情感分类中不同主动学习策略比较研究*

居胜峰，王中卿，李寿山，周国栋
苏州大学 计算机科学与技术学院，苏州 215006
E-mail: {shengfeng.ju, wangzq870305, shoushan.li}@gmail.com; gdzhou@suda.edu.cn

摘 要：近些年来，情感分类在自然语言处理研究领域获得了显著的发展。然而，大部分已有的研究都基于大规模标注样本的分类情况。实际情况下，收集标注样本是一件费时费力的事情。本文在基于少量标注样本的基础上，研究和探讨基于主动学习的情感分类，即主动挑选“优质”的样本进行标注和学习。本文采用了四种不同的学习策略实现主动学习，分别为不确定性、代表性、差异性和特征信息量。实验验证了主动学习对于情感分类的有效性并详细分析了四种策略在基于情感分类的主动学习过程中所发挥的作用。

关键词：主动学习；情感分析；样本选择

A Comparative Study on Different Active Learning Strategies for Sentiment Classification

Ju Shengfeng, Wang Zhongqing, Li Shoushan, Zhou Guodong
NLP Lab, School of Computer Science and Technology, Soochow University, Suzhou 215006
E-mail: {shengfeng.ju, wangzq870305, shoushan.li}@gmail.com; gdzhou@suda.edu.cn

Abstract: Currently, sentiment classification has become a hot research topic in Natural Language Processing area. However, most studies focus on the classification cases when large scale of labeled data are involved for training. In realistic, the data annotation is very time consuming and labor costing. In this paper, we focus on active learning approaches for sentiment classification given only a small set of labeled data, where the other part of labeled data are actively selected for annotating. Four different strategies are proposed. They are uncertainty, representative, diversity, and feature information. Experimental studies shows the effective of active learning for sentiment classification. Furthermore, comparative studies on the four strategies show their respect effect in detail.

Keywords: active learning; sentiment classification; sample selecting

1 引言

随着网络的普及特别是 Web2.0 的兴起和发展，网络上出现了大量带有情感的文本。情感分类（Sentiment Classification）是指对文本进行情感色彩方面的分类任务[7]。情感分类研究已经发展多年，目前主流的情感分类方法一般都基于监督学习[10]。监督学习的一个显著不足是其训练过程中需要大量标注样本。大量标注样本的获取是一件非常费时费力的工作。因此，如何在小规模样本情况下获得理想的情感分类性能是一个非常有实用价值的研究问题。

通常来讲，基于小规模样本的分类方法存在两种方式，即半监督学习（Semi-supervised Learning）和主动学习（Active Learning）。半监督学习的方式是通过充分利用大规模未标注数据里面隐含的分类信息，进而提高分类性能。该方式在情感分类研究中已经渐渐受到广泛的注意[2][7]。主动学习方式是另外一种能够减少标注样本规模的方法。此方式通过主动选择一些“好”的样本进行标注进而参加分类，从而能够在尽可能使用少的标注样本的情况下保持一定的分类效果。相对而言，主动学习方法在情感分类的研究还比较缺乏，这方面的研究才刚刚起步[19]。

本文主要研究主动学习方法在情感分类中的应用。我们假设存在少量已标注的初始样本，在此基础上，分别采用四种不同的策略去选择未标注样本进行标注。这四种策略分别为：不确定性、

* 本研究受国家自然科学基金（61003155，60873150）和模式识别国家重点实验室开发课题基金资助。

代表性、差异性和特征信息量。具体来讲，不确定性是指已有分类器对测试样本分类结果的确信程度；代表性是指在样本能够代表样本所在的集合的程度；差异性是指未标注样本和已有标注样本之间的差异程度；特征信息量是指样本中包含的特征的数目。其中，前三种策略是一般主动学习方法中经常使用的策略[13]，本文给出了他们在情感分类中的具体实现。

本文其他部分安排如下：第二节介绍相关工作；第三节给出本文使用的具体分类和聚类方法简介；第四节详细介绍基于不确定性、代表性、差异性和特征信息量策略的主动学习方法；第五节给出实验结果和分析；第六节给出相关总结。

2 相关工作

近几年来，情感分类渐渐成为自然语言处理研究领域里面的一个研究热点。Pang 等首次将基于监督学习的机器学习方法用于情感分类[10]。后续很多研究旨在使用不同方法去提高监督学习方法的性能，例如抽取主观句[9]，寻找上层分类特征[11]和利用主题部分相关信息[8]。到目前为止，情感分类研究已经在不同文本粒度都进行的深入研究，例如：词语级[3]、短语级[16]、句子级[4][15]和篇章级[7][9][10]。

