基于条件随机场的文本分类模型

张春元
(海南大学 信息科学技术学院,海南 海口 570228)

摘 要:条件随机场(CRFs)是一种十分优秀的统计学习模型,文中尝试将其引入到文本分类之中,提出了一种基于CRFs的文本分类模型。首先通过特征选择将待分类文档和文档类别分别表示成为CRFs的观察序列和状态序列,然后使用文本分类相关领域知识定义特征函数来提取序列之间的关联特征,再采用前向或后向算法评估出给定观察序列条件下各状态序列的概率,据此实现待分类文档的分类。分析表明,这种新模型语义清晰,计算直观,易于融合各种文本分类领域知识,分类效率较高。

关键词:文本分类;条件随机场;序列模型
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2011)07-0077-04

Text Categorization Model Based on Conditional Random Fields

ZHANG Chun-yuan
(Institute of Information Science and Technology, Hainan University, Haikou 570228, China)

Abstract: Conditional random fields (CRFs) is an excellent statistical learning model. Importing it into text categorization, it proposes a text categorization model based on CRFs. Firstly, by choosing feature words, it describes a document as an observation sequence and each category as a state sequence. Then, using domain knowledges related to text categorization, it defines feature functions to extract association features between the sequences. Finally, it uses the forward or backward algorithm to find the probability of each state sequence in a given observation sequence, and uses these probabilities to categorize the document. The analysis shows that this new model has a good semantic interpretation, a strong ability to merge domain knowledge of text categorization, and a high efficiency to categorize documents.

Key words: text categorization; conditional random fields; sequence model

0 引 言

面对日益膨胀的电子文本信息,文本分类作为有效组织和管理它们的重要技术手段,无疑具有重大的研究价值和广阔的应用前景。文本分类理论和技术在发展过程中,十分注重从信息检索、人工智能、统计学、信息论和自然语言理解等相关学科领域吸收营养,例如朴素贝叶斯(Naive Bayes,NB)、支持向量机(Support Vector Machine,SVM)、最大熵模型(Maximum Entropy,ME)、隐马尔可夫模型(Hidden Markov Model,HMM)等分类算法主要源于统计学理论,文档的向量表示方法则借鉴于信息检索中的向量空间模型。

条件随机场(Conditional Random Fields,CRFs)是Lafferty等人[1]于2001年在ME和HMM基础之上提出来的一种十分优秀的统计学习模型,主要用于序列数据的标注和切分,例如自动分词、词性标注、句法分析等,但尚未发现有人将其应用到文本分类之中。我们注意到,作为CRFs产生基础的ME和HMM均被成功引入到文本分类之中,且都取得了不错的分类效果[2-6],为此,文中尝试将CRFs应用到文本分类之中,提出了一种基于CRFs的文本分类模型。该模型将待分类文档和文档类别分别表示成CRFs模型的观察序列和状态序列,然后使用文本分类相关领域知识构造CRFs的特征函数来提取它们的关联特征,再据此评估给定待分类文档条件下各文档类别的概率来完成文档的分类。该模型语义清晰,计算直观,易于融合各种文本分类领域知识,分类效率较高。

1 CRFs模型

CRFs是一种判别式概率无向图学习模型,克服了HMM的独立性假设问题和最大熵马尔可夫模型(Maximum Entropy Markov Model,MEMM)的标注偏置问题,是目前处理序列数据分割与标注问题最好的统计机器学习模型[7]。

收稿日期:2010-12-14;**修回日期**:2011-04-02
基金项目:国家自然科学基金资助项目(60863001)
作者简介:张春元(1973-),男,湖北武汉人,硕士,讲师,研究方向为Web信息检索、Web数据挖掘。

线链条件随机场(Linear Chain Conditional Random Fields,LC-CRFs)是 CRFs 最简单、最常用的一种形式,其结构如图 1 所示,$X=\{x_1,x_2,\cdots,x_N\}$ 为输入的观察序列,$Y=\{y_1,y_2,\cdots,y_N\}$ 为 X 对应的状态序列,y_i 为 x_i 对应的状态值。LC-CRFs 假设各状态结点 $y_1,y_2,\cdots y_N$ 之间存在一阶马尔可夫独立性,通过无向边连接成线性链。在给定观察序列 X 条件下,对于参数为 $\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$ 的 LC-CRFs,状态序列 Y 出现的条件概率定义为:

$$P_\Lambda(Y\mid X)=\frac{1}{Z_\Lambda(X)}\exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1},y_i,X,i)\right) \tag{1}$$

