文章编号：2096-1472(2022)-01-33-04
DOI:10.19644/j.cnki.issn2096-1472.2022.001.008

基于朴素贝叶斯的区域高校图书馆数字资源一站决策算法

顾春燕

(南通大学图书馆，江苏 南通 226001)
✉337927012@qq.com

摘　要：随着智慧图书馆的兴起，可以对图书馆数字资源大数据进行深入挖掘利用，区域高校图书馆数字资源一站式检索必然是进一步增强馆际合作、数据挖掘、资源互享的有效平台。基于朴素贝叶斯的区域高校图书馆数字资源一站式决策算法设计了一种决策树与朴素贝叶斯模型相结合的两层模型方法，通过提取整合区域内各高校图书馆数字资源大数据的特征属性，并利用朴素贝叶斯模型进一步筛选特征属性，从而构建决策树架构，支撑区域高校图书馆数字资源一站式检索。利用基于朴素贝叶斯的区域高校图书馆数字资源一站式决策算法可以实现检索资源过程更加便捷高效，检索结果的准确率呈现翻倍式增长。

关键词：数字资源；朴素贝叶斯；决策树；一站式

中图分类号：TP312　**文献标识码**：A

A One-stop Decision-making Algorithm for Digital Resources of Regional University Libraries based on Naive Bayes

GU Chunyan

(*Nantong University Library*, *Nantong* 226001, *China*)
✉337927012@qq.com

Abstract: With the rise of smart libraries, big data of library digital resources can be deeply excavated and utilized. One-stop retrieval of digital resources in regional university libraries is bound to be an effective platform to further enhance interlibrary cooperation, data mining, and resource sharing. This paper proposes to design a two-layer model method combining decision tree and Naive Bayes model, based on Naive Bayes-based one-stop decision-making algorithm for regional university libraries' digital resources. By extracting and integrating the characteristic attributes of the digital resources big data in various university libraries in the area, and using Naive Bayes model to further filter the characteristic attributes, a decision tree structure can be constructed to support the one-stop retrieval of digital resources in the regional university libraries. The one-stop decision-making algorithm for digital resources in regional university libraries based on Naive Bayes can be realized: the process of retrieving resources is more convenient and efficient, and the accuracy of retrieval results has doubled.

Keywords: digital resources; Naive Bayes; decision tree; one-stop

1 引言(Introduction)

近些年，随着物联网、大数据、云计算、人工智能等新兴技术的发展，“智慧图书馆”成为图书馆界的研究热点，研究者各自从不同的角度对“智慧图书馆”进行探讨。

基金项目：南通大学人文社会科学研究项目“云环境下高校图书馆数字资源高效使用的策略研究”(16w16).

AITTOLA首次提出“智慧图书馆”的概念，他认为“智慧图书馆”是一个不受空间限制且可被感知的移动图书馆[1]。王世伟认为智慧图书馆是以高效、互联、便利为特征，以绿色发展为发展战略，以数字惠民，引导读者智慧阅读，为读者提供全方位一体化的服务为根本追求，实现广阔互联互通与共享融合的未来图书馆发展新模式[2]。未来的发展趋势是基于智能化、网络化、数字化信息技术，实现以人为本、绿色发展、广泛互联的具有高效、便利、互联、智慧等特性的图书馆[3]。

对图书馆数字资源大数据可以进行深入挖掘利用，区域高校图书馆数字资源一站式检索必然是进一步增强馆际合作、数据挖掘、资源互享的有效平台，是未来图书馆实现互联互通、智慧共享的重要途径。各高校图书馆购买的数字资源不同，各数字资源数据库拥有不同的检索平台，导致用户需要不停地切换检索模式，获取所需资源的过程耗时且繁琐，而检索结果会出现重复、不全面的现象。因此，构建区域高校图书馆数字资源一站式检索显得尤为迫切。

2 图书馆数字资源一站式检索研究现状(Research status of one-stop retrieval of library digital resources)

