软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2571−2583 [doi: 10.3724/SP.J.1001.2013.04467]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于聚类的 PU 主动文本分类方法*

刘 露[1,2], 彭 涛[1,2,3], 左万利[1,3], 戴耀康[1]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, USA)

[3](符号计算与知识工程教育部重点实验室(吉林大学),吉林 长春 130012)

通讯作者: 彭涛, E-mail: tpeng@jlu.edu.cn, taopeng@illinois.edu

摘 要: 文本分类是信息检索的关键问题之一.提取更多的可信反例和构造准确高效的分类器是 PU(positive and unlabeled)文本分类的两个重要问题.然而,在现有的可信反例提取方法中,很多方法提取的可信反例数量较少,构建的分类器质量有待提高.分别针对这两个重要步骤提供了一种基于聚类的半监督主动分类方法.与传统的反例提取方法不同,利用聚类技术和正例文档应与反例文档共享尽可能少的特征项这一特点,从未标识数据集中尽可能多地移除正例,从而可以获得更多的可信反例.结合 SVM 主动学习和改进的 Rocchio 构建分类器,并采用改进的 TFIDF(term frequency inverse document frequency)进行特征提取,可以显著提高分类的准确度.分别在 3 个不同的数据集中测试了分类结果(RCV1,Reuters-21578,20 Newsgoups).实验结果表明,基于聚类寻找可信反例可以在保持较低错误率的情况下获取更多的可信反例,而且主动学习方法的引入也显著提升了分类精度.

关键词: PU(positive and unlabeled)文本分类;聚类;TFIPNDF(term frequency inverse positive-negative document frequency);主动学习;可信反例;改进的 Rocchio

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 刘露,彭涛,左万利,戴耀康.一种基于聚类的 PU 主动文本分类方法.软件学报,2013,24(11):2571−2583. http://www.jos.org.cn/1000-9825/4467.htm

英文引用格式: Liu L, Peng T, Zuo WL, Dai YK. Clustering-Based PU active text classification method. Ruan Jian Xue Bao/ Journal of Software, 2013,24(11):2571−2583 (in Chinese). http://www.jos.org.cn/1000-9825/4467.htm

Clustering-Based PU Active Text Classification Method

LIU Lu[1,2], PENG Tao[1,2,3], ZUO Wan-Li[1,3], DAI Yao-Kang[1]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, USA)

[3](Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun 130012, China)

Corresponding author: PENG Tao, E-mail: tpeng@jlu.edu.cn, taopeng@illinois.edu

Abstract: Text classification is a key technology in information retrieval. Collecting more reliable negative examples, and building effective and efficient classifiers are two important problems for automatic text classification. However, the existing methods mostly collect a small number of reliable negative examples, keeping the classifiers from reaching high accuracy. In this paper, a clustering-based method for automatic PU (positive and unlabeled) text classification enhanced by SVM active learning is proposed. In contrast to traditional methods, this approach is based on the clustering technique which employs the characteristic that positive and negative examples should share as few words as possible. It finds more reliable negative examples by removing as many probable positive examples from unlabeled set as possible. In the process of building classifier, a term weighting scheme TFIPNDF (term frequency inverse

* 基金项目: 国家自然科学基金(60903098, 60973040)

收稿时间: 2013-02-28; 修改时间: 2013-07-16; 定稿时间: 2013-08-27

positive-negative document frequency, improved TFIDF) is adopted. An additional improved Rocchio, in conjunction with SVMs active learning, significantly improves the performance of classifying. Experimental results on three different datasets (RCV1, Reuters-21578, 20 Newsgroups) show that the proposed clustering- based method extracts more reliable negative examples than the baseline algorithms with very low error rates and implementing SVM active learning also improves the accuracy of classification significantly.

Key words: positive and unlabeled (PU) text classification; clustering; TFIPNDF (term frequency inverse positive-negative document frequency); active learning; reliable negative example; improved Rocchio

随着网络信息量的迅速增长和万维网信息提取技术的出现,自动文本分类技术已经成为数据发掘领域一项重要的任务.文本分类可以预测一篇文档的类别,并且在现实生活中有着很重要的应用,如过滤垃圾邮件[1,2]、网页分类[3]、提供个性化新闻[4]和用户意图分析[5,6]等等.

在半监督分类中,训练集由少量有标识的正例文档和大量未标识的文档组成.PU 学习技术可以节省收集和标记反例的工作,其目标就是利用未标识的文档来提高分类器的性能.本文就是基于这个目标,分别在两步分类中提出新方法,寻找到更准确的超平面来进行自动文本分类.

在提取反例的过程中,我们采用基于聚类方法来收集可信反例.显然,可信反例就是那些与给定主题不相关的文档.也就是说,它们应该与正例共享尽可能少的特征.聚类是一个在数据分析中常用的机器学习方法.它可以用来聚集相似的文档.受到这个启发,我们首先将正例中的文档聚集成一些小的聚类,然后根据一个事先定义的类半径 r_P,用这些小的聚类来聚集未标识数据集中的文档.在正例集中聚类的目的是提高在未标识文档中聚类的效率.最终,未标识数据集中没有被聚类的文档可被视为可信反例加入可信反例集.采用聚类方法收集反例在我们的非主动学习技术中也取得了良好的效果.本文在构建分类器的过程中提出了一个基于主动学习的方法来进行自动文本分类.首先,应用改进的 Rocchio 算法来构建初始分类器;进而,应用 SVM 主动学习提供一个自动 PU 文本分类的新视角.主动学习研究一个闭环现象,即应该选择什么数据被加入训练集[7].构建分类器时,有少量标识数据和大量未标识数据,当标识的数据被正确选择时,几乎所有的未标识数据都可以被正确地标识.然而,只有当一些未标识的数据不能被改进的 Rocchio 识别或者标识的结果与 SVM 分类冲突时,才由一个人工标识者或者领域专家来手工标识这篇文档.迭代判断不确定性最大的文档类别并更新之前生成的分类器,从而得到最终更精确的分类器.由于 SVMs 随机选择测试数据,主动学习启发式搜索合适的训练数据,这可以使文本分类更加快速、准确.另外,我们采用了 3 个不同的数据集(RCV1,Reuters-21578,20 Newsgroups)作为标准测试数据集,来展示主动学习应用于 PU 文本分类带来的出色效果.

实验结果表明,与现有的可信反例获取方法相比,基于聚类提取可信反例的方法可以在较低错误率的前提下得到更多的可信反例.并且,主动学习的引入也显著提升了分类器的性能.

本文第 1 节介绍 PU 文本分类的相关工作.第 2 节描述基于聚类提取可信反例的方法.第 3 节讨论如何基于主动学习构造 PU 文本分类器.第 4 节介绍实验的环境,分析结果.第 5 节对全文进行总结.

