

# 基于遗传算法的朴素贝叶斯分类

胡为成<sup>1,2</sup>, 胡学钢<sup>1</sup>

(1. 合肥工业大学 计算机学院, 安徽 合肥 230009;

2. 铜陵学院 计算机系, 安徽 铜陵 244000)

**摘要:**朴素贝叶斯分类器是一种简单而高效的分类器,但是其属性独立性假设限制了对实际数据的应用。提出一种新的算法,该算法为避免数据预处理时,训练集的噪声及数据规模使属性约简的效果不太理想,并进而影响分类效果,在训练集上通过随机属性选取生成若干属性子集,并以这些子集构建相应的贝叶斯分类器,进而采用遗传算法进行优选。实验表明,与传统的朴素贝叶斯方法相比,该方法具有更好的分类精度。

**关键词:**数据挖掘;朴素贝叶斯;遗传算法;属性约简;适应度函数

**中图分类号:** TP301

**文献标识码:** A

**文章编号:** 1673-629X(2007)01-0030-03

## Naive Bayes Classification Based on Genetic Algorithms

HU Wei-cheng<sup>1,2</sup>, HU Xue-gang<sup>1</sup>

(1. College of Computer Science, Hefei Technology University, Hefei 230009, China;

2. Department of Computer Science, Tongling College, Tongling 244000, China)

**Abstract:** Naive Bayes classifier is a simple and effective classification method, but its attribute independence assumption makes it unable to express the dependence among attributes in the real world. To avoid the direct influence of feature reduction from data pre-processing on the performance of classification, a new algorithm is introduced in this paper. It makes use of random feature selection to generate several feature subsets from the whole training set, and constructs Bayesian classifiers with the feature subsets, and then optimizes the Bayesian classifiers by using genetic algorithms. Compared with the traditional Naive Bayes methods, the algorithm has better classification precision.

**Key words:** data mining; Naive Bayes; genetic algorithms; feature reduction; fitness function

## 0 引言

数据挖掘(Data Mining, DM)是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含的、事先未知的、潜在有用的信息的处理过程。分类预测是数据挖掘中的重要分支。分类是找出一组能够描述数据集合典型特征的模型,以便对未知变量能做出预测或分类。分类算法的核心部分是构造分类器。

朴素贝叶斯分类器(Naive Bayes Classifiers, 简称NBC)<sup>[1]</sup>由于计算高效、精确度高,并具有坚实的理论基础而得到广泛的应用<sup>[2,3]</sup>。但由于朴素贝叶斯分类器的条件独立性假设,使得所选数据集的条件属性集在预处理时必须进行属性约简。

遗传算法(Genetic Algorithm, GA)是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化

概率搜索算法,具有较强的鲁棒性,其思想简单,应用广泛<sup>[4]</sup>。笔者结合遗传算法,提出一种基于遗传算法的朴素贝叶斯分类算法,在训练集上通过随机属性选取生成若干属性子集,并以这些子集构建相应的朴素贝叶斯分类器,进而采用遗传算法进行优选,从而避免了属性约简的好坏对分类精度的影响。

## 1 朴素贝叶斯分类器

朴素贝叶斯分类器假定特征向量的各分量间相对于决策变量是相对独立的,并使用概率规则来实现学习或某种推理过程,即将学习或推理的结果表示为随机变量的概率分布,这可以解释为对不同可能性的信任程度。它的理论基础就是贝叶斯定理和贝叶斯假设<sup>[2,3]</sup>。

假定随机向量  $x, \theta$  的联合分布密度是  $p(x, \theta)$ , 它们的边际密度分别为  $p(x), p(\theta)$ 。一般情况下设  $x$  是观测向量,  $\theta$  是未知参数向量,通过观测向量获得未知参数向量的估计,贝叶斯定理记作:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = \frac{p(\theta)p(x|\theta)}{\int p(\theta)p(x|\theta)d\theta} \quad (1)$$

从上式可知,对未知向量的估计综合了它的先验信息

收稿日期:2006-04-24

基金项目:安徽省高等学校自然科学研究重点项目(2006kj027A)

作者简介:胡为成(1975-),男,安徽桐城人,讲师,硕士研究生,主要研究方向为数据挖掘、遗传程序设计等;胡学钢,教授,硕士生导师,主要从事数据挖掘、概念格等方向研究。



