基于 Rough Set 的加权朴素贝叶斯分类算法*)

邓维斌[1,2] 王国胤[1] 王 燕[1,2]
(重庆邮电大学计算机科学与技术研究所 重庆 400065)
(西南交通大学信息科学与技术学院 成都 610031)[2]

摘 要 朴素贝叶斯算法是一种简单而高效的分类算法,但其条件独立性假设并不符合客观实际,这在某种程度上影响了它的分类性能。加权朴素贝叶斯是对它的一种扩展。基于 Rough Set 的属性重要性理论,提出了基于 Rough Set 的加权朴素贝叶斯分类方法,并分别从代数观、信息观及综合代数观和信息观的角度给出了属性权值的求解方法。通过在 UCI 数据集上的仿真实验,验证了该方法的有效性。

关键词 朴素贝叶斯,加权朴素贝叶斯,Rough 集,属性重要性,分类

Weighted Naïve Bayes Classification Algorithm Based on Rough Set

DENG Wei-Bin[1,2] WANG Guo-Yin[1] WANG Yan[1,2]
(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065)[1]
(School of Information Science and Technology, Southweast Jiaotong University, Chengdu 610031)[2]

Abstract Naïve Bayes algorithm is an effective simple classification algorithm. Since its conditional independence assumption is not always true in real life, its classification performance is affected to some extent. Weighted naïve Bayes (simply WNB) is an extension of it. Based on the attributes' importance degree theory of rough set, a new weighted naïve Bayes method is proposed. Methods for determining the weights of attributes in the algebra view, informational view and both of them are developed respectively. Simulation results on a variety of UCI data sets illustrate the efficiency of this method.

Keywords Naïve bayes, Weighted naïve bayes, Rough set, Weightiness of attribute, Classification

1 引言

分类是数据挖掘中一个非常重要的问题,在商业中也应用很多。分类的目的是提出一个分类函数或分类模型(常称为分类器)。该模型能把数据库中的数据项映射到给定类别中的某一类。通过分析训练数据样本,产生关于类别的精确描述,可以用来对未知的数据进行分类预测。分类算法的核心部分是构造分类器。在众多分类方法和理论中,朴素贝叶斯(Naïve Bayes,简称 NB)由于计算高效、精确度高,并具有坚实的理论基础而得到了广泛应用[1]。朴素贝叶斯分类基于一个简单的假定:在给定分类特征条件下属性值之间是相互独立的。在现实世界中,这种独立性假设经常是不满足的。因此,针对朴素贝叶斯分类的不足之处,许多学者研究学习贝叶斯网络(Bayes Network)[2~4]来改进其分类性能,然而文[5]已证明:要学习得到一个最优贝叶斯网络是个 NP-hard 问题。如何能既保持朴素贝叶斯计算的简单性,又可以提高其分类性能呢? Zhang Harry 在文[6]中提出了根据属性的重要性给不同属性赋不同权值的加权朴素贝叶斯(Weighted Naïve Bayes,简称 WNB)模型,给出了采用爬山算法和 Monte Carlo 技术确定权值的加权朴素贝叶斯分类方法,并通过实验发现能改进朴素贝叶斯的分类效果[6]。

粗糙集(Rough Set,简称 RS)理论[7]由波兰逻辑学家 Pawlak 教授于 1982 年提出,由于它能有效处理不精确、不一致及不完备信息,近年来在机器学习、数据挖掘、人工神经网络等方面得到了广泛应用研究[8]。本文主要根据 Rough Set 中属性重要性理论,分别从代数观和信息观的角度计算属性的重要性[9,10],并据此确定加权朴素贝叶斯中不同属性的权。从而根据数据本身的特点,提高朴素贝叶斯分类质量。

