文章编号:1672-3961(2014)01-0013-06 DOI:10.6040/j.issn.1672-3961.2.2013.274

基于中文人名用字特征的性别判定方法

于江德[1],赵红丹[1],郑勃举[1],余正涛[2]
(1. 安阳师范学院计算机与信息工程学院,河南 安阳 455000;
2. 昆明理工大学信息工程与自动化学院,云南 昆明 650051)

摘要: 基于中文人名用字具有的较强的性别区分性,提出一种利用朴素贝叶斯分类器对中文人名性别进行判定的方法,该方法将每个中文人名中的第一个字($字_1$)、第二个字($字_2$)、第一和第二个字组合($字_1$ $字_2$)作为区分特征,利用朴素贝叶斯分类方法对该人名所属性别进行判定。在412 775个中文人名语料上采用10重交叉验证法进行训练和测试,对比了依据不同区分特征组合进行性别判定的准确率,分别采用$字_1$,$字_2$,$字_1$ + $字_2$,$字_1$ + $字_1$ $字_2$,$字_2$ + $字_1$ $字_2$,$字_1$ + $字_2$ + $字_1$ $字_2$(全部区分特征)构成的特征组合进行性别判定,平均判定准确率分别为72.75%,86.92%,88.84%,87.37%,89.35%,90.06%,取得的最好平均判定准确率为90.06%。

关键词: 中文人名;性别判定;朴素贝叶斯分类;用字特征;特征组合;区分特征
中图分类号: TP391 **文献标志码:** A

A method of gender discrimination based on character feature of Chinese names

YU Jiang-de[1], ZHAO Hong-dan[1], ZHENG Bo-ju[1], YU Zheng-tao[2]
(1. School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, China;
2. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China)

Abstract: Based on the strong gender discrimination of Chinese names, a method of gender discrimination based on character feature of Chinese names using naïve Bayes classifier was presented. In this method, the first character of each Chinese name (Zi_1), the second character (Zi_2), the first and the second characters (Zi_1Zi_2) were regarded as distinguishing features. The naïve Bayes classification method was used for gender discrimination of Chinese names. Training and testing were done on 412 775 Chinese names corpus using 10 fold cross validation method, and comparative experiments were done according to the different feature combinations, they were Zi_1, Zi_2, $Zi_1 + Zi_2$, $Zi_1 + Zi_1Zi_2$, $Zi_2 + Zi_1Zi_2$, $Zi_1 + Zi_2 + Zi_1Zi_2$(all the distinguishing features). The average accuracy were as followings in turn, 72.75%, 86.92%, 88.84%, 87.37%, 89.35%, 90.06%, of which the best average accuracy was 90.06%.

Key words: Chinese names; gender discrimination; naïve Bayes classification; character feature; feature combination; distinguishing feature

收稿日期: 2013-06-28 **网络出版时间:** 2013-11-22 14:25
网络出版地址: http://www.cnki.net/kcms/detail/37.1391.T.20131122.1425.004.html
基金项目: 国家自然科学基金资助项目(60863011);河南省基础与前沿技术研究计划资助项目(112300410182)
作者简介: 于江德(1971-),男,河南林州人,副教授,博士,主要研究方向为计算语言学,中文信息处理与机器学习等.
E-mail: jiangde_yu@163.com

0 引言

姓名是人类为区分个体而赋予每个人特定的名称符号。人的命名受历史、时代、社会、民族、家庭等诸多文化因素制约[1-2]。中文人名有着极其丰富的文化内涵[3-4]。其所蕴涵的思想理念无比深湛,凝聚了数千年华夏文化的历史积淀,蕴藏着中华民族的智慧和精神,充分显示了中华文明的浓厚底蕴。一个人的名字通常有一定的含义,可以这样说,中文人名与中国文化互为表里,渗透在国人骨子里的文化精髓,比较集中地反映在中文人名上面。中文人名通常具有较强的性别区分性[1,5-6],我们从一个陌生人的名字可推测其是男性或女性,且准确率十之八九,可谓"听其名,知其性"。我们可以从一个陌生人的名字推断其性别,计算机是否可以模拟人的这种智能,或者说我们是否可以设计一套程序实现中文人名性别的判定,这正是本研究的主要工作。该研究对深层次的中文信息处理具有重大意义。例如,可以提高中文人名识别和指代消解的准确率,进而促进中文文本的篇章理解。本研究通过对中文人名语料中男女不同性别的用字情况统计分析,提出1种基于中文人名用字特征的性别判定方法,该方法仅仅根据名字中的用字特征,利用朴素贝叶斯(naïve Bayes,NB)分类方法进行男女性别的判定,并在412 775个中文人名构成的语料上进行训练和测试,实验结果表明,该方法简单可行,取得的最高平均判定准确率达到90.06%。

