计算机科学2000Vol. 27 №. 10

用于数据采掘的贝叶斯分类器研究*)

Studies on Bayesian Classifier in Data Mining

林士敏　田凤占　陆玉昌

(清华大学计算机科学与技术系　智能技术与系统国家重点实验室　北京100084)

Abstract Classification is a basic and important task in data mining and pattern recognism. Bayesian classification is the field of Bayesian learning and Bayesian networks where plenty research works have done and remarkable results have get. In this paper we discuss the basic concepts of Bayesian classification based on Bayesian learning, the principle and using effects of variant Bayesian classifiers, compare their performances, and introduce the research advance and further works.

Keywords Bayesian classifier, Data mining, Machine learning, Knowledge discovery

所谓分类器是一个函数 $f(x)$，它给需要分类的实例 x 赋予类标签 $c_j \in C(j=1,2,\cdots,m)$，实例 x 由一组属性值 $a_1,\cdots,a_n$ 描述，C 是类变量，取有限个值，可看成有限个元素的集合。进行分类首先要构造一个分类器。从预先分类的实例进行有导师学习并建立分类器，是机器学习的中心问题之一。已有的分类器如决策树、决策表、神经网络、决策图和规则等，都可以看成不同的函数表示法。贝叶斯分类器指的是基于贝叶斯学习方法的分类器。

1　最大后验假设(MAP)和优化贝叶斯分类器

在观察到数据之前，根据背景知识或经验确定某个假设空间 H 中的假设 h 成立的概率 $P(h)$，称为 h 的先验概率。D 为训练数据集合，在没有关于哪个假设成立的知识而观察到的 D 的概率，称为 D 的先验概率，记为 $P(D)$。在假设 h 成立的条件下观察到 D 的概率记为 $P(D|h)$。在观察到训练数据 D 的条件下，假设 h 成立的概率 $P(h|D)$称为 h 的后验概率。后验概率反映了训练数据对假设成立概率的影响，它是依赖于 D 的。已知 $P(h)$、$P(D|h)$和 $P(D)$，贝叶斯定理提供了计算一个假设 h 的后验概率的方法，因而成为贝叶斯学习的基石：

$$P(h|D)=P(D|h)P(h)/P(D)$$

通常，学习的任务是：对于给定观察数据 D，在 H 中发现最可能的假设 $h\in H$。任何这样的具有最大可能的假设称为最大后验(MAP，maximum a posteriori)假设，记为 h_{MAP}：

$$\begin{aligned} h_{MAP} &\equiv \arg\max_{h\in H} P(h|D) \\ &= \arg\max_{h\in H} P(D|h)P(h)/P(D) \\ &= \arg\max_{h\in H} P(D|h)P(h) \end{aligned} \quad (1)$$

在没有任何背景知识的情况下，可以假定 H 中所有的假设有相同的先验。于是要想找到最可能的假设，在(1)式中仅仅需要考虑 $P(D|h)$。$P(D|h)$称为给定 h 时数据 D 的似然。任何使 $P(D|h)$最大的假设称为最大似然(ML—maximum likelihood)假设 h_{ML}：

$$h_{ML} \equiv \arg\max_{h\in H} P(D|h) \quad (2)$$

与学习问题不同，分类问题的提法是：给定训练数据，求新实例的最可能的分类。可以对新实例应用MAP假设来回答这个问题，不过用优化贝叶斯分类可以得到更好的结果。所谓优化贝叶斯分类以各自假设的后验为权值，综合所有的假设的预测，来得到新实例的分类。设新实例的分类取某个类集合 C 的任何一个值 c_j，那么新实例分类为 c_j 的概率是：

$$P(c_j|D)=\sum_{h_i\in H} P(c_j|h_i)P(h_i|D)$$

新实例的优化贝叶斯分类是使得 $P(c_j|D)$取最大值的 c_j：

$$\arg\max_{c_j\in V} \sum_{h_i\in H} P(c_j|h_i)P(h_i|D) \quad (3)$$

按照(3)式对新实例进行分类的任何系统称为优化贝

*)国家重点基础研究发展计划项目，国家自然科学基金项目。"九五"国家攀登计划预选项目。林士敏　副教授、访问学者，研究方向：机器学习、数据采掘与知识发现。田凤占　博士研究生，研究方向：数据采掘与知识发现。陆玉昌　教授，研究方向：数据采掘与知识发现，机器学习，知识工程。

