一种增量式 Bayes 文本分类算法

高　洁,吉根林

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要]　文本自动分类是数据挖掘和机器学习中非常重要的研究领域.针对难以获得大量有类标签的训练集问题,提出了基于小规模标注语料的增量式 Bayes 文本分类算法.该算法分两种情况处理:第一种情况是新增样本有类标签,可直接重新计算样本属于某类别的条件概率.第二种情况是新增样本无类标签,则利用现有分类器为其训练类标签,然后利用新样本来修正分类器.实验结果表明,该算法是可行有效的,比 Naïve Bayes 文本分类算法有更高的精度.增量式 Bayes 分类算法的提出为分类器的更新提供了一条新途径.

[关键词]　文本分类,增量学习,Naïve Bayes

[中图分类号]TP393,　[文献标识码]A,　[文章编号]1672-1292-(2004)03-0049-04

0　引言

文本自动分类是数据挖掘和机器学习中非常重要的研究领域,是指在给定分类体系下,根据文本内容自动确定文本类别的过程.目前较为著名的文本自动分类方法有:Naïve Bayes (NB)、最近邻(KNN)、支持向量机(SVM)、决策树(Decision Tree)方法等.NB 方法属于有指导学习,需要对大量的有类标签的训练文档集合进行训练,才能够得到正确的分类器.但在实际运用中,未必能够找到大量的已经正确标注类别的文档集合,而未标注的文档集合却极其丰富.与一般的分类问题相比,文本分类面临一个突出的问题,即属性的个数众多.文本分类中的属性指的是文本中的特征词,特征词的数目是非常庞大的,这些大量的特征词之间不可避免存在某些依赖关系.当引入多个训练实例,由于特征间的关联,基于独立性假设的 NB 误差率必然会增大,Rish[1] 指出 NB 在低熵的情况下会获得更大的精度.另外,对于分类,一般的统计分析方法是将所有训练数据一次读入内存.一方面很有可能造成内存空间的不足;另一方面,当数据信息是分批获得的时候,传统的统计分类方法更显出它的局限性.基于上述问题,本文提出一种基于小规模标注语料的增量式 Bayes 文本分类算法.

1　Naïve Bayes 文本分类

定义 1　设 D 表示训练集,由 N 个训练样本组成,$D = \{d_1, d_2, \cdots d_N\}$;文档 d 由其所包含的特征词表示,即 $d = \{w_1, w_2, \cdots w_n\}$;$C_1, C_2, \cdots C_m$ 表示文本的类别.

NB 方法利用下列 Bayes 公式通过类别的先验概率和词的分布来计算未知文本属于某一类别的概率:

$$P(C_i \mid d) = \frac{P(C_i)\ P(d \mid C_i)}{P(d)} \tag{1}$$

式中:$P(C_i \mid d)$ 为样本 d 属于类 C_i 的概率,$P(d \mid C_i)$ 为类 C_i 中含有样本 d 的概率.在所有 $P(C_i \mid d)$ $(i = 1, \cdots, m)$ 中,若 $P(C_k \mid d)$ 值最大,则文本 d 归为 C_k 类.由于 $P(d)$ 是常数,因此将要求解 $P(C_i \mid d)$ 的问题转化为只要求解 $P(C_i)\ P(d \mid C_i)$.在特征独立性假设前提下:

$$p(C_i \mid d) \approx P(C_i)\ P(d \mid C_i) = P(C_i) \prod_{j=1}^{M} P(w_j \mid C_i) \tag{2}$$

为了避免类概率 $P(C_i)$ 为 0,采用拉普拉斯概率估计得到:

$$\text{类概率 } P(C_i) = \frac{1 + |D_{C_i}|}{|C| + |D|} \tag{3}$$

式中:$|D_{C_i}|$ 为训练集中属于类 C_i 的文档数;$|C|$ 为训练集中类的数目;$|D|$ 为训练集中所包含的文档总数.

采用文档型计算公式计算条件概率 $P(w_j \mid C_i)$,不考虑特征词在文档中出现的频次,仅考虑特

收稿日期: 2004-04-22.
基金项目: 江苏省重点实验室开放基金资助项目(KJS03064).
作者简介: 高洁(1979 -),女,硕士,助教,主要从事数据库和数据挖掘技术的研究. E-mail:scarletg@tom.com
通讯联系人: 吉根林(1964 -),教授,主要从事数据库和数据挖掘技术的教学与研究. E-mail:gji@njnu.edu.cn

征词在文档中是否出现,0 表示未出现,1 表示出现,依下式计算:

条件概率 $P(w_j \mid C_i) =$

$$\frac{1 + C_i \text{ 类文本中出现特征 } w_j \text{ 的文本数}}{2 + |D|} \tag{4}$$

2　增量式 Bayes 文本分类

Bayes 概率是观测者对某一事件发生的相信程度,观测者根据先验信息和现有的统计数据,用概率的方法来预测未知事件发生的可能性,正是这种充分利用先验信息的特性,使之成为增量学习中的最佳选择模型.

