基于主动学习的中文问题分类数据集构建

邱锡鹏，缪有栋，黄萱菁

（复旦大学 计算机科学技术学院，201203 上海）

摘　要：为解决在开放领域问题回答问题中语料规模较小、难以满足问题分类训练需要的问题，用主动学习方法来构建中文问题分类数据集. 根据主动学习的方法进行中文问题类别标注，并且通过主动式特征选择方法来提升性能. 实验结果表明：在使用主动学习方法时可以快速收敛到最佳准确率（85%），在使用人工标注特征下特征集明显的减小. 基于主动学习的标注方法在需要较小人工标注同时取得很好的分类性能，并且在一定程度上还可以明显提高问题分类的准确率.

关键词：主动学习；Passive Aggressive 算法；特征选择；中文问题分类

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：0367－6234(2012)05－0125－04

Constructing Chinese question classification dataset with active learning

QIU Xi-peng, MIAO You-dong, HUANG Xuan-jing

(School of Computer Science, Fudan University, 201203 Shanghai, China)

Abstract: The current corpora of question classification are relatively small and difficult to meet the practical needs of Question Answering system, so that we use active learning methods to construct a Chinese question classification dataset and for question labeling. In addition, we improve the performance of labeling with feature selection. Experimental results show that by using active learning we can quickly converge at the best accuracy (85%) and by using manual tagging we can have small feature set size. The active learning-based labeling method achieved very good classification performance with less manual annotation tagging, which can significantly improve the accuracy of classification to some degree.

Key words: active learning; passive aggressive; feature selection; Chinese question classification

问题分类（Question Classification, QC）是开放领域问题回答（Question Answering, QA）系统的基础和前提，问题分类准确性直接影响整个问答系统的性能[1]. 在 NIST 举办的 TREC QA 评测会议推动下，问题分类的研究已取得很大的进展. 但目前大部分问题分类的研究还集中在英文语料上，在中文问题分类的研究上，由于缺乏大规模的公开中文问题分类数据集，以及中英文的语言区别，因此中文问题分类的性能还达不到英文的水平，这给中文问题回答研究带来了一个主要瓶颈. 因此，标注一个大规模的中文问题分类数据集是中文问答系统研究中非常急迫的工作.

在语料标注中首先需要确定的是标注规范. 目前问题分类语料主要是针对事实类问题进行答案类型的标注，这样无法处理非事实类问题. 本文根据问题类型和答案类型两方面进行标注. 问题类型是定义用户提问的意图，比如“事实类”、“评价类”、“比较类”等. 不同问题类型对应不同的处理方式以及答案生成策略. 答案类型是定义返回答案的类型，比如：“人物”、“歌名”等. 答案类型和问答系统中的其他模块一起配合工作，比如：命名实体识别、文档摘要和答案抽取等. 因此根据 Z. Dong 等[2]的实体分类体系来确定答案类型的标注规范.

在构建数据集的方法中，主动学习方法[3-4]

收稿日期：2010－10－15.
基金项目：国家自然科学基金资助项目(61003091, 61073069).
作者简介：邱锡鹏(1983—)，男，讲师，博士；
黄萱菁(1972—)，女，教授，博士生导师.
通信作者：邱锡鹏，xpqiu@fudan.edu.cn

(Active Learning)已经被证明是一种有效的减少标注工作量有效方法. 主动学习是一种增量式的标注方法,每次只需要人工标注当前模型分类中最不确定的样本,这样可以尽量避免标注重复样本,使得标注样本的差异近可能大. 要标注大规模的数据集,每次按顺序或随机选取样本进行标注的代价相当大,而通过主动学习,每次选取对当前分类模型来说具有最不确定性的样本,会极大程序上降低标注的工作量. 本文采用快速的不确定样本特征选取方式,利用 Passive Aggressive (PA)算法来训练线性分类器,并计算分类的置信度.