1 朴素贝叶斯分类方法

朴素贝叶斯分类方法是目前公认的1种简单有效的分类方法[7-8],它是1种应用基于独立性假设的贝叶斯公式的简单概率分类方法,有着广泛的应用[9-10],如模式识别[11-12]、自然语言处理[13-15]、规划编制[16-18]等领域。在朴素贝叶斯分类方法的研究与应用中,该方法也有许多改进及优化[19-26]。朴素贝叶斯分类方法形式化描述如下:在分类问题中,常常需要把一个事物划分类别。一个事物具有很多特征,把它的众多特征看作一个向量,即 $\boldsymbol{F}=(F_1,F_2,F_3,\cdots,F_n)$,用 $\boldsymbol{F}$ 这个特征向量来表征这个事物。假定有 m 个类别,用集合 $C=\{C_1,C_2,C_3,\cdots,C_m\}$ 表示。朴素贝叶斯分类就是由给定的一个数据样本 F,来求解 F 属于某个类别 C_i 的概率:$P(C_i|F)$。一般情况下,直接计算条件概率 $P(C_i|F)$ 比较困难,而概率 $P(C_i)$,$P(F|C_i)$ 可以从训练数据集中求得。根据贝叶斯公式,

$$P(C_i|F)=\frac{P(C_i)P(F|C_i)}{P(F)},\qquad(1)$$

可以将后验概率 $P(C_i|F)$ 的求解转换为先验概率 $P(C_i)$ 和 $P(F|C_i)$ 的求解。又由于假设表征数据样本 F 的各特征相互独立,所以

$$P(F|C_i)=\prod_{k=1}^{n}P(F_k|C_i)。\qquad(2)$$

由于 $P(F)$ 对于所有类别都相同,显然

$$\arg\max_C P(C_i|F)=\arg\max_C\frac{P(C_i)P(F|C_i)}{P(F)}=\arg\max_C P(C_i)P(F|C_i)。\qquad(3)$$

由以上论述可知,在朴素贝叶斯分类器的结构中(如图1所示),只有一个类节点,其他节点表示分类事物的各个特征属性,每个属性节点有且只有一个父节点,即类节点,且各个属性节点之间是相互独立的。由图1所示结构,根据朴素贝叶斯分类原理,对一个未知类别的样本 F,可以先分别计算出 F 属于每个类别 C_i 的概率 $P(C_i|F)$,然后选择概率最大的作为其类别。

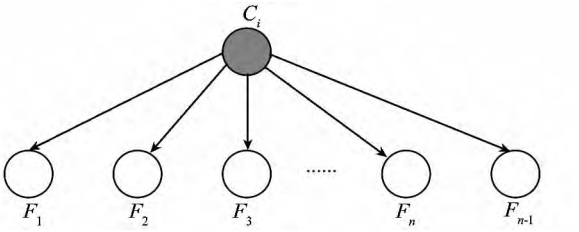

图1 朴素贝叶斯分类器的图形结构

Fig.1 Graphical structure for naïve Bayes classifier

2 基于中文人名用字特征的性别判定

根据中文人名用字特征判定性别是一个典型的二分类问题,本研究利用朴素贝叶斯分类方法根据一个中文人名中的用字特征来判定该人名的性别。首先对中文人名语料中男女不同性别名字的用字特征进行统计分析,然后解析朴素贝叶斯分类方法如何实现中文人名的性别判定,以及实现过程中的两个关键问题:(1)中文人名的表征——区分特征组合;(2)性别判定的依据——条件概率求解。

2.1 中文人名用字的性别区分性

中文人名根据用字多少,可分为单字名、双字名、三字名、三字以上名。统计发现,中文人名以双字名为主,单字名次之,三字名及以上的极其少见。中文人名中传承着浓厚的文化内涵,人名用字具有较强的性别区分性,透过人名便可知其是男性或女

