文章编号:1001-2486(2007)06-0065-05

# 基于自训练EM算法的半监督文本分类*

张博锋,白 冰,苏金树
(国防科技大学 计算机学院,湖南 长沙 410073)

**摘 要**:为了提高计算效率,提出基于自训练的改进EM算法STEM。在每步迭代的E-step中,将中间分类器最有把握对其类别进行预测的未标注样本转移至标注样本集,并应用到M-step中进行下一个中间分类器的训练,从而引入了利用中间结果的自训练机制。文本分类实验表明STEM算法在大部分情况下的分类准确性都高于EM,并通过减少迭代提高了分类器学习的计算效率。

**关键词**:半监督学习;EM算法;自训练;文本分类;naïve Bayes
**中图分类号**:TP181 **文献标识码**:A

# Semi-supervised Text Classification Based on Self-training EM Algorithm

ZHANG Bo-feng, BAI Bing, SU Jin-shu
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

**Abstract**: To improve computation efficiency, an enhanced EM algorithm based on self-training named STEM is proposed. In the E-step of each iteration, the unlabeled sample, whose class can be predicted by the current intermediate classifier with the most confidence, is moved to the labeled set and used in the M-step to train the next intermediate classifier. Therefore the mechanism of self-training by inter-result employing is introduced. Experimentation on text classification indicates that STEM outperforms EM in classification accuracy most of the time and improves the learning efficiency by reducing iterations.

**Key words**: semi-supervised learning; EM algorithm; self-training; text classification; naïve Bayes

利用机器学习的自动文本监督(Supervised)分类是在预先给定的类别(标签)集合下,通过对已标注样本内容特征的学习判定文本的类别,其在自然语言处理与理解、信息过滤与文本挖掘、基于内容的信息安全等领域都有广泛而深刻的背景,是各类监督学习算法如kNN、Rocchio、神经网络、支持向量机及naïve Bayes等研究和应用的经典范例[1-4]。好的分类器需要大量标注(Labeled)样本进行训练,但给出的已标注样本所能提供的信息可能主观而有限;另一方面却有大量更接近样本空间上未知数据分布的未标注样本含有丰富的分布信息。无监督(Unsupervised)学习方法虽然可以在无训练样本的情况下针对样本分布特征进行样本标注,但准确性较差;样本的人工标注需要艰苦而缓慢的劳动,同样制约了整个系统的构建,这就产生了标注瓶颈问题。近年来,利用少量已标注和大量未标注样本训练分类器的半监督学习算法提高了部分分类器的精度,相关研究逐渐引起人们的关注[5]。图1给出了基于监督学习、半监督学习及无监督学习的分类器训练的描述。

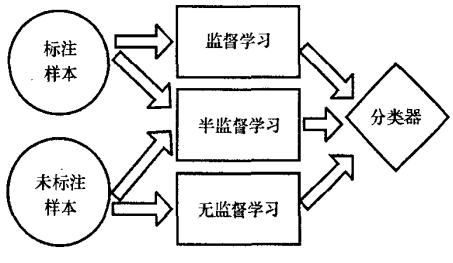


图1 学习方式与数据集的关系
Fig.1 Relation of data sets and learning types

EM(Expectation Maximization)算法是在数据不完整情况下求解最大似然(Maximum Likelihood, ML)或最大后验估计问题(Maximum a Posterior Estimation, MAP)的常用方法[6-7],Nigam在文本分类研究中将它

* 收稿日期:2007-04-18
基金项目:国家自然科学基金重大研究计划资助项目(90604006);教育部高校博士点基金资助项目(20049998027)
作者简介:张博锋(1978—),男,博士生。

用于未标注样本的学习,利用测试样本改进了 Bayes 分类器的分类效果。但在 EM 算法的迭代中,未标注样本总是被设定为"部分地"属于每一个类别,造成中间分类器的更新缓慢,甚至会影响分类精度[8-9]。目前,还没有看到在精度和效率方面针对经典 EM 算法上述不足的优化。

