

基于改进隐马尔可夫模型的文本分类研究

李开荣^{1,2}, 孔照昆¹, 陈桂香¹, 朱俊武¹

(1 扬州大学 信息工程学院, 江苏 扬州 225127; 2 扬州大学 水利科学与工程学院, 江苏 扬州 225127)

摘 要: 将一种改进的隐马尔可夫模型(HMM)应用于文本分类中,在考虑其前向依赖的同时,需考虑状态的后向依赖性.将当前观测值和当前状态对其后一状态的依赖性加入模型的学习,这样的改进模型能有效提高文本信息抽取准确率.在文本分类过程研究中,首先对训练样本进行文本预处理,对 HMM 分类器模型进行参数学习,建立 HMM 分类器后用测试集进行测试并做出性能评价.在性能评价中用改进的评测指标,可针对不同数据集做出准确评价,以及可对比不同分类工作在同一数据集上的性能,大大提高评价质量.

关键词: 隐马尔可夫模型;文本分类;期望交叉熵(ECE); χ^2 统计;TF-IDF 方法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)11-0161-05

Research on Improved HMM-Based Text Categorization

LI Kai-rong^{1,2}, KONG Zhao-kun¹, CHEN Gui-xiang¹, ZHU Jun-wu¹

(1 School of Information Engin, Yangzhou University, Yangzhou 225127, China;

2 School of Hydrant Sci & Engin, Yangzhou University, Yangzhou 225127, China)

Abstract: The application of the improved Hidden Markov Models to text categorization should take the backward dependency as well as forward dependency on states into account. The accuracy of information extraction could be improved by applying the dependency of the current observation value and state on the backward ones into the learning of models. This paper is to preprocess the training samples in text categorization process, to learn the parameters of HMM classifier, establishing one and evaluating its performance through testing set. The improved evaluation criteria could give a fair evaluation of different dataset, make a comparison of different classifiers on the performance of the same dataset and feed back the classifier to improve.

Key words: hidden markov models; text categorization; expected cross entropy(ECE); χ^2 statistic; TF-IDF method

1 引言

在文本分类技术研究领域,各种各样的文本分类算法设计是研究的核心技术,它为用户处理和组织大量文本数据,准确定位所需的信息和分流信息提供捷径.目前,基于集成学习算法^[1]和支持向量机^[2]等的一些算法已经在文本分类中取得了良好的分类效果.

隐马尔可夫模型(hidden markov models, HMM)是一种用参数表示,用于描述随机过程统计特性的概率模型.因其具有成熟的算法以及处理数

据时表现出很好的鲁棒性,被广泛应用于各领域,如语音识别^[3]、处理自然语言^[4-5]、文本分类^[6-8]等领域.文献[6]主要是 HMM 对多页面文档中文本分类的研究,通过整个页面序列提供的上下文信息帮助消歧,提高单页分类精度;文献[7]在考虑中文检索模型中汉语言的特点,针对隐马尔可夫模型存在的近距离依存进行改进,提高模型的检索准确率;文献[8]提出了隐马尔可夫序列分析模型用于自动文本分类的算法,该算法构造的 HMM 分类器一定程度上体现了不同类别文档的语法和语义特征.

本文在阐述了基于 HMM 模型的文本分类过

程后,主要研究改进后的 HMM 模型的训练过程和性能评价.对改进的 HMM 分类器的参数进行调整,给出对 HMM 分类器的性能评价,以实现大量文本信息的智能化处理.

2 改进的隐马尔可夫模型

HMM 是在 Markov 链模型的基础上发展起来的.传统的隐马尔可夫模型具有良好的抽取效果,但其具有前向依赖性,即只考虑对当前状态的依赖、当前状态对前一状态的依赖,忽视可能存在的当前观测值与当前状态对后一状态的依赖性.该改进模型就增加双向考虑,这一过程如图 1 所示.