性。在中文人名中,男性以刚健有力为美,命名注重品格、事业前途,取名时希望他们像山一样屹立(如多用山、峰等字);像金石一样经得起磨炼(如多用鑫、磊、刚等字);取“成、功、栋、伟、建”等字则希望能建功立业,成就一番事业。女性则渴望有花容月貌般的容颜,柔情似水般的性情,美玉般的肌肤等,所以,女性命名中多用“梅、桂、芳、兰、洁、雅、娟、娇、姣、珠、珍、玉”等字。

为了从“量”上对男女人名用字有清晰的认识,本研究首先对 412 775 个中文人名语料数据进行了统计分析,以字为单位,分别统计了人名语料中男性、女性人名用字的字种数(即不同汉字的数量)及出现频次,为论述方便,下面论述中记人名中的第 1、第 2 个字分别为字$_1$、字$_2$,统计时,单字名认为字$_1$为空格,字$_2$为实际的单字名。统计结果显示,本研究实验所用到的中文人名语料中,男性人名中字$_1$的字种共有 2 113 个,字$_2$的字种共有 2 456 个,女性人名中字$_1$的字种共有 1 900 个,字$_2$的字种共有 2 039个。表 1 给出了统计结果中用字频次排在前 30 位的字及频次。由表 1 可知,中文男女人名用字大不相同,男性人名和女性人名用字频次前 30 的字大部分不同。其中,男女字$_1$的前30 个字中有 13 个字相同(包括空格也计算在内),而字$_2$的前 30 个字中只有 2 个字相同,字$_1$ 字$_2$ 双字名的前 30 个名字中没有一个相同。这些统计结果说明: 中文人名中男性女性用字有较强的性别区分性,且字$_2$ 较字$_1$ 更有区分性。另外,从表 1 中字$_1$ 为空格的频次可知语料中共用 21 212 个男性单字人名和 20 713 个女性单字人名。

表 1　中文人名中男女不同性别的用字统计

Table 1　Statistical character in different gender Chinese names

序号	男性		女性		男性		女性		男性		女性	
	字$_1$	频次	字$_1$	频次	字$_2$	频次	字$_2$	频次	双字名	频次	双字名	频次
1		21 212		20 713	军	3 988	英	7 160	志强	211	秀英	372
2	国	5 004	玉	5 589	伟	3 919	霞	5 305	鹏飞	202	桂兰	345
3	文	4 455	秀	5 458	林	3 856	梅	5 278	建军	187	秀云	281
4	建	4 161	小	3 717	杰	3 578	玲	5 147	建国	185	玉梅	277
5	志	3 712	爱	3 683	峰	3 466	兰	5 007	海涛	184	秀梅	269
6	永	3 678	桂	3 393	生	3 355	荣	4 945	胜利	170	桂英	268
7	金	3 579	晓	3 226	明	3 346	丽	4 252	国强	164	桂荣	266
8	新	3 285	艳	2 816	民	2 936	芳	3 863	志刚	157	秀荣	260
9	玉	3 204	素	2 799	华	2 674	红	3 739	建伟	156	玉兰	256
10	明	3 104	淑	2 759	强	2 604	云	3 715	志伟	148	秀兰	255
11	振	3 034	春	2 730	成	2 559	花	3 483	建华	145	兰英	232
12	海	2 747	凤	2 400	龙	2 477	华	3 434	红军	143	秀花	230
13	德	2 435	金	2 208	涛	2 381	珍	3 027	永强	136	红梅	226
14	长	2 378	丽	2 135	亮	2 290	琴	3 019	春生	133	秀珍	221
15	保	2 365	秋	2 024	辉	2 285	芝	2 702	新民	130	凤英	221
16	小	2 362	文	1 996	山	2 192	萍	2 648	海军	129	艳丽	220
17	学	2 140	红	1 980	平	2 000	莲	2 577	红伟	127	爱荣	217
18	红	2 128	瑞	1 793	超	1 986	敏	2 524	建峰	124	秀琴	212
19	广	2 111	新	1 776	祥	1 948	枝	2 403	朝阳	124	桂花	211
20	俊	2 075	雪	1 634	文	1 890	平	2 344	俊杰	121	春玲	203
21	书	2 070	俊	1 596	飞	1 886	娟	2 200	文龙	119	玉珍	203
22	世	1 922	亚	1 577	义	1 847	香	2 162	向阳	118	玉荣	200
23	天	1 811	翠	1 562	阳	1 844	凤	1 931	振宇	115	彩霞	194
24	晓	1 744	海	1 467	东	1 834	菊	1 569	新建	114	红霞	192
25	东	1 740	书	1 454	海	1 808	娜	1 506	光辉	113	丽娟	192
26	庆	1 731	喜	1 437	安	1 743	芬	1 503	晓东	112	玉霞	189
27	中	1 706	慧	1 406	国	1 640	妮	1 468	国安	110	丽娜	187
28	亚	1 686	美	1 349	豪	1 596	娥	1 464	卫东	109	春霞	187
29	子	1 661	梦	1 340	鹏	1 580	燕	1 216	云飞	108	玉玲	186
30	春	1 554	彩	1 283	良	1 546	艳	1 200	家豪	107	秀芝	183

