中的相对位置,简称顶部位置;

(8)水平方向灰度投影直方图曲线中斜率与车道标准线斜率比较接近的曲线段的数量,简称顶部数量。

对于小客和小货车型,其车长、车宽、长宽比一般没有太大差异,如果小货车装载的货物比较平滑,则小货车的顶长比和前后比与小客车型也可能出现比较相似的情况,因此,对其进行分类时,本文选择了顶部长度、顶部位置以及顶部数量这三类特征来构建 C4.5 决策树对其进行分类。

经过对一段视频中小客车和小货车的实测统计,分别为其所对应的顶部长度、顶部位置及顶部数量设置一定的阈值,通过阈值判断后获得的车型样本和属性特征如表 1 所示。

根据式(1)计算可得表 1 所示的样本集的信息熵为

$$Entropy(S) = -(7/13) \log_2(7/13) - (6/13) \log_2(6/13) = 0.995\ 7.$$

$$\begin{aligned} Gain(S, \text{顶部长度}) &= Entropy(S) - (6/13)Entropy(S_{\text{大}}) - (7/13)Entropy(S_{\text{小}}) \\ &= 0.995\ 7 - (6/13)[- (4/6)\log_2(4/6) - (2/6)\log_2(2/6)] - \\ &\quad (7/13)[- (4/7)\log_2(4/7) - (3/7)\log_2(3/7)] \\ &= 0.041\ 4; \end{aligned}$$

$$\begin{aligned} Gain(S, \text{顶部位置}) &= Entropy(S) - (8/13)Entropy(S_{\text{中间}}) - (5/13)Entropy(S_{\text{靠前}}) \\ &= 0.218\ 8; \end{aligned}$$

$$\begin{aligned} Gain(S, \text{顶部数量}) &= Entropy(S) - (10/13)Entropy(S_{\text{少}}) - (3/13)Entropy(S_{\text{多}}) \\ &= 0.317\ 8. \end{aligned}$$

根据式(3) 计算可得表 1 所示的样本集的分裂信息,分别为

$$SplitInformation(S, \text{顶部长度}) = -(6/13) \log_2(6/13) - (7/13) \log_2(7/13) = 0.995\ 7;$$

$$SplitInformation(S, \text{顶部位置}) = -(8/13) \log_2(8/13) - (5/13) \log_2(5/13) = 0.961\ 2;$$

$$SplitInformation(S, \text{顶部数量}) = -(3/13) \log_2(3/13) - (10/13) \log_2(10/13) = 0.779\ 4.$$

最后根据式(4) 计算可得表 1 所示的样本集的信息增益比率分别为

$$\begin{aligned} GainRatio(S, \text{顶部长度}) &= 0.041\ 4/0.995\ 7 = \\ &0.041\ 5; \end{aligned}$$

$$\begin{aligned} GainRatio(S, \text{顶部位置}) &= 0.218\ 8/0.961\ 2 = \\ &0.227\ 6; \end{aligned}$$

$$\begin{aligned} GainRatio(S, \text{顶部数量}) &= 0.317\ 8/0.779\ 4 = \\ &0.407\ 7. \end{aligned}$$

根据信息增益比率的计算结果可知,顶部数量的信息增益比率最大,因此,“小型车”二分为“小货”和“小客”的二分类决策树的根结点选择顶部数量,首先根据顶部数量的“多”或“少”可以将原始样本集{D1, D2, …, D13}二分成{D1, D2, D3, D5, D7, D8, D10, D11, D12, D13}和{D4, D6, D9}两个子集,其中

表 1 车型样本及其属性特征

样本序号	顶部长度	顶部位置	顶部数量	车型
D1	大	中间	少	小客
D2	小	中间	少	小客
D3	小	靠前	少	小货
D4	大	中间	多	小货
D5	小	靠前	少	小货
D6	大	中间	多	小货
D7	小	靠前	少	小货
D8	大	中间	少	小客
D9	小	靠前	多	小货
D10	大	中间	少	小客
D11	大	靠前	少	小客
D12	小	中间	少	小客
D13	小	中间	少	小客

