基于特征类别属性分析的文本分类器分类噪声裁剪方法

王 强[1] 关 毅[1] 王晓龙[1]

摘 要 提出一种应用文本特征的类别属性进行文本分类过程中的类别噪声裁剪 (Eliminating class noise, ECN) 的算法. 算法通过分析文本关键特征中蕴含的类别指示信息, 主动预测待分类文本可能归属的类别集, 从而减少参与决策的分类器数目, 降低分类延迟, 提高分类精度. 在中、英文测试语料上的实验表明, 该算法的 F 值分别达到 0.76 与 0.93, 而且分类器运行效率也有明显提升, 整体性能较好. 进一步的实验表明, 此算法的扩展性能较好, 结合一定的反馈学习策略, 分类性能可进一步提高, 其 F 值可达到 0.806 与 0.943.

关键词 类别属性分析, 类别噪声裁剪, 文本分类
中图分类号 TP391

A Method for Eliminating Class Noise in Text Classification Based on Feature Class Attribute

WANG Qiang[1] GUAN Yi[1] WANG Xiao-Long[1]

Abstract This paper presents a novel algorithm for eliminating class noise based on the analysis of the feature class attribute in text classification. The algorithm can eliminate class noise for classifier by mining the most representative class information of text features, which means that the algorithm can actively prejudge the candidate class labels to unseen documents using the class attribute linked to features and classify them in the candidate class spaces to reduce the number of decisions, retrench time expense, and promote accuracy. The experimental results on Chinese and English corpus show that the algorithm has good performance. The F measure is 0.76 and 0.93, respectively, and the run efficiency of classifier has been improved greatly. A further experiment indicates that the algorithm has good expansibility. Based on a certain feedback learning strategy, the F measure can be further improved to 0.806 and 0.943.

Key words Class attribute analysis, eliminating class noise, text classification

1 引言

文本分类是指按照文本的内容和预先定义的主题类别, 为文本集中的每个文本确定所属主题类别的技术. 由于人工进行文本分类的一致性和正确性很难保证, 而且耗费很大, 因此, 为了满足海量信息处理的需要, 基于统计理论和机器学习的分类方法成为了自动文本分类的主流技术, 典型算法有贝叶斯分类 (Bayes)[1, 2]、k 近邻分类 (kNN)[3, 4]、决策树分类 (Decision tree)[5, 6] 及支持向量机分类 (SVM)[7, 8] 等. 这些算法都采用了归纳学习 (Inductive learning) 的策略, 即使用训练样本, 通过归纳推理将其推广, 并产生一个或一组一般性的概念描述来指导新样本求解. 这种学习策略对样本质量的依赖性较大, 往往会因为训练集中的噪声样本、噪声属性或分类过程中的噪声类别等原因造成分类性能下降.

目前, 解决训练集中的噪声样本、噪声属性对分类算法的影响, 通常采用文档选择[9~11] 与特征选择算法[12, 13]. 文档选择算法选取原来的训练样本集中的一些具有较强代表性的样本作为新的训练样本, 从而达到减少训练噪声样本的目的; 特征选择算法是去除在文本中不能表示信息或表示信息较弱的特征. 本文主要针对分类过程中的类别噪声问题展开研究. 降低这类噪声的基本策略是引入层次分类模型 (Hierarchical text classification)[14, 15], 即通过文档类别层次结构树, 建立分类树的内部节点分类器, 将大类别数的分类问题逐层简化为一个个局部分类问题以提高分类的精度. 但是层次分类算法会受到类别层次树拓扑结构的影响[16], 其构造过程依赖于分类专家的领域知识, 缺少系统化的算法指导, 同时其中间节点的分类精度对系统结果影响较大, 需要复杂的评价与修正机制[17], 这会导致训练与分类过程相对繁琐.

收稿日期 2006-4-24 收修改稿日期 2006-12-18
Received April 24, 2006; in revised form December 18, 2006
国家自然科学基金 (60435020, 60504021) 资助
Supported by National Natural Science Foundation of China (60435020,60504021)
1. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001
1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001
DOI: 10.1360/aas-007-0809

目前应用层次分类方法进行文本分类的研究大多集中于采用系统方法构造分类树及对分类树的中间结果进行评价与修正等方面 (如文献 [15~17]). 而本文则是从分析文本特征的类别属性出发, 提出了一种类别噪声裁剪算法, 其核心是一种动态的层次分类思想. 这种动态层次分类思想, 与文献 [14] 中提到的可动态扩展分类模型, 即动态建立新类别节点的动态性含义不同, 它是通过挖掘文本特征中蕴含的类别属性信息, 利用主动类别预取技术, 动态构造出一个虚拟的分类中间节点, 将候选类别之外的其它路径上的分类器进行剪枝处理, 从而将全部分类模型上的测试文本分类问题转化为一个局部的分类问题, 不再受层次分类中的类别拓扑结构影响. 实验证明, 特征的类别属性分析配以一定的错误检测与修正技术, 能够有效控制分类过程中的类别噪声, 在保证分类精度的前提下减少分类延时, 提高文本分类的服务质量.

2 类别噪声对分类结果的影响

通常的文本分类体系, 将所有的类别放在同一层次上, 即处于同一平面类空间, 在执行分类操作时将测试文本和全部的类别模型进行比较, 择优为文本分配合适的类别, 在这种通过穷举所有的分类模型获得分类结果的方法中, 类别噪声对分类结果的影响主要体现在以下两个方面:

1) 在分类过程中, 除与测试文本真正相关的类别分类器之外, 其它分类器输出的实际上只是一种影响分类结果的噪声信息, 可能会造成测试文本类别的误判. 这种由于分类器噪声所导致的分类错误单纯通过调整分类器很难纠正. 而且在文本分类的实际应用过程中, 往往会遇到类别划分比较细, 类别的数目比较多, 类别间的模糊度比较大的情况, 此时这种噪声信息对分类性能的影响就更为严重.

