基于机器学习的中文书目自动分类研究*

王 昊 严 明 苏新宁

摘 要 面对与日俱增的图书出版量,图书馆编目人员的手工书目分类显得力不从心,如何实现由计算机自动完成图书分类成为数字图书馆建设中亟待解决的关键问题之一。本文尝试将 BP 神经网络和支持向量机等机器学习算法引入到书目分类中,建立了面向中图法的基于机器学习的书目层次分类系统模型,提出了采用特征加权方式描述书目和浅层次分类体系构建的设计思路,并通过大规模实验验证了该模型的可行性和合理性,基本上解决了没有主题标注情况下书目的自动分类问题。图 9。表 5。参考文献 14。

关键词 机器学习 书目自动分类 特征加权 中图法 浅层次分类模型

分类号 TP391

ABSTRACT Books classification by computer has become one of the most critical issues which should been solved immediately in digital library construction because of increasing volume of book publishing. This paper tries to induct the BP nerve net and Support Vector Machine algorithms to bibliography classification, and establishs bibliography hierarchy classification system model based on machine learning faced to the Chinese Library Classification, then proposes the design ideas of describing bibliographies using feature weighted mode and constructing shallow classification system. It verifies the feasibility and rationality of the model by large-scale experiment, and basically solves the case of the bibliography automatic classification without subject labeling, which lays a theoretical foundation for constructing the practical bibliography automatic classification system, and provides factual basis for the wide range application of machine learning methods for the construction of digital libraries. 9 figs. 5 tabs. 14 refs.

KEY WORDS Machine learning. Automatic bibliography classification. Feature weighted. Chinese Library Classification. Shallow classification model.

CLASS NUMBER TP391

1 引言

随着信息技术的发展和成熟,信息自动化技术应用于传统图书馆以建立数字图书馆,实现图书馆资源的数字化、工作的自动化,已经成为目前各级图书馆的首要任务[1]。在图书馆的各项工作中,图书编目是一项繁杂的基础性工作,按照现有的知识分类体系,如中国图书馆分类法(以下简称"中图法"),确定图书的类目号更是编目工作的核心任务之一,也是实现海量图书有效管理的基本前提。目前,中文图书的分类多采用手工方式,或由图书作者或由图书编目人员给出中图法分类号。然而,图书作者给出的分类号带有明显的主观性和非专业性,不利于图书的统一管理;面对巨大的图书出版量,具有专业知识的图书编目人员在手工进行图书分类时显得心有余而力不足,不仅需要大量资金投入,而且严重影响图书馆的工作效率。在这种情况下,将信息自动化技术引入到图书

* 本文系国家社科基金项目"面向语义网本体的知识管理研究"(编号:09CTQ010)的研究成果之一。

编目工作中,由机器来自动完成图书分类成为了数字图书馆建设中亟待解决的关键问题之一。

总结前人的研究成果,目前中文图书自动分类主要有两种方法:一是基于知识库(或主题词表)的图书分类[2],即首先根据图书标引状况建立主题词(或关键词)和类目号的对应关系,形成由主题词、类目号、隶属度3元组所构成的知识库,然后根据待分类图书给出主题词,累加主题词所在类目的隶属度,隶属度总值最高的即为该图书所在的类目;另一种则是基于数据挖掘中的分类功能实现图书分类,即首先构建各类目的主题词(或关键词)向量描述,然后计算待分类图书的主题词向量和各类目主题词向量之间的相关度,相关度最高的即为该图书所属类目。然而这两种方法都存在明显缺陷:①知识库(或主题词向量)需要事先构建,需要经过标引的语料数据提供主题词,并计算其与类目的关联度;②在分类前需要对待分类图书进行标引提取主题词,而中文图书的自动标引本身就是一个技术难题,成为了自动分类难以逾越的障碍。

分析上述两种方法,不难发现:中文图书自动分类的过程可以分为两个步骤,先对已标引数据进行学习建立知识库,可以是主题词典,也可以是类目的主题表示,再利用学习成果指导图书分类。那么,把上述结论泛化,能不能对没有经过主题词提取的数据进行学习,进而实现分类?这实际上就是机器学习(Machine Learning)技术,具体是指采用计算机来模拟人类的学习行为,利用机器学习技术从已知样本中寻找规律,并利用规则对未知数据进行预测[3]。机器学习技术在20世纪90年代被引入到文本分类中,以其自动化程度高、分类速度快、效果好等优点[4],逐步取代词频统计[5]和基于知识工程方法[6]而成为文本分类的主流技术。基于机器学习的文本分类是指通过对现有分类体系和已分类文本的学习,将得到的分类模型作用于未分类文本,根据文本内容获得其类别的过程。

图书资源也可以被认为是篇幅较长的文本信息,因此在文本分类中具有有效作用的机器学习技术在一定的资源环境下也可以适用于图书分类。本文即试图将机器学习分类算法引入到图书自动分类研究中,构建基于特征加权的多层次图书自动分类系统模型,使之能够根据中文图书的内容特征(包括题名、关键字和文摘)自动给出中图法分类号,以解决图书手工分类投入大、效率低、主观性强、片面性等问题,并通过实验证明该模型的准确性和合理性。

2 基于机器学习的图书自动分类系统模型

本节主要讨论采用机器学习技术实现图书自动分类的基本方法,分析在操作过程中需要解决的问题,并针对中文图书特征丰富和中图法层次结构复杂的特点提出基于层次分类器的多层图书自动分类系统模型。

