一种属性相关性的加权贝叶斯分类算法研究

郑 默,刘琼荪

(重庆大学 数理学院,重庆 400030)

摘 要:根据 Rough Set 属性重要度理论,构建了基于互信息的属性子集重要度,提出属性相关性的加权朴素贝叶斯分类算法,该算法同时放宽了朴素贝叶斯算法属性独立性、属性重要性相同的假设。通过在 UCI 部分数据集上进行仿真实验,与基于属性相关性分析的贝叶斯(CB)和加权朴素贝叶斯(WNB)两种算法做比较,证明了该算法的有效性。

关键词:朴素贝叶斯;属性重要度;属性相关;分类

中图分类号:TP301.6 **文献标识码**:A **文章编号**:1674-7720(2011)07-0096-03

Weighted naive Bayesian classification based on attribute correlation

Zheng Mo, Liu Qiongsun

(College of Mathematics and Physics, Chongqing University, Chongqing 400030, China)

Abstract: Based on the theory of rough set, a new naive Bayes method named mutual information-based algorithm for weighted naive Bayes (WCB) was proposed, which synchronously loosen naive Bayes classifier's independence and equal importance of the attribute assumptions. compared with correlated Bayes (CB) and weighted naive Bayes (WNB), simulation results on a variety of UCI data sets illustrate the efficiency of this method.

Key words: naive Bayes; weightiness of attribute; attribute correlation; classification

分类是数据挖掘中一类非常重要的问题,分类算法的核心是构造能快速、有效处理大数据容量、高精确度的分类器。在众多分类算法和理论中,朴素贝叶斯 NB(Naive Bayes)由于计算高效、高精确度,并具有坚实的理论基础而得到了广泛应用。NB 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具有独立性,且每个条件属性对类变量(决策属性)的重要度是相同的。然而,在实际问题中,这些假设往往不能满足。为了保持 NB 的计算既简单,又能提高其分类性能,参考文献[1]提出了一种基于属性相关性分析的贝叶斯分类模型 CB(Correlated Bayes),放宽了属性独立性的假设,当属性间存在相关性时较好地提高了分类性能,但是该模型假定每个属性相对于决策属性重要性相同,当属性相对于决策属性的重要性不相同时,分类效果并没有提高;参考文献[2-4]中提出了根据属性的重要性赋予属性权值的加权朴素贝叶斯 WNB(Weighted Naive Bayes)模型,允许属性之间重要度不相同,较之 NB 模型获得较好的分类效果,但该模型仍基于属性类条件独立假设,当属性间存在相关性时分类效果并不好。综上,上述方法均只侧重改进 NB 方法的某单一假设,并未同时放宽两个假设,现实中数据也常常不能同时满足两个假设。

本文在 CB 模型和 WNB 模型的基础上,以互信息作为度量条件属性相对于决策属性的重要度,提出了集合重要度的概念,并赋予各属性子集权值,同时考虑属性子集内部属性间的相关性,提出属性相关性的加权贝叶斯分类算法(WCB),以达到提高 NB 的分类性能的目的。

1 朴素贝叶斯分类相关模型

1.1 朴素贝叶斯(NB)模型[1]

假设数据集有 p 个属性 $A_1, A_2, \cdots, A_p$;m 个类别 $C_1, C_2, \cdots, C_m$;待分类样本 $X=(x_1, x_2, \cdots, x_p)$,其中 x_i 表示第 i 个属性指标,即 $x_i \in A_i$。判断 X 属于类别 C_i 的概率可由贝叶斯公式计算:$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)} \propto P(X|C_i)P(C_i)$。

假设各属性相对于类别条件独立,即 $P(X|C_i)=\prod_{j=1}^{p} P(x_j|C_i)$。

故 NB 分类模型为:

$$V_{NB}(X)=\arg\max_i P(C_i)=\prod_{j=1}^{p}P(x_j|C_i) \quad (1)$$

1.2 基于属性相关性分析的贝叶斯(CB)模型[1]

对于待分类样本 $X=(x_1,x_2,\cdots,x_p)$，$P(X|C_i)$ 的计算公式为：

$$P(X|C_i)=Corr_x\prod_{j=1}^{p}P(x_j|C_i)$$

其中 $Corr_x$ 用下面公式估计[1]：

$$Corr_x=\sqrt[\beta]{(C_p^2)\prod_{i,j=1}^{p}Corr_{x_i,x_j}},(i<j) \quad (2)$$

故基于属性相关性的 CB 模型为：

$$V_{CB}(X)=\arg\max_i Corr_x P(C_i)\prod_{j=1}^{p}P(x_j|C_i) \quad (3)$$

1.3 加权朴素贝叶斯(WNB)模型[2]

