分层特征计算和错误控制的层次分类方法

吴碧军 李涓子 金 鑫
(清华大学计算机系 北京 100084)

摘 要 中文新闻信息分类标准中,类别数量大。在将其应用于新闻分类时,会出现训练模型大、训练时间长,尤其是当部分类别改变时需要全部重新训练等问题。由于分类标准中类别之间存在层次关系,因此层次分类方法可以作为解决方案。研究层次化的中文新闻分类方法,并从以下两方面改善层次化分类方法的效果:1)分层的新闻特征计算,解决了层次分类中新闻在分类类别下的特征向量的不同表示的问题;2)错误控制,解决了在上一层分类错误的情况下新闻不会分到正确的类别上的情况。实验结果表明,层次分类方法的效果比平面分类的准确度提高了约4%,进行多次特征权重计算的层次分类方法比普通的层次分类的准确度提高了约3%,同时进行错误控制的分类效果比普通层次的分类效果提高了大概3%。

关键词 层次分类,支持向量机,中文信息分类标准,特征计算,错误控制
中图法分类号 TP311 **文献标识码** A

Hierarchical Classification Approach of Hierarchical Feature Selection and Error Control

WU Bi-jun LI Juan-zi JIN Xin
(Department of Computer Science & Technology, Tsinghua University, Beijing 100084, China)

Abstract There are thousands of subjects in Chinese news subject specification. When they are used in news classification, long training time and large model are two key problems we are facing, especially when some of classes are changed. Chinese news subject classification has hierarchical structure and hierarchical can solve the problem partially. We improved the Chinese news hierarchical classification to get better the result from two points of view. 1) Repetitious feature calculation represents news of different layers in hierarchical classification. 2) Use error control to solve the problem that one error classification in upper layer will lead in the error classification of its deeper classes. Our experiments shows that hierarchical classification improves the precision of 4% comparing with flat classification, hierarchical classification with Repetitious feature calculation improves 3% comparing with hierarchical classification, and hierarchical classification with error control improves 3% comparing with hierarchical classification.

Keywords Hierarchical classification, Support vector machine, Chinese news subject classification specification, Feature calculation, Error control

1 引言

在当今信息爆炸的时代,人们每天面对海量的中文新闻信息。如何在海量的新闻信息里面找到自己感兴趣的新闻?对新闻进行分类管理和服务是必须的。但传统的手工分类由于周期长、费用高、效率低,而且需要具有专业知识的人员才能胜任,难以满足当今的实际应用需要。自动分类技术是一个在给定分类体系下利用计算机技术对新闻进行自动判别类别的方法。

据统计,我国新闻信息分类中使用自定义分类法的占大部分,同时有一部分使用中图法、中国新闻资料分类和录像节目资料分类法等。由此可见,我国新闻信息分类方法众多,不利于信息共享。为了实现各个部门之间中文新闻信息的共享,必须具有统一的分类标准,因此中文新闻信息分类标准顺利出台。

中文新闻信息分类标准是一个树状结构,包含23个顶层类和5761个子类别,是一棵十分庞大的树。当面对在中文新闻信息分类标准上的一个分类问题时,按照平面SVM的分类方法,如果不考虑分类中的层次关系,则将非叶子节点和叶子节点看作是平等的,将构建一个包含上千类别的分类器。即使每个类别选取较少个数的训练集,生成的分类器也将非常巨大,同时训练所需的时间也将非常漫长。

对新闻进行层次分类的方法是一个有效的解决途径。首先,根据中文新闻信息分类标准,类别本身具有层次关系,而不是平等的关系;其次,在CPU为Intel双核2.0G,内存2G,操作系统为Windows XP下,经过实验,在每个类别训练集个数一定的情况下,生成的分类器的大小和类别个数基本上呈线性增长,而训练时间和类别个数基本上呈指数增长,实验结

到稿日期:2009-11-26 返修日期:2010-01-20 本文受国家973项目(No. 2007CB310803)资助。
吴碧军(1981—),男,硕士生,主要研究方向为数据挖掘和知识发现,E-mail:wbj@keg.cs.tsinghua.edu.cn;李涓子(1964—),女,教授,主要研究方向为数据挖掘和知识发现;金 鑫(1981—),男,硕士生,主要研究方向为数据挖掘。

