第43卷 第2期 西 安 交 通 大 学 学 报 Vol. 43 №2
2009年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2009

一种应用关联规则森林的改进贝叶斯分类算法

吴宁，柏春霞，祝毅博
（西安交通大学电子与信息工程学院，710049，西安）

摘要：针对朴素贝叶斯分类方法中属性值条件独立假设不适应实际情况的问题，提出了关联规则森林表示法及应用关联规则森林的改进贝叶斯分类算法（ABC算法）. ABC算法利用关联规则挖掘得到满足条件的关联规则，并由此来构造关联规则森林，而规则森林中所有根节点的概率与所有适用的规则置信度连乘，就得到所有属性值的联合概率. 应用UDI数据集对分类器进行了测试，分类结果表明，ABC算法的分类准确率明显高于朴素贝叶斯分类算法，平均提高5%，特别是对属性间有着较强依赖关系的数据集，其分类准确率提高了37%.

关键词：朴素贝叶斯分类；关联规则；联合概率
中图分类号：TP311.13 **文献标志码**：A **文章编号**：0253-987X(2009)02-0048-05

A Modified Bayes Classifier Using Association Rule Forest

WU Ning, BAI Chunxia, ZHU Yibo
(School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: To alleviate the independent assumption on the attribute of the naive Bayes classification, an association rule forest representation and a modified Bayes classifier called ABC are proposed. A data mining method is used to get useful association rules. An association rule forest is constructed from all the resulting useful association rules. Then the joint probability of all the attributes contained in an instance is calculated by multiplying the probabilities of all root nodes with the confidences of all the useful association rules. UDI dataset is used to verify the validation of the ABC. Experimental results show that the ABC has higher classification accuracy, with 5% average improvement, than the naive Bayes one has. Especially, for the dataset containing strong associated attributes, 37% improvement in accuracy is obtained.

Keywords: naive Bayes classification; association rule; joint probability

朴素贝叶斯(NB)分类[1]是一种简单、有效的贝叶斯分类方法，但由于其基于的独立性假设在大多数现实情况中都不成立，因此当属性间存在依赖关系时，其分类性能较差. 树扩张型贝叶斯(TAN)算法[2]采用树形结构来描述属性和类之间的依赖关系，并通过增加属性对之间的关联边来降低该独立性假设，但由于只考虑了两两属性的关联性，使其表示依赖关系的能力受到限制. 网络扩张型贝叶斯(BAN)算法[3]扩展了TAN的结构，允许2个属性间可以形成任意的有向图，使依赖关系表现得更多，然而其所采用的贝叶斯网的学习是一个NP问题[4]. 近几年，又陆续提出了一些在放松条件独立假设下的算法[5-11]，分别从选择关联性强的属性、挖掘1阶或高阶频繁项集以及集成若干频繁项集等角度对朴素贝叶斯分类算法进行了改进，但对属性之间的关联关系没有给予充分的表示，影响了算法的性能.

本文提出了关联规则森林的概念，并提出了一种基于关联规则森林的改进贝叶斯分类(ABC)算法. ABC算法充分利用了关联规则能够反应多个属性间的依赖关系这一特点，来改善朴素贝叶斯算法的分类效果，有效提高了分类准确率.

收稿日期：2008-05-04. 作者简介：吴宁(1961—)，女，副教授，硕士生导师.

1　关联规则森林

1.1　关联规则森林表示方法

关联规则挖掘是用于发现大量数据中项集之间相关性的技术，所得到的相关性用关联规则表示，用支持度和置信度这 2 个参数来描述所挖掘出的规则的有用性和确定性.

项是指事物数据库表中非主键属性的每个不同取值，记为一个属性值. 由多个项构成的集合称为项集[12].

