文章编号 :1006 - 9348(2006)03 - 0222 - 03

文本分类的特征提取方法比较与改进

申红[1],吕宝粮[1],内山将夫[2],井佐原均[2]

(1. 上海交通大学 计算机科学与工程系,上海 200030 ;2. 国立信息与通讯技术研究所计算语言实验室,日本 京都府 619 - 0289)

摘要 :文本的特征提取是文本分类过程中的一个重要环节,它的好坏将直接影响文本分类的准确率。该文介绍了词条的 $\chi 2$ 统计方法(CHI)、词条与类别的互信息(MI)、信息增益(IG)、词条的期望交叉熵(CE)等文本特征提取方法,并对其取词策略进行了改进。为了对这些特征提取方法进行系统地比较,选择了三种代表性的分类器对《读卖新闻》文本数据库进行了分类实验。实验结果表明 $\chi 2$ 统计方法具有最好的准确率,各种改进的特征提取方法都能提高文本分类的准确率。

关键词 :特征提取 ;文本分类 ;互信息 ;支持向量机

中图分类号 :TP391 文献标识码 :A

Comparison and Improvments of Feature Extraction Methods for Text Categorization

SHEN Hong[1], LU Bao - liang[1], Utiyama Masao[2], Isahara Hitoshi[2]

(1. Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200030, China ;
2. Computational Linguistics Group, National Institute of Information and Communications Technology, Kyoto 610 - 0289, Japan)

ABSTRACT : Feature extraction technology is an essential part of text categorization, which affects directly the precision of categorization. This paper introduces four popular feature extraction methods, i. e. a $\chi 2$ - test (CHI), mutual information (MI), information gain (IG), and cross entropy (CE), and proposes corresponding improvements on extracting character. In order to compare these methods comprehensively, we perform simulations on Yomiuri News Corpus using three typical classification algorithms. The experimental results show that the modified feature extraction methods can improve the precision of categorization. In addition, a $\chi 2$ - test method obtains the best classification precision.

KEYWORDS : Feature selection ; Text categorization ; Mutual information ; Support vector machine

1 引言

随着计算机技术和通讯技术的飞速发展,人们可以获得的文本信息越来越多,同时需要投入更多的时间对信息进行组织和整理。文本分类能够改善文本信息杂乱的状况,可以降低查询时间,提高搜索质量,方便用户,从而使快速有效地获取文本信息成为可能。在研究文本分类的过程中,特征提取是最关键的环节之一,具有降低向量空间维数、简化计算、防止过分拟合以及去除噪声等作用,特征提取的好坏将直接影响着文本分类的准确率。

最常用的文本特征表示模型是向量空间模型(Vector Space Model,VSM)。向量空间模型基于这样一个关键假设,即文章中词条出现的顺序是无关紧要的,它们对于文本的类别所起的作用是相互独立的,因此可以把文本看作一系列无序词条的集合。在该模型中,文本空间被视为一组正交词条向量所张成的向量空间[1]。向量的维数往往是惊人的,包含噪声,且特征不明显。特征提取可以看作是从测量空间到特征空间的一种映射或变换。特征提取可以降低特征空间的维数,从而达到降低计算复杂度和提高分类准确率的目的[2]。

2 特征提取方法

对 *VSM* 型的文本样本一般是构造一个特征评估函数,将测量空间的数据映射到特征空间,对特征空间中的特征值进行评估,然后选择合适的词作为样本的特征。特征评估函数通常有下列几种形式 :信息增益(*IG*)、词条与类别的互信息(*MI*)、词条的 χ^2 统计(*CHI*)、及词条的期望交叉熵(*CE*) 等。

1) 信息增益(*IG*) 被广泛地应用于机器学习。名词 W 的 *IG* 值可定义为

$$IG(W) = -\sum_{i=1}^{m} P(C_i)\log P(C_i)$$

基金项目 :国家自然科学基金资助(60375022)
收稿日期 :2004 - 11 - 06

$$+P(W)\sum_{i=1}^{m}P(C_i\mid W)\log P(C_i\mid W)$$

$$+P(W)\sum_{i=1}^{m}P(C_i\mid \overline{W})\log P(C_i\mid \overline{W}) \quad (1)$$