果如图 1 所示；再次，平面分类时，若部分类别改变，则需要重新训练模型文件，而层次分类时只需要训练改变的那些类别对应的模型文件，不需要全部训练，这将节省大量的人力和物力。

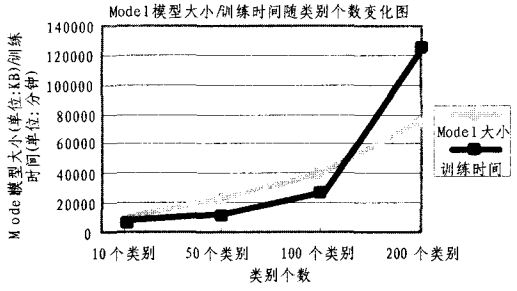

图 1　Model 模型大小和训练时间随类别个数变化图

层次分类方法不但在理论上可行，而且实验证明：层次分类方法的效果比平面分类方法的效果要好，进行多次特征权重计算的层次分类方法比一般的层次分类效果要好，同时，进行错误控制方法对分类效果也有了明显的提高。

本文第 2 节介绍层次分类的相关工作；第 3 节介绍层次分类的方法，包含特征选择、分层特征重新权重计算、错误控制；第 4 节介绍实验；最后为结论。

2　相关研究工作

现在分类方法有很多，主要包括 KNN，Bayes，决策树和支持向量机等。其中支持向量机(SVM)在文本分类上取得了比较好的效果。支持向量机(Support Vector Machines，SVM)是根据统计学习理论中结构化风险最小化的原则提出来的，是一种优秀的基于数据的机器学习算法，在分类问题的应用中取得了良好的效果[1]。同时，T. Joachims 开发了一个 C 语言实现的平面 SVM 分类器[2]。

平面分类认为每个类别都是独立和平等的，而在客观世界中，类别都是具有层次关系的，因此层次分类方法越来越受到重视。类别层次结构包括以下 4 种类型：①虚拟类别树：所有类别被组织成一个树型结构，每个类别至多有一个父类别。文本只能够被分到叶类别中；②类别树：所有类别被组织成一个树型结构，每个类别至多有一个父类别。文本既能够被分到叶类别中，也能够被分到内部类别中；③虚拟有向无环类别图：所有类别被组织成一个有向无环图，文本只能够被分到叶类别中；④有向无环类别图：所有类别被组织成一个有向无环图，文本既能够被分到叶类别中，也能够被分到内部类别中[3]。文献[4]提出利用层次神经网络方法进行层次分类。在面对大规模的类别和海量数据，利用层次 SVM 方法，把复杂的一次分类问题转化为多次的层次分类问题，取得了良好的效果[5,6]。文献[7]提出利用层次 SVM 分类方法进行分类，取得了比平面分类更好的效果。文献[8]提出在 SVM 分类的基础上结合 Bayes 方法，比普通 SVM 层次方法效果要好一点，但没有明显改进。面对在层次分类方法中，上一层错误分类将会导致分类数据永远分不到正确的类别上，错误控制是解决这个问题的基本方法[9]。

在进行特征权重计算中，文献[5]提出利用互信息来计算特征向量；文献[10]提出用类中心点的方法计算特征向量；文献[11]提出用尽量少的词来表示文本，以降低复杂度；文献[12]提出用词组代替单个词计算特征向量。基于中文新闻的具体情况，出现了中文新闻关键字抽取方法[13]，同时文献[13]提出一个在平面分类中已经证明比较好的特征向量计算方法。

上述方法存在着随机选择训练集的问题，特别对于顶层类来说，要保证覆盖所有的子类，同时保证在每个子类上训练集要均衡。为了更好地表现文本的表征意义，单个的词在表达文本语义上都有所欠缺。

本文采取文献[3]中提出的第一种分类类型虚拟类别树进行分类，以 SVM 为基础构架层次分类器把新闻分到叶子节点上。选择训练集时，采取自底向上的方法构建分类器，保证训练集的完整性和均衡性。为了更好地体现文本的表征意义，采用关键字为基础计算特征向量。结合文献[9]和 SVM 分类方法，提出一个面对 SVM 层次分类的错误控制方法。