1 相关工作

PU 文本分类过程主要包含两步:第 1 步,从未标识数据集(用 U 标识)中收集可信反例(也称 RN);第 2 步,使用正例集(用 P 标识)、可信反例集、未标识数据集来迭代构建分类器.国内外很多研究人员对 PU 文本分类问题进行了研究,并获得了很多研究成果,如 S-EM 算法[8],PEBL[9],Roc-SVM[10],Biased-SVMs[11],Weighted Logistic Regression[12],one-class SVMs[13],PSOC[14],PE-PUC[15],SPUL[16]等等.此外,主动学习和迁移学习[17−20]也被应用到 PU 学习中,并且取得了很好的效果.

文献[8]总结了 PU 分类并提出了 S-EM 算法,并结合贝叶斯和 EM 算法来构建分类器,在构建分类器的过程中,试图最大化未标识数据集中可信反例的数量,然而,此时反例集中会包含很多正例文档,由此导致分类器精度的下降.Yu 等人介绍了 PEBL 算法来分类网页[9],然而,生成的可信反例数量太少以至于不能得到最好的分类器.Li 等人结合了 Rocchio 方法和 SVMs 来构建分类器[10],由于 Rocchio 方法没有达到很高的准确度,收集反例的准确度也不是很高.文献[11]中提出了 Biased-SVM 方法,生成的分类器也没有达到很高的精度.Pan 等人提出

了动态分类器(DCEPU)来分类 PU 文本[21].

用于 S-EM 中的 spy 技术[22]将 P 中随机的一组正例加入到 U 中.NB 技术使用朴素贝叶斯分类器收集 U 中的反例[11].Rocchio-SVM 方法用 Rocchio 算法收集可信反例,这种传统的方法未能取得很好的效果[23].它通过构建一个原型向量来构建分类器.1-DNF 作为另一种传统的可信反例收集方法,往往收集到的可信反例较少,导致分类器效果并不理想[9].文献[24]中采用 KNN 算法将未标识数据集中的文档按照其与 k 个最近正例文档的距离进行排序,并且设置一个阈值,未标识数据集中相似度比阈值低的文档被视为可信反例.文献[25]中也介绍了一种基于聚类收集可信反例的方法,他们通过聚类正例和未标识集合中的文档,并通过判断一个聚类中正例的比例来判断这个聚类能否成为可信反例集,但是效率并不高.

Settles 在文献[26]中提出了一些主动学习的分析、关于如何设置变量的总结和一些机器学习相关主题的讨论.Nissim 等人将主动学习作为一种选择性取样的方法来提高分类器的性能[27].Joshi 等人将主动学习应用于多类别分类来处理训练瓶颈,使多类别图像分类更容易[28].Ji 和 Han 在主动学习中描述了一个新的变量最小化的视角并应用于图结构,其中没有特征表示和标签信息[29].Platt 使用一个逻辑连接函数和规范化最大似然值来直接训练核分类器[30].Schohn 等人提供一种启发式方法来启发式地选择训练数据[7].

2 基于聚类收集可信反例

在这一节中,我们针对从 U 集中收集可信反例,提出了一种新方法.可信反例是与给定主题不相关的,也就是说,这些文档应该与正例文档共享尽可能少的特征.聚类是一个聚集相关文档的传统方法,传统的聚类方法可以大致分为划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法等.在同一簇集中的数据或者文档被看作是高度相关的,因此我们应用聚类来收集更多的可信反例.首先,正例集中的文档用合并(自下而上)聚类聚集.为了获得 P 和 U 中的聚类,我们在公式(1)和公式(2)中定义类中心 O_P、类半径 r_P 如下:

$$O_P=\frac{\sum_{i=1}^{m}x_i}{m} \tag{1}$$

$$r_P=\frac{r\times\varphi(m)}{\varphi(m)+1},\ r=\arg\max_{x_k\in P}d(x_k,O_P) \tag{2}$$

其中,我们假定正例集为 $X=\{x_1,x_2,\ldots,x_m\},x_i\in P.m$ 是所有用于聚类的文档总数.$\varphi(m)=\lg(m)+1$,用于调整半径的大小,使其能够得到尽可能多且纯正的反例.整个算法包含两步,即聚类和选择.在聚类过程中,对于每个正例集中的文档,我们首先计算出类中心 O_j 与每个文档距离类中心的距离.根据这个距离,每篇文档可以被分到已存在的类中或者被加入一个新的类中.在选择过程中,首先计算未标识数据集中的每篇文档与 P 中每个聚类的距离.若一篇文档与一个类中心的距离小于半径 r,则这篇文档从未标识集合中移除.选择过程结束后,未标识集合中剩余的文档被当作可信反例集.与传统方法不同,基于聚类提取可信反例并不直接提取反例文档.它从未标识集合中移除尽可能多的可能正例.算法描述如算法 1 所示,其中,第 5 行~第 12 行为寻找类中心过程,第 14 行~第 20 行为提取反例过程.若文档与类中心的距离小于 r_p,则被视为可能的正例,即该文档可以从可信反例集合 RN 中删除.该算法时间复杂度为 $O(mn)$,空间复杂度为 $O(n)$.

算法 1. 基于聚类收集可信反例算法.

输入:P:正例数据集,U:未标识数据集.

输出:RN:可信反例集.

1. **procedure Clustering-Based Method for Collecting Reliable Negative Examples**
2. Assume the positive example set be $X=\{x_1,x_2,...,x_m\}$, $x_i\in P$
3. $O_P\leftarrow\frac{\sum_{i=1}^{m}x_i}{m},r\leftarrow\arg\max_{x_k\in P}d(x_k,O_P),r_P\leftarrow\frac{r\times\varphi(m)}{\varphi(m)+1}$
4. $C_1\leftarrow\{x_1\}$, $O_1\leftarrow x_1$, $numCluster\leftarrow 1$, $z\leftarrow\{x_2,x_3,\ldots,x_m\}$

5. **repeat**
6. **select** one $x_i \in z$ and **find out** the closest O_j to x_i, $O_j \leftarrow \arg\min_{j}^{numCluster}_{k=1} d(x_i, O_k)$
7. **if** $d(x_i, O_j) < r_P$ **then**
8. **add** x_i to class c_j and adjust the center of class j, $O_j \leftarrow \frac{n_j \cdot O_j + x_i}{n_j + 1}$ and $n_j \leftarrow n_j + 1$
9. **else**
10. **add** a new class $C_{numCluster} \leftarrow \{x_i\}$, $numCluster \leftarrow numCluster + 1$, $O_{numCluster} \leftarrow x_i$
11. **end if**
12. **until** z is empty
13. $RN \leftarrow U$
14. **for** $i \leftarrow 1$ to $numCluster$
15. **for** each $x_i \in RN$
16. **if** $d(x_i, O_i)$ **then**
17. $RN \leftarrow RN - \{x\}$
18. **end if**
19. **end for**
20. **end for**
21. **end procedure**

算法 1 基于聚类提取可信反例.$\varphi(m)=\lg(m)+1$.$numCluster$ 表示到目前为止生成的类数量.m 是被应用到聚类当中的文档总数.$C_1, C_2, \ldots, C_m$ 是最终的类.O_j 是类 C_j 的类中心.n 是类 C_j 中的元素个数.

