随机森林理论浅析

董师师[1,2] 黄哲学[1,2]

[1]（深圳市高性能数据挖掘重点实验室 深圳 518055）
[2]（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 随机森林是一种著名的集成学习方法，被广泛应用于数据分类和非参数回归。本文对随机森林算法的主要理论进行阐述，包括随机森林收敛定理、泛化误差界以和袋外估计三个部分。最后介绍一种属性加权子空间抽样的随机森林改进算法，用于解决超高维数据的分类问题。

关键词 随机森林；数据挖掘；机器学习

A Brief Theoretical Overview of Random Forests

DONG Shi-shi[1,2] HUANG Zhe-xue[1,2]

[1](*Shenzhen Key Laboratory of High Performance Data Mining, Shenzhen* 518055, *China*)
[2](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)

Abstract Random Forests is an important ensemble learning method and it is widely used in data classification and nonparametric regression. In this paper, we review three main theoretical issues of random forests, i.e., the convergence theorem, the generalization error bound and the out-of-bag estimation. In the end, we present an improved Random Forests algorithm, which uses a feature weighting sampling method to sample a subset of features at each node in growing trees. The new method is suitable to solve classification problems of very high dimensional data.

Keywords random forests; data mining; machine learning

1 引 言

数据分类是日常生活中常见的一个问题，现实生活中许多领域需要解决的问题本质上是数据分类问题，比如生物信息数据分类，商业客户数据分类等，分类成为数据挖掘领域最重要的研究问题之一。传统的数据分类方法包括聚类算法、贝叶斯分类算法、决策树算法[1-3]、支持向量机算法[4]等。近年来，随着各种数据采集技术的快速发展，实际应用中产生并积累了越来越多的高维复杂数据。传统的数据分类算法对这些高维数据并不适用，研究针对高维数据的分类算法是当前迫切需要解决的难题之一。

决策树是广泛应用的一种分类算法，它是一种树状分类器，在每个内部节点选择最优的分裂属性进行分类，每个叶节点是具有同一个类别的数据。当输入待分类样本时，决策树确定一条由根节点到叶节点的唯一路径，该路径的叶节点的类别就是待分类样本的所属类别。决策树是一种简单且快速的非参数分类方法，一般情况下，还具有很好的准确率，然而当数据复杂或者存在噪声时，决策树容易出现过拟合问题，使得分类精度下降。

随机森林[5]是由美国科学家Leo Breiman将其在1996年提出的Bagging集成学习理论[6]与Ho在1998年提出的随机子空间方法[7]相结合，于2001年发表的一种机器学习算法。随机森林是以决策树为基本分类器的一个集成学习模型，它包含多个由Bagging集成学习技术训练得到的决策树，当输入待分类的样本时，最终的分类结果由单个决策树的输出结果投票决定，如图1所示。随机森林克服了决策树过拟合问题，对

作者简介：董师师，硕士，研究方向为统计学习、数据挖掘，E-mail：ss.dong@siat.ac.cn；黄哲学，博导，研究员，研究方向为高性能计算、数据挖掘、海量复杂数据实时分析及数据库和数据仓库等。

噪声和异常值有较好的容忍性，对高维数据分类问题具有良好的可扩展性和并行性。此外，随机森林是由数据驱动的一种非参数分类方法，只需通过对给定样本的学习训练分类规则，并不需要分类的先验知识。

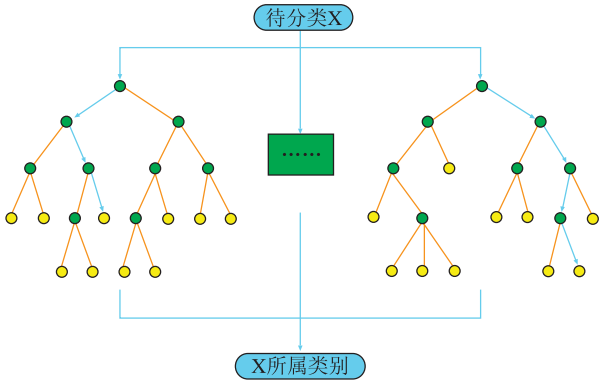

图1 随机森林分类过程

在Breiman提出随机森林算法之后，由于其良好的性能表现，该算法被广泛应用到诸如生物信息[8,9]、医学研究[12,13]、商业管理[10,11]、语言建模[14]、文本分类[15]、经济金融[16]等实际领域，这些领域的问题利用随机森林方法都取得了不错结果。相比于随机森林在应用方面的大量研究，随机森林的理论研究显得不够深入，存在很大不足。而随着针对高维复杂数据的随机森林算法成为新的研究热点，各新算法的性能表现也需要寻求相应的理论支持。因此对随机森林的理论进行研究和发展很有意义。

