一种基于半监督学习的短文本分类方法

张 倩 刘怀亮
(西安电子科技大学经济与管理学院 西安 710071)

【摘要】针对短文本的特征词较少、信息关联性不强以及存在大量样本的标注瓶颈问题,传统的文本分类方法已不能较好地直接适用。将半监督学习思想引入到文本分类过程中,提出一种基于半监督学习的短文本分类方法,通过使用外部网络知识库来扩充短文本特征,构建基于半监督学习的分类模型,使用初始分类器进行迭代自学习实现训练样本中未标注部分的充分利用,从而解决标注瓶颈,提高分类器的性能。对比实验表明,该方法能够提升短文本分类的效果。

【关键词】半监督学习 文本分类 短文本 自训练
【分类号】TP391.1

An Algorithm of Short Text Classification Based on Semi-supervised Learning

Zhang Qian Liu Huailiang
(School of Economics and Management, Xidian University, Xi'an 710071, China)

【Abstract】According to the characteristics of short texts and the bottleneck problem of annotation in dealing with large numbers of unlabeled samples, traditional algorithms of text classification can not be used directly. This paper introduces a method of short text classification based on semi-supervised learning and builds a semi-supervised classification model. It is feasible to accomplish the self-training of the training samples and takes full advantages of the unlabeled parts of training texts by using the initial classifier. The bottleneck problem of annotation is solved and the good performance of classifier is shown. The contrast experiment shows that the algorithm of short text classification based on semi-supervised learning can get better classified effect.

【Keywords】Semi-supervised learning Text classification Short text Self-training

1 引 言

随着即时通信和互联网技术的不断发展,网络中的信息每天都在以一定的速率增长,网络生活中最常见的如:微博及其评论、聊天记录、手机短消息、科技文献摘要、搜索引擎返回的结果和社区论坛中的发帖回复等形式的短文本信息,其中可能隐藏着有价值的信息内容,因而对短文本进行有效的组织分类是非常有必要的。短文本所包含的形式多样,通常是指控制在160字左右的文本,经常以口语化、生活化的不规则形式出现,特征词较少且词与词之间的信息关联性较弱,然而目前较为常用的传统文本分类方法,如基于统计的方法:Naïve Bayes、K-Nearest Neighbor、类中心向量法、回归模型、支持向量机和最大熵模型等[1];基于连接的方法:人工神经网络;基于规则的方法:决策树法和关联规则等[2];这些方法大多是以长文本作为研究对象。考虑到短文本与长文本的不同特点,直接使用传统的方法会在很大程度上影响文本分类的效果[3]。另外,传统的文本分类方法需要对大量的已

收稿日期:2013-01-27
收修改稿日期:2013-02-12

标注样本进行学习训练,人工标注大量无标记短文本的难度较大且耗时耗力。而半监督学习可在已标注样本较少的情况下,结合大量未标注样本进行综合学习来构建性能良好的分类器,从而解决标注瓶颈问题,这在理论与实践上都具有一定的意义。

因此,本文将半监督学习的思想引入文本分类中,改进地使用维基百科对短文本进行特征扩展[4],提出一种基于半监督学习的文本分类方法来实现对短文本的有效分类。该方法通过使用外部知识资源库对短文本进行信息扩充以解决特征稀疏等问题,构建基于监督学习的中间过程初始分类器,经过不断迭代对训练样本中的未标注部分进行半监督自训练,用更新过的训练样本集来构建最终的分类器,达到充分利用大量的未标注样本来提高分类器性能的目的。

2　相关理论研究现状

2.1　短文本分类研究

短文本是近年来网络中较为常见的文本信息形式,存在数量很多,篇幅较短小,含有的特征词较少,词与词之间的关联性不强,对海量的这类文本进行基于语料库的人工标注十分困难,易形成标注瓶颈,利用传统的方法对此类文本进行分类不能达到理想的分类效果,因此已不能满足短文本分类的需要[3]。

