朴素贝叶斯算法的改进与应用

赵文涛[1,2]，孟令军[1]，赵好好[1]，王春春[1]
（1. 河南理工大学 计算机科学与技术学院，河南 焦作 454000；
2. 河南省普通高等学校矿山信息化研究重点实验室，河南 焦作 454000）

摘要：朴素贝叶斯算法是分类算法中最经典、最有影响的算法之一，但仍然存在一些不足之处。针对该算法中下溢问题，对算法基本公式进行了优化改进。针对 NB 算法中准确率问题，结合类别核心词思想和改进后的 TFIDF 算法，提出了一种基于类别核心词和改进型 TFIDF 的朴素贝叶斯 CIT-NB 算法。将改进后的算法应用于新闻数据集文本分类，实验结果表明，CIT-NB 算法的分类性能明显优于原始朴素贝叶斯算法和基于 TFIDF 的分类算法。

关键词：朴素贝叶斯；下溢；核心关键词；TFIDF；文本分类
中图分类号：TP391 **文献标识码**：A **文章编号**：1000－8829(2016)02－0143－05

Improvement and Applications of the Naive Algorithm

ZHAO Wen-tao[1,2], MENG Ling-jun[1], ZHAO Hao-hao[1], WANG Chun-chun[1]
(1. College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China;
2. Opening Project of Key Laboratory of Mine Informatization, Henan Polytechnic University, Jiaozuo 454000, China)

Abstract: Naive Bayes algorithm is one of the most classical and influential classified algorithms, but it still has some aspects to be improved. To solve the problem of underflow, the fundamental formulae is improved. To improve the accuracy of the algorithm, a new algorithm named CIT-NB is proposed based on classified core words and the improved TFIDF algorithm. CIT-NB algorithm is realized in the experiment. Applying the CIT-NB algorithm introduced above to dataset of news classifying, the experimental results indicate that compared with NB algorithm, CIT-NB algorithm performs better in recall and precision, which proves the validity of CIT-NB algorithm obviously.

Key words: naive Bayes; underflow; core words; TFIDF; text classification

文献[1]指出分类问题是数据挖掘领域研究和应用最为广泛的问题之一。目前最常用的分类算法主要有贝叶斯算法、C4.5 决策树算法、支持向量机算法(SVM，support vector machine)等。文献[2]指出 C4.5 决策树算法受到数据集中奇异数据的影响。文献[3]指出 SVM 算法在训练大数据集时速度过慢。朴素贝叶斯(naive Bayes classification)算法由于坚定、严密的数学理论基础成为数据挖掘文本分类领域中最经典的分类算法。

朴素贝叶斯算法中要求属性间相互独立的这一假设在现实中往往难以得到满足。尽管如此，朴素贝叶斯算法还是由于其理论背景、规则简单且易于理解、分类速度快、分类精度高而备受推崇。

本文针对朴素贝叶斯算法溢出问题，对其基本公式进行了修改，并为了提高算法性能，结合类别核心词思想以及使用改进后的 TFIDF(term frequency-inverse ducument frequency)算法计算特征词权重，提出了一种改进的算法基于类别核心词和改进型 TFIDF 的朴素贝叶斯(CIT-NB，Core words and Improved TFIDF Naive Bayes classification)算法，该算法在本文中已经实现，在与传统朴素贝叶斯算法、基于 TFIDF 的朴素贝叶斯算法进行比较时发现，该算法性能在召回率和查准率(精度)方面都有了显著提升。

1 朴素贝叶斯算法简介

1.1 朴素贝叶斯算法简介

收稿日期：2015－05－04
基金项目：河南省科技攻关项目(142402210435)；河南省高等学校矿山信息化重点学科开放实验室开放基金项目(ky2012－02)
作者简介：赵文涛(1963—)，男，硕士，教授，硕士生导师，从事计算机专业教学和数据库技术、信息系统、大数据等方面研究；孟令军，硕士研究生，主要研究方向是数据挖掘，云计算。

朴素贝叶斯算法由于坚定的贝叶斯理论基础被广泛应用于文本分类领域。该算法基于条件独立性假设,即假定给定目标对象的各个属性之间相互独立,互不影响,通过计算目标对象的先验概率,利用贝叶斯定理计算出其后验概率,即该对象属于某类的概率,然后比较后验概率的大小进行决策分类。常用的模型包括多变量伯努利模型和多项式模型,本文采用后者。