本文第三部分主要介绍了Breiman在论文中提出来的随机森林的数学理论，包括随机森林的收敛定理（Convergence Theorem）、泛化误差界（Generalization Error Bound）以及袋外估计（Out-of-bag Estimation）三大部分，其中收敛定理保证了随机森林不会出现过拟合问题，泛化误差界给出了随机森林分类误差率的一个理论上界，袋外估计是在实验中利用袋外数据估计泛化误差界的一种方法。在介绍随机森林理论之前，文中的第二部分对随机森林算法进行了一个简要概述。本文的第四部分介绍了一种随机森林的改进算法——属性加权子空间抽样随机森林算法，它能更好的适用于高维属性数据。第五部分对随机森林方法进行了总结和展望。

2 随机森林算法简介

2.1 决策树

20世纪70年代末期和80年代初期，J.Ross Quinlan提出了ID3决策树算法[1]，Leo Breiman等人提出了CART决策树算法[2]，1993年Quinlan又提出了C4.5决策树算法[3]。这三种算法均采用自上而下的贪婪算法构建一个树状结构，在每个内部节点选取一个最优的属性进行分裂，每个分枝对应一个属性值，如此递归建树直到满足终止条件，每个叶节点表示沿此路径的样本的所属类别。

决策树的一个关键步骤是节点分裂的属性选择，属性选择度量使得每个内部节点按最优的属性进行分裂。属性选择度量是一种启发式方法，理想情况下，根据最优的分裂属性得到的每个划分应该是纯的，也就是落在该划分之下的样本应该是属于同一类[17]，因此可以定义节点的不纯度(Impurity)函数F，函数F越大，节点的不纯度越高，函数F取值为零时，表示该节点处的样本点属于同一个类别。一个最优的分裂属性应该使得父节点与子节点的不纯度增益，即父节点不纯度与子节点不纯度加权和的差值，达到最大，这说明最优属性提供了最多的对分类有利的信息。常用的不纯度函数有Gini指标，信息熵等。

由训练样本集训练出的决策树实际上是对训练样本空间的一个划分，某一类别对应的样本空间子集可以视为一个超立方体的并集，如图2[18]所示，二维情况下是一些矩形的并集，三维情况下是一些长方体的并集：

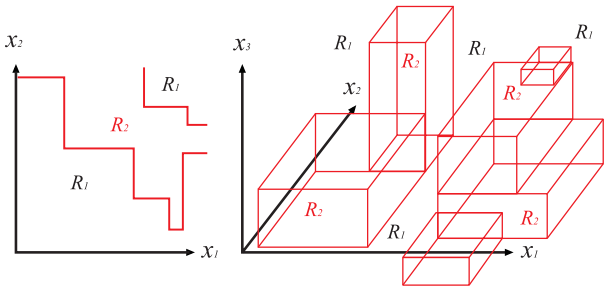

图 2 决策树对样本空间的划分

不同类别对应的样本空间子集是互不相交的，所以这些超立方体并集也是互不相交的，各个超立方体并集并起来，是整个样本空间。

决策树不需要先验知识，相比神经网络等方法更容易解释，但是由于决策树在递归的过程中，可能会过度分割样本空间，最终建立的决策树过于复杂，导致过拟合的问题，使得分类精度降低。为避免过拟合问题，需要对决策树进行剪枝，根据剪枝顺序不同，有事先剪枝方法和事后剪枝方法，但都会增加算法的复杂性。

2.2 集成学习

由于单个分类器往往有分类精度不高，容易出现

过拟合等问题，导致分类器的泛化能力较弱。集成学习是将单个分类器聚集起来，通过对每个基本分类器的分类结果进行组合，来决定待分类样本的归属类别。集成学习比单个分类器更好的分类性能，可以有效地提高学习系统的泛华能力。

集成学习可行有两个前提条件：一是单个基本分类器是有效的，也就是说单个分类器的精度应该大于随机猜对的概率；二是各个基本分类器是有差异的，要达到差异性，可以通过采用不同的训练样本或者不同的训练方法实现。