和样本信息,这是贝叶斯增量学习模型的基础,可理解为:后验知识( $I_1$ )=先验知识( $I_0$ )+样本信息( $S$ )。当新的样本到来时,上面的后验知识变为先验知识,因而是一个利用样本知识来修正当前知识的连续的动态的过程。

朴素贝叶斯分类器将每个训练样本数据分解成一个  $n$  维特征向量  $\mathbf{X}$  和决策类别变量  $C$ ,并假定特征向量的各分量间相对于决策变量是相对独立的。

设特征向量  $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$  表示数据  $n$  个属性 ( $A_1, A_2, \dots, A_n$ ) 的具体取值,类别变量  $C$  有  $m$  个不同的取值  $C_1, C_2, \dots, C_m$ , 即有  $m$  个不同的类别。则:

$$p(\mathbf{X}|C_k) = p(x_1, x_2, \dots, x_n|C_k) = \prod_{i=1}^n p(x_i|C_k) \quad 1 \leq k \leq m \quad (2)$$

由贝叶斯定理知  $\mathbf{X}$  属于  $C_k$  的后验概率为:

$$p(C_k|\mathbf{X}) = \frac{p(\mathbf{X}|C_k)P(C_k)}{p(\mathbf{X})} \quad 1 \leq k \leq m \quad (3)$$

朴素贝叶斯分类器将未知类别的决策变量  $\mathbf{X}$  归属于类别  $C_k$  当且仅当:

$p(C_k|\mathbf{X}) > p(C_j|\mathbf{X})$  对于  $1 \leq j \leq m, j \neq k$  即  $p(C_k|\mathbf{X})$  最大。

由于  $P(\mathbf{X})$  对于所有类别均是相同的,因此:

$$p(C_k|\mathbf{X}) \propto p(\mathbf{X}|C_k)P(C_k) = P(C_k) \prod_{i=1}^n p(x_i|C_k) \quad 1 \leq k \leq m \quad (4)$$

由于类别的事前概率是未知的,因此,可以假设各类别出现的概率相同,即  $P(C_1) = P(C_2) = \dots = P(C_m)$ 。这样求公式(2)的最大转换为求  $p(\mathbf{X}|C_k)$  最大,否则就要求  $p(\mathbf{X}|C_k)P(C_k)$  的最大。可以通过训练样本数据集合估计  $P(C_k)$  和  $p(x_i|C_k)$  ( $1 \leq i \leq n, 1 \leq k \leq m$ ):

$$P(C_k) = s_k/s \quad (5)$$

$$p(x_i|C_k) = s_{ki}/s_k \quad (6)$$

其中,  $s_k$  为训练样本数据集合中类别为  $C_k$  的样本个数,  $s$  为整个训练样本数据集合的容量。 $s_{ki}$  为训练样本数据集合中类别为  $C_k$  且属性  $A_i$  的取值为  $x_i$  的样本个数。

## 2 遗传算法的基本原理

### (1)遗传算法的基本思想。

遗传算法是从代表问题可能潜在解集的一个种群开始的,而一个种群则由经过基因编码的一个数目的个体组成。每个个体实际上是染色体带有特征的实体。染色体作为遗传物质的主要载体,即多个基因的集合,其内部表现(即基因型)是某种基因组合,它决定了个体形状的外部表现。因此,在一开始需要实现从表现型到基因型的映射即编码工作(如二进制编码)。初代种群产生后,按照适者生存和优胜劣汰的原理,逐代演化产生出越来越好的近似解。在每一代,根据问题域中个体的适应度大小挑选个体,并借助于自然遗传学的遗传算子进行组合交叉和变异,产生出代表新的解集的种群。此过程将导致种群像自然进化一样的后代种群比前代更加适应于环境,末代种群中的最优个体经过解码,可作为问题近似最优解<sup>[4]</sup>。