2.1　Bayes 增量学习模型

记参数 θ 为事件发生的先验概率,$P(\theta \mid I_0)$ 为它的概率密度函数,其中 I_0 为观测者的先验信息. 根据 Bayes 规则,新的训练样本 S 的加入,由先验概率密度 $P(\theta \mid I_0)$ 计算后验概率密度 $P(\theta \mid S, I_0)$ 的公式为:

$$P(\theta \mid S, I_0) = \frac{P(S \mid \theta, I_0)\, P(\theta \mid I_0)}{P(S \mid I_0)} = \frac{P(S \mid \theta, I_0)\, P(\theta \mid I_0)}{\int \left(P(S \mid \theta, I_0)\, P(\theta \mid I_0)\right) \mathrm{d}\theta} \tag{5}$$

式(5) 综合了样本信息 S 和先验信息 I_0,这正是增量 Bayes 学习模型的基础,即:

后验信息(I_1) = 先验信息(I_0) + 样本信息(S)

当新的样本到来时,原来的后验信息就变成了先验信息,因此增量 Bayes 学习是一个利用样本信息来修正当前信息的连续动态过程. 但是后验信息必须与先验信息属于同分布,才可以利用后验信息作为进一步实验的先验信息.

定义 2　设样本 S 对参数 θ 的条件分布为 $P(S \mid \theta, I_0)$,先验分布为 $\pi(\theta)$,如果由 $\pi(\theta)$ 和样本 S 决定的后验分布 $P(\theta \mid S, I_0) = \frac{P(S \mid \theta) \times \pi(\theta)}{P(S)}$ 与 $\pi(\theta)$ 同分布,则称 $\pi(\theta)$ 为共轭分布[2].

共轭分布要求先验信息和现在的样本信息有某种共性,这样才能再次转化为同一类型的先验信息.

定义 3　设事件变量 γ 有 $\gamma^1, \gamma^2, \cdots, \gamma^r$ 共 r 个可能的状态,参数向量为 $\theta = (\theta_1, \theta_2, \cdots, \theta_r)$,其中 $\theta_k = P(\gamma = \gamma^k \mid \theta, I_0)$,$k = 1, 2, \cdots, r$,若它们的分布密度满足:

$$P(\theta \mid I_0) = \mathrm{Dir}(\alpha_1, \alpha_2, \cdots \alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1},$$

其中 $\alpha = \sum_{k=1}^{r} \alpha_k$,且 $\alpha_k > 0$.

则称 θ 关于先验信息 I_0 具有 Dirichlet 分布[3]. 其中,$\alpha_1, \alpha_2, \cdots \alpha_r$ 为超参数,当分布均匀时,$\alpha_i = 1$,($i \in 1, 2, \cdots, r$).

如果样本 S 中 $\gamma = \gamma^k$ 出现的次数为 n_k ($k = 1, 2, \cdots, r$),根据共轭分布的定义,它的后验分布(即预测事件的 Bayes 概率) 也服从 Dirichlet 分布:

$$P(\theta \mid S, I_0) = \mathrm{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, \cdots, \alpha_r + n_r), \tag{6}$$

以 θ 的后验条件期望作为它的估计值:

$$\hat{\theta}_k = E\{\hat{\theta}_k \mid S, I_0\} = \frac{\int_0^1 \theta_k \prod_{j=1}^{r} \theta_j^{\alpha_j + n_j - 1} \mathrm{d}\theta_k}{\mathrm{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, \cdots, \alpha_r + n_r)} = \frac{\alpha_k + n_k}{n + \alpha} = \frac{\alpha}{n + \alpha} \cdot \frac{\alpha_k}{\alpha} + \frac{n}{n + \alpha} \cdot \frac{n_k}{n} \tag{7}$$

上式具有明显的统计意义,它表示后验估计是它的先验参数与样本中出现次数的一种加权平均.

2.2　增量式 Bayes 文本分类算法

对于文本分类,增量分类的任务实质是如何根据先验信息和样本信息来确定 $P(C_i)$ 和 $P(w_j \mid C_i)$. 取无信息 Dirichlet 先验,可获得下面的参数估计:

$$P(w_j \mid C_i) = \frac{1 + \text{训练样本中属于 } C_i \text{ 类的含有 } w_j \text{ 的文本数}}{1 + |D_{C_i}|} \tag{8}$$

$$P(C_i) = \frac{1 + |D_{C_i}|}{|C| + |D|} \tag{9}$$

增量式 Bayes 分类主要是针对新增加的样本集 $T = \{t_1, \cdots, t_m\}$ 的处理,分两种情况来讨论:

第一种情况:T 中的每个文本都有类标签. 可直接根据以下公式重新估计条件概率和类概率:

条件概率 $P^*(w_j \mid C_i) =$

$$\frac{1 + \text{训练样本和新增样本中属于 } C_i \text{ 类的含有 } w_j \text{ 的文本数}}{1 + |D_{C_i}| + |T_{C_i}|} \tag{10}$$

$$\text{类概率 } P^*(C_i) = \frac{1 + |D_{C_i}| + |T_{C_i}|}{|C| + |D| + |T|} \tag{11}$$

其中,$|T_{C_i}|$ 表示新增样本集中属于类 C_i 的样本数,$|T|$ 表示新增样本集的大小.

