文章编号:1000-5862(2019)01-0076-08

一种新的样本选择算法及其在文本分类中的应用

万中英,王明文,左家莉,刘长红
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:在保证分类性能的前提下,如何从大量的训练样本集合中选择重要样本子集,是模式分类中的一个重要问题.基于该问题提出了一种新的样本选择算法,并将该算法应用于文本分类,在标准文档集 Reuters-21578、复旦文档集和 20newsGroup 新闻组文档集上进行了实验.实验结果表明:该方法能有效地选取边界样本,且采用 SVM 和 KNN 分类能得到较好的分类结果,尤其是在不均衡文档集上效果更佳.

关键词:边界样本;样本选择;文本分类;支持向量机;K 近邻
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2019.01.13

0 引言

要得到一个较好的训练结果,除了要选择合适的模型外,训练样本的好坏也是至关重要的.由于不同的样本在不同的任务中所起的作用不同,这就要根据需要对训练样本进行选择.

样本选择(Instance/Sample Selection)最早是由 P. E. Hart[1] 针对近邻分类算法提出的,其基本思想是,从原始的样本集 A 中,利用某一规则抽取出一小部分样本,组成新的样本集 B;再在样本集 B 上,利用给定算法 M 训练一个新的分类器,并且使其在样本集 B 和样本集 A 上训练得到的分类性能相当[2].

CNN(Condensed Nearest Neighbor Rule)[1] 方法先是在每个类别中随机挑选合并成样本子集,然后用该样本子集测试,若误分了样本,则直接将该样本加入到样本子集中去,但是该特性也导致了噪声点更容易被保留.RNN(Reduced Nearest Neighbor Rule)方法[3] 的主要思想是样本子集先包含所有的样本,删除 1 个样本之后再对剩下的样本进行分类,若能被正确分类,则删除的样本就被永久删除,该方法恰好能够去除其中的噪声点而且能够保留边界样本.ENN(Edited Nearest Neighbor Rule)[4] 方法,同样能有效地过滤噪声,保留边界样本.除了这些之外,还有 FCNN(Fast Condensed Nearest Neighbor Rule)[5] 等方法,所有这些近邻算法,都是将焦点放在边界样本、噪声点方面.

此外还有许多其他的样本选择方法,如基于聚类的方法[6-7]、基于特定分类器的方法[8-11]、基于智能优化的方法[12-16]、基于组合多个样本选择的方法[14]、基于局部敏感哈希的方法[15] 等.

在训练样本中,边界样本的判定方式以及训练样本中包含边界样本数量的多少对分类的精度起主要作用.考虑到不同类别间相邻的边缘样本应为边界样本,因此本文结合近邻算法提出了一种新的边界样本选择算法.本文将该算法应用于文本分类中,在标准文档集 Reuters-21578、复旦文档集和 20newsGroup 新闻组文档集上进行了实验,选取了边界样本,并得到了较好的分类效果.

1 样本选择算法

1.1 算法思路

文本的表示多采用向量空间模型,而文本就将其看成是 m 维空间上的点.不妨假设在 2 维空间上考虑样本选择问题,如图 1 ~ 图 3 所示.

从图 1 可看出所有的点投影到 y 轴上,点 1 的值最大,点 2 的值最小;将所有的点投影到 x 轴上,点 3 的值最大,点 4 的值最小;而点 1、2、3、4 均在该类的边缘上.由此可见,同一类中,在某 1 维上取得最大值的点或最小值的点一定是在该类的边缘上,而与边缘样本邻近的样本就是边界样本.从第1类

收稿日期:2018-05-22
基金项目:国家自然科学基金(61462045,61462043,61163006)和江西省教育厅科学技术研究(GJJ150354)资助项目.
作者简介:万中英(1977-),女,江西南昌人,副教授,主要从事信息检索、文本挖掘研究.E-mail:libby2005@126.com

的每个边缘样本出发寻找离其最近的 k 个第 2 类样本,再从这 k 个样本出发寻找离其最近的 k 个第 1 类样本,如此反复,直到找不到新样本为止. 图 2 为在 2 维空间中 2 类之间选取的边界样本,其中圆点表示 A 类的样本,三角形表示 B 类的样本, k 取值为 3. 点 1、2、3、4 为 A 类的边缘样本,首先,从点 1 出发找到与其相邻的 k 个 B 类样本,再从 B 类的每一个样本出发找到与其邻近的 k 个 A 类样本; 若没有新的样本加入,则从点 2、3、4 出发重复上面的工作,最后找到的样本为 2 个类的边界样本. 将 2 类问题扩展到多类问题,把其中 1 个类作为一类,其他类作为另一类选取边界样本,对于每个类均采取相同的方法,最后找到的是所有类别间的边界样本. 如图 3 所示,在图 2 的基础上增加 3 个类,分别用方形、心形和十字形来表示.

