一种随机森林的混合算法

曹正凤[1,2],谢邦昌[3],纪　宏[1]

(1.首都经济贸易大学 统计学院,北京 100070;2.北京石油化工学院 经济管理学院,北京 102617;
3.台湾辅仁大学 统计资讯学系,台湾 新北 24205)

摘　要:随机森林(RF)是众多分类算法中精确度较高的算法,但其精确度还有提升的需求。文章通过分析C4.5算法和CART算法的计算过程,比较了两者的异同点,提出了一种新的混合随机森林算法,并使用公共的UCI数据集进行实证分析,实验数据表明该算法可以提高随机森林的精确度。从而,使随机森林算法的应用领域得到了扩大。

关键词:随机森林;混合算法;精确度

中图分类号:O211.6　　**文献标识码**:A　　**文章编号**:1002-6487(2014)04-0007-03

0　引言

随机森林(RF)是一个众所周知的机器学习技术,它的理论基础是统计学习理论,它测试样本重复抽样,随机生成多个决策树,由这些树构造出森林,然后采用简单多数投票法确定分类或预测的结果。它可以处理多变量数据,可以评估变量的重要性,对缺失数据和噪音数据都具有较好的容忍度,具有较高的预测准确率,在医学、蛋白质组学[1]、生态学研究[2]等领域有着广泛的应用。

众所周知,分类算法一个重要的性能指标是精确度,精确度的提高将是分类算法优化研究中永恒的主题,随机森林是众多分类算法中精确度较高的算法,但在查阅大量的研究文献后和经过实验数据检验后,笔者发现其精确度大约在70~80%之间,在某些需要更高精度的领域应用时,该算法还具有进一步优化的现实需求。

基金项目:国家自然科学基金资助项目(71071022)

作者简介:曹正凤(1979-),男,江西九江人,博士研究生,实验师,研究方向:统计理论。
谢邦昌(1962-),男,湖南耒阳人,教授,博士生导师,研究方向:数据挖掘。
纪　宏(1954-),男,北京人,教授,博士生导师,研究方向:国民经济统计。

流平台兼容性较好的开源软件来支持系统的部署。同时基于云计算平台的网络直报系统应兼容既有的网络直报系统,在系统迁移过程中对原有系统不需要进行大规模的改动,可以实现平稳过渡,从而保证关键业务的连续性以及减少系统的迁移成本。

(3)安全性。统计数据网络直报要保证提供可靠安全的数据存储中心,使用数据多副本容错等技术措施来保障服务的高可靠性,使云计算平台比本地计算机更可靠,用户端不必再担心数据丢失、病毒入侵等麻烦。在"云"的另一端,使用严格的权限管理策略以更有效的保障数据安全。另外,要明确权责控制。按照权限管理,对管理员规定执行权限。若对数据进行改动,系统将会自动留下无法消除的痕迹,从而可以根本上杜绝篡改数据的可能性,数据的准确性有了进一步的保证。

(4)成本控制。对基于云的网络直报平台来说,低成本也是一个重要的构建原则。不佳的平台将会消耗更多的服务器、存储、网络设备,从而增加提供冷却的空调数量,消耗大量的电能。首先,要通过优化设计来避免资源浪费。另外,在选择云计算平台时,要考虑环境和空间的布置。传统的服务器,需要占用大量的机架、空间,消耗大量的电缆和辅助材料。而且,空间的占用也会带来管理的困难,增加维护成本。因此应提高部署的密度,采用高密度计算系统。

政府可以与方案提供商合作,由方案提供商规划云计算服务运营体系,并且进行技术改造和业务平台开发。具备基本的雏形之后,可以选择一个市作为试点单位,待各方面发展成熟之后再考虑推广至全省以至全国。另外应注重加强培训,提高统计人员的信息技术应用能力,统计人员要逐步培养成为"熟悉统计科学,掌握云计算技术"的专门人才。

参考文献:

[1]辛金国,王琳燕,韩秀春.网络直报条件下地区统计数据质量影响因素研究[J]调研世界,2010,(8).

[3]王秀花.企业统计报表数据网上直报的几点思考[J].财会与决策.2008,(9).

[4]郑慧勇.关于统计数据中心云计算平台的思考[J].调研与观察,2011,(3).

[5]付瑞平."统计云"先行[J].中国信息化,2011,(8).