第二种情况:T 中的每个文本都没有类标签.

需要对每个文本用现有的分类器为它分配类标签,并同时考虑能够利用这些新文本中含有的有用信息来修正当前的分类器.首先从 T 中选择有助于提高当前分类器精度的文本,把它加入到训练集中,来修正当前的分类参数,直至 T 中的文本全部加入到训练集中.衡量分类器精度的标准是它的分类效果.假定分类为 0 - 1[4] 损失,选择 $t_p \in T$,在当前分类器下获得类标签 C_p,由以下公式计算在 $D^* = D + \{t_p\}$ 下估计测试集 $S = \{s_1, \cdots, s_m\}$ 的分类损失:

$$L(D^*) = \frac{1}{|S| - 1}\sum_{s}(1 - \max p_{D^*}(C_i \mid s)) \quad (12)$$

下面给出增量式 Bayes 文本分类算法.

输入:训练集 $D = \{d_1, \cdots, d_N\}$,

新增实例集 $T = \{t_1, \cdots, t_m\}$,

测试集 $S = \{s_1, \cdots, s_m\}$.

输出:分类器 C.

Step1. 利用训练集 D,学习分类器 C;

$$\text{类概率 } P^*(C_i) = \left\{ \begin{array}{l} \dfrac{|C| + |D|}{|C| + |D| + 1} P(C_i), \text{当 } C_p \neq C_i \\ \dfrac{|C| + |D|}{|C| + |D| + 1} P(C_i) + \dfrac{1}{|C| + |D| + 1}, \text{当 } C_p = C_i \end{array} \right\} \quad (13)$$

$$\text{条件概率 } P^*(w_j \mid C_i) = \left\{ \begin{array}{l} \dfrac{1 + |D_{C_i}|}{2 + |D_{C_i}|} P(w_j \mid C_i), \text{当 } C_p = C_i \text{ 并且 } w_j \notin t_p \\ \dfrac{1 + |D_{C_i}|}{2 + |D_{C_i}|} P(w_j \mid C_i) + \dfrac{1}{2 + |D_{C_i}|}, \text{当 } C_p = C_i \text{ 并且 } w_j \in t_p \\ P(w_j \mid C_i), \text{当 } C_p \neq C_i \end{array} \right\} \quad (14)$$

由于新样本 t_p 的加入,使得先验信息中加入了样本信息.后验信息也就由先验信息和新增样本信息共同确定.通过分析上述公式可知,t_p 加入到训练集后,只有与它相关的项的估计变化较大,而与它无关的项的变化相对较小.所以,在使用加入新样本 t_p 后的 D^* 来分类测试集合 S 的时候,可以忽略与 t_p 无关的项的影响,采用下式:

$$P^*(C_q \mid s_q) = \frac{P(C_q \mid s_q)\prod_{w \in A_q \cap A_p} p^*(w \mid C_q)}{\prod_{w \in A_q \cap A_p} p(w \mid C_q)} \quad (15)$$

其中 $p(w \mid C_q)$ 是训练集 D 训练出来的分类参数,而 $p^*(w \mid C_q)$ 是训练集 D 增加了文本 t_p 之后得到的分类参数.这样,计算只在与 S 中相同的词之间进行,大大地降低测试时的计算复杂度.

3 实验设计

实验的数据主要来自南大小百合 bbs 新闻版块,共有体育、科技、娱乐 3 大类别.

增量式 Bayes 文本分类实验:从训练文档集中选取 9 篇作为初始训练集(3 大类,每 1 类各 3 篇),将剩下的有类标签的文本,看作没有类标签的文本,作为新增文本集(假设新增文本集全部是没有类标签的).按照上文提到的增量式 Bayes 文本分类算法进行分类器训练.

Naïve Bayes 文本分类实验:利用有类标签的训练集采用 Naïve Bayes 分类算法进行分类器训练.

共进行 4 组实验,所有实验均采用相同的特征选择方法文档频率 DF,相同的文本表示,相同的测试集(每类测试集 150 篇,所有测试集与训练集无重叠),采用 F1 值进行性能评价,结果如表 1 所示:

通过实验证明,在具有较少类标签样本的情况下,通过增量学习,可以不断提高分类器的精度.该算法尤其适用于难以获得大量类标签的文本自动分类领域,如 Web 文档的分类.