2.1 系统模型的分析和设计

基于机器学习实现图书自动分类的基本思路:首先分析中文图书的书目数据,从中提取出能够描述图书内容的基本特征和中图法类目号,并根据特征的重要程度赋予不同权值;再以特征向量作为图书的书目表示,构建图书的二元特征矩阵,如图1所示,其中行向量表示书目,列向量表示特征,矩阵值 co_{ij} 则表示书目 i 与特征 j 之间的关联度;在特征矩阵中加入书目的分类信息以形成机器学习的对象,如图2所示,即在图1特征矩阵的基础上增加了一个类目列向量 $cata_i$;然后采用机器学习算法,常用的如决策树、神经网络和支持向量机等对特征+类目矩阵进行学习,获得分类器;最后将分类器作用于待分类图书的特征矩阵,通过自动数据分析即可获得图书的分类情况。整个过程如图3所示。

在图3中,笔者将整个图书自动分类过程分为两个阶段:先学习,后分析。这是采用机器学习方法实现具体应用的基本做法。但是,确定图书的中图类目号不同于一般意义上的文本分类,其中涉及到了特征提取、特征权重设置、机器学习算法选择以及分类方法确定等在具体应用中需要解决的问题。

$$Eigenvector=\begin{bmatrix} co_{11} & co_{12} & \cdots & co_{1j} & \cdots & co_{1m} \\ co_{21} & co_{22} & \cdots & co_{2j} & \cdots & co_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ co_{i1} & co_{i2} & \cdots & co_{ij} & \cdots & co_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ co_{n1} & co_{n2} & \cdots & co_{nj} & \cdots & co_{nm} \end{bmatrix} \begin{matrix} \rightarrow 书目1 \\ \rightarrow 书目2 \\ \cdots \\ \rightarrow 书目i \\ \cdots \\ \rightarrow 书目n \end{matrix}$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

特征1 特征2 … 特征j … 特征m

(其中$i=1,2\ldots\ n; j=1,2,\ldots,m$)

图1 书目的二元特征矩阵

$$Eigen\_Category\_Vertor=\begin{bmatrix} co_{11} & co_{12} & \cdots & co_{1j} & \cdots & co_{1m} & cata_1 \\ co_{21} & co_{22} & \cdots & co_{2j} & \cdots & co_{2m} & cata_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ co_{i1} & co_{i2} & \cdots & co_{ij} & \cdots & co_{im} & cata_i \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ co_{n1} & co_{n2} & \cdots & co_{nj} & \cdots & co_{nm} & cata_m \end{bmatrix} \begin{matrix} \rightarrow 书目1 \\ \rightarrow 书目2 \\ \cdots \\ \rightarrow 书目i \\ \cdots \\ \rightarrow 书目n \end{matrix}$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

特征1 特征2 … 特征j … 特征m 类目

(其中$i=1,2\ldots\ n; j=1,2,\ldots,m$)

图2 书目的特征+类目矩阵

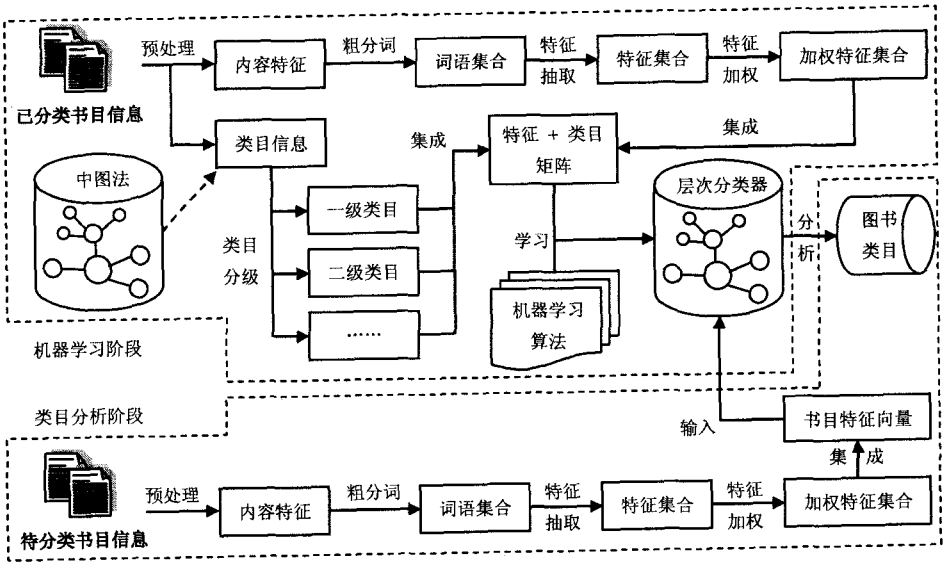

图3 基于机器学习的图书自动分类系统模型

(1)特征提取

提取能够表征图书内容的特征,是实现机器学习的首要前提。特征越能够表达图书的内涵,表明其区分度越高,机器学习的效果也将越好。因此,选择有效的书目特征是实现图书自动分类的关键。在本文中,笔者拟以图书中出现的有意义的词汇作为图书特征。首先采用由中科院计算所研制的 ICTCLAS 分词系统对图书各个著录项的文本进行分词,形成词语集合;然后去除其中所有无意义的词汇,包括停用词、部分高频词和低频词等,形成特征词语集合;对于任意一本图书,如果包含某特征词语,则在该特征上记为1,否则记为0,这样可以将任意图书转化为一个基本特征向量,如公式(1)所示。

$$bibliography = < Co_1 \quad \cdots \quad Co_i \quad \cdots \quad Co_m > (i = 1,2,\cdots,m) \tag{1}$$