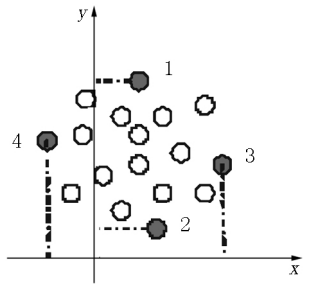

图 1　边缘样本

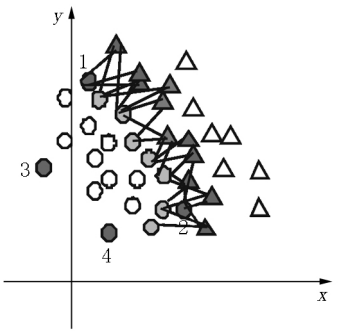

图 2　2 类边界样本的选择

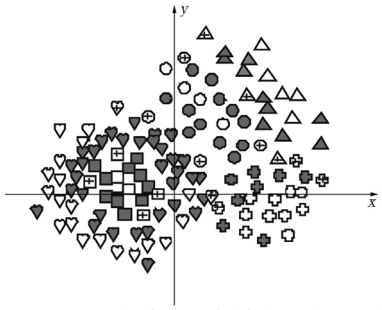

图 3　多类边界样本的选择

1.2　算法描述

基于上述的算法思路,本文提出了完全不依赖于任何分类算法的边界样本选择算法. 设有 n 个 m 维的样本数据,分别属于 $c_1, c_2, \cdots, c_h$ 类. 为了找到这 h 个类的边界,先考虑找 2 个类的边界样本. 如要找 c_1 类的边界样本,把其他类的所有样本看成另一个 c_t 类; 以相同方法再找 c_2 类的边界样本,直到所有类边界样本都找到. 2 个类的边界样本选择算法思想描述如下: (i) 输入 c_1 类样本; (ii) 在 c_1 类中找出每一维的最大值和最小值所对应的样本 $t_1, t_2 \cdots$ 加入到队列 Q 中; (iii) 从队列中取第 1 个样本 t_1 并出队,若 t_1 属于 c_1 类则加入集合 S 中,再计算该样本与另一个类 c_t 中所有样本的欧氏距离,将值最小的前 k 个且不在队列 Q 中的样本加入到队列 Q 中,即与 t_1 最近的另一个类的样本; (iv) 继续从队列中取出下一个样本 t_2 并出队,若 t_2 属于 c_1 类,则步骤同上; 若 t_2 属于 c_t 类,则计算该样本与另一个类 c_1 中所有样本的欧氏距离,将值最小的前 k 个且不在队列 Q 中的样本加入到队列 Q 中; (v) 重复(ii) ~ (iv) ,直到队列 Q 为空为止; (vi) 输出集合 S 中得到的为每个类的边界样本.

接着以相同的方法,从步骤(i) 开始找 c_2 类的边界样本,直到所有类别的边界样本都找完为止. 也可以先计算好 c_1 类样本与 c_t 类样本间的欧氏距离得到矩阵 $\boldsymbol{M}$,$\boldsymbol{M}$ 中的第 i 行表示 c_1 类中第 i 个样本与 c_t 类中样本的距离,$\boldsymbol{M}$ 中的第 i 列表示 c_t 类中第 i 个样本与 c_1 类中样本的距离.

该算法的伪代码如下:

```
Algorithm: 边界样本选择算法
输入: 类别数 c_count; 维数 m; 样本总数 n, c_1,
      c_2, ..., c_h 各类文档;
输出: 所选的所有类的边界样本;
for i = 1 to c_count
    设置 c_i 类的文档类别为 1,其他类别文档
    的类别为 2;
    for j = 1 to m
    找出每一维的最大值和最小值对应的样本
    加入队列 Q 中,且为加入的样本标识为已
    选;
end for
while( Q 不为空)
    将 Q 中第 1 个样本 t 出队;
    if t 的类别号为 1
    将 t 加入集合 S 中;
    计算 t 与类别号为 2 样本的距离;
    将前 k 个距离最小且不被标识的样本加入
    队列 Q 中,并标识加入的样本;
```

end if

if t 的类别号为 2

计算 t 与类别号为 1 的样本的距离;

将前 k 个距离最小且不被标识的样本加入队列 Q 中,并标识加入的样本;

end if

end while

集合 S 中为所选的 c_i 类的边界样本;

end for.

设有 k 个类别,样本总数为 n,样本维数为 m,某一个类 c_i 的样本数为 n_1,则其他样本数是 $n-n_1$. 该算法主要花费的时间是选择样本的时间,即找队列中每个样本的前 k 个距离最小的样本所花费的时间. 最坏的情况是选取了类 c_i 中的所有样本作为边界样本,则花费的时间为 $n_1(n-n_1)$;找出所有类别的边界样本花费为 $kn_1(n-n_1)$;因此该算法的时间复杂度为 $O(kn_1(n-n_1))$. 经典的近邻算法 CNN 除了排序所花时间外,其他所花费时间和 FCNN 的时间复杂度均为 $O(ns)$,其中为选取子集的样本个数. 由于本文的算法是从每个类中分别去选取样本,因而花费的时间较多,但该算法克服了 CNN 算法与读取数据顺序相关的缺点,且不是 FCNN 算法所提出的从类的中心找起而是从类边缘找起,所找到的样本更有可能是边界样本,会减少冗余. 下面将其应用于文本分类,对选取的子集进行分类后的结果作进一步的分析.