(责任编辑/亦　民)

本文考虑建立一个新的算法，该算法通过优化随机森林中单棵决策树的节点分裂算法，达到提高算法精度的目的。关于决策树节点分裂方面的优化内容，学术界研究较多。RuggieriS提出了的EC4.5算法，是对C4.5的改进算法[3]。该算法改进了C4.5算法中的线性搜索，采用二分搜索法。在决策树生成的过程中，EC4.5算法通过牺牲空间换时间的办法，使算法的执行效率得到了提升，但同时，其内存空间需求也相应的增大了[4]。理论界关于决策树节点分裂方面的优化都停留在针对某个算法的优化上，很少有人将多个算法综合在一起进行研究，而单棵决策树节点分裂算法中比较成熟C4.5算法和CART算法，这两个算法在进行节点分裂时，存在一定的区别，比如同一个节点，因为选择的算法不同，选择的属性会不同，正是由于这些差距，使得随机森林算法在精确度上还有提升的可能。本研究旨在建立上述二种算法的混合模型，把二者的优点集中在一起，以期得到性能更佳的分类算法。

1 混合随机森林算法的提出

1.1 C4.5算法

1993年Quinlan提出，采用信息增益率来选择属性，进行决策树的节点分裂，即C4.5算法，其计算过程如下：

①计算出节点的信息熵

$$Info(D)=-\sum_{i-1}^{m}P_i\log_2 P_i \tag{1}$$

其中P_i是D中任意样本属于Ci的概率

②基于按A划分对D的样本分类所需要的期望信息，其公式如下：

$$Info_A(D)=\sum_{j-1}^{v}\left[\left(\frac{|D_i|}{|D|}\right)*Info(D_j)\right] \tag{2}$$

其中，$\frac{|D_i|}{|D|}$充当第j个划分的权重。$Info_A(D)$越小，划分的纯度越高。

③由公式(1)和(2)可得信息增益，定义式为：

$$Gain(A)=Info(D)-Info_A(D) \tag{3}$$

$Gain(A)$表示A的值而导致的信息需求的期望减少。

④描述节点分裂时的信息，其公式为：

$$SplitInfo_A(D)=-\sum_{j-1}^{v}\left[\left(\frac{|D_i|}{|D|}\right)*\log_2\left(\frac{D_j}{D}\right)\right] \tag{4}$$

该值表示数据集D按属性A测试的V个划分产生的信息。

⑤根据信息增益和节点信息，得出信息增益率：

$$GainRatio(A)=\frac{Gain(A)}{SplitInfo_A(D)} \tag{5}$$

节点分裂时，选择增益率最大属性的作为分裂属性。

1.2 CART算法

CART算法，即分类与回归树，它由Breiman于1984年提出，它的基本原理是采用递归分割思想，将当前的样本集一分为二，生成两个子样本集，然后对每个子样本集进行递归分割，直到最后的样本集只有一个样本或者都为同一类别。因此，CART算法生成的决策树一定是二叉树，且结构简洁。节点分裂时，算法使用Gini系数指标来度量数据划分，其计算过程如下：

①计算样本的Gini系数

$$Gini(S)=1-\sum_{i-1}^{m}P_i^2 \tag{6}$$

其中P_i代表类别Cj在样本集S中出现的概率。

②计算每个划分的Gini系数

如果S被分隔成两个子集S1与S2，则此次划分的Gini系数为

$$Gini_{split}(S)=\frac{|S_1|}{|S|}Gini(S_1)+\frac{|S_2|}{|S|}Gini(S_2) \tag{7}$$

其中，|S|是样本集S的样本个数，$|S_1|$、$|S_2|$分别为两个子集S_1与S_2中样本个数。

CART算法通过检查同一层级中，所有变量和该变量对样本集的一切可能的一分为二的的划分，所得到的Gini系数的值，通过比较所有划分的Gini系数的值，来发现最好的划分。在节点分裂时，将每个属性的所有划分按照他们Gini系数来进行排序，节点分裂时，选择Gini系数最小的属性作为分裂属性，并按照其划分办法进行集合的划分。

1.3 C4.5算法和CART算法的共同点

①CART与C4.5算法都是基于信息论的决策树算法

②比较公式(1)和公式(6)，发现两个公式中的Pi是取相同的值，这为两个算法的混合提供了比较好的基础。

③比较公式(2)和公式(7)，发现两个公式中的系数取值也有一定的相似之处，如果公式(2)中，V的值取2时，两者的取值完成一致。

1.4 C4.5算法和CART算法的不同点

①C4.5算法比CART算法多了公式(3)、(4)、(5)三个公式，但这三个公式都可以由公式(2)中的数值取得

②C4.5算法生成的决策树可以是多叉树，而CART算法只能是二叉树。

1.5 混合随机森林

从上述两个算法的异同点可以看出，如果将C4.5算法进行公式(2)的集合划分是，只划分成两个集合，即v=2，此时C4.5算法和CART算法具有较相似的计算过程，在设计时，就可以将这两个算法混合为一个算法，使用该算法生成的随机森林，即为混合随机森林。