2.2 朴素贝叶斯方法实现中文人名的性别判定

对于利用朴素贝叶斯分类方法进行中文人名性别判定问题,可以表述为,给定一些训练样本(x,y),其中 x 表示名字,y 表示性别,首先根据这些已知的样本构建一个能够对实际问题进行准确描述的统计模型 $P(y|x)$,该模型可以用来预测未知人名的性别。构建模型时,首先要解决名字 x 的表征问题,即用哪些区分特征来表征名字 x 便于进行性别判定。然后是性别判定的依据,即条件概率 $P(y|x)$ 的求解问题。

2.2.1 中文人名的表征

对一个中文的姓名,设字$_1$ 为名字中的第 1 个字。字$_2$ 为名字中的第 2 个字。字$_1$ 字$_2$ 就是名字中第 1、第 2 字的组合,对双字名来说就是整个名字。利用朴素贝叶斯分类方法对中文人名性别判定建模中首先要解决人名的表征问题,即设定人名特征来表征该名字。根据 2.1 节的分析,中文人名中的用字特征有较强的性别区分性。本研究利用朴素贝叶斯分类方法对中文人名性别进行判定时,可以根据名字中的用字特征来判定性别,也就是根据 1 个中文名字中不同位置出现的字来判定性别。所以,本研究选取字$_1$、字$_2$、字$_1$ 字$_2$ 作为区分中文人名性别的特征,这 3 个区分特征的任意组合可以表征中文人名,例如"字$_1$ + 字$_2$ + 字$_1$ 字$_2$"是全部 3 个区分特征的组合,这里及后文中的"+"表示多个区分特征组合之意。

为了对这些特征和特征组合对性别判定的作用有个"量"的认识,本研究设计了对比实验。表 2 列出了对比实验中用到的 6 个特征组合。

表 2 特征组合列表

Table 2 List of feature combinations

序号	特征组合及包含的区分特征
1	字$_1$
2	字$_2$
3	字$_1$ + 字$_2$
4	字$_1$ + 字$_1$ 字$_2$
5	字$_2$ + 字$_1$ 字$_2$
6	字$_1$ + 字$_2$ + 字$_1$ 字$_2$

2.2.2 性别判定的依据——条件概率

根据上面的分析,从中文人名判定性别,本质是根据给定的名字 x,求解条件概率 $P(y|x)$。其中 y 的取值有两个{男性,女性}。例如,判定一个姓名为"李秀丽"的性别,需要分别求出 $P(y=\text{男}|x=\text{秀丽})$,$P(y=\text{女}|x=\text{秀丽})$,取条件概率值大的作为该人名对应的性别即可。

然而,给定名字 x,直接根据条件概率的定义式(4)来求解 $P(y|x)$ 难度较大,

$$P(y|x)=\frac{P(y,x)}{P(x)}。\tag{4}$$

根据贝叶斯公式,可以转换为先验概率 $P(y)$ 和 $P(x|y)$(y 分别为男性、女性)的求解。这里假设表征中文人名的特征组合中的各特征相互独立,根据式(2)可得:

$$P(x|y)=\prod_{k=1}^{n}P(x_k|y_i)。\tag{5}$$

在对朴素贝叶斯分类器训练时,可以从中文人名训练语料中用最大似然估计得到以上这些先验概率值。求解公式如下:

$$P(y\text{ 为男性})=\frac{\text{人名语料中男性个数}}{\text{训练语料中人名个数}},\tag{6}$$

$$P(x_i|y\text{ 为男性})=\frac{\text{男性人名中该特征频次数}}{\text{人名语料中男性人名个数}}。\tag{7}$$

女性相关先验概率的求解类似,在此不再赘述。另外由于训练语料有限和数据稀疏现象大量存在,所以,对这些先验概率参数需要进行数据平滑处理,研究采用 Good-Turing(古德-图灵)估计进行平滑处理。