叶斯分类器，给定数据集合 D、假设空间 H 和所有假设的先验概率，在新实例属于各个类的概率中，这个方法取其中概率最大的类作为正确的分类。

一般说来，使用相同的假设空间和先验，贝叶斯优化分类优于其他分类方法，它还有一个独特的性质：它的预测可以对应于不属于假设空间的假设。

2 朴素贝叶斯分类器

朴素贝叶斯分类器是一个简单、有效而且在实际使用中很成功的分类器，其性能可以与神经网络、决策树分类器（例如 C4.5）相比，在某些场合优于其他分类器。

设有变量集 $U=\{A_1,\cdots,A_n,C\}$，其中 $A_1,\cdots,A_n$ 是实例的属性变量，C 是取 m 个值的类变量，假设所有的属性都条件独立于类变量 C。即每一个属性变量都以类变量作为唯一的父节点，就得到朴素的贝叶斯分类器。

分类问题描述为：给定一组训练例 $D_1,\cdots,D_p$，其中 $D_k(k=1,2,\cdots,p)$ 由 U 中的变量组 $(A_1,\cdots,A_n,C)$ 的一组值描述。给定一个新的具有属性 $a_1,\cdots,a_n$ 的实例，判别这个新实例的类别。

使用朴素贝叶斯分类器进行分类的做法是：通过概率计算，从待分类的实例的属性值 $a_1,\cdots,a_n$ 求出最可能的分类目标值。即计算各类 $c_j\in C$ 对于这组属性的条件概率 $P(c_j|a_1,\cdots,a_n)$，其中 $j=1,2,\cdots,m$，并输出条件概率最大的类标签作为目标值。应用贝叶斯定理和条件独立假设：

$$P(c_j|a_1,\cdots,a_n)=\frac{P(a_1,\cdots,a_n|c_j)P(c_j)}{P(a_1,\cdots,a_n)}$$

$$=\alpha\cdot P(c_j)\cdot\prod_{i=1}^{n}P(a_i|c_j)$$

其中 α 是正规化常数。以后验概率作为分类指示，即输出具有最大后验概率的类标签 c_{NB}：

$$c_{NB}=\underset{c_j\in C}{\operatorname{argmax}}P(c_j)\cdot\prod_{i=1}^{n}P(a_i|c_j)c_j \tag{4}$$

其中 c_{NB} 表示朴素贝叶斯网络输出的目标值，常数 α 可以省略。通常(4)式也作为朴素贝叶斯分类器的定义。实际计算时，(4)式中的项 $P(c_j)$ 可以通过计算训练例中 c_j 出现的频率来估计。$P(a_i|c_j)$ 的数目等于属性的数目乘以目标值（即类）的数目，也可以通过计算训练例中出现的频率来估计。

朴素贝叶斯分类器存在的问题是需要一个很强的属性之间条件独立的假设，而这个假设在许多问题中并不成立，如果在这些问题中忽视这一点，理应会引起分类的误差。用图模型来表示时，朴素贝叶斯分类器就是一种特殊的贝叶斯网络，称为朴素贝叶斯网络，能否去掉这个假设，用一般的贝叶斯网络作为分类器而得到更好的分类器呢？这就导致了用一般的贝叶斯网络作为分类器的研究。

3 贝叶斯网络分类器

一般的贝叶斯网络表示了变量集 $U=\{X_1,\cdots,X_n\}$ 的联合概率分布：

$$P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P(X_i|\Pi_{X_i}) \tag{5}$$

其中 Π_{X_i} 表示父节点。也就是说，属性变量可以有一个以上的父节点，因而属性变量之间的依赖关系可以得到表示，这里，学习贝叶斯网络的问题描述为：给定 U 中的一组实例构成的训练集合 $D=\{u_1,\cdots,u_N\}$，找到一个与 D 匹配最好的网络 B。这样，学习贝叶斯网络的问题转化为优化问题，这时类变量和属性变量不加区别。

实际处理这个问题的方法是在可能的网络构成的空间中进行启发式搜索。搜索成功的关键是确定一个合理的评分函数，评价网络对训练数据的匹配程度，以指导搜索。有两种主要的评分函数：贝叶斯评分函数和最小长度原理（MDL—minimal description length）评分函数。它们是渐进正确的，即随着样本数目的增加，得分最高的网络将任意逼近样本的概率分布。下面讨论采用 MDL 评分函数学习贝叶斯网络的一些结果。