关联规则　设 $I=\{i_1,i_2,\cdots,i_m\}$是项的集合，X 和 Y 是 I 中 2 个互不相交的真子集，$X=A_1\wedge A_2\wedge\cdots\wedge A_m$，$Y=B_1\wedge B_2\wedge\cdots\wedge B_n$，其中 $A_i(i\in\{1,\cdots,m\})$和 B_j $(j\in\{1,\cdots,n\})$是项，则关联规则挖掘的就是形如 $X\Rightarrow Y$ 的规则. 其中，$X\Rightarrow Y$ 表示满足 X 中条件的数据库元组多半也满足 Y 中的条件[1]. X 和 Y 分别称为该规则的规则前件和规则后件.

对形如 $X\Rightarrow Y$ 的关联规则，其支持度定义为数据库所有元组中包含 X 和 Y 的元组数[1]，其置信度定义为包含 X 的元组中同时包含 X 和 Y 的元组数，可用下式表示[1]

$$C(X\Rightarrow Y)=P(X\mid Y)=\frac{P(X\cup Y)}{P(X)}$$

式中：$P(X)$为包含项集 X 的元组数；$P(X\cup Y)$为包含项集 X 和 Y 的元组数. 可看出，关联规则能够体现规则前件和后件中所有属性值的联合概率.

为了将多条关联规则融入到关联规则森林中，定义关联规则森林表示法有如下约定：

(1)关联规则中的每个属性值都有惟一对应的节点，即属性值节点；

(2)规则后件中的属性值节点都是规则前件中属性的子节点，在关联规则森林中，没有父节点的节点称为根节点，反之称为非根节点.

1.2　关联规则属性的联合概率计算

假设关联规则集合 R 中有 N 条关联规则，第 i $(i\in\{1,\cdots,N\})$条关联规则可表示为 $X_i\Rightarrow Y_i$，其中 $X_i=A_{i1}\wedge\cdots\wedge A_{im}$，$Y_i=B_{i1}\wedge\cdots\wedge B_{in}$，且 $X_i\cap Y_i=\varnothing$. 设 R 中包含的全部属性值的集合 $E=\bigcup_{i=1}^{N}(X_i\cup Y_i)=\{A_{i1},\cdots,A_{im},B_{i1},\cdots,B_{in}\}$，且规定 R 中所有规则须满足以下 4 个条件：

(1) 每条规则的规则前件和规则后件的属性值集合均不相交，即 $X_i\cap Y_i=\varnothing$；

(2) 任意两条规则的规则后件不相交，即 $Y_i\cap Y_j=\varnothing(i,j\in\{1,\cdots,N\}$且 $i\neq j)$，否则无法根据这 2 条规则计算其属性的联合概率；

(3) 对任意两条规则 $X_i\Rightarrow Y_i$ 和 $X_j\Rightarrow Y_j(i,j\in\{1,\cdots,N\}$，且 $i\neq j)$，有 $X_i\cup Y_i\not\subset X_j\cup Y_j$，否则包含属性多的那条规则可以更好地表示属性间的联系，从而属性少的那条规则就失去了意义；

(4) 任意两条规则 $X_i\Rightarrow Y_i$ 和 $X_j\Rightarrow Y_j(i,j\in\{1,\cdots,N\}$，且 $i\neq j)$，至少有 $X_i\cap Y_j=\varnothing$，或 $X_j\cap Y_i=\varnothing$，否则所构造的森林中会出现回环.

另外，为便于对下述定理的证明，假设这 N 条规则已按如下要求进行排序：对任意值 $i(i\in\{1,\cdots,N-1\})$ 和 j $(j\in\{i+1,\cdots,N\})$，有 $Y_j\cap(X_i\cup Y_i)=\varnothing$.