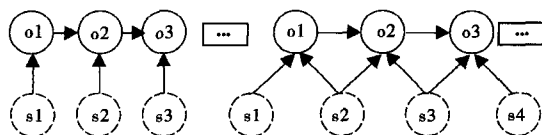


图 1 传统 HMM 与改进的 HMM

将 HMM 描述为一个新六元组

$$K = \{N, M, l, A, B, BB\}^{[9]};$$

(1) N 为状态数,状态集合 $S = (s_1, s_2, \dots, s_N)$, t 时刻的状态为 q_t .

(2) M 为每个状态对应的观测数,观测值集合 $W = (w_1, w_2, \dots, w_M)$.

(3) λ 为模型的初始状态概率矢量, $\lambda_i = P(q_1 = s_i)$, $\sum_{i=1}^N \lambda_i = 1$, $\pi_i \geq 0$.

(4) A 为状态转移概率矩阵, $A = (a_{ij})_{N \times N}$, a_{ij} 是 t 时刻所处的状态 s_i 转移到 $t+1$ 时刻的状态 s_j 的状态转移概率,即

$$\begin{aligned} a_{ij} &= P(q_{t+1} = s_j | q_t = s_i), \\ \sum_{j=1}^N a_{ij} &= 1, \\ a_{ij} &\geq 0, \quad 1 \leq i \leq N \end{aligned} \quad (1)$$

(5) B 为观测值概率矩阵, $B = (b_i(o_l))_{N \times M}$, $b_i(l)$ 表示 t 时刻所处的状态 s_i 输出的观测值为 v_l 的概率,即

$$\begin{aligned} b_i(o_l) &= P(o_l = w_l | q_t = s_i), \\ 1 \leq i \leq N, \quad 1 \leq l \leq M \end{aligned} \quad (2)$$

(6) 改进的模型加入后一时刻的考虑,即 t 时刻的观测值 o_t 的出现取决于当前状态 s_i 及后一时刻 $t+1$ 时刻的状态 s_j . 据此,发射概率矩阵记为 $BB = (bb_{ij}(o_l))_{N \times M}$:

$$\begin{aligned} bb_{ij}(o_l) &= P(o_l = w_l | q_t = s_i, q_{t+1} = s_j), \\ 1 \leq i \leq N, \quad 1 \leq l \leq M \end{aligned} \quad (3)$$

3 基于改进隐马尔可夫模型的文本分类过程

3.1 文本分类过程

文本分类的过程主要包括学习过程和分类过程,如图 2 所示.学习过程可分为训练过程和反馈过程,将训练过程生成 HMM 分类器进行性能评价,通过这种评价——反馈——改进训练过程生成的新分类器对新文本分类,最终输出新文本的类别.

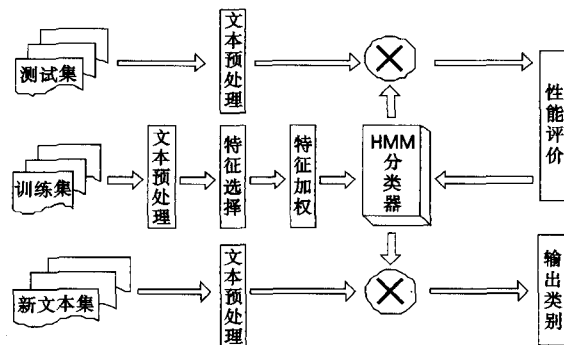


图 2 文本分类过程

3.2 训练过程

3.2.1 文本预处理

假设训练文本集为 $N = \{n_1, n_2, \dots, n_{|N|}\}$, 文本类别集合为 $C = \{c_1, c_2, \dots, c_{|C|}\}$, $|N|$ 为文本集的数目, $|C|$ 是类别集的个数. 文本预处理步骤如下:

① 处理文本中存在的标点符号、数字及大小写,使文本变成只包含词条的字符串;

② 分词处理. 英文处理是以空格或标点将词隔开;中文处理由于词与词之间没有分隔符,所以词的获取需要通过分词技术来实现;

③ 取词根,即将文本中相同或相近含义的词合并成一个语义单位. 主要方法有:基于规则的方法和基于词典的方法;

④ 删除文本中的停用词和稀有词.