2 实验

2.1 仿真实验

为了能看到选取的可视效果,先进行仿真实验. 采用均匀分布生成 4 个类别,每个类 200 个数据,共 800 个数据,如图 4 所示;图 5 是采用边界样本选择方法选取后的结果. 还采用均值为 0,标准差为 1 的正态分布生成了 3 个类别,每个类 200 个数据,共 600 个数据,如图 6 所示;图 7 是选取后的结果. 从图 5 和图 7 可以看出该方法能有效地选取出边界样本.

2.2 标准数据集的实验

本文选用了标准文档集英文路透社文档集 Reuters-21578、20NewsGroup 新闻组和中文的复旦文档集. Reuters-21578 是英文多标签不均衡类,20NewsGroup 新闻组是英文的均衡类,复旦文档集是中文的不均衡类. Reuters-21578 选取了 14 个类

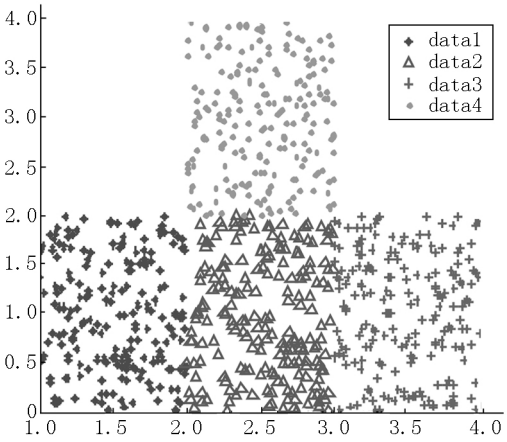

图 4 均匀分布生成的 4 类数据

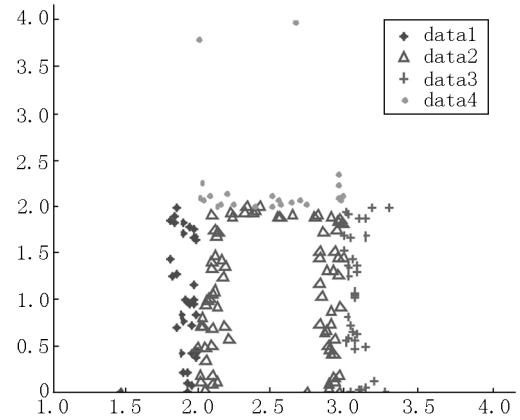

图 5 选取后的结果

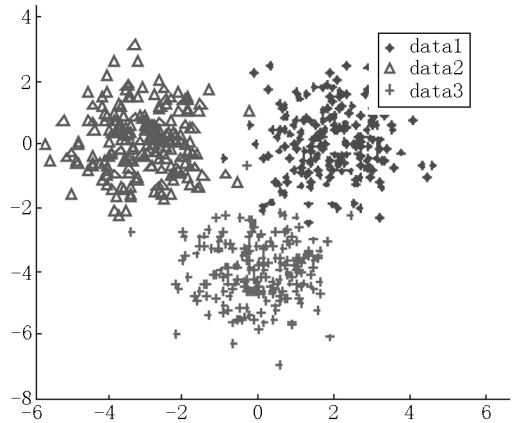

图 6 正态分布生成的 3 类数据

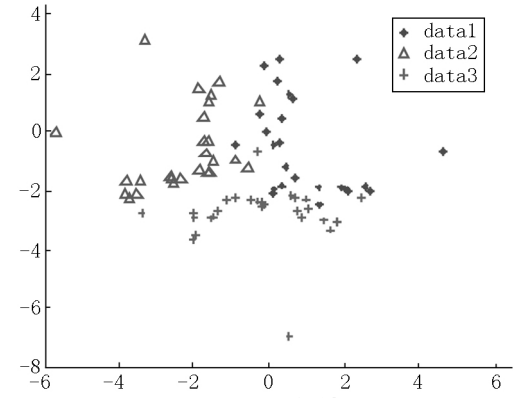

图 7 选取后的结果

别,训练文档数为 7 583,测试文档数为 2 897; 20NewsGroup 新闻组 20 个类别,训练文档数为 11 314,测试文档数为 7 532; 复旦文档集有 20 个类别,训练文档数为 8 214,测试文档数为 5 695. 经过特征选择等预处理后,文档集都选取了 1 000 维的特征数据.

采用宏平均 F_1 值和微平均 F_1 值来评价分类效果,分别用 M_{acF_1} 和 M_{icF_1} 来表示. 下述实验中路透社文档集用 LT 表示,20NewsGroup 新闻组用 NG 表示,复旦文档集用 FD 表示.

2.2.1　SVM 实验　采用 libsvm 分类器的 *c-svc* 多元分类,c 值取 8 000,核函数为径向基函数.

在复旦文档集上进行实验,所有样本经过分类后的 M_{acF_1} 为 0.659 23,M_{icF_1} 为 0.866 46,支持向量数 3 669. 采用本文的方法进行样本选择后,在相同条件下,实验结果如表 1 所示,其中 N_K 是选取的离另一个类最近的样本个数; d_{count} 为最后选取的边界样本个数,$S_{vectors}$ 为支持向量数.