根据式(2)计算可得表 1 所示的样本集的信息增益分别为

子集{D4, D6, D9}中的样本已经全部为小货,成功分出;再选择顶部位置对还未能成功分出的样本子集{D1, D2, D3, D5, D7, D8, D10, D11, D12, D13}进行分类,根据顶部位置“靠前”还是在“中间”,可将该样本子集二分成{D1, D2, D8, D10, D12, D13}和{D3, D5, D7, D11}两个子集,其中子集{D1, D2, D8, D10, D12, D13}中的样本已经全部为小客,成功分出;最后选择顶部长度对样本子集{D3, D5, D7, D11}进行分类,根据顶部长度的相对“大”或“小”,可将样本子集{D3, D5, D7, D11}二分成{D11}和{D3, D5, D7},分别为小客和小货的集合,至此原始样本集已经被成功分类,最终生成的将“小型车”二分为“小货”和“小客”的 C4.5 决策树如图 7 所示。

5.2 车型分类决策树的规则后修剪

决策树代表样本实例属性特征值约束的合取 (conjunction) 的析取 (disjunction) 式. 从树根到树叶的每一条路径对应一组特征测试的合取, 树本身对应这些合取的析取. 因此, 图 1 所示的决策树对应于以下表达式:

小客=[(顶部数量=少)∩(顶部位置=中间)]
∪[(顶部数量=少)∩(顶部位置=靠前)∩(顶部长度=大)] (5)

小货=[(顶部数量=少)∩(顶部位置=靠前)
∩(顶部长度=小)]∪(顶部数量=多) (6)

决策树中从根结点到叶子结点的每一条路径都可以创建一条规则, 即式(5)和式(6)所描述的决策树表达式中的每一个由合取组成的析取项都可以创建一条规则, 如式(5)中第一个由合取组成的析取项所表达的规则可描述为

IF(顶部数量=少)∩(顶部位置=中间)
THEN 车型=小客

根据本文介绍的规则后修剪方法, 如果将式(5)中第一个由合取组成的析取项所表达的规则删去, 显然分类精度会明显降低, 该条规则不能被修剪掉, 而且图 7 所示的决策树已经不能再做任何修剪, 修剪掉任何规则都会使分类精度降低. 因此, 图 7 所示的决策树就是最终生成的能够将“小型车”二分类成“小客”和“小货”的 C4.5 决策树.

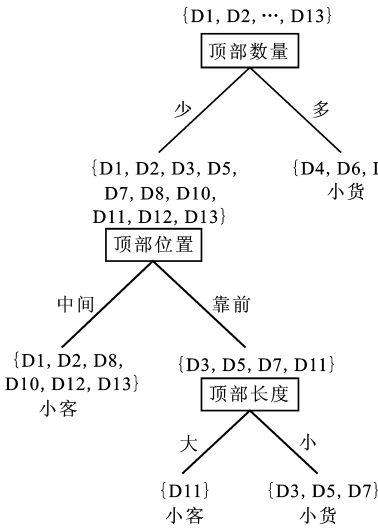


图 7 “小型车”二分为“小货”和“小客”的 C4.5 决策树

6 基于 C4.5 决策树的视频车辆车型分类实验

本文采用大恒 VT142 视频采集卡进行视频图像采集, 并使用 Visual C++ 6.0 进行编程实验. 对

每个 C4.5 决策树车型分类器训练过程均人工选取待分类车型的车辆各 100 辆作为训练样本, 因此, 共使用了 600 辆训练样本为五个六类车辆二分类器进行训练. 分类器训练好后, 再另外人工选取各种车型车辆各 100 辆进行分类效果测试. 车型分类测试结果如表 2 所示.

表 2 C4.5 决策树车型分类测试结果

车型	摩托	小客	大客	小货	大货	拖挂	总体评价
测试样本数	100	100	100	100	100	100	600
正确分类数	92	90	90	75	69	65	481
分类精度/%	92	90	90	75	69	65	80.2

通过表 2 所示的车型分类测试结果可见, 摩托、小客和大客由于其外表面比较固定, 因此这几类车的分类精度比较高; 小货、大货和拖挂会因其装载的货物的不同而投影出差异比较大的投影图像, 进而会导致其投影曲线几何特征差异比较大, 因此这几类车的分类精度较低.