3 改进的 Rocchio 与 SVM 主动学习迭代构建分类器

在构建分类器之前,本文采用一种新的特征权值计算方法 TFIPNDF(term frequency inverse positive-negative document frequency)[31].TFIPNDF 权值计算方法是对 TFIDF(term frequency inverse document frequency)的一种改进方法.为了更好地说明权值计算过程,定义一篇文档 d_i 被表示为 $d_i=\{t_1:x_{i1},\ldots,t_j:x_{ij},t_m:x_{im}\}$.TFIDF 的权值公式为 $TFIDF = f_{ik} \times \log\left(\frac{N}{n_i}\right)$,其中,$f_{ik}$ 表示特征 i 在文档 k 中出现的频率,N 是集合中文档的总数,n_i 表示特征 i 在训练样本集合中出现的文档数.TFIDF 权值计算方法反映了两个思想:(1) 某个特征在一篇文档中出现的次数越多,与文档的主题就越相关;(2) 某个特征出现在越多不同类别的文档中,这个特征就越不能区分不同主题的文档.TFIDF 可以反映一个特征在整个文档集中的重要性,但却没有分别考虑其在正例集和反例集中的区别,无疑是 TFIDF 策略的一个缺陷.如,给定一个训练集包含 20 篇文档(10 篇正例文档和 10 篇反例文档),特征 t 出现在 10 篇文档中.根据 TFIDF 给出的计算公式,即 n_i=10,反向文档频率(IDF)为 $\log\left(\frac{20}{10}\right)$.考虑两种不同的情况:一种是特征 t 在正例集和反例集中分别出现 5 次;另一种是特征 t 在正例中出现 9 次,在反例中出现 1 次.显然,该特征反映了在正例集和反例集中不同的重要性.由此,我们采用一种新的权值计算方法 TFIPNDF.该方法考虑了特征在正例集和反例集中不同的分布情况,弥补了 TFIDF 算法的缺陷,其计算公式如下:

$$TFIPNDF = \begin{cases} f_{ik} \times \frac{P_i}{S_P} \times \log\left(\frac{N}{n_i}\right), & \text{文档} k \in P \\ f_{ik} \times \frac{N_i}{S_N} \times \log\left(\frac{N}{n_i}\right), & \text{文档} k \in RN \end{cases} \tag{3}$$

其中,f_{ik} 表示特征 i 在文档 k 中出现的频率,N 是集合中文档的总数,n_i 表示特征 i 在训练样本集合中出现的文档

数,P_i 表示特征 i 在正例集中出现的次数,N_i 表示特征 i 在反例集中出现的次数,S_P 和 S_N 分别是正例集和反例集的文档总数.

一个数据集中包含不同长度的文档.我们对 TFIPNDF 的权值进行归一化,公式如下:

$$x_{ik}=\begin{cases}\dfrac{f_{ik}\times\dfrac{P_i}{S_P}\times\log\left(\dfrac{N}{n_i}\right)}{\sqrt{\sum_{r=1}^{M}\left[f_{rk}\times\dfrac{P_r}{S_P}\times\log\left(\dfrac{N}{n_r}\right)\right]^2}}, & \text{文档}k\in P\\ \dfrac{f_{ik}\times\dfrac{N_i}{S_N}\times\log\left(\dfrac{N}{n_i}\right)}{\sqrt{\sum_{r=1}^{M}\left[f_{rk}\times\dfrac{N_r}{S_N}\times\log\left(\dfrac{N}{n_r}\right)\right]^2}}, & \text{文档}k\in RN\end{cases} \tag{4}$$

其中,M 是所有特征的数量.在预处理、特征提取和权值计算等一系列工作结束后,我们接下来介绍如何构建分类器.

在 Rocchio 分类器中[32],构造一个原型向量 $\vec{c}_j$ 来总结类 j 中的文档.每个待分类的文档需要与每个类的原型向量计算相似度,这个待分类的文档很有可能属于这个相似度最大的类.原型向量如下构造:

$$\vec{c}_j=\alpha\frac{1}{|c_j|}\sum_{\vec{d}\in c_j}\frac{\vec{d}}{\|\vec{d}\|}-\beta\frac{1}{|D-c_j|}\sum_{\vec{d}\in D-c_j}\frac{\vec{d}}{\|\vec{d}\|} \tag{5}$$

其中,$\|\vec{d}\|$ 代表向量 $\vec{d}$ 的欧几里得距离,D 代表训练集,c_j 是识别出的相关文档的集合,α和β这两个参数用来调整识别相关文档和不相关文档的效果.文献[33]已经给出了两个参数的合理取值,将α和β分别设置为 16 和 4.

大量的距离度量方法已经被提出来,其中最成功的是余弦函数.因此,我们采用余弦函数来计算原型向量和待分类文档向量的相似度.计算公式如下:

$$H_{Rocchio}(\vec{x})=\arg\max_{c_j\in c}\cos(\vec{c}_j,\vec{x})=\arg\max_{c_j\in c}\frac{\vec{c}_j}{\|\vec{c}_j\|}\cdot\frac{\vec{x}}{\|\vec{x}\|} \tag{6}$$

Rocchio 算法的一个缺点是它不能处理一篇文档不属于任何一个给定类别的情况.也就是说,我们需要对 $\cos(\vec{c}_j,\vec{x})$ 设定一个阈值.本文对 Rocchio 算法进行了改进,旨在不仅得到类 c_j 的原型向量,而且得到类半径 R_j.一旦 $\vec{x}$ 和 $\vec{c}_j$ 的距离高于类半径 R_j,待分类的文档就属于类 c_j,也就是 $\vec{x}\in c_j$. 类半径 R_j 的计算方法如下:

$$R_j=\arg\max_{d\notin c_j}\cos(\vec{c}_j,\vec{d})=\arg\max_{d\notin c_j}\frac{\vec{c}_j}{\|\vec{c}_j\|}\cdot\frac{\vec{d}}{\|\vec{d}\|} \tag{7}$$

改进 Rocchio 方法后,确定待分类文档类别的公式如下:

$$H_{Rocchio}(\vec{x})=\arg\max_{c_j\in c}\cos(\vec{c}_j,\vec{x})=\arg\max_{c_j\in c}\frac{\vec{c}_j}{\|\vec{c}_j\|}\cdot\frac{\vec{x}}{\|\vec{x}\|}\ \&\&\ \cos(\vec{c}_j,\vec{x})\geqslant R_j \tag{8}$$

如果没有一个类 c_j 满足上述公式的要求,待分类的文档就不能被分类器正确地识别.也就是说,它不属于任何一个类别.