3 实验与结果分析

3.1 实验数据集

本研究采用的训练语料和测试语料全部是真实的中文人名数据,取自某家省级医院一年内的所有病患的姓名和性别。剔除极少数不可使用的人名数据后,共有 412 775 个中文人名数据。表 3 给出了该人名语料数据的一些统计信息。

表 3 中文人名语料相关统计信息

Table 3 Correlation statistics of the Chinese name corpus

	总个数	单字人名	双字人名	其他
男性人名	237 656	21 212	216 417	27
女性人名	175 119	20 713	154 384	22

为了取得更好的、更可信的实验结果,在已有的人名语料上进行训练测试时,采用 10 重交叉验证(10-fold cross validation)的方法来训练贝叶斯分类器和进行测试。即人名语料被划分为 10 个不相交的组,每次拿出 9 组作为训练集用于训练模型参数,对分类器训练 10 次,每次留出一组作为测试集进行测试。

3.2 性能评估

对中文人名性别判定进行评估时,采用的评估指标是:判定准确率。判定准确率表示在对测试集中全部人名进行的性别判定中,性别判定正确的所

占比率。计算公式如下:

$$\text{判定准确率} = \frac{\text{性别判定正确的人名数}}{\text{测试数据中的人名总数}} \times 100\% 。 \quad (8)$$

3.3　实验及其结果分析

为了验证本研究提出方法的判定性能,设计了相关实验,对比了不同特征组合对性别判定性能的影响,实验关注的是不同特征组合的判定准确率的差异。使用不同特征组合分别在 10 个测试集上进行了测试,依次将这些测试编号为 1 ~ 10,表 4 给出了 10 组测试的判定结果。

表 4　不同特征组合的判定准确率

Table 4　Determination results of different feature combinations　%

测试序号	$字_1$			$字_2$			$字_1+字_2$			$字_1+字_1\ 字_2$			$字_2+字_1\ 字_2$			$字_1+字_2+字_1\ 字_2$		
	男性	女性	综合	男性	女性	综合	男性	女性	综合	男性	女性	综合	男性	女性	综合	男性	女性	综合
1	88.79	51.11	71.95	90.07	83.45	87.11	91.94	85.16	88.91	89.22	82.10	86.04	90.53	87.03	88.97	91.45	87.85	89.84
2	88.97	50.91	72.69	89.97	82.76	86.89	91.77	84.80	88.79	89.76	82.46	86.63	90.71	86.38	88.85	91.38	87.12	89.56
3	89.08	51.15	72.57	90.20	83.03	87.08	92.05	85.09	89.02	90.80	84.03	87.86	91.48	87.55	89.77	92.09	88.36	90.46
4	88.98	50.34	72.51	90.35	82.40	86.96	91.99	84.69	88.88	91.93	85.44	89.16	92.21	87.71	90.29	92.66	88.32	90.81
5	89.08	50.33	73.27	89.94	81.51	86.51	91.83	83.94	88.61	91.97	86.21	89.62	92.10	87.08	90.05	92.52	87.93	90.65
6	89.12	50.31	73.49	90.26	82.05	86.96	92.06	84.02	88.82	91.53	85.16	88.97	92.08	87.07	90.06	92.54	87.79	90.63
7	88.60	50.39	73.24	89.31	82.22	86.46	90.98	84.20	88.26	89.38	83.04	86.83	90.13	86.46	88.65	90.70	87.19	89.29
8	89.06	50.92	73.05	90.00	82.65	86.92	91.89	84.63	88.84	89.19	81.44	85.94	90.44	86.27	88.69	91.21	87.11	89.49
9	89.21	50.78	72.68	90.16	83.84	87.44	92.08	85.37	89.20	89.47	82.13	86.31	90.31	87.61	89.15	91.32	88.13	89.95
10	88.96	50.48	72.06	89.64	83.43	86.91	92.03	85.21	89.03	89.50	82.18	86.29	90.21	87.42	88.98	91.41	87.98	89.91
平均	88.99	50.67	72.75	89.99	82.73	86.92	91.86	84.71	88.84	90.28	83.42	87.37	91.02	87.06	89.35	91.73	87.78	90.06