以往针对“数字资源的一站式检索”的学术研究主要都集中在平台的搭建、分布式数据库检索模型、混合式数据库检索模型、集中式数据库检索模型、基于语义技术的检索模型。

何美琴、陈刚通过构建区域高校图书馆一站式书目检索平台来解决读者在书目检索中遇到的困难，在区域高校资源共享、优势互补的基础上，使读者享受到一站式书目检索带来的快捷方便[4]。杨伟超、刘阳、李淑霞提出构建基于搜索引擎的一站式检索平台，实现在统一的检索界面上，一次检索就能获得所有电子资源的相关文献信息[5]。唐光前提出了一种基于.NET Remoting的分布式异构数据库一站式检索系统模型，向用户提供一步到位的跨库检索服务，可以最大限度地减少检索步骤[6]。张卫华提出了一种基于语义技术的图书馆资源检索模型，增加了本体字典、检索历史抽取库和输出子系统[7]。

不难看出，目前对于图书馆数字资源一站式检索的平台架构、数据库架构的研究已经相对成熟，但较少学者将朴素贝叶斯和决策树算法一起应用于图书馆数字资源一站式检索。如何获取更加高效、更加精确的检索决策算法是本文研究的重点内容。

3 基于朴素贝叶斯的区域高校图书馆数字资源一站式决策算法的整体框架(The overall framework of a one-stop decision-making algorithm for digital resources of regional university libraries based on Naive Bayes)

由于地域、自身办学水平和资金能力的差异，以及各高校重点建设学科的不同，我国高校图书馆数字资源存在资源存储量差距较大、重点学科资源倾斜性较为明显、资源的利用率较低等问题[8]，而区域高校图书馆数字资源一站式检索能有效地解决上述问题。因此，本文提出了一种基于朴素贝叶斯的区域高校图书馆数字资源一站式决策算法。

区域高校图书馆数字资源一站式决策算法的整体框架如图1所示。

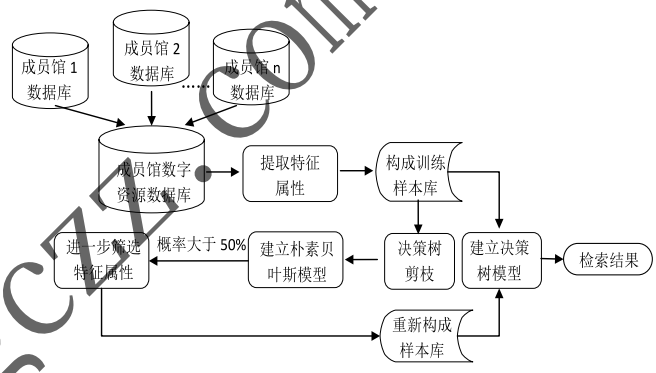

图1 算法框架图

Fig.1 Algorithm framework diagram

(1)由区域内各高校成员图书馆数字资源数据库构成“成员馆数字资源数据库”；

(2)提取各成员馆数字资源特征属性构成训练样本，构建区域高校图书馆数字资源一站式检索决策树模型；

(3)进行决策树剪枝，构建区域高校图书馆数字资源一站式检索朴素贝叶斯模型，计算输出概率，当大于50%时，获取当下所有特征属性重新构成样本库；

(4)筛选后的特征属性构成的新样本库支撑区域高校图书馆数字资源一站式检索。

4 基于朴素贝叶斯的区域高校图书馆数字资源一站式决策算法具体步骤(Specific steps of one-stop decision-making algorithm for digital resources of regional university libraries based on Naive Bayes)

4.1 构建区域高校图书馆数字资源一站式检索决策树模型

(1)提取区域高校图书馆数字资源大数据作为所述C5.0决策树模型的训练样本S，根据该训练样本S获取数字资源特征

属性并将其作为所述决策树模型的输入变量 X_i，每个输入变量 X_i 都有其对应的分类 x_{ij}，其中 $i=1,2,\cdots,n$，j 的值由其对应的 X_i 分类值确定，如表1所示。