传统的分类器往往尽最大的努力来减少分类错误[14,21,34,35].主动学习,也称为查询学习或者最优实验设计,是机器学习的一个分支,更常说的是人工智能的一个分支[26].一种学习算法如果允许启发式地选择它要学习的数据,则往往会用更少的训练次数得到更好的效果.本文中,主动学习的实质就是启发式地选择训练数据,并且重新分类不确定性较大的文档.也就是说,如果数据集中的一篇文档不能被改进的 Rocchio 算法识别,或者它的类别不匹配 SVMs 的分类结果,那么就由一个手工注释者或者一个领域专家来手动标识这篇文档.这个交互的过程可以极大地提高分类的准确度.在主动学习的过程中,文档被交互评估,从而得到一个新的分类器.与我们在 PRL 中的工作相比,本文的文本分类过程克服了 1-DNF 及 1-DNFC 的缺点,利用了聚类方法的特点,得到了更加精确的分类器.

Schölkopf 等人从超平面角度定义了支持向量机[36],且支持向量机已被出色地应用在很多研究领域[37−39].支持向量机通过寻找最大页边距设计一个合理的超平面,从而分类向量空间中的数据.每个样本和超平面的距离是判断这个样本类别的关键.

在这一节中,我们也描述在主动学习过程中如何选择训练数据.从支持向量机的原理中可以看出,在向量空间 S 中,用于分类的超平面只与支持向量 sv 相关.给出样本 x 和超平面 h_i 的距离 $d(x,h_i)$,选择样本的策略就是找到一个样本 x_0,使得它和超平面的距离最小,即 $d(x_0,h_i)=\min\{d(x,h_i)\},x\in S$.也就是说,这个样本应该尽可能地离超平面近一些.主动学习采样的收敛性已经在文献[40]中得到证明,本文只考虑支持向量机的二元分类情况.主动学习分类算法描述如算法 2 所示,其时间复杂度为 $O(n)$,空间复杂度也为 $O(n)$.

算法 2. 主动学习算法.

输入:训练集 T(P:正例数据集,RN:可信反例集,U:未标识数据集),ns:取样中样本数量.

输出:分类器 f.

1. **procedure Active Learning**
2. Initialize training sample set I_0, $I_0\leftarrow Sub_p\cup Sub_n$ //Sub_p(Sub_n) is the subset of P(RN)
3. $T_0\leftarrow T-I_0$, $i\leftarrow 1$
4. **repeat**
5. $b_i\leftarrow\varnothing$
6. **find** the optimal classification hyperplane h_i of training sample set I_{i-1} based on SVMs
7. $sv\leftarrow$support vector of hyperplane h_i
8. $dp\leftarrow$the distance between the support vector of positive samples and sv
9. $dn\leftarrow$the distance between the support vector of negative samples and sv
10. **choose** ns samples from T_{i-1} which are closest to h_i and $d(x,h_i)<dp$ and $d(x,h_i)<dn$ /* $x\in T_{i-1}$, $d(x,h_i)$ is the distance between x and h_i */
11. $b_i\leftarrow$the ns samples
12. label the ns samples correctly
13. **for** each $x\in T_{i-1}-b_i$
14. $g(x)=\sum_l y_l a_l k(x_l,x)-b$ /* a_l is a coefficient, y_l is a classification label (1,−1), x_l is a support vector, $k(x_l,x)$ is the kernel function of SVMs. b is a classification threshold */
15. $h_i(x)=\begin{cases}+1, \text{ when } (|g(x)|\geqslant dp \ \&\& \ g(x)>0) \text{ or } (g(x)>0 \ \&\& \ H_{Rocchio}(x)=+1)\\ -1, \text{ when } (|g(x)|\geqslant dn \ \&\& \ g(x)<0) \text{ or } (g(x)<0 \ \&\& \ H_{Rocchio}(x)=-1)\end{cases}$
16. **if** x can not be identified by improved Rocchio or its category does not match the result of SVMs, **then**
17. **label** x manually
18. $b_i\leftarrow b_i\cup x$
19. **end if**
20. **end for**
21. $I_i\leftarrow I_{i-1}\cup b_i$, $T_i\leftarrow T_{i-1}-b_i$
22. $i\leftarrow i+1$
23. **until** T_i is empty or b_i is empty
24. $f\leftarrow h_i$
25. **return** f
26. **end procedure**

4 实验与结果

在这一节中,我们使用本文提到的技术构建了一个 PU 文本分类系统,并且在 3 个测试集上测试了我们的方法.首先,我们将在第 4.1 节定义度量标准以评估所提出的方法.然后将在第 4.2 节中介绍 3 个数据集.为了测试我们方法的有效性,第 4.3 节将介绍用来参照的算法.最终,所有的实验结果将在第 4.4 节中给出.

4.1 度量标准

为了评估 PU 文本分类器的性能,两个最常用的度量标准(准确率和召回率)被用来直接反映系统的可行性.文本分类中的准确率是被正确分类的正例文档数量占被分类为正例的文档数量的比例.这个度量标准用来衡量系统是否可以很好地剔除不相关类别的文档.召回率,也称为查全率,是被正确分类的正例文档数量占实际上正例文档数量的比例.这个度量标准反映了系统找全所有相关文档的能力.我们假定 T 是测试集中所有类别相关文档的集合,C 是被分类器判断为类别相关文档的集合.因此,我们用公式定义准确率和召回率如下:

$$precision=\frac{|C\cap T|}{|C|}\times 100\% \tag{9}$$

$$recall=\frac{|C\cap T|}{|T|}\times 100\% \tag{10}$$

由于这两个标准并不相关,用两个标准评估不同的分类器是困难的.因此,采用准确率和召回率的调和平均值 F-Measure[41]来作为分类器的评估标准.

根据我们对准确率和召回率重要性的不同需求,定义 F-Measure 值如下:

$$F\text{-}Measure=\frac{(\beta^2+1)precision\times recall}{\beta^2\times precision+recall} \tag{11}$$

其中,β是一个反映准确率和召回率相对重要程度的权值.显然,若β>1,则召回率的重要性大于准确率;反之亦然.本文将β设定为常量 1.

我们也定义了错误率 ERR,从找到的可信反例数量和错误率的角度,将我们的方法与 CPUE[25]、改进的 1-DNF[14]和 1-DNF[9]进行比较.假定 $RN(P_t)$是可信反例中的正例集合,P_t 是未标识数据集中正例的集合,错误率 ERR 如下计算:

$$ERR=\frac{|RN(P_t)|}{|P_t|} \tag{12}$$

4.2 数据集

本文将在 3 个数据集上测试提出的方法:Reuters Corpus Volume 1(RCV1),Reuters-21578 和 20 Newsgroups.下面将分别介绍这 3 个数据集.