国内外的学者分别通过扩展 Web 语义核函数[5]、构建资源描述框架来表示结构化概念网络[6],从而实现对短文本的分类。蔡月红等[7]通过使用属性选择技术构建半监督分类模型。同时,也有学者将注意力集中到使用或构建本体等来进行短文本的特征词扩展,通过扩充短文本的语义信息来弥补短文本特征稀疏的问题[8-10]。史伟等[11]对微博进行抽取和情感分析,从而建立模糊情感本体。但由于引入或构建的外部资源库可能存在适用领域范围小、可扩展性较差等因素,很难适应互联网发展的进度。文献[4]和文献[12]提出通过引入具有覆盖领域知识面较广的在线百科全书来对短文本进行分类和聚类,扩充了短文本的信息,但该类方法没有考虑到在已标注样本较少时的学习问题以及训练样本中未标注部分的价值。传统文本分类与基于对训练样本集进行扩展的短文本分类工作流程如图 1 所示,本文采用图 1(b)的思想对短文本进行特征扩展。

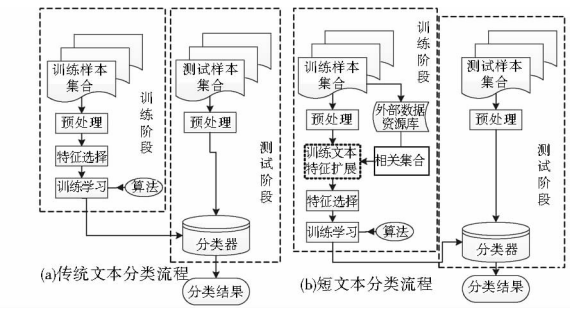

图 1　传统文本分类流程与短文本分类流程

2.2　半监督学习研究

从已有的分类研究成果可以看出,大多数方法是基于监督学习的[4,8,9],虽然已取得较好的效果,但是为了保证分类器具有良好的泛化能力,这类方法的使用需要以存在大量已标注语料为前提。如图 2 所示,基于监督学习的传统分类方法主要是通过对训练样本集中的已标注部分进行学习训练,推导出相应的关系模型,再利用模型对测试样本进行预测判断。该方法忽略了样本集合中数量颇为丰富的未标注部分的存在价值。而半监督学习分类方法的主要思想则是通过将已标注样本和未标注样本综合利用来进行分类器的训练,既保证了训练速度又可以提高分类的效果。

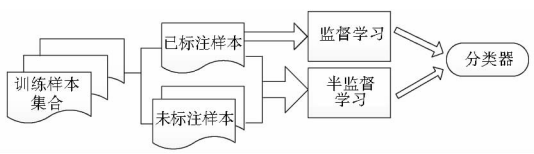

图 2　传统监督学习与半监督学习对比

半监督学习的相关研究起始于 20 世纪中期,在解决自然语言处理中关于文本分类相关问题的过程中,随着利用未标注数据 EM 算法的提出[13]与理论框架的成功构建[14],Shahshanani 等[15]较早开始了对半监督学习的研究工作。发展至今,半监督理论已经在许多领域得到应用,如:网页分类、文本聚类、标签生成、数字图像检索等,是机器学习中重要的研究部分[16]。目前,国内外研究中关于半监督分类学习算法主要包括:基于生成式模型的方法[17]、基于自学习和协同训练的方法[18]、密度变化的方法[19]、直推式支持向量机的方法[20]和基于图的方法[21]等。其中,基于自训练的半监督学习方法可以在具有少量标注数据且无误标记样本的情况下进行,并达到较好的训练效果,操作实现简单。Nigam 等[22]通过使用 EM 算法对未经标注的文本

进行学习,构建半监督文本分类模型。张博锋等[23]引入中间结果提出基于自训练的半监督学习算法 STEM,提高分类器学习计算效率。文献[24]则是借助切边权重统计方法对自标记的样本集进行修正。本文就是基于自训练的方法来进行半监督学习。