Bagging[6]是由Breiman根据统计中Bootstrap思想提出的一种集成学习算法，它从原始样本集中可重复抽样得到不同的Bootstrap训练样本，进而训练出各个基本分类器。假设原样本集的样本容量大小为N，每次有放回抽取的Bootstrap样本大小也为N，那么每个样本未被抽中的概率约为 $\left(1-\frac{1}{N}\right)^N$，当$N$很大时，这个概率值趋于 $1/e \approx 0.368$，这表明每次抽样时原样本集中约有37%的样本未被抽中，这一部分未被抽中的样本称为袋外数据。Breiman指出对于决策树等不稳定（即对训练数据敏感）的分类器，能提高分类的准确度。此外Bagging算法可以并行训练多个基本分类器，可以节省大量的时间开销，这也是该算法的优势之一。

2.3　随机森林

随机森林是以K个决策树 $\{h(X,\theta_k),k=1,2,\cdots,K\}$ 为基本分类器，进行集成学习后得到的一个组合分类器。当输入待分类样本时，随机森林输出的分类结果由每个决策树的分类结果简单投票决定。这里的 $\{\theta_k,k=1,2,\cdots,K\}$ 是一个随机变量序列，它是由随机森林的两大随机化思想决定的：

（1）Bagging思想：从原样本集X中有放回地随机抽取K个与原样本集同样大小的训练样本集 $\{T_k,k=1,2,\cdots,K\}$（每次约有37%的样本未被抽中），每个训练样本集T_k构造一个对应的决策树。

（2）特征子空间思想：在对决策树每个节点进行分裂时，从全部属性中等概率随机抽取一个属性子集（通常取 $\lfloor log_2(M)\rfloor+1$ 个属性，M为特征总数），再从这个子集中选择一个最优属性来分裂节点。

由于构建每个决策树时，随机抽取训练样本集和属性子集的过程都是独立的，且总体都是一样的，因此 $\{\theta_k,k=1,2,\cdots,K\}$ 是一个独立同分布的随机变量序列。

训练随机森林的过程就是训练各个决策树的过程，由于各个决策树的训练是相互独立的，因此随机森林的训练可以通过并行处理来实现，这将大大提高生成模型的效率。随机森林中第k个决策树 $h(X,\theta_k)$ 的训练过程如下图3所示。

将以同样的方式训练得到K个决策树组合起来，就可以得到一个随机森林。当输入待分类的样本时，随机森林输出的分类结果由每个决策树的输出结果进行简单投票（即取众数）决定。

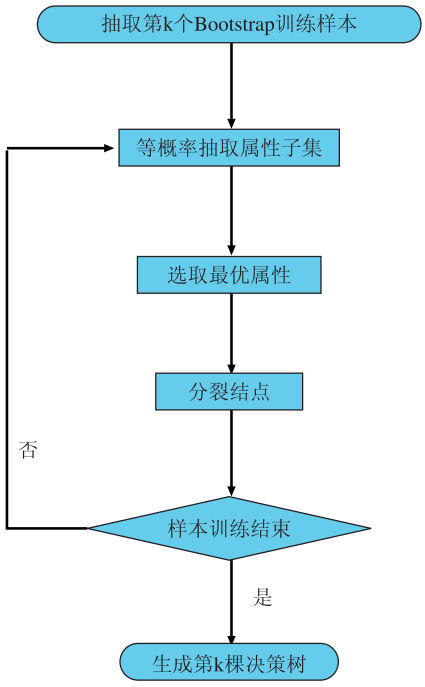

图3　随机森林中单个决策树训练过程

随机森林是一种有效的分类预测方法，它有很高的分类精度，对于噪声和异常值有较好的稳健性，且具有较强的泛化能力。Breiman在论文中提出了随机森林的数学理论，证明随机森林不会出现决策树的过拟合问题，推导随机森林的泛化误差界，并且利用Bagging产生的袋外数据给出泛化误差界的估计。

3　随机森林理论要点

设随机森林可表示为 $\{h(X,\theta_k),k=1,2,\cdots,K\}$，其中$K$为随机森林所包含的决策树个数。原样本集$T$为 $\{(x_i,y_i),x_i\in X,y_i\in Y,i=1,2,\cdots,N\}$，其中$N$为样本容量，$X$中的元素为$M$维的属性向量，$Y$包含$J$个不同类别。第$k$次抽取的Bootstrap训练样本集记为$T_k$。

