维普资讯 http://www.cqvip.com

基于主动学习的文档分类

覃刚力[1] 黄 科[2] 杨家本[1]

(清华大学自动化系 北京 100084)[1] (清华大学计算机系 北京 100084)[2]

Active Learning Based Text Categorization

QIN Gang-Li[1] HUANG Ke[2] YANG Jia-Ben[1]

(Department of Automation, Tsinghua University, Beijing 100084)[1] (Department of Computer Science, Tsinghua University, Beijing 100084)[2]

Abstract In the field of text categorization, the number of unlabeled documents is generally much gretaer than that of labeled documents. Text categorization is the problem of categorization in high-dimension vector space, and more training samples will generally improve the accuracy of text classifier. How to add the unlabeled documents of training set so as to expand training set is a valuable problem. The theory of active learning is introducted and applied to the field of text categorization in this paper, exploring the method of using unlabeled documents to improve the accuracy of text classifier. It is expected that such technology will improve text classifier's accuracy through adopting relatively large number of unlabelled documents samples. We brought forward an active learning based algorithm for text categorization, and the experiments on Reuters news corpus showed that when enough training samples available, it's effective for the algorithm to promote text classifier's accuracy through adopting unlabelled document samples.

Keywords Active learning, Text categorization, VSM, Machine learning

1 引言

随着 Internet 快速普及和发展，使得网络上的电子文档数量激增。用户在享受它所提供的大量信息的同时，也越来越感到被庞大复杂的信息所淹没。然而网络上的文档数据并不是被有组织地管理，而仅仅是一个大的无序数据集合。在网络上寻找自己需要的信息常常要花费大量的时间和精力。如果对这些文档数据进行良好的索引和归纳，将有助于把文档数据系统化，以便进行有效的检索。文档分类就是这样的一个解决方法，它按照一定的标准体系，对文档进行自动标记，对一个待识别文档给予一个或多个类别标记。近年来文档分类技术引起了相当的研究兴趣，这些研究大多采用文档的向量空间模型。

文档的向量空间模型最初由 Salton 提出[1]并在 SMART 系统中应用，研究者在这个向量空间模型基础之上应用和发展了多种特征提取技术和分类器技术，这些技术大多可归为统计识别方法和机器学习方法，比如，最近邻方法、Bayes 方法、决策树方法、神经网络方法、符号学习算法等[2~5]。现有研究提出了多种启发式方法。比如，从数据稀疏的区域选择下一个样本[6]；从当前分类器可靠性较低的区域选择下一个样本[7]；从对原有分类器影响大的区域选择样本[8]等。从试验来看，这些方法取得的效果还不是很明显。

主动学习(Active Learning)是一种学习者主动选取样本进行学习的方法[9]。其基本出发点是，在机器学习领域，学习者有能力或者需要主动地影响或选择训练样本，从而希望学习结果能沿一条更好的道路进化。本文对主动学习的方法进行了研究，提出了基于主动学习的文档分类算法。通过试验，对主动学习在文档分类中应用的一些关键因素进行了分析和比较。

2 文档分类的一般过程

对文档进行分类时，通常需要对初始文档进行预处理，其中可能包括：(1)对于 HTML 文档，去除 HTML 标记；(2)去除文档中的功能词(stop words)，比如，英文中的“the”，“and”等被认为无助于表现文档内容的词；(3)对于中文，进行分词；对于英文，从一个单词的各种变体中提取词干(Stemming)

