文章编号:1000-5862(2014)03-0278-04

基于粗糙集和随机森林算法辅助糖尿病并发症分类研究

聂 斌[1],王 卓[2],杜建强[1],朱明峰[1],林剑鸣[1],艾国平[1],熊玲珠[1]
(1. 江西中医药大学计算机学院,江西 南昌 330004;2. 南昌大学软件学院,江西 南昌 330047)

摘要:提出基于粗糙集和随机森林算法辅助糖尿病并发症分类. 首先,运用简化的分明矩阵法对属性约简,产生新的决策信息系统;其次,采用随机森林算法对该新信息系统生成随机森林,实现分类;最后,通过糖尿病并发症临床诊断数据子集测试. 实验表明该方法有效性,并优于直接用随机森林算法分类.

关键词:粗糙集;随机森林;糖尿病;并发症
中图分类号:TP 391 **文献标志码**:A

0 引言

随机森林[1-5]采用随机选择样本和随机选择特征,具有许多优点:比单棵的决策树更稳健,泛化性能好;具有较好的抗噪声能力,抗异常值,自动处理缺失值以及能应对不平衡数据;对规模参数不敏感,对于不相关和冗余特征不敏感,能应对特征比分类少的情况;对于多种资料,可以产生高准确度的分类;不必担心过度拟合;能够估计哪个特征在分类中更重要,也可侦测偏离者;在训练过程中,能够检测到特征间的互相影响;袋外数据 K 折交叉验证可提供高度可靠的评估模型;算法容易理解、快速. 随机森林成功应用于文本分类[6]、基因识别[7]、目标定位[8]、流量分类[9]等. 然而数据的高噪对随机森林分类有较大影响,而糖尿病并发症数据来源于临床确诊病例等,具有高噪性特点.

粗糙集理论[10]是一种处理模糊和不确定性知识的数学工具,其主要思想是在保持分类能力不变的前提下,通过知识约简[11-20],导出问题的决策或分类规则. 粗糙集的约简去噪功能,对数据预处理起到较好的作用. 在实际应用中,粗糙集结合随机森林[21],发挥了 2 种理论算法的优势,取得了较好的效果. 本文结合粗糙集约简和随机森林的优点,用于辅助糖尿病并发症分类,并通过实验数据测试该方法的有效性.

1 粗糙集基本理论

1.1 信息系统

信息系统[10,17]可以定义为 $S=(U,A\cup D)$,其中 U 是一个非空、有穷、称为全域的个体的集合,A 是非空、有穷条件属性集合,即对于属性 $a\in A$,有 $a:U\to V_a$,其中 V_a 是属性 a 的值集;集合 $V=\bigvee_{a\in A}V_a$ 称为属性 A 的值区域,D 称为决策属性集. 表 1 为信息系统,$U=\{1,2,3,4,5,6\}$,条件属性集 $A=\{$albuminuria,retina,skin,angina,hypertension,myopia$\}$,决策属性集 class 为{eye,kidney,angiocarpy}.

表 1 糖尿病并发症数据子集信息表

No	albuminuria	retina	skin	angina	hypertension	myopia	class
1	0	1	1	0	0	1	eye
2	0	1	1	0	0	1	eye
3	0	1	0	0	1	0	eye
4	1	0	1	0	1	0	kidney
5	1	0	0	0	0	1	kidney
6	0	0	0	1	1	1	angiocarpy

注:albuminuria,蛋白尿症;retina,视网膜病变;skin(skin-eruptions),皮肤溃烂;angina,心绞痛;hypertension,高血压;myopia,近视;class(diabetic-complication),糖尿病并发症类型;1,不正常;0,正常;eye,糖尿病性眼病;kidney,糖尿病性肾病;angiocarpy,糖尿病性心血管病.

收稿日期:2013-11-02
基金项目:国家自然科学基金(81160424),江西省自然科学基金(20122BAB205083)和江西省教育厅科技计划(GJJ11541,GJJ13014)资助项目.
作者简介:聂 斌(1972-),男,江西峡江人,讲师,主要从事中医信息学、数据挖掘和人工智能方面的研究.