3.1　收敛定理

定义随机森林的泛化误差PE^*如下：

$$PE^* \overset{\text{def}}{=} P_{X,Y}\left(av_k I(h(X,\theta_k)=Y)-\max_{j\neq Y} av_k I(h(X,\theta_k)=j)<0\right) \quad (1)$$

它度量了随机森林对给定样本集的分类错误率。Breiman证明了如下的收敛定理：

定理：当随机森林中决策树的个数$K\to\infty$时，有如下的几乎处处收敛[1]关系成立：

$$PE^* \xrightarrow{a.s.} P_{X,Y}\left(P_\theta(h(X,\theta)=Y)-\max_{j\neq Y}P_\theta(h(X,\theta)=j)<0\right) \tag{2}$$

证明：比较(1)式右边和(2)式右边，要证明(2)式，只需证明对任意的j，在$(\theta_1,\theta_2,\cdots,\theta_k,\cdots)$的取值空间中存在一个零测集$C$，使得在$C$之外，对于所有的$x$，成立：

$$\frac{1}{K}\sum_{k=1}^{K}I(h(\theta_k,x)=j)\xrightarrow{K\to\infty}P_\theta(h(\theta,x)=j)$$

对于一个给定的训练集和固定的θ，使得$h(\theta,x)=j$成立的所有x构成一个超立方体的并集（见图2）。考虑各个$h(\theta_k,x)=j$对应的超立方体的并并集，最后只有R个这样的超立方体并集，记为$S_1,\cdots,S_R$。R是有限的，因为被分为第j类的x构成原训练集的一个子集，而原训练集是一个有限集合，其所有可能的子集有且仅有为2^N个，故$R\leqslant 2^N$。定义$\varphi(\theta)=r$如果$\{x:h(\theta,x)=j\}=S_r$。设K_r表示K个决策树中$\varphi(\theta_k)=r$的次数，其中$k=1,2,\cdots,K$，则：

$$\frac{1}{K}\sum_{k=1}^{K}I(h(\theta_k,x)=j)=\frac{1}{K}\sum_{r=1}^{R}K_r\,I(x\in S_r)$$

等式左边表示K个决策树中将x归为第j类的次数，如果将x被某个决策树归为第j类，那么一定存在$r\in\{1,2,\cdots,R\}$使得$x\in S_r$，将$x\in S_1,x\in S_2,\cdots,x\in S_R$的次数相加，即等式右边对$K_r$求和，结果同样是$K$个决策树中将$x$归为第$j$类的次数。当$K\to\infty$时，由Borel强大数定律[2]可得：

$$\frac{K_r}{K}=\frac{1}{K}\sum_{k=1}^{K}I(\varphi(\theta_k)=r)\xrightarrow{a.s.}P_\theta(\varphi(\theta)=r)$$

不收敛的$(\theta_1,\theta_2,\cdots,\theta_k,\cdots)$构成其取值空间中的一个零测集$C$，使得取值空间除去$C$之外，对任意的$j$，处处成立：

$$\frac{1}{K}\sum_{k=1}^{K}I(h(\theta_k,x)=j)\ \to\ \sum_{r=1}^{R}P_\theta(\varphi(\theta)=r)I(x\in S_r)=P_\theta(h(\theta,x)=j)$$

注意大数定律只说明不收敛的$(\theta_1,\theta_2,\cdots,\theta_k,\cdots)$是存在的，并不能指出$(\theta_1,\theta_2,\cdots,\theta_k,\cdots)$具体取什么值的时候不收敛。

3.2 泛化误差界

定义随机森林的余量函数$mr(X,Y)$为：

$$mr(X,Y)\overset{\text{def}}{=}P_\theta(h(X,\theta)=Y)-\max_{j\neq Y}P_\theta(h(X,\theta)=j) \tag{3}$$

余量函数反映了样本X对应的正确类别Y的票数超过其他得票最多类别的票数的程度，余量函数的值越大，表明随机森林分类的置信度越高。

记：

$$\hat{j}(X,Y)=arg\max_{j\neq Y}P_\theta(h(X,\theta)=j) \tag{4}$$

余量函数$mr(X,Y)$可改写为：

$$\begin{aligned}mr(X,Y)&=P_\theta(h(X,\theta)=Y)-P_\theta\left(h(X,\theta)=\hat{j}(X,Y)\right)\\&=E_\theta\left(I(h(X,\theta)=Y)-I\left(h(X,\theta)=\hat{j}(X,Y)\right)\right)\end{aligned}$$

