DOI:10.13245/j.hust.2013.s2.018

KNN 和 SVM 并行结合的算法

李胜东[1] 吕学强[2] 施水才[2] 石俊涛[1]

(1 廊坊燕京职业技术学院计算机工程系，廊坊 河北 065200；
2 北京信息科技大学网络文化与数字传播北京市重点实验室，北京 100101)

摘要 根据话题跟踪的定义和特点，分析了 K 最近邻(KNN)算法和支持向量机(SVM)算法的优缺点，发现它们的优缺点具有互补的可能性，提出了 KNN 和 SVM 并行结合的算法作为话题跟踪算法，设计了话题跟踪实验，实验结果证明了新算法作为话题跟踪算法，考虑了话题跟踪的特点，利用了 KNN 算法和 SVM 算法的理论优势而避免了理论的缺陷，处理话题跟踪问题时具有很好的话题跟踪效果.

关键词 支持向量机；K 最近邻；并行算法；话题跟踪；话题检测

中图分类号 TP391 **文献标志码** A **文章编号** 1671-4512(2013)S2-0113-04

KNN and SVM Parallel algorithm

Li Shengdong[1] *Lü Xueqiang*[2] *Shi Shuicai*[2] *Shi Juntao*[1]

(1 Department of Computer Engineering, Langfang Yanjing Polytechinc Institute, Langfang 065200, Hebei China; 2 Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract According to the definition and characteristics of topic tracking, the advantages and disadvantages of the K-nearest neighbor (KNN) algorithm and support vector machines (SVM) algorithm were analyzed. It is found that their advantages and disadvantages have the possibility of complementary, proposes parallel with KNN and SVM algorithm as topic tracking algorithm, designs topic tracking experiments, and experimental results prove that the new algorithm as the topic tracking algorithm considers the characteristics of the topic tracking, takes advantage of theoretical advantages of the KNN algorithm and SVM algorithm while avoiding their theoretical defect and has a good topic tracking effect in dealing with the problems of topic Tracking.

Key words support vector machines; K-nearest neighbor; parallel algorihm; topic tracking; topic detection

现在社会快速地跨入网络时代，由于网络信息量太大，与一个话题相关的信息往往孤立地出现在许多不同的地方，甚至出现在不同的时间. 如果仅仅通过这些孤立的信息，人们对某个话题就很难做到全面地把握. 在这种背景下，研究人员就开始关注一种新的技术，它就是话题跟踪技术. 这种技术就是追踪事件后继发展动态的信息智能获取技术，它可以使人们从整体上了解一个事件的全部细节以及该事件与其他事件之间的联系[1].

在话题检测与跟踪(TDT)研究中，话题跟踪研究在本质上等价于一种受监督的分类研究[2]，其关键技术是文本分类. 目前，分类效果最好且应用最广泛的文本分类算法[3-6]是 KNN 文本分类算法和 SVM 文本分类算法，但它们在话题跟踪研究中都不能完全满足要求，故本研究提出了一种新的话题跟踪算法.

收稿日期 2013-07-25.

作者简介 李胜东(1984-)，男，硕士研究生，E-mail：lsd_6@126.com.

基金项目 网络文化与数字传播北京市重点实验室开放课题资助项目(ICDD201105，ICDD201205)；国家自然科学基金资助项目(61271304)；北京市教委科技发展计划重点资助项目暨北京市自然科学基金 B 类重点资助项目(KZ201311232037)；2013 年河北省高等学校科学技术研究自筹资金资助项目(Z2013162).

1 话题跟踪的特点

a. 小规模训练语料. 话题跟踪提供给实验的训练语料非常少,通常是 N_t 个新闻报道($N_t=\{1,2,4\}$),这些语料都是关于同一个话题,话题跟踪的任务是希望在后继的新闻报道流中检测出所有关于该话题的新闻报道.

b. 时间特性. 输入话题跟踪实验的新闻报道按照时间排序,依次判定是否与给定话题相关;一个新闻报道出现以前,必须对它的前一个新闻报道做出判定. 为了满足这个特性,在每个新闻报道进入话题/报道表示模型前,先对所有报道按照时间排序,然后将这些符合时间要求的报道以队列的形式顺次进入话题跟踪算法.

这两个特点是话题跟踪的特点,也是话题跟踪与文本分类的本质区别.

2 KNN 和 SVM 算法

KNN 文本分类算法的基本思想[7]是:文本内容被形式化为特征空间中的加权特征向量. 对于一篇测试报道,计算它与训练报道集中每个报道的相似度,找出 K 个最相似的报道,根据加权距离和,判断测试报道所属的话题类别.

支持向量机的理论基础[8]是 VC 维和结构风险最小原理,即统计学习理论. 它的基本思想是使用简单的线性分类器划分话题/报道特征空间. 对于在当前特征空间中线性不可分的模式,则使用一个核函数把话题/报道特征空间映射到一个高维空间中,使得新闻报道能够线性可分.