其中:

$$Co_i = \begin{cases} 1 & bibliography \text{ 各著录项文本中包含词汇 } term_i \\ 0 & \text{否则} \end{cases}$$

(2)特征权重设置

图书著录信息不同于一般文本信息,它一般是由多个能够描述图书内容的著录项所组成,包括题名、关键词和文摘等,从这些类型的文本中都能提取与图书内容相关的特征词汇。但是各著录项对图书内容的表达能力不同,对揭示图书主题的贡献也各不相同。因此从各个著录项中抽取出来的特征词语在描述图书内容时就显示出了不同力度。一般认为,关键词是对图书内容的深度标引,基本上反映出图书的主要内容,根据关键词实现图书自动分类也是目前最常用的手段,它是最能体现图书主题的著录文本;题名是图书内容的深度浓缩,书中所论述的内容理论上要求围绕题名展开,从题名中提取出来的特征词语对图书主题的揭示也具有很大的作用;文摘也是对图书内容的概括,但其篇幅相对题名要大得多,因此其对图书主题的揭示就不如题名集中,从中提取出来的词语相对来说揭示性较低。因此,笔者根据各著录项对图书主题揭示程度的不同,为从中抽取出来的特征词语设置不同的权重。图书的特征向量被修正,如公式(2)所示。

$$bibliography = < Co_1 \quad Co_2 \quad \cdots \quad Co_i \quad \cdots \quad Co_m > (i = 1,2,\cdots,m) \tag{2}$$

其中:

$$Co_i = WeightT \times SumT_i + WeightS \times SumS_i + WeightA \times SumA_i$$

($WeightT$、$WeightS$、$WeightA$ 分别表示题名、关键词和摘要的权重;$SumT_i$、$SumS_i$、$SumA_i$ 则表示第 i 个特征词分别在当前图书的题名、关键词和摘要中出现的次数)

在本文中,$WeightT$、$WeightS$、$WeightA$ 分别被设置为3、4、1。例如,特征词"数据挖掘"在某图书的题名中出现1次,关键词中出现1次,文摘中出现4次,那么该图书的"数据挖掘"特征值为 $1 \times 3 + 1 \times 4 + 4 \times 1 = 11$。

(3)机器学习算法选择

常用的机器学习方法包括决策树、人工神经网络、支持向量机、贝叶斯分类等等。决策树算法(Decision Tree, DT)[7]试图通过对现有特征及类目数据的学习,以信息增益作为衡量标准,最终形成以特征的逻辑判断作为内部节点和类目作为叶子节点的二叉树或多叉树,进而可以根据未知数据特征值的大小在决策树上选择合理分类路径(类目),典型的决策树算法有 ID3[8] 和 C45[9]。神经网络(Nerve Net, NN)试图模拟人类思维,通过不断学习降低判断的错误率,进而形成能够对指定环境下状态进行自动判断的神经网络模型,最常用的神经网络分类算法是 BP(Back Propagation)神经网络(BP-NN)[10],是一种按误差逆传播算法训练的多层前馈网络,该算法能够对大规模的样本语料进行充分学习,从而实现对未知数据的分类和预测。支持向量机(Support Vector Machine, SVM)[11]则是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的一种模式识别方法,它使用数学方法和优化技术,通过对有限样本的学习,在学习精度和识别能力之间寻求最佳折衷以构建分类器,是目前应用范围较广、具有较好识别能力的分类方法。贝叶斯分类算法是一类利用概率统计知识进行分类的有效算法,它通过某对象的先验概率,

利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类，朴素贝叶斯（Naive Bayes，NB）[12]是其中比较典型的贝叶斯分类算法，该算法假设对象的各种特征相互独立。本文拟采用 BP－NN 和 SVM 作为书目自动分类的选择算法。

（4）分类方法确定

一般来说，可以采用两种方法来实现文本分类，即单层分类法（flat classification）和层次分类法（hierarchical classification）[13]。单层分类法，是指将数据中出现的所有类目均认为是独立类目，而忽略了类目间的相互关系，在分类时将文本归入到置信度最高的相关类目中。这种分类方法比较适合类目数量不多，类目间关系比较单一的数据环境。一旦分类体系中类目数量达到较大规模或类目相互之间具有层次关系时，其计算的复杂度将急剧上升，而对类目的区别能力则显著下降，分析的准确度迅速下滑。因此该方法并不适用于类目规模较大的分类环境。层次分类法，则是根据类目之间的层次关系，将一个复杂的分类任务分解为若干层规模较小的分类，逐层归类待分类文本。这种方法将复杂问题简单化，不仅极大降低了计算复杂度，而且适用于规模较大分类体系的自动分类。层次分类法将整个分类体系转化为如图 4 所示的树形结构，而分类问题就被转化为从根节点出发寻找叶节点的过程。首先，通过数据训练为每一个内部节点建立 m 元分类器，m 值即为当前节点子节点的个数；进行数据分析时，从根节点开始，使用 m 元分类器将测试实例分配到相应子结点所在的子树中，如此逐层分析，直到测试实例到达某叶节点，标志整个分类过程结束。图 4 中虚线箭头标出了测试实例层次分类的方向。鉴于本文采用的中图法分类体系类目繁多、结构相对复杂，因此笔者采用层次分类法来建立分类模型，进行实验分析。