本文提出了基于自训练(Self-training)的 EM 算法,称为 STEM。STEM 会在每一步的迭代中将当前中间分类器分类把握性最大的文本直接加入到标注样本集合,这种自训练的方式可以通过减少迭代加快 EM 算法的迭代速度,同时最终分类器的准确性也得到提高。

## 1 Naïve Bayes 文本分类方法

设类别(标签)的集合为 $C=\{c_1,c_2,\cdots,c_{|C|}\}$,特征(或词)的集合为 $V=\{v_1,v_2,\cdots,v_{|V|}\}$,已标注的样本集合为 $L=\{(d_1,l_1),(d_2,l_2),\cdots,(d_{|L|},l_{|L|})\}$,其中 $l_i\in C$,称为 $d_i$ 的类别(或标签)。在监督学习的条件下,样本集合 $D=L$。所有文本由一个混合的多项式参数模型(Mixture of Multinomials)$\theta$ 产生,每组混合成分组(Mixture Component)都是 $\theta$ 的互不相交子集,并与每个类别一一对应。

长度为 $|d_i|$ 的文本 $d_i$ 的生成方法如下[8-9]:首先根据类别的分布 $P(c_j|\theta)$ 从 $C$ 中选取一个标签 $c_j$ 作为文本的类别,即选择了对应的文本生成的混合成分组。然后根据特征关于类别 $c_i$ 的分布 $P(v_t|c_j;\theta)$ 从 $V$ 中进行 $|d_i|$ 次满足 naïve Bayes 假设并与次序无关的特征选取,假设第 $k$ 次选中特征 $v_{d_{i,k}}$ 进行记录,得到

$$P(d_i|c_j;\theta)=P(|d_i|)\prod_{k=1}^{|d_i|}P(v_{d_{i,k}}|c_j;\theta) \tag{1}$$

即 $d_i$ 根据混合成分组中参数确定的分布 $P(d_i|c_j;\theta)$ 生成,其关于 $\theta$ 的似然函数为:

$$P(d_i|\theta)=\sum_{j=1}^{|C|}P(c_j|\theta)P(d_i|c_j;\theta) \tag{2}$$

每组混合成分中的所有参数实际定义了一个特征集合上的多项式分布,即每个参数定义了一个特征出现的概率,记为 $\theta_{tj}\equiv P(v_t|c_j;\theta)$ 且 $\sum_t P(v_t|c_j;\theta)=1$;另外一类参数定义了不同成分的混合权重,即类别概率,记为 $\theta_j\equiv P(c_j|\theta)$,因此可以记参数 $\theta\equiv\{\theta_{t,j},\theta_j|t=1,\cdots,|V|;j=1,\cdots,|C|\}$。

Naïve Bayes 文本分类器的监督学习过程即通过已标注的样本集合 $D=\{(d_1,l_1),(d_2,l_2),\cdots,(d_{|L|},l_{|L|})\}$ 估计混合模型参数的过程,对 $\theta_{tj}$ 和 $\theta_j$ 的估计分别为:

$$\hat{\theta}_{tj}\equiv P(v_t|c_j;\hat{\theta})=\frac{1+\sum_{i=1}^{|D|}N(v_t,d_i)P(c_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N(v_s,d_i)P(c_j|d_i)} \tag{3}$$

$$\hat{\theta}_j\equiv P(c_j|\hat{\theta})=\frac{1+\sum_{i=1}^{|D|}P(c_j|d_i)}{|C|+|D|} \tag{4}$$

其中,$N(v_t,d_i)$ 表示特征 $v_t$ 在文本 $d_i$ 中出现的次数,根据 $c_j$ 与 $d_i$ 的归属关系,$P(c_j|d_i)\in\{0,1\}$。