Step2. If $T = \Phi$, Return C; Else go on;

Step3. 令 $l = 0$;对 T 中的每一个元素 $t_p \in T$,重复:

(1) 利用 C,分类 t_p,得到类标签 C_p,

(2) 利用 $D + \{< t_p, C_p >\}$,对 S 中的每一个元素分类,并计算分类损失之和 L,

(3) if $L > l$ then $l = L$, $x = t_p$, $C = C_p$;

Step4. $D = D + \{< t_p, C_p >\}$, $T = T - x$, go to Step1.

算法主要是利用无标签文本中的有用信息修正分类参数得到更精确的分类器.因为该算法是一个嵌套循环的过程,若每次都直接完成上述操作,其复杂程度是相当高的.而 Bayes 本身具有增量学习特性,使得某些操作可以增量地进行,从而大大地降低了算法的复杂度.下面给出上述算法中的增量学习公式.

假设选择 t_p 加入到训练集中,分类器分类 t_p 为 C_p,根据 Dirichlet 先验分布特性重新给出如下的类概率和条件概率公式:

表 1　增量式 Bayes 分类器与 Naïve Bayes 分类器的性能比较

	主题类(训练样本集)	F1 值	
		增量式 Bayes (每类 150 篇)/ %	Naïve Bayes (每类 150 篇)/ %
实验一	体育类(100 篇)	91.1	90.2
	科技类(100 篇)	90.2	89.4
	娱乐类(100 篇)	90.8	90.2
实验二	体育类(300 篇)	96.2	92.8
	科技类(300 篇)	95.7	91.3
	娱乐类(300 篇)	95.8	92.6
实验三	体育类(100 篇)	91.1	90.2
	科技类(200 篇)	92.6	90.2
	娱乐类(300 篇)	95.8	92.6
实验四	体育类(300 篇)	96.2	92.8
	科技类(200 篇)	92.6	90.2
	娱乐类(100 篇)	90.8	90.2

4　结束语

NB 分类由于它的简单和易于实现受到越来越多的关注. 基本的 NB 算法需要对有类标签的大量文本进行训练,但实际中很难得到较为完备的训练文本库. Dempster *et al* 于 1977 年提出了 EM(Expectation-Maximization)算法的理论框架,将迭代最大似然估计缺失数据的技术形式化. 在文本分类中, Nigam 等人将 EM 算法引入到 NB 分类[5],在分类器训练中可以处理未标识文本. 首先利用由少量有类标签的文本组成的原始训练集合,初始化 Bayes 分类器的参数,然后利用 EM 算法自动增加训练量以获得局部最优解,调整 Bayes 分类器. EM 算法一个最大的缺陷就是计算量大,收敛缓慢. 增量式 Bayes 算法和 EM 方法的最大不同在于前者通过最小化当前分类器的分类损失,来选择有利于提高分类器性能的文本加入训练集,即保证每次加入训练集的都是相对最优的,因为这些文本本身也含有对分类类别有用的信息,所以如何有效地利用这些信息进行增量分类成为问题的关键.

增量式 Bayes 算法的提出,为分类器的更新提供了一种新的途径. 新的更加完备的文本信息出现,需要对原有的分类器进行更新,才能保证分类的性能得到不断提高.

[参考文献]

[1] Rish I. An empirical study of the naïve Bayes classifier[C]. IJCAI-01 workshop on "Empirical Methods in AI" Technical reports, 2001.215.

[2] SAMUEL KOTZ. Modern Bayesian Statistics[M]. George Washington University Press, 2000.109.

[3] Geiger D, Heckeman D. A characterization of the Dirichlet distribution with applicable to learning Bayesian network[A]. Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence[C]. Montreal, QU, 1995.196-207.

[4] Dominigos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss[J]. Machine Learning, 1997, 29(2-3):103-130.

[5] Kamal Nigam, *et al*. Learning to classify the text from labeled and unlabeled documents[A]. Proc 15th National Conference on Artificial Intelligence[C]. Wisconsin, 1998.792-799.

Incremental Bayes Text Categorization Algorithm

GAO Jie, JI Genlin

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

Abstract: Automatic text categorization is an important research field in data mining and machine learning. An incremental Bayes text categorization algorithm based on small labeled documents is presented to solve the difficult problem involving getting labeled training documents. The algorithm can process two cases: the labeled and unlabeled incremental documents. Directly computing the probability of the samples of a certain class is the processing method for labeled documents. The unlabeled documents are labeled first by using the original classification, and then the new classification is trained from the incremental documents. The experimental results show that this algorithm is feasible and effective, providing a new method for updating of classification.

Key words: text categorization, incremental learning, Naïve Bayes

[责任编辑:刘健]