将文档进行预处理以后，每个文档就变换为向量空间中的一个高维向量，这个向量的每一维是文档中出现的一个字或词，而其权重则是这个词在文档中出现的频率。然而，对于不同的词，其重要性不仅取决于它在这一篇文档中的频率，而且和它在整个文档集中的分布有关系。对于那些广泛分布于各类文档的词，其分类的作用显然要小。所以，通常会对这个初始的文档向量进行调整，对于不同的词给予不同的权值。一个常用而有效的加权方案是 IDF 指标值[1]，第 i 个词的 IDF 指标值计算如下：

$$idf_i = \log(N/n_i) \quad (1)$$

其中，N 是训练集中文档的数目，n_i 是第 i 个词在训练集中出现的文档的个数。idf 值意味着，如果一个词出现的文档数目越少，那么它对区别各个类别的作用越大，它的权重也就越大。将此指标值对文档向量进行加权，那么在文档 j 中第 i 个词的权重调整为：

$$Weight_{ji} = tf_{ji} \times idf_i \quad (2)$$

其中，tf_{ji}是第 i 个词在文档 j 中出现的频率。

在获得每一个文档的向量表示以后，下一步就是分类器的训练。一种有效而简单的分类器是最小距离分类器。在这种方法中，对于每一个类别从其训练样本集中求取这个类别的

覃刚力 博士研究生，研究领域为机器学习以及多智能体系统的建模、学习和进化机制等。**黄 科** 硕士研究生，研究方向为数据挖掘，模式识别，人工智能。**杨家本** 教授，研究领域为知识系统、复杂系统建模、优化以及多智能体系统的研究、开发和应用。

维普资讯 http://www.cqvip.com

代表向量，以此代表向量作为类中心。在分类阶段，根据待分类文档与哪一个类中心向量距离最小而将其归于哪一类。其中，距离的度量可以采用向量点积。这个分类以及评价过程可以表示如下：

(1)求取类中心，对于第 C_i 类，其类中心向量 $Center_i$ 的计算公式为：

$$Center_i=\frac{1}{N_i}\sum_{j=1}^{N_i}Doc_{ij} \quad (3)$$

其中，N_i 是第 C_i 类中文档的数目，而 Doc_{ij} 是类别为 C_i 的第 j 个文档向量。

(2)对待分类文档 Doc_x 进行分类，其类标签 $Label_x$ 按照下式计算：

$$Label_x=\arg\max_j Sim(Center_j,Doc_x) \quad (4)$$

其中，相似度 Sim 的计算可以采用向量点积。

(3)分类性能的计算可以采用分类精度来表示，对于第 i 类，其分类精度为：

$$Accuracy_i=\frac{\text{分到第 i 类中且属于第 i 类的文档数目}}{\text{分到第 i 类的文档数目}} \quad (5)$$

而对于整个分类器的分类精度，可以利用以下公式计算：

$$Accuracy_{total}=\frac{\text{所有待分类文档中被正确分类的文档的数目}}{\text{所有待分类文档的数目}} \quad (6)$$

3 主动学习的框架及其在文档分类中的应用

主动学习的一般过程如下：为求取从模式空间 X 到类别空间 Y 的映射 X→Y，给定训练集 $\{(x_i,y_i)\}_{i=1}^{n}$，其中 $x_i\in X$，$y_i\in Y$。学习者根据一定的标准反复从训练集中选择新的输入 $\tilde{x}$，观察其输出 $\tilde{y}$，并将新的实例 $(\tilde{x},\tilde{y})$ 加入到训练集中，以改善学习效果。也就是希望实例 $(\tilde{x},\tilde{y})$ 的加入使得分类器的经验风险最小化。

由此可见，主动学习过程中基本的问题是如何选择下一个样本 $\tilde{x}$。在文档分类领域，和未标注的文档集相比，已标注类别的文档数量通常较少，从而数据稀疏，据此所训练的分类器精度不高，而实际上同时存在着大量的未标注样本。然而基于统计模型的传统模式识别方法在少量训练样本下所构成的分类器通常是不可靠的，所以如何有效利用这些大量未标注的文档来影响分类器的构造，提高文档分类的精度是一个值得研究的课题[10]。因此，人们针对这一问题提出了基于最大似然和 Bayes 分类器的方法[11]，基于混合高斯模型的方法[12]等，这些方法取得了一定程度的效果。作为一种尝试和比较，我们考虑将主动学习的框架引入文档分类领域，利用这种方法逐步地将未标注文档引入到分类器的构造中来，从而利用未标注文档来提高文档分类的精度。我们提出一个依据分类器信度进行主动学习的文档分类系统流程，其基本的步骤如下：