在分类过程中通过计算文本的后验概率 $P(c_j|d_i;\hat{\theta})$,最终将 $d_i$ 分入使得后验概率最大的类别,由 Bayes 公式:

$$P(c_j|d_i;\hat{\theta})=\frac{P(c_j|\hat{\theta})P(d_i|c_j;\hat{\theta})}{\sum_{r=1}^{|C|}P(c_r|\hat{\theta})P(d_i|c_r;\hat{\theta})}=\frac{P(c_j|\hat{\theta})\prod_{k=1}^{|d_i|}P(v_{d_{i,k}}|c_j;\hat{\theta})}{\sum_{r=1}^{|C|}P(c_r|\hat{\theta})\prod_{k=1}^{|d_i|}P(v_{d_{i,k}}|c_r;\hat{\theta})} \tag{5}$$

因此,naïve Bayes 的监督学习和分类过程实质上是采用了两个不同的 MAP,其学习过程即寻找分类器

$$\hat{\theta}\equiv\arg\max_\theta P(\theta|L) \tag{6}$$

而分类过程即寻找类别或标签

$$c^* = \arg\max_{c_j} P(c_j \mid d_i; \hat{\theta}) \tag{7}$$

## 2 基于 EM 算法的未标注样本学习

EM 算法可以在数据不完整的情况下迭代求解 MAP 问题[6]。当 naïve Bayes 分类器的训练中引入未标注集合 $U$ 后,整个样本集合 $D = L \cup U = \{(d_1, l_1), (d_2, l_2), \cdots, (d_{|L|}, l_{|L|}), d_{|L|+1}, d_{|L|+2}, \cdots, d_{|L|+|U|}\}$ 可看成一个部分数据 $d_{|L|+1}, d_{|L|+2}, \cdots, d_{|L|+|U|}$ 的标签信息 $l_{|L|+1}, l_{|L|+2}, \cdots, l_{|L|+|U|}$ 缺失的不完整数据集。

EM 算法将未标注样本集 $U$ 结合进 naïve Bayes 的学习[8-9]。首先仅采用标注集合 $L$ 进行初始化训练,得到第一个中间分类器 $\hat{\theta}$,接下来在 E-step 中根据 $\hat{\theta}$ 中的参数值计算所有类关于每个未标注样本的后验概率 $P(c_j \mid d_i; \hat{\theta})$,随后在 M-step 中利用包括了标注和未标注样本的训练集 $D$ 以及 $P(c_j \mid d_i; \hat{\theta})$ 训练新的中间分类器 $\hat{\theta}$。EM 步骤一直迭代,直到 $\hat{\theta}$ 收敛。Dempster 证明了这样的每一轮迭代都会比上一轮得到更加具有相似性的参数估计[6]。

在每一轮中间分类器的参数估计中,式(3)和(4)因有未标注样本的参与,改为:

$$\hat{\theta}_{ij} \equiv P(v_t \mid c_j; \hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} N(v_t, d_i) P(c_j \mid d_i; \hat{\theta})}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(v_s, d_i) P(c_j \mid d_i; \hat{\theta})} \tag{8}$$

$$\hat{\theta}_j \equiv P(c_j \mid \hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} P(c_j \mid d_i; \hat{\theta})}{|C| + |D|} \tag{9}$$

其中,对于 $d_i \in L$,仍旧有 $P(c_j \mid d_i; \hat{\theta}) \in \{0,1\}$,而对于 $d_i \in U$,要注意 $P(c_j \mid d_i; \hat{\theta}) \in [0,1]$,即在不能完全确定未标注文本归属的情况下,认为文本对每个类的训练都有贡献,即文本依照权重 $P(c_j \mid d_i; \hat{\theta})$ 部分地属于每个类别。

可以用隐变量来表示未标注文本缺失的标签信息,对文本 $d_i \in D$,定义 $|C|$ 个隐变量 $z_{ij}$:

$$z_{ij} = \begin{cases} 1, & \text{if } l_i = c_j \\ 0, & \text{otherwise} \end{cases}, \quad j = 1,2,\cdots,|C|$$