3 KNN 和 SVM 并行结合的算法

根据 KNN 和 SVM 算法的基本思想,KNN 是基于传统的统计理论,它的特点是需要比较多的训练语料训练模型;SVM 算法是基于统计学习理论,它的特点是在小规模训练语料条件下能够得到全局最优解. 根据话题跟踪的定义和特点,为了充分利用 KNN 和 SVM 对训练语料的特点,提出了 KNN 和 SVM 并行结合的算法作为话题跟踪算法. 该算法先初始化训练报道向量的规模,设为 u(令 u 大于 K 最近邻值). 如果训练报道向量的实际规模 V 小于 u,算法认为训练语料比较少,不足以训练一个有效的 KNN 模型,但可以训练一个有效的 SVM 模型,故这个新算法调用 SVM 算法,处理测试报道向量;否则,这个新算法调用 KNN 算法,处理测试报道向量. 详细算法思想和步骤如下.

步骤 1 对于训练集和测试集中的报道,通过向量空间模型,得到训练集中的报道向量和测试集中的报道向量.

步骤 2 计算训练报道向量规模 V,测试报道向量规模 n,然后判断训练报道向量规模 V 是否大于系统初始化训练报道向量的规模 u,如果 $V>u$,跳到步骤 3;否则,跳到步骤 7.

步骤 3 对于每一个测试报道向量,根据定义 1,计算该测试报道与训练集中每篇报道的相似度.

步骤 4 根据报道相似度,在训练报道集中选出与该测试报道最相似的 K 篇报道.

步骤 5 根据定义 2,计算该测试报道在 K 最近邻文本中的权重.

步骤 6 比较 K 最近邻的权重,将报道分到权重最大的那个近邻所在的那个话题中,然后判断测试报道向量规模 n 是否等于 0,如果为 0,算法结束;否则,算法跳到步骤 2.

步骤 7 把训练集的报道向量映射到一个矩形区域.

步骤 8 将矩形区域划分成若干小区域,使每个小区域至多含有一个报道向量.

步骤 9 对每个包含报道向量的小区域边界,根据报道向量的类别标定方向,一种为顺时针方向,一种为逆时针方向.

步骤 10 合并矩形边界,同向线段保留,逆向线段抵消,获得若干个封闭折线组成的分类超平面,并以链的形式存储超平面.

步骤 11 输入测试集的报道向量,计算该报道向量关于以上分类曲线的围绕数,根据围绕数判定该报道向量所在的话题类别,然后判断测试报道向量规模 n 是否等于 0,如果为 0,算法结束;否则,算法跳到步骤 2.

4 实验

为了测试 KNN 和 SVM 并行结合的算法作为话题跟踪算法的性能,设计了两部分话题跟踪实验:第一部分实验测试 KNN 算法作为话题跟踪关键技术时的话题跟踪性能;第二部分实验测试 KNN 和 SVM 并行结合的算法作为话题跟踪关键技术时的话题跟踪性能.

在实验中,所采用的分词程序是中科院计算

所提供的ICTCLAS[7]；语料是中科院计算所提供的1.415 0×10⁴篇中文新闻报道文本文档[9-10]，共分两个层次：第一个层次是12个主题；第二个层次是60个话题；实验结果评测标准是话题检测与跟踪(TDT)评测方法[1].

4.1　实验结果

在实验1中，话题跟踪关键技术为KNN算法，特征空间维数为1 000. 通过调整K最近邻值，得到不同K最近邻值条件下的TDT评测结果. 本次试验使用60个话题进行测试，共测试了60次，然后按照每个主题进行评测，最终得到了12个TDT评测[1]结果，如表1所示.

表1　实验1中的TDT评测结果

主题	10	20	30	50	60
人才	0.210 0	0.218 1	0.070 7	0.072 6	0.077 0
体育	0.061 3	0.069 4	0.070 7	0.072 6	0.077 0
卫生	0.297 1	0.289 4	0.282 1	0.264 3	0.254 9
地域	0.652 6	0.640 0	0.678 7	0.712 2	0.779 2
环境	0.248 6	0.255 4	0.252 1	0.262 3	0.265 0
房产	0.090 4	0.101 6	0.108 6	0.102 9	0.109 4
教育	0.300 8	0.294 6	0.292 5	0.288 8	0.277 3
汽车	0.096 1	0.119 6	0.112 2	0.103 0	0.097 4
电脑	0.202 6	0.218 1	0.225 0	0.225 2	0.227 9
科技	0.596 7	0.587 7	0.592 1	0.576 5	0.588 5
艺术	0.700 9	0.702 7	0.719 2	0.730 9	0.741 4
金融	0.318 8	0.330 7	0.347 5	0.351 3	0.349 3

在实验1中，根据TDT评测结果评测话题跟踪性能，然后根据话题跟踪性能评估KNN算法作为话题跟踪关键技术的性能. TDT评测结果越小，说明话题跟踪性能越好，也表明KNN作为话题跟踪关键技术的性能越好. 根据表1，当K最近邻值为10时，平均的TDT评测结果最小，其值为0.3147. 这说明了KNN算法作为话题跟踪关键技术在K最近邻值为10处有最好的性能. 根据表1中的平均值，计算KNN算法的各种情况下的平均评测结果为0.323 0，即为该算法在处理话题跟踪问题时的综合性能.