从表 1 可看出,在 $N_K = 6$ 时达到最佳,且精度高于采用所有样本进行 SVM 实验得到的结果,而样本数只有 4 746 个,支持向量数只要2 972个. 且当 $N_K = 1$ 时,M_{acF_1} 高于所有样本分类下的结果,M_{icF_1} 的值略低于,但不到 0.1%. 由此可见,针对不均衡类的文档集本文的算法不仅减少了样本数量,而且提高了分类效率.

在 Reuters-21578 上进行实验,所有样本经过分类后的 M_{acF_1} 为 0.757 4,M_{icF_1} 为 0.890 0,支持向量数为 300 6 个. 采用本文的方法进行样本选择后,在相同条件下,实验结果如表 2 所示.

表 1　FD 选取边界样本后的 SVM 分类结果

N_K	1	2	3	4	5	6	7
d_{count}	3 988	4 289	4 576	4 782	4 967	5 104	5 258
M_{acF_1}	0.662 2	0.662 5	0.665 9	0.663 3	0.664 0	0.665 3	0.660 9
M_{icF_1}	0.865 7	0.866 6	0.867 1	0.865 8	0.867 3	0.868 6	0.867 4
$S_{vectors}$	2 790	2 840	2 876	2 926	2 956	2 972	3 001

表 2　LT 选取边界样本后的 SVM 分类结果

N_K	1	2	3	4	5	6	7
d_{count}	3 384	3 543	3 631	3 702	3 760	3 802	3 849
M_{acF_1}	0.765 6	0.762 5	0.758 3	0.755 4	0.753 8	0.753 9	0.757 0
M_{icF_1}	0.888 3	0.889 4	0.888 3	0.888 0	0.887 6	0.886 9	0.888 0
$S_{vectors}$	2 293	2 425	2 477	2 527	2 562	2 573	2 593

从表 2 可看出,当 $N_K = 2$ 时效果达到最佳,且精度与采用所有样本进行 SVM 实验得到的结果相当. 当 $N_K = 1$ 时,M_{acF_1} 高于所有样本分类下的结果,M_{icF_1} 的值略低于,但不到 0.1%,选择的样本数不到原来的一半,由此可见,本文的算法在多标签不均衡的文档集上也是行之有效的.

在 20NewsGroup 上进行实验,所有样本经过分类后的 M_{acF_1} 为 0.741 3,M_{icF_1} 为 0.746 4,支持向量数为 6 480. 采用本文的方法进行样本选择后,在相同条件下,实验结果如表 3 所示.

从表 3 可看出,当 $N_K = 6$ 时 M_{acF_1} 达到最佳; 且精度略低于采用所有样本进行 SVM 实验得到的结果,但也没超过 0.3%. 由此可见,本文的算法在均衡文档集上效果不明显. 但当 $N_K = 1$ 时,在精度相差 1.5% 的范围内,还是能有效地减少样本的数量.

表 3　NG 选取边界样本后的 SVM 分类结果

N_K	1	2	3	4	5	6	7
d_{count}	6 341	7 171	7 841	8 332	8 731	9 045	9 280
M_{acF_1}	0.729 7	0.735 1	0.735 6	0.738 4	0.739 2	0.739 3	0.737 6
M_{icF_1}	0.733 9	0.740 0	0.740 3	0.743 2	0.743 8	0.744 2	0.742 0
$S_{vectors}$	4 847	5 320	5 647	5 834	5 960	6 048	6 110

从上述 3 个标准文档集的实验结果可看出,当 $N_K = 1$ 时其精度就与采用所有样本的实验结果相当,甚至更好. 说明采用本文的方法选取的边界样本包含了更多的分类信息,对分类的贡献也最大; 且选择算法在整个过程不依赖于任何的分类器,而选取的样本直接用与 SVM 分类就能达到较好的分类效果.

2.2.2　KNN 实验　本文还进行了 KNN 实验. 因为路透社文档集是一个不均衡的多标签的文档集,因

此有些类别中的文档数相差较大,如所选最大类 earn 类的训练文档数为 2 877,而最小类 gnp 类的训练文档数为 101;且有较多文档同时属于多个类别,如 corn 类别中的文档几乎同时属于 grain 类. 这些都将影响分类效果,在 KNN 分类算法中选取分类结果最好的 K 值. 路透社文档集选取 $K=30$ 进行分类,而 20NewsGroup 新闻组和复旦文档集选取 $K=5$ 进行分类. 其分类结果如表 4 ~ 表 6 所示.

表 4 FD 选取边界样本后的 KNN 分类结果

N_K	1	2	3	4	5	6	7
d_{count}	4 195	4 324	4 440	4 543	4 647	4 746	4 843
M_{acF_1}	0.538 8	0.534 9	0.539 8	0.541 1	0.543 3	0.546 7	0.546 9
M_{icF_1}	0.795 7	0.794 9	0.798 2	0.800 1	0.802 9	0.804 3	0.804 5

注:采用所有样本进行 KNN 实验的 M_{acF_1} 为 0.554 21,M_{icF_1} 为 0.816 16.