定理　设有 N 条关联规则包含的全部属性值的集合为 E，第 i 条规则的置信度为 C_i，构造的关联规则森林中所包含的 M 个根节点对应的属性值集合 $S=\{D_1,\cdots,D_M\}$，第 j 个根节点对应的属性值 $D_j(j\in\{1,\cdots,M\})$的概率为 $P(D_j)$，则 E 中包含的全部属性值的联合概率为

$$P(E)=\Big(\prod_{j=1}^{M}P(D_j)\Big)\Big(\prod_{i=1}^{N}C_i\Big)\qquad(1)$$

证明　当 $N=1$ 时，对于第一条规则 $X_1\Rightarrow Y_1$，由置信度定义可知 $P(X_1\cup Y_1)=P(X_1)C_1$. 由于仅有一条规则，因此 $N=1,M=m$，即 $E=\{A_{11},\cdots,A_{1m},B_{11},\cdots,B_{1n}\}$，$S=\{A_{11},\cdots,A_{1m}\}$. 又由于没有规则来反映 X_1 中属性值间的联系，故可认为 X_1 中的 m 个属性值是相互条件独立的，从而有

$$P(E)=P(A_{11},\cdots,A_{1m},B_{11},\cdots,B_{1n})=P(A_{11},\cdots,A_{1m})C_1=\Big(\prod_{j=1}^{m}P(D_j)\Big)\Big(\prod_{i=1}^{1}C_i\Big)=\Big(\prod_{j=1}^{M}P(D_j)\Big)\Big(\prod_{i=1}^{N}C_i\Big)$$

所以，$N=1$ 时式(1)成立.

假设 $N=k$ 时，用 $T(k)$来表示当前关联规则森林中根节点的个数，即 $M=T(k)$，$S=\{D_1,\cdots,D_{T(k)}\}$，则有 $P(E)=\Big(\prod_{j=1}^{T(k)}P(D_j)\Big)\Big(\prod_{i=1}^{k}C_i\Big)$.

当 $N=k+1$ 时，将第 $k+1$ 条关联规则加入到前 k 条关联规则构造的关联规则森林中. 由于对这 N 条规则的限制，第 $k+1$ 条关联规则的规则后件 Y_{k+1}中包含的属性值均不可能成为父节点，而规则前件 X_{k+1}可划分为 2 个集合 U 和 V，其中 U 中的属性值在前 k 条规则中没有出现过，V 中的属性值在前 k 条规则中出现过. 由于没有任何关联规则来反映 U 中属性值与前面出现过的属性值相关，因此可以认为 U 中属性值和前面出现过的属性值相互独

立. 设前 k 条规则中出现过的属性值集合为 W_k，故

$$P(W_{k+1}) = P(W_k \cup X_{k+1} \cup Y_{k+1}) =$$
$$P(W_k \cup U \cup V \cup Y_{k+1}) =$$
$$P(W_k \cup U \cup Y_{k+1}) = P(W_k \cup U)C_{k+1} =$$
$$P(W_k \cup U)C_{k+1} = P(W_k)P(U)C_{k+1} =$$
$$\left(\prod_{j=1}^{T(k)} P(D_j)\right)\left(\prod_{i=1}^{k} C_i\right)P(U)C_{k+1}$$

由于 U 中属性值在关联规则森林中没有父节点，所以根节点集合 $S=T(k)\cup U$. 故可假设 $U=\{D_{T(k)+1},\cdots,D_{T(k+1)}\}$，即有 $P(U)=\prod_{t=T(k)+1}^{T(k+1)} P(D_t)$，则 $k+1$ 条规则中所有属性值的联合概率为

$$P(E) = P(W_{k+1}) =$$
$$\left(\prod_{j=1}^{T(k)} P(D_j)\right)\left(\prod_{i=1}^{k} C_i\right)\left(\prod_{t=T(k)+1}^{T(k+1)} P(D_t)\right)C_{k+1} =$$
$$\left(\prod_{j=1}^{T(k+1)} P(D_j)\right)\left(\prod_{i=1}^{k+1} C_i\right)$$

综上所述，对于任意的 $N>0$，式(1)都成立.

证毕.

因此，通过关联规则构造出关联规则森林并使用式(1)，可得到属性的联合概率.