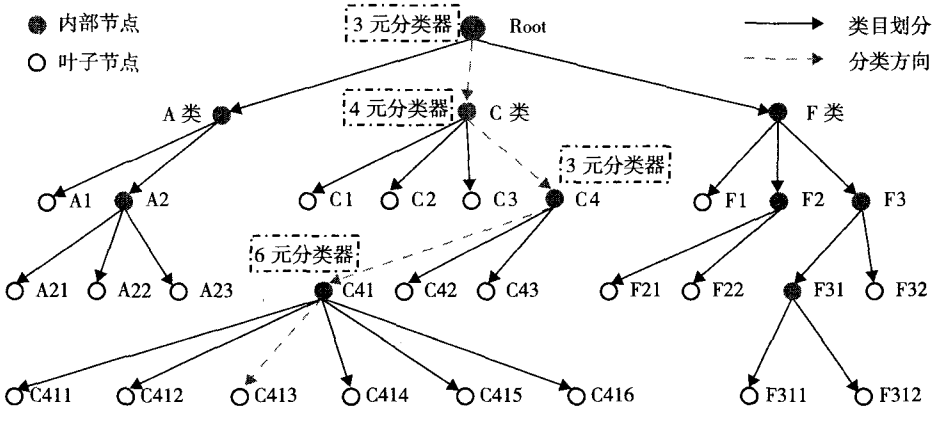

图 4　面向大规模分类体系的层次分类方法

2.2　面向机器学习的浅层次分类体系

为了进行实验分析，笔者抽取了南京大学图书馆收藏的部分图书信息建立了一个书目数据集（简称 DB_ACFNX），其中图书来自 5 大类：A 类（马列主义、毛泽东思想、邓小平理论）、C 类（社会科学总论）、F 类（经济）、N 类（自然科学总论）和 X 类（环境科学、安全科学），共计 23981 条书目记录。DB_ACFNX 中的书目均采用中图法进行标记，每一本图书均简化为只有一个类目号（如果涉及多个类目号的，则以第一个类目为准）；中图法类目号采用字母和阿拉伯数字混合编排的标记方式，其中字母放在类目

号首位,表示大类,中图法中一共定义了 A - Z 共 22 个大类;其后跟随整数,一般来说,每一位数字表示一个类层次,每 3 位数字后加"."隔开,也有个别类目繁多的大类标记后面紧跟字母表示下位类,如 T 大类的下位类;基于中图法的图书分类中,还存在大量的复分现象,例如《中国西北地区气候与生态环境概论》一书的类目号为 X171.1(2),其中(2)即为地区复分号,表示"中国"。复分现象的存在虽然使得图书分类更加确切,但是也给类目自动分配带来了相应的困难。为了尽量简化实验环境,排除干扰,本文暂时不考虑复分号。

通过上述对中图法分类号的简化处理,笔者分析了 DB_ACFNX 中书目数据的类目情况。在所有书目数据中,共涉及 1846 个类目号,根据图 4 所示的分类方式,可形成具有 9 层深度的层次结构。图 5 展示了面向中图法不同层次深度的书目分布(实线)和类目分布(虚线)情况。从图中可以发现,中间层次的书目和类目的数量都远远高于高层次和低层次,这使得书目数据呈现出一种分布不平衡、高层次和低层次数据稀疏的现象,这也表明在类目自动分配时,应该尽可能集中在 3 - 7 层;图 6 中实线展示了各层次类目的平均含书量,在第 6 层次已经下降到了 10 以下,虚线则描述了各个层次上稀疏类目(书目数量少于 4)所占的比例,在第 6 层次这个比例上升到了 70% 以上。由此可见,就 DB_ACFNX 语料库而言,层次 1 - 5 的书目数据可能具有较理想的学习效果,而在层次 6 - 9 中,由于类目中可训练语料的减少以及稀疏类目的增多,将导致这些层次的类目区分度明显下降。