记：

$$rmg(\theta,X,Y)=I(h(X,\theta)=Y)-I\left(h(X,\theta)=\hat{j}(X,Y)\right) \tag{5}$$

余量函数$mr(X,Y)$可进一步改写为：

$$mr(X,Y)=E_\theta rmg(\theta,X,Y) \tag{6}$$

由收敛定理知道$PE^*\xrightarrow{a.s.}P_{X,Y}(mr(X,Y)<0)$，要估计泛化误差$PE^*$的上界，可转化为估计$P_{X,Y}(mr(X,Y)<0)$的上界。假定$E_{X,Y}mr(X,Y)>0$，这种假定是合理的，因为$E_{X,Y}mr(X,Y)$表示随机森林对各个样本分类结果的期望，大于零表示随机森林分类对样本的分类是可信的。利用契比雪夫不等式[3]可得：

$$\begin{aligned}&P_{X,Y}(mr(X,Y)<0)=P_{X,Y}(mr(X,Y)-E_{X,Y}mr(X,Y)<-E_{X,Y}mr(X,Y))\\&<P_{X,Y}(\left|mr(X,Y)-E_{X,Y}mr(X,Y)\right|>E_{X,Y}mr(X,Y))\\&\leq\frac{var_{X,Y}(mr(X,Y))}{E_{X,Y}mr(X,Y)^2}\end{aligned}$$

于是可得：

$$PE^*\leq\frac{var_{X,Y}(mr(X,Y))}{E_{X,Y}mr(X,Y)^2} \tag{7}$$

定义随机森林单棵决策树的分类强度s、决策树之间的相关性$\bar{\rho}$为：

$$s=E_{X,Y}mr(X,Y) \tag{8}$$

---

[1] 几乎处处收敛：设ξ和$\{\xi_n\}$是定义在概率空间(Ω,F,P)上的随机变量序列，若存在零测集Ω_0，即$\Omega_0\in F,P(\Omega_0)=0$，对$\forall\omega\in\Omega/\Omega_0$，有$\xi_n(\omega)\to\xi(\omega)$，称$\xi_n$几乎处处收敛于$\xi$，记为$\xi_n\xrightarrow{a.s.}\xi$。

[2] Borel强大数定律：设$\{\xi_n\}$是概率空间(Ω,F,P)上的独立同分布随机变量序列，$P(\xi_n=1)=p$，$P(\xi_n=0)=1-p$，$0<p<1$，记$S_n=\sum_{k=1}^{n}\xi_k$，则$\frac{S_n}{n}\xrightarrow{a.s.}p$。

[3] Chebychev不等式：对任一随机变量X，若期望EX和方差DX存在，则对$\forall\varepsilon>0$，有$P\{|X-EX|\geqslant\varepsilon\}\leqslant\frac{DX}{\varepsilon^2}$。

$$\bar{\rho}=\frac{E_{\theta,\theta'}(\rho(\theta,\theta')sd(\theta)sd(\theta'))}{E_{\theta,\theta'}(sd(\theta)sd(\theta'))} \tag{9}$$

通过如下计算可以得到一个由s和$\bar{\rho}$表示的PE^*的上界。

注意对独立同分布的随机变量θ,θ'，有$(E_\theta f(\theta))^2=E_{\theta,\theta}f(\theta)f(\theta')$。则$var_{X,Y}(mrX,Y)$计算如下：

$$\begin{aligned}
&var_{X,Y}mr(X,Y)=E_{X,Y}(mr(X,Y))^2-\left(E_{X,Y}mr(X,Y)\right)^2\\
&=E_{X,Y}\left(E_\theta rmg(\theta,X,Y)\right)^2-\left(E_\theta E_{X,Y}rmg(\theta,X,Y)\right)^2\\
&=E_{X,Y}E_{\theta,\theta'}\left(rmg(\theta,X,Y)rmg(\theta',X,Y)\right)-E_{\theta,\theta'}\left(E_{X,Y}rmg(\theta,X,Y)E_{X,Y}rmg(\theta',X,Y)\right)\\
&=E_{\theta,\theta'}\left(E_{X,Y}\left(rmg(\theta,X,Y)rmg(\theta',X,Y)\right)-E_{X,Y}rmg(\theta,X,Y)E_{X,Y}rmg(\theta',X,Y)\right)\\
&=E_{\theta,\theta'}\left(cov_{X,Y}\left(rmg(\theta,X,Y)rmg(\theta',X,Y)\right)\right)\\
&=E_{\theta,\theta'}(\rho(\theta,\theta')sd(\theta)sd(\theta'))
\end{aligned}$$

