一种利用关联规则的改进朴素贝叶斯分类算法①

陈朝大[1]　梁柱勋[2]　郑士基[3] (1.广东技术师范学院天河学院 广东 广州 510540;
2.西安交通大学 陕西 西安 710049; 3.广东江门市新会区有线广播电视网络中心 广东 江门 529100)

摘　要: 朴素贝叶斯分类是一种简单而高效的分类模型，然而条件独立性假设在现实中很少出现，致使其性能有所下降。通过引入关联规则，从两方面来改善朴素贝叶斯分类的性能。一方面，通过对关联规则的挖掘，发现条件属性之间的关联关系，并且利用这种关联关系弱化朴素贝叶斯的独立性假设；另一方面，通过关联规则的置信度，给朴素贝叶斯加权。

关键词: 分类模型; 朴素贝叶斯; 数据挖掘; 置信度; 关联规则

Modified Naive Bayes Classifier Using Association Rules

CHEN Chao-Da[1], LIANG Zhu-Xun[2], ZHENG Shi-Ji[3]
(1.Tianhe College of Guangdong Polytechnical Normal University, Guangzhou 510540, China; 2.Xi'an JiaoTong University, Xi'an 710049, China; 3. JiangMen City XinHui Cable TV Station, Jiangmen 529100, China)

Abstract: Naive Bayes classification is a kind of simple and effective classification model. However, the performance of this model may be poor due to the assumption on the condition independence. By introducing association rules, this classification model can be improved in two way. On the one hand, the associated relationship between condition attributes can be found out through association rules mining, in order to weaken the independent assumption. On the other hand, Naive Bayes is weighted by computing the confidence of association rules.

Keywords: classification model; naive bayes; data mining; confidence level; association rules

1　引言

分类是数据分析中一个非常重要的过程。随着人们对数据挖掘方面的研究，提出了多种分类的方法，如决策树、关联规则、贝叶斯、神经网络、遗传算法、基于案例的推理等[1]。其中朴素贝叶斯分类算法以其简单、高效与准确等特点，得到了广泛的研究与应用。

2　朴素贝叶斯分类器

2.1 朴素贝叶斯的基本原理

朴素贝叶斯分类基于贝叶斯定理。朴素贝叶斯是将未知分类的一个 n 维样本 $X=\{x_1,x_2,...x_n\}$分配给类集合 $C=\{C_1,C_2,C_3,...,C_m\}$中的一个 C_i。问题可转换成，给定元组 X，分类法将预测 X 属于具有最高后验概率的类。也就是说，朴素贝叶斯分类法预测属于类 C_i，当且仅当[2]

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

根据贝叶斯定理，可知

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (2)$$

其中 $P(C_i|X)$是指在 $X=\{x_1,x_2,...x_n\}$的情况下，事件 C_i 发生的概率，$P(C_i)$指事件 C_i 发生的概率。

由于对于每个 Ci，P(X)的取值都是一样的，所以只需要求得 $P(X|C_i)P(C_i)$的最大值就可以得到对应的

① 收稿时间:2010-03-14;收到修改稿时间:2010-05-04

$P(C_i|X)$的最大值，从而求出相应的 C_i 完成分类。

然而，在条件属性与类别属性非常多的情况下，求解 $P(X|C_i)$变得非常困难。为了让计算变得简单，朴素贝叶斯假设项集 X 中的条件属性两两相互独立，这样可得下述公式：

$$P(X \mid C_i) = \prod_{j=1}^{n} P(x_j \mid C_i) \quad (3)$$

朴素贝叶斯分类器完成的工作就是计算

$$c(X) = \arg\max_{C_i \in C} P(C_i) \prod P(x_j \mid C_i) \quad (4)$$

其中 argmax 表示 c(x)的值应等于使上式得到最大值的 C_i。

2.2 朴素贝叶斯的缺点

虽然朴素贝叶斯分类器的原理非常简单直接，并且可以与一些较复杂的分类算法相媲美，但是它却忽略了两个重要的因素，使其的性能受到影响。

1) 朴素贝叶斯分类器的条件独立假设在实际情况中很少发生，这一要求过于苛刻，可能导致其性能下降。

2) 朴素贝叶斯分类器假设每个条件属性对类别属性的影响是一样的，然而在现实中，每个条件属性对类别属性的影响并不是一样的。

下文引入关联规则，针对上述朴素贝叶斯分类器存在的问题各自进行改进。

3 关联规则

设 $I=\{I_1,I_2,...,I_m\}$是项的集合，关联规则是形如 A=>B 的蕴涵式，其中 $A=\{A_1,A_2,...,A_k\} \subset I$，$B=\{B_1,B_2,...,B_k\} \subset I$，并且 A 与 B 不相交。A 称为前件，B 称为后件。关联规则中的两个重要概念是支持度与置信度。关联规则 A=>B 的支持度表示数据库的数据元中，同时包含集合 A 与 B 的概率 $P(A \cup B)$。其置信度表示同时包含 A、B 的数据元与只包含 A 的数据元的百分比 P(B|A)。