RCV1 数据集是目前使用最广泛的服务于文本分类的标准数据集.它是来自于路透社的新闻语料库,收集了 804 414 篇文档,用于训练集和测试集.该测试集按树形层次分布,根节点包含 3 个大类别:topic,industry 和 region.我们选取其中一个 topic 子集进行实验.在 topic 这一层中包含 4 个分支:CCAT(corporate/industrial),ECAT(economics),GCAT(government/social)和 MCAT(market).在这 789 670 篇文档中,本文只在每类中采用 3 000 篇文档,其中 70%被用作训练集,剩余的 30%被用作测试集.

Reuters-21578 数据集包含从路透社新闻组收集的 21 578 篇文档.在这 135 个主题 21 578 篇文档中,我们选取其中 10 个类别 9 980 篇文档作为本文的测试集和数据集.仍将每类 70%作为训练集,30%作为测试集,具体的数量见表 1.

20 Newgroups 是由 Ken Lang 等人收集的数据集,包含 20 000 篇新闻文档,分布在 20 个类别中.本文只选取其中的 10 个类共 9 840 篇文档进行实验.另外,我们使用 20news-bydate.tar.gz 版本的数据集作为我们实验的训练集和测试集.训练集和测试集的数量已经由提供者提前确定.具体每类的文档数量见表 2.

Table 1 Number of documents of each topic in Reuters-21578

表 1 Reuters-21578 中每个类别文本数量

类别	数量
Acq	2 369
Corn	237
Crude	578
Earn	3 964
Grain	582
Interest	478
Money	717
Ship	286
Trade	486
Wheat	283

Table 2 Number of documents of each topic in 20 Newsgroups

表 2 20 Newsgroups 中每个主题文本数量

类别	数量
comp.graphics	973
comp.os.ms-windows.misc	985
comp.sys.ibm.pc.hardware	982
comp.sys.mac.hardware	963
comp.windows.x	988
sci.crypt	991
sci.electronics	984
sci.med	990
sci.space	987
soc.religion.christian	997

4.3 参照的方法

许多 PU 分类算法和可信反例提取方法被应用于自动文本分类,为了评估本文提出的文本分类方法的有效性,我们分别为两步分类方法选取了不同的参照方法.在提取可信反例阶段,我们选取了 CPUE、改进的 1-DNF 和 1-DNF 算法,从提取可信反例数量和错误率两个角度对本文提出的方法进行评估.在构造分类器阶段,我们分别选取高效的 PSOC[14],PEBL[9]和 OCS(one-class SVM)[36]这 3 种算法作为参照算法.

4.4 结 果

为了评估我们方法的性能,首先从收集到的反例数量和错误率的角度测试本文提出方法的性能.表 3~表 5 列出了 3 个数据集中分别用基于聚类提取反例方法、CPUE、改进的 1-DNF 和 1-DNF 提取反例的数量和相应的错误率.根据文献[14],我们将改进的 1-DNF 中的参数λ设置为 0.2.结果表明,基于聚类提取可信反例的数量明显高于其他 3 种算法.虽然 1-DNF 算法可以得到最低的错误率,但是它获得了最少的可信反例,不能达到很好的效果.基于聚类提取可信反例的算法在错误率上明显好于 CPUE、改进的 1-DNF 和 1-DNF.因此结果表明,本文提出的基于聚类提可信反例的方法可以在较低错误率的前提下识别出更多的反例.

Table 3 Number of reliable negative examples and the error rates of four kinds of methods in Reuters Corpus Volume 1

表 3 Reteurs Corpus Volume 1 中 4 种方法收集的反例个数和相应的错误率

类别	1-DNF		改进的 1-DNF		CPUE		聚类	
	RN	ERR (%)	RN	ERR (%)	RN	ERR (%)	RN	ERR (%)
CCAT	354	0.29	1 596	1.19	1 898	1.07	2 245	0.72
ECAT	301	0	1 956	0.89	2 115	0.78	2 447	0.45
GCAT	374	0.47	2 539	1.47	2 819	1.34	3 308	0.96
MCAT	311	0	2 416	1	2 666	0.89	3 193	0.65

Table 4 Number of reliable negative examples and the error rates of four kinds of methods in Reuters-21578

表 4 Reteurs-21578 中 4 种方法收集的反例个数和相应的错误率

类别	1-DNF		改进的 1-DNF		CPUE		聚类	
	RN	ERR (%)	RN	ERR (%)	RN	ERR (%)	RN	ERR (%)
Acq	176	0	773	1.09	881	1.04	1 257	0.68
Corn	218	0	785	1.79	897	0.92	1 376	0.67
Crude	121	0	456	0	535	0.85	880	0.59
Earn	214	0.16	835	0.30	926	0.27	1 281	0.18
Grain	144	0	806	0	971	0.84	1 360	0.44
Interest	276	0	1 143	0	1 296	1.03	1 707	0.75
Money	258	0	1 220	0	1 332	2.05	1 584	1.16
Ship	376	1.57	1 430	1.77	1 578	1.76	1 900	1.62
Trade	113	0	379	0	509	0	831	0
Wheat	222	0	917	0	1 056	0	1 396	0

Table 5 Number of reliable negative examples and the error rates of four kinds of methods in 20 Newsgroups

表 5 20 Newsgroups 中 4 种方法收集的反例个数和相应的错误率

类别	1-DNF		改进的 1-DNF		CPUE		聚类	
	RN	ERR (%)	RN	ERR (%)	RN	ERR (%)	RN	ERR (%)
comp.graphics	117	0	439	1.79	678	1.72	906	1.11
comp.os.ms-windows.misc	111	0	412	1.22	703	1.09	1 110	0.81
comp.sys.ibm.pc.hardware	120	0	593	1.78	783	1.59	1 041	0.92
comp.sys.mac.hardware	125	0.64	590	1.82	698	1.35	1 354	0.67
comp.windows.x	115	0	523	1.22	750	0.99	948	0.73
sci.crypt	101	0	412	1.24	637	1.12	801	0.81
sci.electronics	126	0.56	561	1.85	716	1.77	1 077	1.58
sci.med	117	0	390	1.36	543	1.18	1 171	0.76
sci.space	118	0	419	1.75	602	1.61	1 169	0.88
soc.religion.christian	121	0	389	1.80	550	1.64	960	1.17

一个 PU 文本分类器的性能在很大程度上取决于它能否以低错误率获得尽可能多的可信反例.在 1-DNF 算法中,该算法只考虑一个特征项在正例文档集和未标识文档集中出现的频率之差,而没有考虑其在正例文档中本身的频率.例如,我们想要搜索关于“game”的文档,而特征“blue”在正例文档集中出现的频率是 0.3%,但在未标识文档集中出现的频率是 0.1%.显然,“blue”不是一个正例特征,这与 1-DNF 算法得出的结果冲突.在这种情况下,正例特征会大量增加,相应地,反例数量会大量减少,甚至为 0.改进的 1-DNF 算法对 1-DNF 算法进行了改进.该算法不仅考虑了特征在正例文档集中和未标识文档集中的频率差,而且考虑了特征在正例文档集中本身的频率.也就是说,只有当同时满足两个条件时,一个特征项才能被看作正例特征项:

- 首先,该特征在正例文档集中出现的频率大于其在未标识文档集中出现的频率;
- 其次,该特征在正例文档集中出现的频率应大于一个固定的阈值.