1.2　属性约简

对于一个决策信息系统,条件属性中的各个属性间存在某种程度的关联或依赖,相对于决策属性,某些条件属性可能是冗余的. 对属性的约简[17]是用最基本的条件属性集代替原来的条件属性集,而不影响原有的功能.

定义 1　设 R 是一个等价关系族,$r \in R$,若 $I_{ND}(R) = I_{ND}(R - \{r\})$,则称 r 在 R 中是可被约去的知识; 若 $P = R - \{r\}$ 是独立的,则 P 是 R 中一个约简.

定义 2　若任一个 $r \in R$ 是 R 中不可约去的,则等价关系族 R 是独立的; 否则 R 是相关的.

定义 3　R 中所有不可约去的关系称为核,由它构成的集合称为 R 的核集,记成 $C_{ORE}(R)$.

简化分明矩阵属性约简法[17,22],是一边从信息系统 $S = (U, A \cup D)$ 中提取关于属性值是分明的属性并构造成分明合取范式,一边作这种逻辑公式的等价变换,直接得到最小析取范式,它对应于该信息表的约简.

2　随机森林分类

2.1　随机森林分类理论简介

随机森林采用随机选择样本和随机选择特征,生成多个决策树$\{h(\boldsymbol{x}, \boldsymbol{\theta}_k)\}$组成的分类器,其中$\{\boldsymbol{\theta}_k\}$为相互独立且同分布的随机向量. 最终由所有决策树投票综合决定输入向量 $\boldsymbol{x}$ 的最终类标签.

为了构造 k 棵树,需要先产生 k 个随机向量 $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_k$,这些随机向量 $\boldsymbol{\theta}_i$ 是相互独立的,并且是同分布的. 随机向量 $\boldsymbol{\theta}_i$ 用于构造决策分类树 $h(\boldsymbol{x}, \boldsymbol{\theta}_i)$,简化为 $h_i(\boldsymbol{x})$. 在构造树的过程中,按照节点不纯度最小的原则从特征中随机选取一个特征进行分支生长.

设 I 为示标函数,$n_{h_i,c}$ 是树 h_i 对类 C 的分类结果,n_{h_i} 是树 h_i 的叶子节点个数,对测试样本 $\boldsymbol{x}$,预测类标签 c_p 为

$$c_p = \arg\max_c \left(\frac{1}{k} \sum_{i=1}^{k} I\left(\frac{n_{h_i,c}}{n_{h_i}} \right) \right). \tag{1}$$

2.2　Gini 不纯度

设 $p(w_i)$ 为节点 n 上属于 w_j 类样本个数占训练样本总数的频率,$G(n)$ 为节点 n 的 Gini 不纯度,

$$G(n) = \sum_{i \neq j} p(w_i) p(w_j) = 1 - \sum_{i \neq j} p^2(w_j). \tag{2}$$

当节点 n 上分类数据全部来自于同一类别时,则此节点的不纯度 $G(n) = 0$; 如果分类数据服从均匀分布,则不纯度较大.

3　基于粗糙集的随机森林分类方法

3.1　算法思想

基于有放回属性方法的自助法重采样技术生成多个树分类器及回归树,其步骤如下: (i) 采用简化的分明矩阵法约简,生成新的决策信息表,作为原始训练样本. (ii) 从容量为 N 的原始训练样本集合中采取有放回抽样的方法随机抽取 k 个自助样本集. (iii) 每个自助样本集是每棵分类树的全部训练数据. 每个自助生长成为单棵分类树,并采用递归方法依次向下生长. 在树的每个节点处,从 M 个属性中随机选取 m 个属性(取 $m = \sqrt{M}$). 按照节点不纯度最小的原则从这个 m 特征中选取一个属性进行分支生长. 在整个森林的生长过程中,m 将保持恒定. (iv) 每棵分类树充分生长,并按"Gini 不纯度"下降差作为停止生长原则,不进行剪枝操作. (v) 根据生成的多个树分类器对新的数据进行预测,分类结果按每棵树分类器的投票多少来定.

