第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)15—0193—02 文献标识码：A 中图分类号：TP391

基于粗糙集约简的决策林构建方法

王名扬[1,2]，胡清华[2]，于达仁[2]
(1. 东北林业大学信息与计算机工程学院，哈尔滨 150040；2. 哈尔滨工业大学航天学院，哈尔滨 150001)

摘 要：针对如何提高决策林的分类精度问题，提出一种基于粗糙集约简构建决策林的技术，包括基于逐次数据约简构建粗糙决策林和基于遗传算法构建粗糙决策林。对3个UCI数据集的验证表明，基于遗传算法构建的粗糙决策林获得了更好的分类效果。
关键词：决策林；粗糙集；约简；遗传算法

Construction Method of Decision Forests Based on Rough Set Reduction

WANG Ming-yang[1,2], HU Qing-hua[2], YU Da-ren[2]
(1. College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040;
2. School of Astronautic, Harbin Institute of Technology, Harbin 150001)

【Abstract】This paper proposes a technique to construct decision forests based on rough set reduction to enhance the classification performance of decision forests. It includes two methods: one is based on sequentially data reduction to construct rough decision forests, the other is based on genetic algorithm to construct rough decision forests. Experimental results in three data sets of UCI show that the rough decision forests constructed by genetic algorithm get better classification performances.
【Key words】decision forests; rough set; reduction; genetic algorithm

1 概述

决策林是近年来兴起的一种分类机制，它将多个决策树进行融合以提高分类精度。相对于单个决策树而言，决策林通常具有较高的预测和泛化能力，因此，在实际中得到了广泛的应用。决策林的构建一般有如下2种方法：

(1)重采样法。这种方法利用原始样本集的子集训练决策树，有2种划分样本子集的方法：Bagging[1]和Boosting[2]。其中，Bagging是一种随机选取样本子集的方法。而在Boosting方法中，那些在上一个样本子集中被错分的样本将以较大的概率出现在下一个样本子集中。

(2)子空间法。这种方法将原始样本集的特征空间划分为几个子空间，并分别利用这些特征子空间训练决策树。文献[3]提出了一种随机选择子空间的方法，每个特征子空间大小相等，其中包含的特征都是从原始特征空间中随机选择出来的。这种方法构建的决策林称为随机决策林(Random Decision Forests, RaDF)。

随着特征数量的增加，特征间的组合会呈指数形式增长，在众多的特征子空间中寻找最合适的特征子空间非常困难，同时，决策林的分类性能不仅取决于单个决策树的分类精度，而且也取决于决策树间的差异性。差异性大的决策树能够提供更多的信息，因此，决策树间的差异性对决策林的分类性能至关重要，由差异性大的决策树构建的决策林往往比差异较小的决策树构建的决策林更能够提高分类的精度。

本文提出一种构建决策林的机制。由于这种机制基于粗糙集理论，因此称为粗糙决策林(Rough Decision Forests, RoDF)。这种机制首先通过粗糙集约简产生一系列的特征子集，每个特征子集都将被作为训练子空间用于训练决策树。根据粗糙集理论，每个粗糙约简子集都具有与原始数据集相同的分类能力。因此，粗糙决策林保证了每个决策树都具有较高的分类精度。但是，在实际应用中，粗糙集产生的约简有几十个甚至上百个。因此，在构建决策林时，选择合适的约简或者合适的决策树是一个非常基本且重要的问题。

本文提出了2种选择约简或者决策树的方法，分别从最大化决策树的输入差异和最大化决策树的输出差异2个角度分析决策树间的差异性对决策林性能的影响。

2 构建决策林的2种方法

要构建粗糙决策林，首先需要对原始数据集进行约简。基于粗糙集理论的维度约简技术已经非常成熟，研究人员提出了各种不同的贪心算法来搜索粗糙约简集合[4-5]。本文利用粗糙集约简软件ROSETTA(可以从http://rosetta.lcb.uu.se/general/download/上下载)对数据集进行约简。然后，每个约简被作为训练特征子空间去训练决策树，常见的训练决策树方法有如下几种：CART, ID3, C4.5, See5.0。这些方法在不同的应用中均取得了较好的性能，本文利用CART技术训练决策树。最后，选择合适的约简或者决策树，将它们融合起来构建决策林。本文提出了2种选择约简或者决策树的方法。

2.1 逐次约简方法

决策林的性能不仅取决于系统中单个决策树的分类能力，而且与它们间的差异性有关。通过这种方法能够产生完全不同的训练特征子空间，从而保证每个决策树都具有完全不同的输入，以此来提高决策树间的差异度。(1)该方法递归

基金项目：东北林业大学青年科研基金资助项目(07024)
作者简介：王名扬(1980－)，女，在职博士研究生，主研方向：智能信息处理，数据挖掘；胡清华，副教授、博士；于达仁，教授、博士生导师
收稿日期：2009-01-12 **E-mail：**wangmingyang@hcms.hit.edu.cn

地对数据集进行约简，以产生完全不同的最小约简。在该方法的每一步，已经产生的最小约简包含的特征都要在原始特征空间中排除掉，并利用剩余的特征空间产生一个最小约简。(2)利用每一步得到的最小约简训练决策树，由于每个最小约简包含的特征都完全不同，因此每个决策树都具有完全不同的输入特征空间。(3)将训练得到的决策树进行融合，通过多数投票法得到最终的分类结果。该方法的示意图见图 1。图中，S 代表原始的特征空间，它包含了原始数据集中所有的特征；R_i 代表在第 i 步产生的最小约简；T_i 代表由最小约简 R_i 训练出的决策树。