其中 C_i（其中 $i=1,\ldots,m$）表示第 i 类，$P(C_i)$ 是在训练样本集中样本是 C_i 类的概率，$P(W)$ 是名词 W 在训练样本集中出现的概率，$P(C_i\mid W)$ 是名词 W 出现的前提下样本是 C_i 类的概率，$P(C_i\mid \overline{W})$ 是名词 W 不出现的前提下样本是 C_i 类的概率。*IG* 值越高，对分类预测提供的信息就越多。通过设定阀值，可以将 *IG* 值小于阀值的名词删除掉，从而降低特征空间维数[3]。

2）词条与类别的互信息（*MI*）所用的评估函数为

$$MI(W,C_j)=\log(P(W\mid C_j)P(W)) \quad (2)$$

其中 C_j 是类别，$P(W)$ 是名词 W 在训练样本集中出现的概率，$P(W\mid C_j)$ 是样本是 C_j 类的前提下名词 W 出现的概率。

需要选取互信息量最大的名词作为特征词，这样的词在某个类中的出现概率大，而在其他类中出现的概率小。这是因为互信息量越大，名词和类别之间同时出现的概率也越大。它的另一种表示形式是：

$$MI_{avg}(W)=\sum_{i=1}^{m}P(C_i)MI(W,C_i) \quad (3)$$

如果训练样本数为 N，一篇文章最大的名词量为 k，总的类别数为 m，总的名词量为 V，那么，算法实现可分两步：*a*）将文本模型转换成类别和与之对应的名词向量的形式，时间复杂度为 $O(Nmk)$，空间复杂度为 $O(mV+Nk)$；*b*）在 *a*）的基础上求 *MIavg*，其时间复杂度为 $O(mV)$，空间复杂度为 $O(mV)$。这样，求 *MIavg* 的时间复杂度为 $O(Nmk+mV)$。

3）词条的 χ^2 统计（*CHI*）方法是指在词和类间没有独立性，并可类比为一个自由度的 χ^2 分布。其评估函数为

$$\chi^2(W,C)=\frac{n(n_{11}\times n_{12}\times n_{21}\times n_{22})^2}{(n_{11}+n_{12})\times(n_{21}+n_{22})\times(n_{11}+n_{21})\times(n_{12}+n_{22})} \quad (4)$$

$\overline{W}$ 表示除了名词 W 之外的其他名词，$\overline{C}$ 表示除了类别 C 之外的其他类别。那么名词 W 和类别 C 的关系就有 4 种：(W,C)，$(W,\overline{C})$，$(\overline{W},C)$，$(\overline{W},\overline{C})$。用 n_{11}，n_12，n_{21}，n_{22} 分别表示这 4 种情况的频度，总数 $n=n_{11}+n_{12}+n_{21}+n_{22}$。同时要求 $n_{11}\times n_{22}>n_{21}\times n_{12}$，否则名词和类别之间就是相斥的关系。

χ^2 - 统计量的值越高，名词和类别之间的独立性就越小。它的一种转换形式为

$$\chi^2_{avg}(W)=\sum_{i=1}^{m}P(C_i)\chi^2(W,C_i) \quad (5)$$

CHI 方法的时间复杂度为 $O(mNk+mV)$，空间复杂度为 $O(mV)$。

4）交叉熵，也称为 *KL* 距离：

$$CE(W)=-\sum_{i=1}^{m}P(C_i\mid W)\log\frac{P(C_i\mid W)}{P(C_i)} \quad (6)$$

交叉熵反映了文本类别的概率分布和在出现了某个特定名词的条件下文本类别的概率分布之间的距离，名词 W 的交叉熵越大，对文本类别分布的影响也越大[4]。其时间复杂度为 $O(mNk+mV)$，空间复杂度为 $O(mV)$。

上述几种特征评估函数各有优缺点。*IG* 计算量相对其它几种方法较大；对于 *MI* 方法，在相同的条件概率下，稀有名词会比一般词获得更高的得分；*CHI* 方法基于 $\chi 2$ 分布，如果这种分布被打破，则对低频词不可靠[5]。另外，公式（3）、（5）和（6）中都为取和运算，容易抹煞掉某类特有的特征词。考察公式（3）、（5）和（6），我们可以看出，当某词是多个类的特征词，那么它的评分会很高；但是，如果某个词是某个类的特有特征词，则它最后的评分会被中和掉。