### (2)交叉操作和变异操作。

交叉运算是指对两个相互配对的染色体按某种方式相互交换部分基因,从而形成两个新的个体。变异模拟了生物进化过程中的基因突变现象,变异算子是以一定的概率改变遗传基因的操作。对个体进行变异,可以保持群体的多样性,增加了自然选择的余地,并使遗传算法跳出局部极值点。为了避免早熟现象,采用自适应方法动态调节交叉概率和变异概率,使得交叉概率  $p_c$  和变异概率  $p_m$  能够随适应度自动改变<sup>[5]</sup>。当种群各个体适应度趋于一致或者趋于局部最优时,使  $p_c$  和  $p_m$  增加;而当群体适应度比较分散时,使  $p_c$  和  $p_m$  减少。同时,对于适应度高于群体平均适应度的个体,对应于较低的和,使该解得以保护进入下一代;而低于平均适应度的个体,相对应于较高的  $p_c$  和  $p_m$ ,使该解被淘汰掉。选取的交叉概率  $p_c$ 、变异概率  $p_m$  计算表达式如下:

$$p_c = \begin{cases} p_{c1} - \frac{(p_{c1} - p_{c2})(f' - f_{avg})}{f_{max} - f_{avg}}, & f' \geq f_{avg} \\ p_{c1}, & f' < f_{avg} \end{cases} \quad (7)$$

$$p_m = \begin{cases} p_{m1} - \frac{(p_{m1} - p_{m2})(f_{max} - f)}{f_{max} - f_{avg}}, & f \geq f_{avg} \\ p_{m1}, & f < f_{avg} \end{cases} \quad (8)$$

式中,  $p_{c1} = 0.9$ ,  $p_{c2} = 0.6$ ,  $p_{m1} = 0.1$ ,  $p_{m2} = 0.001$ ,  $f'$  为要交叉的两个个体中较大的适应度值,  $f$  为要变异个体的适应度值。

### (3)选择操作。

在群体中选取优胜的个体,淘汰劣质的个体的操作称作选择。根据染色体适应度值的大小选择适应性更强的染色体生成新的种群。因此适应度值越大,被选中的概率就越大。

### (4)终止条件。

若经过若干代运算后,仍没有满足用户给定阈值的规则,则结束,输出结果。

## 3 基于遗传算法的朴素贝叶斯分类算法(GABC)

### (1)GABC 编码方式。

文中采用传统的二进制编码方式,每条染色体由一组二进制位构成,长度为数据库中随机属性的个数,每个二进制位依次与数据库中的一个属性相对应。若某个二进制位为 1,则表示数据库对应的属性参与构建朴素贝叶斯分类器。

### (2)GABC 适应度函数。

适应度通常用来度量群体中各个个体在优化计算中有可能达到或接近于找到最优解的优良程度。适应度函数是用来评估个体的适应度,即区分群体中个体好坏的标准。衡量朴素贝叶斯分类器分类效果除了分类精度要高之外,还应考虑分类误差在实例空间中的分布程度,即差异度。文中适应度函数设为:

$$F = R + \lambda D \quad (9)$$



式中,  $R$  为 NBC 在验证集上的分类精度,  $D$  为 NBC 在验证集上的差异度,  $\lambda$  为决定差异度影响的系数。

(3)GABC 算法。

① 采用分层随机取样方法将数据库分成训练集和验证集;

② 随机生成  $S$  个随机属性子集;

③ 将  $S$  个随机属性子集对应的 NBC 作为初始种群, 采用遗传算法优选;

④ 调整  $\lambda$ , 重复步骤③。

算法中, 交叉算子采用两点交叉, 即交换父本两个基因位间的部分, 产生相应的后代; 选择算子采用轮盘赌选择法。另外, 每一代遗传群体中适应度最好的 5% 个体不参加交叉和变异, 自动保留到下一代。

4 实 验

实验在 4 个数据集(来自 UCI 机器学习数据库)上进行<sup>[6]</sup>。对每个数据集, 采用分层随机抽样, 训练集数据占 70%, 验证集占 30%。所谓分层随机抽样, 就是从每个类的实例中随机抽样, 以便结果集中实例的类分布与初始集大致相同。这种方法常常比简单随机取样有更好的精度评估<sup>[7]</sup>。

本实验使用 VC++ 6.0, 在内存为 512MB, CPU 为奔腾 1.7G 的微机上进行。实验中, 遗传种群的规模为 100, 遗传算法执行的最大代数为 150。实验结果如表 1 所示。