2) 平面分类方法中的分类响应时间与参与分类决策的模型数目呈线性关系, 当类别数目比较多时, 分类器处理的时间、空间耗费就会随之增大, 这就降低了文本分类的使用效率. 无论是单分类系统 (分类器只输出一个最相关类别) 还是多分类系统 (分类器可以输出多个相关类别), 与测试文本真正相关的类别数目都是有限的, 大量的处理器时间与内存空间实际都花费在了处理噪声类别的信息上, 导致分类器的使用代价上升, 响应时间下降.

本文提出了一种基于文本特征的类别属性进行分类器类别噪声裁剪的方法, 此方法充分挖掘文本关键特征中蕴含的类别信息对分类的指示作用, 较好地解决了以上两个问题.

3 基于文本特征类别属性分析的类别噪声裁剪算法

3.1 文本特征的类别属性分析

为了对文本特征的类别属性进行定量分析, 实现对分类过程中类别噪声的裁剪, 本文引入了以下一些概念. 给定 n 维向量 $\boldsymbol{X} = (w_1, w_2, \cdots, w_n)$ 为一个文本, $c_j = (X_1^j, X_2^j, \cdots, X_{n_i}^j)(j = 1, 2, \cdots, m)$ 为包含 n_i 个文本、具有类别标识 c_j 的文本集, 定义:

定义 1. 设特征 w_i 在 c_j 中出现的词频数为 T_{ij}, 文档频数为 d_{ij}, 则 w_i 对 c_j 贡献率 Sw_{ij} 为

$$Sw_{ij} = \frac{fw_{ij} \cdot \log(dw_{ij} + 1.0)}{\sqrt{\sum_{i=1}^{n} [fw_{ij} \cdot \log(dw_{ij} + 1.0)]^2}} \tag{1}$$

其中: $fw_{ij} = T_{ij}/L_i$, L_i 是 w_i 出现在所有类别中的总频数; $dw_{ij} = d_{ij}/D_i$, D_i 是出现 w_i 的总文本数.

定义 2. 给定特征选取的阈值 $MinRatio$ ($MinRatio$ >0), 对于任意的特征 w_i, 定义其特征评分 $\mathrm{Imp}(w_i)$ 如下

$$\mathrm{Imp}(w_i) = \sqrt{\frac{\sqrt{\sum_j (Sw_{ij} - \bar{S}_i)^2}}{\sum_j Sw_{ij}}} \tag{2}$$

如果 $\mathrm{Imp}(w_i) > MinRatio$, 则称 w_i 为重要特征, 予以保留; 否则为噪声特征, 予以滤掉.

式 (2) 中, $\bar{S}_i = \sum_j Sw_{ij}/m$. $\mathrm{Imp}(w_i)$ 通过计算 Sw_{ij} 的方差来评价 w_i 的重要性, $\mathrm{Imp}(w_i)$ 越大, 表明 w_i 在不同类别之间的代表性差异越大, 其分类价值也就越大.

定义 3. 设 w_i 为重要特征, 给定类别代表性的阈值 $LowRatio$ ($LowRatio$ >0), 定义满足以下条件的类别集合 $Cl_i = \{c_j | Sw_{ij} \geq LowRatio, j = 1, 2, \cdots, m\}$ 为特征的代表类别集合.

定义 1 与定义 2 为衡量特征的重要性提供了一个标准, 这样就可以根据特征的重要程度, 筛选出对分类具有重要价值的特征数据, 过滤掉噪声特征. 同时, 重要特征 w_i 在不同类别中的贡献率是不同的, 这就意味着它在不同类别中对分类所起的作用也是不同的, 特征的类别贡献率越大, 它对这个类别进行分类时的指示意义也就越强, 定义 3 提供了根据特征的类别贡献率选择特征的代表类别集合的方法. 通过定义 1~3, 可以实现在特征选择的过程中对特征类别属性的挖掘.

3.2 类别噪声裁剪方法

本节讨论在文本特征的类别属性分析的基础上,

对分类器类别噪声进行裁剪的方法, 以解决在第 2 节中提出的两个问题. 具体的解决过程可描述为: 对于任意的待测文本 $\boldsymbol{X}_i$, 设 CL_i 是出现在 $\boldsymbol{X}_i$ 文档关键句中的文本特征构成的类别代表集合的并集, 被称为候选类别集合. 如果类别 $c_j \in CL_i$, 则保留 c_j 并调用 c_j 分类器进行类别验证; 如果类别 $c_j \notin CL_i$, 则认为类别 c_j 与测试文本无关, 可直接裁剪掉 c_j.

这种方法可以选择出现在文本中的关键句子、文本摘要甚至是标题中的特征作为文本的关键特征, 并以此来构造候选类别集合. 一种比较简单的策略是选择文本标题中的关键特征, 这主要是考虑到文本标题对文本类别主题具有预示作用[18]. 因为标题是作者根据所期望的突出主位的要求安排信息的结果, 它引导读者以某种特定的方式解读文章, 左右读者对整篇文章的理解. 同时标题和正文又是一个有机的整体, 标题统领整篇文章, 是正文内容的最高概括和抽象, 正文内容是对标题的解构与阐释.