3.2.2 特征选择

文本特征选择可以提高文本分类的准确率和运行速度. 常用方法有:信息增益 (IG)、期望交叉熵 (ECE)、 χ^2 统计量等.

假设类别 $c_j \in C$, 词 t 为原始特征集合中的一个特征. 通过期望交叉熵的方法来选取特征词,期望交叉熵只考虑特征在文本中的发生情况. 词 t 与类 c_j 的期望交叉熵 $CE(t)$ 的评估函数为:

$$CE(t) = P(t) \sum_{j=1}^m P(c_j | t) \log \frac{P(c_j | t)}{P(c_j)} \quad (4)$$

其中, c_j 为第 j 类文档集合, $m = 1, 2, \dots, |C|$,

m 为类别数, $P(c_j | t)$ 为文档中含有特征词 t 时属于 c_j 类的概率, $P(c_j)$ 为训练集中文本属于 c_j 类的概率, 需要选取特征项的信息增益最大的词作为特征词, 选好的特征词构成特征集 W .

3.2.3 特征加权(特征抽取)

由于词与类别之间没有独立性, 它们之间的关联度越高, 则独立性越小, 因此就可以区别不同类别的不同特征词. 我们采用信息增益的方法进行评估, 其评估函数为:

$$IG(t) = P(t) \sum_j P(c_j | t) \log \frac{P(c_j | t)}{P(c_j)} + P(\bar{t}) \sum_j P(c_j | \bar{t}) \log \frac{P(c_j | \bar{t})}{P(c_j)} \quad (5)$$

其中 c_j 为第 j 类文档集合. 对于特征词 t 和文档类别 c_j , IG 考虑的是 c_j 类中特征词 t 出现和不出现的概率来确定该特征词 t 对 c_j 类的信息增益. 特征项信息增益值越高, 则词与类别之间的独立性越小, 选择信息增益大的词作为特征词.

通过设定阈值, 按照升序排列, 就可以找到与 c 类最关联的特征词集合 $W_c = \{w_{c_1}, w_{c_2}, \dots, w_{|c|}\}$, 所以得到类集合 C 的特征集为 $W_C = \{W_1, W_2, \dots, W_{|C|}\}$.

3.2.4 建立类别集的 HMM 分类器

在类 c 的 HMM 分类器中, S_0 和 S_i 为初始状态和结束状态, 中间状态集合为 $S_c = \{S_1, S_2, \dots, S_{|c|}\}$, 输出观测值集合即为特征词集合 $W_c = \{w_{c_1}, w_{c_2}, \dots, w_{|c|}\}$. 由于状态转移是向该类别特征逐步逼近, 表现为状态为顺序输出, 就有 $P(S_{i+1} | S_i) = 1$, 得到类 c 分类器的状态转移概率矩阵 A_c .

状态 S_i 的观测值分布为 $b_{ij}^{(c)}(k) = P_c(w_k | S_i)$, w_k 为特征集 W 第 k 个特征词, 而状态 S_i 所对应的输出概率为 w_i , 所以

$$b_i^{(c)}(k) = P_c(w_k | w_i) = \frac{P_c(w_k, w_i)}{P_c(w_i)} = P_c(w_k, w_i) \times \frac{1}{P_c(w_i)} \quad (6)$$

$P_c(w_k, w_i)$ 为衡量 w_k 与类 c 中的特征词 w_i 的相关度, 可使用 χ^2 统计计算:

$$P_c(w_k, w_i) = \frac{\chi^2(w_k, w_i)}{\sum_{i=1}^{|C|} \chi^2(w_k, w_i)} \quad (7)$$

其中,

$$\chi^2(w_k, w_i) = m(m_{11} \times m_{22} - m_{12} \times m_{21})^2 (m_{11} + m_{12})^{-1} (m_{21} + m_{22})^{-1} (m_{11} + m_{21})^{-1} (m_{12} + m_{22})^{-1} \quad (8)$$

其中 $m_{11}, m_{12}, m_{21}, m_{22}$ 分别表示 w_k 和 w_i 的 4 种情况的频数, 这 4 种情况分别为: 类 c 包含 w_k 和 w_i 的概率, 类 c 包含 w_k 而不包含 w_i 的概率, 类 c 不包含 w_k 而包含 w_i 的概率, 类 c 既不包含 w_k 也不包含 w_i 的概率; 其中 $m = m_{11} + m_{12} + m_{21} + m_{22}$ 为总数. $\sum_{i=1}^{|C|} \chi^2(w_k, w_i)$ 为归一化分母.