针对这些情况，我们提出了下面的改进方案。当测试数据和训练数据都是相同的几类，但是这几类分布不完全一样时，可以这样取词，*a*）对每一类取 *MI* 或者 χ^2 值最大的 k 个名词，这样构成一个特征向量空间。用公式可分别表示如下：

$$\bigcup_{i=1}^{m}\{w\mid MI(w,C_i)\text{ 是 }\{MI(W,C_i)\}\text{ 中最大的前 }k\text{ 个。}\} \quad (7)$$

$$\bigcup_{i=1}^{m}\{w\mid \chi^2(w,C_i)\text{ 是 }\{\chi^2(W,C_i)\}\text{ 中最大的前 }k\text{ 个。}\} \quad (8)$$

b）对 *CE* 方法也作了相应的改进，先将公式（6）变形为

$$CE(W,C)=-P(C\mid W)\log\frac{P(C\mid W)}{P(C)} \quad (9)$$

其中 $P(C)$ 是样本为 C 类的概率，$P(C\mid W)$ 是在出现名词 W 的前提下样本是 C 类的概率。然后按如下公式取词

$$\bigcup_{i=1}^{m}\{w\mid CE(w,C_i)\text{ 是 }\{CE(W,C_i)\}\text{ 中最大的前 }k\text{ 个。}\} \quad (10)$$

改进方法的时间和空间复杂度和原来的方法是相同的。这样的取词策略可确保多类的特征词和单类特有的特征词都能兼顾，并能克服测试样本集和训练样本集分布不一所带来的负面影响。

3 实验设计及其结果分析

本实验采用《读卖新闻》数据库[6]。该数据库含有从 1987 年到 2001 年的读卖新闻，总共有 2190521 篇文章。我们使用 1996 年到 2001 年的文本作为研究对象，其中 913118 个文本作为训练样本，181863 个文本作为测试样本。在本实验中，我们使用其中最大的前 5 类，各个类的样本个数分布如表 1 所示。

表 1 样本数据表

样本类别	样本数据集	
	训练样本	测试样本
犯罪事件	103607	24374
体育新闻	79726	17610
亚太事务	41374	5943
南北美洲	36275	6109
健康新闻	35932	7004
总计	296914	61040

表2　实验结果测试准确率表

特征提取方法	M^3 – SVM	SVM	k – NN
CHI *	92.34	92.29	77.84
CHIavg	90.09	90.24	51.50
MI *	88.13	88.32	72.67
MIavg	84.75	85.16	80.53
CE *	76.28	77.79	66.17
CE	72.00	69.49	58.26

此实验所采用的特征提取方法有：MIavg(公式3)，CHIavg(公式5)，CE(公式6)，改进的MI(公式7)，改进的CHI(公式8)和改进的CE(公式10)。提取的特征向量维数为5000。

我们所采用的分类器有SVM，k – NN和M^3 – SVM[7]。其中SVM、k – NN均为文本分类中常用的分类器，一般情况下，SVM的分类准确率要高于k – NN。另外，M^3 – SVM是针对大规模数据而设计的一种超并列、模块式分类器，它采用‘分而治之’的思想，将大规模问题分解为容易求解的小问题[8]。

在本文的实验中，SVM和M^3 – SVM所使用的惩罚参数C取64，高斯核函数的σ取0.125；k – NN中的k取值范围为3到600。k – NN的实验结果代表在不同的k值下最高的准确率。CHI *、MI *和CE *表示本文所提出的改进方法。实验结果如表2和表3所示，表2为测试样本在不同的特征提取方法和不同的分类器下的分类测试准确率，表3为在不同分类器下的训练和测试时间，并行训练时间是指解决一个二类子问题所需要的最长时间，串行训练时间指训练所有分类器需要的时间，另外k – NN只有测试时间。

表3　实验结果训练及测试时间表

特征提取方法	M^3 – SVM			SVM			k – NN	
	训练时间(s)		测试时间(s)	训练时间(s)		测试时间(s)	训练时间(s)	测试时间(s)
	串行	并行		串行	并行			
CHI *	112177	13430	9647	124262	20700	6933	–	63487
CHIavg	183147	22633	13248	195282	32727	9140	–	69835
MI *	40868	5221	5833	44604	7813	4286	–	49730
MIavg	278464	32642	21439	292388	39915	17866	–	53231
CE *	197346	23763	6869	222464	33565	5249	–	42447
CE	530454	65350	23069	565386	95589	13524	–	47709