表 5 LT 选取边界样本后的 KNN 分类结果

N_K	1	2	3	4	5	6	7
d_{count}	3 384	3 543	3 631	3 702	3 760	3 802	3 843
M_{acF_1}	0.699 1	0.697 7	0.695 5	0.699 4	0.698 5	0.693 3	0.694 8
M_{icF_1}	0.811 0	0.817 2	0.824 8	0.837 9	0.837 9	0.833 1	0.833 1

注:采用所有样本进行 KNN 实验的 M_{acF_1} 为 0.703 38,M_{icF_1} 为 0.847 92.

表 6 NG 选取边界样本后的 KNN 分类结果

N_K	1	2	3	4	5	6	7
d_{count}	6 341	7 171	7 841	8 332	8 731	9 045	9 280
M_{acF_1}	0.593 4	0.586 6	0.595 7	0.600 8	0.604 1	0.608 2	0.609 6
M_{icF_1}	0.568 0	0.548 9	0.560 6	0.567 4	0.571 4	0.576 0	0.577 9

注:采用所有样本进行 KNN 实验的 M_{acF_1} 为 0.628 80,M_{icF_1} 为 0.610 99.

从表 4 和表 6 可看出,随着 N_K 取值的增加精度也随之增加,且逐渐接近采用所有样本的分类结果;从表 5 可看出,当 $N_K=4$ 时达到最好精度. 此后,随着样本的增加反而有所下降,这是因为 LT 集是一个不均衡的多标签的文档集,最大的类别文档数为 2 877,而最小的类文档数只有 101. 当 N_K 达到一定值时,小类的文档基本都已加入样本集,随着 N_K 值的增大,加入的都是大类的文档,这样反而会影响小类的精度,从而影响整个文档集的结果.

从实验结果可看出,KNN 的结果虽然不如 SVM 好,可也说明边界样本包含了更多的分类信息,对分类的贡献也是最大的;在 LT 和 FD 的 KNN 实验中发现通过本文方法选取的样本时进行 KNN 实验就已经能够达到与所有样本进行 KNN 实验的结果;而在 NG 文档集中的实验结果较差是因为它是一个均衡类,它仅包含边界样本信息,进行 KNN 分类实验是不够的,它还应该加入内部样本,且样本数据越多越能达到较好的效果.

2.2.3 选取样本数目分析 根据本文方法选取各类样本数目及与样本总数的比较如图 8 ~ 图 13 所示. 图中纵坐标为选取的文档数目;横坐标为选取的离另一个类最近的样本个数 N_K;图例中用 $c_i(i=1,2,3,\cdots,20)$ 来代表不同的类别,其后括号里的为该

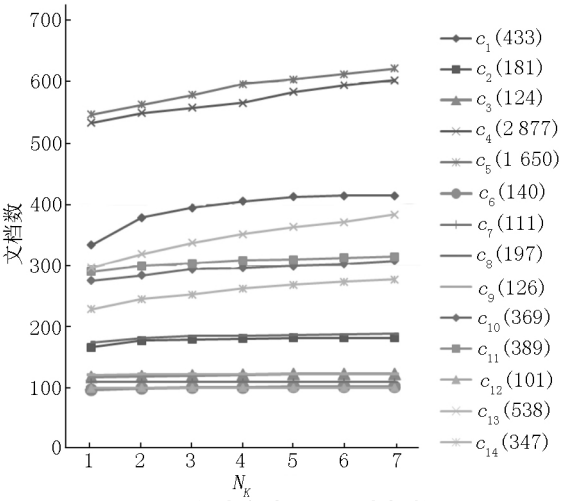

图 8 IT 集中各类别选取文档数目

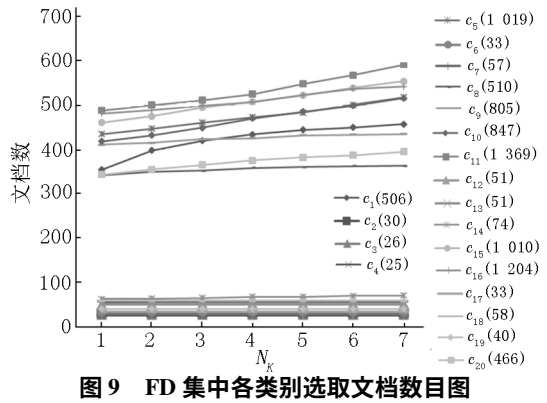

图 9 FD 集中各类别选取文档数目图

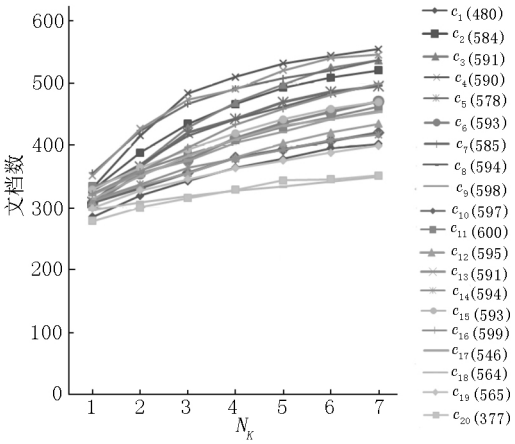

图 10　NG 集中各类别选取文档数目图

类的原文档数目.