一直以来，主动学习方法一直是机器学习领域关注的的一个重要研究分支。已有的主动学习算法大致可以分三类：第一类是选择最能减小当前分类器对分类误差的样本作为候选集，如基于误差减少的抽样方法[12][14][18]；第二类是选择当前分类器测试结果中最不确定的样本作为候选样本，如不确定性抽样方法[6]；第三类是根据多个分类器对于样本类别预测差异程度来选择候选样本，如询问委员会方法（QBC 方法）[1][5][20]。

在情感分类研究中，主动学习方法的使用才刚刚起步。文献[17]中运用基于深度置信网络方法实现主动学习的方法选取初始样本，即在大规模的未标注样本中选取合适的样本作为种子样本进而进行半监督学习。文献[2]中采用不确定性策略主动选择歧义性样本，集合直推式学习方法构建情感分类器。在这篇相关文章中，主动学习的策略都比较单一，仅仅作为其他方法的一个辅助方法。本文实现了多种策略的主动学习方法，并系统研究了这些方法在情感分类中的应用。此外，本文的研究主要集中在给定少量种子样本的基础上，运用主动学习的方法挑选更多合适的样本，进而进行人工标注及参与进一步学习。

3 分类和聚类方法简介

在本节中，我们简述在后续章节中所使用的相关机器学习技术，分别为最大熵分类方法和 K-均值聚类方法。

3.1 最大熵分类

最大熵分类方法是基于最大熵信息理论，其基本思想是为所有已知的因素建立模型，而把所有未知的因素排除在外。也就是说，要找到一种概率分布，满足所有已知的事实，但是让未知的因素最随机化。

在最大熵模型下，预测条件概率 $P(c \mid d)$ 的公式如下：

$$P(c \mid d) = \frac{1}{Z(d)} \exp(\sum_i \lambda_{i,c} F_{i,c}(d,c))$$

其中 $Z(d)$ 是归一化因子。$F_{i,c}$ 是特征函数，定义为：

$$F_{i,c}(d,c') = \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0, & \text{otherwise} \end{cases}$$

在情感分类中，我们所用的特征主要是词特征。

3.2 K-均值聚类

K-均值聚类（k-Means Clustering）是一种简单但有效的非监督实时聚类算法。该算法原理简单，即在最小误差函数的基础上将数据划分为预定的类数 K。聚类过程中，我们采用余弦相似度方式实现两个向量之间的距离衡量。

4 主动学习方法

本文分别实现基于不确定性、代表性、差异性和特征信息量的主动学习方法，并将它们应用到情感分类研究。

1）不确定性（Uncertainty）是指已有分类器对测试样本分类的确信度。通常来说，分类器无法确定的样本可以认为其包含现有的分类模型所不具备的分类信息。所以不确定性是衡量样本重要性的一个重要信息。为了将不确定性引入主动学习中，首先训练种子样本训练获得一个分类模型，并通过分类模型对未标注样本进行分类测试，然后选取分类结果中类别后验概率最接近 0.5（分类模型认为是不确定性最高的样本）作为候选样本并进行人工标注，最后把这些标注好的样本加入到种子样本集中。基于不确定性的主动学习的算法描述如图 1 所示。

算法 1：基于不确定性（Uncertainty）的主动学习算法

输入：标注样本 L，未标注样本 U

输出：新的标注样本 L

程序：

循环 N 次

1）从 L 中学习分类器 F

2）使用 F 对未标注样本 U 进行分类，获得 U 中每个样本针对 L 的不确定性

3）选择 U 中最不确定的 n 个样本

4）把 n 个样本从 U 中取出，放入 L 中

图1 基于不确定性的主动学习算法

由于不确定性是主动学习策略里面最重要的的一个指标，后续的其他策略实现中，都是以不确定性为基础的。

2）代表性（Representative）是指样本能够代表样本空间分布的程度。为了能够选择具有代表性的样本，我们首先通过聚类方法将未标注样本分配到各个小类中，再在各个小类中选择样本。可以认为被选中的样本能够很好的代表这个小类的样本分布情况，进而体现出其代表性。此算法在算法 1 的基础上结合了聚类算法，在循环前对未标注样本进行聚类，而且第 3 步挑选不确定性最高的样本改成挑选每类中的不确定性最高的样本。