实验结果表明，我们提出的取词策略对三种传统的特征提取方法都有效。表2的数据显示了改进方法的准确率比传统的方法都好，其中CHI *方法在M^3 – SVM和SVM下的准确率最高，MIavg方法在三个分类器中的准确率的差距最小。另外，就分类器而言，k – NN的实验效果最差，SVM与M^3 – SVM的实验结果相近，这说明在大规模文本分类中M^3 – SVM与SVM具有几乎相同的准确率，但是，M^3 – SVM具有训练时间短、容易并行实现的优点[7]。表3的数据显示了改进方法在同一个分类器下训练和测试所用的时间也比传统方法要少，其中MI *方法在M^3 – SVM和SVM下的训练和测试用的时间是最少的，比传统方法有了较大的提高，另外由于k – NN只有测试过程，所以测试用的时间比同样数据在其他分类器下用的测试时间要高很多。

4　结束语

本文比较了四种在文本分类中常用的特征提取方法，并提出了取词策略的改进，实验结果说明，改进的方法提高了分类准确率。对于不同的特征提取方法，我们还比较了k – NN，SVM和M^3 – SVM分类器的优劣。实验结果表明，SVM与M^3 – SVM优于k – NN，且具有几乎相同的准确率。

作为进一步的工作，我们将对全《读卖新闻》数据库进行仿真实验，并寻求适合大规模文本分类的特征提取方法和与之相应的分类算法。

致谢

我们向对本文的工作给予帮助的上海交通大学仿脑计算实验室的博士生文益民，范志刚和连惠诚表示感谢。

参考文献：

[1] 王灏，黄厚宽，田盛丰. 文本分类实现技术[J]. 广西师范大学学报(自然科学版)，2003，21(1)，173 – 179，

[2] 秦进，陈笑蓉，汪维家，陆汝占. 文本分类中的特征抽取[J]. 计算机应用，2003，23(2)，45 – 46

[3] Yi – Ming Yang，Jan O Pederson. A Comparative Study on Feature Selection in Text Categorization [C]. Proc. of 14th International Conference on Machine Learning (ICML – 97)，1997，412 – 420.

[4] 黄萱菁，等. 独立于语种的文本分类方法[J]. 中文信息学报. 2000，14(.6)，1 – 7

[5] T E Dunning. Accurate methods for the statistics of surprise and coincidence [J]. Computational Linguistics，1993，19(1)，61 – 74.

[6] M Utiyama and H Isahara. Large – scale text categorization (in Japanese) [C]. 9th Annual Meeting of the Association (Japan) for Natural Language Processing，2003，385 – 388.

[7] B L Lu，K A Wang，M Utiyama and H Isahara. A part – versus – part method for massively parallel training of support vector machines [C]，Proc. of International Joint Conference on Neural Networks (IJCNN04)，Budapest，Hungary，July 26 – 29，2004. 735 – 740.

[8] B L Lu and M Ito. Task Decomposition and Module Combination Based on Class Relations：A Modular Neural Network for Pattern Classification [J]. IEEE Trans. Neural Networks，1996，10(5)，1244 – 1256.

[作者简介]

申　红(1977 –)，女(汉族)，湖北人，硕士生，主要研究领域为自然语言处理，文本分类；

吕宝粮(1960 –)，男(汉族)，山东人，工学博士，教授，博士生导师，IEEE高级会员，主要研究领域为仿脑计算机理论与模型，机器学习，自然语言处理；

内山将夫(1976 –)，男，日本人，工学博士，研究员，主要研究领域为统计学习，自然语言处理；

井佐原均(1956 –)，男，日本人，工学博士，研究员，主要研究领域为自然语言处理。

文本分类的特征提取方法比较与改进

万方数据 WANFANG DATA 文献链接

作者： 申红， 吕宝粮， 内山将夫， 井佐原均， SHEN Hong， LU Bao-liang， Utiyama Masao， Isahara Hitoshi

作者单位： 申红,吕宝粮,SHEN Hong,LU Bao-liang(上海交通大学,计算机科学与工程系,上海200030)， 内山将夫,井佐原均,Utiyama Masao,Isahara Hitoshi(国立信息与通讯技术研究所计算语言实验室,日本,京都府,619-0289)

刊名： 计算机仿真 ISTIC PKU

英文刊名： COMPUTER SIMULATION

年，卷(期)： 2006,23(3)

被引用次数： 17次

参考文献(8条)

1. 王灏;黄厚宽;田盛丰 文本分类实现技术 2003(01)