在实验2中，话题跟踪关键技术为KNN和SVM并行结合的算法；$u=60$，若训练语料规模大于60，则该算法调用KNN算法($K=35$)，否则调用SVM算法；通过调整特征空间维数，得到不同特征空间维数(1 000，2 000，3 000，5 000，6 000)条件下的TDT评测结果. 本次试验使用60个话题进行测试，共测试了60次，然后按照每个主题进行评测，最终得到了12个TDT评测结果，如表2所示.

同理，根据表2，特征空间维数为6 000时，平均的TDT评测结果最小，其值为0.207 7，这说明了当特征空间维数为6 000时KNN和SVM并行结合的算法有最好的性能. 根据表2中的平均值，计算KNN和SVM并行结合的算法在各种情况下的平均评测结果为0.228 7，即为该算法在处理话题跟踪问题时的综合性能.

表2　实验2中的TDT评测结果

主题	1 000	2 000	3 000	5 000	6 000
人才	0.170 0	0.157 6	0.173 7	0.163 1	0.150 8
体育	0.024 1	0.014 3	0.018 8	0.016 1	0.013 6
卫生	0.199 1	0.188 1	0.179 9	0.179 2	0.168 7
地域	0.664 7	0.643 0	0.591 9	0.602 2	0.524 8
娱乐	0.190 4	0.179 2	0.169 5	0.171 4	0.167 7
房产	0.047 5	0.048 1	0.036 7	0.033 4	0.034 0
教育	0.210 4	0.205 6	0.202 4	0.189 3	0.190 2
汽车	0.071 7	0.065 6	0.066 0	0.054 0	0.051 5
电脑	0.129 6	0.123 0	0.096 2	0.085 4	0.082 5
科技	0.429 1	0.379 6	0.378 0	0.368 8	0.343 3
艺术	0.653 1	0.606 5	0.578 4	0.561 1	0.563 2
金融	0.248 6	0.236 9	0.217 8	0.203 7	0.202 7

4.2　实验分析

通过对比表1和表2，KNN和SVM并行结合的算法在处理话题跟踪问题时的最好评测结果(0.207 7)和平均评测结果(0.228 7)都比KNN算法的最好评测结果(0.314 7)和平均评测结果(0.323 0)小，即其处理话题跟踪问题时的最好性能和综合性能都比KNN算法好，这说明了基于统计学习理论的SVM算法弥补了基于传统统计理论的KNN算法不适宜处理小规模训练语料的缺陷，而KNN算法弥补了SVM算法不适宜处理大规模训练语料的缺陷，故以这两个算法为基础提出的KNN和SVM并行结合的算法作为话题跟踪算法，具有良好的话题跟踪性能.

参　考　文　献

[1] Nist. The 2004 topic detection and tracking (TDT2004) task definition and evaluation plan [EB/OL]. [2012-12-20]. http://www.itl.nist.gov/iad/mig/tests/tdt/2004/TDT04.Eval.Plan.v1.2.pdf.

[2] 夏迎炬，黄萱菁，胡恬，等. 自适应信息过滤中使用少量正例进行阈值优化[J]. 软件学报，2003，14(10)：1697-1705.

[3] Dasarathy B V. Nearest neighbor (NN) norms：NN pattern classification techniques [M]. Las Alamitos：IEEE Computer Press，1991：321-329.

[4] Rocchio J J. Relevance feedback in information retrieval [M]. SMART Retrieval System-Experi-

ments in Automatic Document Processing, 1971: 41-48.

[5] McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification[C]. Menlo Park: AAAI-98 Workshop, 1998: 41-48.

[6] Safavian S R, Landgreble D. A survey of decision tree classifier methodology [J]. IEEE Trans Systems, Man and Cybernetics, 1991, 21: 1-58.

[7] **魏登萍,王挺,王戟. 融合描述文档结构和参引特征的Web服务发现**[J]. **软件学报**, 2011, 22(9): 2006-2019.

[8] **何径舟,王厚峰. 基于特征选择和最大熵模型的汉语词义消歧**[J]. **软件学报**, 2010, 21(6): 1287-1295.

[9] **谭松波,王月粉. 中文文本分类语料库**-TanCorpV1.0 [EB/OL]. [2012-12-20]. http://www.searchforum.org.cn/tansongbo/corpus.htm.

[10] Tan Songbo, Cheng Xueqi, Ghanem M M, et al. A novel refinement approach for text categorization [C]// Proceedings International Conference on Information and Knowledge Management, 2005: 469-476.