3.2　算法流程图

算法流程图如图 1 所示.

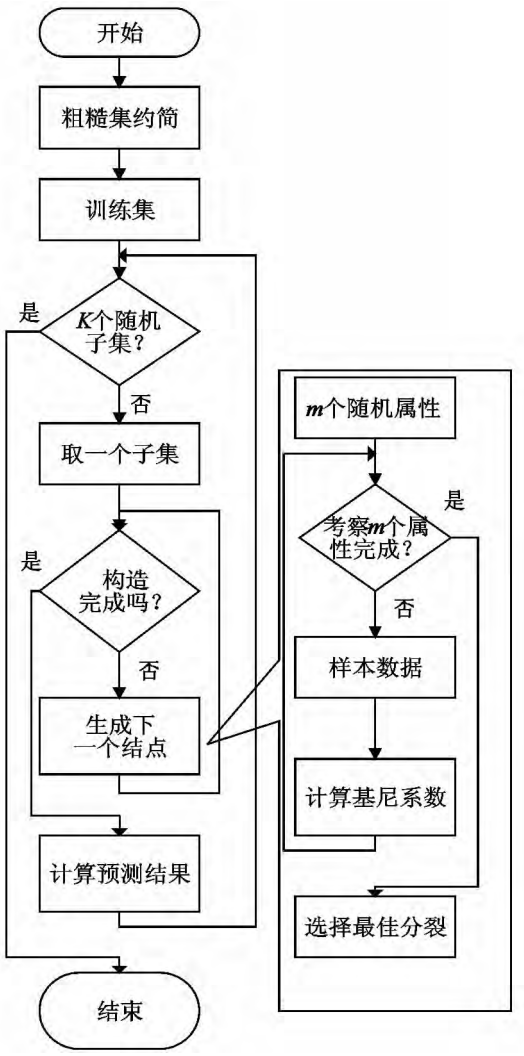


图 1　算法流程图

4 实验结果和实验验证

4.1 实验验证

采用表 1 数据为训练集,表 2 数据为测试集.

表 2 糖尿病并发症数据测试信息表

No	album-inuria	retina	skin	angina	hypert-ension	myo-pia	class
1	0	1	0	0	1	1	eye
2	1	0	1	0	1	0	kidney
3	1	0	0	0	0	1	kidney
4	0	0	0	1	1	1	angio-carpy

实验验证步骤:(i) 采用简化的分明矩阵法对表 1 进行约简,形成新的训练集信息表;(ii) 采用随机森林算法分别对训练集和约简后的信息分类,并进行自测和对测试集预测,运用 R 语言实现;(iii) 分析与比较实验结果.

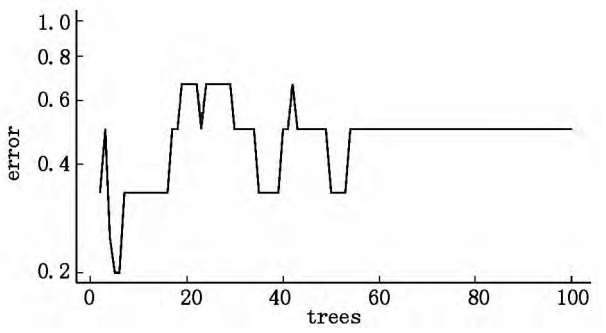
4.2 实验结果

(i) 采用简化的分明矩阵法对表 1 进行约简,得到决策信息表如表 3 所示.