朴素贝叶斯算法具体描述:

设有样本数据集 $D=\{D_1,D_2,\cdots,D_n\}$,每一个样本数据的属性集 $X=\{X_1,X_2,\cdots,X_d\}$,类变量 $Y=\{Y_1,Y_2,\cdots,Y_m\}$,D 可以分成 Y_m 个类别。将 X 和 Y 视为随机变量,则称 $P(Y|X)$ 即为 Y 的后验概率,$P(Y)$ 为 Y 的先验概率,由贝叶斯定理可知,后验概率可以由先验概率 $P(Y)$、类条件概率 $P(X|Y)$、和证据 $P(X)$ 表示:

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

对于式(1)中的类条件概率 $P(X|Y)$ 有两种不同的计算方法:朴素贝叶斯分类器和贝叶斯信念网络,本文使用朴素贝叶斯分类器计算类条件概率 $P(X|Y)$。正如上文所说,朴素贝叶斯分类基于条件独立性假设,给定类别标号 y 的情况下,条件独立性假设可以形式化地表示为

$$P(X|Y=y)=\prod_{i=1}^{d}P(X_i|Y=y) \tag{2}$$

由式(1)、式(2)可知,在条件独立性假设的前提下,通过给定 Y,计算每个条件概率 $P(X|Y)$ 就可以求出类条件概率,进而求出后验概率:

$$P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}P(X_i|Y)}{P(X)} \tag{3}$$

在比较后验概率 $P(Y|X)$ 大小时,$P(X)$ 固定不变,那么只需要知道式(3)分子相对大小即可,即

$$Y_{\max}=\operatorname{argmax}P(Y|X)=\operatorname{argmax}P(Y)\prod_{i=1}^{d}P(X_i|Y) \tag{4}$$

1.2　朴素贝叶斯问题分析

朴素贝叶斯算法是分类领域最经典、最常用的分类算法之一,但仍然存在许多不足,具体表现在以下几个方面:

① 式(4)右端大量条件概率相乘而可能出现的朴素贝叶斯算法下溢问题。具体而言,大量条件概率值都是小于 1 的数,在高维属性集中甚至各个特征词的条件概率趋近于 0,甚至一个或多个特征词后验概率值在计算机处理后取 0,这样,大量趋近于 0 的条件概率相乘容易出现下溢,这样会极大地降低 NB 算法的分类性能。

② 从朴素贝叶斯算法基本公式来看,文本的类别与特征词的词频有很大关系,词频越大,说明该词越具有代表性。但事实上有可能并非如此,有些低频特征词可能更能代表类别特征,即词频并不足以完全代表某个词在分类中的重要程度,这样也会在某种程度上降低算法的分类性能。

③ 文本分类时,特征空间由预处理后的特征词组成,初始特征空间达到数万维,计算量过大,特征词评估难度较大。此外,朴素贝叶斯算法在假设特征词权重全部一致(为①),这会赋予对决策分类无关或影响较小的、受到噪声污染的特征词的影响,从而降低分类器性能。

2　朴素贝叶斯算法改进

本文提出的基于类别核心词和改进型 TFIDF 的朴素贝叶斯 CIT-NB 算法针对上面提到的朴素贝叶斯算法的问题,逐个解决改善,旨在提高分类器性能。

2.1　解决下溢问题

研究式(4)可以发现,在对后验概率进行比较时,会有大量条件概率相乘,不排除下溢情况出现,这就会导致后验概率比较结果的不确定性,从而降低算法性能。因此这里对式(4)取对数避免下溢情况出现,这样式(4)可转换为

$$Y_{\max}=\operatorname{argmax}P(Y|X)=\ln P(Y)+\sum_{k=1}^{r}\ln P(X_i|Y) \tag{5}$$