在 EM 算法的迭代过程中,E-step 首先通过当前的中间参数计算 $z_{ij}$ 的期望值,这个期望值表示了每个未标注样本在多大程度上可能是某个类别的样本;在 M-step 中 $z_{ij}$ 将被代入下一轮最大化参数的估计中,每个未标注的样本都以自己可能的部分参与每个分类器的训练。基于上面的讨论,EM 算法可以表示为以下形式:

- E-step: Set $\hat{z}^{(t+1)} = \mathrm{E}[z \mid D; \hat{\theta}^{(t)}]$;
- M-step: Set $\hat{\theta}^{(t+1)} = \arg\max_{\theta} P(\theta \mid D; \hat{z}^{(t+1)})$。

在实际的计算中,E-step 利用当前中间参数 $\hat{\theta}$ 计算所有的 $P(c_j \mid d_i; \hat{\theta})(d_i \in U)$,此时有 $z_{ij}$ 的数学期望 $\mathrm{E}[z_{ij} \mid D; \theta] = P(c_j \mid D; \theta) = P(c_j \mid d_i; \theta) \in [0,1]$;对于 $d_i \in L$,$P(c_j \mid d_i; \hat{\theta}) \in \{0,1\}$ 由其已知的标签确定,不需要进行具体的计算。而 M-Step 正是在未标注文本部分属于每个类别的情形下对参数 $\hat{\theta}$ 的估计,即对一个中间分类器的训练。

## 3 基于自训练的 EM 算法 STEM

在上述 EM 算法的每轮迭代中,如图 2(a)所示,始终设定未标注样本 $d_i$ 依照其权重 $P(c_j \mid d_i; \hat{\theta})$ 部分地属于每个类别,即使在分类器对其类别 $c^*$ 非常有把握的情况下,在下一轮计算中 $d_i$ 也只能贡献占其 $P(c^* \mid d_i; \hat{\theta})$ 的部分参与与 $c^*$ 有关的分类器的训练,同时必须考虑其属于其他每个类别的 $P(c_j \mid d_i; \hat{\theta})$ 的可能性。EM 算法对未标注样本不作"激进"的判断和调整,就可能会使得算法的收敛速

度变慢;在算法的设置中,由式(8)和(9),$P(c_j \mid d_i;\hat{\theta})$始终不能为0,因此本应属于其他类别的未标注文本信息也可能会以 $P(c_j \mid d_i;\hat{\theta})$的部分对与其他类别有关的分类器训练产生干扰,虽然这种干扰随着迭代的进行逐渐减轻。同时半监督学习中有一种自训练方式,即分类器在每一轮的训练过程中可以将上一轮分类的肯定结果加入到当前标注样本集,用自己产生的结果再次训练自己,这种方法也可以取得很好的训练效果[10]。

本文将自训练方法部分引入 EM 算法,每次迭代中根据 E-step 的结果直接标注最有把握的一个或多个样本,即将它们直接从未标注样本集转入到标注样本集,M-step 将在新的训练集下进行中间分类器的训练。这样,每一轮的迭代都会使未标注样本集 $U$ 不断缩小,从而使得迭代速度加快,同时提前终止不同类别之间样本对与其他类别有关分类器训练的信息干扰,因为此时对当前刚被加入标注样本集的文本 $d_i \in L$,如图2(b)所示,会根据其类别归属直接设置 $P(c_j \mid d_i;\hat{\theta}) \in \{0,1\}$,它不再部分地属于所有类别,而是完整属于最可能被归入的那个类别,其信息就不会再介入其他类别训练的计算中。