2. 秦进;陈笑蓉;汪维家;陆汝占 文本分类中的特征抽取[期刊论文]-计算机应用 2003(02)

3. Yi-Ming Yang;Jan O Pederson A Comparative Study on Feature Selection in Text Categorization[外文会议] 1997

4. 黄萱菁 独立于语种的文本分类方法[期刊论文]-中文信息学报 2000(06)

5. T E Dunning Accurate methods for the statistics of surprise and coincidence 1993(191)

6. M Utiyama;H Isahara Large-scale text categorization (in Japanese) 2003

7. B L Lu;K A Wang;M Utiyama;H Isahara A part-versus-part method for massively parallel training of support vector machines[外文会议] 2004

8. B L Lu;M Ito Task Decomposition and Module Combination Based on Class Relations:A Modular Neural Network for Pattern Classification[外文期刊] 1996(05)

本文读者也读过(4条)

1. 秦进.陈笑蓉.汪维家.陆汝占 文本分类中的特征抽取[期刊论文]-计算机应用2003,23(2)

2. 黄秀丽.王蔚 一种改进的文本分类特征选择方法[期刊论文]-计算机工程与应用2009,45(36)

3. 朱靖波.陈文亮.ZHU Jing-bo.CHEN Wen-liang 基于领域知识的文本分类[期刊论文]-东北大学学报（自然科学版）2005,26(8)

4. 呼声波.刘希玉 网页分类中特征提取方法的比较与改进[期刊论文]-山东师范大学学报（自然科学版）2008,23(3)

引证文献(17条)

1. 李新福 组合降维技术在中文网页分类中的应用[期刊论文]-计算机工程与应用 2007(24)

2. 李兆翠.刘培玉.周洪利 基于贝叶斯方法的客户端邮件过滤器的设计与实现[期刊论文]-信息技术与信息化 2007(3)

3. 郭颂.马飞 文本分类中信息增益特征选择算法的改进[期刊论文]-计算机应用与软件 2013(8)

4. 陈钒.冯志勇 语言自然节奏在文本分类中的研究与应用[期刊论文]-计算机工程与应用 2012(30)

5. 陈吕强.朱颢东.伏明兰 使用类内集中度和分层递阶约简的特征选择方法[期刊论文]-计算机工程与应用 2010(30)

6. 夏晶晶.朱颢东 基于特征辨别能力和分形维数的特征选择方法[期刊论文]-微型机与应用 2010(7)

7. 王明令 中文文本分类中特征提取方法的比较与改进[期刊论文]-兰州工业高等专科学校学报 2010(6)

8. 林啟锋.蒙祖强.陈秋莲 结合同义向量聚合和特征多类别的KNN分类算法[期刊论文]-计算机科学 2013(12)

9. 王培涌. 陈好刚. 王树峰 一种改进的中文文本特征选择方法[期刊论文]-现代计算机（专业版） 2009(12)

10. 郑雅婷. 张鹰 Web文本挖掘技术在网上购物中的应用[期刊论文]-牡丹江师范学院学报（自然科学版） 2008(4)

11. 张元虹. 郭剑毅. 龚华明. 薛征山 基于DF与LSA相结合的降维法的文本分类系统的研究[期刊论文]-山西电子技术 2008(4)

12. 张玉芳. 王勇. 刘明. 熊忠阳 新的文本分类特征选择方法研究[期刊论文]-计算机工程与应用 2013(5)

13. 熊忠阳. 蒋健. 张玉芳 新的CDF文本分类特征提取方法[期刊论文]-计算机应用 2009(7)

14. 李建林 一种基于PCA的组合特征提取文本分类方法[期刊论文]-计算机应用研究 2013(8)

15. 张素琪. 刘恩海. 贺亚. 董永峰 基于改进的免疫克隆支持向量机网页分类研究[期刊论文]-计算机工程与科学 2011(12)

16. 丁霄云. 刘功申. 孟魁 基于一类SVM的不良信息过滤算法改进[期刊论文]-计算机科学 2013(z2)

17. 肖可. 奉国和 1999～2008年国内文本分类研究文献计量分析[期刊论文]-情报学报 2010(4)

引用本文格式：申红. 吕宝粮. 内山将夫. 井佐原均. SHEN Hong. LU Bao-liang. Utiyama Masao. Isahara Hitoshi 文本分类的特征提取方法比较与改进[期刊论文]-计算机仿真 2006(3)