面向短文本的动态组合分类算法

闫 瑞[1,2],曹先彬[1,2],李 凯[3]
(1. 中国科学技术大学计算机科学技术系,安徽合肥 230027;2. 安徽省计算与通讯软件重点实验室,安徽合肥 230027;
3. 解放军保密委员会技术安全研究所,北京 100091)

摘 要: 短文本分类是网络内容安全的一种主要方法. 然而,短文本固有的关键词特征稀疏和样本高度不均衡等特点,使得难以直接使用现有针对长文本的分类算法. 本文提出了一种针对短文本的动态组合分类算法. 首先构造出一种树状组合分类器结构,可有效缓解短文本特征稀疏和样本高度不均衡对分类性能的影响;进一步,提出了一种动态调整策略来训练组合分类器,可以根据样本的分布特点自适应地调整分类器的组合结构. 测试实验表明,相对于传统的单一分类方法和集成分类方法,动态组合分类算法在短文本分类中可以获得更好的准确率和召回率.

关键词: 短文本分类; 组合分类器; 动态调整策略; AdaBoost 算法
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2009) 05-1019-06

Dynamic Assembly Classification Algorithm for Short Text

YAN Rui[1,2], CAO Xian-bin[1,2], LI Kai[3]
(1. *Department of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui* 230027, *China*;
2. *Key Laboratory of Software in Computing and Communication, Hefei, Anhui* 230027, *China*;
3. *Institute of Technical Security, PLA, Beijing* 100091, *China*)

Abstract: Short text classification is a key technology in network content security application. However, the sparse features and unbalanced data of the short text make the traditional text classification method incompetent for short text classification. This paper proposed a dynamic assembly classification method for short text classification. In this method, a treelike assembly classifier was constructed to support the classification, which reduced the impact of the sparse features and unbalanced data of the short texts. Further, a dynamic adjusting strategy was presented in the construction procedure, which adjusted the combinational structure of the classifier in an adaptive way. The experimental results show that, comparing with the traditional classifiers such as single classifier and ensemble classifier, the proposed assembly classifier gets better precision rate and recall rate.

Key words: short text classification, assembly classifier, dynamic adjusting strategy, adaBoost.

1 引言

网络内容安全是近来网络安全研究的一个重要领域,主要研究基于网络内容(主要是文本内容)的安全性问题. 它涉及到针对文本内容的获取、分类、聚类、话题发现与跟踪等关键技术;其中,分类是实现网络内容安全的一个主要手段. 随着即时通信、聊天室、BBS、Email、博客等网络交流平台的不断涌现,针对这类应用背景下特定文本形式的分类方法研究已经成为相关领域的一个重点[1-2].

在这类应用背景中,待处理的文本在形式上已经发生了很大变化,一般称之为短文本. 短文本的固有特点对相应的分类算法设计提出了更高的要求,现有文本挖掘领域已取得较大成功的文本分类方法还难以直接引用;因此,针对短文本的有效分类算法已经成为目前亟待解决的一个研究难题.

短文本分类面临的难点主要有:

(1) 短文本关键词特征稀疏. 与一般成句子的长文本相比,短文本的关键词特征稀疏(每个短文本中一般只含有数十个甚至几个关键词),难以充分挖掘出特征之间的关联性.

(2) 样本高度不均衡. 短文本应用背景(如网络内容安全)需要处理海量的文本流,而其中真正关注的检测对象(如敏感话题、事件)在数量上只占很小的一部分. 目前,专门针对短文本分类的研究工作还较少. 在方法上,主要还是直接采用用于长文本的分类算法,包括基

收稿日期:2008-05-08; 修回日期: 2008-08-02
基金项目:国家 973 课题(No. 2004CB318109),国家 863 课题(No. 2007AA11Z240),教育部新世纪优秀人才支持计划(No. NCET-07-0787)

于单一分类器的方法如 KNN[3]、Bayes[4]、SVM[5,6]以及基于集成学习的方法(以 Boosting 为代表[7~10]). 由于短文本中关键词特征稀疏,且不同的特征对分类结果影响的程度差异很大,而这些用于长文本分类的单一分类器方法都依据词频特征的相似性来分类,因此直接用在短文本分类中很难取得很好的效果. 在另一方面,基于集成学习的 Boosting 方法近年来在长文本分类中得到了广泛关注[7~10]. 该方法针对每个特征设计一个弱分类器,然后线性组合多个弱分类器得到一个强分类器. 该方法在用于文本分类时,通过为每个弱分类器赋予一个合适权重,得到每个词频特征的权重,因而适合于解决短文本分类面临的特征稀疏问题. 但是,这一方法还无法处理短文本分类面临的样本高度不均衡问题;因为在解决数据不均衡的分类问题时,集成分类器容易把大量的其它文本错分为关注的检测对象. 因此,基于集成分类器的方法还需要进一步改进才能更好地解决短文本分类问题.