1 已有研究

在英文问题分类方面, X. Li 等[5-6]构建了一个简单的英文问题分类数据集,采用了 Winnow network 算法进行问题分类,准确率达到 82%,在使用了一些词的特征以及一些语义特征后,准确率达到 89%. D. Zhang 等[7]使用基于语法树核的方法,先进行句法分析,然后用支撑向量机(Support Vector Machine, SVM)进行分类,使得准确率提高到 90% 左右. Z. Huang 等[8-9]利用英文问句的特点,结合英文疑问词(wh-word,比如:who/what/where)、中心词(Head word)以及该词的同义词和上位词等特征,在这些特定的特征下,采用 SVM 或最大熵方法进行分类,准确率在 89% 左右,在减小特征情况下准确率并没有显著下降.

在中文问题分类方面, 张宇等[10]采用了针对问题分类问题提出了改进的贝叶斯分类方法来改进分类,准确率为 72.4%. 文勖等[11]加入了句法结构特征,在有效的分词和句法分析数据上,得到了 73% 的准确率.

2 基本算法及框架

2.1 Passive Aggressive(PA)算法

PA 算法是一种在线学习算法,其基本思想是为了保证更新后的分类器尽量保留以前的信息,每次总是选取与原有分类器参数向量最接近的新向量,并且利用合适的损失函数作为更新后的向量在当前样本点上的惩罚,而这个惩罚是每次控制分类器更新程度的重要参数.

假设:x 为样本;y 为对应的类别;$\boldsymbol{\Phi}(x,y)$ 为定义在(x,y) 上的特征向量. 每个样本x的预测类别为

$$\hat{y} = \underset{y}{\operatorname{argmax}}(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,y)).$$

式中: $\boldsymbol{w}$ 为权重向量,文献[12]中提供了 3 种不同的优化准则来进行学习更新权重的策略.

在 PA 算法中,目标函数定义为

$$\boldsymbol{w}_{t+1} = \underset{w}{\operatorname{argmin}} \frac{1}{2} \| \boldsymbol{w} - \boldsymbol{w}_t \|^2 + C\xi, \quad (1)$$

$$\text{s.t. } l(\boldsymbol{w},\boldsymbol{\Phi}(x,y)) \leqslant \xi,\ \xi \geqslant 0.$$

式中 $l(w,\boldsymbol{\Phi}(x,y))$ 为损失函数,其定义为

$$l(w,\boldsymbol{\Phi}(x,y)) = \begin{cases} 0, & \gamma(\boldsymbol{w};\boldsymbol{\Phi}(x,y)) > 1; \\ 1 - \gamma(\boldsymbol{w};\boldsymbol{\Phi}(x,y)), & \text{otherwise.} \end{cases}$$

式中: $\gamma(\boldsymbol{w};\boldsymbol{\Phi}(x,y)) = \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,y) - \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,\tilde{y})$, $\tilde{y} = \underset{z \neq y}{\operatorname{argmax}}(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,z))$. Y 为样本的真实类别;$\tilde{y}$ 为分类器预测的最佳类别(除真实类别外).

对于式(1),使用拉格朗日算法来求解最优化问题. 这样 $\boldsymbol{w}$ 的更新方法为

$$\boldsymbol{w} = \boldsymbol{w}_t + a^* (\boldsymbol{\Phi}(x,y) - \boldsymbol{\Phi}(x,\tilde{y})).$$

式中: $a^* = \min\left(C, \dfrac{l_t}{\| \boldsymbol{\Phi}(x,y) - \boldsymbol{\Phi}(x,\tilde{y}) \|^2}\right)$;

l_t 为当前第 t 次迭代时的损失值.

PA 算法的具体训练流程如图 1 所示.

假设给定 n 个训练样本 (x_i,y_i), $i = 1,2,\cdots,n$.

初始化: $w_1 = 0$.