用 $MDL(B|D)$ 表示给定训练数据集 D 时，贝叶斯网络 B 的 MDL 评分函数，那么

$$MDL(B|D)=1/2\cdot\log N|B|-LL(B|D) \tag{6}$$

其中 $|B|$ 表示网络参数的个数，$1/2\cdot\log N$ 是网络每一个参数使用的比特数（信息单位）。第一项表示网络 B 的描述长度，即描述网络需要的比特数。第二项是给定 D 时 B 的对数似然函数，它表示基于概率分布 P_B，描述 D 所需要的比特数：

$$LL(B|D)=\sum_{i=1}^{N}\log(P_B(u_i)) \tag{7}$$

对数似然的统计含义是：对数似然越高，B 对数据集 D 中的概率分布的拟合就越好。

使用以上方法，可以从训练数据集合得到一个表示变量 $A_1,\cdots,A_n,C$ 的联合概率分布 $P(A_1,\cdots,A_n,C)$ 的贝叶斯网络 B，然后用它作为分类器对新的实例进行分类。设新的实例的属性集为 $a_1,\cdots a_n$，计算后验概率 $P(c_j|a_1,\cdots,a_n)$，其中 $c_j\in C$，并以有最大后验概率的类标签作为输出。这个方法分类的性能取决于贝叶斯网络学习的渐进正确性，即给定大的数据集，学习得到的网络能以小的误差逼近领域变量的概率分布。

Nir Friedman 等人用美国加州大学的 Irvine 知识库的25个数据库的数据对这个问题进行了深入的研究。实验表明，对于某些数据集合，贝叶斯网络分类器的性能明显优于朴素贝叶斯分类器，而对另一些数据

集合，则不如朴素贝叶斯分类器，主要表现在分类的准确性下降，也就是说，实际应用时可能遇到这样的情况：学习过程得到MDL评分很好的网络，作为分类器的性能却很差。究其原因，是当变量属性增多时，类变量对于属性变量的条件概率的偏差在MDL评分函数中反映迟钝，而这对分类却有决定意义，这使得贝叶斯网络分类器在属性很多时性能变差。实验表明，当超过15个属性时，贝叶斯网络分类器的性能不如朴素贝叶斯分类器。对这些网络做深入分析的结果显示，在两个数据库上表现相当差的网络，对分类有影响的相关属性的数目相当小。在数据库包含的35～36个属性中，分类器在进行分类时只依赖于其中的5个属性。

所谓相关属性的划分基于"马尔科夫毯"概念。马尔科夫毯的定义是：在一个给定的网络结构 G 中，变量 X 的马尔科夫毯包括 X 的父节点(双亲)、X 的子女节点和 X 的子女节点的父节点。马尔科夫毯中的节点有如下重要的性质：给定变量 X 的马尔科夫毯，X 条件独立于网络中的其他变量。即 X 的条件概率分布只依赖于马尔科夫毯中的节点。而不依赖于网络中在马尔科夫毯以外的节点。

也就是说，一般的贝叶斯网络在分类时进行了特征选择，只检查类变量 C 的马尔科夫毯中的属性(节点)，而忽略了其他属性。通常这样做是有用的，被忽略的确实是与分类无关的属性。可是，正像实验所表明，这样做也可能忽略了对于分类至关重要的属性。

4 扩展的朴素贝叶斯网络分类器TAN

基于以上分析结果，Nir Friedman等提出了扩展的朴素贝叶斯网络分类器，其基本思想是在朴素贝叶斯分类器的基础上，在属性之间增添连接弧，以消除朴素贝叶斯分类器关于条件独立的假设。这样的弧称为扩展弧。从节点 A_i 到 A_j 的扩展弧表示属性 A_j 对分类的影响也取决于 A_i 的值。可能有这种情况：待分类实例的属性值 a_i 和 a_j 对分类的影响都不大，$P(a_j|c)$和 $P(a_i|c)$的值低，但 $P(a_j|c,a_i)$的值却高，这时朴素贝叶斯分类器会过度降低实例属于类 c 的概率，而扩展的贝叶斯网络分类器却可以避免这一点。