从上述的方法可以看出, 通过为每个测试文本构造一个候选类别集合, 可以初步判定未出现在候选类别集合中的类别为与测试文本无关的类别, 将其直接裁剪掉. 其实质是一种动态层次分类的思想, 如图 1 所示, 即在平面分类方法的基础上, 利用候选类别集合动态构造出一个虚拟的分类中间节点, 将候选类别之外的其它路径上的分类器进行剪枝处理, 从而将全部分类模型上的测试文本分类问题转化为一个局部的分类问题. 同时, 此方法中的中间节点是虚拟的、动态生成的, 不受层次分类中的类别拓扑结构影响.

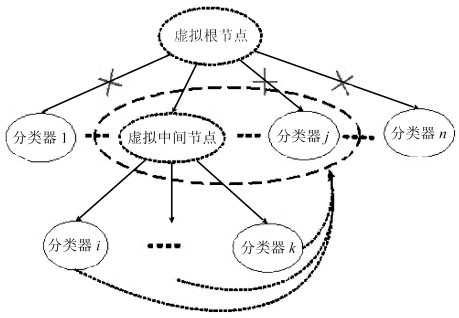

图 1　类别噪声裁剪示意图

Fig. 1　The illustration of eliminating class noise

但是虚拟中间节点同样可能会出现错误的判断, 导致后续的分类器无法输出正确结果, 因此有必要引入中间节点的分类错误检测与修正机制. 有两种方案可供选择: ROH (Recovery oriented error handling) 和 Error masking[17]. ROH 方法是在层次分类过程中, 记录前驱节点中具有高分类置信度的内部节点 HAC (High confidence ancestor), 当系统检测到文本是在一条错误的路径中被处理时, 可从 HAC 点重新选择另一条路径执行分类操作. Error masking 方法则是采用一种竞争投票机制, 利用多个分类器在层次分类路径上执行分类, 选择被多数分类器认可的、优势大的路径作为正确的分类路径输出. 这种算法需要若干个具有一定差异性的分类器, 并且在分类过程中执行多次分类, 分类代价较高. 在此借鉴 ROH 方法, 引入分类可信度判别机制. 但与 ROH 方法不同, 本文所实现的动态层次分类是一种二层的层次分类, 其根节点 (唯一的 HAC 节点) 是虚拟的, 无法计算其概率, 因此实现策略调整为: 为每个分类器定义决策判别阈值 $Th(c_j)$, 只有当分类器输出的分类值大于其相应的类别阈值时, 才认为其分类输出是合理的, 否则就拒绝其输出的结果. 如果所有虚拟中间节点指定的分类器输出都被拒绝, 则需要重新构造虚拟节点, 其构造方法是把出现在测试文本中的所有特征的代表类别集合的并集作为新的候选类别集合, 以新虚拟节点指定的分类器输出作为最终的分类结果. 如图 1 中的虚线所示. $Th(c_j)$ 设置策略见 3.4 节.

这种基于特征的类别属性进行类别噪声裁剪的方法, 由于在平面分类体系中利用中间虚拟节点进行了分类剪枝以去除类别噪声, 可以在保证分类准确率的前提下大幅度降低单一测试文本的分类器调用次数, 使分类系统整体的响应时间得到提升, 而且分类类别的数目越多, 这种分类器效率的提高就越显著.

3.3　文本分类器类别噪声裁剪算法

输入: 测试文本 $\boldsymbol{X}_i$
输出: 测试文本的主题类别集合 T
功能: 应用 ECN (Eliminating class noise) 算法为测试文本分配主题类别

1. $T = \phi$; //存储输出类别的集合
2. $VirtualNode = \phi$; //标题特征构成的候选类别集合
3. $ReVirtualNode = \phi$; //所有特征构成的候选类别集合
4. FOR EACH $w_i \in \boldsymbol{X}_i$
5.　IF w_i appears in Title
6.　　$VirtualNode = VirtualNode \cup Cl_i$
7.　ELSE
8.　　$ReVirtualNode = ReVirtualNode \cup Cl_i$
9. END;
10. FOR EACH $t \in VirtualNode$
11.　IF $Classifier_t(\boldsymbol{X}_i) > Th(c_t)$
12.　　Insert t into T
13. END;
14. IF $T = \phi$
15.　FOR EACH $t \in ReVirtualNode$
16.　　IF $Classifier_t(\boldsymbol{X}_i) > Th(c_t)$

17. Insert t into T

18. END;

19. Output T; //如果 $T = \phi$, 则输出分类决策值最大的分类器所属的主题类别

算法讨论: 算法的时间复杂度主要体现在第 4、10 与 15 步, 通常文本向量比较稀疏, 因此与文本特征相关的循环可忽略不计. 设分类体系有 n 个主题类别, m 篇测试文本, 在不使用类别噪声裁剪算法时, 其时间复杂度为 $O(m \times n)$. 而使用类别噪声过滤后, 因为参与分类比较的类别数 $k \ll n$, 所以其时间复杂度为 $O(m \times k)$.

3.4 参数估计

下面讨论类别噪声裁剪算法中涉及到的参数 $MinRatio$、$LowRatio$ 和 $Th(c_j)$ 的确定问题.

1) 确定 $MinRatio$: $MinRatio$ 是特征选择的阈值, 其值的确定采用标准的 5 折 1 交叉验证的方法: 将训练文本随机分为 5 份, 进行 5 轮测试, 在每一轮测试中选择 4 份作为训练集, 1 份作为测试集, 测试不同的 $MinRatio$ 值对分类性能的影响. 将 5 轮性能测试的结果取平均值, 选择性能最优的参数值, 最终设定 $MinRatio$ 的值为 0.6.