CPUE 通过同时对正例文档集和未标识文档集中的文档进行聚类获得可信反例,这个过程是相对耗费时间和影响分类器准确度的.其错误率增加的原因在于,所有正例文档被直接用于与未标识文档集进行聚类.许多未标识文档被错误地当作非可信反例来对待.相应地,该算法会将可信反例的数量减少很多.本文提出的提取可信反例的方法在文档和类中心之间定义了一个半径,并通过移除未标识文档中尽可能多的“可能正例”来收集可信反例.未标识文档集中剩余的文档被当作可信反例.从收集的可信反例数量和错误率的角度来看,本文提出的方法获得了较好的性能.

为了验证 TFIPNDF 算法的有效性,我们采用不同的权值计算方法来比较 TFIPNDF 算法给分类效果带来的影响.如图 1 所示,通过观察比较平均 *F*-Measure 值在 3 个不同数据集中每个类别上的区别,我们发现,TFIPNDF 算法优于传统的 TFIDF 算法.

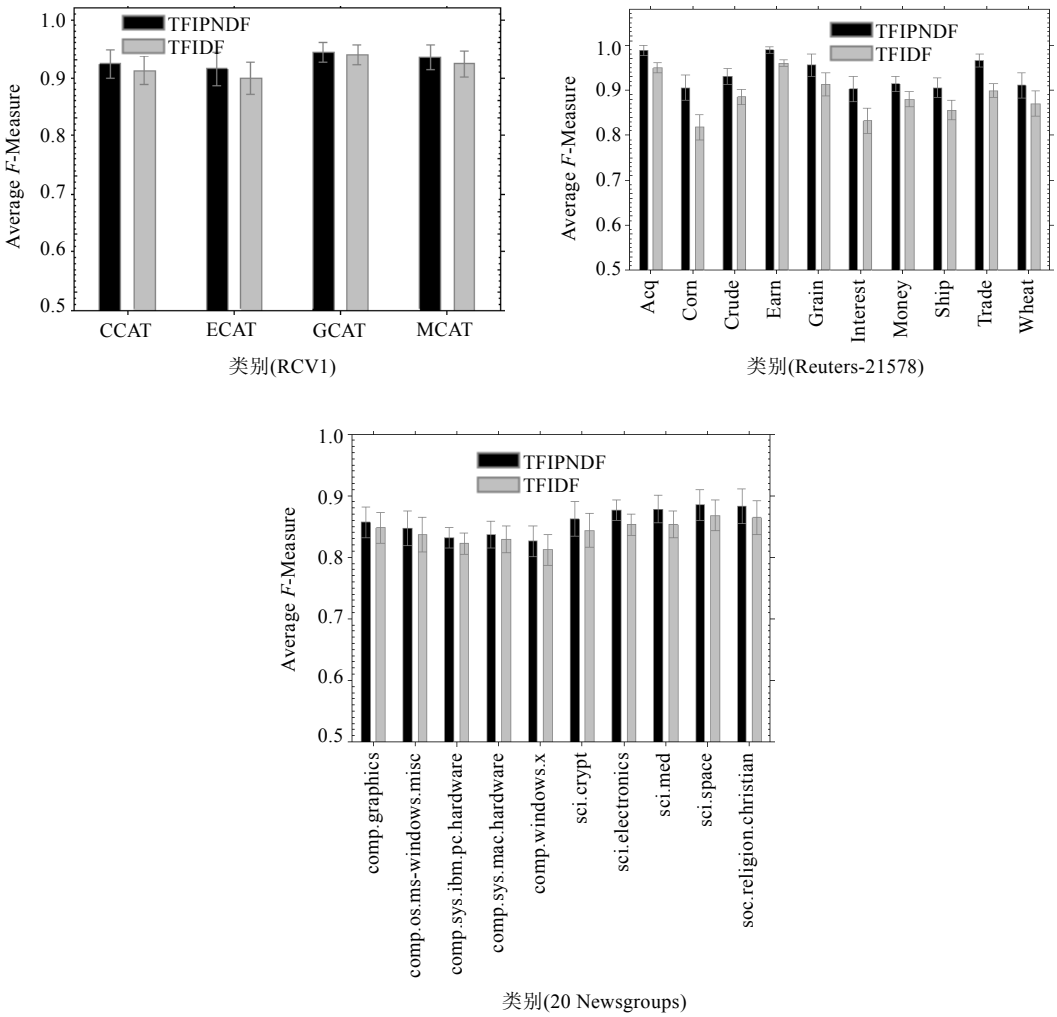

Fig.1 Comparison of average *F*-Measures achieved by our clustering-based method using TFIPNDF and TFIDF weightings for each dataset

图 1 基于聚类的方法分别采用 TFIPNDF 和 TFIDF 在每个数据集中获得的平均 *F*-Measure 值的比较

对本文提出的方法最重要的评估就是 SVM 主动学习对分类效果带来的影响.我们分别用 3 种算法(PSOC, PEBL 和 OCS)和我们提出的主动学习分类方法在 3 个不同数据集上进行测试,比较分类效果.为了清楚地表现实验结果,图 2 用 *F*-Measure 值动态地描述了本文方法在不同数据集上的表现.实验结果表明,本文提出的主动学习分类方法的平均 *F*-Measure 值明显高于其他 3 种方法.

在主动学习算法中,我们首先选取了距离超平面最近的 *ns* 个样本进行标记.由于它们距离超平面最近,所以最容易标记.也就是说,分类器更容易确定其类别.与之前经典的分类算法相比,本文提出的方法不仅引入了启发式的方法来标记文本,减少了标记时间,增加了结果的准确度,而且领域专家的引入也可以大大增加分类器的准确性.