代入分类强度s、相关性$\bar{\rho}$，注意到$E\xi^2\geq(E\xi)^2$，可得：

$$\begin{aligned}
&var_{X,Y}mr(X,Y)=\bar{\rho}(E_\theta sd(\theta))^2\leq\bar{\rho}E_\theta\left(sd(\theta)\right)^2=\bar{\rho}E_\theta var_{X,Y}\left(rmg(\theta,X,Y)\right)\\
&=\bar{\rho}E_\theta\left(E_{X,Y}\left(rmg(\theta,X,Y)\right)^2-\left(E_{X,Y}rmg(\theta,X,Y)\right)^2\right)\\
&=\bar{\rho}\left(E_\theta E_{X,Y}\left(rmg(\theta,X,Y)\right)^2-E_\theta\left(E_{X,Y}rmg(\theta,X,Y)\right)^2\right)\\
&\leq\bar{\rho}\left(E_\theta E_{X,Y}\left(rmg(\theta,X,Y)\right)^2-\left(E_\theta E_{X,Y}rmg(\theta,X,Y)\right)^2\right)\\
&\leq\bar{\rho}(1-s^2)
\end{aligned}$$

综上可得，随机森林的泛化误差界为：

$$PE^*\leq\frac{\bar{\rho}(1-s^2)}{s^2} \tag{10}$$

由此看出，随机森林的泛化误差界与单个决策树的分类强度s成负相关，与决策树之间的相关性$\bar{\rho}$成正相关，即分类强度s越大，相关性$\bar{\rho}$越小，则泛化误差界越小，随机森林分类准确度越高。这也启发我们，对随机森林模型进行改进时，可以从两方面着手：一是提高单棵决策树的分类强度s，二是降低决策树之间的相关性$\bar{\rho}$。

3.3　袋外估计

由于Bagging方法每次从原样本集T中随机抽取Bootstrap训练样本T_k时，约有37%的样本没有被选中，这一部分未被选中的袋外数据可用于估计随机森林的单棵决策树分类强度s、决策树之间的相关性$\bar{\rho}$，进而可得到随机森林泛化误差界的估计。Breiman在论文中指出袋外估计是无偏估计，袋外估计与用同训练集一样大小的测试集进行估计的精度是一样的。

(1)分类强度s的估计

将(3)式代入(8)式可得：

$$s=E_{X,Y}\left(P_\theta(h(X,\theta)=Y)-max_{j\neq Y}P_\theta(h(X,\theta)=j)\right) \tag{11}$$

令：

$$Q(x,j)=\frac{\sum_{k=1}^{K}I(h(x,\theta_k)=j;(x,y)\notin T_k)}{\sum_{k=1}^{K}I((x,y)\notin T_k)} \tag{12}$$

其中y是x对应的正确类别，由x可唯一确定y，$(x,y)\notin T_k$表明原训练集中的(x,y)未被抽中来参与构建第k个决策树，属于袋外数据。$Q(x,j)$为概率$P_\theta(h(x,\theta)=j)$的袋外估计，用$Q(x,y)$、$Q(x,j)$估计$P_\theta(h(x,\theta)=y)$、$P_\theta(h(x,\theta)=j)$，再关于所有的样本点(x,y)取平均，即可得到分类强度s的估计：

$$s=\frac{1}{N}\sum_{i=1}^{N}\left(Q(x_i,y_i)-\max_{j\neq y_i}Q(x_i,j)\right)$$

(2)相关性$\bar{\rho}$的估计

由于已知：

$$\bar{\rho}=\frac{var_{X,Y}mr(X,Y)}{(E_\theta sd(\theta))^2} \tag{13}$$

分别估计$var_{X,Y}\ mr(X,Y)$和$E_\theta\ sd(\theta)$，代入可得$\bar{\rho}$的估计。

设$Q(x,j)$如(12)式表示$P_\theta\ (h(x,\theta)=j)$的袋外估计，令：

$$D_1(\theta_k)=\frac{\sum_{n=1}^{N}I(h(x_n,\theta_k)=y_n;(x_n,y_n)\notin T_k)}{\sum_{n=1}^{N}I((x_n,y_n)\notin T_k)} \tag{14}$$

$$D_2(\theta_k)=\frac{\sum_{n=1}^{N} I(h(x_n,\theta_k)=\hat{j}(x_n,y_n);(x_n,y_n)\notin T_k)}{\sum_{n=1}^{N} I((x_n,y_n)\notin T_k)} \quad (15)$$