2) 确定 $LowRatio$: 每一个特征对不同的类别都会有相应的贡献值, 设定 $LowRatio$ 的目的是为特征选择出其贡献率较大的类别 (即认为是与特征联系紧密的类别). 由于文本特征的类别属性同样存在多类或兼类的现象, 采用最大值的方法会丢失重要的特征类别信息. 而如果特征对某一类别的贡献率高于其对所有类别贡献率平均值的话, 通常可以认为该特征对这一类别的贡献是较大的. 因而本文是根据特征对所有类别贡献率的平均值来确定 $LowRatio$ 值 (其中 l 为特征 w_i 的类别贡献率 $Sw_{ij} \neq 0$ 的类别个数)

$$LowRatio = \text{avg}(Sw_{ij}) = \frac{1}{l}\sum_{j=1}^{m} Sw_{ij} \quad (3)$$

3) 确定 $Th(c_j)$: $Th(c_j)$ 是分类器决策判别的阈值, 可提供分类错误检测与修正, 其值的确定仍然基于标准的 5 折 1 交叉验证方法.

定义 4. 设 $d \in D(c_j)$ 为训练文本集合中 c_j 类的一个文本, 在 c_j 类分类器 $Classifier_{c_j}$ 作用下的输出值为 $Classifier_{c_j}(d)$, 定义阈值 $Th(c_j)$ 的上界 $TruePosTh(c_j)$ 为

$$TruePosTh(c_j) = \min_{d \in D(c_j)} Classifier_{c_j}(d) \quad (4)$$

定义 5. 同理, 定义阈值 $Th(c_j)$ 的下界 $FalPosTh(c_j)$ 为

$$FalPosTh(c_j) = \max_{c' \neq c_j, d \in D(c')} Classifier_{c_j}(d) \quad (5)$$

定义 4 中定义的 $TruePosTh(c_j)$, 是对训练集中所有属于 c_j 类别的文本执行分类操作后输出结果的最小值, 定义 5 中的 $FalPosTh(c_j)$, 是对训练集中所有非 c_j 类别的文本执行分类操作后输出结果的最大值. 所设置的 $Th(c_j)$ 值, 应小于 $TruePosTh(c_j)$, 以避免出现本应属于 c_j 类别的文本被拒识的情况, 同时 $Th(c_j)$ 的值也不应太小, 至少应大于 $FalPosTh(c_j)$, 以防止非 c_j 类别的文本被错分到 c_j 类中. 在本文中规定: 如果 $FalPosTh(c_j) < TruePosTh(c_j)$, 则设置 $Th(c_j) = FalPosTh(c_j)$; 否则将 $TruePosTh(c_j)$ 与 $FalPosTh(c_j)$ 中的值按升序排列, 选择所有分类值的中值点作为 $Th(c_j)$ 的初始值, 通过分类器 F 量度指标来逐步调整, 最终使分类器 F 量度最大的值即可设置为 $Th(c_j)$ 的最终值.

4 实验结果分析

4.1 实验设置

本文在中英文数据集上对文本分类器类别噪声裁剪算法进行了实验. 中文分类测试体系采用中图法分类体系, 训练集与测试集分别采用 2003、2004 年 863 文本分类评测的 3 600 篇文本. 英文分类测试采用 Reuters-21578 数据集, 去除其中 TOPIC 为空或者 $< BODY >< /BODY >$ 标签不包含文本内容的所有文本, 其中训练集包含 6 574 篇文本, 测试集包含 2 315 篇文本.

实验中, 中文文本的分词使用实验室的 ICSU 系统完成, 利用特征的类别属性分析算法提取特征, 以 Okapi[19] 权值计算公式计算特征词权重, 分类器采用一对多的线性核函数 SVM 算法. 中文文档的评测采用 863 文本分类评测标准, 英文文档的评测采用 Reuter-21578 提供的评测软件. 评测中采用宏、微平均 (Macro & micro average) 策略, 利用 F 值进行实验对比.

为了考查 ECN 算法过滤噪声以及对分类性能的影响, 实验首先从分类体系中随机抽取若干类别为实验对象, 然后依次从当前分类体系的剩余类别中按比例随机抽取若干类别加入到分类系统中 (即不断向系统中加入类别噪声), 评估在不同比例的噪声类别情况下, ECN 算法与 Non-ECN (Non-eliminating class noise)、IG (Information gain) 和 CHI ($\chi^2 - test$) 等算法的分类性能差异. 实验随机进行 5 轮测试, 以 5 组实验数据的平均值为依据分析 ECN 算法对分类性能的改进情况.