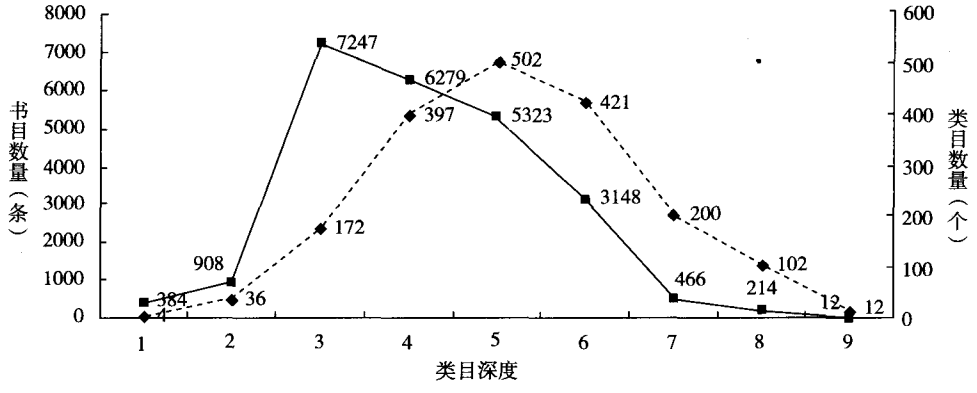

图 5 基于层次深度的书目和类目分布

综上所述,在层次分类法中,一方面,随着层次深度的不断增加,层次中包含的书目数量将会不断减小,而机器学习通常需要一定规模的训练语料进行学习,深层次小规模的数据量将会导致机器学习不充分,所获得模型的区分能力将严重下降,有必要将各层次小类目进行合并;另一方面,机器学习的理论基础是概率统计方法,在单层分类正确率上只可能无限接近于 100%,而始终无法达到,于是随着层次深度的增加,分类的综合正确率将会逐渐下降,这也决定了面向机器学习的层次分类体系不宜过大,以保证自动分类的实用性。为此,在基于机器学习的层次分类中需要对中图法的分类体系进行适当调整,压缩分类层次。鉴于实验数据量的限制以及中图法层次类目的特点,笔者将层次分类和单层分类方法相结合,建立了面向机器学习的 3 层分类体系,即:前两层采用中图法的分层方法,使得小类目尽可能分散在各个

大类目中;将中图法中所有3层以下(包括3层)的小类目合并,形成第3层,以保证第3层具有足够的训练语料。转化后的层次分类体系如图7所示。

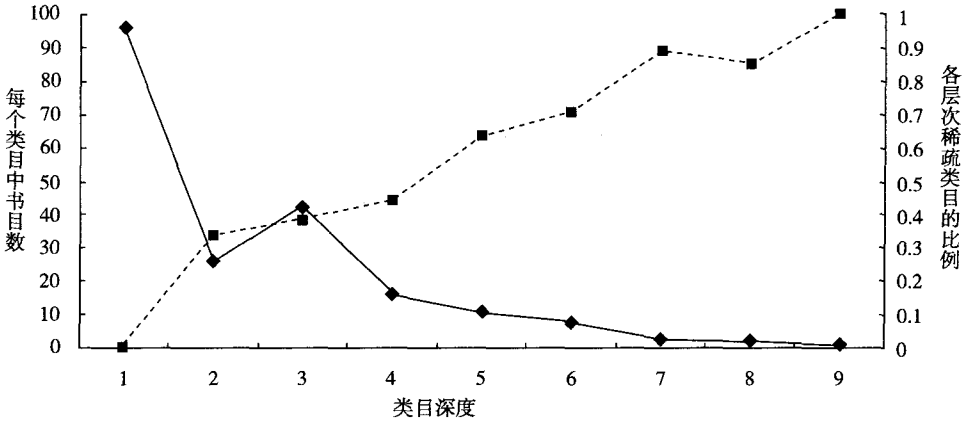

图6 各层次中每个类目含书量和稀疏类目所占比例

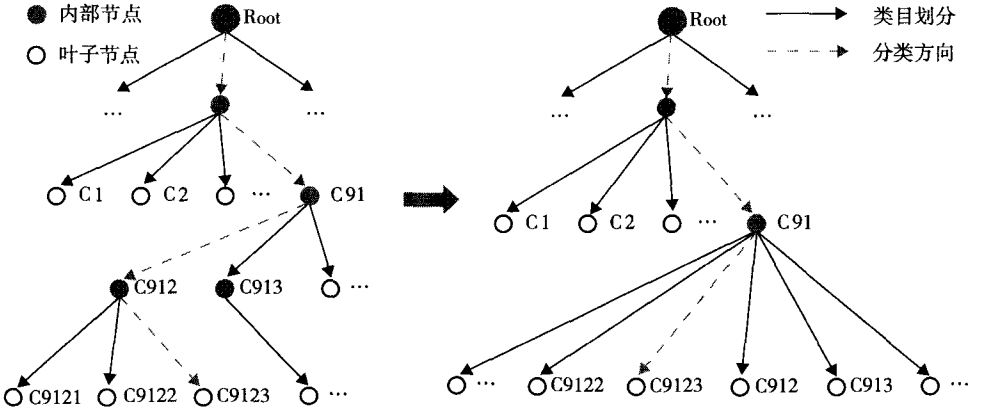

图7 中图法分类体系转化为面向机器学习的层次分类体系

3 基于关键词的书目自动分类实验分析

关键词是对书目内容的集中反映,在很多情况下均视其为书目主题,根据关键词判断图书类目是合理的做法。笔者以关键词作为描述书目的特征集合,为每个书目构建关键词向量,形成书目×关键词的二元关联矩阵,结合书目的类目信息,作为数据对象来进行单层图书自动分类实验。

在DB_ACFNX中,标注了关键词的书目共有16501条,其中A大类3096条,C大类8083条,F大类1154条,N大类1706条,X大类2462条,特征则有2161个。笔者把所有可用书目数据分为两个部分:随机抽取15000条数据作为训练样本,在学习算法的作用下获得大类分类器;其他的1501条数据作为分析样本,通过分类器