研究式(5)右边可以发现条件概率存在值为 0 的情况,这会造成分类结果的不正确性,从而影响算法性能,因此这里通过 Laplace 平滑技术避免该值为 0,可简单地进行加 1 处理,该条件概率计算如下:

$$P(X_i|Y)=\frac{T_k+1}{\sum_k T_k+T_C} \tag{6}$$

式中,T_k 为特征项在该类出现的次数,代表该类所有特征项总数;T_C 为平滑因子,代表所有特征项总数。

2.2　使用类别核心词思想

从贝叶斯公式直观来看,文本的类别与特征词的词频有很大关系,词频越大,说明该词越具有代表性。但事实上,很多时候类别核心词(标题、摘要和关键词)更能代表类别特征,即词频并不足以完全代表某个词在分类中的重要程度。因为人工分类文本时,人们更看重核心词。即认为在标题、摘要和关键词部分出现的特征词的重要程度更高,它们能更好地代表该类文本的类别特征,这些词即类别核心词。

基于此,可以做出改进,主要思路先在文本分词的基础上选出每类的类别核心词;然后计算类别核心词相对于所有特征词的重要程度;最后在朴素贝叶斯分类方法中,将类别核心词的重要程度融合进来,对测试文本进行分类。实验表明,这种基于类别核心词的改

进后的算法效率较高,准确性也较好。

2.3 使用改进后的 TFIDF 算法计算权重

朴素贝叶斯算法认为所有条件属性对于决策分类的重要性一样,即权重都是 1,这样会让许多与分类无关的、受到噪声污染的属性与其他关键属性具有同样的权重值,因此,只是简单统计词频是不合理的。基于此,众多科研人员将各种文本赋权算法应用于朴素贝叶斯分类器的构造中。文献[4]提出了赋予不同属性不同权值的加权朴素贝叶斯 WNB(weighted naive bayes)模型,文献[5]提出的 APNBC(adjusted probability naive bayes classifier)模型则通过二次加权调整后验概率后进行分类。TFIDF 是用来评估一个关键词对于语料库或文件集中某份文件的重要程度的常用算法,单词在某份文件中的 TF-IDF 值越大,则表示该词对该文件的贡献值越大,反之,则该词对该文件影响越小。文献[6]和文献[7]等都将 TFIDF 算法应用到朴素贝叶斯算法的构造中。

但是将传统 TFIDF 算法应用到朴素贝叶斯分类器有一个缺点,没有考虑到类间、类内词汇的分布情况,即某词汇若在某类文档中高频率出现,则该词汇的分类性能显然很强,但是,根据经典 TFIDF 算法,若某词汇在很多文档中高频率出现,则其 IDF 较小,反之,IDF 较大。这同样会降低 NB 算法的分类性能。这对分类器决策性能有较大影响,所以本文将词汇分布考虑在内,采用张玉芳等在文献[8]中提出的改进 TFIDF 算法,即对 IDF 计算方法进行如下改进,假设在某类文档中的 t 个文档都存在某特征词,除了该类别的文档,数据集中的另外 h 个文档也含有该特征词,则 IDF 进行如下计算:

$$IDF = \log(t \times N/n) = \log\lfloor t \times N/(t+h) \rfloor \tag{7}$$

2.4 解决维度过大问题

对文本进行处理时,通常需要采用向量空间模型进行描述,进行文本分类处理时候,将其向量化处理为由 n 个特征项 $X_1, X_2, \cdots, X_i, \cdots, X_n$ 构成的向量空间。其中特征项可以是字、词、词组或短语等。而要确定组成特征项,首先需要对文本内容进行分词处理。分词处理后得到的特征项组成的向量空间通常维度较高,影响分类方法的性能,因此需要从中选取对分类有效的特征项构成特征项库,达到对向量空间降维,提高分类效率的目的,这个过程被称为特征提取。互信息是统计学用于描述两个变量相关性的方法,在被用于文本特征统计的时候,是对特征项与类之间共享信息的度量。

采用适当的方法进行特征提取,建立有效的特征项库常用的特征提取方法有信息增益(IG,information gain)、互信息(MI,mutual information)、CHI 估计和文档频次(DF,document frequency)等。它们都是按一定的评估方法计算出每个特征项的度量值,再选取排列靠前的 N 个特征项作为特征项库,互信息定义如下:

$$I(X_i, Y) = \log \frac{p(X_i \mid Y)}{p(X_i)} \approx \frac{A \times N}{B \times C} \tag{8}$$