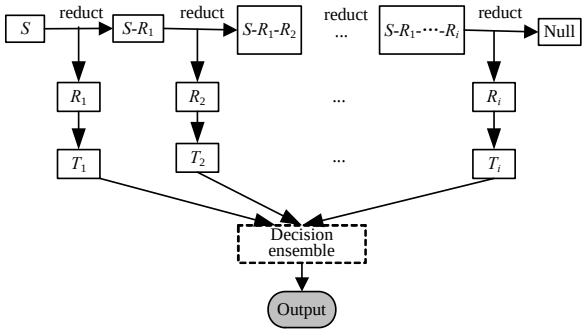

图 1　逐次数据约简算法构建决策林的示意图

2.2　基于遗传算法的最优决策树组合选择

在构建决策林时，从众多的决策树中选择合适的决策树组合是一个典型的组合优化问题。遗传算法是处理这类问题的一个有效方法。经过一系列的遗传、交叉和变异操作后，遗传算法能够在群体中找到适应能力强的优秀个体。

本文利用遗传算法在众多的决策树中选择最优的决策树组合，使该组合能产生最优的分类性能。(1)利用粗糙集约简技术从原始样本集中找到所有的约简；(2)每个约简都被用作训练决策树的训练子空间；(3)利用遗传算法从所有训练得到的决策树中选择最优的决策树组合；(4)通过多数投票方法对选出的决策树的结果进行汇总，得到最终的分类结果。

遗传算法可以定义为一个 8 元组：

$$SGA = (C, E, P_0, M, \Phi, \Gamma, \Psi, T)$$

其中，C 为对个体的编码，采取的是二进制编码；E 为适应度函数，本文利用系统的分类精度作为适应度函数；P_0 为初始群体；M 为群体的大小；Φ 为选择算子；Γ 为交叉算子；Ψ 为突变算子；T 为终止条件。当达到规定的最大迭代数或者进一步迭代对精度改善没有太大帮助时，算法就停止。具体过程如图 2 所示。

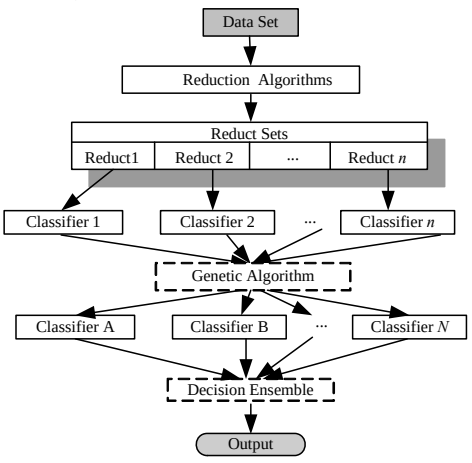

图 2　基于遗传算法选择决策树构建决策林的示意图

3　实验

在划分子空间的方法中，随机子空间方法比较常用。本文以随机子空间法构造的随机决策林为基准对上述 2 种构建方法的分类能力进行了验证。

表 1 给出了从 UCI 数据集中选择的 3 组数据样本的基本信息，包括样本的数量、属性的数量和根据原始样本集得到的约简数量。可以看出，在 3 个数据集中，粗糙集约简得到的约简数量最少为 135 个，最多为 211 个，说明在实际中，每个数据集都可能产生大量的约简。如果仅仅利用其中的一个约简进行预测，将损失那些隐含在其他约简中的重要信息。本文提出的粗糙决策林通过将多个约简训练得到的决策树进行融合，充分考虑了多个约简中包含的重要信息。

表 1　数据样本基本信息

数据集名称	缩写	样本	属性	约简
Ionosphere database	Ionos	351	35	194
Wisconsin diagnostic breast cancer	WDBC	569	32	211
Wine recognition data	Wine	168	13	135

利用这 3 组数据集对本文的 2 种构建粗糙决策林的方法与随机决策林方法进行了对照。在实验中，利用 CART 算法训练决策树，其中，将每个特征子集对应样本的 2/3 作为训练集，剩余的 1/3 作为检验集。最终得到的结果如表 2 所示。其中，Orig.代表根据原始数据集训练得到的决策树的分类精度；RaRDF 代表由随机选择子空间方法构建的随机决策林的分类精度；SeRDF 表示由逐次数据约简方法构建的粗糙决策林的分类精度；GARDF 表示由遗传算法选择方法构建的粗糙决策林的分类精度。

表 2　不同方法的分类结果

数据集	Orig.	RaRDT	SeRDT	GARDT
Ionos	0.813	0.921	0.980	0.990
WDBC	0.876	0.905	0.935	0.971
Wine	0.708	0.830	0.896	1.000
Aver.	0.799	0.885	0.937	0.987

从表 2 可见，对所有数据集而言，直接由原始数据集训练得到的决策树的分类效果最差。基于遗传算法选择方法构建的粗糙决策林获得了最好的分类精度，分类精度相对于原始决策树增长了约 20%。由逐次数据约简方法构建的决策林相对于随机决策林也取得了较好的分类效果，但略差于基于遗传算法的粗糙决策林构建方法。这说明，用于构建决策树的特征子空间之间的差异对决策林的分类效果有非常大的影响，但并不是绝对影响。为更清晰地展示实验结果，图 3 给出了随机决策林及 2 种粗糙决策林构建方法的分类结果。

图 3　3 种决策林构建方法的分类精度

从前面的分析可知，基于逐次数据约简的方法保证了决策树输入间最大的差异性，但没有获得最好的分类效果。为

（下转第 197 页）