后即可自动获得大类类目。表1显示的书目×(关键词特征+类目)矩阵作为训练样本,根据公式(1),当书目包含某关键词时,其相应的矩阵值为1,否则为0。

表1 书目×(关键词特征+类目)二元关联矩阵

书目号	…	毛泽东思想	管理信息系统	企业管理	自然灾害	环境保护	可持续发展	…	类目
…	…	…	…	…	…	…	…	…	…
0000042117	…	0	1	0	0	0	0	…	C
0000078094	…	1	0	0	0	0	0	…	A
0003460570	…	0	0	1	0	0	0	…	F
0000081180	…	0	0	0	0	0	1	…	N
0003149268	…	0	0	0	0	1	0	…	X
00033496120	0	0	1	0	0	…	X		
…	…	…	…	…	…	…	…	…	…

在C大类的8083条书目中,一共可以分为15个子类,包括没有子类的空类目;同时根据中图法的分类规则,把编号为'9'的子类拆分为'91'、'92'、'93'、'95'、'96'和'97',和其他子类并列为C大类下的二级类目;此外,从中可以提取特征840个,用于描述C类书目。本实验以C类中7000条书目作为训练样本,余下1083条数据则作为分析样本。

针对上述的实验数据,笔者采用BP-NN和SVM算法,分别以一级类目和C大类下的二级类目作为单层分类体系,进行了自动分类实验。实验结果如表2和表3所示。

表2 面向主题基于机器学习的图书自动分类实验结果

算法	一级类目			二级类目(C类)			二层分类综合值(%)
	正确数	错误数	正确率(%)	正确数	错误数	正确率(%)	
BP-NN	1405	96	93.60	822	261	75.90	71.04
SVM	1412	89	94.07	985	98	90.95	85.56

表3 基于SVM的一级类目自动分类实验结果

类目	正确识别数	应该识别数	实验识别数	正确率(%)	召回率(%)	F1值(%)
A大类	114	123	115	99.13	92.68	95.80
C大类	910	918	972	93.62	99.13	96.30
F大类	6	14	24	25.00	42.86	31.58
N大类	134	144	138	97.10	93.06	95.04
X大类	248	302	252	98.41	82.12	89.53

从总体上看：①无论采用 SVM 算法，还是 BP－NN 算法，单层图书自动分类的正确率都已经超过了 75%，两层分类的综合正确率也都超过了 70%，表明机器学习方法在浅层图书分类中具有实用价值；②在当前的实验环境下，SVM 算法的图书分类效果优于 BP－NN，在训练语料的规模达到一定程度后，这种优势明显减弱；③当训练语料达到较大规模时，两种算法的分类正确率均超过了 93%，可以说达到了很好的分类效果。然而当训练语料的规模下降，分类类目数增多时，两种算法的正确率都出现了明显的下滑，特别是 BP－NN，下滑态势非常明显，说明 BP－NN 更适合于大型的学习场合，大规模的训练语料使得神经网络具有更好的区分度，而 SVM 则保持在一个比较稳定的水平，训练语料规模的减少对 SVM 模型的分类效果有影响，但影响不大。

从各类目的识别情况来看：①大部分类目的分类结果都比较理想，4 个大类的分类 F1 值都达到或接近了 90%，其中 3 个大类的 F1 值甚至超过了 95%，可见它们的正确率和召回率都已经达到了相当高的水平；②其中 F 大类的分类效果很差，一方面可能是因为 F 大类的训练语料较少，分类器没有训练到最佳状态，更重要的则可能是测试语料中该类目的图书很少（数据稀疏），被识别为该类目的可能性被极大地降低。

4 基于特征加权的书目自动分类实验分析

关键词（或主题）标引是自然语言处理研究中的难题，F1 值一般徘徊在 30%—50% 之间，即便采用在序列标注中具有很强识别率的条件随机场（CRFs）模型，其 F1 值也仅仅达到 51.25%[14]，还远远没有达到实用程度。因此，基于关键词实现图书分类存在一定的条件限制，在 DB_ACFNX 中就有 7480 条书目没有标注关键词。本文提出了基于特征加权的文本自动分类方法，以解决没有标注关键词情况下的图书分类问题。

笔者以题名、关键词和摘要作为本次实验书目的特征来源。先对各特征来源进行分词处理，然后基于停用词表排除其中的无意义词汇，进而根据一定的选择标准确定特征词。特征词选择的标准为：①特征词在每一书目中权重均大于 1，在书目中权重为 1 的词可以认为对当前书目没有描述意义；②特征词在所有书目中权重和大于 3，表明该特征词在类目中的出现不是偶然现象，而是具有一定普遍性。经过筛选，共获得特征词 5487 个，分布在 23912 条书目中，取其中 20000 条书目作为训练语料，其他 3912 条书目作为测试语料，以 SVM 作为分类算法，中图法一级类目作为分类目标。实验结果如表 4 前部所示。

为了进一步验证基于特征加权图书分类方法的正确性和合理性。笔者对二级类目 C 类中的书目分类也进行了实验分析。特征词仅要求在每一书目中权重大于 1，以保证具有足够的特征数量以区分书目。这样可获得特征词 5393 个，分布在 9828 条 C 类书目中，以其中的 9000 条数据作为训练样本，828 条数据作为测试样本，以 C 类下 15 个子类为分类目标。实验结果如表 4 后部所示，最后可计算出二层分类的综合正确率。

表 4 基于特征加权的图书自动分类实验结果

算法	一级类目			二级类目（C 类）			二层分类综合值（%）
	正确数	错误数	正确率（%）	正确数	错误数	正确率（%）	
SVM	3753	159	95.94	757	71	91.43	87.72

