基于SVM新闻文本分类的研究

张国梁　肖超锋
（中国科学技术大学 信息学院）

摘　要：网络新闻自动分类已经成为当下的热点问题，SVM分类算法是文本分类中应用较为成熟的一种方法。文章针对SVM文本分类中特征选择和核函数选择的两个重要问题，在新闻文本实验环境下进行了探讨，结果表明使用互信息特征选择法且特征数在4000左右，使用SIGMOID核函数的情况下准确率与召回率均可达到97%的分类效果。
关键词：支持向量机；文本分类；特征选择；核函数选择

Research of News-Text Classification Based on SVM

Zhang Guoliang　Xiao Chaofeng
(University of Science and Technology of China)

Abstract: Today, the automatic classification of network news has become the hot research topic, the classification algorithm SVM (Support Vector Machine) is a relatively mature method on text classification. Under the experimental environment of news texts, this paper discusses two important issues about the feature selection and kernel function selection in text classification based on SVM, the results show that using IM (Mutual Information) feature selection method with about 4000 features and using SIGMOD kernel function, both the recall rate and precision of the classification can achieve 97%.
Key words: SVM (support vector machine); text classification; feature selection; kernel function selection

0 引言

随着信息技术的快速发展，网络新闻成为人们重要的信息来源之一，而网络新闻信息量大，组织混乱等特点阻碍着人们信息的获取。传统的人工分类整理的方法变得几乎不可能，机器自动文本分类突显出其重要作用。不同于其它类型的文本，新闻文本需要在尽量短的长度下讲述一个鲜明的事实。SVM分类法被认为是文本分类中效果较优秀的一种方法。目前，基于SVM文本分类存在的一些问题是：1)特征值的选择，文献[1]表明，在英文测试集上信息增益与开方检验的效果最好。2)核函数的选择，核函数作为SVM解决非线性可分问题的一种手段，并没有切实可行的指导原则。本文将通过实验方式确定在新闻文本领域中特征选择和核函数选择对于SVM法分类效果的影响。

1 特征选择算法

文本分类中维数过高的特征向量始终是不容忽视的一个因素，包含大量的无关特征，一方面降低了训练的速度，另一方面干扰了分类的准确度。特征选择就是准确识别并保留具有相当区分度的分类特征项，降低特征向量的维度。特征选择是影响分类性能的一个重要因素，常用选择算法有文档频率法、互信息法、开方检验法与信息增益法四种[1]。

(1) 文档频率法(DF)。文档频率法直接以特征项的文档频率作为参数进行选择，能够很好区分类别特征的词是中频词，高频词往往是一些通用词，而低频词对类别的贡献较小。

(2) 信息增益法(IG)。信息增益是信息论中的重要概念，衡量特征能够为分类系统带来多少信息，信息增益越大，其所包含的分类信息量越大。信息增益的计算公式如下：

$$IG(t)=p(t)\sum_{i}p(Ci|t)\log\frac{p(Ci|t)}{p(Ci)}+p(\bar{t})\sum_{i}p(Ci|\bar{t})\log\frac{p(Ci|\bar{t})}{p(Ci)} \quad (1)$$

其中，$p(t)$为训练集中包含特征项t的文档数与总文档数之比；$p(C_i)$为训练集中C_i类文档数与总文档数之比；$p(C_i|t)$为C_i类中包含t的文档数与C_i类中总文档数之比。

(3) 互信息法(MI)。互信息同样来源于信息论，度量两个信息量间的关联程度。互信息的理论依据是特征词与类别关联度越高，对类别的贡献度就越高。互信息的计算公式如下：

$$MI(t,Ci)=\log\frac{p(Ci|t)}{p(Ci)} \quad (2)$$

其中，$p(C_i)$为训练集中C_i类的文档数与总文档数之比，$p(C_i|t)$为C_i类中包含t的文档数与C_i类中总文档数之比。

(4) 开方检验(CHI)。开方统计量同样是描述特征项与类别之间的关联程度，开方统计值大则表示特征项与类别之间的独立性小，对分类的贡献就大。开方检验的计算如式（3）所示：

$$\chi(t,c)=\frac{(AD-BC)}{(A+B)(C+D)} \quad (3)$$

其中A为类c中包含特征项的文档个数；C为类c中不包含特征项的文档个数；B为非类c中包含特征项的文档个数；D为非类c中不包含特征项的文档个数。

2 SVM核函数

文本分类是一类非线性可分的问题，SVM使用高维映射处理方法将非线性可分的问题转化为线性可分，重点在于选择合适的核函数，常用核函数有：

(1) 线性核函数
$$K(x,y)=x\bullet y \quad (4)$$
(2) 多项式核函数
$$K(x,y)=[x\bullet y+1]^q \quad (5)$$
(3) RBF核函数
$$K(x,y)=\exp\{-r\|x-y\|^2\} \quad (6)$$
(4) Sigmoid核函数
$$K(x,y)=\tanh\{r(x\bullet y)+c\} \quad (7)$$

实践表明，核函数的选取对分类效果有巨大的影响。SVM中核函数的选取没有确定的指导方法，不同数据情况偏向不同的核函数，RBF在不同的应用中取得较为均衡的效果，文本分类中选择线性核函数的效果较好。

3 实验

实验目的是为了在真实的应用背景下，以优化文本分类性能为目标，获得特征选择与SVM核函数选择的一些

指导。本文选择Sogou开放实验平台提供的新闻分类语料库[5]，来源于Sohu新闻网站保存的大量经过编辑手工整理与分类的新闻语料与对应的分类信息，每篇新闻的长度在100-300个字左右，属于短新闻。实验测试财经类与非财经二分类，从语料库中抽取1000篇财经类新闻，抽取非财经类新闻1000篇，共2000篇文档作为样本集。