3 用于短文本分类的半监督学习算法

3.1 基本思想

由于短文本存在缺少描述信息、特征较为稀疏以及因存在大量未标注样本而造成的标注瓶颈等缺陷,直接使用传统文本分类方法已不能满足短文本分类的需要,因此对此类文本需要进行预先处理。一些研究提出通过引入特征扩展的方式来解决这类问题,文献[4]借助外部网络知识库维基百科进行概念抽取并建立语义概念集合以实现测试样本的语义特征扩展,再用基于监督学习的传统文本分类算法构建分类器实现文本分类,但该方法没有考虑到已标注数据较少时存在的学习问题。本文通过使用基于维基百科的方法对训练样本进行语义特征扩展,并结合半监督学习的思想提出一种用于短文本分类的半监督学习方法。与文献[4]不同之处在于,本文是通过扩展训练样本集的语义特征,并限制选取扩展特征词的个数,以减少扩展后引入过多噪声而造成的效果不明显,然后使用基于自训练的半监督学习方法,充分利用未标注样本改进分类器性能。

3.2 短文本特征扩展

按照图 1(b)所示的短文本分类流程,为实现文本的特征扩展,需要从外部知识库中将页面描述内容与类别之间的相关关系提取出来,并使用统计规律和分类信息将其量化,建立语义相关概念集合,然后将该集合作为语义信息基础对短文本进行扩展,并计算扩展后特征词的权重。本文使用维基百科来构建语义相关概念集合[4],对训练样本的文本特征词进行语义扩展,重新计算特征词的权重并合并相同的特征词,将权重值进行排序后选取前 λ 个对应的特征词作为关键词代表,从而将文本表示成一个新的二元组形式。

3.3 基于半监督学习的分类算法

(1)基于自训练的半监督学习

自训练(Self - Training)算法[25]是半监督学习中比较常见的方法之一,首先对少量的已标注样本进行监督学习训练,然后再将数量较多的未经标注的样本加入到训练所得的初始分类器中进行预测,预测得出的数值越大代表分类取得的置信度越高,将置信度较高的文本连同其分类标注一起加入到训练集中作为新的训练样本集进行再学习,迭代训练直到满足条件为止。自训练在方法的使用上具有以下特点:

①封装性

该方法通过对同一集合中的未标注数据进行自我预测,实现对已标注数据集的扩大,不断迭代直到所有未标注数据更新为带有标注的数据为止,自始至终是一个围绕自我内部进行迭代重复学习的过程。

②开放性

从宏观的角度,自训练的方法可以被看成一个学习的框架,通过选择不同的分类方法嵌入使用,能够比较方便地应用到不同情况中从而更好地解决问题,体现了此方法的开放性,本文就是将传统分类算法 KNN 与半监督 Self - Training 算法结合使用的。

需要注意的是,自训练方法在迭代过程中,如果初始的训练样本集中已标注样本数量过少,则可能会出现错误标注,并通过迭代使错误逐渐被放大,最终导致错误累积。文献[24]和文献[26]通过使用切边权重统计和对原始特征空间进行重采样来降低数据噪声。

(2)改进的基于 K 近邻自训练短文本分类算法

为了方便描述,给定类别(标签)集合 $C=\{c_1,c_2,\cdots,c_{|c|}\}$,训练样本集合 R 由已标注样本集合 L 和未标注样本集合 U 组成。

$$\begin{cases}\text{已标注样本集合 } L=\{(d_1,l_1),(d_2,l_2),\cdots,(d_l,l_l)\},l_i\in C\\ \text{未标注样本集合 } U=\{d_{l+1},d_{l+2},\cdots,d_{l+u}\}\\ S=l+u,\text{通常 } l\ll u\end{cases}$$

对短文本进行基于半监督的学习,是为了充分利用大量存在于训练样本中的未标注部分,通过迭代实现已标注样本集的不断扩大,提高分类器的准确率。本文根据半监督学习的理论和 Zhu 等[27]提出的最近邻(KNN 算法)进行改进,计算文本间的相似度。采用 KNN 算法生成的初始分类器对未标注部分进行学习时,要从未标注样本集合 U 中随机抽取 ξ 个样本组成未标注样本子集,逐步进行自训练学习和预测。本文使用的半监督分类学习算法流程如下:

Input:已标注样本集 L,未标注样本集 U,测试样本集 T

Output:测试样本集 T 的各个标记

①对训练样本集合 R 中的已标注部分 L 和未标注部分

U 进行语义特征扩展。

②从 U 中随机抽取 ξ 个样本组成子集 U_ξ，使用传统的 KNN 算法对已标注样本集合 L 进行训练，得到一个初始的中间分类器 Classifier 1。

③使用 Classifier 1 对 U_ξ 执行预测，找出 K 个近邻，将属于同一个类别的文本相似度相加求和并用相似度和较大的那个类的类标对其进行标注。

④将已经确定标注的子集 U_ξ 连同其类标记加入到已标注样本集合 L 中生成新的已标注样本集合 L'，并且从 U 中删除 U_ξ，即为 $L' = L + U_\xi$，$U' = U - U_\xi$。

⑤迭代循环步骤② - 步骤⑤，直到满足条件 U = Ø 时停止。

⑥迭代结束后，使用更新生成的 L' 来训练最终的分类器。

算法流程如图 3 所示：

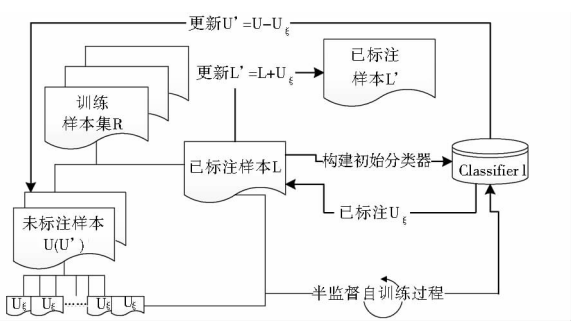

图 3　半监督学习算法流程

4　基于半监督学习的短文本分类流程描述

本文提出的基于半监督学习的短文本分类方法基本流程如图 4 所示：

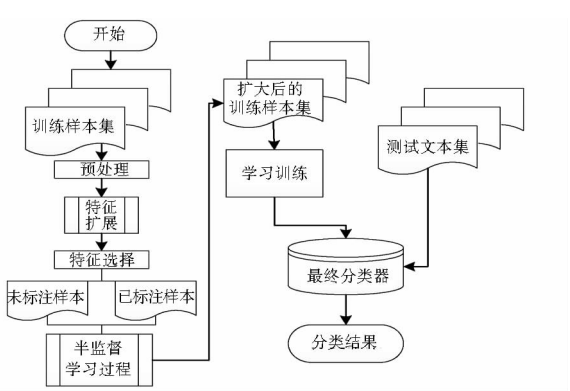

图 4　基于半监督学习的短文本分类流程

(1)对短文本进行预处理，即分词和去停用词。将文本用一组原始特征词的词条 $d = \{t_1, t_2, \cdots, t_n\}$ 进行表示，通过计算特征词权重对所有训练样本中的文本用一个二元组表示特征向量，即 $d_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \cdots, (t_{im}, w_{im})\}$，$t_i$ 表示第 i 个文本中的特征词，w_i 表示词的权重，由 TF - IDF 权值公式进行计算，这样就将文本表示成结构化的形式。

(2)对训练样本中已标注样本与未标注样本进行特征扩展，形成新形式的特征向量，并计算扩展特征词的权重，合并相同的项后选取权重值排序中前 λ 个来表示文本 $d_i = \{(V_{i1}, W_{i1}), (V_{i2}, W_{i2}), \cdots, (V_{i\lambda}, W_{i\lambda})\}$，其中 $V_{i\lambda}$、$W_{i\lambda}$ 分别表示扩展后的特征词及其权重。在此基础上对训练集使用半监督学习方法，利用少量的已标注样本构建初始分类器，再结合随机抽取的一定数量的未标注样本一起进行自学习，迭代过程中将经过中间分类器标注的未标注样本与其标注一起加入到已标注样本集中，并将其从未标注样本中删除，迭代结束后会得到一个更新后的扩大版训练样本集，最后再对这个新的训练样本集进行学习得到最终的分类器，并对测试样本进行分类，输出分类结果。