对比表2和表4,不难发现:①同样基于SVM算法,无论是一级还是二级类目,后者的正确率都比前者高,导致这个结果的原因可能是因为后者的训练语料的规模比前者大,但不可否认的是,在没有标注关键词的情况下(基于特征加权的图书分类方法虽然也以关键词作为特征来源,但是对关键词进行了切分),也可以直接基于机器学习的方法实现图书分类目的;②在保证训练语料规模的情况下,基于特征加权的图书分类方法的正确率均超过或接近92%,在训练语料规模增大的情况下,这个比率还会变大;③二层分类综合正确率也超过了87%,说明基于特征加权的研究思路在书目层次分类中具有一定的实用价值。

为了进一步验证训练语料规模对分类效果的影响,笔者针对不同规模的训练语料分别进行了SVM机器学习,获得的分类器的分类正确率变化情况如图8所示。很明显,在测试语料完全相同的情况下,训练语料规模的减少将会降低分类器的性能,而且下降的趋势也越来越明显。因此,在训练实用的书目分类器时,一定要保证有充分的训练语料,以保证分类器的性能。

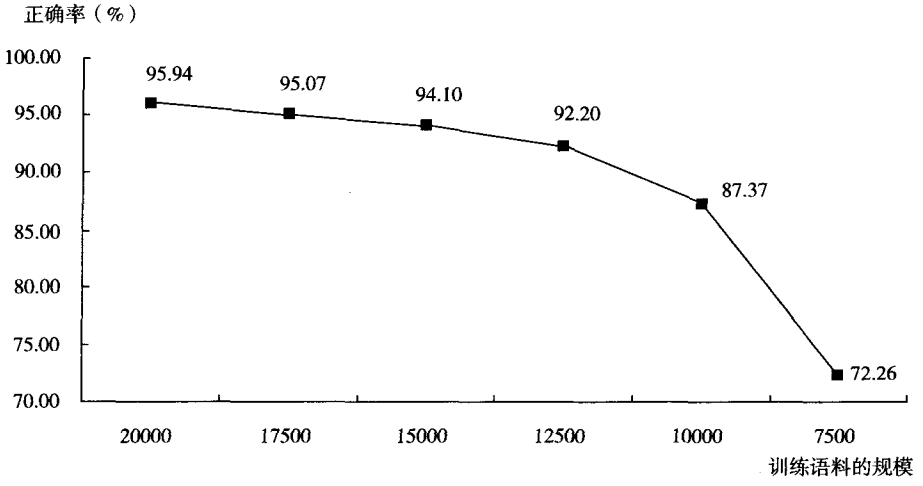

图8 训练语料规模对分类器分类效果的影响

在上面的实验中,关键词始终作为书目的特征来源,因此高正确率可能是由于关键词作为特征造成的。那么,在关键词完全没有标注时,仅以题名和摘要作为特征词来源的情况下,SVM算法的分类性能又将如何?为此,笔者对特征来源的不同组合进行了实验对比,结果如表5所示。在训练语料规模不变的情况下,由于特征来源的减少,所获得的分类器的分类正确率从95.94%下降到了88.40%,进而再降至87.80%,表明在特征来源减少的情况下,对书目主题的揭示程度也随之降低,书目的特征描述向量无法充分表示书目;关键词对书目主题的揭示能力明显强于摘要,由于缺少关键词,分类器的分类正确率由原来的95%以上下降至90%以下,可见关键词的存在能够极大地提高书目分类正确率;在只有题名作为特征来源的情况下,分类正确率接近88%,可见即使某书目只能提供题名,在3层分类体系中,将该书目正确分类的可能性也能接近70%。

表5 特征来源不同对分类器分类效果的影响

算法	特征来源	训练语料规模	测试语料规模	一级类目		
				正确数	错误数	正确率(%)
SVM	题名+关键词+摘要	20000	3912	3753	159	95.94
SVM	题名+摘要	20000	3785	3346	439	88.40
SVM	题名	20000	3736	3280	456	87.80

5 浅层次书目自动分类实验分析

前文实验均采用层次分类法来实现图书自动分类,然而从实验结果不难发现:随着分类层次深度的增加,分类的正确率随之降低,可以预想当层次增加到一定深度时,分类正确率将会下降到一个较低的水平;而且由于数据来源的限制,当类目下降到第3层时,数据量已经比较小了,例如DB_ACFNX中最多的C91类目中所有数据仅为3557条。为此,笔者提出了将中图法转化为3层次分类体系的解决方案。本节实验试图计算某指定类目的3层分类综合正确率,以验证浅层次书目自动分类方法的有效性。

笔者选择样本数据量最大的C91类目进行第3层分类实验,将C91以下所有类目作为第3层分类目标。将停用词以外的所有词语作为特征词,这样可获得5620个特征词,分布在3557本图书中,取其中3300条数据作为学习样本,剩下257条作为测试样本,C91下连同不分类的情况共有42个类目。最终正确分类了224条,分类正确率达到了87.16%。