4.2 中图分类法数据集结果分析

表 1 显示的是在中图法数据集上 ECN 算法过滤噪声类别的情况统计. 表中各列的含义如

下：$Noise$ 表示系统中噪声类别占全部类别的比例, $Filtered$ 表示被过滤的类别占全部类别的比例、$Filtered\ \&\ Noise$ 表示被过滤掉的类别中噪声类别占全部类别的比例, NEP (Noise elimination precision) 表示类别噪声裁剪的准确率, TCC 表示候选类别集合中包含正确类别的比例, 分别定义如下 (测试文本 d_i, $i = 1, 2, \cdots, n$)：

$$Noise = \frac{1}{n}\sum_{i=1}^{n}\frac{\|M_i\|}{\|U\|} \tag{6}$$

$$Filtered = \frac{1}{n}\sum_{i=1}^{n}\frac{\|F_i\|}{\|U\|} \tag{7}$$

$$Filtered\,\&\,Noise = \frac{1}{n}\sum_{i=1}^{n}\frac{\|F_i \cap M_i\|}{\|U\|} \tag{8}$$

$$NEP = \frac{1}{n}\sum_{i=1}^{n}\frac{\|F_i \cap M_i\|}{\|F_i\|} \tag{9}$$

$$TCC = \frac{1}{n}\sum_{i=1}^{n}\frac{\|\bar{F}_i \cap G_i\|}{\|G_i\|} \tag{10}$$

其中, F_i 定义为对文本 d_i 分类时被裁剪掉的类别集合, $\bar{F}_i$ 是 F_i 集合的补集, M_i 是对文本 d_i 分类时系统中全部类别噪声的集合, G_i 是包含文本 d_i 的正确类别的集合, U 是系统中所有类别的集合 $U = M_i \cup G_i$.

表 1　中图法数据集上类别噪声裁剪结果分析

Table 1　Class noise elimination results on the CLC datasets

$Noise$	$Filtered$	$Filtered\,\&\,Noise$	NEP	TCC
0.667	0.501	0.471	0.945	0.9119
0.75	0.603	0.579	0.967	0.9042
0.80	0.643	0.624	0.974	0.9046
0.90	0.737	0.727	0.987	0.9
0.933	0.83	0.82	0.988	0.8484
0.95	0.812	0.803	0.989	0.8216
0.967	0.82	0.812	0.991	0.7595
0.972	0.825	0.817	0.991	0.7106

表 1 显示, 在中文数据集上, 当 $Noise$ 值为 66.7% 时, ECN 算法可过滤掉 50% 左右的类别, 并且其正确性 (NEP 值) 达到 94.5%, 剩余的候选类别中包含正确类别的比值 TCC 达到 91.19%；随着类别数目增加, 系统中的类别噪声比例不断增大, ECN 算法可过滤掉的类别比率 $Filtered$ 与正确率 NEP 呈上升趋势, 分别达到 82.5% 与 99.1%, 但是由于引入了更多与目标类别模糊度大的类别信息, TCC 值通常会下降. $Filtered$ 与 NEP 值的增加说明系统可以正确过滤掉大量的噪声类别, 同时较大程度提高分类速度, 而 TCC 值的下降, 是否会对分类性能产生较大的影响呢? 这可从图 2 显示的系统 $Macro\ F$ 值随类别噪声比例变化的情况分析图中找到答案.

图 2 的横坐标是系统中不同比例下的类别噪声值, 纵坐标是系统的 $Macro\ F$ 值. 从图中可以看到, 随着系统中的类别噪声比例不断增大, 四种分类算法的性能都呈下降趋势, 但 ECN 算法却始终保持不同程度的优势. 即使是在 $Noise$ 为 90% 的情况下, ECN 算法的 $Macro\ F$ 值仍可达到 0.74, 比 Non-ECN 与 CHI 算法分别提高 7 个百分点与 10 个百分点, 其提升分类性能的效果还是相当明显的. 在最坏的情况下, 即加入所有类别使系统的类别噪声比例达 97.2% 时, ECN 算法的 $Macro\ F$ 值下降到 0.498, 但与 IG、Non-ECN 相比仍然高出 1.5 和 4.4 个百分点, 此时系统的 TCC 降到最低值 0.71, 也就是说此时 ECN 算法的分类准确率的上限为 0.71, 仍然远远高出系统实际的分类准确率, 因此 TCC 值的下降不会对分类性能产生较大影响. 同时, 这 20 个百分点左右的错分文档, 即使在全空间内进行分类搜索, 也不会产生正确的分类结果. 然而采取 ECN 算法, 通过对分类特征类别语义的进一步细划, 过滤掉与目标类别模糊度大的类别信息, 通常可以获取正确的分类类别.

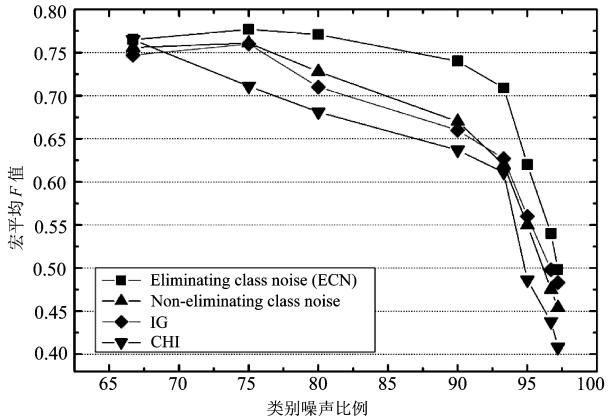

图 2　中图法数据集上 ECN 算法宏平均 F 值随类别噪声比例变化情况分析

Fig. 2　The analysis of $Macro\ F$ value changing with the class noise on CLC dataset

图 3 探讨针对所有的类别, 分类器决策判别阈值 $Th(c_j)$ 对分类结果的影响. 图中的柱状图是当分类器阈值 $Th(c_j)$ 取不同值时, 分类器的使用频率随左侧坐标系的变化情况；而折线表示当 $Th(c_j)$ 取不同值时, 分类系统 $Micro\ F$ 值随右侧坐标系的变化情况. 当 $Th(c_j)$ 取 0 时, 系统不采用任何错误检测与修正机制, 此时系统共调用了 31 838 次 SVM 分类器完成分类操作, 由于存在分类的主动预取错误, 其分类的 $Micro\ F$ 值仅为 0.732；随着 $Th(c_j)$ 逐渐变大, 分类器的使用频率逐渐变大, 其分类性能呈现小幅上升又逐渐下降的趋势, 最终停留在某一值上不再发生变化. 如当 $Th(c_j)$ 取值为 1 时, 系统

