基于属性选择的改进加权朴素贝叶斯分类算法①

王行甫, 杜 婷

(中国科学技术大学 计算机学院, 合肥 230022)

摘 要: 朴素贝叶斯分类算法简单且高效, 但其基于属性间强独立性的假设限制了其应用范围. 针对这一问题, 提出一种基于属性选择的改进加权朴素贝叶斯分类算法(ASWNBC). 该算法将基于相关的属性选择算法(CFS)和加权朴素贝叶斯分类算法(WNBC)相结合, 首先使用CFS算法获得属性子集使简化后的属性集尽量满足条件独立性, 同时根据不同属性取值对分类结果影响的不同设计新权重作为算法的加权系数, 最后使用 ASWNBC 算法进行分类. 实验结果表明, 该算法在降低分类消耗时间的同时提高了分类准确率, 有效地提高了朴素贝叶斯分类算法的性能.

关键词: 属性选择; 朴素贝叶斯分类; 权重; 相关性; 关联性

Improved Weighted Naive Bayes Classification Algorithm Based on Attribute Selection

WANG Xing-Fu, DU Ting

(School of Computer Science, University of Science and Technology of China, Hefei 230022, China)

Abstract: Naive Bayes Classification is simple and effective, but its strong attribute independency assumption limits its application scope. Concerning this problem, an improved WNBC algorithm is proposed based on attribute selection. This algorithm combines CFS algorithm with WNBC algorithm, it firstly uses CFS algorithm to get an attribute subset so that the simplified attribute subset can meet conditional independency; meanwhile, the algorithm's weighting coefficient is designed on that different attribute values have different influences on the classification result. Finally, it uses ASWNBC algorithm to classify datasets. The experimental results show that the proposed algorithm improves the classification accuracy with lower time consumption, therefore heightens the performance of NBC algorithm.

Key words: attribute selection; naive Bayes classification (NBC); weight; dependency; relevance

朴素贝叶斯分类算法发源于古典数学理论, 有着坚实的理论基础, 其理论基础贝叶斯理论是由英国数学家 Thomas Bayes 在前人知识积累的基础上首次归纳总结出来的一个数学理论体系. 与其他分类算法相比, 朴素贝叶斯算法[1](Naive Bayes Classification NBC)是目前公认的一种相对简单且高效的分类算法, 因其稳定的分类效率被广泛应用于自然语言处理、机器学习、机器人导航、模式识别等领域.

NBC 算法是一种基于概率的分类方法, 该算法假设一个属性对给定类的影响独立于其他属性. 理论上, 当满足此假设时, NBC 算法与其他分类算法相比具有最小的误差率. 但此假设在实际应用中往往无法满足, 为提高分类精度, 研究人员提出多种加权朴素贝叶斯分类算法[2](WNBC). 如陈朝大等[3]提出的利用关联规则改进 NBC 算法, 通过关联规则的置信度给朴素贝叶斯加权; 张步良[4]提出的基于分类概率的 NBC, 使用朴素贝叶斯分类成功的概率作为加权系数; Jie Lin 等[5]提出的基于粒子群算法的 WNBC 算法, 通过粒子群算法的自动搜索功能对现有数据和信息进行学习, 以数据集中所有数据各自权重的平均值作为加权系数. 相比较 NBC 算法, 上述加权朴素贝叶斯分类算法在一定程度上均提高了分类准确率, 但由于增加了加权系数

① 基金项目:国家科技重大专项(2012ZX10004-301-609);国家自然科学基金(61272472,61232018,61202404);安徽省教学研究计划 2010
收稿时间:2014-12-02;收到修改稿时间:2015-01-26

的计算过程，使得分类消耗时间相应提高.

为了在降低分类消耗时间的同时提高分类准确率，本文将基于相关的属性选择算法[6](CFS)和加权朴素贝叶斯分类算法(WNBC)进行结合，提出一种基于属性选择的改进加权朴素贝叶斯算法(ASWNBC). 使用 CFS 算法进行属性选择得到良好的属性子集，然后在属性子集上进行分类，有效地降低了分类消耗时间. 另外，根据属性不同取值对分类结果影响不同设计合理的权值作为算法的加权系数，提高了分类准确率. 对 UCI 上 10 个数据集进行分类的结果表明，本文提出的 ASWNBC 算法有效地结合了 CFS 算法的简化性和 WNBC 算法的高效性，在降低分类消耗时间的同时提高了分类准确率.

1 朴素贝叶斯分类法

贝叶斯分类法的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，具有最大后验概率的类则为该对象所属的类. NBC 算法是在贝叶斯分类法的基础上提出的，该算法满足一个简单的假定，即在给定目标值时属性值之间相互条件独立.