5　实验及其结果分析

5.1　实验准备

本文实验中的概念集合采用文献[4]中所提取的相关概念，使用由搜狐、新浪、腾讯等各大网站的帖子评论中搜集的短文本构建的语料库，包含经济、娱乐、教育、体育、医疗、计算机和社会 7 个类别，平均每个类别有 2 000 个语料。随机从每个类别中抽取 20% 作为固定的测试样本 T，剩余的 80% 语料作为训练样本(L ∪U)，随机抽取其中 10 000 个语料作为未标注样本集，其余作为已标注样本来训练初始分类器。K 值取 15，ξ 取 100。采用准确率(Precision)、召回率(Recall)和 F_1 值进行分类效果评估。

5.2　实验结果分析

实验分为三组：实验一采用传统的文本分类方法对语料进行分类，即直接使用 KNN 算法；实验二使用文献[4]的方法对文本进行分类；实验三使用本文提出的基于半监督学习的方法进行分类。实验的开始阶段需要利用中国科学院计算技术研究所研制的 ICTCLAS 软件对语料进行分词、去停用词处理，特征选择采用信息增益的方法。实验结果如表 1 所示。

表 1　实验结果对比

类别	实验一(%)			实验二(%)			实验三(%)		
	准确率	召回率	F_1	准确率	召回率	F_1	准确率	召回率	F_1
经济	63.721	67.380	65.499	69.983	70.210	70.096	70.825	71.484	71.153
娱乐	60.250	69.300	64.459	76.890	70.308	73.452	79.921	72.300	75.920
体育	79.728	75.050	77.318	79.328	78.392	78.857	84.143	85.250	84.693
教育	67.322	71.183	69.199	68.466	71.063	69.740	67.321	71.216	69.214
医疗	64.671	67.432	66.023	67.643	69.355	68.488	70.157	71.382	70.764
计算机	75.243	76.824	76.025	84.243	79.710	81.914	87.634	82.211	84.836
社会	60.139	62.200	61.152	62.783	63.681	63.229	67.221	72.934	69.961
平均值	67.296	69.910	68.525	72.762	71.817	72.254	75.317	75.254	75.220

根据表 1 的数据绘制折线图如图 5 所示：

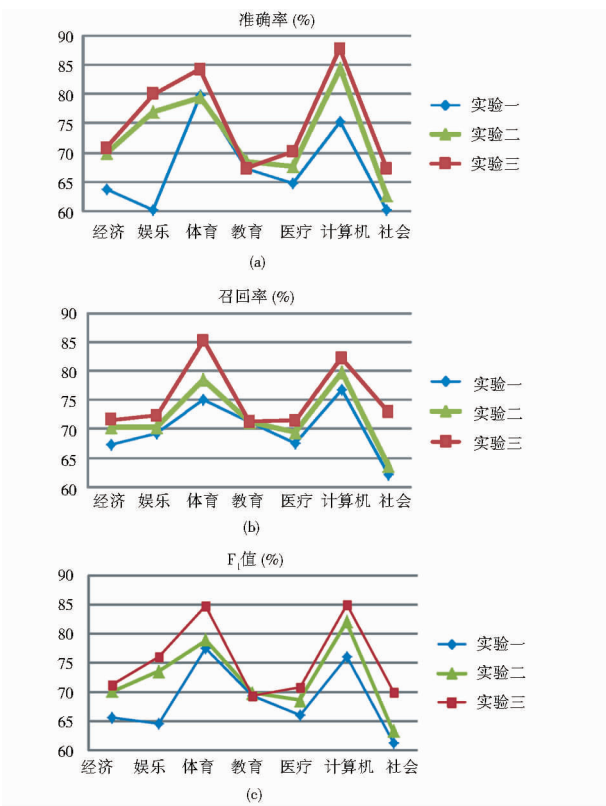

图 5　准确率、召回率和 F_1 值对比

可以看出，使用本文提出的基于半监督学习的分类方法对短文本进行分类的准确率、召回率和 F_1 值都比其他方法有所提高。由此可说明，短文本分类时进行训练样本的扩展，并使用扩展后已标注部分构建初始分类器，对未标注部分进行自学习预测，充分实现了未标注样本的价值，提升了分类的效果。但是在实验过程中对于个别类别的判断还存在不足，造成分类效果不明显，原因可能是在构建初始分类器时，每个类别的已标注样本所占比例不均匀或者是出现了噪声问题并通过迭代不断扩大。总体来说，使用该方法构建的分类器性能从准确率、召回率和 F_1 值指标上有所提高。