2 朴素贝叶斯分类与加权朴素贝叶斯分类

2.1 朴素贝叶斯模型

贝叶斯分类是基于贝叶斯公式,即:$P(C|X)=\frac{P(X|C)\times P(C)}{P(X)}$。其中:$P(C|X)$为条件 X 下 C 的后验概率,$P(C)$为 C 的先验概率,$P(X|C)$为条件 C 下 X 的后验概率,$P(X)$表示 X 的先验概率。

为叙述方便,对符号作如下约定:用大写字母表示变量,C 表示类别变量,A 表示属性变量,假定共有 m 个属性变量,$A=\langle A_1,A_2,\cdots,A_m\rangle$;用小写字母表示变量取值,$Val(C)=\{c_1,c_2,\cdots,c_l\}$,$Val(A)=\{a_{i1},a_{i2},\cdots,a_{ik}\}$分别表示类别变量和属性变量的值域;用 X 表示待分样本集,$x=\langle a_1,a_2,\cdots,a_m\rangle$表示待分类样本;用 T 表示训练样本集,$t_i=\langle a_1,a_2,\cdots,a_m,c_l\rangle$表示训练实例。在朴素贝叶斯中假设各属性相对于类别条件独立,则有:

*)本课题得到国家自然科学基金(No. 60373111)、新世纪优秀人才支持计划、重庆市教委科学技术研究项目(040505)资助。邓维斌 硕士生,主要研究领域为智能信息系统;王国胤 博士,教授,博导,主要研究领域包括 Rough 集理论、神经网络、机器学习、数据挖掘等;王 燕 博士生,主要研究领域为智能信息处理和数据挖掘。

$$P(a_1,a_2,\cdots,a_m|c_l)=\prod_{i=1}^{m}P(a_i|c_l)$$

从而后验概率公式为：$P(c_l|x)=\frac{P(c_l)}{P(x)}\prod_{i=1}^{m}P(a_i|c_l)$，测试样本(E)被分在后验概率最大的类中，由于 $P(x)$ 为一常数，则朴素贝叶斯分类模型为[1]：

$$V_{nb}(E)=\underset{c}{\operatorname{argmax}}\ P(c)\prod_{i=1}^{m}P(a_i|c)$$

2.2 加权朴素贝叶斯模型及分类过程

(1)加权朴素贝叶斯模型

由于在实际中比较难于满足朴素贝叶斯条件独立性的假设，可给不同的属性赋不同的权值使朴素贝叶斯得以扩展，则加权朴素贝叶斯模型为[6]：$V_{wnd}(E)=\underset{c}{\operatorname{argmax}}\ P(c)\prod_{i=1}^{m}p(a_i|c)^{w_i}$，其中 w_i 代表属性 A_i 的权值，属性的权值越大，该属性对分类的影响就越大。加权朴素贝叶斯的关键问题就在于如何确定不同属性的权值。

(2)加权朴素贝叶斯的分类过程

Step1 数据预处理：将训练样本和待分类样本进行补齐和离散化；

Step2 判断：如果是分类任务，则转 Step5，如果是训练任务则转 Step3；

Step3 概率表及权值参数的学习：

Step3.1 概率参数学习：扫描所有训练样本，计算所有的先验概率 $P(a_{ik}|c_l)$，即在类别 c_l 中属性 A_i 的第 k 种取值的概率；以及 $P(c_l)$，即取值为类别 c_l 的概率；

Step3.2 权值参数学习：扫描所有训练样本，计算属性 A_i 的权值 w_i；

Step4 生成加权朴素贝叶斯概率表及属性权值列表，即所需的加权朴素贝叶斯分类器；

Step5 分类：调用概率表及属性权值列表，得出分类结果。

3 属性重要性的权值求解

3.1 代数观下属性重要性的权值求解

定义 1(属性重要性的代数观点定义)[9] 设 $T=(U,R,V,f)$ 是一个决策表系统，其中 $R=C\cup D$，C 是条件属性集合，$D=\{d\}$ 是决策属性集合，且 $A\subset C$，F 是属性集 D 导出的分类，则对任意属性 $a\in C-A$ 的重要性 $SGF(a,A,D)$ 定义为：