2 应用关联规则森林的改进贝叶斯分类算法

朴素贝叶斯分类方法是将未知的样本 $X=\{x_1, x_2, \cdots, x_N\}$，分配给 M 个类 $G_1, G_1, \cdots, G_M$ 中的某个类 G_i ($i\in\{1,\cdots, N\}$)，其中 G_i 满足以下条件：当且仅当在样本所有属性同时存在的条件下，该类的后验概率最高，即 $P(G_i|X)$ 最大. 根据贝叶斯定理，样本 X 属于类 G_i 的概率 $P(G_i|X)$ 可表示为

$$P(G_i \mid X) = \frac{P(X \mid G_i)P(G_i)}{P(X)}$$

式中：$P(X)$ 是 X 的先验概率，其值为常数；$P(G_i)$ 可由训练得到. 因此，求解概率 $P(G_i|X)$ 的问题转化成了求解 $P(X|G_i)$ 的问题. 为了简化计算，朴素贝叶斯分类算法作了与实际情况不相符的独立性假设. 考虑到属性间的依赖性，ABC 算法引入了关联规则森林来计算 $P(X|G_i)$.

$P(X|G_i)$ 的含义为 X 中的所有属性值在类 G_i 中同时存在的概率. 利用关联规则挖掘可以得到有用的关联规则，进而构造出关联规则森林，然后用式(1)来计算 X 中所有属性的联合概率.

在执行 ABC 算法前，先要提取有用的规则，即对有 M 个类别 $G_1, G_2, \cdots, G_M$ 的数据集 E 进行图 1 所示的处理. 求解频繁项集时采用 FP-tree[9] 算法，并允许限定生成的高频项集的最高项数，这是由于关联规则森林可以用若干个低项数高频项集的规则来反映高项数高频项集规则所反映的属性间依赖关系.

在求取规则过程中，根据置信度的定义，置信度最大的规则的后件必然只有一个属性. 另外，由于规则后件仅为类别属性时，在计算过程中对联合概率的计算没有贡献($P(G_i|G_i)=1$)，所以在程序实现中只有 2 种备选规则：①规则的后件只有一个非类别属性；②规则的后件有 2 个属性(其中一个为类别属性). 在 m 确定的情况下，频繁项集产生规则集合的时间复杂度成线性关系.

ABC 算法的描述如图 2 所示，其流程图如图 3

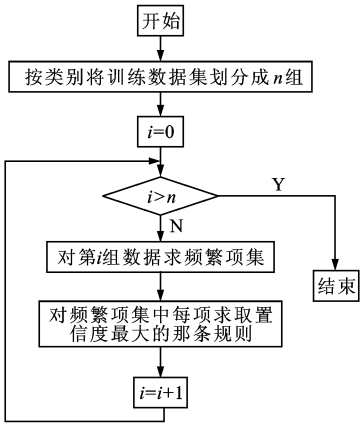

图 1 关联规则提取的流程图

输入：测试样本s_t；类型集合H；关联规则集合R；条件概率集合L.
输出：s_t所属的类别L.
变量：最大联合概率max P；H中的第i个类型值G_i；R中的元素r_{ij}，表示第i个类型分组提取的第j个关联规则；计数变量i,j；有用规则集合R'；当前先验概率p.
(1) max $P\leftarrow 0$
　//逐一计算并比较相对于所有类型的先验概率，取其中的最大值作为最大联合概率
(2) for each G_i in H
(3)　for each r_{ij} in R
(4)　　if r_{ij} is useful　//若r_{ij}中的属性均在s_t所包含的属性集合中则为有用规则
(5)　　insert r_{ij} into R'
(6)　　endif
(7)　endfor
(8)　$p \leftarrow$ calc$P(s_t, R', L)$　//通过联合概率函数calc$P$求得当前类G_i的先验概率p
(9)　if $p>$max P　//若当前先验概率较大则取其值作为最大联合概率，并记录当前类
(10)　max $P\leftarrow p$
(11)　$L\leftarrow G_i$
(12)　endif
(13) endfor
(14) return L　//返回最终取得最大联合概率值的类型

图 2 ABC 算法描述

所示. 图中的 calc$P$ (S_t, R', L)用于计算联合概率，其联合概率计算流程如图 4 所示.

