一种用于大规模数据集的决策树采样策略

赵国强,王会进

(暨南大学 信息科学技术学院,广东 广州 510632)

摘 要:为提高大规模数据集生成树的准确率,提出一种预生成一棵基于这个数据集的决策树,采用广度优先遍历将其划分为满足预定义的限制的数据集,再对各数据集按照一定比例进行随机采样,最后将采样结果整合为目标数据集的数据采样方法。通过对一UCI数据集进行采样,并用现有决策树算法实验证明,该采样方法优于传统随机采样方法,基于该采样方法的生成树准确率有所提高。

关键词:决策树;样本选取;广度优先遍历

中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2010)21-0005-02

A sampling strategy on decision tree for large data sets

ZHAO Cuo Qiang, WANG Hui Jin

(College of Information Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract: To raise the accuracy of decision trees on extensive data sets, proposed a new kind of way to sample on data sets. Pre-generated a decision tree using some fast decision tree algorithms, divide the decision tree into some data sets in predefined limit by BSF manner, then sample on every set in random, integrate all sets into target set. Experiment on an UCI data set show that the ratio of average correct rates is higher than traditional random sample.

Key words: decision tree; sample selection; BFS

随着信息爆炸时代的到来,人们常常要面对海量的数据分析和处理任务,而且这些数据还在以几何级数的速度增加。同时,在现实中这些海量数据往往是高维而稀疏的,且存在着大量的冗余。因而能对数据进行有效地采样,且保持其准确率的处理方法成为人工智能、机器学习、数据挖掘等领域的重要研究课题之一。

决策树[1]算法由于其易于理解等特点被广泛应用于机器学习和数据挖掘中。然而由于决策树算法采用的是贪心策略,这就决定了其生成的决策树只是局部最优而非全局最优。同时一个决策树算法的成功在于生成基于给定的数据集下最高准确率的生成树。但是由于面对的数据集是海量的,所以如果简单地运用决策树生成算法,不仅需要大量的计算,而且无法保证低错误率和低偏差。所以有必要在真正进行挖掘前进行数据采样,以期有效地提高准确率。

本文提出一种结构化的采样技术,运用现有决策树算法对整个数据集生成决策树,然后对生成的决策树进行后加工,再基于生成的多个数据集进行随机取样,最后,整合取样后的样本生成目标样本集。

1 决策树算法

决策树技术(Decision tree)是用于分类和决策的主要技术,决策树学习是以实例为基础的归纳学习方法,通过一组无序、无规则的实例推理出决策树表示形式的分类规则。决策树是运用于分类的一种类似于流程图的树结构,其顶层节点是树的根节点,每个分枝代表一个测试输出,每个非叶子节点表示一个属性的测试,每个叶节点代表一个类或一个类的分布。决策树进行分类主要有两步:第一步是利用训练集建立一棵决策树,建立决策树模型;第二步是利用生成完毕的决策树模型对未知数据进行分类。

由于决策树算法具有良好的预测性和易理解性,所以被广泛研究和应用。目前,有许多好的决策树算法,如ID3、C4.5[2]、CART[3]等。ID3算法采用贪心(即非回溯的)方法,决策树以自顶向下递归的分治方法构造。通过对一个训练集进行学习,生成一棵决策树,训练集中的每一个例子都组织成属性-属性值对的形式,例子的所有

属性都为离散属性。而 C4.5 是由 ID3 演变来的，其核心思想是利用信息熵原理，使用信息增益率(Gain Ratio)的信息增益扩充，使用分裂信息(Split Information)值将信息增益规范化，递归地构造决策树分支，完成决策树。本文的实验中生成预决策树时将用该算法。CART (Classification And Regression Tree)算法采用一种二分递归分割的技术，将当前的样本集分为两个子样本集，使得生成的决策树的每个非叶子节点都有两个分支。因此，CART 算法生成的决策树是结构简洁的二叉树。同时，CART 算法考虑到每个节点都有成为叶子节点的可能，对每个节点都分配类别。分配类别的方法可以用当前节点中出现最多的类别，也可以参考当前节点的分类错误或者其他更复杂的方法。

当然也有一些非常好的针对大数据集的决策树算法，如 SPRINT、SLIQ 等，然而由于生成的树过于庞大，给理解它带来了一定困难。虽然还有一些相关的剪枝技术，但其中也伴随着由于过度剪枝而降低精确度的问题，使得其无法接近最优。

2 采样方法

本文提出一种基于预生成决策树的机构化的采样方法。首先通过现有的任意一种快速的决策树生成算法生成一棵决策树；之后对生成的决策树进行后加工，再基于生成的数据集进行随机取样；最后，整合取样后的样本集生成目标样本。