其中:$f_j(y_{i-1},y_i,X,i)$ 为特征函数,由用户自行定义;λ_k 是 $f_k(y_{i-1},y_i,X,i)$ 的权值,也称模型参数,经训练学习求取;$Z_\Lambda(X)=\sum_Y\exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1},y_i,X,i)\right)$ 为归一化因子。

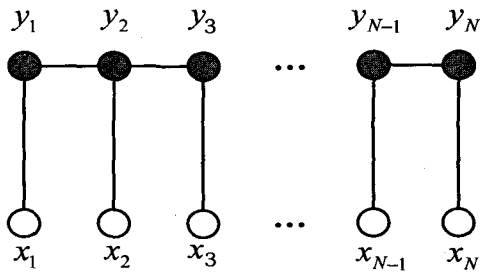

图 1 线链条件随机场模型结构图

从公式(1)不难看出,建立 LC-CRFs 模型关键要解决四方面问题:观察序列的表示;状态序列的表示;特征函数的定义;模型参数的估计。其中特征函数的定义尤为关键,其定义得合适与否将直接影响到整个 CRFs 模型的使用效果。借助特征函数的定义,用户不但可以将观察值与状态值之间的对应关系、状态值之间的转移关系整合到 CRFs 中来,而且还可以将相关领域知识引入到 CRFs 中来。CRFs 对特征函数的定义具有非常强的灵活性和包容性,允许用户从不同角度定义多个特征函数。

从理论上讲,CRFs 的研究及应用主要归结为三个基本问题的求解[8]:

① 学习问题:给定训练集 $\{X^{(s)},Y^{(s)}\}_{s=1}^{S}$,求使 $\{Y^{(s)}\}_{s=1}^{S}$ 出现的可能性为最大的模型参数集合 Λ;

② 解码问题:给定 Λ、X,求 X 最可能对应的状态序列 $Y^*=\arg\max_Y P_\Lambda(Y\mid X)$;

③ 评估问题:给定 Λ、X、Y,求 $P_\Lambda(Y\mid X)$。

其中:学习问题是 CRFs 建模过程中需要解决的关键问题,Λ 一般采用 CG、GIS 或 L-BFGS 等算法对训练集迭代学习求解;解码问题通常用来实现有序数据序列的标注和切分,是 CRFs 研究与应用的热点,Y^* 一般采用 Viterbi 算法求解;评估问题则很少有人关注,$P_\Lambda(Y\mid X)$ 一般采用前向算法或后向算法求解。

2 基于 CRFs 的文本分类模型

2.1 CRFs 文本分类模型设计

CRFs 的评估问题可以看作是对给定观察序列与状态序列对应程度的一种评估。如果将待分类文档、文档类别分别表示成 CRFs 的观察序列 X 及状态序列 Y,二者之间的关联特征通过 CRFs 的特征函数提取,那么 $P_\Lambda(Y\mid X)$ 可以看作是待分类文档与文档类别之间的关联度的一种度量,文本分类问题就可以归结为 CRFs 的评估问题进行处理。

基于这一思想,提出了一种基于 LC-CRFs 的文本分类模型,具体设计过程如下:

1)特征粒度选取。

同 NB、KNN、SVM 等分类模型一样,选用何种粒度的特征项来表示文本文档和文档类别也是建立 CRFs 分类模型首先需要明确的问题。常用的特征单位有词、N-Gram、词组和概念,从现有研究成果来看,选用词为单位进行分类效果较好[5,9],因此文中选取词为特征单位。

2)文本文档表示。

在 CRFs 分类模型中,通过特征选择方法将待分类文本文档表示成为观察序列进行处理。在众多特征选择方法中,信息增益(Information Gain,IG)是一种非常有效的维数约简方法[9],其值反映了特征项的类别区分能力,具体计算公式为:

$$IG(t_i)=-\sum_{j=1}^{m}P(c_j)\log P(c_j)+P(t_i)\sum_{j=1}^{m}P(c_j\mid t_i)\log P(c_j\mid t_i)+P(\bar{t_i})\sum_{j=1}^{m}P(c_j\mid \bar{t_i})\log P(c_j\mid \bar{t_i}) \tag{2}$$