对于 $\frac{1}{P_c(w_i)}$, 采用改进的 TFIDF 方法^[10], 则

$\frac{1}{P_c(w_i)} = \text{IDF} = \log\left(\frac{m}{m+k}\right) \times N$, m 表示类 c 中包含特征词 w_i 的文档数; k 表示除类 c 外包含特征词 w_i 的文档数; N 为总文档数. 改进的 TFIDF 方法的思路为: 如果类 c 中包含特征词 w_i 的文档数量大, 而在其他类中包含特征词 w_i 的文档数量小的话, 则 w_i 能够代表类 c 的特征. 由式(6)得到观测输出概率矩阵 B_c .

在改进的模型中用最大似然法来估计参数模型, BB 计算如下:

$$bb_{ij}^{(c)}(k) = \frac{\text{Emission}(i, j, k)}{\sum_{i=1}^M \text{Emission}(i, j, k)} \quad (9)$$

其中, $\text{Emission}(i, j, k)$ 表示在训练样本中状态 S_i 与 S_j 为前后状态, S_i 输出符号 w_k 的次数.

所以, 类 c 的 HMM 分类器的模型可表示为 $K = \{S_0, A_c, B_c, BB_c\}$. 那么在文本类别集合中, $HMM_C = \{HMM_1, HMM_2, \dots, HMM_{|C|}\}$.

3.3 性能评价

3.3.1 宏观平均值(Macro-average)

首先分别求得分类器对各类别分类的准确率 p 和召回率 r . 然后, 在此基础上计算分类器对文本集中所有类别分类的准确率和召回率的平均值.

宏平均准确率和宏平均召回率的计算公式:

$$\text{Macro}_p = \frac{\sum_{i=1}^{|C|} p_i}{|C|} \quad (10)$$

$$\text{Macro}_r = \frac{\sum_{i=1}^{|C|} r_i}{|C|} \quad (11)$$

其中, C 为文本集中类别集, $|C|$ 为类别集 C 中类别的数目.

数据集不同时评测数值会产生剧烈波动. 现考虑如下三因素: ①训练集中的文档比例 dm , 当训练集中类 C_i 文档比例增加, 分类器向该类别偏离; ②类半径 R , 随类半径的增大, 分类质量下降; ③类间

距,类别间的距离 width,随着类间距的减小,分类质量下降.根据如上 3 个结论评测指标^[11]为:

$$\text{macro_average_}F_1 = \theta \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{F_{1i} \log(R_i)}{dm_i} \quad (12)$$

为避免 macro_average_ F_1 值过小的情况,加入参数 θ 为调整参数.

3.3.2 微观平均值(Micro-average)

微平均准确率和微平均召回率的计算公式:

$$\text{Micro_}p = \frac{\sum_{i=1}^{|C|} a_i}{\sum_{i=1}^{|C|} a_i + \sum_{i=1}^{|C|} b_i} \quad (13)$$

$$\text{Micro_}r = \frac{\sum_{i=1}^{|C|} a_i}{\sum_{i=1}^{|C|} a_i + \sum_{i=1}^{|C|} c_i} \quad (14)$$

微平均值的计算公式如下:

$$\text{Micro_}F_1 = \frac{2 \cdot \text{Micro_}p \cdot \text{Micro_}r}{\text{Micro_}p + \text{Micro_}r} \quad (15)$$

其中 a 为分类器正确判别文本数, b 为不属于该类却判别为该类的文本数, c 为属于该类却判别为非该类的文本数.

宏平均是对类的平均,容易受到较小类别的影响;而微平均是对文本的平均,易受到大类别的影响.