式中,$p(X_i)$ 是训练集中特征词 X_i 出现的概率,$p(X_i \mid Y)$ 是类别 Y 中 X_i 出现的概率,为了方便计算,A 为类别 Y 和 X_i 同时出现的频率,N 为训练集文本的频数,B 为训练集中出现特征词 X_i 的文本频数,C 为训练集中的 Y 类文本频数。但是式(8)没有体现出本文提出的特征词权重,做如下改进:

$$I'(X_i, Y) = p(X_i)p(Y)p(X_i \mid Y)I(X_i, Y) \tag{9}$$

将词频与特征词在每个类中的分布考虑进去,如果特征词互信息值高于阈值,则选取该特征词作为特征空间中的词,否则移除该词。这有效降低了特征空间维度和噪声词的影响。

3 CIT-NB 算法

本文提出的基于类别核心词和改进型 TFIDF 的朴素贝叶斯 CIT-NB 算法主要包括数据集的预处理阶段、训练阶段和分类阶段,其中,训练阶段通过使用部分数据集构造分类器,是整个算法的核心所在。

3.1 预处理过程

步骤 1:以分类效果为依据,删掉冗余属性,选择出训练集中属性集的优化组合,使用中科院 ICTCLAS 汉语分词系统[9]对优化后的文本内容进行分词,得到 Train 表。

步骤 2:如果属性优化组合里含有连续属性,对连续属性的特征值进行离散化,文献[10]提供的数据预处理工具 PreProcessor 进行离散化,离散化的参数选用默认设置。对于数据集中的缺失项,按照缺失项出现次数最多的值补齐。

步骤 3:使用分词工具对训练集中的标题、摘要和关键词进行分词,得到 Train_k 表,其中所有特征词即类别词。

步骤 4:分别对 Train 表和 Train_k 表进行词频统计,将 Train 表中频数小于 4 的词移除,根据式(9)对 Train 表所有词语进行互信息计算,移除互信息值较小的词语,将剩下高于阈值的词语作为特征词条,同时降低噪声污染。

3.2 训练过程

步骤 1:首先利用改进后的 TFIDF 算法对类文本中的特征项 计算 TFIDF 值 $N(w_x, d_i)$。

步骤 2:计算类文本中类别词的个数 n_m 及 Y_m 类文本特征词的总个数 n'_m。

步骤 3:计算 Y_m 类文本类别词词频之和

$\sum_{j=1}^{n}\sum_{i=1}^{d}N(w_x,d_i)$ 及该类所有类别词平均词频 $f_m=\frac{\sum_{j=1}^{n}\sum_{i=1}^{d}N(w_x,d_i)}{n_m}$。

步骤 4：计算类所有特征词词频之和 $\sum_{j=1}^{n'}\sum_{i=1}^{d}N(w_x,d_i)$ 及平均词频 $f'_m=\frac{\sum_{j=1}^{n'}\sum_{i=1}^{d}N(w_x,d_i)}{n'_m}$。

步骤 5：计算 Y_m 类类别词重要度 $k=\frac{f_m}{f_m'}$。

步骤 6：判断特征词 x 是否属于 Y_m 类类别词，如果是，则 x 权重为 $w_x=N(w_x,d_i)\times\frac{k}{1-k}$；否则，$w_x=N(w_x,d_i)$。

步骤 7：最终根据特征词的权值进行先验概率的计算和分类器的构造。

4　性能分析

在 CIT-NB 算法中，在原始 NB 算法的基础上，通过对特征词赋权值、提取类别核心词重要度等手段，明显提高了分类器性能。

4.1　实验环境、数据集和评价指标

本文实验环境：联想 Z470 机器，Intel Core i3-2410M，2.3 GHz CPU，2GB 内存，750 GB 硬盘，Windows 7 操作系统，JAVA 编程语言，Eclipse-4.3.2 开发环境。

本文使用的实验数据集是从复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供的中文语料库的 20 个分类中取出 7 个类（Art，History，Space，Mine，Law，Sports，Energy）进行文本分类，其中训练语料和测试语料分别有 9804、9833 篇文档，从每类中选出 200 个文档，共计 1400 个文档，每类中 150 个文档页面作为训练集，50 个页面作为待测试页面。