不同的属性根据其分类重要性赋不同的权值，即 WNB 模型为：

$$V_{WNB}(X)=\arg\max_i P(C_i)\prod_{j=1}^{p}P(x_j|C_i)^{w_j} \quad (4)$$

其中，w_j 表示属性 A_j 的权值，属性的权值越大，该属性对分类的影响就越大。

CB 和 WNB 模型针对不同侧面对 NB 模型进行了改进。本文根据粗糙集属性重要度理论，提出了属性集合重要度的概念；基于 CB 与 WNB 模型的分类思想，构造了一种基于属性相关性的 WCB 模型。

2 属性相关性的加权贝叶斯分类(WCB)模型

2.1 属性集合重要度

定义 1(条件熵[3])设条件属性 $A_i=(a_{i1},\cdots,a_{in})$，其中 $a_{ik}\in A_i$，决策属性 $C=(C_1,\cdots,C_m)$，则决策属性 C 相对于属性 A_i 的条件熵定义为：

$$H(C|A_i)=-\sum_{k=1}^{n}P(a_{ik})\sum_{j=1}^{m}P(C_j|a_{ik})\log_2 P(C_j|a_{ik}) \quad (5)$$

定义 2(基于信息论的属性重要度[4-5])设 $S=<U,A\cup C,V,f>$ 是一个决策表系统，其中 U 是对象的集合，A 是条件属性集合，C 是决策属性集合，属性子集 $B\subseteq A$，则对任意属性 $A_j\in A-B$，该属性的重要度定义为：

$$SGF(A_j,B,C)=H(C|B)-H(C|B\cup\{A_j\}) \quad (6)$$

特别地，若取 $B=\phi$，则 $SGF(A_j,C)=H(C)-H(C|A_j)$，此时属性 A_j 相对于决策属性 C 的重要度也就是属性 A_j 和 C 的互信息。

定义 3(属性集合重要度)设条件属性集 $A=(A_1,A_2,\cdots,A_p)$，决策属性 C，则属性集 A 相对于决策属性 C 的重要度定义为集合内每个属性相对于决策属性 C 的重要度的均值，即：

$$SGF(A,C)=\left(\sum_{i=1}^{p}SGF(A_i,C)\right)/Card(A) \quad (7)$$

2.2 WCB 分类模型

令 $E_1,\cdots,E_s$ 是属性集 $\{A_1,A_2,\cdots,A_p\}$ 的一个划分，且集合内部属性间相关性较强，集合之间属性近似独立。待分类样本 $X=(x_1,x_2,\cdots,x_p)$(或记为 $(e_1,\cdots,e_s)$，$e_j\in E_j$)属于类 C_i 的概率可表示为：

$$P(C_i|X)=P(C_i)P(x_1,\cdots,x_p|C_i)/P(X)=P(C_i)\prod_{j=1}^{s}P(e_j|C_i)$$

对于给定样本，$P(X)$ 是确定的，故：

$$P(C_i|X)\propto P(C_i)\prod_{j=1}^{s}P(E_j|C_i) \quad (8)$$

针对每个属性集 E_j，根据式(7)赋予权重 w_j，即：

$$w_j=\left(\sum_{A_k\subseteq E_j}SGF(A_k,C)\right)/Card(E_j) \quad (9)$$

式(8)修改为：

$$P(C_i|X)\propto P(C_i)\prod_{j=1}^{s}P(E_j|C_i)^{w_j} \quad (10)$$

对于集合 E_j 内部，由于属性间具有一定的相关性，根据 CB 模型分类思想，有：

$$P(E_j|C_i)=P(x_{j1},\cdots,x_{jl}|C_i)=Corr_{E_j}\times\prod_{k=1}^{l}P(x_{jk}|C_i) \quad (11)$$

其中 $Corr_{E_j}$ 由式(2)计算。

结合式(10)，式(11)，故 WCB 分类模型为：

$$V_{WCB}(X)=\arg\max_i P(C_i)\prod_{j=1}^{s}P(E_j|C_i)^{w_j}=$$

$$\arg\max_i P(C_i)\prod_{j=1}^{s}Corr_{E_j}\times\prod_{k=1}^{l}P(x_{jk}|C_i)^{w_j} \quad (12)$$