$$SGF(a,A,D)=r_{A\cup\{a\}}(F)-r_A(F)$$

其中 $r_A(F)$ 表示属性子集 A 对集合簇 F 的近似分类质量。$SGF(a,A,D)$ 的值越大，说明在已知 A 的条件下，属性 a 对于决策 D 就越重要[8]。这里，若取 $A=\phi$，则分别求出每个属性 A_i 对决策属性 D 的重要性表示为 $SGF(a_i,D)$。在朴素贝叶斯分类算法中，每个属性的重要性相等，权值均为 1，加权朴素贝叶斯属性的权值应根据其重要性与所有属性重要性的数学期望相比较进行重新分配。则代数观下属性权值可定义为：

定义 2(代数观下属性权值定义) 设 $SGF(a_i,D)$ 表示属性 A_i 在代数观下相对于决策属性 D 的重要性，共有 m 个条件属性，则属性 A_i 的权值为：

$$w_i=\frac{SGF(a_i,D)}{\frac{1}{m}\sum_{i=1}^{m}SGF(a_i,D)}$$

3.2 信息观下属性重要性的权值求解

定义 3(属性重要性的信息论观点定义)[9] 设 $T=(U,R,V,f)$ 是一个决策表系统，其中 $R=C\cup D$，C 是条件属性集合，$D=\{d\}$ 是决策属性集合，且 $A\subset C$，则对任意属性 $a\in C-A$ 的重要性 $SGF(a,A,D)$ 定义为：

$$SGF(a,A,D)=H(D|A)-H(D|A\cup\{a\})$$

其中 $H(D|A)$ 表示属性集 D 相对于属性集 A 的条件熵。若取 $A=\phi$，则 $SGF(a,A,D)=H(D|A)-H(D|\{a\})$，称为条件属性 a 和决策 D 的互信息，记为 $I(a,D)$[8]。则：

$$I(a,D)=\sum_{a_i\in a}\sum_{d_j\in d}P(a_i,d_j)\log_2\frac{P(a_i,d_j)}{P(a_i)P(d_j)}$$

$I(a,D)$ 的值越大，说明属性 a 对于决策 D 就越重要[9]。与代数观下权值计算类似，可根据互信息求信息观下属性重要性的权值：

定义 4(信息观下属性权值定义) 设 $I(a_i,D)$ 表示条件属性 A_i 与决策属性 D 的互信息，共有 m 个条件属性，则属性 A_i 的权值为：

$$w_i=\frac{I(a_i,D)}{\frac{1}{m}\sum_{i=1}^{m}I(a_i,D)}$$

3.3 综合代数观和信息观属性重要性的权值求解

文[9]指出：属性重要性的代数定义和信息定义具有互补的特性，即：属性重要性的代数定义考虑的是该属性对论域中确定分类子集的影响，而信息定义考虑的是该属性对于论域中不确定分类子集的影响。同时证明了属性重要性在代数观和信息观下并非具有一致性。

为了综合考虑属性对论域中确定分类子集和不确定分类子集的影响，现综合代数观和信息观的属性重要性，将其属性权值定义为：

定义 5(综合代数观和信息观的属性权值定义) 设 w_{1i} 和 w_{2i} 分别代表条件属性 A_i 在代数观和信息观下的属性重要性，可得属性 A_i 在综合代数观和信息观下的权值 w_i 为：

$$w_i=\frac{w_{1i}+w_{2i}}{2}$$

3.4 属性权值计算举例

如表 1，一个决策表有条件属性：Outlook，Tem(temperature)，Hum(humidity)，Windy，决策属性 d：

表 1 一个决策表

U	Outlook(a_1)	Tem(a_2)	Hum(a_3)	Windy(a_4)	d
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Sunny	Mild	High	False	N
8	Sunny	Mild	Normal	False	N
9	Sunny	Mild	Normal	True	P
10	Rain	Mild	Normal	False	P
11	Overcast	Hot	High	False	N
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