NBC 算法[1]的工作过程如下:

1) 设 A 表示训练样本的属性集，有 n 个属性 $A_1, A_2, \ldots, A_n$; C 表示类集合，有 m 个类 $C_1, C_2, \ldots, C_m$. 每个数据样本 X 用一个 n 维特征向量来描述 n 个属性的值，即: $X=\{x_1, x_2, \ldots, x_n\}$，其中$x_i \in A(1 \le i \le n)$. 确定数据的特征属性，获得训练样本.

2) 对训练样本集进行统计，计算得到每个特征属性在各类别的条件概率估计即:

$$\begin{aligned} &P(A_1|C_1), P(A_2|C_1), \cdots, P(A_n|C_1); \\ &P(A_1|C_2), P(A_2|C_2), \cdots, P(A_n|C_2); \\ &\vdots \\ &P(A_1|C_m), P(A_2|C_m), \cdots, P(A_n|C_m); \end{aligned}$$

3) 对每个类别计算后验概率，取最大后验概率项作为样本所属类别，即若某个样本 X 利用朴素贝叶斯分类法判断属于类 C_i，当且仅当满足条件:

$$P(C_i|X) > P(C_j|X) \quad 1 \le j \le m, j \ne i$$

根据贝叶斯定理可知:

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

NBC 算法中 P(X)对每个类别均可视为常数，当类别的先验概率未知时，常假定类别的概率相等即 $P(C_1) = P(C_2) = \cdots = P(C_m)$，因此，将判别 $P(C_i \mid X)$ 最大转化为判别 $P(X \mid C_i)$最大即可.

在 NBC 算法中假定各条件属性相互独立，可以得出:

$$P(X|C_i) = \prod_{j=1}^{n} P(x_1|C_i) \times \cdots \times P(x_n|C_i) \tag{2}$$

所以，样本 X 属于 C 中某个类别需满足:

$$c(X) = \text{argmax} P(C_i) \prod_{j=1}^{n} P(x_j|C_i) \quad (1 \le j \le n, 1 \le i \le m) \tag{3}$$

根据公式(3)，NBC 算法即可将样本对象划分到后验概率最大的类别中，从而完成分类.

从 NBC 算法的工作过程可知，整个工作过程简单高效. 但在许多实际问题中，NBC 算法的独立性假设并不成立，若忽视这一点，得到的分类结果往往不准确.

2 现有的加权朴素贝叶斯分类算法

2.1 加权朴素贝叶斯模型

NBC 算法基于条件独立性假设，认为每个属性对类属性影响相同，但事实并非如此，有些属性对分类影响大而有些属性对分类影响较小. 如果把与分类无关的、冗余的以及被噪声污染的属性和其他属性视为同等地位，将会导致分类的准确率下降. 为提高 NBC 算法的准确率扩大其适用范围，人们将各种属性加权算法和 NBC 算法相结合，根据各属性对分类影响的大小赋予不同的权重，将 NBC 算法扩展为加权朴素贝叶斯算法(WNBC). 其中，加权朴素贝叶斯分类算法的模型[2]大多为:

$$c(X) = argmax P(C_i) \prod_{j=1}^{n} P(x_j|C_i)\, W_{A_k,C_i}$$

$$(1 \le j \le n, 1 \le i \le m) \tag{4}$$

公式(4)中，W_{A_k,C_i} 代表属性 A_k 的权值. 属性的权值越大表明该属性对分类的影响越大.

2.2 属性加权算法

Harry 等[2]提出了加权朴素贝叶斯分类模型，该模型根据条件属性对决策所起的作用赋给它们不同的权重. 在此基础上，研究人员提出了多种属性权值计算方法，比如文献[2]中分别采用爬山算法、信息增益和蒙特卡罗技术来确定属性的权值.

近几年的研究文献中，研究人员又提出了基于相关系数的加权朴素贝叶斯分类模型. 文献[7]基于相关系数的加权朴素贝叶斯分类算法[7](WNB-CC)，根据条件属性和决策属性之间的相关程度越高条件属性对分类的重要性应越大的原理，使用相关系数公式(5)计算第 i 个条件属性的权重系数.

$$W_{A_i} = |\rho(X_i, C)| = \left|\frac{\mathrm{cov}(X_i, C)}{\sqrt{D(X_i)D(C)}}\right| \quad (5)$$

文献[7]中得出 WNB-CC 算法在平均性能上较 NBC 算法有所提高，但是相关系数是对两个随机变量相关程度的度量，现用的度量公式并不能针对所有的分布情况都准确地描述其相关性.