3　本文的方法

分类问题一般分为训练和标注两部分。在本文所描述的层次分类方法中，在训练部分，在层次分类体系下，根据人工分类标注，对每个非叶子节点构建一个分类器。在构建下一层分类器时，为了更好地表现在当前分类类别下的新闻，需要重新计算特征向量和关键字抽取。在标注部分，根据层次分类体系，自顶向下，一直分到叶子节点为止。同时，在标注部分必须进行重新计算特征向量。具体如图 2 所示。

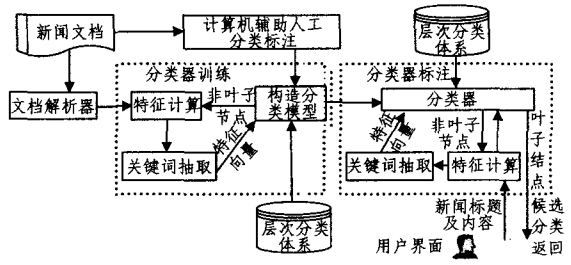

图 2　层次分类步骤图

层次分类方法主要包括特征选择、分层特征权重计算和错误控制 3 部分。

3.1　特征选择

特征选择作为分类的基础，对分类效果具有很大的影响。相比于一般文本分类采用一元词表示特征，一元词的表征意义不够丰富，同时数目也比较多。考虑到中文新闻的特点，对新闻的表征意义来说，关键词是更有意义的语言单位。关键词包括一元词、二元词和三元词。中文中最基本的词是一元词，而二元词和三元词是一元词相互合并得来的。比如说，经济和航母都是一元词，当这两个词连接在一起出现时，就成二元词经济航母，相比简单的两个一元词经济和航母更能表示新闻。同时，通过二元词和三元词，我们可以用较少的特征表达相同的新闻信息，自然就降低了算法的复杂性。

3.2　分层特征权重计算

对新闻进行特征向量表示，包括新闻预处理、关键字抽取和特征计算 3 个方面。新闻预处理包含分词、词性标注，然后去掉停用词和虚词，得到一个候选的关键词列表；在候选的关键词列表中，根据词频计算一个权值，再根据权值进行大小排序后得到一个关键词列表；最后利用特征权重的方法计算得出特征向量。图 3 描述了这个过程。

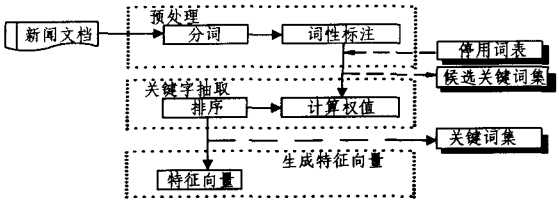

图 3　特征向量生成步骤

在利用 SVM 方法对新闻分类中，特征向量的选择是一个非常重要的方面，对分类的效果有很大的影响。特征权重计算方法包括 TF/IDF、互信息和信息增益等各种不同方法。通过研究，式(1)[6]所示为改进的 TF/IDF 方法对新闻分类比其他方法具有更好的分类效果。

$$W(t,\vec{d})=\frac{tf(t,\vec{d})*\log(N/n_t+0.01)}{\sqrt{\sum_{t\in\vec{d}}[tf(t,\vec{d})*\log(N/n+0.01)]^2}} \quad (1)$$

式中，$W(t,\vec{d})$为词 t 在文本$\vec{d}$中的权重，$tf(t,\vec{d})$为词 t 在文本$\vec{d}$中的词频，N 为训练文本的总数。

在中文新闻信息分类标准中，有上千个类别，几万乃至几十万的特征，如果采用平面分类特征向量生成方法，肯定不能满足层次分类的需求，同时会导致效率比较低。

在层次分类中，在进行上一层分类后，特征向量不能准确表示在当前分类类别下的新闻，而需要对特征向量进行重新计算，降低区分上一层类别之间的词的权重，增加区分当前类别之间词的权重。同时，对特征向量进行降维操作，在上一层分类中分到同一类别的新闻必然有其共性，这些共性的存在使得下一层分类器变得困惑。特征重新选择不但降低了计算复杂度，而且使下一层的分类器变得更加准确。

在图 4 中，对层次分类的第一层进行分类，政治、政策和党派这类词对区分新闻是否属于政治类有很大的关系；经济、企业、贸易和通货膨胀这类词对区分新闻是否属于经济类有很大的关系；体育、比赛、冠军和足球这类词对区分新闻是否属于体育类有很大的关系。但是，对第二层进行分类时，在区分体育类下面足球、篮球和排球的类别时，特征向量需要重新计算，降低体育、比赛和冠军这些词的权值，而增加足球、篮球和排球这些词的权重。同时考虑到它们都属于体育类，具有一定共性，进行降维操作是很合理的。