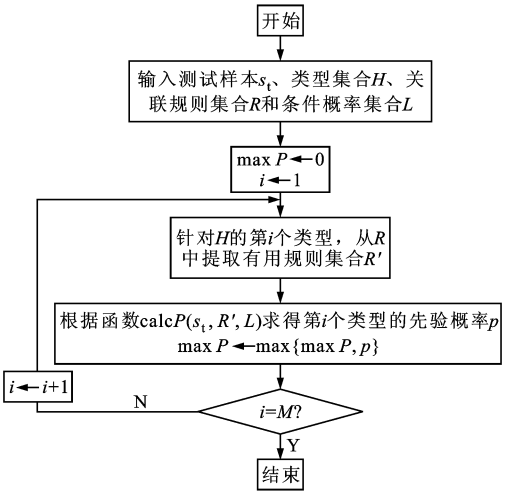

图 3　ABC 算法流程图

3　实验结果与分析

采用 UCI 中的 10 个离散化数据集对 ABC 分类算法进行了测试(对有缺省属性值的数据集，将缺失的属性值指定为一特定值). 求取频繁项集的支持度取 20%，选择规则的置信度取 60%，数据集中 70%数据为训练数据，30%数据为测试数据. 对于属性数和类别数少的数据集，重复 50 次，而对于属性数和或类别数比较大的数据集(zoo)则重复 20 次. 另外，由于 zoo 的属性比较多，产生频繁项集所需时间较长，故设置频繁项最高项为 5. 实验结果如表 1 所示.

从实验结果可看出，大部分情况下 ABC 算法的准确率都比 NB 算法高，这是因为这些数据集中的属性间有着较强的依赖关系. shuttle-landing-control 数据集是一些关于太空船是否能自动着陆的数据，其属性之间的依赖性非常强，如风向、风力、飞机稳定性等属性之间有着直接的联系，因此 ABC 算法的准确率比 NB 算法提高很多，而 post-operative patient 中的属性 CORE-STBL (体内温度稳定性)

输入：测试样本s_t;有用规则集合R';所有属性在各个类型下的条件概率集合L.
输出：s_t中所有属性的联合概率p。
变量：根节点属性值集合S;非根节点属性值集合$\bar{S}$;集合L中值用l_i表示,i为基数变量.
(1) purify R'　//清除不满足前面提到的4个条件的规则
(2) initialize S as s_t set　//初始化S为s_t中包含的属性值结合
(3) initialize $\bar{S}$ as empty set　//初始化$\bar{S}$为空集合
(4) p←0
　//根据R'中的关联规则构造关联规则森林,求出规则森林的值即为联合概率p
(5) for each rule in R'
　　//将出现在规则后件且属于S中的属性值从S去除并添加到$\bar{S}$中
(6)　for each element in the consequence set　//对规则后件中的属性值逐一排查
(7)　　if elements in S
(8)　　delete the element from S
(9)　　insert the element into $\bar{S}$
(10)　endif
(11)　endfor
　　//将出现在规则前件且不属于$\bar{S}$的属性值添加到S中
(12)　for each element in the antecedent set　//对规则前件中的属性值逐一排查
(13)　　if elements in $\bar{S}$
(14)　　insert the element into S
(15)　endif
(16)　endfor
(17)　p←p×Confidence of rule
(18)endfor
　//求联合概率,即通过规则置信度乘积与根集合中各属性值条件概率的乘积相乘所得
(19)for i←1 to size of S
(20)　　p←p×l_i
(21)endfor
(22)return p

图 4　联合概率计算

表 1　ABC 和 NB 算法的准确率比较

数据集名称	数据集特征				准确率/%	
	类别数	属性数	实例数	有无缺省	NB	ABC
contraceptive method choice	3	9	1 473	N	51.144 5	51.338 7
breast-cancer	2	9	699	Y	95.801 8	96.078 7
contact-lenses	3	4	24	N	66.439 7	67.189 7
tic-tac-toc	2	9	958	N	65.329 5	66.527 7
zoo	7	17	101	N	84.225 1	86.014 9
post-operative patient	3	8	90	Y	41.052 8	43.403 9
hayes-roth(training)	3	5	132	N	73.085 6	76.239 7
balance-scale	3	4	625	N	70.034 6	70.034 6
haberman	2	4	306	N	53.542 7	56.769 8
shuttle-landing-control	2	6	15	N	42.242 9	79.781 0