2.3 WCB 模型的构造步骤

(1)对训练样本进行缺失处理和离散化处理。

(2)分类器的构造。

①扫描训练样本集，统计训练集中，类别 C_i 的个数 d_i 和类 C_i 中属性 A_k 取值为 a_{ik} 的实例个数 d_{ik}，构成统计表；

②对训练属性集进行聚类，并由式(9)计算属性子集 E_j 的权重 w_j；

③计算所有的先验概率 $P(C_i)=d_i/d$，由式(2)和式(11)计算条件概率 $P(E_j|C_i)$，形成概率表；

④式(2)中，选择控制参数 $\beta\in[0,0.3]$，取步长 $h=0.01$，选取训练效果最优的 β 值构建分类器；

(3)分类，对于样本 X，调用概率表和构建好的分类器，得出分类结果。

3 实验分析

为了验证 WCB 算法的分类效果，本文选用 UCI[6]机器学习库中的 8 个数据集进行算法测试，以分类正确率作为算法优劣的主要评价指标。在相同的试验环境下，利用 MATLAB 编程分别实现了 WNB 算法、CB 算法和本文提出的 WCB 算法。数据集中连续属性进行离散化处

理。由于 Letter-Recognition、kr-vs-kp 和 MushRoom 数据集样本容量或属性个数较多，一次测试需要较长时间，采用分割数据集的方法进行测试，取 2/3 的数据作为训练集，1/3 数据作为测试集。其余数据集均采用 10 折交叉验证，取 10 次的平均值作为实验的测试结果。实验结果如表 1 所示。

表 1 实验数据集及分类效果

数据集 Data Set	实例数 Size	类 Class	属性 Attribute	正确率/% WNB	CB	WCB
Breast_Cancer	699	2	9	0.942 8	0.982 8	0.989 9
Contraceptive_Method_Choice	1 473	3	9	0.934 7	0.645 5	0.943 4
kr-vs-kp	3 196	2	36	0.865 0	0.937 7	0.938 5
Letter-Recognition	19 999	26	16	0.892 9	0.904 8	0.907 7
MushRoom	8 124	2	22	0.998 9	0.998 9	0.998 0
Nursery	12 960	5	8	0.910 8	0.956 8	0.956 3
Solar_Flare	1 389	2	10	0.807 75	0.812 8	0.843 1
tic-tac-toe	958	2	9	0.698 2	0.790 98	0.831 5
Average accuracy				0.881 4	0.878 8	0.926 0

仿真实验表明，WCB 算法在大部分数据集上分类正确率高于 CB 算法和 WNB 算法，由于本文构造的 WCB 算法兼顾了不同的属性相关性和属性重要性，更能反映真实情况，并克服了 CB 算法和 WNB 算法的不足。因为本算法既要调用属性集合重要度子函数，又要选取控制参数 β，所以算法运行的时间比 CB 算法和 WNB 算法稍长。当属性子集重要度 w_j 均为 1 时，WCB 算法即为 CB 算法；当向量相关度系数 $Corr_{E_j}$ 均为 1 时，WCB 算法则为 WNB 算法。因此，本文提出的算法不会比二者分类效果差。

实验都采用了 UCI 标准数据集，因此实验结果具有一定的可比性。

本文提出的 WCB 算法放宽了 NB 的两个假设，同时考虑属性相关性和属性重要性，进一步扩展了现有贝叶斯分类算法，有效地提高了分类效果。同时本文提出一种属性集合重要度的计算方法，随着属性相关性和属性重要度研究的发展，还可以使用其他度量属性相关性或属性重要性的方法，寻找分类效果更好的 WCB 算法是今后的研究方向。

参考文献

[1] 章舜仲，王树梅，黄河燕，等.基于属性相关性分析的贝叶斯模型[J].情报学报，2007，24(2)：58-65.

[2] HARRY Z, SHENG S L. Learning weighted naive bayes with accurate ranking [A]. Fourth IEEE International Conference on Data Mining (I CDMπ 04)[C]. Brighton, UK. 2004：567-570.

[3] 邓维斌，黄蜀江，周玉敏.基于条件信息熵的自主式朴素贝叶斯分类算法[J].计算机应用，2007，27(4)：888-891.

[4] 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.

[5] 曾黄麟.粗集理论及其应用(修订版)[M].重庆：重庆大学出版社，1998.

[6] NEWMAN D J, HETTICH S, BLAKE C L, et al. UCI repository of machine learning databases [EB/OL]. http://www.ics.uci.edu/mlearn/MLRepository. html, 1998.

(收稿日期：2010-11-14)

作者简介：

郑默，女，1985 年生，硕士，主要研究方向：智能计算、数据挖掘。

刘琼荪，女，1956 年生，教授，主要研究方向：智能计算、数据挖掘、应用统计。