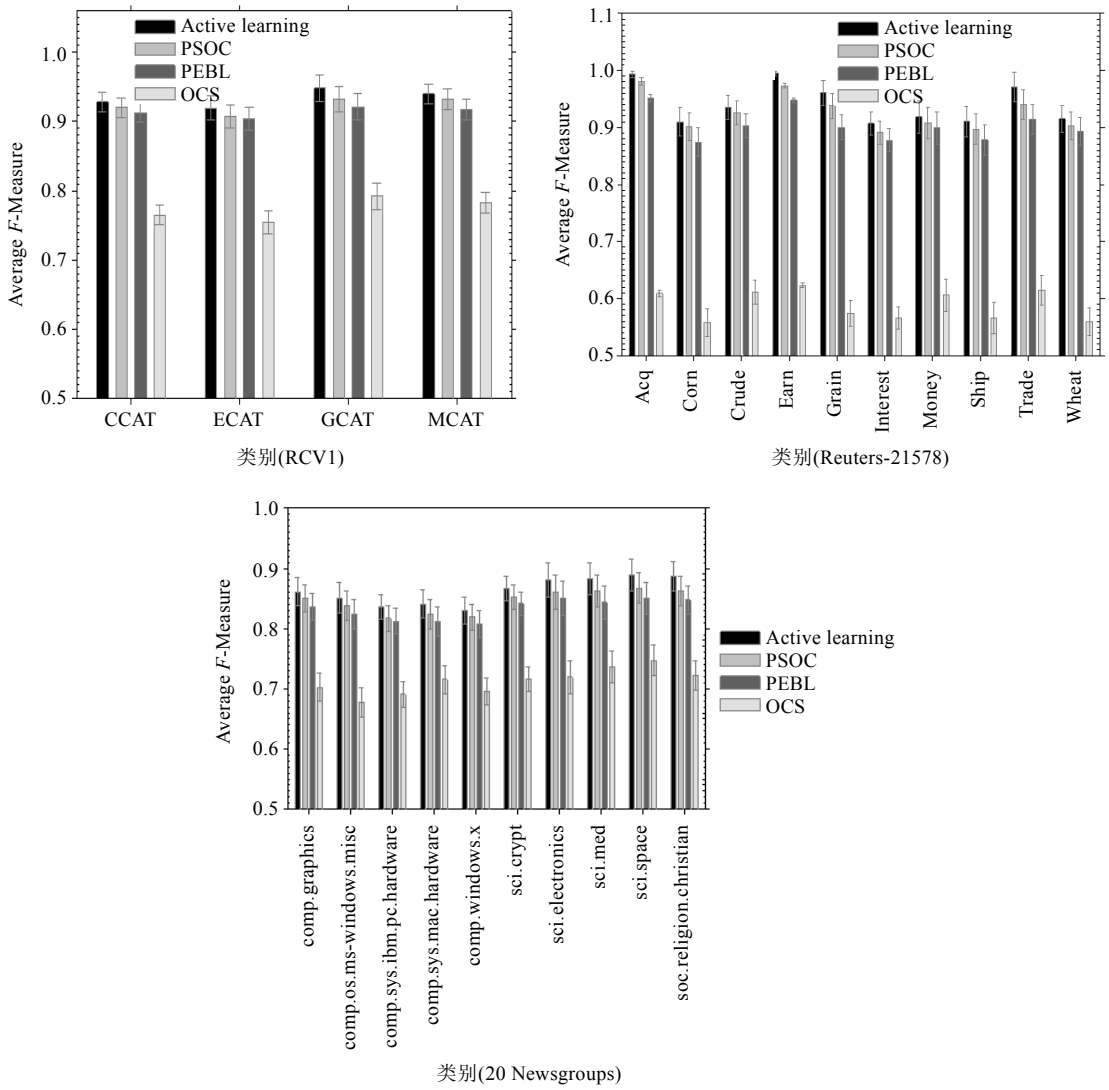

Fig.2 Performance of four text classifying methods for each topic on three datasets

图 2 4 种分类方法在 3 个数据集中每一类的分类表现

5 结 论

本文提出了一种基于聚类的提取可信反例方法和基于主动学习构建分类器的方法.传统的收集可信反例的方法都是直接从未标识集合中提取可信反例,如 1-DNF 和改进的 1-DNF 等.CPUE 方法虽然也使用基于聚类的方法收集可信反例,但未能达到很高的效率和准确率.本文提出的方法从未标识数据集中尽可能多地移除可能的正例,剩下的文档被看作可信反例.这个方法既获得了更多的可信反例,也保持了很低的错误率和高准确率.在构建文本分类器时,首先通过定义类半径这个新概念来改进 Rocchio 分类器,进而 SVM 主动学习方法被应用于构建分类器过程中.由于反复迭代识别不确定性较大的文档,从而使分类器准确率得到显著提高.实验结果表明,本文提出的构建分类器的方法比许多优秀分类器的准确性更高.因此,本文的方法是可行而有效的.

致谢 NIST 授权并提供了 RCV1 数据集,对本文实验部分有很大的帮助,在此表示感谢.

References:

[1] Liu W, Wang T. Online active multi-field learning for efficient email spam filtering. Knowledge and Information Systems, 2012, 33(1):117−136. [doi: 10.1007/s10115-011-0461-x]

[2] Fumera G, Pillai I, Roli F. Spam filtering based on the analysis of text information embedded into images. Journal of Machine Learning Research, 2006,7:2699−2720.

[3] Qi XG, Davison BD. Web page classification: Feature and algorithms. ACM Computing Surveys, 2009,41(2):Article 12. [doi: 10.1145/1459352.1459357]

[4] Anotonellis I, Bouras C, Poulopoulos V. Personalized news categorization through scalable text classification. Frontiers of WWW Research and Development-APWEB, Lecture Notes in Computer Science, 2006,3841:391−401. [doi: 10.1007/11610113_35]

[5] Hu M, Liu B. Mining and summarizing customer review. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2004. 168−177. [doi: 10.1145/1014052.1014073]

[6] Kim S, Hovy E. Determining the sentiment of opinions. In: Proc. of the Int'l Conf. on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004. [doi: 10. 3115/1220355.1220555]

[7] Schohn G, Cohn D. Less is more: Active learning with support vector machines. In: Proc. of the 17th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, Inc., 2000. 839−846.

[8] Liu B, Lee WS, Yu PS, Li XL. Partially supervised classification of text documents. In: Sammut C, Hoffmann AG, eds. Proc. of the 19th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, Inc., 2002. 387−394.

[9] Yu H, Han JW, Chang KCC. PEBL: Positive example based learning for Web page classification using SVM. In: Proc. of the Knowledge Discovery and Data Mining. New York: ACM, 2002. 239−248. [doi: 10.1145/775047.775083]

[10] Li XL, Liu B. Learning to classify texts using positive and unlabeled data. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, Inc., 2003. 587−592.

[11] Liu B, Dai Y, Li XL, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Proc. of the 3rd IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2003. 179−186. [doi: 10.1109/ICDM.2003.1250918]

[12] Lee WS, Liu B. Learning with positive and unlabeled examples using weighted logistic regression. In: Proc. of the 20th Int'l Conf. on Machine Learning. 2003. 448−455.