不采取任何主动预取技术, 对每个测试文档都需要调用 36 个分类器执行分类操作, 其分类器的使用频率为 36 × 3600=129 600 次, 但由于引入了比较多的类别噪声, 其分类器的 $Micro\ F$=0.744 却处于相对低点. 实际系统是在 $Th(c_j)$ 取值为 0.001, 即当分类器的使用频率为 34 957, 其分类性能达到最高, $Micro\ F$ 值为 0.756. 这说明通过穷举分类模型的方法, 虽然保证了候选类别集的准确率, 但并不能保证系统性能最优. 而 ECN 算法, 结合错误检测与修正机制, 可以有效地去除分类过程中的类别噪声, 在提高分类精度的前提下大幅减少分类器的使用频率, 提高文本分类系统的使用效率.

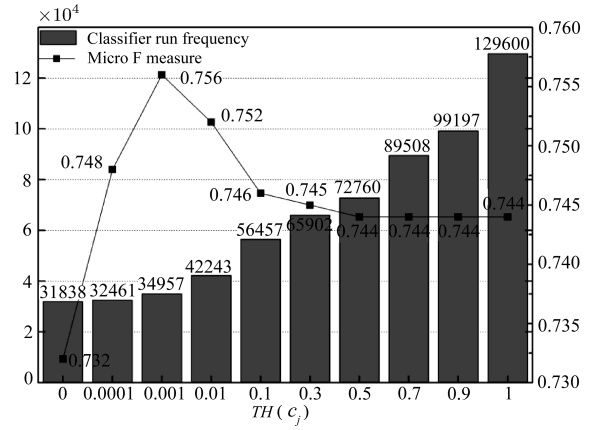

图 3 中图法数据集上类别噪声裁剪算法对分类器使用频率及分类性能的影响情况分析

Fig. 3 The trend analysis of ECN′s classifier run frequency and $Micro\ F$ measure on CLC dataset

4.3 Reuters-21578 数据集结果分析

表 2 显示的是在 Reuters-21578 数据集上 ECN 算法过滤噪声类别的情况.

表 2 Reuters-21578 数据集上类别噪声裁剪结果分析

Table 2 Class noise elimination results on the Reuters-21578 dataset

Noise	*Filtered*	*Filtered & Noise*	*NEP*	*TCC*
0.50	0.5	0.494	0.989	0.9885
0.66	0.556	0.55	0.991	0.982
0.75	0.548	0.534	0.971	0.9455
0.80	0.607	0.597	0.982	0.9478
0.857	0.645	0.636	0.988	0.9391
0.875	0.677	0.671	0.991	0.9478
0.889	0.694	0.689	0.992	0.9483
0.90	0.7	0.69	0.987	0.9027

表 2 显示, 在 Reuters-21578 数据集上, ECN 算法表现出与中文数据集上相类似的趋势, 即随着系统中的类别噪声比例不断增大, $Filtered$ 与 NEP 均呈上升趋势, 但是 TCC 的值通常会下降. 不同的是, 此时 TCC 值明显高于中文数据集上的结果. 造成这一差异的原因, 主要是因为中文文本在转换为特征词向量时需要经过分词处理, 分词一旦出现偏差就会影响类别语义的识别, 而英文文本由于采用 Term 为基本处理单元, 通常不存在分词错误问题, 同时在英文 Term 的处理上使用了词根还原技术, 进一步提高了 Term 对类别的指向作用.

图 4 显示的是 Reuters-21578 数据集上分类系统 $Macro\ F$ 值随类别噪声比例变化的情况分析. 从图中看到, 以 $Noise$ 等于 80% 为例, ECN 算法与 Non-ECN、CHI 算法相比, 其 $Macro\ F$ 值可以分别提高 6∼10 个百分点, 表现出了较好的性能, 然而这种分类性能提高的幅度会随着更多与目标类别模糊度大的类别的加入而逐渐下降, 但即使是在 $Noise$ 为 90% 的情况下, ECN 算法仍然可以保持相对较高的分类性能.

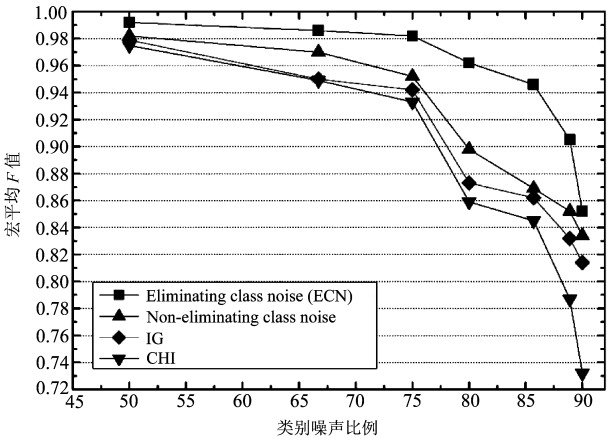

图 4 Reuters-21578 数据集上系统宏平均 F 值随类别噪声比例变化情况分析

Fig. 4 The analysis of $Macro\ F$ value changing with the class noise on Reuters-21578 dataset

图 5 探讨英文数据集上分类器决策判别阈值 $Th(c_j)$ 对分类结果的影响. 从图中看到, 随着 $Th(c_j)$ 值的变化, 其分类器的使用频率呈上升的趋势并达到最大值, 而标识分类系统性能的 $Micro\ F$ 指标则先上升后下降, 表现出了与中文数据集上相类似的趋势. 通过对图 4 与图 5 的分析, 证明在英文数据集上, ECN 算法同样可以取得较好的实验效果, 该算法可以有效去除分类过程中的类别噪声, 提高分类性能. 同时, 中、英文数据集上的实验也都表明, 不使用错误检测与修正机制, ECN 算法并不足以保证其高分类性能, 因此系统引入了阈值策略. 该策略可有效检测动态层次分类过程中的分类出错情况并及时修正, 提升系统性能.