另外，在粗糙集理论的基础上，研究人员开始对基于粗糙集的加权朴素贝叶斯分类算法进行研究，如文献[8]中提出了一种基于属性序约简的加权朴素贝叶斯算法(WNB-AOD)[8]. 首先使用基于属性序和分治法的属性约简算法进行属性约简，得到约简规则建立新的决策表，然后构建加权朴素贝叶斯模型进行分类. 从文献[8]中可知，WNB-AOD 算法在一定程度上提高了分类准确率，但是其建立新决策表的过程较为复杂，导致算法的分类消耗时间较大.

3 基于属性选择的改进加权朴素贝叶斯算法

3.1 基于相关的属性选择算法

3.1.1 属性选择

在数据集的属性中，有些属性对分类影响较大，而有些属性对分类影响很小. 用属性关联度表示一个属性和类属性间的相关性，它反映这个属性对分类结果影响的程度；用属性冗余度表示一个属性和其他属性之间相关性，它反映这个属性和其他属性间的依赖度[9]. 对数据集进行属性选择，主要希望得到一个属性子集，使得属性子集中的属性和类属性总体相关性较大，属性间的冗余度较小.

3.1.2 CFS 算法

CFS 算法是基于相关的属性选择算法，它基于相关的启发式属性评估函数对属性进行排序选择，是一个简单过滤算法. 该算法从测度属性间的简单相关开始，然后基于复合检验设计原理[6]构造启发式属性评估函数，最后将此函数用于评估和选择属性以及属性子集.

基于类似于统计检验理论复合检验的设计原理，CFS 算法中生成一个能评价属性子集分类能力的评估函数[10]:

$$\mathrm{Merit}_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (6)$$

在公式(6)中, Merit_s 是一个包含 k 个属性的属性子集 S 的启发式评估值，$\overline{r_{cf}}$表示 S 中属性和类属性间相关测度 r_{cf} 的均值，$\overline{r_{ff}}$表示 S 中属性间交互相关性的均值. 式中的分子可视为属性和类属性间关联性大小的度量，分母可视为子集属性间冗余性大小的度量.

上述评估函数在设计上从属性和类属性间关联的强度和属性间相互影响导致的冗余性两个方面进行考虑，因此该评估函数具有与类属性高度相关而与其他属性彼此不相关或相关度低的特性，可以很好的剔除不相关的属性，得到良好的属性子集，进而为分类算法提供可靠高效的数据子集.

3.2 加权系数的改进

常用 WNBC 算法对每个属性进行加权时，通常没有考虑到各属性的不同取值对分类的影响. 如程克非等[11]提出的一种特征加权算法，对各属性的每个取值采取相同的加权值，并没有体现出各属性的不同取值对分类的影响. 为获得准确的加权系数，本文根据属性不同取值对分类结果影响的不同，采用一种新的权值计算方法.

设T_{A_k}表示属性A_k的取值个数，$T(A_k = v)$表示属性A_k取值为 v 的样本对象个数，$T_{(A_k=v\cap C_i)}$表示属性 A_k 取值为 v 且属于类 C_i 的样本对象个数. 根据各属性的不同取值对分类的影响设计权值，加权系数公式表示为:

$$W_{A_k,C_i} = \frac{T_{A_k} + \frac{T_{(A_k=v\cap C_i)}}{T_{(A_k=v)}}}{T_{A_k}} \quad (7)$$

公式(7)根据每个属性的不同取值对分类的影响计算权值，W_{A_k,C_i}不仅反映了每个属性对分类的影响程度，而且量化了每个属性及其类别之间的关联. 因此，将W_{A_k,C_i}作为 WNBC 算法的加权系数更加合理准确，有助于提高分类准确率.

3.3 ASWNBC 算法

在实际数据集中常存在冗余数据、不相关数据等，若使用数据集中的全部属性进行分类，将会增加算法的复杂性，使训练的分类过度拟合，降低分类效率[12]. 为克服这一问题，提出一种基于属性选择的 WNBC 算

法(ASWNBC), 即在使用 WNBC 算法进行分类前, 对属性集进行属性选择.

ASWNBC 算法结合了属性选择的简化性和加权朴素贝叶斯的高效性. 一方面对数据的冗余属性进行基于属性相关的属性选择处理, 得到良好的属性子集, 降低算法的时间和空间复杂度; 另一方面根据属性取值不同对分类结果影响的不同, 量化数据的关联属性作为加权系数, 提高分类的准确率. 该算法的具体流程如下:

1) 预处理数据

对数据集进行规范化、离散化处理, 对缺失数据进行填充, 使得到的样本数据具有足够的信息.