对于 $t = 1,2,\cdots,$ 重复步骤 1) ~ 5):

1) 随机选取一个样本(x_t,y_t);

2) 预测: $\tilde{y} = \underset{r \neq y_t}{\operatorname{argmax}}(\boldsymbol{w}_t \cdot \boldsymbol{\Phi}(x_t,r))$;

3) 计算当前的损失: $l_t = (\boldsymbol{w}_t \cdot \boldsymbol{\Phi}(x_t,\tilde{y})) - (\boldsymbol{w}_t \cdot \boldsymbol{\Phi}(x_t,y_t))$;

4) 令: $\tau_t = \min\left(C, \dfrac{l_t}{\| (\boldsymbol{w}_t \cdot \boldsymbol{\Phi}(x_t,\tilde{y})) - (\boldsymbol{w}_t \cdot \boldsymbol{\Phi}(x_t,y_t)) \|^2}\right)$;

5) 更新 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \tau_t(\boldsymbol{\Phi}(x_t,y_t) - \boldsymbol{\Phi}(x_t,\tilde{y}))$.

图 1 Passive aggressive 算法的训练过程

2.2 主动学习

主动学习是一种半监督学习算法,其核心思想是通过标注最少量的样本,并使得分类模型的性能可以快速提高,减少人工标注的工作量. 近年来主动学习方法已经被广泛应用于需要大量人工标注的数据集构建工作中.

在主动学习中,首先标注两个数据集:训练数据集 L 和测试数据集 T. 初始训练集 L 仅仅包含少量的样本,未标注的样本都被分配到未标注数据集 U.

首先使用训练数据集 L 中的样本学习出一个分类模型. 用这个模型对未标注数据集 U 中的样本进行预测,并选取出前 k 条预测置信度最低的样本进行人工标注,并从 U 中删除,标注的样本后加入 L 中,并重新训练分类模型. 并不断重复上述过程. 测试集 T 用来检验当前模型的预测效果. 当新的模型在测试集 T 上的性能增加小于一定阈值时,就停止标注或重新选取测试集 T.

为了选取最不确定的样本,定义一个预测置信度 R. 设 y_1 和 y_2 分别为两个当前模型预测分数最高的类. 标签为

$$y_1 = \operatorname{argmax}_y(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,y)),$$

$$y_2 = \operatorname{argmax}_{y \neq y_1}(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,y)).$$

令 $R = \dfrac{\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,y_1)}{\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(x,y_2)}$. R 值越小意味着当前分类模型对于得分最高的前两个类的区分能力越低,预测不确定性越高.

对于未标注集合 U 的所有数据,本文都使用当前分类模型计算相应的置信度,排序找出置信度最小的前 i 个样本进行人工标注,从 U 中删除并加入到训练集 L 中. 标注流程图如图 2 所示.

输入:训练集 L,未标注集 U,测试集 T.
当未标注集合不为空集时,重复下列步骤 1) ~ 5):
1) 在当前训练集 L 上使用 PA 算法训练分类模型;
2) 在测试集上计算当前分类模型的总体准确率;
3) 如果准确率变化差异小于预设的阈值,则中止循环;
4) 对未标注样本集 U 中的所有样本计算相应的置信度,并选取置信度最低的前 i 个样本,并从 U 中删除;
5) 对这 i 个样本进行人工标注,并加入训练集 L.
输出:样本集合 L.

图 2　基于主动学习的数据标注流程

3　基于主动学习的问题分类标注方法

在问题分类中,提取复杂特征增加了处理时间但并不能有效地改进模型性能. 并且当处理口语或网络问句时,深层语法分析不能取得很好的结果,这样利于语法特征反而会降低性能. 因此根据先验知识提取一些有效特征有效提高模型性能.

本文提取线索词 (hint word)作为特征. 当一个问句中出现多个线索词上时,将多个线索词之间的共现关系作为扩展特征.