结合上节的实验结果,可获得如图9所示的分类层次与训练样本及正确率之间的关系图。可以看出:①随着分类层次的加深,单层分类的正确率逐渐下降,可能是由于分类的训练样本数逐层下降所导致;②单层分类正确率逐层下降的趋势并不明显,当学习样本数下降到3300条时,仍能保证有87%以上的正确率,因此在训练样本规模上最好控制在5000条以上,以保证分类具有较好的效果(接近或超过90%);③3层分类的综合值仅为76.45%。由于单层分类的正确率不可能达到100%,多层次分类的正确率会逐层下降,即使每层分类的正确率都达到了96%,那么3层之后正确率会下降至88.47%,因此在实际应用中,类目深度不宜设置过大,同时也要保证每层类目中的类目数不宜过多。在中图法分类体系中,本文提出的设置3层分类是比较合理的选择,即如果单层分类正确率能够达到97%,那么就能够保证图书分类的正确率在90%以上。

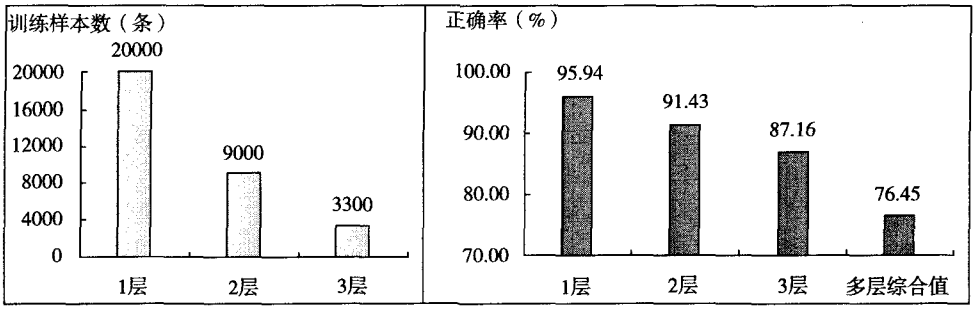

图9 分类层次与训练样本及正确率之间的关系

6 结语

本文在提出采用机器学习方法实现浅层次书目分类模型的基础上,以部分类目的图书作为实验对象,对模型进行了实验论证和分析。笔者认为:①以特征加权的方式来体现特征词对书目的贡献率是可行的方法;②在各种机器学习算法中,SVM 在面向大规模训练语料的环境下具有较好的分类效率和效果;③面向中图法的层次分类中,类目深度不宜过大,控制在3层可以获得较好的分类效果。在实验数据量并不是非常充分的情况下,基于特征加权的书目浅层次分类取得了76.45%的分类正确率,说明该模型具有一定的实用价值。

在文本的实验基础上,笔者认为可以针对中图法中不同类目或子类目建立实用的 n 元分类器,用于图书馆自动编目工作中,促进数字图书馆事业的发展;机器学习方法作为人工智能的一个重要方向,不仅可以用于书目自动分类研究中,图书馆中所有具有选择性的工作都可以考虑采用机器学习方法,例如书目主题词的确定、虚拟参考咨询服务中自动问答系统的实现等等;此外,中图法中存在大量的图书复分现象,为了简化实验环境,本文没有考虑为书目自动添加复分号的方法,这也可以作为我们进一步的研究方向,以完善面向中图法的书目自动分类系统。

参考文献:

[1] 王昊. 基于本体的 CSSCI 学术资源网络模型构建及应用[D]. 南京:南京大学, 2008.
[2] 何琳, 侯汉清, 白振田, 等. 基于标引经验和机器学习相结合的多层自动分类[J]. 情报学报, 2006, 25(6): 725 - 729.
[3] Tom M M. 机器学习[M]. 曾华军, 张银奎, 等, 译. 北京: 机械工业出版社, 2003.
[4] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1 - 47.
[5] Maron M. Automatic indexing: An experimental inquiry[J]. Journal of the Association for Computing Machinery, 1961, 8(3): 404 - 417.
[6] Gennari J H, Musen M A, Fergerson R W, et al. The evolution of protégé: An environment for knowledge-based systems development[J]. International Journal of Human-Computer Studies, 2003, 58(1): 89 - 123.
[7] 王桂芹, 黄道. 决策树算法研究及应用[J]. 电脑应用技术, 2008(72): 1 - 7.
[8] Quinlan J R. Induction of decision tree[J]. Machine Learning, 1986,1(1):81 - 106.
[9] Quinlan J R. C4.5: Programs for machine leaning [M]. Los Altos, California: Morgan Kaufmann Publishers, Inc. ,1993.
[10] Hecht-Nielsen R. Theory of the back propagation neural network[C]. In Proceedings of International Joint Conference on Neural Networks, IEEE, 1989, 1: 593 - 603.
[11] Cortes C, Vapnik V. Support-vector network[J]. Machine Learning, 1995(20): 273 - 297.
[12] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997 (29) : 131 - 163.
[13] Wang Jun, Lee Meng-Chen. Reconstructing DDC for interactive classification [C]. Conference on Information & Knowledge Management, 2007 (CIKM07).
[14] 章成志, 苏新宁. 基于条件随机场的自动标引模型研究[J]. 中国图书馆学报, 2008(5): 89 - 94, 99.

王 昊 南京大学信息管理系讲师。通讯地址:江苏省南京市汉口路22号。邮编:210093。
严 明 解放军南京政治学院基础部教授。通讯地址:江苏省南京市中山北路305号。邮编:210003。
苏新宁 南京大学信息管理系教授、博士生导师。通讯地址:江苏省南京市汉口路22号。邮编:210093。

(收稿日期:2010-05-14)