从图 8 和图 9 可看出,文档数在 100 以下的小类别在 $N_K = 1$ 时,该类别文档已基本被选入; 文档数在 100 ~ 200 之间的类别,类别文档也基本在 $N_K = 2$时全部被选入,而文档数在 200 个以上的类别,所选文档数会随着 N_K 值的增加而逐渐增加. 而对于均衡类别的 20newsGroups 文档集,从图 10 可以看出,其所有类别都会随着 N_K 值增加而所选文档数会逐渐增加. 从图 11 ~ 图 13 可以看出,在复旦文档集和路透文档集中所选文档数目的增加速度相对于 20newsGroups 文档集来说要慢. 这是因为前 2 个文档集是不均衡类,后面加入的都是大类别的文档,而 20newsGroups 文档集的每个类别的文档数都在增加.

为了分析由于大类别文档数的增加是否会对小类别的精度产生影响,图 14 ~ 图 16 将每个类选取边界后,将测试集进行 SVM 实验的 F_1 值的变化情况描述出来.

从图 14 ~ 图 16 可看出,大部分的类别随着 N_K 值的增大其 F_1 值基本保持稳定,而影响整个文档集精度的主要是几个小类别文档. 图14中FD集中的

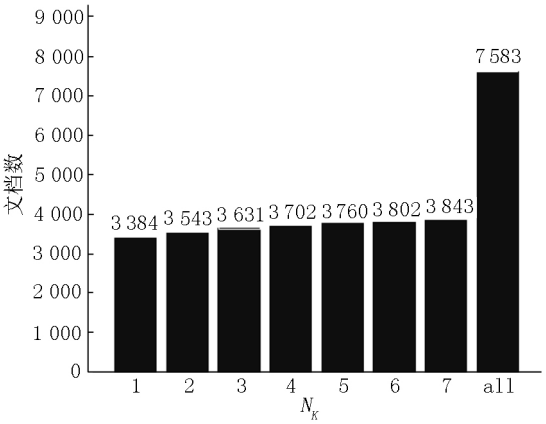

图 11　IT 集每次选取文档集数目及总文档数

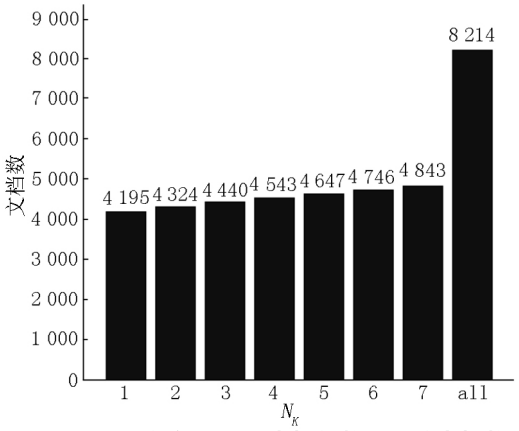

图 12　FD 集每次选取文档集数目及总文档数

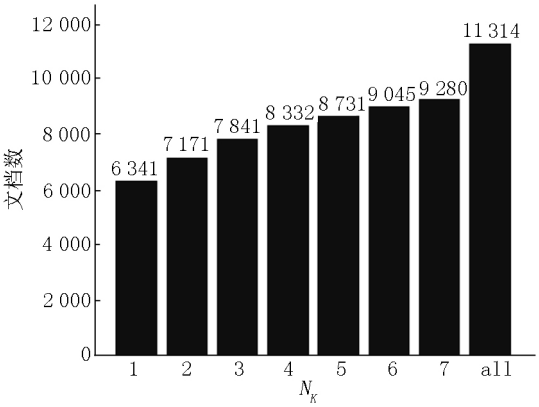

图 13　NG 集每次选取文档集数目及总文档数

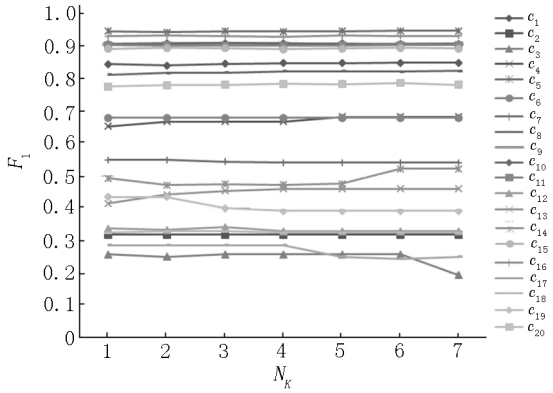

图 14　FD 文档集各类别 F_1 值的变化情况

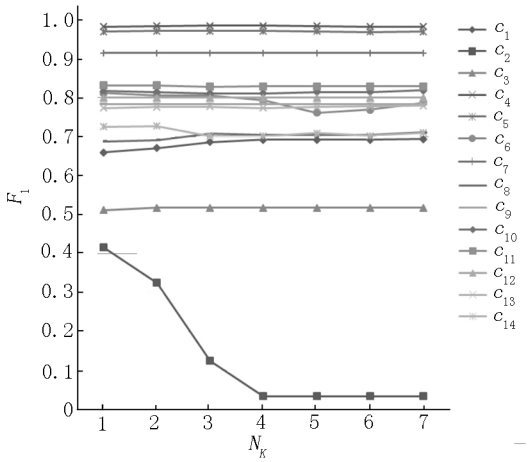

图 15　IT 文档集各类别 F_1 值的变化情况

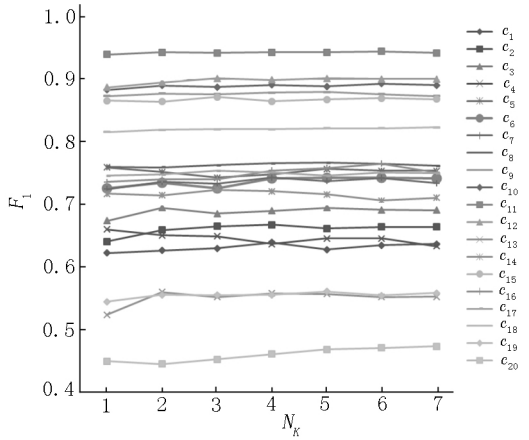

图 16 NG 文档集各类别 F_1 值的变化情况

c_3 类(26 个样本) 在 $N_K=7$ 时精度明显下降; c_{19} 类(40 个样本) 在 $N_K=3$ 时精度先有所下降,再趋于稳定; c_{17} 类(33 个样本) 在 $N_K=5$ 时精度先有所下降,再趋于稳定; 还有 c_7(57 个样本) 和 c_{12}(51 个样本) 也是类似的情况. 由此可见,随着 N_K 值的增大,大类别文档的加入对小类别的精度是有一定影响的. 图 15 中 IT 集中 c_2 类(181 个样本) 的精度下降幅度比较大,这是因为 IT 集是一个不均衡多标签的类别,刚好 c_2 类的样本大部分有 2 个标签以上,随着大类别文档的加入就严重影响了其精度; c_{20} 类(347 个样本) 在 $N_K=3$ 时精度先有所下降,之后趋于稳定; c_6 类(140 个样本) 在 $N_K=3$ 时精度有所下降,之后趋于稳定. 图 16 中 NG 集中类别基本能趋稳定,c_4(590 个样本) 和 c_1(480 个样本) 有些波动; c_{14} 类(594 个样本) 精度稍有所下降.