基于主动学习的文档分类算法

Step1：对文档集进行预处理，构造文档向量并按照式(2)采用 tf * idf 的方案调整向量权值；

Step2：基于训练集按照式(3)计算出每一类别的类中心向量；

Step3：对测试集向量分类，设定每一次主动学习要选取的文档数目 k；

Step4：循环进行主动学习：

while 训练集还未包括所有文档 DO

1)对于每一个未被用于主动学习的文档 Doc_i，计算它和各个类中心的相似度。将 Doc_i 和各类中心的相似度按照降序排列，得到 $Sim_{i1},Sim_{i2},\cdots,Sim_{in}$；

2)计算文档 Doc_i 的分类可靠度：$rel_i=\frac{Sim_{i1}}{Sim_{i2}}$；

3)将未被用于主动学习的文档按照 rel 值从高到低降序排列，选择 rel 值最高的 k 个文档加入到训练集；

4)基于新训练集按照式(3)重新计算出每一类别的类中心向量；

5)利用新的分类器对测试集中从未用于主动学习的文档进行分类；

END；

Step5：利用最后所得的分类器对测试集进行分类；

Step6：计算整个分类器分类精度 $Accuracy_{total}$。

在这个流程中，我们不是从未标注文档中每一次选取一个文档进行主动学习，而是每一次选取了 k 个文档，这是因为考虑到系统的效率。我们选取的是分类可靠度最大的那些文档，直观地认为将这些文档加入训练集有助于提高分类器的精度。

4 试验结果与分析

本试验所采用的数据集是路透社 1987 年的财经新闻集(Reuters-21578[13]中的一部分)，其中共有两万多篇英文新闻，每一篇新闻包括标题、时间、新闻内容以及由人工归类的本篇新闻所属的类别信息。我们从出现频率最高的 7 类中每类取了约 500 篇的文档作为试验用数据集，其中训练集和测试集的划分在试验中作为参数是可变的，将除去作为训练用的文档都作为测试文档。

a)主动学习的试验结果

为了比较主动学习算法的效果，我们计算了仅利用训练集训练分类器进行分类的结果和在学习过程中随机选择样本进行学习的结果。设定每次用于学习的样本数目 k 为 50，在训练集数目不同的情况下所得到的结果比较如图 1。

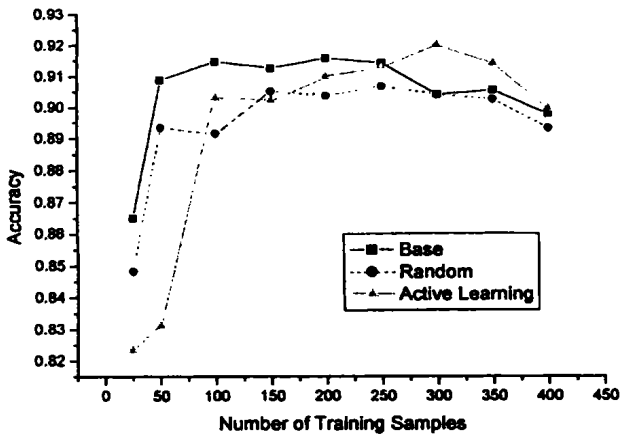

图 1 主动学习的性能比较

图 1 中，横坐标为每一类别初始训练集中文档的数目，纵坐标为测试集的分类精度。Base 曲线表示未利用未标注样本来修改分类器时的分类精度曲线；Random 曲线表示从未标注样本中随机选取样本的分类结果来逐步修改分类器的精度变化曲线；Active Learning 曲线表示从未标注样本中用主动

维普资讯 http://www.cqvip.com

学习算法来逐步修改分类器得到的精度变化曲线。

从图中精度结果对比可知，在训练样本数目比较少的时候，也就是分类器不是很可靠的时候，在未标注样本中依据分类结果主动选择样本不如随机选择样本来修改分类器，而这种随机选择又不如不对分类器进行修改。这和预计是相符合的，因为这时用于训练分类器的样本数目少，分类器并不可靠，利用一个不可靠的分类器给出的结果来对分类器进行改造，结果只能是分类器的性能越改越差。而主动学习由于其对分类器结果的依赖程度深，所以这时利用主动学习算法来选择样本也就比随机选择所导致的分类器精度下降要大。