7 结束语

为了克服常见分类器存在的训练时间长、编程和运算复杂、难以满足实时性等问题, 本文提出了基于 C4.5 决策树的视频车辆车型分类算法, 并以二分“小型车”为“小客车”和“小货车”两类的 C4.5 决策树构造为例, 对决策树结点属性特征选择和规则后修剪这两个决策树构造关键步骤作了详细介绍. 构造好的 C4.5 决策树能够被表示为多个 if-then 规则, 容易编程实现, 运算复杂度低, 实时性较好. 本文最后对整个车型分类算法进行了实验验证, 实验结果证实了该算法的有效性.

参考文献:

[1] Hong Zheng, Li Pan, Li Li. A morphological neural network approach for vehicle detection from high resolution satellite imagery[J]. Neural Information Processing, 2006, 4233(5949):99-106.

[2] WU S T, TOMMY W S. Self-organizing and self-evolving neurons: A new neural network for optimization[J]. IEEE Transactions on Neural Networks, 2007, 18(2): 385-396.

[3] Xiaobin L, Hui F, Jianmin X. Automatic vehicle classification based on video with BP neural networks [C]//4th International Conference on Wireless Communications, Networking and Mobile Computing. IEEE, 2008:1-3.

[4] Thi T H, Robert K, Lu S, et al. Vehicle classification

at nighttime using eigenspaces and support vector machine[C]//Congress on Image and Signal Processing. IEEE,2008;422-426.

[5] Yang W L, Wang Y, Vahdat A, et al. Kernel latent SVM for visual recognition[C]//Advances in Neural Processing Systems. IEEE, 2012: 809-817.

[6] 葛广英. 基于 SVM 的车型检测和识别算法[J]. 计算机工程, 2007, 33(6):6-8,11.

[7] Comes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[8] Kornaropoulos E M, Tsakalides P. A novel KNN classifier for acoustic vehicle classification based on alpha-stable statistical modeling [C]//IEEE Workshop Statistical Signal Processing. IEEE,2009:1-4.

[9] Judea Pearl. Reverend Bayes on inference engines: a distributed hierarchical approach [C]// Proceedings AAAI National Conference on AI. Pittsburgh, Pa, 1982;133-136.

[10] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1):81-106.

[11] Quinlan J R. C4. 5: programs for machine learning [M]. Boston: Morgan Kaufmann Publishers, 1993.

[12] 姚亚涛, 邢留涛. 决策树 C4. 5 连续属性分割阈值算法改进及其应用[J]. 中南大学学报(自然科学版), 2011, 42 (12): 372-376.

作者简介:

卞建勇 男,(1980-), 博士,副教授. 研究方向为计算机视觉及嵌入式系统应用. E-mail:bjyong977@126. com.

徐建闽 男,(1960-),博士,教授. 研究方向为交通控制及交通信息处理.

(上接第 103 页)

[7] Jesús L F, Arturo H M, Carlos G R, et al. Robust nonlinear adaptive control of a "boost" converter via algebraic parameter identification[J]. IEEE Trans Ind Electron, 2014, 61(8): 4105-4114.

[8] Abdennadher K, Venet P, Rojat G, et al. A real-time predictive-maintenance system of aluminum electrolytic capacitors used in uninterrupted power supplies[J]. IEEE Trans Ind Appl , 2010, 46(4): 1642-1644.

[9] Yao K, Tang W J, Hu W B, et al. A current-sensor-less online ESR and C identification method for output capacitor of buck converter[J]. IEEE Trans Power Electron, 2015, 30(12):6993-7005.

作者简介:

陈 晨 女,(1992-),博士研究生. 研究方向为 DC-DC 转换器的数字控制与参数辨识.

E-mail:chenchen_hust@163. com.

童乔凌 男,(1981-),博士,副教授,博士生导师. 研究方向为超大规模集成电路设计.

闵 闰 男,(1989-),博士. 研究方向为 DC-DC 电力电子系统的建模、分析与控制.

吕 典 男,(1995-),博士研究生. 研究方向为高频 DC-DC 转换器的设计、建模及控制.