6　结　语

本文将半监督学习的思想引入到处理短文本分类问题中，针对短文本存在的对大量未标注样本的标注瓶颈、特征项较少、信息关联较弱等问题，提出一种基于半监督学习的文本分类方法，并阐述了相关的理论及其研究现状、基本思想和具体的方法。通过对训练样本进行特征扩展，并使用已标注样本构建的初始分类器对大量未标注样本进行分类预测，通过不断迭代将未标注样本转化为已标注样本，最后使用更新过的扩大训练样本进行最终分类器的训练。对比实验表明本文提出的方法对短文本进行分类是可行的，并能够使分类效果得到提升。

参考文献：

[1] 蒲筱哥. Web 自动文本分类技术研究综述[J]. *情报学报*, 2009, 28(2): 233 - 241. (Pu Xiaoge. A Literature Review on Web Automated Text Categorization Technology[J]. *Journal of the China Society for Scientific and Technical Information*, 2009, 28(2): 233 - 241.)

[2] 苗夺谦, 卫志华. 中文文本信息处理的原理与应用[M]. 北京:清华大学出版社, 2007. (Miao Duoqian, Wei Zhihua. The Theory and Application for Chinese Text Information Processing [M]. Beijing: Tsinghua University Press, 2007.)

[3] 王细薇, 沈云琴. 中文短文本分类方法研究[J]. *现代计算机: 专业版*, 2010(7): 28 - 31. (Wang Xiwei, Shen Yunqin. Research on Chinese Short Text Classification Method[J]. *Modern Computer*, 2010(7): 28 - 31.)

[4] 范云杰, 刘怀亮. 基于维基百科的中文短文本分类研究[J]. *现代图书情报技术*, 2012(3): 47 - 52. (Fan Yunjie, Liu Hua-

iliang. Research on Chinese Short Text Classification Based on Wikipedia[J]. *New Technology of Library and Information Service*, 2012(3): 47 - 52.)

[5] Yih W T, Meek C. Improving Similarity Measures for Short Segments of Text[C]. In: *Proceedings of the 22nd National Conference on Artificial Intelligence*. 2007: 1489 - 1494.

[6] 林小俊, 张猛, 暴筱, 等. 基于概念网络的短文本分类方法[J]. *计算机工程*, 2010, 36(21): 4 - 6. (Lin Xiaojun, Zhang Meng, Bao Xiao, et al. Short - text Classification Method Based on Concept Network[J]. *Computer Engineering*, 2010, 36(21): 4 - 6.)

[7] 蔡月红, 朱倩, 孙萍, 等. 基于属性选择的半监督短文本分类算法[J]. *计算机应用*, 2010, 30(4): 1015 - 1018. (Cai Yuehong, Zhu Qian, Sun Ping, et al. Semi - supervised Short Text Categorization Based on Attribute Selection[J]. *Journal of Computer Applications*, 2010, 30(4): 1015 - 1018.)

[8] 宁亚辉, 樊兴华, 吴渝. 基于领域词语本体的短文本分类[J]. *计算机科学*, 2009, 36(3): 142 - 145. (Ning Yahui, Fan Xinghua, Wu Yu. Short Text Classification Based on Domain Word Ontology[J]. *Computer Science*, 2009, 36(3): 142 - 145.)

[9] 王盛, 樊兴华, 陈现麟. 利用上下位关系的中文短文本分类[J]. *计算机应用*, 2010, 30(3): 603 - 606. (Wang Sheng, Fan Xinghua, Chen Xianlin. Chinese Short Text Classification Based on Hyponymy Relation[J]. *Journal of Computer Applications*, 2010, 30(3): 603 - 606.)