和 BP-STBL(血压稳定性)之间依赖性很强,因此也有较明显的提高,而 balance-scale 数据集中属性间关系非常弱,没有直接的联系,因此 ABC 算法与 NB 算法的分类结果相等. 研究分析表明,随着属性间依赖关系的增强,ABC 算法的分类效果明显优于 NB 分类.

4 结束语

本文提出了关联规则的规则森林表示方法和联合概率计算公式,把联合概率应用到朴素贝叶斯分类中,提出了一种改进的贝叶斯分类算法. 实验结果表明,本文算法弥补了朴素贝叶斯分类的缺陷,提高了分类准确度. 针对数据集属性很多时因产生频繁多项集致使分类费时较长这一问题,本文算法采用了限定频繁项的最高项数的策略,在分类精度没有明显改变的情况下,有效减少了分类时间,保证了算法的实用性.

参考文献:

[1] HAN Jiawei, KAMBER M. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京: 机械工业出版社, 2006.

[2] FRIEDMAN N, GEIGER D, GOLDSZMIT M. Bayesian network classifier[J]. Machine Learning, 1997, 29: 131-163.

[3] KEOGH E J, PAZZANNI M J. Learning augmented Bayesian classifiers: a comparison of distribution-based and classification-based approaches[C]// Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 1999: 225-230.

[4] CHICKERING D M, GEIGER D, HECKERMAN D. Learning Bayesian networks is NP-complete[M]// Learning from Data: Artificial Intelligence and Statistics. Berlin, Germany: Springer-Verlag, 1996: 121-130.

[5] WEBB G I, BOUGHTON J R, WANG Zhihai. Not so naive Bayes: aggregating one-dependence estimators [J]. Machine Learning, 2005, 58 (1): 5-24.

[6] MARTINEZ M, SUCAR L E. Learning an optimal naive Bayes classifier [C]// Proceedings of International Conference on Pattern Recognition. Las Alamitos, CA, USA: IEEE Computer Society Press, 2006: 1236-1239.

[7] 石洪波,王志海,黄厚宽,等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报, 2004, 15(2): 193-199.
SHI Hongbo, WANG Zhihai, HUANG Houkuan, et al. A restricted double-level Bayesian classification model [J]. Journal of Software, 2004, 15(2): 193-199.

[8] MERETAKIS D, WUTHRICH B. Extending naive Bayes classification using long itemsets [C]// Proceedings of 5th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 1999: 165-174.

[9] ABELLÁN J, CANO A, MASEGOSA A R, et al. A semi-naive Bayes classifier with grouping of cases [C]// Proceedings of European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Berlin, Germany: Springer-Verlay, 2007: 477-488.

[10] 眭俊明, 姜远, 周志华. 基于频繁项集挖掘的贝叶斯分类算法[J]. 计算机研究与发展, 2007, 44(8): 1293-1300.
XU Junming, JIANG Yuan, ZHOU Zhihua. Bayesian classifier based on frequent item sets mining [J]. Journal of Computer Research and Development, 2007, 44 (8): 1293-1300.

[11] HAN Jiawei, PEI Jian, YIN Yiwen, et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach [J]. Data Mining and Knowledge Discovery, 2004, 8 (1): 53-87.

[12] 何军,刘红岩,杜小勇. 挖掘多关系关联规则[J]. 软件学报, 2007,18(11): 2752-2765.
HE Jun, LIU Hongyan, DU Xiaoyong. Mining of multi-relational association rules [J]. Journal of Software, 2007, 18(11): 2752-2765.

(编辑 刘杨 苗凌)