[13] Manevitz LM, Yousef M. One-Class SVMS for document classification. The Journal of Machine Learning Research, 2001,2: 139−154.

[14] Peng T, Zuo WL, He FL. SVM based adaptive learning method for text classification from positive and unlabeled documents. Knowledge and Information Systems, 2008,16(3):281−301. [doi: 10.1007/s10115-007-0107-1]

[15] Yu S, Li CP. PE-PUC: A graph based PU-learning approach for text classification. Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, 2007,4571:574−584. [doi: 10.1007/978-3-540-73499-4_43]

[16] Xiao YS, Liu B, Yin J, Cao LB, Zhang CQ, Hao ZF. Similarity-Based approach for positive and unlabeled learning. In: Walsh T, ed. Proc. of the 22nd Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2011. 1577−1582.

[17] Wu J, Lu MY. Asymmetric semi-supervised boosting scheme for interactive image retrieval. ETRI Journal, 2010,32(5):766−776. [doi: 10.4218/etrij.10.1510.0016]

[18] Li ZM. Li L, Liu YJ, Bao JW. An improved method for support vector machine-based active feedback. In: Proc. of the 2008 3rd Int'l Conf. on Pervasive Computing and Applications, Vol.1. 2008. 389−393. [doi: 10.1109/ICPCA.2008.4783617]

[19] Zhou ZH, Chen KJ, Dai HB. Enhancing relevance feedback in image retrieval using unlabeled data. ACM Trans. on Information Systems, 2006,24(2):219−244. [doi: 10.1145/1148020.1148023]

[20] Sheng LY, Ortega A. Graph based partially supervised learning of documents. In: Proc. of the 2011 IEEE Int'l Workshop on Machine Learning for Signal Processing. 2011. 1−6. [doi: 10.1109/MLSP.2011.6064566]

[21] Pan SR, Zhang Y, Li X. Dynamic classifier ensemble for positive unlabeled text stream classification. Knowledge and Information Systems, 2012,33(2):267−287. [doi: 10.1007/s10115-011-0469-2]

[22] Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 2nd ed., Heidelberg: Springer-Verlag, 2011.

[23] Cooley R, Mobasher B, Srivastava J. Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1999,1(1):5−32. [doi: 10.1007/BF03325089]

[24] Zhang BZ, Zuo WL. Reliable negative extracting based on KNN for learning from positive and unlabeled examples. Journal of Computers, 2009,4(1):94−101. [doi: 10.4304/jcp.4.1.94-101]

[25] Zhang BZ, Zuo WL. A novel reliable negative method based on clustering for learning from positive and unlabeled examples. In: Proc. of the AIRS 2008. LNCS 4993, Heidelberg: Springer-Verlag, 2008. 385−392. [doi: 10.1007/978-3-540-68636-1_37]

[26] Settles B. Active learning literature survey. Technical Report, 1648, University of Wisconsin-Madison, 2010.

[27] Nissim N, Moskovitch R, Rokach L, Elovici Y. Detecting unknown computer worm activity via support vector machines and active learning. Pattern Analysis and Application, 2012,15(4):459−475. [doi: 10.1007/s10044-012-0296-4]

[28] Joshi AJ, Porikli F, Papanikolopoulos NP. Scalable active learning for multiclass image classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012,34(11):2259−2273. [doi: 10.1109/TPAMI.2012.21]

[29] Ji M, Han JW. A variance minimization criterion to active learning on graphs. In: Proc. of the 15th Int'l Conf. on Artificial Intelligence and Atatistics (AISTATS). 2012. 556−564.

[30] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Proc. of the Advances in Large Margin Classifiers. Cambridge: MIT Press, 1999. 61−74.

[31] Peng T, Liu L, Zuo WL. PU text classification enhanced by term frequency-inverse document frequency-improved weighting. Concurrency and Computation: Practice and Experience, Published Online: 10 MAY 2013. [doi: 10.1002/cpe.3040]

[32] Denis F, Gilleron R, Tommasi M. Text classification from positive and unlabeled examples. In: Proc. of the Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU). 2002.

[33] Buckley C, Salton G, Allan J. The effect of adding relevance information in a relevance feedback environment. In: Croft WB, van Rijsbergen CJ, eds. Proc. of the Int'l ACM SIGIR Conf. New York: Springer-Verlag, 1994. 292−300. [doi: 10.1007/978-1-4471-2099-5_30]

[34] Chen L, Guo G, Wang K. Class-Dependent projection based method for text categorization. Pattern Recognition Letters, 2011, 32(10):1493−1501. [doi: 10.1016/j.patrec.2011.01.018]

[35] Mesleh AM. Feature sub-set selection metrics for arabic text classification. Pattern Recognition Letters, 2011,32(14):1922−1929. [doi: 10.1016/j.patrec.2011.07.010]

[36] Schölkopf S, Platt J, Shawe J, Smola A, Williamson R. Estimating the support of a high-dimensional distribution. Technical Report, MSR-TR-99-87, Microsoft Research, 2001.

[37] Bouguila N. Hybrid generative discriminative approach for proportional data modeling and classification. IEEE Trans. on Knowledge and Data Engineering, 2012,24(12):2184−2202. [doi: 10.1109/TKDE.2011.162]

[38] Anguita D, Ghio A, Oneto L, Ridella S. In-Sample and out-of-sample model selection and error estimation for support vector machine. IEEE Trans. on Neural Networks and Learning Systems, 2012,23(9):1390−1406. [doi: 10.1109/TNNLS.2012.2202401]

[39] Yu HF, Hsieh CJ, Chang KW, Lin CJ. Large linear classification when data cannot fit in memory. ACM Trans. on Knowledge Discovery from Data, 2012,5(4):Article 23. [doi: 10.1145/2086737.2086743]

[40] Park JM, Hu Y. On-Line learning for active pattern recognition. IEEE Signal Processing Letters, 1996,3(11):301−303. [doi: 10.1109/97.542161]

[41] Croft WB, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice. Boston: Addison Wesley, 2009.

刘露(1989－),女,辽宁大连人,硕士生,主要研究领域为数据挖掘,Web 挖掘,信息检索,机器学习.
E-mail: liulu0804@gmail.com

左万利(1957－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据挖掘,Web 挖掘,机器学习,自然语言处理,信息检索.
E-mail: wanli@jlu.edu.cn

彭涛(1977－),男,博士,副教授,主要研究领域为数据挖掘,Web 挖掘,机器学习,自然语言处理,信息检索.
E-mail: tpeng@jlu.edu.cn

戴耀康(1989－),男,硕士生,主要研究领域为数据挖掘,Web 挖掘,信息检索.
E-mail: mrdyk@126.com