其中:$P(c_j)$ 为训练集中 c_j 类文档出现的概率,$P(t_i)$、$P(\bar{t_i})$ 分别为训练集中含有、不含有特征项 t_i 的文档的概率,$P(c_j\mid t_i)$、$P(c_j\mid \bar{t_i})$ 分别为训练集中含有、不含有特征项 t_i 的文档属于 c_j 类的概率,m 为文档类别数。

从待分类文档 d 的原始特征项中选取 IG 值最大的若干个特征项,按 IG 值升序排列即为 d 在 CRFs 中的观察序列表示,记为 $X^{(d)}=\{t_1^{(d)},t_2^{(d)},\cdots,t_N^{(d)}\}$。由公式(2)可知:$d$ 中各特征项的 IG 值只与训练集有关,而与 d 本身无关;如果 d 中某特征项在整个训练集中不存在,则其 IG 值为 0。因此,可事先从训练集的原始特征项中按 IG 值大小选取若干个特征项降序排列构建一个基本特征集 T,这样 d 进行特征选择时只需查询其原始特征项在 T 中出现的情况即可,可有效提高特征选择的效率。

3)文档类别表示。

文档类别表示是从训练集的原始特征中为每一文档类别选择一组最能反映该文档类别统计特性的特征项来作为其状态序列,选用 χ^2 统计量来完成这一工作。设 A、B 分别为 c_j 类、非 c_j 类训练集中含有特征项 t_i 的文档数量,C、D 分别为 c_j 类、非 c_j 类训练集中不含有特征项 t_i 的文档数量,则 t_i 和文档类别 c_j 的关联度可表示为:

$$\chi^2(t_i,c_j)=\frac{(A+B+C+D)\times(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{3}$$

通过公式(3)从训练集的原始特征中选取与文档类别 c_j 最相关的若干个特征,按 χ^2 值升序排列即为 c_j 在 CRFs 模型中的状态序列表示,记为 $Y^{(c_j)}=\{t_1^{(c_j)},t_2^{(c_j)},\cdots,t_n^{(c_j)}\}$。考虑到训练集的原始特征项数量众多,为了加快特征选择的速度,只从基本特征集 T 中选择文档类别的有效特征项。

4)特征函数定义。

经上述表示后,CRFs 分类模型中的观察结点和状态结点之间的对应实际上是待分类文档和文档类别的有效特征词之间的对应,因而很自然地想到可以定义一个特征函数来提取"词-词"特征,通过词间关联度来评估文档和文档类别的相关度。对任一文档类别 c_j,其词-词特征函数定义为:

$$f^{(c_j)}(y_i,x_i)=f^{(c_j)}(t_i^{(c_j)},t_i^{(d)})=\mathrm{sim}^{(c_j)}(t_i^{(c_j)},t_i^{(d)}) \tag{4}$$

其中:$t_i^{(c_j)}$、$t_i^{(d)}$ 分别为 y_i 和 x_i 对应的特征词;$\mathrm{sim}^{(c_j)}(t_i^{(c_j)},t_i^{(d)})$ 为 c_j 类文档中 $t_i^{(c_j)}$ 与 $t_i^{(d)}$ 的关联度,可以通过 Similarity Thesaurus[10]、$\chi_c^2(w_k,w_i)$ 统计量[2]、互信息等方法测算,此处通过构建词-词关联矩阵[11]来度量。设基本特征集 T 中特征词 $t_g(t_g\in T,g=1,2,\cdots,|T|)$ 在 c_j 类训练集 D_{c_j} 的文档 $d_h(d_h\in D_{c_j},h=1,2,\cdots,|D_{c_j}|)$ 中出现的次数为 $tf_{t_g,h}$,则 c_j 类的词-词关联矩阵定义为:$R_{c_j}=[r_{u,v}]_{|T|\times|T|}$,其中 $r_{u,v}$ 为特征词 t_u 与 t_v 的关联度,计算公式为:

$$r_{u,v}=\left(\sum_{h=1}^{|D_{c_j}|}(tf_{t_u,h}\times tf_{t_v,h})\right)\Big/\left(\sum_{h=1}^{|D_{c_j}|}tf_{t_u,h}^2+\sum_{h=1}^{|D_{c_j}|}tf_{t_v,h}^2-\sum_{h=1}^{|D_{c_j}|}tf_{t_u,h}\times tf_{t_v,h}\right) \tag{5}$$