本文使用召回率和查准率（精度）两个指标对实验的分类结果进行评价，具体定义如下：

$$\text{查准率}(P)=\frac{A}{A+B}\quad \text{召回率}(R)=\frac{A}{A+C}$$

其中，A，B，C，D 的含义和关系如表 1 所示。

表 1　参数及其含义

	相关页面	不相关页面
检索	A	B
未被检索	C	D

A 指的是算法正确分类的文档数，B 是算法检索到的不相关页面数，C 指的是未被检索到的相关页面数，D 则指的是未被检索的不相关页面。查准率和召回率是广泛应用于信息检索领域的两个指标，查准率反映出算法的精度。$F=\frac{2(P\times R)}{(P+R)}$，通过 F-measure 指标来综合比较算法的分类性能，F-measure 越大，分类性能越好。

4.2　实验结果

各类训练文本 k 值如表 2 所示。

表 2　各类训练文本 k 值

分类	Art	History	Space	Mine	Law	Sports	Energy
k 值	0.6923	0.7120	0.6824	0.7217	0.6937	0.7439	0.7824

分别使用原始 NB 法、TFIDF-NB 法和 CIT-NB 法对数据集进行分类，结果如表 3 所示。

表 3　NB 算法、TFIDF-NB 算法 CIT-NB 算法性能对比

算法		Art	History	Space	Mine	Law	Sports	Energy
NB	P	0.8963	0.9012	0.9085	0.9124	0.8379	0.9129	0.9258
	R	0.8672	0.8796	0.8823	0.9122	0.9312	0.9224	0.9371
	F	0.8815	0.8903	0.8952	0.9123	0.8821	0.9176	0.9314
TFIDF-NB	P	0.9033	0.9264	0.9062	0.9388	0.8865	0.9397	0.9477
	R	0.8943	0.9071	0.9102	0.9323	0.9471	0.9482	0.9480
	F	0.8988	0.9166	0.9082	0.9355	0.9158	0.9439	0.9476
CIT-NB	P	0.9381	0.9410	0.9210	0.9321	0.9044	0.9417	0.9577
	R	0.9147	0.9211	0.9102	0.9458	0.9628	0.9526	0.9582
	F	0.9263	0.9309	0.9156	0.9389	0.9327	0.9471	0.9580

F 指标如图 1 所示。

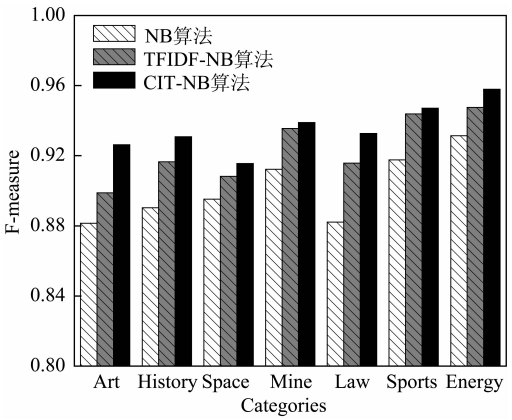


图 1　F-measure 对比图

4.3　结果分析

结合表 2 和表 3 可以看出，Energy 类文本数据的 k 值最大，从表 3 可以看出，三种方法对数据进行分类时 Energy 类 F 值最大，这说明其文本的类别核心词重要度越高，该类文本的特征也越明显，其区分文本的能力越强，从某种程度而言，分类器性能与类别核心词有关。该算法也更适合于有类别核心词的文本分类。

从图 1 可以看出，CIT-NB 算法的分类性能明显优于 NB 算法和 CIT-NB 算法，这是由于 CIT-NB 算法综合考虑类别核心词以及类内、类间的特征项分布，类内

分布越均匀,类间分布越不均匀,核心类别词频率越高,对类的分别能力越强,贡献值越大,分类效果就越好。

5 结束语

本文在朴素贝叶斯理论基础上,避免了数据下溢问题的同时,又结合类别核心词思想和改进后的TFIDF算法对特征词加权的方法对原始朴素贝叶斯算法进行了改进,提出了CIT-NB算法。实验数据证明,CIT-NB算法性能取得了显著的提高。

但是,本文在进行实验时,从语料库中的20个类别中选择了7个类别,用的数据集较小。再好的算法遇到大数据都会出现一定的问题。随着数据量的爆炸式增长,算法必须能够处理大数据集。在所知范围内,已经有比较多朴素贝叶斯算法的并行化实现[11-13],但是并没有出现本文提出CIT-NB算法的并行化实现,接下来将结合信息增益、数据倾斜和Hadoop等相关方面对此进行进一步深入研究。

参考文献:

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. 数据挖掘导论(中文版)[M]. 范明,范宏建,等,译. 北京:人民邮电出版社,2011:139-141.