表1 决策树模型中输入变量 X_i 的分类表

Tab.1 Classification table of input variables X_i in decision tree model

输入变量	每个变量对应的分类						
X_1	x_{11}	x_{12}	x_{13}	x_{14}			
X_2	x_{21}	x_{22}					
…	…	…	…	…	…	…	…
X_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}			

(2)为了解决训练样本S自身信息熵偏大的问题，采用信息增益率来替代信息增益进行最佳分支变量的选择，即根据所述训练样本S的信息增益率 $Gains(X_i)$ 选择所述决策树模型的最佳分支变量。该信息增益率 $Gains(X_i)$ 计算公式如下：

$$Info(S)=-\sum_{k=1}^{n}\left(\left(freq(C_k,S)/|S|\right)\cdot\log_2\left(freq(C_k,S)/|S|\right)\right)$$

$$Info(X_i)=-\sum_{j}\left(\left(\left|x_{ij}\right|/|X_i|\right)\cdot Info\left(x_{ij}\right)\right),(i=1,2,\cdots,n)$$

$$Info(x_{ij})=-\sum_{k=1}^{n}\left(\left(freq\left(C_k,x_{ij}\right)/\left|x_{ij}\right|\right)\cdot\log_2\left(freq\left(C_k,x_{ij}\right)/\left|x_{ij}\right|\right)\right)$$

$$Gain\left(X_i\right)=Info\left(S\right)-Info\left(X_i\right),(i=1,2,\cdots,n)$$

$$Gains\left(X_i\right)=Gain\left(X_i\right)/Info\left(X_i\right),(i=1,2,\cdots,n)$$

其中：

C_k 为区域高校图书馆数字资源检索结果的决策树输出变量，$k=1,2,\cdots,n$，C_1、C_2、…、C_K 分别对应成员馆1数据库、成员馆2数据库、……、成员馆n数据库；

$|S|$ 为训练样本S的样本总数；

$freq(C_k,S)$ 为训练样本S中属于成员馆 C_k 的样本数量；

$freq(C_k,x_{ij})$ 为训练样本S中包含输入变量 X_i 分类值为 x_{ij} 的属于成员馆 C_k 的样本数量；

$|X_i|$ 为训练样本S中包含输入变量 X_i 的样本数量；

$|x_{ij}|$ 为训练样本S中包含输入变量 X_i 分类值为 x_{ij} 的样本数量；

$Info(S)$ 为训练样本S的信息熵，$Info(X_i)$ 为高校图书馆数字资源特征属性 X_i 在训练样本S中的条件熵，$Info(x_{ij})$ 为高校图书馆数字资源特征属性分类 x_{ij} 在训练样本S中的条件熵；

$Gain(X_i)$ 为高校图书馆数字资源特征属性 X_i 的信息增益。

(3)提取区域高校图书馆数字资源大数据并从下向上进行决策树后剪枝，其关键是误差的估计及剪枝标准的设置。该决策树后剪枝具体方法如下：

采用置信区间法，给定置信水平 $1-\alpha$，假设第 r 个节点上的 A_r 个诊断预测中有 B_r 个是错误的，则错误率为 B_r/A_r，且通过正态分布表得到第 r 个节点上的区域高校图书馆数字资源检索结果的误差 μ_r 满足：

$$P\left\{\left|\frac{\frac{B_r}{A_r}-\mu_r}{\sqrt{\frac{B_r}{A_r}\left(1-\frac{B_r}{A_r}\right)/A_r}}\right|\leqslant Z_{\alpha/2}\right\}=1-\alpha$$

令 $E_r=B_r/A_r$，μ_r 的置信区间为：

$$\left[E_r-Z_{\alpha/2}\sqrt{\frac{E_r(1-E_r)}{A_r}},E_r+Z_{\alpha/2}\sqrt{\frac{E_r(1-E_r)}{A_r}}\right]$$

假设父节点上的诊断结果的误差为 μ，每个子节点占该分支样本量的比例为 θ_r，对父节点拥有的所有子节点的诊断结果的误差进行加权计算：

$$\varphi=\sum_{r=1}^{k}\theta_r\mu_r$$

当所有 μ_r 都取最小值时，如果 $\varphi>\mu$，则将该父节点的所有子节点都修剪去除(如图2所示，其中node2、node6、node8的子节点都已修剪去除)。