文档和各文档类别的相关度还可以通过文档拥有类别有效特征词的数量多少来进行评估,因此可基于词频权重同时结合 χ^2 统计量定义一个特征函数来提取"类别词"特征。对任一文档类别 c_j,其类别词特征函数定义为:

$$f^{(c_j)}(y_i,X)=f^{(c_j)}(t_i^{(c_j)},X)=tf_{t_i^{(c_j)}}\Big/\max_k\{tf_{t_k^{(c_j)}}\}\times\chi^2(t_i^{(c_j)},c_j) \tag{6}$$

其中:$tf_{t_i^{(c_j)}}$ 为类别 c_j 的有效特征词 $t_i^{(c_j)}$ 在 X 代表的待分类文档 d 中出现的次数,$\max_k\{tf_{t_k^{(c_j)}}\}$ 为类别 c_j 的有效特征词在 d 中出现的最高次数。需要说明的是,此时 X 应该采用 c_j 的类别特征项集合 $\{t_1^{(c_j)},t_2^{(c_j)},\cdots,t_n^{(c_j)}\}$ 或者 d 的原始特征项来表示,采用 d 的有效特征项反而不便于"类别词"特征的提取。

5)模型参数估计。

由公式(4)、(6)可知,CRFs 分类模型的特征函数定义均是面向具体文档类别的,因而模型参数也是类别相关的,每个文档类别都有相应的模型参数集合,各类模型参数集合估计时需选用相应类别文档训练集进行学习。例如对文档类别 c_j 的特征函数,需选用该类训练集 D_{c_j} 来求解 Λ_{c_j}。各模型参数集合的迭代求解拟采用 L-BFGS 算法,因其比 CG 算法和 GIS 算法具有更快的收敛速度[12]。

2.2 CRFs 文本分类模型应用

设文档类别集合为 $C=\{c_1,c_2,\cdots,c_m\}$,按 2.1 节所述经训练学习建立的 CRFs 文本分类模型的基本特征集为 T,任一文档类别 c_j 的状态序列表示为 $Y^{(c_j)}$,词-词关联矩阵为 R_{c_j},模型参数为 Λ_{c_j},则待分类文档 d 应用 CRFs 文本分类模型分类步骤如下:

1)以词为特征单位对 d 进行预处理,从 T 中按顺序选取 $|Y^{(c_j)}|$ 个出现在 d 中的特征项升序排列即为 d 的观察序列表示,记为 $X^{(d)}$。

2)利用前向或后向算法,依次计算在给定 d 条件下各文档类别的概率 $P_{\Lambda_{c_j}}(Y^{(c_j)}\mid X^{(d)})(j=1,2,\cdots,m)$。在计算过程中,"词-词"特征值直接查找 R_{c_j} 获取;采用类别词特征函数在进行特征提取时,其 $X^{(d)}=Y^{(c_j)}$。

3)如果分类为非兼类分类问题,则 d 的类别号为 $c=\mathrm{argmax}_{c_j}P_{\Lambda_{c_j}}(Y^{(c_j)}\mid X^{(d)})$;如果分类为兼类分类问题,设类别阈值为 ε,则 d 属于所有 $P_{\Lambda_{c_j}}(Y^{(c_j)}\mid X^{(d)})>\varepsilon$ 的类别。

3 模型分析

在 CRFs 文本分类模型中,观察序列和状态序列分别采用文档和类别的有效特征词集来表示,一个观察序列代表着一篇文档,一个状态序列代表着一个文档类别,$P_\Lambda(Y\mid X)$ 反映了文档和类别之间的关联程度,整个模型具有良好的语义解释。模型对文档和类别关联度的计算过程比较直观,首先通过特征函数的定义确定要提取哪些类型的关联特征,然后对各类训练集学习提取出各类别下的关联特征并通过模型参数估计算出这些关联特征的权值,在此基础上,根据待分类文档中各类关联特征出现的情况即可评估该文档与各类别的关联度。词-词特征函数、类别词特征函数