综合分析表 4 中的数据可以得出如下结论: (1) 中文人名中,$字_2$ 较 $字_1$ 更具有性别区分性,单独 $字_2$ 特征的性别判定准确率达到 86.92%,而 $字_1$ 的判定准确率只有 72.75%,相差 14.17%。这也验证了表 1 中的统计结果,$字_2$ 较 $字_1$ 更具有性别区分性。(2) 除了只有 $字_1$ 特征的性别判定准确率差别较大外,其余 5 组特征向量的判定准确率比较接近,差别在 4% 以内。性别判定准确率从低到高依次为 $字_1$,$字_2$,$字_1+字_1\ 字_2$,$字_1+字_2$,$字_2+字_1\ 字_2$,$字_1+字_2+字_1\ 字_2$,最好的是包含全部区分特征的特征组合。(3) 对所有特征组合,男性人名的判定准确率都高于女性的准确率,特别是 $字_1$ 特征组合的差别更大,这点也可以从图 2 直观地看到。

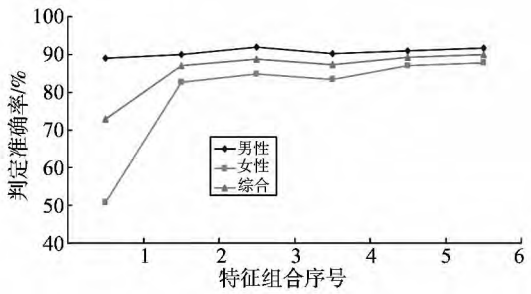

图 2　男性女性判定结果的比较

Fig. 2　Determination results comparison for male and female

4　结语

中文人名具有较强的性别区分性,本研究通过对大量中文人名的统计分析,对男女不同性别的中文人名用字特征有了更清晰的认识,在此基础上,提出一种基于中文人名用字特征的性别判定方法,并利用朴素贝叶斯分类方法设计开发了一套程序实现了中文人名的性别判定。在 412 775 个中文人名数据上的实验结果表明,该方法简单可行。该方法的提出与实现将对中文信息深层次处理具有重要意义。中文人名除了具有较强的性别区分性之外,还具有时代、地域、家庭等特性,下一步将获取更多的中文人名语料,分析中文人名和出生年代、地域、家庭等的相关性。

参考文献:

[1] 郑淑花. 汉语人名用字的统计分析 [J]. 皖西学院学报, 2010, 16(1): 113-116.
ZHENG Shuhua. A statistical analysis for the character used in Chinese names [J]. Journal of West Anhui University, 2010, 16(1): 113-116.

[2] 李益德,李学杰. 汉语人名用字管窥 [J]. 周口师范学

院学报,2005,22(1):114-117.
LI Yide, LI Xuejie. Research on character used in Chinese names [J]. Journal of Zhoukou Normal University, 2005, 22(1):114-117.

[3] 侯一麟. 汉语人名音、意、字说略[J]. 清华大学学报:哲学社会科学版,1995,10(1):84-86.
HOU Yilin. Study and review on the sound, meaning, character of Chinese names [J]. Journal of Tsinghua University: Philosophy and Social Sciences, 1995, 10(1):84-86.

[4] 段新和. 论文化角度下汉语人名的修辞意蕴[J]. 湖北工业大学学报,2009,24(6): 115-117.
DUAN Xinhe. On the rhetorical implication of Chinese names from the cultural perspective [J]. Journal of Hubei University of Technology, 2009, 24(6):115-117.

[5] 于芳. 汉语人名研究述评[J]. 南平师专学报,2006,25(3):76-78.
YU Fang. Study and review on the Chinese names [J]. Journal of Nanping Teachers College, 2006, 25(3):76-78.

[6] 钱进. 姓名用字的性别差异统计分析[J]. 常州工学院学报,2004,17(5):60-62.
QIAN Jin. Statistical analysis of gender difference in names [J]. Journal of Changzhou Institute of Technology, 2004, 17(5):60-62.

[7] 程玉虎,仝瑶瑶,王雪松. 类相关性影响可变选择性贝叶斯分类器[J]. 电子学报,2011,39(7):1628-1633.
CHENG Yuhu, TONG Yaoyao, WANG Xuesong. A selective Bayesian classifier based on change of class relevance influence [J]. Acta Electronic Sinica, 2011, 39(7):1628-1633.