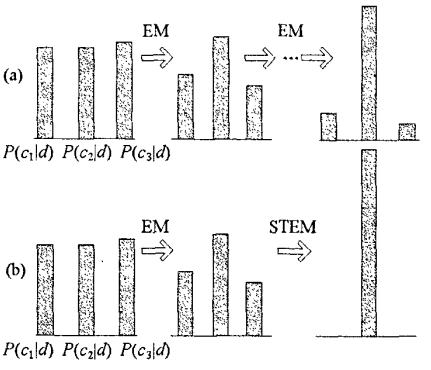


图2 EM与STEM过程示意图

Fig.2 Depiction of the EM and STEM process

表1给出了 STEM 算法的主要步骤。不失一般性,在本文讨论的 STEM 中,在每一轮 E-step 中将会挑出所有 $z_{ij}$ 中最大的 $z_{i^*j^*}$,并将 $d_{i^*}$ 从 $U$ 中剔除并将 $(d_{i^*}, l_{j^*})$归入集合 $L$,同时将 $z_{i^*j}(j \neq j^*)$置0而 $z_{i^*j^*}$ 置1。M-step 在新的更加完整的训练集上进行中间分类器 $\hat{\theta}$ 的估计。事实上每一轮的迭代并不只局限于仅选出一个未标注样本进行标注,在实际应用中可以依照不同的原则,如文本分类中常用的基于排位的 RCut、基于比率的 PCut 以及基于得分的 SCut 等阈值方法[11],在不影响分类效果的情况下,选中多个可以标注的未标注样本,使得程序的效率更高。这是进一步需要详细研究的工作,这里不再进行具体的讨论。

**表1 STEM 算法**

Tab.1 Algorithm of STEM

- Set $t=0$.
- Initialize $\hat{\theta}^{(0)} = \arg\max_{\theta} P(\theta \mid L)$.
- While $U \neq \emptyset$ do:
  - E-step: Set $\hat{z}^{(t+1)} = \mathrm{E}[z \mid D;\hat{\theta}^{(t)}]$.
    - Set $(i^*, j^*) = \arg\max_{(i,j)} \{z_{ij} \mid d_i \in U\}$.
    - Set $L = L \cup \{(d_{i^*}, l_{j^*})\}$.
    - Set $U = U \setminus \{d_{i^*}\}$.
    - For $j=1$ to $|C|$: Set $z_{i^*j}=0$.
    - Set $z_{i^*j^*}=1$.
  - M-step: Set $\hat{\theta}^{(t+1)} = \arg\max_{\theta} P(\theta \mid D;\hat{z}^{(t+1)})$.
  - Set $t = t+1$.
- Output $\hat{\theta}^{(t)}$.

STEM 算法在每一轮迭代中的中间分类器会给最有把握判断其类别的某些未标注样本打上标签并加入标注样本集合,用自训练机制辅助 EM 方法,为后续步骤的中间分类器训练提供了更多质量较高的标注样本,从而控制并减少了算法迭代的最大步骤并有效提高分类器的训练性能。

## 4 实验与分析

实验采用著名的语料 20 Newsgroups[8-9],它共包含了20个类别,超过20 000篇样本基本均匀地分布于类别之间。对文本的预处理包括去停词(Stop Word)和采用 IG 方法的特征选择等[1]。采用与文献[9]

类似的方法,随机地从每个类别选取 20% 的文本作为固定测试样本,随后再从剩余的文本中随机选取 10 000篇作为固定的未标注样本集。

首先比较 STEM、EM 以及普通的 naïve Bayes 算法的分类效果。通过不断变化标注样本集的大小,得出算法的微平均精度 (Micro-average Accuracy)[1] 与标注样本集大小间的关系曲线,如图 3 所示。可以看出,两种半监督学习方法在引入未标注样本后,都在一定程度上提高了分类器的准确性,特别是在标注样本数目较少的情况下,性能有超过 10% 以上的差距。除了标注样本数量最少的情况下,STEM 的分类准确性较 EM 算法略低外,在绝大部分情况下 STEM 的分类准确性都高于 EM,例外的产生的一个可能原因是由于标注样本数量过于稀少,因此产生的中间分类器的准确性较差,在初期的误判率相对较高,分类器在自训练中学习了被自己错误标注的样本,此时比较保守的训练方法占优,但随着标注样本数量的增加,中间分类器的准确性一旦有所提高,这一影响便会大大降低。