表 3 约简后的糖尿病并发症数据训练信息表

No	album-inuria	retina	skin	angina	class
1	0	1	1	0	eye
2	0	1	1	0	eye
3	0	1	0	0	eye
4	1	0	1	0	kidney
5	1	0	0	0	kidney
6	0	0	0	1	angioc-arpy

(ii) 随机森林算法直接对表 1 数据生成随机森林. 随机森林算法直接对表 1 数据生成随机森林,其分类效果由袋外数据估计的错误率、生成分类决策树的数量和错误率的对应关系,以及对新数据的预测来评价. 通常认为袋外数据估计的错误率越低,表明生成的随机森林越好,分类效果越好;错误率低及生成较少分类决策树时的效果好,或者说错误率低并受生成决策树数量多少影响较小时,效果好;对新数据预测准确度越高,分类效果越好. 袋外数据估计,eye-糖尿病性眼病,kidney-糖尿病性肾病,angio-carpy-糖尿病性心血管病,共 3 类疾病 6 个病例,其中 1 例 angiocarpy 分类结果错误,3 例 eye 分类结果正确,2 例 kidney 中有 1 例错误,总体分类结果错误率为 33.33%,即准确度为 66.67%;图 2 为随机森林对未约简数据分类时分类决策树的数量与错误率对应图,表明当生成分类决策树的数量约为 5 时,错误率约为 20%,另外有几个平台期错误率保持在 33% 左右,当分类决策树的数量约为 55 时,错误率在 50% 以上. 该随机森林分类对表 2 数据预测,预测结果准确率 100%.



注:横坐标 trees 为随机森林中分类决策树多少,纵坐标 error 为袋外数据估计的错误率,实线表明两者对应的结果.

图 2 随机森林对未约简数据分类时树的数量与错误率对应图

(iii) 随机森林算法对约简表 3 数据生成随机森林. 随机森林算法直接对表 3 数据生成随机森林,其分类效果由袋外数据估计的错误率、生成分类决策树的数量和错误率对应关系,以及对新数据的预测来评价. 为袋外数据估计,共 6 例新数据,其中只有 1 例 angiocarpy 分类结果错误,总体分类结果错误率为 16.67%,即准确度为 83.3%;图 3 为随机森林对约简数据分类时决策树的数量与错误率对应图,表明当分类决策树的数量为 7 ~ 10 时错误率约为 5%,另外有几个平台期错误率保持在 5% 左右的较低水平,当树的数量约为 65 时,错误率在 5% 左右恒定;对表 2 预测结果准确率为 100%,考虑到篇幅,图略.

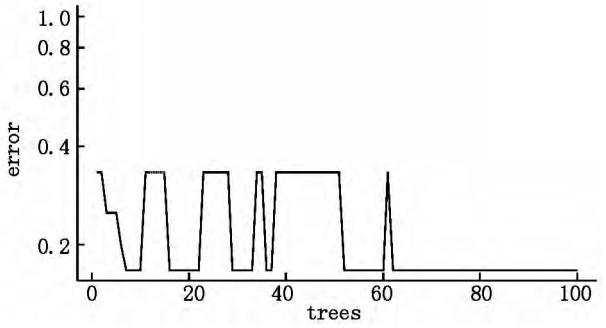


图 3 随机森林对约简数据分类时树的数量与错误率对应图

(iv) 实验结果比较. 随机森林能对糖尿病数据子集实现分类;经粗糙集约简后,采用随机森林分类的准确度总体从 66.7% 提高到 83.3%,并且无论分类决策树数量为多少时的各个时期相对保持在较高准确度,决策树达到 65 左右,错误率稳定在 5% 较低平台期,准确度达 95% 左右.

表 4 分类结果比较 %

方法	训练集袋外数据估计		测试集最小错误率
	预测准确率	总体错误率	
随机森林	33.33	20	100
粗糙集和随机森林	16.67	5	100

5　结论

本文结合粗糙集理论和随机森林算法实现分类,借助 R 语言实现. 通过糖尿病并发症子集数据测试表明该方法有效性,并优于直接用随机森林算法分类. 下一步工作将继续研究粗糙集理论和随机森林算法有机融合,对复杂数据分类.

6　参考文献

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

[2] Berk R A. Random Forests [EB/OL]. [2013-06-11]. http://link.springer.com/search? query = Random + Forests

[3] 雷震. 随机森林及其在遥感影像处理中应用研究 [D]. 上海: 上海交通大学,2012.