为此,本文提出了一种针对短文本的动态组合分类算法. 该算法首先构造出一种树状组合分类器结构来支持分类,并进一步提出了一种动态调整策略来训练组合分类器. 与已有的基于单一分类器或简单集成分类器的方法相比,该方法可以根据样本的分布特点自适应地调整分类器的组合结构,从而有效缓解短文本特征稀疏和样本高度不均衡对分类性能的影响. 基于中文聊天室环境的测试实验表明,该算法在用于短文本分类时,可以获得较好的准确率和召回率.

2 相关工作

短文本分类算法与传统用于长文本的分类算法有共通之处,但也由于短文本的自身特点而需要进行针对性研究和设计. 目前,分类方法一般可以分为基于单一分类器的和基于集成的两种.

基于单一分类器的分类算法比较普遍,常用的有 KNN[3],Bayes[4]、SVM[5,6]. 这类方法的基本思想是首先分析训练数据集的特点,为每个类别产生一个相应数据集的准确描述或模型;然后利用类别的描述或模型对测试数据集进行分类. 这一类方法在传统的长文本分类中应用得非常广泛,并取得了较大的成功[3~6],而在短文本分类应用方面的工作还相对较少,典型工作如:Eiman M. Elnahrawy 将 KNN、Bayes 和 SVM 三种分类方法用于聊天室背景下的短文本分类,并对三种方法的性能进行了比较分析[11]. 得出的结论是:KNN 方法训练和分类速度最快,但是分类效果一般;Bayes 方法速度接近 KNN 方法并且分类效果最好;而 SVM 方法速度最慢,效果介于 KNN 和 Bayes 之间. Sarah Zelikovitz 提出了一种针对短文本分类的改进算法,针对训练样本数量较少的难点,提出引入背景知识的方法来提高分类效率,即为每一个类别引入大量相关的长文本数据进行训练[12];由于加入了大量的文本数据进行分类,因此该方法的训练时间较长. 总之,基于单一分类器的分类算法在用于短文本的分类时,由于关键词特征稀疏,难以为每个类别得到一个数据集的准确描述,因此效果不够理想.

基于集成学习的组合分类器是近年来随着机器学习领域的研究进展引入到文本分类中的一种新方法. 它的基本思想是通过多个弱分类器的综合评判,来提高分类准确率. 目前,这一方法在长文本分类中已取得了一定成功[7~10]. 鉴于其在长文本分类中的良好效果,将之引入到短文本分类中是自然的;然而,相对于长文本,短文本的样本数据不均衡性程度更为严重,需要关注的检测对象相对其它样本在数量上只占很小的部分,组合分类器容易把大量的其它文本错分为关注的检测对象;因此,还需要针对短文本样本高度不均衡特点,研究合适的组合分类器结构生成方法,以便得到结构更为合理的组合分类器来支持短文本的有效分类.

3 面向短文本的动态组合分类算法

3.1 树状组合分类器的结构与分类过程

我们提出的组合分类器如图 1 所示,它是一种树状结构,可以看作是几个子组合分类器和单分类器的组合,树中每一个节点都是一个分类器. 考虑到在实际构造过程中,得到的组合分类器可能是森林结构,包含多棵树,因此我们在训练开始时,加入一个特殊的根节点分类器(对所有的输入样本都判断为正),训练得到的组合分类器作为它的孩子,这样可以保证最终得到的组合分类器在结构上仍然是一颗树.