图 4 比较了在不同大小的未标注样本集合下 EM 算法和 STEM 算法的迭代次数,这里 EM 算法采用了文献[8]中的终止条件。可以看出 STEM 算法的迭代次数基本与未标注样本的数目相同,并在一定程度上低于 EM 算法的迭代次数,因此,STEM 的计算效率也要高于 EM。

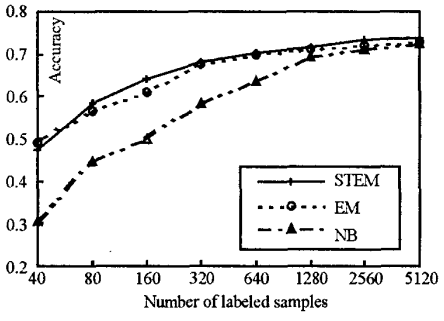


图 3 精度比较
Fig.3 Comparison of accuracy

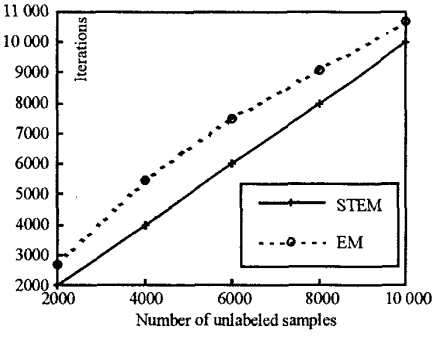


图 4 迭代次数比较
Fig.4 Comparison of iterations

## 5 结论

本文提出了基于自训练的 EM 算法 STEM。STEM 在迭代过程中会将当前中间分类器判断把握性最大的某些未标注样本从未标注样本集中剔除并打上相应标签加入到标注样本集合,这种自训练的方式为后续步骤的中间分类器训练提供了更多的标注样本,通过未标注集的不断减小加快迭代速度,也不断消除不同类别之间样本的信息干扰。实验证实 STEM 算法在大部分情况下都得到了高于 EM 的分类准确性,并大大提高了计算过程的效率。

## 参考文献:

[1] Sebastiani F. Machine Learning in Automated Text Categorization [J]. ACM Computing Surveys, 2002, 34(1):1 - 47.
[2] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术进展 [J]. 软件学报, 2006, 17(9): 1848 - 1859.
[3] 王建会, 王洪伟, 申展, 等. 一种实用高效的文本分类算法 [J]. 计算机研究与发展, 2005, 42(1): 85 - 93.
[4] 黄萱菁, 夏迎炬, 吴立德. 基于向量空间模型的文本过滤系统 [J]. 软件学报, 2003, 14(3): 435 - 442.
[5] Zhu X. Semi-supervised Learning Literature Survey [R]. Technical Report TR 1530, Computer Sciences, University of Wisconsin-madison, December 2006.
[6] 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 (Methodological), 1997, 39(1): 1 - 38.
[7] 孙广玲, 唐降龙. 基于分层高斯混合模型的半监督学习算法 [J]. 计算机研究与发展, 2004, 41(1): 156 - 161.
[8] Nigam K. Using Unlabeled Data to Improve Text Classification [D]. Ph. D. Dissertation, Computer Science Department, Carnegie Mellon University, May 2001.
[9] Nigam K, McCallum A, Mitchell T. Semi-supervised Text Classification Using EM [C]//Semi-supervised Learning, MIT Press: Boston, 2006.
[10] Nigam K, Ghani R. Analyzing the Effectiveness and Applicability of Co-training [C]//Proceedings of the 2000 ACM CIKM, McLean, US, 2000: 86 - 93.
[11] Yang Y A. Study on Shareholding Strategies for Text Categorization [C]//Proceedings of 2001 ACM SIGIR, New Orleans, US, 2001: 137 - 145.