当训练集的规模超过整个数据集的一半，也就是 250 个时，主动学习算法所获得的分类精度开始超过随机选择和不修改分类器所获得的精度。这是因为随着训练集的增大，分类器的可靠性提高，依靠分类结果的可靠性来主动选择未标注样本进行学习有助于提高分类器的精度。应该看到，不管训练集占整个数据集的多大百分比，随机选择样本进行学习的结果都不如不选择样本进行学习。

在上面的试验结果曲线中，单条曲线的分类精度最大值不是出现在训练集样本最大的精况，这是因为每次所用的测试集不同。试验设置了每类 500 个文档样本，除了用于训练集的样本以外，剩下的样本作为测试样本，所以在训练集样本不同的情况下测试集也不同。上面的试验在于比较不同训练集的情况下不同方法之间的效果差别，单条曲线的变化意义不大。

b)学习次数变化时的试验结果

上面的试验中测试集的所有的文档都先后被用于主动学习。但有理由相信，按照 rel 值的指标来选择文档，先入选的文档其可靠性要高于后入选的文档，所以如果不是利用测试集中所有的文档进行主动学习，而是只选择其中一部分 rel 值高的文档进行主动学习，其最后的分类精度可能比利用所有的测试文档进行学习的精度要高。为此我们测试了在不同初始训练集的情况下采用不同数目的文档进行学习的结果。

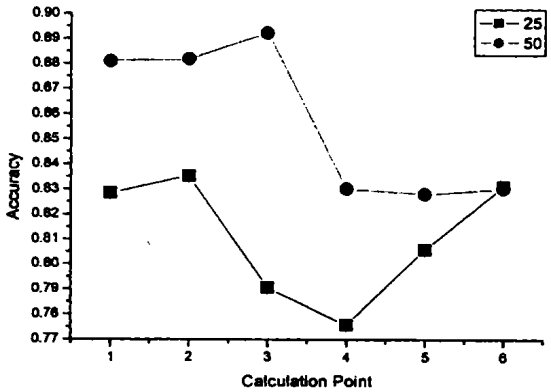

图 2　训练集小的情况下的结果

试验设置与上一节相同，每次选取 50 个分类可靠性高的测试文档进行主动学习，在用于学习的测试文档数目首次超过测试集文档数目的 1/6，2/6，3/6，4/6，5/6 以及所有测试集文档都被用于学习的时候分别测试分类器在所有测试样本上的分类精度。图 2～4 显示了初始训练样本集数目不同的试验结果，其中横坐标的数字表示在用于主动学习的文档数目达到测试集文档数目的六分之几的时候分类器的分类精度(如 1 表示用于主动学习的文档数目首次超过测试集文档的 1/6 的时候分类器的分类精度，其余类推)。纵坐标表示相应分类器的分类精度，各条曲线对应初始训练集大小不同的情况，每条曲线对应的初始训练集的大小由图标表示出来。图 2 显示的是在训练样本数少的时候的结果，图 3 显示的是在训练样本数目比较大而且分类精度在学习过程中出现局部最大值的情况，图 4 显示的是在主动学习过程中分类精度不出现局部最大值的情况。