具体算法是：首先对整个数据集采用一种快速的决策树生成算法生成决策树。然后采用广度优先遍历该生成树，当遍历的节点所包含的样本量等于预定义的限制时终止，将遍历过的节点所包含的样本存于数据集 $Si(i=1\sim n)$。如此反复，直到遍历过所有节点为止。由此便产生了 n 个数据集，然后再随机地从这 n 个数据集中随机取样本，其中每个集内所取样本的数量 K 由以下公式决定：$K=M\times|Si|/|\sum iSi|$。其中 M 表示目标样本大小，$|Si|$表示数据集 Si 中样本的个数，$|\sum iSi|$表示样本总个数。最后再将随机取得的所有样本整合为目标样本集。该算法采样的过程如下所示：

(1)用现有决策树算法对整个数据集建立决策树。

(2)Do

Do

广度优先算法遍历生成树；

从左到右整合兄弟节点；

While 节点包含样本的个数<预定义限制；

将整合好的样本存于集合 Si；

i++；

While 遍历完所有节点；

(3)对每一个集合 $Si(i=1\sim n)$进行大小为 K 的随机采样，其中 $K=M\times|Si|/|\sum iSi|$；

(4)整合(3)中采集得到的所有样本生成目标样本集。

3 实验

选取 UCI 数据集[4]中的大型数据集“census-income”作为实验对象。该数据集包括 199 523 个样本，共包括 41 个属性，其中 8 个是连续性的。同时对于连续属性的样本先做了离散化，以节省计算时间。

选用 C4.5 算法作为预先生成树的算法，产生的树共有 1 821 个节点，其中叶子节点为 1 661 个，错误率为 0.042 8。其中在进行树的广度优先遍历时的预定义的集合大小为 30 000。得到的生成树如下：

```
capital_gains='(-∞~57 ]'
|dividends_from_stocks='(-∞~0.5 ]'
| | weeks_worked_in_year='(-∞~0.5 ]':
| | weeks_worked_in_year='(0.5~51.5 ]':
| | weeks_worked_in_year='(50.5~+∞ ]'
| | | capital_losses='(-∞~1881.5 ]'
| | | | sex=Female:
| | | | sex=Male:
| | | capital_losses='(1881.5~+∞ ]'
| dividends_from_stocks='(0.5~+∞ ]'
capital_gains='(57~+∞ ]':
```

采用常用的随机采样方法对数据集“census-income”进行大小为 10 000 的采样 5 次，之后采用经典的决策树算法 C4.5、CART 进行决策树的生成，其树的规模及准确率如表 1 所示。同时对该数据集合采用文中所提出的采样方法进行大小为 10 000 的采样 5 次，并用决策树算法 C4.5、CART 进行决策树的生成，其树的规模及准确率如表 2 所示。

表 1 随机取样 10 000 个的结果

样本集	C4.5		CART	
	树的规模	准确率/%	树的规模	准确率/%
样本 1	92	94.84	25	94.62
样本 2	57	94.58	17	94.58
样本 3	79	94.57	15	94.6
样本 4	142	94.62	23	94.66
样本 5	83	94.55	29	94.67
平均值	90.6	94.63	21.8	94.63

表 2 采用结构化的采样方法采样 10 000 个的结果

样本集	C4.5		CART	
	树的规模	准确率/%	树的规模	准确率/%
样本 1	178	94.53	15	94.78
样本 2	35	94.64	21	94.68
样本 3	105	94.73	17	94.65
样本 4	110	94.76	15	94.71
样本 5	28	94.62	19	94.83
平均值	91.5	94.66	17.4	94.73

由表 1、表 2 比较可知，新的采样方法在生成树的准确率方面比 C4.5 算法和 CART 算法都有所提高，

特别是对CART算法有较大的提高。

随机采样的方法是在对较大规模的数据库进行数据挖掘时常用的方法，然而由于决策树生成算法是贪婪算法，其只能找出局部最优解，所以简单的随机采样方法不能对准确率的提高起到作用。本文提供的新的采样方法通过用现有决策树快速生成预决策树的方法，有效利用已生成的知识结构，再对预决策树进行更加具有平衡性的采样进而形成目标数据集。实验证明，该采样方法与随机采样方法相比，准确率有一定提高。

参考文献

[1] QUINLAN, J R. Induction of decision tree [J]. Machine Learning, 1986,1(1):81-106.

[2] QUINLAN, J R. C4.5: Programs for machine learning[R]. Morgan Kaufmann Publishers, Inc., 1993.

[3] MACHOVA, K. BARCAK, F. BEDNAR, P. A bagging method using decision trees in the role of base classifiers [J]. Acta Polytechnica Hungarica, 2006,3(2): 121-132.

[4] NEWMAN D. UCI KDD Archive. [http://kdd. ics. uci. edu]. Irvine, CA: University of California, Department of Information and Computer Science, 2005.

(收稿日期:2010-07-11)

作者简介:

赵国强，男，1984年生，硕士研究生，主要研究方向：计算机软件与理论。