根据表 1，按以上三种方法计算的属性权值如表 2。

4 仿真实验

为了验证前述基于 Rough Set 的加权朴素贝叶斯方法的

效果，选用了 UCI 机器学习数据库(ftp://ftp.ics.uci.edu/pub/machine-learning-databases)中的 15 个数据集作为测试数据，这些数据集的基本信息如表 3。实验的硬件环境为：赛扬 1.7G CPU，256M 内存；软件环境为：Windows 2000 系统，SQL Server 2000 数据库。

表 2 对表 1 中条件属性重要性的权值计算表

	Outlook(a_1)	Tem(a_2)	Hum(a_3)	Windy(a_4)
方法 1	2.0	0.0	0.2857	1.7143
方法 2	2.2108	0.7148	0.9069	0.1673
方法 3	2.1054	0.3574	0.5963	0.9408

注：方法 1，2，3 分别为：代数观、信息观及综合代数观和信息观的权值计算方法。

表 3 所用测试数据集的描述

编号	数据集	条件属性数	决策类数	样例数	有无缺省数据	有无连续值属性
1	Balance-scale	4	3	625	N	N
2	Ballon-Adult-Stretch	4	2	20	N	N
3	Breast-Cancer-Wisconsin	10	2	699	Y	N
4	Car-Evaluation	6	4	1728	N	N
5	Iris	4	3	150	N	Y
6	Lung-Cancer	56	3	32	Y	N
7	Mushroom	22	2	8124	Y	N
8	Pima-Indians-Diabetes	8	2	768	N	Y
9	House-Votes-84	16	2	435	N	N
10	Zoo	18	7	101	N	N
11	Glass	10	7	214	N	Y
12	Wine	13	3	178	N	Y
13	Hepatitis	19	2	155	Y	Y
14	Letter	16	26	20000	N	Y
15	Chess(KR-VS-KP)	36	2	3196	N	N

测试步骤为：

Step1 用重庆邮电大学计算科学技术研究所研发的“基于 Rough Set 的智能数据分析系统(RIDAS)[11]”对数据集进行预处理(用“条件组合补齐算法”进行补齐，用“基于属性重要性”的方法进行离散化)；

Step2 如果某个属性的取值个数与样本数相等时，它对分类的作用很小，于是先将这样的属性去掉[12]。在所用的 15 个数据集中有 3 个这样的属性：Breast-Cancer 中的 Sample Code Number 属性，Zoo 中的 Animal Name 属性和 Glass 中的 Id Number 属性；

Step3 将数据集平均分成 5 份，首先用其中 4 份作为学习数据，学习得到贝叶斯概率表及属性权值。之后调用加权朴素贝叶斯公式对余下的 20%样本进行测试，每个数据集轮流以上实验 5 次，取 5 次结果的平均值作为这个数据集的测试结果。分别用本文介绍的算法和文[6]提出的爬山算法确定权值的加权朴素贝叶斯(从文[6]知：采用 Monte Carlo 技术与爬山算法的计算过程和结果基本相同，故本文不再与之比较)进行测试，几种方法的测试正确率和训练权值的时间如表 4(accuracy 表示测试正确率，time 表示训练得到属性权值的时间)。

取 0.95 的置信度对几种方法的实验结果用双尾 t-检测进行比较[15]，结果如表 5。其中 w-t-l 分别表示当前行所在方法比当前列所在方法性能优、相当于列的数据集个数，如第一行第一列中的“5-10-0”表示 AWNB 分类性能比 NB 方法优的数据集有 5 个、性能相当的有 10 个、性能不足的有 0 个。

表 5 几种方法的测试结果比较

		NB	AWNB	IWNB	SWNB
测试正确率比较	AWNB	5-10-0			
	IWNB	7-8-0	4-9-2		
	SWNB	8-7-0	6-9-0	5-10-0	
	HCWNB	7-8-0	4-10-1	3-11-1	1-12-1
		AWNB	IWNB	SWNB	
计算权值时间比较	IWNB	3-2-10			
	SWNB	0-0-15	0-4-11		
	HCWNB	0-1-14	0-6-9	0-8-7	