2) 获得属性子集

采用 CFS 算法对数据的属性集进行选择, 用 Best First 搜索策略[13]探测整个搜索空间. 如果连续 5 次扩展子集, 使用属性评估函数 Merit$_s$ 得到的属性子集评估值均没有改进, 搜索停止并获得属性子集.

3) 计算加权系数

根据每个属性取值的不同对分类的影响程度, 按照公式(7)计算权值W_{A_k,C_i}作为 WNBC 算法的加权系数.

4) 构建加权朴素贝叶斯模型

统计数据集里的样例生成模型的网络结构进行参数学习, 并根据加权系数W_{A_k,C_i}构建加权朴素贝叶斯模型.

5) 获得分类结果

使用加权朴素贝叶斯模型在获得的属性子集上进行训练, 输出最终的分类结果.

4 实验结果和分析

为了验证算法的有效性, 对算法进行实验测试和评估. 实验使用的数据集来自 UCI 数据库, 使用的实验平台为: 硬件环境 CPU 4.0 GHz, 软件环境 Windows7, 实验工具 Matlab7.0.

实验过程中, 采用从 UCI 上选择的 10 个数据集, 为提高实验结果的可靠性, 对数据集中数据的排列顺序进行打乱, 并对每个数据集轮流实验 5 次, 取 5 次实验的平均结果作为最终测试结果.

首先根据文献[1,7,8], 分别使用 NBC 算法、WNB-CC 算法、WNB-AOD 算法对 10 个数据集进行训练分类. 此时并未对数据集的属性进行选择, 三种算法在数据集全部属性的前提下进行分类的准确率和消耗时间如表 1 所示. 其次根据本文提出的 ASWNBC 算法流程, 首先对 10 个数据集中属性值为连续性的进行离散化处理; 然后, 使用 CFS 算法对数据的属性集进行选择得到属性子集, 属性选择结果如表 2 所示; 最后, 使用 ASWNBC 算法对数据集进行训练分类, 分类的准确率和消耗时间如表 3 所示. 根据表 1 和表 3 的分类结果, 将 ASWNBC 算法和 NBC 算法、WNB-CC 算法、WNB-AOD 算法的分类消耗时间和分类准确率进行对比并分析实验结果.

表 1 NBC、WNB-CC 和 WNB-AOD 算法分类结果

数据集	样例个数	全部属性	类别	NBC		WNB-CC		WNB-AOD	
				准确率(%)	时间(ms)	准确率(%)	时间(ms)	准确率(%)	时间(ms)
German	1000	15	2	75.50	105	76.20	157	77.52	182
Autos	205	25	1	56.20	54	54.87	72	55.23	103
Satimage	6435	37	6	82.30	259	84.40	302	85.32	412
Waveform21	5000	19	2	79.72	184	81.93	268	80.85	310
Mushroom	8124	22	1	95.84	288	96.54	357	97.52	429
Vehicle	951	18	4	88.03	92	90.75	138	89.21	170
Segment	2310	18	7	90.06	133	91.16	184	92.76	225
Vote	450	16	2	89.37	82	90.02	111	90.95	138
Cleveland	403	13	2	75.73	63	75.05	92	76.25	85
Lymphography	148	19	4	84.17	79	82.05	88	79.15	82

表 2 CFS 算法属性选择结果

数据集	样例个数	全部属性个数	类别	选择属性个数
German	1000	15	2	7
Autos	205	25	1	5
Satimage	6435	37	6	14
Waveform21	5000	19	2	10
Mushroom	8124	22	1	4
Vehicle	951	18	4	9
Segment	2310	18	7	12
Vote	450	16	2	5
Cleveland	403	13	2	6
Lymphography	148	19	4	10

表 3 ASWNBC 算法分类结果

数据集	样例个数	选择属性	类别	准确率(%)	时间(ms)
German	1000	7	2	77.03	75
Autos	205	5	1	55.12	40
Satimage	6435	14	6	84.25	128
Waveform21	5000	10	2	81.94	119
Mushroom	8124	4	1	96.72	167
Vehicle	951	9	4	90.34	71
Segment	2310	12	7	93.45	83
Vote	450	5	2	92.12	61
Cleveland	403	6	2	77.62	47
Lymphography	148	10	4	80.41	55

从表 1 中分类结果可知, 相比较 NBC 算法, 直观上可看出 WNB-CC 算法、WNB-AOD 算法提高了大部分数据集的分类准确率, 但同时也增加了分类消耗时间. 从表 2 属性选择结果可知, ASWNBC 算法中使用 CFS 算法进行属性选择, 去除了与类属性关联性低但和其他属性冗余度高的属性, 使 10 个数据集选择后的属性数均少于原来的属性数, 得到良好的属性子集, ASWNBC 算法在属性子集上进行训练分类得到分类结果表 3. 对照表 1 和表 3 的分类结果可知, 相比较 NBC 算法、WNB-CC 算法、WNB-AOD 算法, ASWNBC 算法分类结果中各数据集分类消耗时间较低, 大部分数据集的分类准确率有所提高.