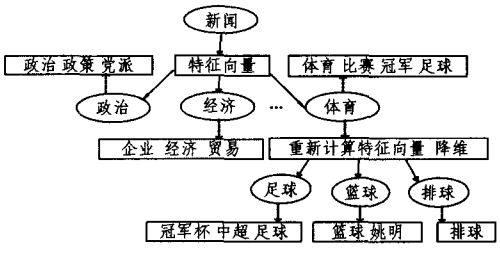

图 4　特征向量重新计算示意图

在式(1)特征向量选取公式中，由于在子类别进行重新分类时，选择它的子类对应的新闻集作为其文本集，公式中对应参数都会相应改变。具体来说，对区分上一层类别的关键词的 IDF 值会显著增加，对区分当前类别的词的 IDF 值会显著减少，重新计算的特征向量更能代表在当前分类类别下新闻的分类表示。同时根据层次的深度，按照递减的规则进行降维。

3.3　错误控制

由于层次分类存在这样的问题：当一个文档在某一层分类错误后，文档就永远都不可能分到正确的类别上。当新闻在某一层上分类错误时，在错误类的子类上继续分类将会导致在这些子类上得到的权值都不高，而在正确的类别的子类上得到的权值都比较高。为此，我们在某一层上多选择几个类别，在每个类别下再进行分类。然后对这些权值通过某种方法计算出一个新的权值，从而根据权值大小来确定它属于某个类别。这些方法包括累加法和累乘法等。

本文进行错误控制的方法可以表示如下：

用 value[i]数组表示 SVM 分类器得出的第 $i+1$ 个值，classname[i]为第 $i+1$ 值对应的类别。

```
If value[1]−value[2]
  Annotation classname[1];
Else {
  For(i=0 ;i<countCandidateClass ;i++ ){
    If CandidateClass[i] is leaf
      tempValue[number++]=value[i] * value[i] ;
    else {
      Nextlayervalue[] = SVMClassification;
    for(int j=0;j<countCandidateClass;j++){
        tempValue[number++]=value[i] * Nextlayervalue
                                        [i];
    }
  }
}
Sort(tempValue);
Annotation classname[max tempValue];
}
```

图 5 对错误控制方法的步骤做了更直观的描述。

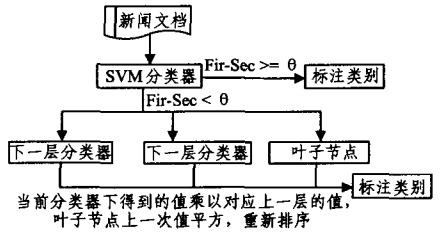

图 5　错误控制步骤图

4　实验

4.1　实验数据及评测方法

本文采取的分类体系是中文新闻信息分类标准。标准的建立是为了满足全球中文传媒多样化需求，提高新闻信息的交换能力，实现中文新闻信息在全球华文传媒乃至更大范围内的交流和共享，提高我国新闻宣传报道质量，拓展新闻信息服务新领域，为深层次的信息挖掘提供技术基础，为广大受众提供更加快捷、丰富、优质的新闻信息服务。它以政治、经济、文化作为分类大纲，包括 23 个顶层类、5000 多个叶子节点类。

本节将通过实验来比较平面分类和层次分类，普通层次分类和进行特征重新选择的层次分类，普通层次分类和进行错误控制的层次分类方法的分类效果。由于中文新闻信息分类标准中类别十分庞大，本文只选取其中的一部分来实验，包

括 23 个顶层类、146 个二层类和 179 个叶子节点。本文随机选择训练集和测试集，而对于训练集和测试集选择问题，在平面分类中，对 179 个叶子节点相对应类别选取 100 篇新闻作为其训练集、10 篇作为测试集；在层次分类中，对 23 个顶层类，每个类别选取 150 篇新闻作为训练集、10 篇作为测试集；对第二层的每个类别选取 80 篇新闻作为训练集、10 篇作为测试集。在选择训练集和测试集的过程中，自底向上选择，以保证每个类别的训练集和测试集的完整性和均衡性。