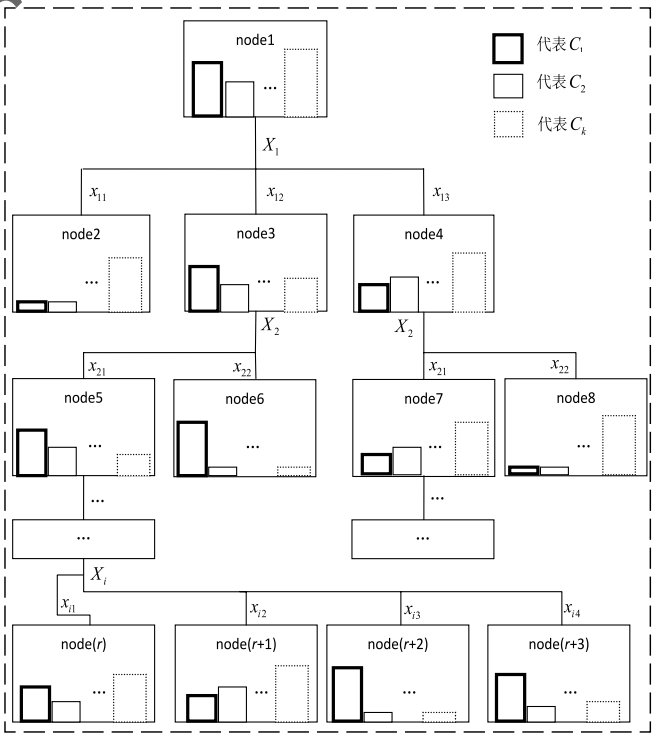

图2 区域高校图书馆数字资源一站式检索决策树剪枝流程图

Fig.2 Flow chart of one-stop retrieving decision-making tree pruning for digital resources of regional university libraries

4.2 构建区域高校图书馆数字资源一站式检索朴素贝叶斯模型

(1)从区域高校图书馆数字资源大数据中提取包含上述决策树模型筛选后的特征属性数据，并重新构成训练样本D，提取上述决策树模型中所有输出变量为C_1(成员馆1)的节点(以图2为例，提取以后的结果如图3所示)。

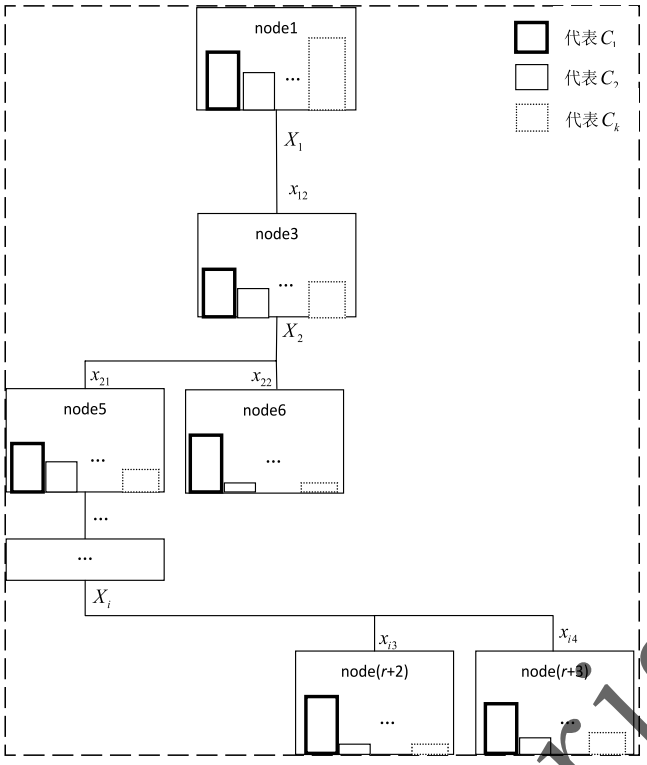

图3 区域高校图书馆数字资源一站式检索算法中重新提取特征属性流程图

Fig.3 Flow chart of re-extracting characteristic attributes from one-stop retrieval algorithm for digital resources of regional university libraries