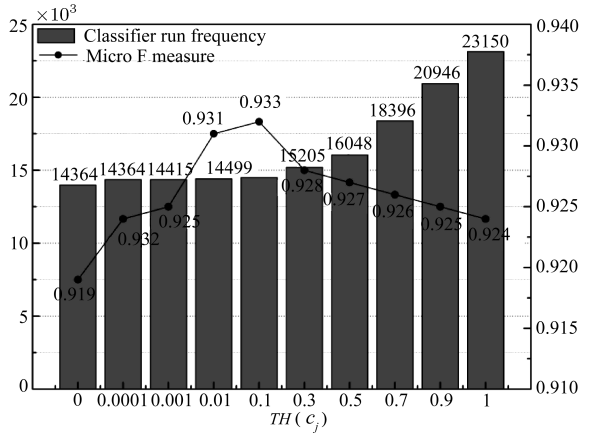

图 5　Reuters 数据集上类别噪声裁剪算法对分类器使用频率及分类性能的影响情况分析

Fig. 5　The trend analysis of ECN′s classifier run frequency and *Micro F* measure on Reuters-21578 dataset

4.4　ECN 算法的总体性能比较

表 3 将 ECN 算法与 Non-ECN、IG 及 CHI 的总体分类性能进行了对比. ECN 算法在中、英文数据集上的宏、微平均 F 值分别为 0.756 与 0.759, 0.86 与 0.933, 均获得了较其它算法更优的分类性能. 中文数据集上的成绩超过了 2004 年 863 文本分类评测的最佳结果, 英文数据集上的成绩也与 Reuters-21578 上的最好成绩基本相当[20]. 实验中同时将 ECN 算法与传统的层次分类算法 HC (Hierarchical classification) 进行了实验对比, HC 的错误修正采用 ROH 策略. 由于实验中采用的中图法与 Reuters-21578 数据集都是平面分类方式, 因此 HC 算法的实验采用人工方法设置分类层次树为两层结构. 在中图法数据集上, 设置 A-K 共 11 个类别为一组, 建立中间节点为社会科学类, N-X 共 25 个类别为一组, 建立中间节点为自然科学类; 在 Reuters-21578 数据集上, 设置 acq、corn、crude、earn、grain 共 5 个类别为一组, 建立中间节点 A, interest、money-fx、ship、trade、wheat 共 5 个类别为一组, 建立中间节点 B. 结果显示, HC 与 ECN 算法相比, 其中英文数据集上的 $Macro\ F$ 值分别低了 1.8 与 1 个百分点, 算法分层处理的思想并没有体现出优势, 其原因主要是分类结构树的层次设置对分类结果所造成的影响. 针对不同的分类体系, 其分类树应该设置几个层次, 应该采用什么标准将不同的类别放置于同一个类别层次等问题, 目前都还缺乏系统的设置方法, 不容易找到最优设置.

ECN 算法由于可以基于特征的类别属性直接裁剪噪声类别信息, 因此其在反馈学习过程中具有特别的优势. 表 3 中第 1 列显示了 ECN 与 FL (Feedback learning) 结合后系统的分类性能情况. 其核心思想是：从测试样本标题中选取最具有代表性的 N-gram 字串, 利用该字串的类别属性信息确定该文档的候选类别, 然后用分类器进行类别判定. 在选取 N-gram 字串的方法上采用了词汇链构造算法, 所使用的中、英文词汇资源分别为 WordNet 和 HowNet. 通过对获取的词汇链进行评价 (根据词汇链长度与链中词汇的频度), 选取最具代表性的词汇链, 将出现在链中的标题字串作为反馈学习得到的新特征加入到扩展特征词集中. 这样, 不必重新训练分类模型, 只利用新学习到的特征词汇及其类别语义, 就可使中、英文数据集的 $Macro\ F$ 值分别达到 0.804 与 0.874, 较反馈学习前又提高了 5 和 1.4 个百分点.

表 3　ECN 算法分类性能比较

Table 3　The performance of ECN Algorithm

		ECN+ FL	ECN	HC	Non-ECN	IG	CHI
CLC	*Macro F*	0.804	0.756	0.738	0.739	0.73	0.711
	Micro F	0.806	0.759	0.74	0.744	0.732	0.714
Reuters	*Macro F*	0.874	0.86	0.85	0.856	0.857	0.809
	Micro F	0.943	0.933	0.923	0.928	0.926	0.876

5　结论

基于特征类别属性分析的类别噪声裁剪算法能够有效地处理文本分类问题, 该算法通过识别文本中关键特征所蕴含的主题类别信息, 能够主动预测待分类文本可能归属的类别, 从而有效减少不相关类别参与决策引入的类别噪声, 在保证分类精度的前提下提高文本分类的响应时间, 提高文本分类技术的使用效率. 此外, 该算法具有较好的可扩展性, 可以通过基于错误实例的相关反馈学习或手工添加类别语义词的方法, 进一步提高文本分类的精度.