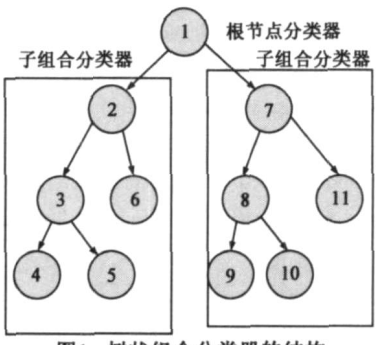


图1 树状组合分类器的结构

为表达方便,从下文开始,我们称树中每个节点所表示的分类器叫做节点分类器.

假设短文本空间共有 N 个类别,我们为每个类别训练一个组合分类器 $C_i = 1, 2, \cdots, N$,得到多个组合分类器. 基于这种组合分类器,对短文本的分类过程如图 2 所示. 其中,待分类短文本 st_i 的分类过程为:采用自顶向下的方法,如果 st_i 被组合分类器一个节点分类器判断为 True,则此节点分类器的所有子节点继续判断,否则其子节点拒绝继续判断,并返回 False;若存在一条

从根到叶子节点的路径,且路径上所有节点分类器都判断此短文本为 True,则短文本 st_i 属于类别 c_i,否则 st_i 不属于类别 c_i.

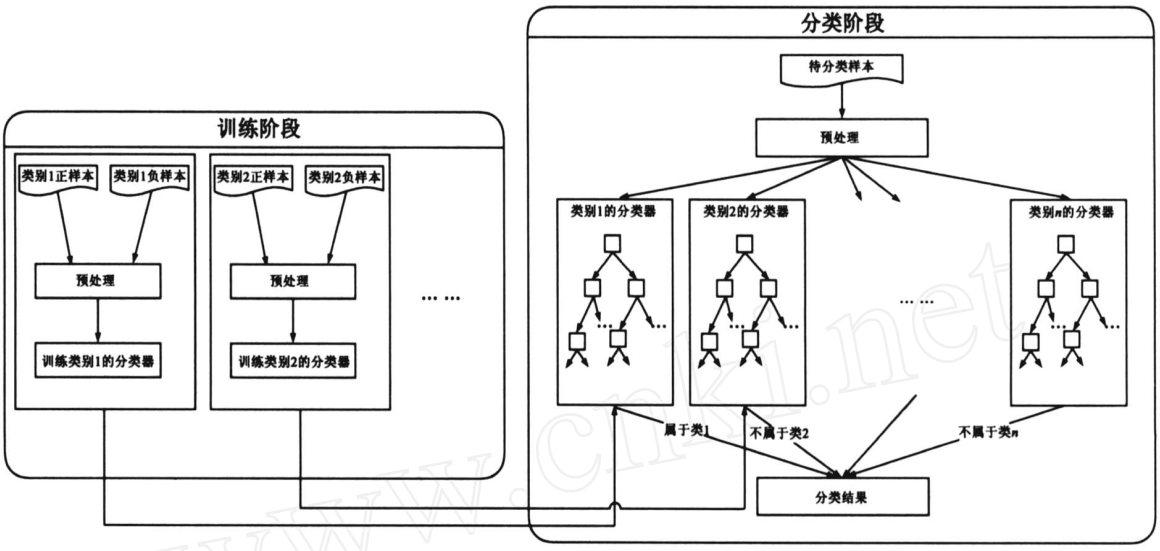


图2　树状组合分类器的训练及分类过程

3.2　树状组合分类器的动态自适应训练

图 2 同时给出了树状组合分类器的基本的训练过程,主要包括两个部分:节点分类器的训练和组合分类器的训练. 在本文算法中,每训练完毕一个节点分类器,就被加入到组合分类器中;而每个节点分类器的功能又取决于其在组合分类器中的位置,故节点分类器的训练和组合分类器的构造是在同一个过程中完成的.

在上述训练过程中,我们规定每个节点分类器都用 AdaBoost 方法训练得到. 至此,我们面临的问题就是:如何通过训练得到适合短文本分类的树状组合分类器的结构. 这一问题的解决涉及到以下几个关键技术:(1) 根据短文本样本的分布特点,如何构造树状组合分类器的结构自适应调整策略;(2) 如何为一个节点分类器选择合适的训练样本;(3) 在训练过程中,若发现当前待分问题比较复杂时,需要将之划分为两个子分类问题来解决. 这时如何设计合理的样本分割策略,下面分别介绍.