(2)自上向下获取每个节点所经过的特征属性分类x_{ij}，并定义第r个节点所拥有的特征属性集合Y_r为：$Y_r=\{y_1,y_2,\cdots,y_m\}$，其中$m$是对应节点所拥有的特征属性的个数。

(3)再利用贝叶斯公式得出第r个节点上输出变量为成员馆1的可能性概率，则$P(C_1|y_1\cdot y_2\cdot y_3\cdots y_m)$为：

$$P(C_1|y_1\cdot y_2\cdot y_3\cdots y_m)=\frac{P(C_1)\cdot P(y_1|C_1)\cdot P(y_2|C_1)\cdots P(y_m|C_1)}{\sum_{k=1}^{n}P(C_k)\cdot P(y_1|C_k)\cdot P(y_2|C_k)\cdots P(y_m|C_k)}$$

$$P(C_k)=freq(C_k,D)/|D|,(k=1,2,\cdots,n)$$

$$P(y_m|C_k)=\frac{1}{P(C_k)}\left(freq(C_k,y_m)/\,freq(C_k,D)\right),(k=1,2,\cdots,n)$$

其中：

$|D|$为训练样本D的样本总数；

$freq(C_k,D)$为训练样本D中属于成员馆C_k的样本数量；

$freq(C_k,y_m)$为训练样本D中包含输入变量y_m的属于成员馆C_k的样本数量。

(4)当$P(C_1|y_1\cdot y_2\cdot y_3\cdots y_m)$大于50%时，获取第$r$个节点上的所有特征属性构成新样本库，新样本库将直接作为检索数据库提供检索。

5 决策树与朴素贝叶斯模型相结合算法的优点(Algorithm advantages of combining decision tree and Naive Bayes model)

(1)该算法首先基于区域高校图书馆数字资源的大数据，采用决策树模型来预测检索结果，并将信息增益率作为选择最佳分支变量的依据，提高了分类的精度；然后采用朴素贝叶斯模型进一步筛选特征属性，对检索结果进行概率计算，经过上一层模型的预处理，检索结果更加精确；同时采用决策树和朴素贝叶斯两层模型的新思路进行数字资源的一站式检索，摆脱了以往一层模型检索结果区间大、范围广、较为模糊的缺陷。

(2)该算法的决策树模型是利用训练样本自顶向下构造的，而后再从下向上剪枝，都是通过节点关联，利于结构化编程的实现。同时，算法中的朴素贝叶斯模型的数学计算方法更利于计算机的处理，实现起来很容易。

(3)该算法构建区域高校图书馆数字资源一站式检索C5.0决策树模型，其是C4.5应用于大数据集的分类算法，提高了执行效率，减少了内存使用。同时，C5.0决策树模型规则十分直观，在面对数据遗漏和输入字段很多的问题时非常稳健，并且它通常不需要很多的训练次数。

6 结论(Conclusion)

本文提出了一种基于朴素贝叶斯的区域高校图书馆数字资源一站式检索。该方法设计了一种决策树与朴素贝叶斯模型相结合的两层模型方法，通过提取区域高校图书馆数字资源大数据中的特征属性，构建区域高校图书馆数字资源一站式检索决策树模型，然后根据训练样本的信息增益率选择所述决策树模型的最佳分支变量，接着从下向上进行决策树后剪枝，最后构建区域高校图书馆数字资源一站式检索朴素贝叶斯模型来进一步筛选特征属性构成样本库，从而实现数字资源的一站式检索。

决策树和朴素贝叶斯两层模型相结合，使得检索结果更加全面精确。基于朴素贝叶斯的区域高校图书馆数字资源一站式决策算法可以让资源相对丰富的高校扶持资源相对匮乏

(下转第40页)