实验1. 特征选择算法对于分类效果的影响

在实验1中选取线性核函数，针对每个特征选择算法，试着代入不同的阈值，选取最好的结果进行对比。实验时将样本集分成10份做交叉实验，选5组实验结果作为对比，见表1。由表1可得到以下结论：DF的准确率要偏低一些，其余三种方法的准确率基本相当；MI与DF法的召回率要远高于IG与CHI的特征选择法，DF要略高于MI；DF的F1指标要略好于MI，而远高于另外两种方法。在新闻文本识别中，低频词数量较多，而文档长度相对较小，待测试的文本中缺少很多特征项，造成MI与CHI方法准确率虽然很高，但漏掉了很多相关文本。

表1 特征选择实验对比

特征算法 \ 实验序号		1	2	3	4	5
DF	准确率	1.00	0.966	0.99	0.97	0.95
	召回率	0.74	0.85	0.85	0.88	0.92
	F1	0.85	0.90	0.91	0.92	0.93
IG	准确率	1.0	1.0	1.0	1.0	1.0
	召回率	0.38	0.35	0.57	0.31	0.42
	F1	0.55	0.52	0.73	0.47	0.59
MI	准确率	1.0	1.0	1.0	0.99	1.0
	召回率	0.72	0.77	0.83	0.79	0.84
	F1	0.83	0.87	0.91	0.88	0.91
CHI	准确率	1.00	1.00	1.00	0.98	0.98
	召回率	0.34	0.47	0.58	0.41	0.58
	F1	0.51	0.64	0.73	0.58	0.73

实验2. 特征数的选择

实验1中发现，特征数的选取对分类效果的影响很大，但IG与CHI特征数并不完全可控，而且其最好的分类效果也不理想，因此此次实验中只针对MI与DF进行相关实验。实验结果分别见表2与表3。

表2 DF特征数实验

特征数	正确数	召回数	召回率	准确率	F1
1114	92	104	0.920	0.885	0.902
1597	94	102	0.940	0.920	0.931
2489	93	100	0.930	0.930	0.930
3152	91	96	0.910	0.948	0.929
6032	84	86	0.84	0.977	0.903
10906	72	74	0.720	0.973	0.828

表3 MI特征数实验

特征数	正确数	召回数	召回率	准确率	F1
2744	74	74	0.740	1.000	0.851
3668	97	99	0.970	0.980	0.975
4695	100	110	1.000	0.909	0.952
5615	81	81	0.810	1.000	0.895
8272	72	72	0.720	1.000	0.837
11554	64	64	0.640	1.000	0.780

从表2与3发现：DF的召回率要略好于MI；而MI的准确率要略好于DF；DF的F1指标平均要略好于MI，但MI在4000左右时分类效果较为显著；从宏观上看，两种方法都会在特征数取到某段时分类效果达到一个峰值；分类效果并不随着特征数的增多而变好，相反特征数太多时分类效果反而很差；另外，MI对特征数较DF对特征数敏感。

实验3. 核函数的选择

从上述实验中，发现DF与IM在新闻文本分类中效果较好，本实验分别就两种特征选择方法对SVM四种核函数进行了选择，特征数选取其最优时的值，分别得到表4与表5的实验结果。

从表4与表5中得到结论：POLY核对于文本分类不可用；RBF核的召回率要好于另外两种核，但其准确率却极差；在DF特征选择时，LINEAR核的分类效果要远好于SIGMOID核；在IM特征选择时，SIGMOID的特征选择略好于LINEAR核，最终结果表明IM与SIGMOID核的组合效果显著。

表4 DF时核函数实验

核函数	正确数	召回数	召回率	准确率	F1
LINEAR	95	102	0.941	0.931	0.941
POLY	0	0	0	—	—
RBF	100	193	1.000	0.518	0.683
SIGMOID	69	89	0.690	0.775	0.730

表5 IM时核函数实验

核函数	正确数	召回数	召回率	准确率	F1
LINEAR	97	101	0.970	0.960	0.965
POLY	0	0	0	—	—
RBF	100	188	1.000	0.532	0.694
SIGMOID	97	100	0.970	0.970	0.970

4 结论

本文研究了在文本分类环境下，基于SVM文本分类的特征选择与核函数选择对于分类效果的影响，取得了一些重要结论。实验结果表明，短小文本自动分类中，互信息与文档频率法做特征选择的效果最好；分类效果并不随着特征数的增多而变好，而是分别在4000和2000左右时达到最好；此应用环境中选取线性核函数与SIGMOID核函数的分类效果明显好于多项式核函数与径向基核函数。

参考文献：

[1] Yang Y M, Pederson J O. A comparative study on feature selection in text categorization[C]// Proc. of 14th International Conference on Machine Learning, 1997: 412-429.
[2] Tan A-H, Yu P. A comparative study on Chinese text categorization methods[C]//PRICAI 2000 Workshop on Text and Web Mining. Melbourne, 2000: 24-35.
[3] 贾洄, 梁久祯. 基于支持向量机的中文网页自动分类[J]. 计算机工程, 2005,31(10): 145-147
[4] Vapnik V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.
[5] 搜狗实验室.文本分类语料库[EB/OL]. http://www.sogou.com/labs/dl/c.html

作者简介：

张国梁，中国科学技术大学，硕士
手机：15956909540
电子信箱：zhgl1987@mail.ustc.edu.cn
通信地址：安徽省合肥市中国科学技术大学西区11#228室（230027）