综上所述,通过该文提出的方法选取边界后小类别文档基本能选入边界样本集合,且当 N_K 值达到一定值时精度基本能趋于稳定. 随着大类别文档的加入对某些小类别的精度有所影响,但总体来看还是比较稳定的,从而更进一步地说明本文所提出的方法能有效地选取边界样本,且对小类别样本效果更显著.

2.2.4 **对比实验** 为了进一步检验本文算法的有效性,将经典样本选择算法 CNN 和 FCNN 算法应用于文本分类,并与本文提出的算法进行对比. 表 7 为最后所选取的样本数; 表 8 为各方法进行样本选择后再进行 SVM 分类后的 F_1 值 . 其中 ALL 表示选取所有样本的方法,MIS 表示本文的提出的算法.

表 7 各方法所选文档数

类别	CNN	FCNN	MIS	ALL
FD	2 856	2 887	4 746	8 214
IT	2 825	6 164	3 543	7 583
NG	4 953	10 026	9 046	11 314

表 8 各方法选取样本后进行 SVM 分类的 F_1 值

类别		CNN	FCNN	MIS	ALL
FD	M_{acF_1}	0.650 4	0.648 4	0.665 3	0.659 2
	M_{icF_1}	0.853 6	0.855 1	0.868 6	0.866 4
IT	M_{acF_1}	0.749 4	0.753 6	0.762 5	0.757 4
	M_{icF_1}	0.882 1	0.887 9	0.889 4	0.890 0
NG	M_{acF_1}	0.722 3	0.740 3	0.739 3	0.741 3
	M_{icF_1}	0.726 8	0.745 1	0.744 2	0.746 4

从表 7 可以看出,除 FD 数据集外 CNN 选取的样本数最少,FCNN 选取的样本数最多,本文提取的方法选取的样本数处于 2 者之间. 从表 8 可以看出,经过选取样本后进行 SVM 分类的结果,本文提出的方法在不均衡文档集 FD 上性能是最佳的; 在多标签不均衡文档集 LT 上 M_{acF_1} 的值是最佳的,M_{icF_1} 的值与选取所有样本的值相当; 在均衡文档集 NG 上选取所有样本的性能是最佳的. 但从总体来看这些算法的性能区别不大,而本文提出的算法在不均衡文档集上性能更显著. 造成这种结果的原因: 一方面是由于本文的算法是从边缘样本出发,有针对性. 由于 CNN 算法样本的选择是依次进行的,所以越排在前面的样本比排在后面的样本被选择的可能性更大,使得选择的样本代表性差且可能含有冗余,而在不均衡文档集中小类别样本很有可能是排在后面的. FCNN 算法是从类的中心开始选择样本,而小类别与大类别的中心可能距离很近,从而造成错误的选择. 另一方面是因为本文的算法是分别从每个类别中选择样本,就能尽可能地将每个类的边界样本找出来,而不会忽略小类别. 从这 2 个方面就可以说明为什么本文的算法在不均衡文档集上性能更显著.