$$P(B \mid A) = \frac{P(A \cup B)}{P(A)} \quad (5)$$

关联规则揭示了属性之间的关联关系。

3.1 关联规则的挖掘规则

关联规则的挖掘算法有很多种，本文关联规则按照如下几点挖掘：

1) 关联规则是一个二项集，也就是前件和后件都只有一个属性，这样做有利于得到各条件属性之间的相关性；

2) 关联规则的前件必须是条件属性，后件可以是条件属性或者是类别属性。

按照以上规则，给定支持度阈值 s 和置信度阈值 c。利用 Apriori 算法产生形如$\{x_1,x_2\},\{x_3,c_1\}$的只含两个属性的频繁项集。再根据置信度公式(5)与 s 计算出关联规则。例如下面是一个给定所需集合，得出的可能的关联规则的一个例子：

输入：条件属性的集合 $X=\{x_1,x_2,x_4,x_4\}$

类别属性的集合 $C=\{c_1,c_2\}$

训练数据集合 DataSet

输出：关联规则 $x_1 => x_3$, $x_2 => x_3$, $x_4 => c_1$

3.2 应用关联规则弱化独立性假设

在现实中，条件属性之间并不是完全独立的，属性之间存在着一定的联系。一种计算属性之间相关性的方法是(3)中提出的 x_2，这种方法需要对每两个属性的所有的可能取值求其频度，时间复杂度较高。所以本文提出一种更为简便的方法，就是利用挖掘出来的关联规则，计算属性之间的相关性。根据上述挖掘关联规则的规定，关联规则可分成两类，I 类是只包含条件属性的规则，II 类是包含类别属性的规则[3]。我们用 I 类中关联规则的置信度 c 来近似前件与后件的相关性。例如，$x_1 => x_3$ 的置信度为 40%，意味着在出现 x_1 的元组中，出现 x_3 的概率为 40%。当这一关联规则的置信度高达某一阈值时，我们就可以预测，当 x_1 在元组中出现时，那么 x_3 也很有可能在该元组中出现。这时，我们就可以说 x_1 与 x_3 的相关性很强，并称 x_3 是 x_1 的冗余属性。既然 x_3 相对于 x_1 是冗余的，那么我们当然可以把 x_3 从样本集合中删除。这样一来，就可以减少集合中的属性个数，并提高了计算 $P(X|C_i)$的准确性。当然我们并不需要让置信度为 1 时，才能称属性之间具有强相关性。如可以把阈值设置成 80%，当某一关联规则的置信度高达这一阈值时，就可以说后件是前件的冗余属性。

删除某些属性后计算 $P(X|C_i)$，该属性就不会出现在连剩中。这也可以把它当作被剩 1 代替了，这也是符合实际的。因为如果 x_3 是 x_1 的冗余属性，那么当 x_1 出现了，那么 x_3 出现的概率就很大，可近似于 1。

对于那些没有在关联规则中出现的属性，如上例

中的 x_4，就没有在任何一条关联规则中出现，那么我们就可认识 x_4 与其它属性的相关性很弱，近似地说，x_4 与其它属性是相互独立的。

改进后的项集 X 的后验概率计算公式如下：

$$P(X \mid C_i) = \prod_{k=1}^{m} P(x_k \mid C_i) \qquad (6)$$

其中，$x_k \in redAttrSet$，m 是 RedAttrSet 中的属性的个数。RedAttrSet 删除冗余属性后的约简属性集。

属性的约简算法可描述如下：

输入：训练数据集 DataSet，关联规则集 RuleSet，属性样本 AttrSet，置信度阈值 value

输出：约简属性样本 RedAttrSet

//按照前述规则，从数据集中挖掘出 I 类关联规则

```
RuleSetI = DigRuleI(DataSet);
for each rule in RuleSet
c = compute_confidece(rulei);  //计算每个关联规则的置信度
if  c < value then
rulei.post->RedAttrSet;  //如果置信度小于阈值，相关性弱，不必删除后件属性
end if                   //rule.pre 规则的前件，rule.post 规则的后件
rulei.pre->RedAttrSet    //把前件属性加入约简属性集
end for
end
```

3.3 应用关联规则对传统贝叶斯加权

在传统的朴素贝叶斯分类算法中，假设每个条件属性对类别属性的影响是一样，即连剩时所有 xi 的后验概率的系数或权值都一样。然而现实中，有些条件属性对类别属性的影响很大，有些却相对较少，而不相等的。针对这一问题，Harry[4]等提出了一种基于不同条件属性对类别属性的重要性的不同，赋予各属性不一样的权值的加权朴素贝叶斯 WNB(Weigthed Naive Bayes)。这一方法的主要问题是怎样决定各属性的权值。这种加权朴素贝叶斯的计算公式如下：

$$c(X) = \arg\max_{C_i \in C} P(C_i) \prod P(x_j \mid C_i)^{a_j} \qquad (7)$$

其中 α_j 是条件属性 x_j 的权值。具体思想：当某一条件属性 x_j 对类别属性 C_i 的影响较大时，在计算 $P(x_j|C_i)$ 时，给予该项一个较低的权值，反之给予较高的权值。这样就可以改进朴素贝叶斯分类算法中，忽略不同条件属性对类别属性的不同影响的问题，提高了朴素贝叶斯的分类精度。本文根据关联规则，给出一种计算各属性权值的方法。