分别通过文档特征词和类别特征词的词间关联度、文档拥有类别特征词的数量来评估文档和文档类别的相关度，两个定义都比较容易理解。

CRFs文本分类模型是CRFs的一个应用实例，自然拥有CRFs诸多特点。CRFs具有强大的特征融合能力，允许用户从不同层面定义多个特征函数以完成特定的任务。文中定义了两个特征函数来分析文档和类别的相关度，这两个特征函数可以单独使用，也可组合使用，在此基础上，还可以从其它角度继续定义新的特征函数来支持文档的分类，CRFs为各种文本分类领域知识的整合应用提供了一个良好的框架。与不少分类方法假设文档特征相互独立相比，CRFs的另一个优点便是允许状态值之间存在一阶或高阶马尔可夫性，这也更合乎文档特征的实际。

CRFs的不足之处是模型参数的估计需要花费较长时间来进行学习训练，但对CRFs文本分类模型来说，模型参数的学习训练只需经历一次，其所花费的时间即使较长但不用重复，因而分类所需时间主要集中在待分类文档的预处理、特征选择和分类判定等环节上。在模型设计过程中，对模型分类效率的提高给予了充分考虑：在特征选择上，通过预先构建基本特征集来加快特征选择效率，待分类文档无需计算即可表示成观察序列；在模型参数估计上，选用的是L-BFGS算法；条件概率评估时，“词-词”特征值也无需计算，只查询学习训练阶段就建立起来的词-词关联矩阵即可。

综上所述，CRFs文本分类模型具有以下特点：

1）语义清晰，计算直观，易于理解；

2）特征融合能力强，易于融合各种文本分类领域知识；

3）分类效率较高。

4 结束语

文中将文本分类问题转换成为CRFs评估问题进行处理，提出了基于CRFs的文本分类模型，给出了具体建模方法和分类应用步骤。分析表明，CRFs文本分类模型具有诸多优点，是一种非常有前景的文本分类模型。

下阶段的工作主要包括：通过实验来检验CRFs文本分类模型的分类效率和效果，与其它经典文本分类模型进行比较；对现有文本分类知识进行吸收整合，定义更为有效的特征函数；尝试定义转移特征函数来分析类别特征词之间的关联关系，以改进分类效果；研究基于本体表示的CRFs文本分类模型。

参考文献：

[1] Lafferty J, Mccallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of 18th International Conference on Machine Learning, 2001. San Fransisco: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[2] 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.

[3] 罗双虎，欧阳为民. 基于隐Markov模型的文本分类[J]. 计算机工程与应用，2007，43(30)：179-181.

[4] Yi K, Beheshti J. A hidden Markov model-based text classification of medical documents[J]. Journal of Information Science, 2009, 35(1): 67-81.

[5] 李荣陆，王建会，陈晓云，等. 使用最大熵模型进行中文文本分类[J]. 计 算机研究与发展，2005，42(1)：94-101.

[6] Nigam K, Lafferty J, Mccallum A. Using maximum entropy for text classification[C]// Proceedings of the IJCAI-99 Workshop on Information Filtering, 1999. San Fransisco: Morgan Kaufmann Publishers Inc, 1999: 61-67.

[7] 黄健斌，姬红兵，孙鹤立. 基于混合跳链随机场的异构Web记录集成方法 [J]. 软件学报，2008，19(8)：2149-2158.

[8] Feldman R, Sanger J. 文本挖掘(英文版)[M]. 北京：人民邮电出版社，2009.

[9] 陆 旭. 文本挖掘中若干关键问题研究[M]. 北京：中国科学技术大学出版社，2008.

[10] Qiu Y, Frei H. Improving the Retrieval Effectiveness by a Similarity Thesaurus[R]. Zurich: ETH Zurich, Department of Computer Science, 1994.

[11] Baeza-yates R, Ribeiro-neto B. 现代信息检索[M]. 王知津，贾福新，郑红军，等译. 北京：机械工业出版社，2005.

[12] Sha F, Pereira F. Shallow parsing with conditional random fields[C]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003. New Jersey: Association for Computational Linguistics, 2003: 131-141.

（上接第76页）

2004.

[10] Web Service Architecture, W3C Working Draft[EB/OL]. 2003-08-08. http://www.w3.org/TR/2003/Wd-ws-arch-20030808.

[11] OpenH323 project[EB/OL]. 2007. http://www.h323.org/.

[12] O'OHERTYP. JAIN SIP Tutorial[EB/OL]. 2004. http://java.sun.com/produncts/jain.