使用 Logistic 回归模型进行中文文本分类

李新福[1],赵蕾蕾[1],何海斌[1],李 芳[2]
LI Xin-fu[1],ZHAO Lei-lei[1],HE Hai-bin[1],LI Fang[2]

1.河北大学 数学与计算机学院,河北 保定 071002
2.河北大学 人文学院,河北 保定 071002
1.College of Mathematics and Computer,Hebei University,Baoding,Hebei 071002,China
2.College of Humanities,Hebei University,Baoding,Hebei 071002,China
E-mail:mc_lxf@126.com

LI Xin-fu,ZHAO Lei-lei,HE Hai-bin,et al.Using Logistic regression model for Chinese text categorization.Computer Engineering and Applications,2009,45(14):152-154.

Abstract: In this paper,Logistic regression model is used for Chinese text categorization.The categorization performance of this method is analyzed using different approaches for text feautre generation,different dimension of features and different documents set.Moreover,its classification performance is compared to linear SVM classifier in experiments.The experiments results show that its perfromance is comparable with linear SVM classifier.It's a promising method for text categorization.
Key words: Logistic regression model;support vector machines;text categorization;features

摘 要:使用 Logistic 回归模型进行中文文本分类,通过实验,比较和分析了不同的中文文本特征、不同的特征数目、不同文档集合的情况下,基于 Logistic 回归模型的分类器的性能。并将其与线性 SVM 文本分类器进行了比较,结果显示它的分类性能与线性SVM方法相当,表明这种方法应用于文本分类的有效性。
关键词:Logistic 回归模型;支持向量机;文本分类;特征
DOI:10.3778/j.issn.1002-8331.2009.14.046 **文章编号**:1002-8331(2009)14-0152-03 **文献标识码**:A **中图分类号**:TP391

1 引言

随着互联网的快速发展,为了能够有效地组织和处理网络上的文本信息,需要按照其内容自动分类。自动文本分类是将自然语言文本自动指定至一个或几个预定义的文本类别中的方法,它是文本信息处理的一项关键技术。文本表示和分类器设计是自动文本分类的关键。现有的分类方法主要是基于向量空间模型和统计理论、机器学习方法,主要有 Bayes[1]、KNN[2]、Boosting[3]、SVM[4]等。文献[5]中使用英文分类语料,对常用的分类方法进行了比较研究,KNN、SVM较其他分类方法有较高的分类准确性和稳定性。KNN 分类器在训练文档增加时,其分类时间将快速增加。SVM 本质上是一种二分类, 如果使用 SVM 分类器实现多类分类,必须构造多个分类器。当类别数较多时,分类时间较长,并且 SVM 分类器的训练时间也比较长。而且,基于向量空间模型的文本表示方式,文档向量数据具有高维性和稀疏性。

Logistic 回归属于广义线性模型中的一种,目标是估计样本的后验概率[6]。Logistic 回归得到样本对每个类别的隶属度,或者说样本的后验概率。本文使用 Logistic 回归模型进行文本分类,它的训练时间和分类时间相对较短。在特征选择方面,充分利用词性信息、语义信息[7]。通过实验发现,基于 logistic 模型的文本分类的分类准确率与 SVM 相当。表明了这种方法应用于文本分类的有效性。

2 Logistic 回归模型

Logistic 回归模型是用概率估计来进行分类的。假如需要研究的某一事件发生与否,引入一个潜变量 y^*,它表示这一事件发生的可能性,它的值域为全体实数,其值越大表示这一事件发生的可能性越大,而把 0 作为一个分界点,大于 0 表示发生, 否则为不发生。Logistic 回归模型广泛应用在辅助医疗诊断、灾害气象预测、经济预测等系统中。

对于文本分类问题,一个文档分到某个类别中假如看做是一个事件,文档的特征即文档中出现的词语、概念看成影响这个分类事件发生与否的因素。给定一个文档集合,利用回归分析研究文档特征和文档类别之间的内在关联,从而预测文本文

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60772073);河北省自然科学基金(the Natural Science Foundation of Hebei Province of China under Grant No.F2006001020);河北省教育厅科研基金(the Foundation of Education Bureau of Hebei Province,under Grant No.2005347);河北大学科研基金(the Fundation of Hebei University,under Grant No.Y2004045)。
作者简介:李新福(1970-),男,博士,副教授,主要研究方向为智能图文处理;赵蕾蕾(1982-),女,硕士生,主要研究方向为中文信息处理;何海斌(1978-),男,助教,主要研究方向中文信息处理;李芳(1980-),女,硕士生,主要研究方向为中文信息处理。
收稿日期:2008-03-18 **修回日期**:2008-05-19