[4] Salford Systems. What are the advantages of RandomForests? [EB/OL]. [2013-06-11]. http://www.salford-systems.com/en/products/randomforests/faqs/item/134-what-are-the-advantages-of-randomforests?

[5] Liu Miao,Wang Mingjun,Wang Jun,et al. Comparison of random forest,support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar [J]. Sensors and Actuators B: Chemical,2013,177: 970-980.

[6] 张华伟,王明文,甘丽新. 基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(3): 139-143.

[7] 郭颖婕,刘晓燕,郭茂祖,等. 植物抗性基因识别中的随机森林分类方法 [J]. 计算机科学与探索,2012(1): 67-77.

[8] 刘足华,熊惠霖. 基于随机森林的目标检测与定位 [J]. 计算机工程,2012,13: 5-8.

[9] 张建,武东英,刘慧生. 基于随机森林的流量分类方法 [J]. 信息工程大学学报,2012(5): 621-625.

[10] Zdzisław Pawlak. Rough sets [J]. International Journal of Computer and Information Science,1982,11(5): 341-356.

[11] 苗夺谦,胡桂荣. 知识约简的一种启发式算法 [J]. 计算机研究与发展,1999,36(6): 681-684.

[12] 官礼和,王国胤. 决策表属性约简集的增量式更新算法 [J]. 计算机科学与探索,2010,4(5): 436-444.

[13] 陈林,邓大勇,闫电勋. 基于属性重要度并行约简算法的优化 [J]. 南京大学学报: 自然科学版,2012,48(4): 376-382.

[14] 王名扬,于达仁,胡清华. 基于粗糙集约简的多分类器系统构造方法 [J]. 计算机工程与应用,2010,46(3): 49-50,93.

[15] 蒙祖强,史忠植. 一种新的启发式知识约简算法 [J]. 小型微型计算机系统,2009,30(7): 1249-1255.

[16] Yao Yiyu,Zhao Yan. Discernibility matrix simplication for con-structing attribute reducts [J]. Information Sciences,2009,179(7): 867-882.

[17] 刘清. Rough 集及 Rough 推理 [M]. 北京: 科学出版社,2001.

[18] 田卫东,周创德,胡学钢,等. 基于简化分辨矩阵的粗糙集属性约简算法 [J]. 计算机科学,2008,35(3): 209-212.

[19] 周丽,吴根秀,晏伟峰,等. 一种基于区分矩阵的实值属性约简算法 [J]. 江西师范大学学报: 自然科学版,2011,36(2): 135-139.

[20] 吴根秀,王功,纪军,等. 多值信息系统的基于相似度的粗糙集模型 [J]. 江西师范大学学报: 自然科学版,2011,36(1): 88-90.

[21] Yeh C C,Lin Fengyi,Hsu C C. A hybrid KMV model,random forests and rough set theory approach for credit rating [J]. Knowledge-Based Systems,2012,33,166-172.

[22] 聂斌,王命延,于海雯,等. 基于 Rough 集从临床数据中提取诊断规则 [J]. 南昌大学学报: 理科版,2008,32(2): 193-197.

The Study on Classification of Secondary Complications of Diabetes Based on Rough Set and Random Forest

NIE Bin[1],WANG Zhuo[2],DU Jian-qiang[1],ZHU Ming-feng[1],LIN Jian-ming[1],AI Guo-ping[1],XIONG Lin-zhu[1]

(1. School of Computer,Jiangxi University of Traditional Chinese Medicine,Nanchang Jiangxi 330004,China;
2. Software School,Nanchang University,Nanchang Jiangxi 330047,China)

Abstract: The paper study on classification of secondary complications of diabetes based on rough set and random forest. First,using the simplified matrix method for attribute reduction,clearly have new decision information system; second,by random forest algorithm to generate random forests for the new information system,and gain the classification result. The method has been proved feasible and effective after test the database,it´s classification result are better than directly with random forest algorithm.

Key words: rough sets; random forest; diabetes; complications　　(责任编辑: 冉小晓)