3.2.1　树状组合分类器的结构自适应调整策略

树状组合分类器在训练时,要求能够根据样本的分布特点,自适应地调整树的分支数和层数,以提高分类的准确率;因此,树状组合分类器的结构不是事先确定的,而是在训练的过程中来自动地确定. 首先给出一个定义:

定义 1　训练难度(The difficulty of training) 定义为 $\text{trian\_error\_rate} \times \ln(n)$,反映一个分类器的训练难易程度. 其中,trian_error_rate 为 AdaBoost 分类器测试训练样本的错误率,n 为 AdaBoost 分类器选择的节点分类器的个数.

树状组合分类器的结构自适应调整策略遵循如下准则,实际上也就是具体的训练过程:

(1) 树状组合分类器在刚开始训练时,只含有一个无识别能力(对任意样本都判断为 True) 的根节点分类器. 树只有一个根节点,在训练中通过不断给叶子节点增加子节点的方法来更新树.

(2) 树上的每条从根节点到叶子节点的路径上的所有节点分类器组成一个级联分类器,当它对测试数据的准确率和召回率满足要求时,我们就停止在此叶子节点后添加新节点的操作,并称此节点为 stopExpandNode.

(3) 每次训练从选中的一个非 stopExpandNode 叶子节点 N 开始,首先确定它在树状组合分类器中的位置,然后根据它的位置,确定训练它的样本集合.

(4) 使用 AdaBoost 方法训练选定的叶子节点 N 分类器,计算得到这个分类器的训练难度. 如果训练难度较小,把刚训练得到的分类器作为节点 N 的唯一子节点加入树中;否则抛弃这个分类器,把该训练样本集合分割成两个训练样本子集,用每个训练样本子集单独训练得到一个分类器,作为节点 N 的两个子节点加入到树中.

(5) 对树进行测试,判断新加入的节点是否是 stopExpandNode. 当树中的所有叶子节点都是 stopExpandNode 时,我们就构造好了最终的树状组合分类器.

上述过程可用伪代码描述如下:

算法:结构自适应调整策略
输入:
Samples:训练样本 $S=\{$正样本,负样本$\}$
θ:训练难度阈值
$f(c_i)$:训练难度函数
$g(s_i)$:样本分割策略函数
输出:一个树状组合分类器 T
方法:
(1) 初始化树状组合分类器 T 的根节点分类器 N,该节点对任意样本都判断为 True
(2) while 存在可扩展叶子节点 i
(3) 构建训练分类器 C_i 的样本集合 $S_i=\{$正样本,负样本$\}$
(4) 采用 AdaBoost 算法训练分类器 C_i
(5) if $f(c_i)<\theta$ then
(6) 将分类器 c_i 作为 i 的子节点插入到树 T 中
(7) else
(8) 割样本:$g(S_i)\rightarrow\{s_1,s_2,\cdots,s_n\}$,为每个子样本训练分类器,得到$(c_1),c_2,\cdots,c_n$,并将 $c_1,c_2,\cdots,c_n$ 作为 i 的子节点插入到树 T 中
(9) 返回树 T

3.2.2 节点分类器训练样本的选择

分类器的功能是由训练样本决定的,同时树状组合分类器中的每个节点分类器的功能取决于它在树中的位置;因此,每个节点分类器的训练样本需要特别的处理.

对应上述的树状组合分类器训练过程,假设待训练的节点分类器是 c,它的父节点是 N,那么 c 使用的正样本集合与训练 N 时使用的正样本集合相同,记为 P,以保证不降低(或降低很小)分类的准确率.

得到负样本集合则相对复杂一些. 在当前生成的树状结构中,可以唯一地确定一条从根节点到节点 N 的路径. 可以把这条路径上的所有节点分类器组合看作是一个串联的级联分类器. 使用这个级联分类器对总的负样本集合测试,选择$|P|$个被错误判断的负样本作为待训练分类 C 器的负样本集合.

3.2.3 样本分割策略的设计

在训练树状组合分类器过程中,当发现当前的分类问题比较复杂时(即针对它训练得到的节点分类器的训练难度大于给定阈值),我们把这个分类问题分为两个独立的子分类问题,每个子问题分别设计一个节点分类器来加以解决. 上述分类问题的划分取决于能否把训练样本分为两个相对独立的部分,每部分的样本含有相同或相似的特征,因此需要设计出合理的样本分割策略.