那么$D_1(\theta_k)$、$D_2(\theta_k)$分别是概率$P_{X,Y}(h(X,\theta_k)=Y)$、$P_{X,Y}\left(h(X,\theta_k)=\hat{j}(X,Y)\right)$的袋外估计。

估计$var_{X,Y}\,mr(X,Y)$如下：

$$var_{X,Y}mr(X,Y)=E_{X,Y}\left(P_\theta(h(X,\theta)=Y)-\max_{j\neq Y}P_\theta(h(X,\theta)=j)\right)^2-s^2$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left(Q(x_i,y_i)-\max_{j\neq y_i}Q(x_i,j)\right)^2-\left(\frac{1}{N}\sum_{i=1}^{N}\left(Q(x_i,y_i)-\max_{j\neq y_i}Q(x_i,j)\right)\right)^2$$

估计 $E_\theta sd(\theta)$ 如下：

$$E_\theta sd(\theta)=E_\theta\sqrt{var_{X,Y}rmg(\theta,X,Y)}$$

根据(5)式可得$rmg(\theta,X,Y)$的分布律如下：

$rmg(\theta,X,Y)$	p
1	$p_1:=P_{X,Y}(h(X,\theta)=Y)$
-1	$p_2:=P_{X,Y}\left(h(X,\theta)=\hat{j}(X,Y)\right)$
0	$p_3:=1-p_1-p_2$

于是可以计算$rmg(\theta,X,Y)$的方差为：

$$var_{X,Y}rmg(\theta,X,Y)$$
$$=E_{X,Y}\left(rmg(\theta,X,Y)\right)^2-\left(E_{X,Y}rmg(\theta,X,Y)\right)^2$$
$$=p_1+p_2-(p_1-p_2)^2$$

将$var_{X,Y}rmg(\theta,X,Y)$、$D_1(\theta_k)$、$D_2(\theta_k)$代入$E_\theta sd(\theta)$，可得：

$$E_\theta sd(\theta)=E_\theta\sqrt{p_1+p_2-(p_1-p_2)^2}$$
$$=\frac{1}{K}\sum_{k=1}^{K}\sqrt{D_1(\theta_k)+D_2(\theta_k)-\left(D_1(\theta_k)-D_2(\theta_k)\right)^2}$$

4 随机森林改进方法

在处理海量数据时，由于高维属性数据中，不同的类别与不同的属性子集相关，且大量属性与类别无关，对分类没有帮助，这样的情况下如果采用随机森林算法，在决策树的每个结点处，等概率地随机抽取属性子集，会导致在属性子集中，包含许多与分类无关的属性，最终将降低决策树的分类强度，使得泛化误差界增大，从而随机森林的精度下降。此时如果要控制分类错误率，需要增大在每个节点抽取的属性的个数，使得抽取到更多的有用属性，而这将带来较大的计算开销。

许保勋在其论文中提出了一种属性子集的加权抽样方法[19]，提高了与分类相关的属性的抽取概率。具体的，在每个待分类的节点处，计算各个可选属性与类别的相关性大小，据此赋予属性不同的抽取概率，相关性大的属性被抽取的概率越大，此时的属性子集是通过非等概率随机抽样得到的。这种抽样方法的好处是得到的属性子集中，包含与分类有关的属性的概率增大，从而使得构建的决策树有更好的分类强度。

该算法通过计算χ^2统计量来赋予各属性不同的抽取概率。设类别集Y有p个不同取值$\{y_1,y_2,\cdots,y_q\}$，属性A有p个不同取值$\{a_1,a_2,\cdots,a_p\}$，val_{ij}表示样本集中满足$A=a_i$ 且Y=y_j的样本个数，记：

$$val_{i.}=\sum_{j=1}^{q}val_{ij}\ ,\ val_{.j}=\sum_{i=1}^{p}val_{ij}\ ,\ val=\sum_{i=1}^{p}\sum_{j=1}^{q}val_{ij}$$