档所属类别。考虑二分类情况,设自变量为 $X_i=\{x_1,x_2,\cdots,x_k\}$表示第 i 个文档的特征,y_i^* 表示文档 DOC_i 属于文档类别的可能性,$Y_i=1$ 表示属于,$Y_i=0$ 表示不属于该类别。响应变量 y^* 和自变量 X_i 之间的关系描述如下:

$$y_i^*=w_0+\sum_{k=1}^{N}w_k x_{ik}+\varepsilon_i \tag{1}$$

其中:ε_i 表示随机因素的影响,如果 ε_i 有 Logistic 分布,那么文档属于该类别的概率 P_i 为:

$$P(Y_i=1|X_i)=P\{\varepsilon_i\leqslant(w_0+\sum_{k=1}^{N}w_k x_{ik})\}=\frac{1}{1+e^{-\varepsilon_i}}=\frac{1}{1+e^{-(w_0+\sum_{k=0}^{N}w_k x_{ik})}} \tag{2}$$

则事件发生比可以表示为:

$$odds=\frac{p_i}{1-p_i}=e^{w_0+\sum_{k=1}^{N}w_k x_{ik}} \tag{3}$$

对事件发生比取自然对数,则 Logistic 模型表示为:

$$\log ity=\ln(odds)=w_0+\sum_{k=1}^{N}w_k x_{ik}=\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}+b \tag{4}$$

Logity 与文档特征之间成线性关系。如果令 $\boldsymbol{w}=[\boldsymbol{w}^{*\mathrm{T}},b]^{\mathrm{T}}$,并将文档向量作扩维处理 $\boldsymbol{x}^*=[\boldsymbol{x}^{\mathrm{T}},1]^{\mathrm{T}}$,文档类别用 $y_i\in\{1,-1\}$表示,可以将 logity 表示为:$\log ity=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^*$,在 Logistic 回归中,通过最大化下面的条件似然函数估计参数 w,通过取负号,也就是最小化下面的泛函:

$$\min l(w)=\sum_{i=1}^{N}\ln(1+e^{-y_i\boldsymbol{w}^{\mathrm{T}}x_i}) \tag{5}$$

并加入正则化项:

$$\min\ l(w)=\sum_{i=1}^{N}\ln(1+e^{-y_i\boldsymbol{w}^{\mathrm{T}}x_i})+1/2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \tag{6}$$

求解回归参数是一个无约束的凸优化问题,一般采用牛顿-拉夫逊方法(Newton-Raphson)求解。由上式可以看出,求解此问题需要用到所有的训练样本,用 Newton-Raphson 方法优化时,每次迭代都需要计算维矩阵的逆矩阵,对于文本分类来说,文本特征的高维性,在训练样本集的规模较大时,计算量将会很大,针对此问题,提出了一些近似算法,例如 BFGS 变尺度优化算法[8]、共轭梯度混合法[9]等,在每次迭代时计算目标函数的梯度,空间复杂度为 O(N);每次迭代解决两个样本的凸优化问题的 SMO 算法应用到 Logistic 回归[10],迭代次数明显增加。本文采用基于信赖域型牛顿法(Trust Region NewtonMethod)[11]求解回归参数。

3　基于 Logistic 回归模型的文本分类方法

3.1　文本分类模型

文本分类通常指在给定的分类体系下,根据文本的内容将一篇文本分到一个或是几个定义好的文本类别里。文本分类由训练过程和测试过程组成。分类器训练过程中,首先将训练文本用特征描述,得到训练文档向量集合,用该集合训练分类器,优化分类器参数,建立文本分类模型;测试过程中,将测试文本用训练时的文本特征描述,得到测试文档向量,用训练的文本模型对测试文本类型进行预测,并对分类性能进行评价。

与英文等西方语种书写方式不同,中文的词与词之间没有明显的切分标记,所以中文文本分类时,通常先进行分词处理。为了对 Logistic 回归在中文文本分类中的性能作出评价,分别使用关键词和语义概念作为文本的特征进行测试。关键词采用中科院计算所最新的汉语词法分析系统(ICTCLAS3.0)的分析结果,语义概念特征生成方法如下所述。