本文采用了一种简单易行的样本分割策略,描述如下:

(1) 把刚训练好的节点分类器的所有弱分类器随机地分成两个子集合 $s1,s2$,且满足 $s1\cap s2=\phi$,$|s1|=|s2|$;

(2) 对正样本集合中的每个样本,如果 $s1$ 中判断结果为 True 的弱分类器数目比 $s2$ 中判断结果为 True 的弱分类器数目多,则把此样本归为第一类,否则归为第二类.

因为每个弱分类器是根据一个特征训练得到的,所以若一个弱分类器对一个样本判断为 True,则表示此样本具有相应的特征. 这样,第一类样本就含有较多的 $s1$ 中弱分类器所含有的特征,而第二类样本含有较多的 $s2$ 中弱分类器所含有的特征. 所以就满足了把具有相同或相似的特征样本集合作为一类的要求.

3.3 树状组合分类器的特点

在机器学习中,分类器组合的方式主要有并联[7~10]和串联[13]两种. Boosting 算法从本质上说就是一个并联结构的组合分类器,它可以提高分类的准确率;但是并联结构需要综合多个节点分类器的结果,这必然会降低分类速度,而且它也不能有效解决样本不均衡问题. 另一方面,串联结构的组合分类器将多个单分类器从上到下链状排列,一个短文本只有通过上一个分类器的"认可"才能被下一个分类器分类,当且仅当一个短文本被所有节点分类器"认可"才被确认为某一类;这种方法可以减少误分,并且速度也较快. 但其召回率将随层数的增加而显著下降.

本文提出的树状组合分类器兼取了并联和串联结构的优点. 它能根据分类问题的难度,自适应地决定采用并联或串联结构来组织节点分类器,以达到最优的分类精度. 与已有的单一分类器、并联或串联组合分类器比较,它具有如下特点:

(1) 它的层次化结构、以及逐步求精原理使得它可以有效解决样本不均衡问题.

(2) 它能够根据分类问题的难度,把一个分类问题分为两个子问题,每个子问题分别使用一个单一分类器来解决;这样就降低了分类难度,克服了串联组合分类器召回率低的缺点.

(3) 它遵照"早拒绝"的原则,即一个待分类对象只有被上层的分类器判为 True,才会被下层的分类判断,从而保证了分类的速度和准确率.

(4) 它具有较好的通用性. 当分类任务不存在样本不均衡问题时,它可退化为一个能根据分类问题难度而自动确定单一分类器个数的并联组合分类器;当分类任务并不复杂到非要分为多个子问题时,它可退化为一个串联组合分类器;当分类问题并不太复杂且不存在样本不均衡问题时,它就是一个单一分类器.

4　实验与分析

为验证本文提出的动态组合分类器的性能,我们将该分类算法与基于单一分类器的 SVM 算法和基于集成学习的 AdaBoost 算法进行对比分析.

4.1　实验数据

由于目前短文本数据尚无国际公认的标准测试语料库,我们从 www. xiaonei. com 的群组聊天室中取出中文聊天数据作为数据源进行实验. 实验中将一个简短聊天数据片段(即一个帖子及其回复)作为一个短文本处理. 我们准备了 1384 篇短文本共四个话题领域的数据,表 1 给出了这一数据集的统计信息.

表 1　实验数据

话题类别	短文本数目
武侠	467
NBA	381
武器	248
游戏	292

4.2　评估指标

我们使用准确率(*precision*)、召回率(*recall*)和测试值作为分类器的评价标准,其中,准确率为分类正确文本数与总文本数的比率,即:$precision = \frac{\text{分类正确的评文本数}}{\text{总文本数}}$;召回率为采用分类正确文本数与分类应有的文本数的比率,即:$recall = \frac{\text{分类正确的文本数}}{\text{分类应有的文本数}}$;$F_1$ 测试值综合考虑了准确率和召回率,其数学公式如下:

$$F_1\ \text{测试值} = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}}.$$

4.3　实验结果与分析

实验 1:与常用的 SVM、AdaBoost 方法的比较

本实验中,我们分别使用了 *SVM* 方法,AdaBoost 方法和本文提出的树状组合分类器方法对语料进行测试,分类结果如图 3 所示:

从分类结果可以看出,与 SVM 方法和 AdaBoost 方法相比,树状组合分类器方法对短文本分类的效果较好. 由于基于单一分类器的 SVM 方法在很大程度上依赖文本的词频和关键词的相关性进行分类;而一般的聊天数据(短文本数据)都具有关键词稀疏和关联度低的特点,对一篇短文本数据进行分词后并去掉停词后,通常只剩下数十个甚至几个关键词. 因此,在对整个短文本空间进行词频统计和权重计算时,很难为每个关键词赋予一个合适的权重使其对短文本分类都具有较好的结果. 而本文提出的动态组合分类算法不依赖于词频统计,而是根据短文本空间关键词的本身特征,动态地构造了一棵描述该特征的组合分类器. 由于这一组合分类器具有将弱分类器提升为强分类器的天然特性,因此该组合分类器比单一分类器具有更好的分类性能.

AdaBoost 算法虽然考虑了短文本关键词稀疏及关联度低的特性,采用集成学习的方法训练得到符合短文本空间特征的分类器,但是 AdaBoost 对每个待分类的样本均采用同一个分类器的参数描述进行分类(即分

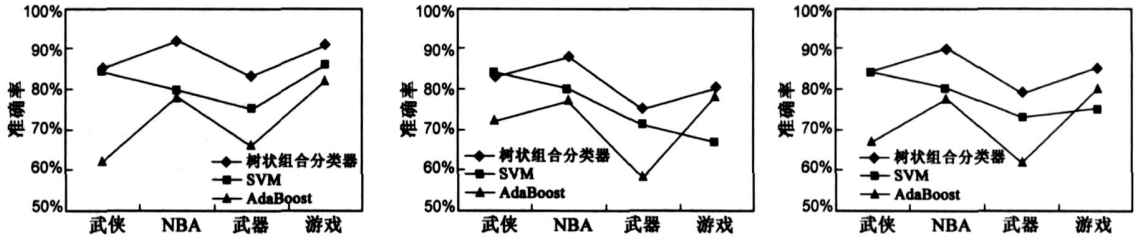


图3　三种分类器的性能比较

类器只有一条搜索路径. 若该路径对待检测短文本 st_i 的检测结果为 False,那么 st_i 就不属于该分类器对应的文本类别),因而没有考虑到短文本样本的高度不均衡性. 而本文提出的动态组合分类算法是一个树状结构的组合分类器,从根节点到叶子节点的每条路径都对待检测短文本 st_i 进行检测,只要有一条路径对 st_i 检测为 True,则 st_i 就属于该分类器对应的类别;反之,只有当所有的路径对 st_i 的检测都为 False 时,st_i 才不属于这个分类器对应的类别. 在理想情况下(即树的结构为一棵满 N 叉树),该分类器具有 N^{k-1}(k 为树高)个参数描述,可以很大程度上缓解由于短文本样本高度不均衡带来的分类困难,从而提高了分类器的召回率. 从实验结果中我们可以看出,本文提出的动态组合分类算法的召回率高于 SVM 方法和 AdaBosot 方法.

实验 2:关键参数对分类器性能的影响

在本文提出的分类算法中,训练难度阈值 θ 是一个很重要的参数,θ 的取值将决定分类器的分类性能. 为了获得 θ 的大小影响分类器性能的规律,我们分别对该参数进行了多组实验,实验结果如图 4 所示. 从图中我们可以看到,当 θ 取值范围在 1.3 - 1.6 时,都取得了较好的分类效果;其中 θ 在取值为 1.5 时,分类器具有最好的性能.

由于训练难度阈值 θ 的大小决定了训练样本集合的划分,因此,若 θ 取值过小,分类器则会将训练样本集合分割成大量的小样本集合进行训练,分类器会过度拟和样本数据,从而导致分类准确率下降;若 θ 取值

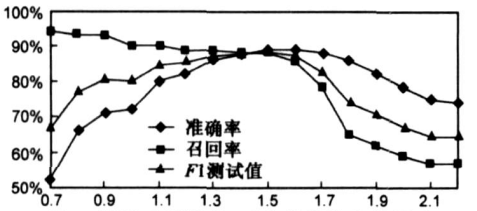


图4 训练难度阈值θ对分类器性能的影响

过大,最终生成的组合分类器器则退化为串行级联分类器,同样导致分类准确率下降.