[10] 白秋产, 金春霞. 概念属性扩展的短文本聚类算法[J]. *长春师范学院学报*, 2011, 30(5): 29 - 33. (Bai Qiuchan, Jin Chunxia. Short Text Clustering Algorithm Based on Concept Feature Expansion[J]. *Journal of Changchun Normal University*, 2011, 30(5): 29 - 33.)

[11] 史伟, 王洪伟, 何绍义. 基于微博平台的公众情感分析[J]. *情报学报*, 2012, 31(11): 1171 - 1178. (Shi Wei, Wang Hongwei, He Shaoyi. Study on Public Sentiment Based on Microblogging Platform[J]. *Journal of the China Society for Scientific and Technical Information*, 2012, 31(11): 1171 - 1178.)

[12] Banerjee S, Ramanathan K, Gupta A. Clustering Short Texts Using Wikipedia[C]. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2007: 787 - 788.

[13] Day N E. Estimating the Components of a Mixture of Normal Distributions[J]. *Biometrika*, 1969, 56(3): 463 - 474.

[14] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. *Journal of the Royal Statistical Society: Series B*, 1977, 39(1): 1 - 38.

[15] Shahshanani B M, Landgrebe D A. The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1994, 32(5): 1087 - 1095.

[16] 秦飞. 基于半监督学习的文本分类研究[D]. 成都:西南交通大学, 2010. (Qin Fei. Research on Document Classification Algorithm Based on Semi - supervised Learning[D]. Chengdu: Southwest Jiaotong University, 2010.)

[17] Nigam K, McCallum A, Mitchell T. Semi - supervised Text Classification Using EM[A]//Semi - supervised Learning[M]. Boston: MIT Press, 2006.

[18] 侯翠琴,焦李成. 基于图的 Co - Training 网页分类[J]. *电子学报*, 2009, 37(10): 2173 - 2180. (Hou Cuiqin, Jiao Licheng. Graph Based Co - Training Algorithm for Web Page Classification[J]. *Acta Electronica Sinica*, 2009, 37(10): 2173 - 2180.)

[19] 郑海清, 林琛, 牛军钰. 一种基于紧密度的半监督文本分类方法[J]. *中文信息学报*, 2007, 21(3): 54 - 60. (Zheng Haiqing, Lin Chen, Niu Junyu. A Closeness - based Semi - supervised Text Classification Method[J]. *Journal of Chinese Information Processing*, 2007, 21(3): 54 - 60.)

[20] Vapnik V N. Statistical Learning Theory[M]. Wiley - Interscience, 1998.

[21] Blum A, Chawla S. Learning from Labeled and Unlabeled Data Using Graph Mincuts[C]. In: *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, USA. 2001: 19 - 26.

[22] Nigam K, McCallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents Using EM[J]. *Machine Learning*, 2000, 39(2 - 3): 103 - 134.

[23] 张博锋,白冰,苏金树. 基于自训练 EM 算法的半监督文本分类[J]. *国防科技大学学报*, 2007, 29(6): 65 - 69. (Zhang Bofeng, Bai Bing, Su Jinshu. Semi - supervised Text Classification Based on Self - training EM Algorithm[J]. *Journal of National University of Defense Technology*, 2007, 29(6): 65 - 69.)

[24] 陈才扣, 喻以明. 半监督邻近鉴别分析[C]. 见:*2010 年第三届计算智能与工业应用国际学术研讨会*, 2010:435 - 438. (Chen Caikou, Yu Yiming. Semi - supervised Neighborhood Discriminant Analysis[C]. In: *Proceedings of the 3rd International Conference on Computational Intelligence and Industrial Application*, 2010: 435 - 438.)

[25] Zhu X J. Semi - Supervised Learning Literature Survey [R/OL]. [2013 - 01 - 13]. http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf.

[26] Zhou D, Zhang C S. Semi - supervised Learning Using Random Subspace Based Linear Embedding Repulsion Graph[C]. In: *Proceedings of the 31st Chinese Control Conference*. 2012: 3676 - 3680.

[27] Zhu X J, Goldberg A B. Introduction to Semi - Supervised Learning[M]. San Rafael, CA: Morgan and Claypool Publishers, 2009: 9 - 19.

(作者 E - mail:zqvictory2011@yeah.net)