4　实　验

4.1　数据收集

本文从“百度知道”和“搜搜问问”两个问答网站下载一百万条娱乐相关的问句. 然后将利于主动学习的方法标注. 每次进行标注时,让两个标注者进行标注,当两个标注者标注结果不一致时,让一个校对人员进行最终确认.

4.2　主动抽取特征的实验效果以及分类器选择实验

为了验证线索词特征的有效性, 与 n-gram 特征进行对比.

首先进行特征选取,采用图 3 中的实验流程,设定初始样本集大小和每次增加的样本数量都为 100. 实验中,在 30 次循环后准确率收敛,此时抽取特征为 400 个特征. 使用这 400 个特征进行后续实验.

图 3 是 PA 与感知器两个分类器的结果比较. 从图 3 中明显可以发现 PA 算法比感知器的性能有很大提升. 因此之后的实验都采用 PA 算法.

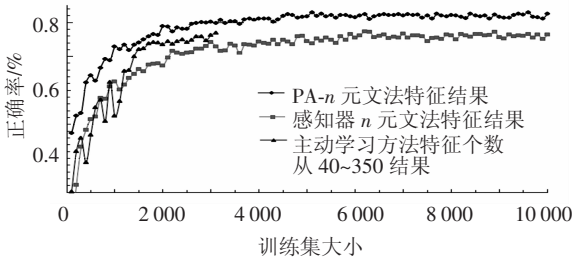

图 3　PA 算法和感知器铁问题分类性能比较

在感知器的 n-gram 特征和主动抽取线索词特征的结果对比中看出通过加入线索词特征可以快速提升分类模型的效果.

对于采用相同算法时,使用线索词特征的收敛速度更快,并且分类性能更好. 在最终的线索词库中,大部分为疑问词或明显指示词的不同变换形式.

因此在数据构建过程中,用 PA 算法进行模型训练. 这里使用 2 种特征 (n-gram 特征(n = 3)、线索词特征)进行比较. 并比较主动学习和随机抽取样本两种不同的标注方法进行训练,一共 4 组方式作为横向比较. 将这4 组方式选取同样的初始训练集,之后每次根据不同选取样本策略选取 100 条数据进行标注.

4 组实验的性能比较如图 4 所示. 基于主动学习的两组方法在收敛速度上明显快于另外两组,当准确率达到 80% 时,采用基于主动学习的方法只需要基于随机采样的 1/2 数据量,这就减少了 1/2 的数据标注工作量. 因此采用基于主动学习标注方法在减少人工方面是十分显著的.

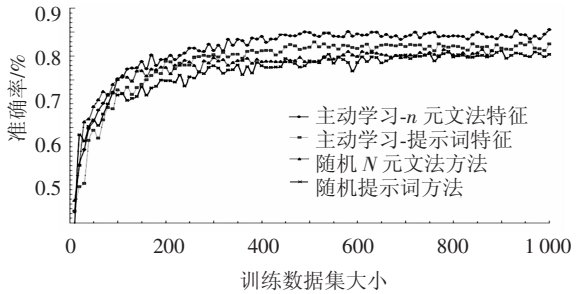

图 4　4 种抽取特征方式的比较

在使用线索词作为特征时,错别字会降低分类性能. 例如:“歌 - > 哥”,“专辑 - > 专集”等. 在以后研究中,需要进一步针对这个问题进行研究.

4.3 构建中文问题分类数据集

经过基于主动学习的数据集标注,最终形成了包含 12 309 条问句的中文问题回答数据集.

4.3.1 问题类型

在所有问句中,目前包含 11 种问句形式,统计各类的数据量和样例问题如表 1 所示.

每一类都有自己显著的问句形式并且类与类之间没有显著的相关性,可作为分类器的类别进行学习.

4.3.2 答案类型

相对于问题类型,答案类型是协助问答系统进行答案抽取的一个重要依据. 目前数据集中包含答案类型可以分为 7 大类和 66 个小类. 表 2 是 66 个小类的数据分布情况.