[8] 王双成,杜瑞杰,刘颖. 连续属性完全贝叶斯分类器的学习与优化[J]. 计算机学报,2012,35(10):2129-2138.
WANG Shuangcheng, DU Ruijie, LIU Ying. The learning and optimization of full Bayes classifiers with continuous attributes [J]. Chinese Journal of Computers, 2012, 35(10):2129-2138.

[9] BARBARA F, PICTERS I, LINDA C, et al. When learning naive Bayesian classifiers preserves monotonicity [J]. Lecture Notes in Computer Science, 2011(6717): 422-433.

[10] DENIS F, MAGNAN C N. Efficient learning of naive Bayes classifier under class-conditional classification noise [C]//ACM International Conference Proceeding Series, Proceedings of the 23rd International Conference on Machine Learning. Pennsylvania: [s. n.], 2006: 265-272.

[11] 蔺志青,郭军. 贝叶斯分类器在手写汉字识别中的应用[J]. 电子学报,2002,30(12):1-4.
LIN Zhiqing, GUO Jun. An application of Bayesian classifier in the recognition of handwritten Chinese character [J]. Acta Electronic Sinica, 2002, 30(12):1-4.

[12] PENG H, LONG F, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and minredundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8):1227-1238.

[13] 饶丽丽,刘雄辉,张东站. 基于特征相关的改进加权朴素贝叶斯分类算法[J]. 厦门大学学报:自然科学版,2012,51(4):682-685.
RAO Lili, LIU Xionghui, ZHANG Dongzhan. An improved weighted naïve Bayes classification algorithm using feature correlation [J]. Journal of Xiamen University: Natural Science, 2012, 51(4):682-685.

[14] 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.
LUAN Fang, YUAN Junying. Naïve Bayes Chinese text classification based on core words of class [J]. Journal of Shandong University: Natural Science, 2006, 41(3): 46-49.

[15] 周国强,崔荣一. 基于朴素贝叶斯分类器的朝鲜语文本分类的研究[J]. 中文信息学报,2011,25(4): 16-19.
ZHOU Guoqiang, CUI Rongyi. Research on Korean text categorization based on naive Bayesian classifier [J]. Journal of Chinese Information Processing, 2011, 25(4):16-19.

[16] WEBB G I, BOUGHTON J R, WANG Z. Not so naive Bayes: aggregating one-dependence estimators [J]. Machine Learning, 2005, 58(1):5-24.

[17] ZHOU X, LIU K Y, WONG S C. Cancer classification and prediction using logistic regression with Bayesian gene selection [J]. Journal of Biomedical Informatics, 2004, 37(4):249-259.

[18] 宫秀军,刘少辉,史忠植. 一种增量贝叶斯分类模型[J]. 计算机学报,2002,25(6):646-650.
GONG Xiujun, LIU Shaohui, SHI Zhongzhi. An incremental Bayes classification model [J]. Chinese Journal of Computers, 2002, 25(6):646-650.

[19] 吕昊,林君,曾晓献. 改进朴素贝叶斯分类算法的研究与应用[J]. 湖南大学学报:自然科学版,2012,39(12):56-61.
LÜ Hao, LIN Jun, ZENG Xiaoxian. Research and application of improved naive Bayesian classification algorithm [J]. Journal of Hunan University: Natural Sciences, 2012, 39(12):56-61.

[20] 王中锋,王志海. 基于条件对数似然函数导数的贝叶斯网络分类器优化算法[J]. 计算机学报,2012,35(2):364-374. (下转第23页)

1426-1428.

[7] HADLEY G R. Multistep method for wide-angle beam propagation [J]. Optics Letters, 1992, 17(24): 1743-1745.

[8] 张华. 三维全矢量广角束传播方法的研究 [M]. 长春: 吉林大学,2003.

[9] AGRAWAL A, SHARMA A. Recent wide-angle beam propagation methods: an assessment [J]. Lasers and Electro-Optics, 2005, 11: 969-970.

[10] MASOUDI H M, AKOND M S. Stable time-domain beam propagation method for modeling ultrashort pulse propagation in dispersive optical structures [J]. IEEE Photonics Technology Letters, 2012, 24: 769-771.