[2] Buades A, Coll B, Morel J M. A non-local algorithm for image denoising[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005:60-65.

[3] Wegener D, Mock M, Adranale D, et al. Toolkit-based high-performance data mining of large data on mapreduce clusters [C]//IEEE International Conference on Data Mining Workshops. 2009:296-301.

[4] Webb G I, Pazzan M J. Adjusted probability Naive Bayesian induction [C]//Proceedings of the 11th Australian Joint Conference on Artificial Intelligence. 1998:285-295.

[5] Zhang H, Sheng S L. Learning weighted Naive Bayes with accurate ranking[C]// Fourth IEEE International Conference on Data Mining. 2004:567-570.

[6] 郑炜,沈文,张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J]. 西北工业大学学报,2010,28(4):622-627.

[7] 袁方,苑俊英. 基于类别核心词的朴素贝叶斯中文文本分类[J]. 山东大学学报(理学版),2006,41(3):46-49.

[8] 张玉芳,彭时名,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.

[9] 邸鹏,段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理,2014,29(1):71-75.

[10] 邓甦,付长贺. 四种贝叶斯分类器及其比较[J]. 沈阳师范大学学报(自然科学版),2008,26(1):31-33.

[11] Chu C T, Kim S K, Lin Y A, et al. Map-reduce for machine learning on multicore [C]//20th Annual Conference on Neural Information Processing Systems. 2007:281-288.

[12] 曾青华,袁家斌,张云渊. 基于Hadoop的贝叶斯过滤Mapreduce模型[J]. 计算机工程,2013,39(11):57-60.

[13] 卫洁,石洪波,冀素琴. 基于Hadoop的分布式朴素贝叶斯文本分类[J]. 计算机系统应用,2012,21(2):210-213.

□

征 稿 启 事

《测控技术》以国防及工业领域中的测试测量与控制技术为核心,涉及多个学科领域,如计算机技术、电子技术、自动控制技术、传感器及仪表技术、网络与通信技术、自动测试技术和虚拟与仿真技术等。本刊开辟的栏目和内容主要有:专家论坛、综述、技术专题、仪表与传感器、数据采集与处理、控制系统、总线与网络、软件纵横、理论与实践、科技动态。

欢迎业界专家、教授、中青年学者及工程技术人员投稿,其论文应反映当前国内外测控技术及其应用方面的先进理论、技术成果或发展趋势;有实际应用背景,并有相应的实验结果;未在国内外期刊上公开发表过的有独创内容的论文。主要内容和结论要明确。具体要求如下:

★ 文题醒目切题,简明扼要,中英文对照;

★ 给出中英文摘要、关键词。摘要应包括研究目的、过程和方法、主要结果和结论,要准确反映论文要点,一般在200~300字左右,关键词5~8个;

★ 综述性文章字数要求在5000字左右,其他技术论文要求在4000字左右,典型图表一般不超过6幅,图表中如有英文一定要译成中文。外文字母分清大、小写,物理量用斜体,文中科技词语、计量单位与符号应符合国家公布的标准要求;

★ 给出必要的参考文献,并在文内用上角标标出。新理论、新观点应给出出处;

★ 来稿一律不退,请自留底稿,超过3个月未接到通知者请与编辑部联系,录用稿编辑部有删改权,属保密部分,请作者慎重处理,文责自负。录用稿件将酌情收取少量版面费,稿费从优;

★ 来稿请注明所投栏目;若课题受到资助,请注明资助项目名称及代号;

★ 给出作者简介,包括:姓名、性别、出生年、籍贯、职称、学位、研究方向、现从事的研究工作、曾获得的奖项;并给出详细通信地址、邮编及电话,以便联系。

详情请登录:www.mct.com.cn

通信地址:北京亦庄经济技术开发区经海二路29号院9号楼五层《测控技术》杂志社

邮政编码:101111 电话:010-65676316,65665486 传真:010-65670337