3.2　语义概念特征生成方法

文本特征的提取是文本自动处理的重要基础,在实验语料中,训练集用词有十几万个,由于文章长短不同,而平均文档用词数仅 258 个,因此形成的文档关键词向量空间具有高维性和稀疏性。为了缩减文本向量空间的维数,提高后继处理的性能,在依据词频来提取特征的基础上,对语义特点进行分析并通过加权的方法来确定文本语义概念特征。

语义分析需要一定的语义知识资源作为基础。WordNet 和知网(HowNet)[12]分别是描述英、中文概念的语义知识库。知网是一个以概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系的常识知识库,它是一个网状的有机的知识系统。在知网中每一个词语的概念及其描述形成一个记录,每一个记录都主要包含多项内容。其中每个词语由 DEF 来描述其概念定义,DEF 的值由若干个义原及其与主干词之间的语义关系描述组成,义原是知网中最基本的、不易于再分割的意义的最小单位。

词语与词语之间存在着同义、近义、反义等关系,词语本身也存在着一词多义的现象,另外,词语出现在文中的不同位置也有不同的语义特点。为了用语义概念表示文本,首先要确定句子中的词语在这个句子中所表达的语义。也就是需要进行语义消歧。语义消歧是解决如何在给定上下文语境中确定多义词的义项的问题。目前语义消歧方法主要有 Bayes、最大熵、基于词典等[13]。本文采用基于文献[14]的消歧策略,对经过分词和词性标注后的句子进行语义消歧,并在每个词后面标注上相应的语义编号。将语义作为文本的特征,从而形成相应的文档向量用于文本分类。

4　实验及结果分析

实验数据集 1 采用了复旦大学的中文分类语料,该语料包括训练语料库和测试语料库,所有的语料共分成了 20 个不同的类别。通过分析得到了不同类别语料的分布信息如表 1 所示。数据集 2 用网页分类数据集来源于北大天网的"中文网页分类竞赛"的训练网页集,共包含 15 304 个简体中文网页,分为 11 个类别[15]。

表 1　数据集 1 文档分布

类别	训练/篇	测试/篇	类别	训练/篇	测试/篇
1 航空	640	642	11 经济	1 600	1 601
2 能源	32	33	12 法律	51	52
3 电子	27	28	13 医药	51	53
4 通讯	25	27	14 军事	74	76
5 计算机	1 357	1 357	15 政治	1 024	1 026
6 矿产	33	34	16 体育	1 253	1 254
7 交通	57	59	17 文学	33	34
8 艺术	740	742	18 教育	59	61
9 环境	1 217	1 218	19 哲学	44	45
10 农业	1 021	1 022	20 历史	466	468

本文采用了两种评价指标:准确率 P 和召回率 R。两个指标的定义如下:$P=a/(a+b)$,$R=a/(a+c)$,其中:a 分类正确的文档数;b 分入该类而实际不属于该类的文档数;c 属于该类的文档

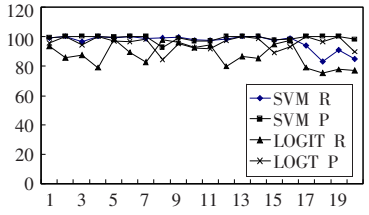

图1 选用语义特征时训练集1上的分类效果

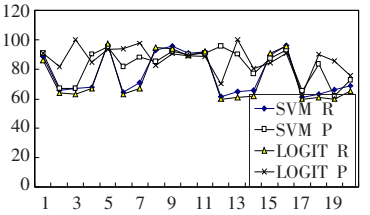

图2 选用语义特征测试数据集1上的分类效果

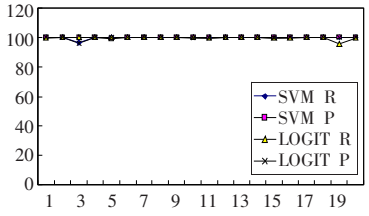

图3 选用关键词特征训练集1上的分类效果

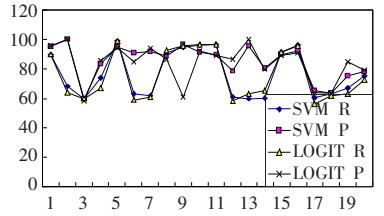

图4 选用关键词特征测试集1上的分类效果

数而判为不属于该类的文档数。分类平均正确率为:$P=(\sum P_i)/C$。SVM分类器采用线性核函数,惩罚因子取值为1。

数据集1上的实验结果见图1~图4,Logistic和SVM分类的准确率和召回率相差不多。语义特征维数1 400维而关键词特征高达159 684维,两种文本表示特征的分类结果中,语义概念特征相对稍微低1个百分点左右,可能是由于语义分析消歧效果不够理想,下一步工作中将加强语义消歧模块功能。