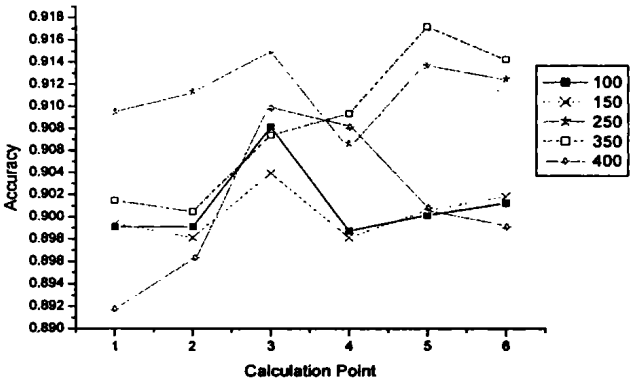

图 3　训练集较大且有局部最大值的结果

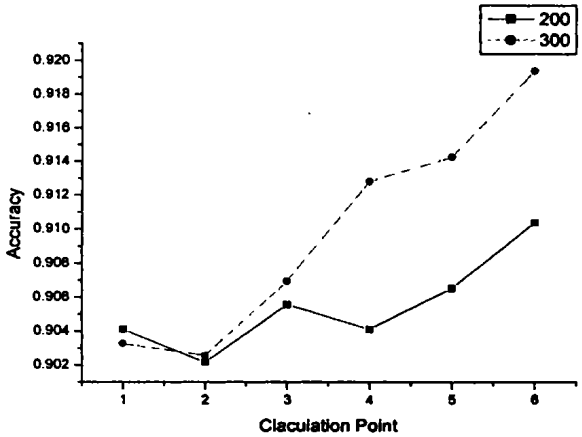

图 4　训练集较大且没有局部最大值的结果

从试验结果可知，在大部分情况下，并不是将所有的训练集样本都用于学习对分类有利。分类器的精度提高随着主动学习的文档数目的增加先增加后减小。也就是说，后面加入训练集的文档对于分类器的精度是有损害的。这和预计情况是一致的，因为文档被选中得越迟，说明分类器对其分类结果的可靠度越低，所以其加入可能会对分类器性能选成损害。从上面的结果也不难看出，一般初始训练集的数目越大，那么分类器越可靠，在测试集中能被用于主动学习且使分类器精度提高的文档数目就越多。

c)分类可靠性和分类精度分析

在前面的试验中，选择测试集中的样本进行主动学习是以分类可靠度 rel 作为指标的，但是在前面的试验中并未对分类可靠度的合理性进行试验。分类可靠度高是否就意味着分类精度高，这正是下面的试验所要回答的问题。

在试验中，在不同的初始训练集大小的情况下分别按照分类可靠度 rel 的标准选择测试样本进行主动学习，直至所有的测试样本都被用于主动学习，然后用最后所得分类器对所有的测试样本进行分类得到最终的分类结果。根据最终的分类结果，将测试文档按照分类可靠度 rel 的大小进行降序排列，然后将所有的测试文档均匀地分为 6 个区段，使得每个