3）差异性（Diversity）是指所选的未标注样本和已有标注样本的差异程度。直觉上，待选样本应该尽可能同已有标准样本有一定差异，从而保证新的分类信息的加入。为了体现待选样本同已有标注样本的差异性，首先计算未标注集合中每个样本距离种子样本中心的距离，距离越远代表差异性越大，然后选取分类不确定性和差异性最大的样本作为候选样本进行人工标注，把这些标注好的样本也加入到种子样本集中。我们根据如下公式计算未标注样本的差异性：

$$diversity = \lambda distance + (1-\lambda) uncertainty$$

其中，*distance* 代表未标注样本距离种子样本中心的距离，*uncertainty* 代表未标注样本的不确定性，$0<\lambda<1$。此算法在算法 1 的基础上排序条件改为差异性，在循环前先计算每个未标注样本与已标注

样本中心点的距离，而且第 3 步挑选不确定性最高的样本改成挑选差异性最高的样本。

4）特征信息量（Feature Information）特征信息量是指样本中包含的特征的数目。如果样本包含的特征太少，一方面该样本可能包含的分类信息太少，另外一方面这些样本往往是一些噪声样本。相反，包含大量特征的样本可能会带有大量分类信息。因此，应该尽量倾向于选择包含大量特征的样本，即特征信息量大的样本。此算法在算法 1 的基础上增加特征数作为评选条件，在循环前先计算每个未标注样本的特征数，而且第 3 步挑选不确定性最高的样本时也要求样本的特征数要大于一个预定值。

5 实验结果与分析

实验采用的数据来自亚马逊[1]英文网站的产品评论，包括四个领域：书籍、DVD、电子和厨房，每个领域各有 6500 篇评论。实验过程中，我们选择 100 个样本作为已有标注的种子样本，400 个样本作为测试样本，其他剩余样本作为未标记样本。分类算法使用 MALLET 机器学习工具包[2]里面的最大熵分类器，其所有参数设置为初始默认值。

在情感分类中，通常使用准确率（accuracy，*acc.*）作为分类效果的衡量标准，准确率的计算公式如下：

$$acc. = \frac{number\ of\ correctly\ classified\ samples}{total\ number\ of\ all\ samples}$$

我们分别使用四种策略主动选择 100、900 和 1800 个样本三种情况进行测试，并作相应的分析。除了上述的四种策略，我们同时给出随机抽取样本方式的分类结果。为了更清楚地描述实验结果，我们分别用将随机（Random）选择样本进行标注的方法称为 RAND，将基于不确定性（Uncertainty）的主动学习方法称为 UNCE，将基于代表性（Representative）的主动学习方法称为 REPR，将基于差异性（Diversity）的主动学习方法称为 DIVE，将基于特征信息量（Feature Information）的主动学习方法称为 FEAT。

表 1 是对于每类文档加 100 篇的分类结果。从表中可以看出，相对于随机选择方法，每一种主动学习策略都明显有优势。在四种策略中，不确定性和特征信息量两个策略表现最佳，相对于随机方法都有超过 2 个百分的提高。另外，各种方法在 DVD 领域的效果都不佳，但都比较稳定。

表 1　加入 100 个样本后的分类结果

	书籍	DVD	电子	厨房	平均值
RAND	0.667	0.681	0.685	0.731	0.691
UNCE	0.693	0.698	0.698	0.770	0.715
REPR	0.669	0.693	0.698	0.741	0.700
DIVE	0.658	0.693	0.675	0.765	0.698
FEAT	0.713	0.698	0.728	0.763	0.726

表 2 是对于每类文档加 900 篇的分类结果。从表中数据可以看出，各方法相对于随机选择样本进行标注的方法的提高与取 100 篇时的情况基本相同。其中对于基于特征信息量的主动学习方法相对于其他方法的提高也趋于稳定。原因可能是随着样本增多，信息量渐渐饱和，所以选取文本时对于特征的多少影响已经不大。

[1] http://www.amazon.com/

[2] http://mallet.cs.umass.edu/

表2 加入900个样本时的分类结果

	书籍	DVD	电子	厨房	平均值
RAND	0.735	0.730	0.757	0.808	0.758
UNCE	0.768	0.758	0.780	0.835	0.785
REPR	0.747	0.737	0.795	0.830	0.777
DIVE	0.760	0.760	0.773	0.828	0.780
FEAT	0.785	0.740	0.805	0.813	0.786