实验表明, 采用类别裁剪算法构造分类候选集合的方法, 仍然会有一定程度的错判, 导致分类不准确, 之所以没有对分类性能产生太大的影响, 是因为这些被错分的文本, 即使采用全部分类器进行判别仍然得不到正确的分类结果. 下一步的工作, 主要是进一步深入研究虚拟中间点分类错误的检测与修正技术, 从而使类别裁剪技术更加有效.

References

1 Mitchell T M. *Machine Learning*. New York: McGraw Hill, 1996. 112∼141

2 Sebastiani F. Text categorization. In: Proceedings of Text Mining and Its Applications to Intelligence, CRM and

Knowledge Management. Southampton, UK, WIT Press, 2005. 109∼129

3 Masand B, Linoff G, Waltz D. Classifying news stories using memory based reasoning. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, ACM Press, 1992. 59∼65

4 Lam W, Ho C Y. Using a generalized instance set for automatic text categorization. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR′98). Melbourne, Australia, ACM Press, 1998. 81∼89

5 Quinlan J R. Induction of decision trees. *Machine Learning*, 1986, **1**(1): 81∼106

6 Apte C, Damerau F, Weiss S. Text mining with decision rules and decision trees. In: Proceedings of the Workshop with Conference on Automated Learning and Discovery: Learning from Text and the Web. Pittsburgh, USA, 1998. 487∼499

7 Kwok J T Y. Automatic text categorization using support vector machine. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR′98). Melbourne, Australia, ACM Press, 1998. 347∼351

8 Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of European Conference on Machine Learning (ECML). Berlin, Germany: Springer, 1998. 137∼142

9 Zhu X Q, Wu X D, Chen Q J. Eliminating class noise in large datasets. In: Proceedings of the 20th International Conference on Machine Learning (ICML 2003). Washington DC, USA, 2003. 920∼927

10 Li Rong-Lu, Hu Yun-Fa. A density-based method for reducing the amount of training data in KNN text classification. *Journal of Computer Research and Development*, 2004, **41**(4): 539∼545
(李荣陆, 胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法. 计算机研究与发展, 2004, **41**(4): 539∼545)

11 Wang Qiang, Wang Xiao-Long, Guan Yi, Xu Zhi-Ming. A research on text categorization based on the fusion of K-NN and SVM. *Chinese High Technology Letters*, 2005, **5**: 19∼24
(王强, 王晓龙, 关毅, 徐志明. K-NN 与 SVM 相融合的文本分类技术研究. 高技术通讯, 2005, **5**: 19∼24)

12 Yang Y M, Pedersen J. A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML97). San Francisco, USA: Morgan Kaufmann Publishers, 1997. 412∼420

13 Yan J, Liu N, Zhang B Y, Yan S C, Chen Z, Cheng Q S, Fan W G, Ma W Y. OCFS: optimal orthogonal centroid feature selection for text categorization. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR′2005). Salvador, Brazil, ACM Press, 2005. 122∼129

14 Choi B, Peng X G. Dynamic and hierarchical classification of web pages. *Online Information Review*, 2004, **28**(2): 139∼147

15 McCallum A, Rosenfeld R, Mitchell T, Ng A. Improving text classification by shrinkage in a hierarchy of classes. In: Proceedings of the 15th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1998. 359∼367

16 D′Alessio S, Murray K, Schiaffino R. The effect of topological structure on hierarchical text categorization. In: Proceedings of the Sixth Workshop on Very Large Corpora at COLING-ACL. Montreal, Canada, 1998. 66∼75

17 Cheng C H, Tang J, Fu A W, King I. Hierarchical classification of documents with error control. In: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hong Kong, China, 2001. 433∼443

18 Zhang Jia-Min. The meta-functional perspective of title prediction. *Foreign Language Education*, 2004, **25**(6): 36∼39
(张加民. 标题预示性的元功能视角. 外语教学, 2004, **25**(6): 36∼39)

19 Ault T, Yang Y M. KNN at TREC-9. In: Proceedings of the Ninth Text Retrieval Conference (TREC-9). Maryland, USA, 1999. 127∼134

20 Debole F, Sebastiani F. An analysis of the relative hardness of reuters-21578 subsets: research articles. *Journal of the American Society for Information Science and Technology*, 2005, **56**(6): 584∼596

王 强 哈尔滨工业大学智能技术与自然语言处理研究室博士研究生. 主要研究方向为文本挖掘和机器学习算法. 本文通信作者.
E-mail: qwang@insun.hit.edu.cn
(**WANG Qiang** Ph. D. candidate at Intelligent Technology and Natural Language Processing Laboratory, Harbin Institute of Technology. His research interest covers text mining and machine learning. Corresponding author of this paper.)

关 毅 哈尔滨工业大学计算机科学与技术学院教授. 主要研究方向为问答系统、统计语言处理及文本挖掘.
E-mail: guanyi@insun.hit.edu.cn
(**GUAN Yi** Professor at School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers QA, statistical language processing, and text mining.)

王晓龙 哈尔滨工业大学计算机科学与技术学院教授. 主要研究方向为人工智能、机器学习、计算语言学与中文信息处理.
E-mail: wangxl@insun.hit.edu.cn
(**WANG Xiao-Long** Professor at School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers artificial intelligence, machine learning, computational linguistics, and Chinese information processing.)