本文对查准率、查全率和 F1 测试值进行评估。查准率：所有判断的新闻文档中与人工分类结果吻合的新闻文档所占的比率；查全率：与人工分类结果应有的新闻文档中分类系统吻合的新闻文档所占的比率。查准率和查全率反映了分类质量的两个不同方面，两者必须综合考虑，不可偏废。F1 测试值：查准率×查全率×2/(查准率+查全率)。

4.2 实验结果及分析

4.2.1 平面分类与层次分类的效果比较

为了比较平面分类和层次分类的效果，本文做如下假设：在平面分类中，构建一个虚拟的层次关系，对于 179 个叶子节点我们从底向上返回到顶层的 23 类中(比如说，顶层类 A 有子类 A1，A2 等，本来属于 A1 的新闻把它错分到 A2 中；但在平面分类第一层中，仍然认为这个分类按照层次关系是正确的)，在各个子类别中随机选择数量一定的训练集。分类效果比较如表 1 所列。

表 1 平面和层次分类效果比较图

	查准率	查全率	F1 测试值
平面分类第一层	80.14%	76.28%	78.16%
层次分类第一层	82.28%	82.93%	82.60%
平面分类第二层	90.96%	89.78%	90.37%
层次分类第二层	92.52%	91.54%	92.03%

在第一层的各个类上 Precision 比较如图 6 所示。

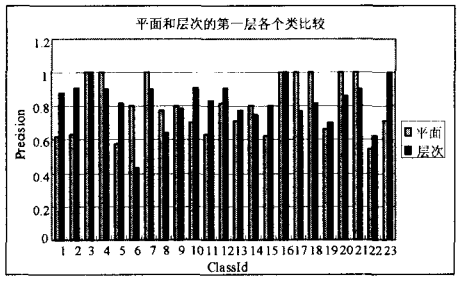

图 6 查准率图

在第一层的各个类上 Recall 比较如图 7 所示。

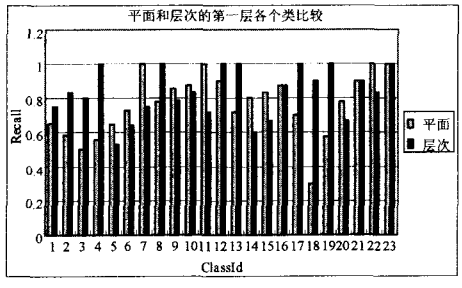

图 7 查全率图

在第一层的各个类上 F1 测试值比较如图 8 所示。

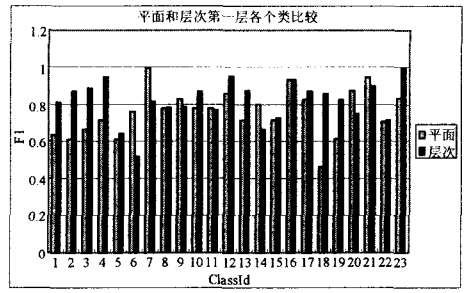

图 8 F1 值图

由于在层次分类中，一篇新闻在某个层次(非叶子节点)分错以后，它将永远也分不到正确的类别上(解决方案在下面的章节错误控制中)，为了不影响比较的效果，在第二层的比较当中，我们过滤掉错误的分类和顶层类为叶子节点的分类，以保证第一层结果分类正确。

由于普通层次分类方法考虑了类别之间的层次关系，采用分而治之自顶向下的方法，增加了模型的分类能力。相比于平面分类方法，在时间复杂度上也有明显的改进。同时从实验结果可以看出，层次分类第一层和平面分类第一层相比在查准率、查全率和 F1 值上分别提高了 2.14%，6.65%和 4.44%，在第二层上分别提高了 1.56%，1.76%和 1.66%。在层次分类和平面分类的第一层的 23 个类别上，层次分类在大部分类别上的查准率、查全率和 F1 值都比平面分类要高。

4.2.2 普通层次分类与重新计算特征权重层次分类的效果比较

根据式(1)，当训练集发生改变时，对区分上一层类别的关键词的 IDF 值会显著增加，对区分当前类别词的 IDF 值会显著减少，重新计算的特征向量更能代表当前分类器下新闻的分类表示。分类正确度比较如表 2 所列。