表3 加入1800个样本时的分类结果

	书籍	DVD	电子	厨房	平均值
RAND	0.765	0.737	0.790	0.824	0.779
UNCE	0.785	0.760	0.828	0.855	0.807
REPR	0.775	0.735	0.830	0.855	0.799
DIVE	0.793	0.745	0.790	0.843	0.793
FEAT	0.775	0.755	0.828	0.853	0.803

表3是对于每类文档加1800篇的分类结果。随着文本数量的变大，各种方法分类效果的提高已经基本一致。对比表2和表3，我们大致可以发现，加入900个样本时分类器的分类效果已经超过随机取样本1800篇时的分类效果。因此，在情感分类中，主动学习方法可以大大将少样本的标注代价。

6 结语

本文主要研究基于主动学习的情感分类方法。我们分别实现和比较了四种不同的选择策略去选择"好"的样本进行标注。这四种策略分别为不确定性、代表性、差异性和特征信息量。实验结果表明，基于不确定性和特征信息量的策略最为成功，综合考虑这两种因素可以在尽量少的标注样本情况下明显提高分类性能。

实验结果表明加入代表性和差异性并没有能得到我们预想的结果。我们认为可能的原因是用余弦距离计算相似度在情感分类中并不是一个好的方法，而且各个策略直接本身存在一定的渗透，彼此不完全独立。今后的工作是寻找一种能很好表示文本间的情感的相似度的计算方法，并综合考虑这四种因素提高情感分类性能。

参 考 文 献

[1] Argamon-Engleson S and I. Dagan. Committee-Based Sample Selection For Probabilistic Classifiers. Journal of Artificial Intelligence Research. 1999, 11, 335-360.

[2] Dasgupta S and N. Vincent. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In Proceedings of ACL. 2010.

[3] Esuli A. and F. Sebastiani. Determining the Semantic Orientation of Terms through Gloss Classification. In Proceedings of CIKM. 2005.

[4] Kim S. and E. Hovy. Determining the Sentiment of Opinions. In Proceedings of COLING. 2004.

[5] Kothari R, V. Jain. Learning from Labeled and Unlabeled Data using a Minimal Number of Queries. IEEE Transactions on Neural Networks, 2003, 14(6), 1496-1505.

[6] Lewis D and W Gale. A Sequential Algorithm for Training Text Classifiers. In Proceedings of 17th International Conference on Research and Development in Information Retrieval. 1994.

[7] Li S, C. Huang, G. Zhou, and S. Lee. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment

Classification. In Proceedings of ACL. 2010.

[8] McDonald R., K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured Models for Fine-to-coarse Sentiment Analysis. In Proceedings of ACL. 2007.

[9] Pang B. and L. Lee. 2004. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In Proceedings of ACL. 2004.

[10] Pang B., L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP. 2002.

[11] Riloff E., S. Patwardhan, and J. Wiebe. Feature Subsumption for Opinion Analysis. In Proceedings of EMNLP. 2006.

[12] Roy N and A. McCallum. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In Proceedings of ICML. 2001.

[13] Shen D, J. Zhang, J. Su, G. Zhou and C. Tan. Multi-criteria-based Active Learning for Named Entity Recognition. In Proceeding of ACL. 2004.

[14] Tong S and E. Chang. Support Vector Machine Active Learning for Image Retrieval. In Proceedings of the ACM International Multi-media Conference and Exhibition. 2001.

[15] Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of ACL. 2002.

[16] Wilson T., J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. Computational Linguistics, 2009, 35(3), 399-433.

[17] Zhou S, Q. Chen, and X. Wang. Active Deep Networks for Semi-Supervised Sentiment Classification. In Proceedings of COLING. 2010.

[18] 宫秀军, 孙建平, 史忠植. 主动贝叶斯网络分类器. 计算机研究与发展, 2002, 39(5), 574-579.

[19] 龙军, 殷建平, 祝恩, 赵文涛. 主动学习研究综述. 计算机研究与发展, 2008, 45, 300-304.

[20] 徐杰, 施鹏飞. 图像检索中基于标记与未标记样本的主动学习算法. 上海交通大学学报, 2004, 38(12), 2068-2072.

[21] 赵悦, 穆志纯, 潘秀琴, 李霞丽. 一种基于半监督主动学习的动态贝叶斯网络算法. 信息与控制, 2007, 36(2), 8-20.