5 结束语

本文提出了一种针对短文本的动态组合分类算法.该算法首先构造出一种树状组合分类器结构来支持分类,并进一步提出了一种动态调整策略来训练组合分类器.与已有的基于单一分类器或简单集成分类器的方法相比,该方法可以根据样本的分布特点自适应地调整分类器的组合结构,基于聊天室环境的测试实验表明,该算法在用于短文本分类时,可以有效缓解短文本特征稀疏和样本高度不均衡对分类性能的影响,获得了较好的准确率和召回率.

下一步工作包括:(1)在目前的算法中,对样本的分割采用的还是基于随机策略的方法,这不足以将待分割的样本分割成两个最优的划分.接下来可以进一步研究使用聚类的方法发现待分割的短文本中隐含的规律,以获得更高的准确率和召回率.(2)现在的实验数据规模还比较小,还需要在更大规模的真实数据环境下进行进一步验证和完善(特别是在大规模数据分类时的时间花费需要满足实用化要求).(3)这种分类器组合结构的数学基础也将是一个有意义的研究任务.

参考文献:

[1] Bengel J, Gauch S, Mittur E, Vijayaraghavan R. Chattrack: Chat room topic detection using classification [A]. Proceedings of the 2nd Symposium on Intelligence and Security Informatics [C]. Berlin Germany: Springer-Verlag, 2004. 266 - 277.

[2] Haichao Dong, Siu Cheung Hui, Yulan He. Structural analysis of chat messages for topic detection [J]. Online Information Review, 2006, 5(30): 496 - 516.

[3] Yang Y, Liu X. A re-examination of text categorization methods [A]. Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York, USA: ACM, 1999. 42 - 49.

[4] Fried N, Geiger D, Goldszmidt M. Bayesian network classifiers [J]. Machine Learning, 1997, 29 (2-3): 131 - 163.

[5] Vapnic V. The nature of statistical learning theory [M]. New York: Springer, 1995. 138 - 170.

[6] Joachims T. Text categorization with support vector machines: Learning with many relevant features [A]. Proceedings of the 10th European Conference on Machine Learning [C]. Berlin Germany: Springer-Verlag, 1998. 137 - 142.

[7] Schapire R E, Singer Y. Boostexter: A boosting-based system for text categorization [J]. Machine Learning, 2000. 39(2-3): 135 - 168.

[8] Kim Y H, Hahn S Y, Zhang B T. Text filtering by boosting naive bayes classifiers [A]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York, USA: ACM, 2000. 168 - 175.

[9] James G Shanahan. Boosting support vector machines for text classification through parameter-free threshold relaxation [A]. Proceedings of the 12th international conference on Information and knowledge management [C]. New York, USA: ACM, 2003. 247 - 254.

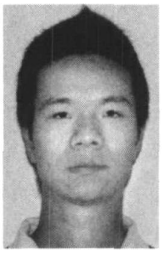
[10] Bloehdorn S, Hotho A. Text classification by boosting weak learners based on terms and concepts [A]. International Conference on Data Mining, Brighton [C]. UK: IEEE Pness, 2004. 331 - 334.

[11] Elnahrawy E M. Log-based chat room monitoring using text categorization: A comparative study [A]. Proceedings of The IASTED International Conference on Information and Knowledge Sharing [C]. St. Thomas, US Virgin Islands: acta press, 2002. 111 - 115.

[12] Sarah Zelikovitz. Improving short-text classification using unlabeled data for classification problems [A]. Proceedings of the Seventeenth International Conference on Machine Learning [C]. San Francisco, USA: Morgan Kaufmann Publishers Inc, 2000. 1191 - 1198.

[13] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features [A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition [C]. Kauai Marriott, Hawaii: Proc. of CVPR, 2001. 511 - 518.

作者简介:



闫 瑞 男,1984 年 1 月生于安徽省亳州市.现为中国科学技术大学计算机系硕士研究生.主要从事信息安全的研究.

E-mail: yanrui06@mail.ustc.edu.cn

曹先彬 男,1969 年 1 月生于安徽省巢湖市.博士、教授、博士生导师.主要从事计算智能、信息安全、智能交通系统等的研究.

E-mail: xbcao@ustc.edu.cn