通过对几种方法测试结果的比较分析，可以得出以下结论：

(1)代数观、信息观以及综合代数观和信息观的加权朴素贝叶斯均能较好地改进朴素贝叶斯的分类效果；

(2)当条件属性数和分类数较少时，基于代数观的加权朴素贝叶斯效果较好；

(3)当条件属性数较多时，基于信息观的加权朴素贝叶斯效果较好；

(4)综合代数观和信息观的方法平均识别正确率较基于代数观和信息观高，与爬山算法相当；

(5)就训练间来看，本文所用方法均比爬山算法优。

结论及展望 虽然朴素贝叶斯分类方法由于其时间复杂度低、精度较高，在分类算法中得到较大重视，但其条件独立性假设在很大程度上限制了它的使用。本文从 Rough Set 理论中的属性重要性出发，提出了以属性重要性作为权值的加权朴素贝叶斯分类方法，并分别从代数观、信息观及综合代数观和信息观的角度探讨了权值求解方法。最后以 UCI 中的 15 个数据集为测试数据，通过实验比较了朴素贝叶斯分类、爬山算法求解权值的加权朴素贝叶斯和基于 Rough Set 的加权朴素贝叶斯分类的分类效果。实验表明：本文所提出的方法可以根据数据本身的特点在较少的训练时间下提高朴素贝叶斯的分类精度。其中代数观方法适用于条件属性和分类数较少的数据集；信息观方法适用于条件属性较多的数据集；综合代数观和信息观的方法平均识别率较前两者高。本文中的算法在时间效率上均比文[6]中的爬山算法和 Mente Carlo 方法优。

由于在实验过程中进行了数据的补齐和离散化的预处理，然而在对原始信息系统进行预处理的同时或多或少地改变了原始信息，如何直接对具有遗失值和连续值属性的信息系统进行加权朴素贝叶斯分类将是我们下一步研究的方向。

致谢 对在实验过程中鄢羽同学给予的大量帮助表示感谢。

参考文献

1 史忠植. 知识发现. 北京：清华大学出版社，2002.169～220
2 Cooper G. Computational complexity of probablitisitic inference using Bayesian belief networks. Artificial Intelligence，1990，42：393～405
3 Friedman N，GeigerD，Goldszmidt M. Bayesian network Classifiers. Maching Learnig，1997，29：131～163
4 王双成，苑森淼. 具有丢失数据的贝叶斯网络结构学习研究. 软件学报，2004，15(7)：1042～1048
5 Chickering D M. Learning Bayesian networks is NP-complete[A]. Learningfrom Data：AI and Statistics V[M]. New York：Springer，1996.121～130

(下转第 219 页)

和 i3 上的运行时间之比在 10%以内。

结论 本文把布尔函数中变量熵的概念进行扩展并应用于 OBDD 变量排序算法的设计。熵作为特征值的优点是能够较好地反映变量的全局特征，有利于获得稳定的算法性能。基于熵的变量排序算法通过选择性地搜索更有希望获得较好结果的变量序空间降低了算法运行时间，提高了算法的性能。实验表明：与模拟退火算法和遗传算法相比，基于熵算法的时间效率明显改善，在获得相当结果的情况下，时间仅为相应算法的 80.84%和 29.79%。进一步的工作包括寻找更为系统的变量序特征分析方法，并在此基础上设计更为有效的变量排序算法。