表 1 GABC 与 NBC 分类精度的比较

数据集	记录数	属性数	NBC	GABC		
				$\lambda=0$	$\lambda=0.5$	$\lambda=1$
Kr-vs-kp	3 196	36	74.75%	80.36%	80.17%	79.65%
Mushroom	8 124	22	73.49%	79.76%	80.57%	74.86%
Blance	625	4	69.45%	71.57%	69.87%	68.98%
Breast-cancer	286	9	70.84%	73.59%	72.07%	71.45%

由表 1 可知, GABC 在大多数的情况下, 都能比传统的 NBC 取得较好的效果, 只有在 Blance 和 Breast-cancer 等属性数较少的数据集中, 两者的分类精度相近, 这说明基于遗传算法的朴素贝叶斯分类是有很有效的。而且, 对同

一个数据集而言, GABC 的分类精度随着  $\lambda$  的变化而不同, 并从数据集 Mushroom 的实验结果来看, 有时适当地考虑差异度的影响, 可进一步提高分类能力。

5 结 语

基于遗传算法的朴素贝叶斯分类方法不仅避免了属性约简对分类精度的影响, 而且充分考虑了分类误差在实例空间中的分布程度。实验表明, 与传统的朴素贝叶斯方法相比, 该方法具有更好的性能。下一步的研究方向应是:

(1) 如何在一个给定领域里自动选取  $\lambda$  的优化值问题。

(2) 先对遗传算法进行改进, 再和朴素贝叶斯相结合。

改进方法主要包括: 改进遗传算子; 将遗传算法与其他优化算法相结合, 构造混合遗传算法; 使用并行遗传算法等方法。

参考文献:

[1] Written I H, Frank E. Data Mining: Practical Machine learning Tools and Techniques with Java Implementation[M]. Seattle: Morgan Kaufmann Publishers, 2000: 265-314.

[2] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

[3] 朱 明. 数据挖掘[M]. 合肥: 中国科学技术大学出版社, 2002.

[4] 王小平, 曹立明. 遗传算法——理论、应用与软件实现[M]. 西安: 西安交通大学出版社, 2002.

[5] 武兆慧, 张桂娟, 刘希玉. 基于模拟退火遗传算法的关联规则挖掘[J]. 计算机应用, 2005, 25(5): 1009-1011.

[6] Blake C L, Merz C J. UCI repository of machine learning database[EB/OL]. 1998-12-30[2002-09-28]. <http://www.ics.uci.edu/mllearn/MLRepository.html>.

[7] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Mellish C. Proceedings of IJCAI95. San Mateo: Morgan Kaufmann, 1995: 1137-1143.

(上接第 11 页)

模型, 分析它们各自的优缺点。与 DAC 和 MAC 模型相比, RBAC 模型具有明显的优势, 然而仍存在动态性不强等缺陷, TBAC 模型恰可弥补这一缺陷。针对模型的动态性和角色的生命周期约束, 文中结合 RBAC 与 TBAC 模型, 提出了一种 T-RBAC 模型。该 T-RBAC 模型层次分明、授权灵活、维护方便。结合静态授权工具和工作流程引擎, 实现了 B/S 结构 ERP 系统的动态权限控制, 并成功应用于东莞某印刷企业。

参考文献:

[1] 徐日佳, 赵敬中. 一种改进的 RBAC 模型的研究与应用[J].

微机发展, 2005, 15(8): 95-97.

[2] 邓集波, 洪 帆. 基于任务的访问控制模型[J]. 软件学报, 2003, 14(1): 76-79.

[3] 沈海波, 洪 帆. 基于企业环境的访问控制模型[J]. 计算机工程, 2005, 31(14): 144-146.

[4] SEJONG O H, PARK S. An Improved Administration Method on Role-Based Access Control in the Enterprise Environment [J]. Journal of Information Science and Engineering, 2001, 17: 921-944.

[5] 金稼玲, 杨材堂. 基于 T-RBAC 的企业权限管理方法[J]. 计算机工程, 2004, 30(19): 93-95.

[6] 王军强, 杨宏安. 管理信息系统权限控制的组件化研究与实现[J]. 计算机工程与应用, 2005(5): 173-175.