4 基于改进 HMM 文本信息抽取模型

本文以改进的 HMM 模型为基础,对文本信息段进行抽取处理,如图 3 所示.

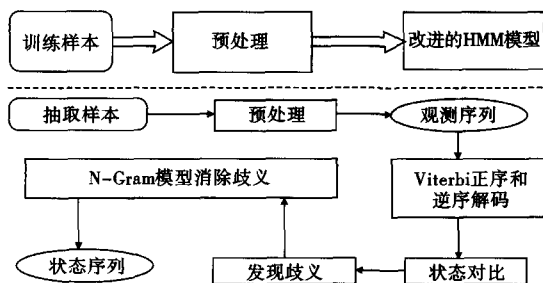


图 3 文本信息段抽取处理过程

其中,在改进后的模型中需通过 Viterbi 解码算法进行双序解码, Viterbi 算法描述如下:

变量:

$$\delta_i(i)(j) = \max_{q_1, q_2, \dots, q_{i-1}} P(q, q_i = S_i, q_{i+1} = S_j, O | K) \quad (16)$$

初始化:

$$\delta_1(i)(j) = \lambda_i b_{ij}(o_1) a_{ij} b_j(o_2), \quad 1 \leq i \leq N \quad (17)$$

递归:

$$\delta_{t+1}(i)(j) = \left[\max_{1 \leq k \leq N} \delta_t(k)(i) a_{ki} \right] b_{ij}(o_{t+1}) b_j(o_{t+2}) \quad (18)$$

终止:

$$P^* = \max_{1 \leq i, j \leq N} [\delta_T(i)(j)] \quad (19)$$

对于观测值序列 $O = (o_1, o_2, \dots, o_T)$, 用 Viterbi 算法解码时, $P(q_{i+1} | q_1, q_2, \dots, q_i) = P(q_{i+1} | q_i)$, 也就是 o_1 的状态确定 o_2 的状态. 对于逆序观测序列 $O' = (o_T, o_{T-1}, \dots, o_1)$, $t+1$ 时刻的状态 q_{t+1} 由时刻 t 的状态 q_t 来决定, 即 o_1 的状态由 o_2 的状态决定. 因此, 观测值序列 $O' = (o_T, o_{T-1}, \dots, o_1)$ 的解码过程可以描述为: 对于观测值序列 $O = (o_1, o_2, \dots, o_T)$, t 时刻的状态 q_t 由时刻 $t+1$ 的状态 q_{t+1} 决定, 即 $P(q_i | q_1, \dots, q_i, q_{i+1}) = P(q_i | q_{i+1})$, 也就是后向依赖性. 利用 Viterbi 算法选择对应的状态序列 $S = (q_1, q_2, \dots, q_i)$, 使 S 能够合理地解释观测序列 O , 即处理解码问题解开隐含部分, 在优化准则下找到最优状态序列.

在给定观测值序列中, 其双序解码后可得两状态序列 $S = (s_1, s_2, \dots, s_i, \dots, s_N)$ 和 $S' = (s'_1, s'_2, \dots, s'_j, \dots, s'_N)$, 两序列 s_i 和 s'_j 应有相同的观测值, 其中 $i+j-1=N$. 若状态对比后 s_i 和 s'_j 的结果不一致, 则提取 s_i 及其前的 $n(n=\min(i, j))$ 个状态和 s'_j 及其后的 n 个状态构建新状态序列 Qs 和 Qs' . 再采用 N-Gram 模型获取 $P(Qs)$ 与 $P(Qs')$, 将概率较大的状态序列作为最终解码序列.