上述 II 类关联规则中，每个规则的后件都是类别属性[5]。根据这些关联规则，可得到不同的条件属性对类别属性的影响程度或重要性。其依据就是每条 II 类规则的置信度。如上例中的关联规则 $x_4 => c_1$，当其置信度为 50%时，表明当条件属性 x4 在元组中出现时，c_1 也在该元组出现的概率是 50%。当这个置信度较高时，也就可以说 x_4 对类别属性 c_1 具有较高重要性，这样我们就有理由在计算 $P(x_4|c_1)$时，赋予该项一个相应的权值，使 $P(x_4|c_1)$的概率变大，从而体现出条件属性 x_4 对类别属性 c_1 的重要性[6]。由于概率 P 和置信度 c 的取值范围都是[0, 1]，如果希望使概率 P 变大，那么相应地其权值 α 就要取得较小值。无论权值 α 取得多小，也不会使概率 P 的值大于 1。根据以上分析，α 的取值如下：

$$a = 1 - c \qquad (8)$$

c 是 II 类关联规则的置信度。当置信度 c 较高，条件属性对类别属性的影响较大。所以让 1 减去其置信度，使其权值 α 取较小，从而其概率 P 相应变大；反之则使 α 取较大。

对于没有在 II 类关联规则中出现的条件属性，如上例中的 x_1，x_2，x_3 均未在 II 类关联规则中出现，本文简单地令其权值 $\alpha = 1$。因为在挖掘关联规则时，是根据支持度与置信度阈值来挖掘的。当条件属性没出现在 II 类关联规则时，表明该条件属性对类别属性的影响较少，让其权值 $\alpha = 1$ 也是可以理解的[7]。

计算各个条件属性的算法描述如下：

输入：训练数据集 DataSet，关联规则集 RuleSet，属性样本 AttrSet。

输出：每个条件属性的权值 α。

//按照前述规则，从数据集中挖掘出 II 类关联规则

```
RuleSetII = DigRuleII(DataSet);
for each rule in RuleSetII  //对于出现在 II 类
```

规则中的条件属性，计算其权值

```
c = compute_confidence(rulei);
rulei.pre.α = 1 - c;
end for
for each Atrr not in RuleSetII   //对于没有出现在 II 类规则中的条件属性，权值置为 1
α = 1;
end for
```

根据以上两步基于关联规则的改进，从理论上可提高朴素贝叶斯分类器的分类精度。

3 性能比较

从(4)与改进后的朴素贝叶斯分类公式(9)可知，改进后的计算公式中，样本 X 对于每个类 C_i 所得到的后验概率都大于或等于改进前的计算结果。这是因为：(1)所有置信度大于阈值的冗余属于的 $P(x_k|C_i)=1$；(2)总有 $P(x_k|C_i)^{\alpha_j} \geq P(x_k|C_i)$，其中 $0<P(x_k|C_i)<1$，$0 \leq \alpha_j \leq 1$。这样使得在(4)式中较大的 X 后验概率 $P(C_i|X)$，在(9)在越大；而小的后验概率则变化不大。把大的后验概率放大有利于提高分类算法的精度[8]。

$$C(X)=\arg\max_{C_i \in C} P(C_i)\prod_{k=1}^{m} P(x_k \mid C_i)^{a_j} \qquad (9)$$

4 结语

综上所述，通过引入关联规则，从两方面来改善朴素贝叶斯分类的性能。一方面，通过对关联规则的挖掘，发现条件属性之间的关联关系，并且利用这种关联关系弱化朴素贝叶斯的独立性假设；另一方面，通过关联规则的置信度，给朴素贝叶斯加权。

参考文献

1 谈恒贵,王文杰,李游华.数据挖掘分类算法综述.微型机与应用, 2005(2):4—9.

2 Jiawei HAN, Micheline KAMBER 著;范明,孟小峰译.数据挖掘概念与技术.机械工业出版社, 2006:201—202.

3 王峻.一种基于属性相关性度量的朴素贝叶斯分类模型.安庆师范学院学报, 2007,13(2):14—16.

4 Zhang H, Sheng S. Learning weighted Naive Bayes with accurate ranking. Proceedings of the 4th IEEE International Conference on Data Mining, 2004:567—570.

5 吴宁,柏春霞,祝毅博.一种应用关联规则森林的改进贝叶斯分类算法.西安交通大学学报, 2009,43(2):48—52.

6 班桦,吴耿锋,吴绍春.分布式数据挖掘中间层.计算机工程与设计, 2006,4:661—663.

7 张明卫,王波,张斌,朱志良.基于相关系数的加权朴素贝叶斯分类算法.东北大学学报, 2008,29(7):953—955.

8 Li G, Qi X, Wang X, et al. A linear-time algorithm for computing translocation distance between signed genomes. CPM, 2004.