建立节点分裂混合算法，提升算法的精确度，混合算法模型如下：

$$\Phi(\alpha)=\beta_1 Gini_{split}(S)-\beta_2 GainRatio(A) \tag{8}$$

其中：

$0\leqslant\beta_i(i=1,2)\leqslant 1$不能二个同时为0，也不能同时为1，即边界点上，只能为(1,0)、(0,1)组合；

$Gini_{split}(S)$为CART算法分类的计算公式(2)；

$GainRatio(A)$为C4.5算法分类的计算公式(7)。

在节点分裂时，选择$\Phi(\alpha)$值最小的划分进行分裂。

2 实证分析

根据公式(8),使用JAVA语言编程实现混合算法,选取UCI数据集中的训练文件Adult Data Set,该数据集是巴里·贝克尔根据1994年的人口普查数据库中的数据,预测个人年收入是否超过5万元,数据集下载地址为:http://archive.ics.uci.edu/ml/datasets/Adult。该数据集为二分类数据。

2.1 数据结构

该数据集中,结果分类为两类,分别为">50K"和"<=50K"。

连续性变量共有6个,分别为:age,fnlwgt,education-num,capital-gain,capital-loss,hours-per-week。

离散型变量共有8个,分别为:workclass,education,marital-status,occupation,relationship,race,sex,native-country

数据集显示图略。

2.2 数据预处理

首先将结果属性值定义为"yes"和"no",便于程序的设计。对离散型属性,不需要做处理就可以直接使用,对于连续性属性变量就需要进行预处理,首先把连续型属性的值划分成间距相同的区间,即实现连续性变量的"离散化"。其"离散化"的过程如下:

①寻找该连续型属性的最小值和最大值,并分别赋值为Min和Max变量;

②随机设置变量N,将区间【Min,Max】分成N+1个区间,取得这些区间中的等分断点Ai(Ai=Min+(Max-Min)/N*i,其中,i=1,2,……,N);

③针对每个Ai,分别【Min,Ai】和(Ai,Max】(i=1,2,……,N)两个区间上,对应的Gini值,并排序;

④选取序列中Gini值最大的等分断点Ak作为该连续型属性的断点,把属性值设置为【Min,Ak】和(Ak,Max】两个区间值,分别记为MinAk,MaxAk。

将数据进行上述的整理后,去掉缺失数据,得到训练集1505个样本,测试集1200个样本。样本示意图略。

2.3 实验分析

将预处理后的样本,引入到程序中,程序运行结果如下。

表1 混合随机森林算法测试结果表

序号	β_1	β_2	精确度	运行时间(毫秒)
1	1	0	0.7475	558344
2	0.9	0.1	0.7515	580968
3	0.8	0.2	0.7555	609828
4	0.7	0.3	0.7495	591157
5	0.6	0.4	0.7515	601344
6	0.5	0.5	0.7515	607844
7	0.4	0.6	0.7515	628140
8	0.3	0.7	0.7495	641438
9	0.2	0.8	0.7515	643281
10	0.1	0.9	0.7495	649406
11	0	1	0.7475	671157

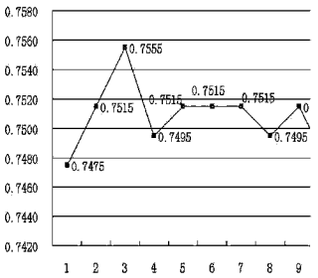

图1 混合随机森林算法测试结果示意图

从表1和图1可以看出:

(1)C4.5算法和CART算法精确度相当;

(2)使用混合随机森林算法后,算法的精确度整体上升;

(3)使用混合随机森林算法后,算法的时间复杂度也随之上升;

(4)当β_1的值为0.8,β_2的值为0.2时,程序的精确度高超过单独使用C4.5算法或CART算法的精确度,因此可以使用混合算来提升随机森林算法的精确度,可以达到某些特定领域的需求。

3 结束语

综上所述,本文提出的混合随机森林算法,可以提高算法的精确度,在某些对精确度要求高的领域得到使用。同时,由于C4.5算法使决策树多向分叉,但CART只能进行二向分叉,因此二者结合在一起,只能使用二向分叉,这使得算法在应用时或许存在一定的问题,需要在今后的研究中进一步拓展,以期取得更好的成果。

参考文献:

[1]G. Izmirlian. Application of the Random Forest Classification Algorithm to a Seldi-tof Proteomics Study in the Setting of a Cancer Prevention Trial[J]. Annals of the New York Academy of Sciences,2004,1020.

[2]D. R. Cutler, et al.Random Forests for Classification in Ecology[J]. Ecology,2007.,88(1).

[3]Ruggieri S. Efficient C4.5.IEEE Transactions on Knowledge and Data Engineering[Z] .2002.

[4]栾丽华,吉根林.决策树分类技术研究[J].计算机工程,2004,30(9).

(责任编辑/亦 民)