若属性A与类别Y独立，那么有：

$$P(A=a_i,Y=y_j)=P(A=a_i)\cdot P(Y=y_j)$$

即：

$$\frac{val_{ij}}{val}=\frac{val_{i.}}{val}\cdot\frac{val_{.j}}{val}$$

检验属性A与类别Y的独立性可以构建χ^2统计量：

$$\chi^2(A,Y)=\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(val_{ij}-t_{ij}\right)^2}{t_{ij}}$$

其中 $t_{ij}=\frac{val_{i.}\times val_{.j}}{val}$

χ^2统计量越大，相关性越大，表示属性对分类越有帮助。构建决策树时，在每个待分裂的节点处，计算各个特征对应的χ^2统计量，为了将权重作为特征抽取的概率，需要将χ^2统计量进行规范化，特征A_i对应的权重为：

$$w_i=\frac{\sqrt{\chi^2(A_i,Y)}}{\sum_i\sqrt{\chi^2(A_i,Y)}}$$

实验结果表明，改进后的算法，在处理高维属性数据时，分类精度明显好于原随机森林算法，对低维数据处理结果虽然没有明显提高，但是也不会比原随机森林算法的结果差。

5 结 语

随机森林是现在研究较多一种数据挖掘算法，由于其良好的性能表现，在现实生活中也获得了广泛的应用。目前多数研究都是针对随机森林算法做出改进，对随机森林理论的研究尚不够深入。本文主要对

Breiman在论文中提出的随机森林理论进行解释，包括收敛定理、泛化误差界以及袋外估计三大部分，并介绍了一种随机森林的改进算法。下一步的工作是希望对这种改进算法的在效能上的提高进行理论上的说明。

参 考 文 献

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

[2] Breiman L, Friedman J H, Olshen R A, et al. Classification and regression trees. Monterey, CA: wadsworth & brooks/cole advanced books & software, 1984.

[3] Quinlan J R. C4.5: Programs for machine learning [J]. Morgan Kaufmann Publishers, 1993: 302.

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

[5] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[6] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24 (2): 123–140.

[7] Ho T. The random subspace method for constructing decision forests [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20 (8): 832-844.

[8] Chen X, Liu M. Prediction of protein—protein interactions using random decision forest framework [J]. Bioinformatics, 2005, 21(24): 4394-4400.

[9] Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes [J]. Bioinformatics, 2010, 26(2): 250-258.

[10] Ward M, Pajevic S, Dreyfuss J, et al. Short-term prediction of mortality in patients with systemic lupus erythematosus: classification of outcomes using random forests [J]. Arthritis and Rheumatism, 2006, 55(1): 74-80.

[11] Kim S, Lee J, Ko B, et al. X-ray image classification using random forests with local binary patterns [C] // In proceedings of the 9th International Conference on Machine Learning and Cybernetics. Qingdao, China: IEEE Computer Society, 2010: 3190-3194.

[12] Ying W, Li X, Xie Y, et al. Preventing customer churn by using random forests modeling [C] // In proceedings of the 7th IEEE international Conference on Information Reuse and Integration. Las Vegas, USA: IEEE Computer Society, 2008: 429-434.

[13] Xie Y, Li X, Ngai E, et al. Customer churn prediction using improved balanced random forests [J]. Expert Systems with Applications, 2009, 36(3): 5445-5449.

[14] Oparin I, Glembek O, Burget L, et al. Morphological random forests for language modeling of inflectional languages [C] // In proceedings of the 2nd IEEE Workshop on Spoken Language Technology. Goa, India: IEEE Computer Society, 2008: 189-192.

[15] 张华伟, 王明文, 甘丽新. 及于随机森林的文本分类模型研究 [J]. 山东大学学报: 理学版, 2006, 41(3): 5-9.

[16] 方匡南, 朱建平. 基于随机森林方法的基金超额收益方向预测与交易策略研究 [J]. 经济经纬, 2010(2): 61-65.

[17] Han J W, Kamber M. 数据挖掘概念与技术 [M]. 范明, 孟小峰译. 第二版. 北京: 机械工业出版社, 2001.

[18] Richard O D. 模式分类 [M]. 李宏东, 姚天翔等译. 第二版. 北京: 机械工业出版社, 2003: 321.

[19] Xu B X, Huang Z X, Williams, et al. Classifying very high-dimensional data with random forests built from small subspaces [J]. International Journal of Data Warehousing and Mining, 2012, 8 (2): 44-63.