整体分类效果见表2、表3,在两个不同来源的数据集上的Logistic与SVM分类效果近似,说明了基于Logistic分类方法的适应性和有效性。数据集2上测试集的分类正确率相对于数据集1偏低,与网页数据的特性有关。

表2 数据集1分类平均正确率 (%)

语义特征				关键词特征			
训练集		测试集		训练集		测试集	
SVM	LOGIT	SVM	LOGIT	SVM	LOGIT	SVM	LOGIT
98.86	95.39	83.17	82.59	99.95	99.745	84.98	83.89

表3 数据集2分类平均正确率 (%)

语义特征				关键词特征			
训练集		测试集		训练集		测试集	
SVM	LOGIT	SVM	LOGIT	SVM	LOGIT	SVM	LOGIT
99.92	98.08	68.56	70.11	100	100	66.29	67.56

5 结束语

本文使用Logistic回归模型进行中文文本分类。通过实验,比较和分析了关键词、语义特征、不同文档集合的情况下,基于Logistic回归模型的分类器的性能。并将其与线性SVM文本分类器进行了比较,结果显示它的分类性能与线性SVM方法相当,表明了这种方法应用于文本分类的有效性。如果能进一步提高语义分析的准确率和考虑文本数据分布不均衡性,将得到更好的分类结果。

参考文献:

[1] Rennie J D M,Shih L,Teevan J,et al.Tackling the poor assumptions of Naïve Bayes text classifiers [C]//Proceedings of the Twenti-eth International Conference on Machine Learning,2003,2:616-623.

[2] Chiang J H,Chen Y C.Hierarchical fuzzy-KNN networks for news documents categorization[C]//10th IEEE International Conference on Fuzzy Systems,2001(2):720-723.

[3] Sebastiani F,Nazionale C,Valdambrini N.An improved boosting algorithm and its application to text categorization[C]//Proceedings of the Ninth International Conference on Information and Knowledge Management,2000:78-85.

[4] Zhang Hao,Berg A C,Maire M,et al.SVM-KNN:Discriminative nearest neighbor classification for visual category recognition[C]//IEEE Computer Society Conference on HHComputer Vision and Pattern Recognition,2006:2126-2136.

[5] Yang Y.An evaluaton of statistical approaches to text categorization[J].Information Retrieval,1999,1(1):76-78.

[6] 王济川,郭志刚.Logistic回归模型方法及应用[M].北京:高等教育出版社,2001.

[7] 邹娟,周经野,邓成.一种基于语义分析的中文特征值提取方法[J].计算机工程与应用,2005,41(36):164-166.

[8] 赵凤治.数值优化中的二次逼近法[M].北京:科学出版社,1994.

[9] Komarek P,Moore A.Fast robust logistic regression for large sparse datasets with binary outputs[C]//Proceedings of the Ninth International Workshop on Artifical Intelligence and Statistics,2003:197-204.

[10] Keerth S S,Duan K B,Shevade S K,et al.A fast dual algorithm for kernel logistic regression[J].Machine Learning,2005,61(1):151-165.

[11] Lin C J,Weng R C,Sathiya Keerthi S.Trust region Newton methods for large-scale logistic regression[C]//Proceedings of the 24th International Conference on Machine Learning,2007,3:561-568.

[12] 董振东.知网[EB/OL].http://www.keenage.com.

[13] 谈文蓉,符红光,刘莉,等.一种基于贝叶斯分类与机读词典的多义词排歧方法[J].计算机应用,2006,26(6):1389-1391.

[14] Chen Hao,He Ting-ting,Ji Dong-hong,et al.An unsupervised approach to Chinese word sense disambiguation based on Hownet[J].Computational Linguistic and Chinse Language Processing,2005,10(4):473-482.

[15] 李新福.组合降维技术在中文网页分类中的应用[J].计算机工程与应用,2007,43(24):169-171.