为了更加准确地分析实验结果, 根据表 1 和表 3 对 10 个数据集的分类准确率和消耗时间求平均值, 统计获得平均分类准确率和消耗时间, 对照结果如表 4 所示. 从表 4 的对照结果可看出, 与 NBC 算法相比, ASWNBC 算法的分类准确率平均提高了约 1.21%, 分类消耗时间平均降低了约 49.3ms. 相比 WNB-CC 算法和 WNB-AOD 算法, ASWNBC 算法的分类准确率平均提高了约 0.5%, 分类消耗时间平均降低了约 110ms, 分类准确率有所提高并且在很大程度上降低了分类消耗时间. 对上述几种算法的分类准确率和分类消耗时间进行综合比较可知, ASWNBC 算法具有良好的性能, 可以有效地提高分类准确率, 降低分类消耗时间.

表 4 平均准确率和消耗时间对照表

使用算法	平均准确率(%)	平均消耗时间(ms)
NBC	81.69	133.9
WNB-CC	82.30	176.9
WNB-AOD	82.48	213.6
ASWNBC	82.90	84.6

5 结语

本文提出了一种基于属性选择的加权朴素贝叶斯算法(ASWNBC), 在算法中使用 CFS 算法进行属性选择获得良好的属性子集, 降低了属性间的冗余度. 同时, 根据属性不同取值对分类结果的不同影响计算权值, 使 WNBC 算法的加权系数更加合理. 对 UCI 数据库中的 10 个数据集进行训练分类的结果表明, 本文算法具有良好的性能, 可以有效地提高分类准确率, 降低分类消耗时间. 但是, 不同数据集属性之间的冗余度和属性与类属性间的关联度不尽相同, 当使用 CFS 算法进行属性选择时, 若去除的属性和类属性间的关联度较大, 此时将得到效果较差的属性子集, 从而导致分类准确率下降. 下一步将重点研究属性选择算法, 使选择后的属性间尽量满足条件独立性假设, 进一步提高算法性能.

参考文献

1 Jiawei H, Kamber M. Data mining: concepts and techniques . San Francisco, CA. itd: Morgan Kaufmann. 2001, 5.

2 Zhang H, Sheng S. Learning weighted naive Bayes with accurate ranking. 2004 Fourth IEEE International Conference on.Data Mining(ICDM'04). IEEE. 2004. 567–570.

3 陈朝大,梁柱勋,郑士基.一种利用关联规则的改进朴素贝叶斯分类算法.计算机系统应用,2010,19(11):106–109.

4 张步良.基于分类概率加权的朴素贝叶斯分类方法.重庆理工大学学报(自然科学版),2012,26(7):81–83.

5 Lin J, Yu J. Weighted naive bayes classification algorithm based on particle swarm optimization. 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN). IEEE. 2011. 444–447.

6 Hall MA. Correlation-based Feature Selection for Machine Learning[Thesis]. The University of Waikato, 1999.

7 张明卫,王波,张斌,等.基于相关系数的加权朴素贝叶斯分类算法.东北大学学报(自然科学版),2008,29(7):952–955.

8 Zhou X, Luo K. A weighted naive Bayes algorithm based on the attribute order reduction. Advanced Materials Research, 2013, 718: 2108–2112.

9 魏浩,丁要军.一种基于相关的属性选择改进算法.计算机应用与软件,2014,31(8).

10 AJD Ślęzak. Rough set methods for attribute clustering and selection. Applied Artificial Intelligence, 2014, 28(3): 220–242.

11 程克非,张聪.基于特征加权的朴素贝叶斯分类器.计算机仿真,2006,23(10):92–94.

12 Zhang S, McCullagh P, Callaghan V. An efficient feature selection method for activity classification. 2014 International Conference on Intelligent Environments (IE). IEEE. 2014. 16–22.

13 Kishimoto A, Fukunaga A, Botea A. Evaluation of a simple, scalable, parallel best-first search strategy. Artificial Intelligence, 2013, 195: 222–248.