参 考 文 献

1 Bryant R E. Graph-based algorithms for Boolean function manipulation[J]. IEEE Transactions on Computers, 1996, 35(8): 677～691
2 Bollig B, Wegener I. Improving the variable ordering of OBDDs is NP-complete[J]. IEEE Transactions on Computers, 1996, 45(9): 933～1002
3 Friedman S J, Supowit K J. Finding the optimal variable ordering for binary decision diagrams[J]. IEEE Transactions on Computers, 1990, 39(5): 710～713
4 Ebendt R, Günther W, Drechsler R. An Improved Branch and Bound Algorithm for Exact BDD Minimization[J]. IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, 2003, 22(12): 1657～1663
5 Ebendt R, Günther W, Drechsler R. Combining ordered best-first search with branch and bound for exact BDD minimization [C]. ASP-DAC, 2004. 875～878
6 Malik S, Wang A R, Brayton R K, Sangiovanni-Vincentelli A. Logic verification using binary decision diagrams in a logic synthesis Environment[A]. In: Proceedings of International Conference on Computer-aided Design[C], 1988. 6～8
7 龙望宁，杨士元，闵应骅，等. 基于重量分析的 OBDD 变量排序算法[J]. 计算机学报，1997，20(8)：702～710
8 贝劲松，边计年，薛宏熙，等. OBDD 变量排序的自适应选择算法. 计算机辅助设计与图形学学报[J]，1999，11(5)：412～416
9 Thornton M A, Williams J P, Drechsler R, et al. SBDD Variable Reordering Based on Probabilistic and Evolutionary Algorithms [A]. In: Proceedings of IEEE Pacific rim conference[C], 1999. 381～387
10 Jain J, Moundanos D, Bitner J, et al. Efficient variable ordering and partial representation algorithms[A]. In: Proceedings of the 8th International Conference on VLSI Design[C], 1995. 81～86
11 Fujita M, Matunaga Y, Kakuda T. On the variable ordering of binary decision diagrams for the application of multi-level logic synthesis[A]. In: Proceedings of European Design automation Conference[C], 1991. 50～54
12 Rudell R. Dynamic variable ordering for ordered binary decision diagrams[A]. In: Proceedings of International Conference on Computer-aided Design[C], 1993. 42～47
13 Bollig B, Löbbing M, Wegener I. Simulated annealing to improve variable orderings for OBDDs[A]. In: Proceedings of International Workshop on Logic Synthesis[C], 1995
14 Drechsler R, Becker B, Gockel N. A genetic algorithm for variable ordering of OBDDs[J]. IEE Proceedings of Computers and Digital Techniques, 1996, 143(6): 364～368
15 Collaborative Benchmarking Laboratory. 1993 LGSynth Benchmarks. North Carolina State University, Department of Computer Science, 1993
16 Somenzi F. CUDD: CU Decision Diagram Package, Release 2.3.1, 2001

(上接第 206 页)

6 Harry Z, Sheng S. Learning Weighted Naïve Bayes with Accurate Ranking. In: Fourth IEEE International Conference on Data Mining (ICDM'04), 2004. 567～570
7 Pawlak Z. Rough set. International Journal of Computer and Information Sciences, 1982, 11(5): 341～356
8 王国胤. Rough 集理论与知识获取. 西安：西安交通大学出版社，2001. 1～147
9 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简. 计算机学报，2002，25(7)：759～766
10 Wang Guoyin, Zhao Jun, An Jiujiang, Wu Yu. A Comparative Study of Algebra Viewpoint and Information Viewpoint in Attribute Reduction, Fundamenta Informaticae, 2005, 68(3): 289～301
11 Wang G Y, Zheng Z, Wu Y. RIDAS—A Rough Set Based Intelligent Data Analysis System. In: First IEEE International Conference On Machine Learning and Cybernitics(ICMLC2002), 2002
12 Jiang Liangxiao, Zhang Huajie, Cai Zhi hua, Su Jiang. Evolution Naïve Bayes. In: Proceedings of International Symposium on Intelligent Computation and its Application (ISICA 2005), 2005. 344～350
13 Huang Jin, Lu Jingjing, Ling C X. Comparing Naïve Bayes, Decision Tree, and SVM with AUC and Accuracy. In: IEEE International Conference on Data Mining(ICDM2003), 2003. 553～556
14 Huang Jin, Ling C X. Using AUC and Accuracy in Evaluting Learning Algorithms. In: Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(3): 299～310
15 Nadeau C, Bengio Y. Inference for the generalization error. In Advances in Neural Information Processing Systems MIT Press, 1999, 12: 307～313