表 2 重新计算特征权重和错误控制的分类效果

	查准率	查全率	F1 测试值
普通层次分类	92.52%	90.28%	91.39%
重新计算特征权重的层次分类	94.82%	93.98%	94.4%
普通层次分类(前三类)	93.08%	92.82%	92.95%
错误控制的层次分类(前三类)	95.21%	96.54%	95.87%

从实验结果可以看出，重新计算特征权重的层次分类与普通层次分类相比在查准率、查全率和 F1 值上分别提高了 2.30%，3.70%和 3.01%。

4.2.3 普通层次分类与进行错误控制层次分类的效果比较

为了尽可能提高分类效果，提供多个结果供用户选择。本文在某一层上多选择几个类别，在每个类别下再进行分类。对这两个权值通过累乘法得到新的权值，从而确定它属于某个类别。在实验中，本文取前 3 个分类类别作为分类结果，即本来的类别为这 3 个类别之中的一个就认为分类正确。通过利用错误控制的方法后，分类性能有了明显提高，如表 2 所列。

从实验结果可以看出，重新计算特征向量的层次分类和普通层次分类相比在查准率、查全率和 F1 值上分别提高了 2.13%，3.72%和 2.92%。

总之，本节实验证明了层次分类比平面分类的效果要好，在查准率、查全率和 F1 值方面都有明显提高。为了更好地表征当前类别下的文本，通过对特征进行重新选择在分类效果上取得了良好的结果。为了解决上一层类别上的错误分类

(下转第 180 页)

$C^e(x_1,x_2,\cdots,x_l,0.5,\alpha_1,\alpha_2,\cdots,\alpha_l)$
$=\min(1,\max(0,\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_lx_l-(l-1)e))$

下限组合　突变组合

$C^e(x_1,x_2,\cdots,x_l,0,\alpha_1,\alpha_2,\cdots,\alpha_l)$
$=ite\{0|x_1,x_2,\cdots,x_l<e;1|x_1,x_2,\cdots,x_l>e;e\}$

文献[14,15]中都讨论了二元加权零级泛组合运算模型，但现有的二元加权泛组合运算模型是针对命题真值的线性加权或指数加权，虽然都能满足工程应用的需要，但线性加权不能满足泛组合运算应该在全局上取值的要求，指数加权不仅计算复杂且只能用于[0,1]标准区间[13]，因此本文提出了生成元加权零级泛组合运算模型，能满足泛组合运算在全局上取值的要求，且易于扩展到任意区间，扩大了适用范围。在实际组合决策应用时，可根据情况选择所使用的加权方式。

结束语　由于复杂系统中存在各种不确定性及相互关系，经典的二值逻辑对它来说显得太刚性，因此有研究者转而寻求“非标准逻辑”或“柔性逻辑”。模糊逻辑通过承认命题真值的连续可变性，打破了人们长期以来的“二值观”。而泛逻辑学则更进一步，它在承认命题真值连续可变性的基础上，分析命题之间关系的连续可变性，提出了“广义相关性”和“广义自相关性”两个重要的概念，将命题连接词运算模型定义为由相关性所控制的算子簇，实现了连接词运算模型的柔性化。其中泛组合命题连接词运算模型是为了满足连续值逻辑中综合决策的需求而提出的，仅有二元模型，在实际应用中受限，迫切需要多元模型。本文在原有的等权二元泛组合运算模型的基础上，修正了二元泛组合运算模型，提出了多元泛组合模型及生成元加权零级泛组合运算模型，并严密推导证明了它们的特殊算子。这些工作不仅完善了原有的泛逻辑命题连接词理论，而且为泛组合的应用提供了丰富的模型选择。

参 考 文 献

[1] Klement E P, Navara M. Propositional fuzzy logics based on Frank T-norms: A comparison[M]// Dubois D, Klement E P, Prade H, eds. Fuzzy Sets, Logics and Reasoning about Logics. Applied Logic Series 15. Dordrecht, Netherlands: Kluwer Academic Publishers, 1999: 17-38

[2] 吴望名. 参数 Kleene 系统中的广义重言式[J]. 模糊系统与数学, 2000, 14(1): 1-7

[3] 王国俊, 兰蓉. 系统 H_α 中的广义重言式理论[J]. 陕西师范大学学报: 自然科学版, 2003, 31(2): 1-11

[4] He H C, Wang H, Liu Y H, et al. Principle of Universal Logics [M]. Beijing: Science Press, 2005