3 总结

为解决海量文本分类计算效率问题,缩小数据处理规模的样本选择是解决途径之一. 而大多数的样本选择算法并未应用于文本分类,且依赖于具体的分类器,本文针对这一问题提出了不依赖于任一分类算法的边界样本选择算法,该算法在标准文档集 Reuters-21578、复旦文档集和 20newsGroup 新闻组文档集上进行了实验,选取了边界样本. 选取样本后进行的 SVM 实验结果与所有样本进行实验的结果相当,甚至有所提高. 在 KNN 实验中选取样本后进行的实验结果与所有样本进行实验的结果相当,而在 NG 文档集中结果略低,这是因为均衡类中边界样本包含了大部分的分类信息,但对 KNN 算法来

说还是不够的,还须加入内部样本信息. 本文对经典的样本选择算法 CNN 和 FCNN 方法进行了对比实验,实验结果表明本文提出的算法取得了较好的结果,尤其是在不均衡文档集上. 从总的实验结果来看,本文提出的算法能有效地选取边界样本,并更进一步说明边界样本对分类的精度起主要作用.

4　参考文献

[1] Hart P E. The condensed nearest neighbor rule [J]. IEEE Transaction on Information Theory,1968,14(5):15-516.

[2] 李畅. 基于边界样本选择的支持向量机 [D]. 石家庄: 河北大学,2014.

[3] Gates G W. The reduced nearest neighbor rule [J]. IEEE Transactions on Information Theory,1972,18(3):431-433.

[4] Wilson D L. Asymptotic properties of nearest neighbor rules using edited data [J]. IEEE Transaction on Systems,Man and Cybernetics,1972,2(3):408-421.

[5] Angiulli F. Fast nearest neighbor condensation for large data sets classification [J]. IEEE Transactions on Knowledge and Data Engineering,2007,19(11):1450-1464.

[6] Tambouratzis T. Counter-clustering for training pattern selection [J]. The Computer Journal,2000,43(3):177-190.

[7] Lyhyaoui A,Ynez M M,Mora I. Sample selection via clustering to construct support vector-like classifiers [J]. IEEE Transactions on Neural Networks,1999,10(6):1474-1480.

[8] 杨宏晖,王芸,孙进才,等. 融合样本选择与特征选择的 AdaBoost 支持向量机集成算法 [J]. 西安交通大学学报,2014,48(12):63-68.

[9] Ramesh B,Sathiaseelan J G R. An advanced multi class instance selection based support vector machine for text classification [J]. Procedia Computer Science,2015,57:1124-1130.

[10] 周玉,朱安福,周林,等. 一种神经网络分类器样本数据选择方法 [J]. 华中科技大学学报: 自然科学版,2012,40(6):39-43.

[11] 胡小生,钟勇. 基于边界样本选择的支持向量机加速算法 [J]. 计算机工程与应用,2017,53(3):169-173.

[12] Yang Honghui,Zhou Xin,Wang Yun,et a1. A new adaptive immune clonal algorithm for underwater acoustic target sample selection [EB/OL]. [2017-03-11]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6718810.

[13] Anwar I M,Salama K M,Abdelbar A M. Instance selection with ant colony optimization [J]. Procedia Computer Science,2015,53(1):248-256.

[14] Marcin Blachnik. Ensembles of instance selection methods based on feature subset [J]. Procedia Computer Science,2014,35:388-396.

[15] Álvar Arnaiz-González,José-Francisco Díez-Pastor,Juan J Rodríguez,et al. Instance selection of linear complexity for big data [J]. Knowledge-Based Systems,2016,107(C):83-95.

[16] Sun Wei,Lin Aiping,Yu Hongshan,et al. All-dimension neighborhood based particle swarm optimization with randomly selected neighbors [J]. Information Sciences An International Journal,2017,405(C):141-156.

The New Boundary Sample Selection Method and Its Application in the Text Classification

WAN Zhongying,WANG Mingwen,ZUO Jiali,LIU Changhong

(School of Computer Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: On the premise of ensuring the classification performance,how to select an important sample set from a large number of training sample sets has become an important issue in the pattern classification. Aiming at this problem,a new sample selection algorithm is proposed and applied to text categorization. Experiments are carried out on the standard document set Reuters-21578,Fudan document set and 20 news group document set. The experimental results show that the proposed method can effectively select the boundary samples,and the SVM and KNN classifiers can get better classification results,especially on the unbalanced document set.

Key words: boundary samples; sample selection; text classification; SVM; KNN

(责任编辑: 冉小晓)