维普资讯 http://www.cqvip.com

区段上的文档数目相同，计算每个区段上的文档分类精度，从而得到文档分类精度对于 rel 值的变化关系。

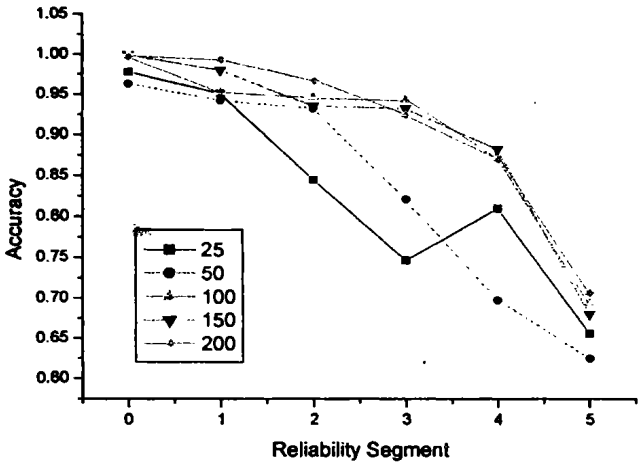

图 5　初始训练集小于数据集一半的结果

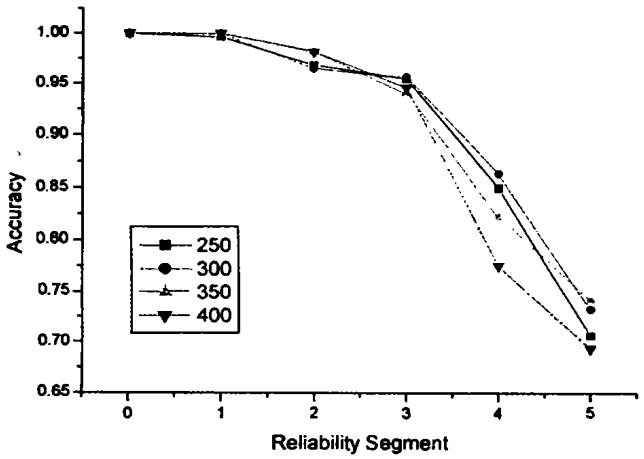

图 6　初始训练集大于数据集一半的结果

在上面的两张图中，不同的曲线表示初始训练集大小不同的情况下所得的结果。横坐标表示按照分类可靠度 *rel* 值划分的不同区段，0 表示 *rel* 值最高的 1/6 部分文档，1 表示 *rel* 值次高的 1/6 部分文档，依次类推。由试验结果不难看出，*rel* 值和分类精度之间存在明显的单调关系，即 *rel* 值越高，则分类精度越高，*rel* 值越低，则分类精度越低。这就说明在前面的试验中使用 *rel* 值作为选择测试集样本进行主动学习的指标是合理的。

结论　本文介绍了主动学习的基本思想及其在文档分类领域的应用。在实际的文档分类应用中，大量的文档是未经标注的文档样本，在传统的方法中这些样本对于分类器是没有影响的，如何充分利用这些大量的未标注样本来改进分类器的性能是一个值得研究的课题。本文在不同的训练集大小的情况下进行了主动学习的试验，分析了主动学习次数的不同对结果的影响，选择文档标准的影响。通过试验发现，主动学习的效果对于分类器的可靠性依赖比较大，在训练集样本多且分类器比较可靠的时候利用主动学习通常能改善分类器的分类精度，反之，如果训练集中文档数目不多，分类器并不可靠的时候利用主动学习很可能会降低分类器的分类精度。如何在分类器不是很可靠的情况下利用未标注样本是将来一个值得研究的方向。

参 考 文 献

1　Salton, Gerard. Introduction to modern information retrieval. Auckland: McGraw-Hill, 1983

2　Aas K, Eikvil L. Text Categorisation: A Survey, Rapport Nr. 941, June, 1999. ISBN 82-539-0425-8

3　Wiener E, Pedersen J O, Weigend A S. A Neural Network Approach to Topic Spotting. In: Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval

4　Yang Yi-ming. An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, 1999, 1(1-2): 67～88

5　Larkey L S, Croft W B. Combing Classifiers in Text Categorization. In: Proc. of SIGIR-96, 19th ACM Intl. Conf. on Research and Development in Information Retrieval

6　Lenden A, Weber F. Implementing inner drive by competence reflection. In: Proc. 2nd Int. Conf. On Simulation of Adaptive Behaviror, MIT Press, 1993

7　Thrun S, Moller K. Active exploration in dynamic environments. Advances in Neural Information Processing Systems 4, Morgan Kaufmann

8　Cohn D, Atlas L, Ladner R. Training Connectionist Networks with Queries and Selective Sampling. Advances in Neural Information Processing Systems 2, Morgan Kaufmann

9　Cohn D A, Ghahramani Z, Jordan M I. Active Learning with Statistical Models. Advances in Neural Information Processing System 7, MIT Press, 1995

10　Goldman S, Zhou Yan. Enhancing Supervised Learning with Unlabeled Data. In: The Seventeenth Intl. Conf. on Machine Learning, 2000

11　Nigam K, Mccallum A K, et al. Text Categorization from Labeled and Unlabeled Documents using EM. Machine Learning. Kluwer Academic Publishers, 2000

12　Nigam K P. Using Unlabeled Data to Improve Text Classification: [CMU PhD Thesis]. 2001

13　http://www.research.att.com/～lewis/reuters21578.html