[5] 张小红. 基于 T-模与伪 T-模的逻辑系统及其代数分析[D]. 西安: 西北工业大学, 2005

[6] 鲁斌, 何华灿. 泛模糊逻辑控制器研究[J]. 计算机工程与应用, 2003, 39(16): 13-16

[7] Mao M Y, Chen Z C, He H C. A New Uniform Neuron Model of Generalized Logic Operators Based on [a, b][J]. International Journal of Pattern Recognition and Artificial Intelligence, 2006, 20(2): 159-171

[8] Jin Y, He H C, Lü Y T. Ternary Optical Computer Principle [J]. Scince in China, Ser F, 2003, 46(2): 145-150

[9] 刘丽, 何华灿, 贾澎涛. 泛逻辑控制模型研究[J]. 计算机工程, 2007, 33(19): 7-9

[10] Zadeh L A. Fuzzy Sets[J]. Information and Control, 1965, 8(3): 338-353

[11] 应明生. 模糊逻辑的紧致性[J]. 科学通报, 1998, 43(4): 379-383

[12] 王万森, 何华灿. 基于泛逻辑学的逻辑关系柔性化研究[J]. 软件学报, 2005, 16(5): 754-760

[13] 贾澎涛. 基于柔性逻辑的时间序列数据挖掘研究[D]. 西安: 西北工业大学, 2008

[14] 刘丽. 基于柔性逻辑的智能控制研究[D]. 西安: 西北工业大学, 2007

[15] 付利华. 复杂系统的柔性逻辑控制理论及应用研究[D]. 西安: 西北工业大学, 2005

(上接第 168 页)

将导致永远分不到正确的类别，错误控制方法达到了预期的效果。

结束语　本文针对平面分类在类别庞大的情况下会出现分类器巨大、训练时间过长等问题，说明了层次分类的必要性，阐述了平面分类和层次分类的异同点，给出了平面分类和层次分类在分类性能上的比较，并提出了用分层特征权重计算和错误控制的方法对分类方法进行优化。实验证明层次分类方法比平面分类方法的分类效果要好，同时分层特征权重计算和错误控制对层次分类效果有显著的提高。

本文提出的层次分类方法是对现有的平面分类的一个改进。为了更好地提高层次分类效果，在特征重新计算和错误控制方面还需要进一步改进。

参 考 文 献

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer, 2000: 1-300

[2] Svmlight J T. An implementation of Support Vector Machines (SVMs) in C[EB/OL]. http://svmlight.joachims.org/

[3] Sun Aixin, Lim Ee-Peng. Hierarchical text classification and evaluation[C]// Proceedings of the 2001 International Conference on Data Mining. 2000: 521-528

[4] Ruiz M E, Srinivasan P. Hierarchical neural networks for text categorization[C]// Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99). 1999: 281-282

[5] Dumais S, Chen H. Hierarchical classification of Web content [C]// Proceedings of the 23rd ACM Int. Conf. on Research and Development in Information Retrieval. 2000: 256-263

[6] Dekel O, Keshet J, Singer Y. Large margin hierarchical classification[C]// Proceedings of the 21st ICML. 2004: 27-34

[7] Cai Lijuan. Hierarchical Document Categorization with Support Vector Machines[C]// CIKM04. 2004: 78-86

[8] Cesa-Bianchi N. Hierarchical Classification : Combining Bayes with SVM[C]// Proceedings of the 23rd ICML. 2006: 177-184

[9] Cheng C, Tang J, Fu A Wai-chee, et al. Hierarchical Classification of Documents with Error Control[C]// PAKDD. 2001: 433-443

[10] Susan G. Training a Hierarchical Classifier Using Interdocument Relationships[J]. Journal of the American Society for Information Science and Technology, 2009, 60(1): 47-58

[11] Koller D, Sahami M. Hierarchically classifying documents using very few words[C]// Proceedings of the Fourteenth International Conference on Machine Learning. 1997: 170-178

[12] Mladenic D, Grobelnik M. Feature Selection for classification based on text hierarchy[C]// Proceedings of the Workshop on Learning from Text and the Web. 1998

[13] 樊绮娜. 中文新闻领域的关键词抽取方法研究与实现[D]. 北京: 清华大学, 2006