表 1 中文问题分类数据集的问题类型统计

问句形式	标注总数	样例问句
推荐类	3 159	请推荐一首好听的歌
需求类	2 154	谁有《忘情水》的 CD
事实类	4 941	这首歌是谁唱的?
枚举类	1 487	刘德华的歌有哪些?
评价类	449	林宥嘉的歌好听吗?
方法类	3	怎么下载?
关系类	11	他和作者什么关系?
描述类	98	他最近忙什么?
是非类	3	他跟他的朋友谁大?
比较类	1	他和姚明谁更高?
原因类	3	为什么我的电脑没有声音?

表 2 中文问题分类数据集的答案类型统计

答案类型统计					
歌名:3 131	电脑术语:36	军用器材:2	器官及排泄物:4	国家:47	星座:5
网名: 620	专辑名:616	电视台:16	活动节目名:1	百分数: 6	动物:2
歌词:447	数量:551	事件:148	文艺作品名:365	关系:11	代码:68
网址:562	品牌:29	其他机构:17	其他物理度量:53	电视节目:56	书号:1
信息类:456	年龄:219	时间长度:67	数码产品及家电:5	频率:1	游戏名:1
名称:2 582	陆地:2	食材:12	交通运输工具:1	服装: 7	职位:1
电台频道:3	体育术语:4	方位:2	艺术术语: 123	人物类:2	剧情:12
曲谱:158	植物:2	颜色:3	日常生活用品:3	金额: 155	职业:12
时间:738	序号:74	奖项:48	赢利性机构: 234	位置类:10	语言:10
建筑物:12	地址:113	别称:12	不确定类型:1	医疗术语:1	种族:2
教育机构:29	其他术语:14	非实体:1	国家行政单位:339	其他:1	乐器:43

5 结 语

本文提出一种基于主动学习的构建中文问题分类数据集方法,并从问题类型和答案类型两个方面定义了一套中文问题类别规范. 实验显示该方法有效地减少了标注样本的工作量.

参考文献:

[1] VOORHEES E M. Overview of the TREC 2003 question answering track [C]//Proceedings of the Twelfth Text Retrieval Conference (TREC 2003). [S. l.]: Computer Science Bibliography, 2003: 54 – 68.

[2] DONG Z, DONG Q. HowNet and the computation of meaning [M]. [S. l.]: World Scientific Publishing Company, 2006.

[3] FREUND Y, SEUNG H S, SHAMIR E, *et al.* Selective sampling using the query by committee algorithm [J]. Machine Learning, 1997,28(2/3): 133 – 168.

[4] COHN D A, GHAHRAMANI Z, JORDAN M I. Active learning with statistical models[J]. Journal of Artificial Intelligence Research, 1996, 4(1): 129 – 145.

[5] LI X, ROTH D. Learning question classifiers [C]// Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 1 – 7.

[6] LI X, ROTH D. Learning question classifiers: the role of semantic information[J]. Natural Language Engineering, 2006, 12(3): 229 – 249.

[7] ZHANG D, LEE W S. Question classification using support vector machines[C]//Proceedings of the 26th ACM SIGIR Conference in Information Retrieval. New York, NY: ACM, 2003:26 – 32.

[8] HUANG Z H, THINT M, QIN Z C. Question classification using head words and their hypernyms [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2008: 927 – 936.

[9] HUANG Z H, THINT M, CELIKYILMAZ A. Investigation of question classifier in question answering [C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2009: 543 – 550.

[10] 张宇,刘挺,文勖. 基于改进贝叶斯模型的问题分类[J]. 中文信息学报, 2005, 19(2):100 – 105.

[11] 文勖,张宇,刘挺,等. 基于句法结构分析的中文问题分类 [J]. 中文信息学报, 2006,20(2):33 – 39.

[12] CRAMER K, DEKEL O, KESHET J, *et al.* Online passive-aggressive algorithm[J]. Journal of Machine Learning Research, 2007, (7): 551 – 585.

(**编辑** 张 红)