5 结束语

本文旨在将改进的隐马尔可夫模型应用于文本分类中. 改进后的隐马尔可夫模型同时考虑了正序和逆序解码模型, 将双序进行综合, 增强了抗干扰性能, 使得解码更加精确. 并采用宏观平均值和微观平均值这种评价标准来对整个文本集分类进行性能分析, 该测评指标可准确评价不同训练集模型建立的质量, 指导逐步完善改进 HMM 模型. 在 HMM 分类器模型的建立中, 设定观测输出概率矩阵时, 引入 χ^2 统计和改进的 TFIDF 方法, 反映出特征词的语义关联性, 使得文本分类过程更稳定, 分类更精确. 而在这个过程中还存在着一些不足和需要修改的地方: (1) HMM 分类器模型参数的学习是基于统计特性, 而没有反映出文本本身词的语义信息; (2) 模型

的构建需要大量的训练数据.当数据数量不够大时,总有可能出现0观测值的现象,因此需要采用一些平滑算法;(3)本文的研究仅是从理论入手,尚缺少完善的实验数据.

参考文献:

- [1] 蒋宗礼,徐学可.一种基于集成学习与类指示器的文本分类方法[J].北京工业大学学报,2010,36(4):546-553.
- [2] Shao Fu bo, He Guo ping, Zhang Xin. An improved algorithm for multiclass text categorization with support vector machine[C]// International Symposium on Computational Intelligence and Design. 2008:336-339.
- [3] Rabiner L E A. Tutorial on hidden markov models and selected application in speech recognition[J]. Proceedings of The IEEE, 1989,79(2):257-286.
- [4] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization [C]// Proceedings of the Eighteenth Conference on Artificial Intelligence. [s. l.]:[s. n.],2000:584-589.
- [5] Freitag D, McCallum A. Information extraction with

HMMs and shrinkages [C] // Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction,[s. l.]:[s. n.],1999:31-36.

- [6] Frasconi P, Soda G, Vullo A. Hidden Markov models for text categorization in multi-page documents [J]. Journal of Intelligent Information Systems, 2002, 18(2):195-217.
- [7] 董云耀,钱如栏.一种改进的基于隐马尔可夫的信息检索模型[J].杭州电子科技大学学报,2009,29(4):46-49.
- [8] 杨健,汪海航.基于隐马尔可夫模型的文本分类算法[J].计算机应用,2010,30(9):2348-2350,2361.
- [9] 梁吉光,田俊华,姜杰.基于改进 HMM 的文本信息抽取模型[J].计算机工程,2011,37(20):178-179,182.
- [10] 张玉芳,彭时明,吕佳.基于文本分类 TFIDF 方法的改进与应用[J].计算机工程,2006,32(19):76-78.
- [11] 龚笔哄,彭波.对文本分类评测方法稳定性的研究[J].模式识别与人工智能,2008,21(1):12-17.

作者简介:

李开荣 男,(1963—),博士研究生,副教授.研究方向为智能信息处理、知识管理与决策支持.

(上接第 160 页)

本文仅仅探讨了具有自适应泄漏弹性的伪熵函数,下一步中我们将考虑基于伪熵函数具体的自适应泄漏弹性公钥加密方案.

参考文献:

- [1] Halderman J A, Schoen S D, Heninger N, et al. Lest we remember: Cold boot attacks on encryption keys [C]// In USENIX Security Symposium, SAN JOSE CA,2008:45-60.
- [2] Kocher P C. Timing attacks on implementations of diffie-hellman, ras, dss, and other systems[C]// CRYPTO, Springer-Verlay,1996:104-113.
- [3] Kocher P C, Jaffe J, Jun B. Differential power analysis[C]// In CRYPTO, 1999:388-397.
- [4] D A Osvik, A. Shamir, E. Tromer. Cache attacks

and countermeasures: The case of aes[C] // In CT-RSA, San Jose CA USA;IEEE,2006:1-20.

- [5] M. Braverman, A. Hassidim, Y. T. Kalai. Leaky Pseudo-Entropy Functions [C] // Computer Science Beijing, China:[s. n.] 2011.
- [6] C. Peikert and B. Waters. Lossy trapdoor functions and their applications [C] // In STOC. Victoria, British Columbia, Canada;IEEE, 2008:187-196.
- [7] J. Damgard and M. Jurik. A generalization, a simplification and some applications of Paillier's probabilistic public-key system[C] // In PKC, 2001, Che Ju Island, Korea;IEEE, 1992:119-136.

作者简介:

陈浩 男,(1987—),硕士研究生.研究方向为网络与信息安全.