[11] MASOUDI H M, AKOND M S. Efficient iterative time-domain beam propagation methods for ultra short pulse propagation: analysis and assessment [J]. Lightwave Technology, 2011, 29: 2475-2481.

[12] MASOUDI H M. A novel nonparaxial time-domain beam-propagation method for modeling ultrashort pulses in optical structures [J]. Lightwave Technology, 2007, 25: 3175-3184.

[13] BEKKER V, SEWELL P, BENSON T M, et al. A wide-angle alternating-direction implicit finite-difference beam propagation method [J]. Lightwave Technology, 2009, 27: 2595-2604.

[14] BEKKER V, SEWELL P, BENSON T M, et al. A wide-angle alternating-direction implicit finite-difference BPM [C] //Transparent Optical Networks, ICTON 9th International Conference. [S. l]: [s. n.], 2007: 250-253.

[15] LE K Q, BIENSTMAN P. Three-dimensional higher-order Padé approximant-based wide-angle beam propagation method using complex Jacobi iteration [J]. Electronics Letters, 2010, 46: 241-242.

[16] LU Y Y, HO P L. Beam propagation method using a $[p-1)/p]$ Padé approximant of the propagator [J]. Optics Letters, 2002, 27(9): 683-685.

[17] CHUI S L, LU Y Y. A propagator-θ beam propagation method [J]. IEEE Photonics Technology Letters, 2004, 16(3): 822-824.

[18] LU Y Y. Some techniques for computing wave propagation in optical waveguides [J] Communications in Computational Physics, 2006, 1(6): 1056-1075.

[19] YEVICK D, THOMSON D. Complex Padé approximants for wide-angle acoustic propagators [J]. Joural of The Acoustical Socienty of America, 2000, 108(6): 2784-2790.

[20] BURDEN Richard L. 数值分析 [M]. 7 版. 北京: 高等教育出版社, 2001.

[21] BAKER G A. Padé Approximants, part I: basic theory [M]. Massachusetts: Addison-Wesley Publishing Company, 1981.

[22] BAKER G A. Essentials of Padé approximants [M]. New York: Academic Press Inc, 1975.

[23] 宋贵才. 光波导原理与器件 [M]. 北京: 清华大学出版社,2012.

[24] HO P L, LU Y Y. Improving the beam propagation method for TM polarization [J]. Optical and Quantum Electronics, 2003, 35(4): 507-519.

[25] HO P L, LU Y Y. A mode-preserving perfectly matched layer for optical waveguides [J]. IEEE Photonics Technology Letters, 2003, 15(9): 1234-1236.

(编辑: 胡春霞)

(上接第 18 页)

WANG Zhongfeng, WANG Zhihai. An optimization algorithm of Bayesian network classifiers by derivatives of conditional log likelihood [J]. Chinese Journal of Computers, 2012, 35(2): 364-374.

[21] 张鹏,唐世渭. 朴素贝叶斯分类中的隐私保护方法研究 [J]. 计算机学报, 2007, 30(8): 1267-1276.
ZHANG Peng, TANG Shiwei. Privacy preserving naive Bayes classification [J]. Chinese Journal of Computers, 2007, 30(8): 1267-1276.

[22] YANG Y, WEBB G I. Weighted proportional k-interval discretization for naive-Bayes classifiers [J]. Lecture Notes in Computer Science, 2003(2637): 501-512.

[23] HSU Chungchian, HUANG Yanping. Extended naive Bayes classifier for mixed data [J]. Expert Systems with Applications, 2008, 35(3): 1080-1083.

[24] MERLER S, CAPRILE B. Parallelizing AdaBoost by weights dynamics [J]. Computational Statistics & Data Analysis, 2007, 51(5): 2487-2498.

[25] 邓维斌,王国胤,王燕. 基于 Rough Set 的加权朴素贝叶斯分类算法 [J]. 计算机科学,2007,34(2): 204-206.
DENG Weibin, WANG Guoyin, WANG Yan. Weighted naive Bayes classification algorithm based on rough set [J]. Journal of Computer Science, 2007, 34(2): 204-206.

[26] 石洪波, 王志海. 一种限定性的双层贝叶斯分类模型 [J]. 软件学报,2004,15(2): 193-200.
SHI Hongbo, WANG Zhihai. A restricted double-level bayesian classification model [J]. Journal of Software, 2004, 15(2): 193-200.

(编辑: 胡春霞)