表 4 实验结果-测试正确率及训练时间表

编号	数据集	NB	AWNB		IWNB		SWNB		HCWNB	
		accuracy(%)	accuracy(%)	time(s)	accuracy(%)	time(s)	accuracy(%)	time(s)	accuracy(%)	time(s)
1	Balance	77.12±2.82	83.12±1.12	1.15±0.07	76.72±3.60	1.32±0.06	81.88±4.43	1.40±0.14	80.72±2.05	1.37±0.10
2	Ballon	90.0±13.69	75.0±17.68	0.12±0.01	80.0±11.18	0.21±0.07	80.0±11.18	0.33±0.07	90.0±13.69	0.25±0.02
3	Breast-Cancer	92.44±2.58	85.0±3.51	0.61±0.05	95.86±2.90	1.03±0.04	95.78±2.32	1.49±0.33	95.96±2.81	1.57±0.15
4	Car	80.5±4.84	88.74±6.55	3.26±0.09	87.7±4.44	3.36±0.14	87.76±5.38	3.66±0.21	84.92±6.04	4.19±0.59
5	Iris	92.32±6.69	94.86±1.43	0.15±0.06	91.32±0.86	0.24±0.03	95.32±3.98	0.33±0.02	94.72±4.19	0.32±0.01
6	Lung-Cancer	40.0±10.46	45.0±6.85	3.19±0.07	55.0±6.85	3.50±0.45	55.0±6.85	5.54±0.28	47.5±5.59	7.35±0.17
7	Mushroom	69.46±4.16	71.76±18.33	38.30±0.42	78.8±1.94	25.81±1.17	80.14±3.15	48.80±0.71	83.64±6.07	53.66±8.67
8	Pima-Indians	72.04±1.99	74.87±2.04	0.41±0.02	74.76±3.46	0.64±0.15	76.26±2.85	0.74±0.13	74.72±1.14	0.75±0.04
9	House-Vote	49.66±7.95	52.64±6.35	1.58±0.02	46.44±4.91	1.55±0.14	53.7±5.04	2.24±0.42	50.32±8.07	2.11±0.85
10	Zoo	87.2±7.53	85.0±3.54	0.65±0.02	92.0±5.70	1.47±0.24	89.6±5.32	1.62±0.17	90.0±7.07	1.54±0.01
11	Glass	64.52±8.75	56.58±7.60	0.63±0.01	67.78±2.01	1.48±0.23	65.84±2.20	1.56±0.20	66.52±6.45	1.62±0.01
12	Wine	81.88±5.89	85.88±1.86	0.75±0.01	86.14±0.74	1.08±0.19	88.62±1.47	1.25±0.11	85.16±4.34	1.14±0.03
13	Hepatitis	80.5±2.67	83.96±3.48	1.01±0.01	84.35±1.10	1.21±0.23	84.26±1.04	1.43±0.29	84.48±1.01	1.79±0.05
14	Letter	63.06±0.52	64.63±2.43	33.10±1.59	65.43±1.69	30.92±1.48	68.83±2.25	58.74±0.52	66.27±4.17	89.43±2.56
15	Chess	70.64±9.10	74.58±3.77	31.62±0.77	75.14±3.97	9.94±0.48	76.4±2.88	34.11±1.09	80.28±2.23	88.48±5.47
	平均值	75.42	76.11	7.77	78.5	5.58	79.96	10.88	79.88	17.04

注：AWNB, IWNB, SWNB, HCWNB 分别表示代数观、信息观、综合代数观和信息观及爬山算法加权朴素贝叶斯方法。