如何增添一组最好的扩展弧是一个难处理的问题，因为这相当于学习以类变量 C 为根节点的最好的贝叶斯网络。为了解决这个问题，Nir Friedman等对增添扩展弧加以某些限制，提出一种"树扩展朴素贝叶斯网络"(tree-augmented naive Bayesian network)作为分类器，简称TAN。这种网络中，类变量没有父节点，每一个属性变量以类变量和最多一个属性变量为父节点。于是，每一个属性变量可以有一个指向自己的扩展弧。这种网络具有扩展的朴素贝叶斯网络分类器的优点(如计算的简单性和鲁棒性等)，性能优于朴素贝叶斯分类器，而且可以通过有效的学习来建立。

学习TAN分类器是一个人机交互的过程，其方法基于1968年Chow和Liu提出的学习树结构的贝叶斯网络的方法，该方法将构造最大似然树的问题简化为在一个图中寻找最大权重跨度(spanning)树的问题，使得所选择的弧构成一颗树，而且附属于选择弧的权重之和为最大。这个算法使用互信息函数：

$$I_P(X;Y)=\sum_{X,Y}P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)}$$

来衡量属性 X 提供给属性 Y 的信息量。构造TAN分类器的过程使用条件互信息函数：

$$I_P(X;Y|Z)=\sum_{X,Y,Z}P(X,Y,Z)\log\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}$$

代替Chow和Liu方法中的互信息函数，条件互信息函数衡量的是当属性 Z 的值已知时，属性 X 提供给属性 Y 的信息量。TAN分类器构造过程如下：

(1)通过训练集计算属性对之间的条件互信息 $I_{P_D}(A_i;A_j|C)$，$i\neq j$。

(2)建立一个以 $A_1,\cdots,A_n$ 为顶点的完全无向图，顶点 A_i 到 A_j 标以权重 $I_{P_D}(A_i;A_j|C)$。

(3)建立一个最大权重跨度树。

(4)选择一个根结点，将所有边的方向设置成由根结点向外指，把无向树转换为有向树。

(5)增加一个类变量结点 C，并加上从 C 到每一个属性结点 A_i 的弧。

建立最大权重跨度树的方法是：首先把边按权重由大到小排序，之后遵照被选择的边不能构成回路的原则，按照边的权重由大到小的顺序选择边。这样由所选择的边构成的树便是最大权重跨度树，Nir Friedman等证明，这样得到的TAN贝叶斯网络具有最大似然，而且计算的时间复杂度为 $O(n^2\cdot N)$，其中 n 是属性的个数，N 是实例的个数。

TAN贝叶斯网络的一个扩展是多网分类器，即将数据集合按照不同的类分组，对不同的类增添不同的扩展弧和不同的树结构，对每一个类 c_i 都建立一个关于属性变量 $A_1,\cdots,A_n$ 的贝叶斯网络，称为 c_i 的局部网。局部网的集合连同 C 的先验 $P(C)$就称为贝叶斯多网分类器。这种分类器的性能与TAN分类器相当。

5 朴素贝叶斯分类器的提升

提高朴素贝叶斯分类器的性能的另一种方法是"提升"(Boosting)，提升是一种通用的方法，由Freund和Schapire于1995年提出，其主要思想是从训练例学习一系列的分类器，每一个分类器根据前一个分类器错误分类的实例，对训练例的权重进行修正，再学习新的分类器。例如，学习得到分类器 H_i 后，将其错误分

类的训练例的权重提高，然后再从改变权重的训练例中学习下一个分类器 H_{t+1}，此过程重复 T 次，最后得到的分类器输出各个 H_t 的输出的加权平均，权重对应于 H_t 在其训练集的分类准确性。算法如下：

令权重 $w_i^{(1)}=1/N$，其中 $i=1,\cdots,N$ 是实例的序号，T 为循环次数，$t=1$ 到 T 执行：

* 对于权重 $w_i^{(t)}$，建立假设 $H^{(t)}:X\to[0,1]$
* 令 $H^{(t)}$ 的误差为 $\varepsilon^{(t)}=\sum_{i=1}^{N} w_i^{(t)}|y_i-h_i^{(t)}(x_i)|$
* 令 $\beta^{(t)}=\varepsilon^{(t)}/(1-\varepsilon^{(t)})$ 且 $w_i^{(t+1)}=w_i^{(t)}(\beta^{(t)})^r$，其中 $r=1-|y_i-h_i^{(t)}(x_i)|$
* 正规化 $w_i^{(t+1)}$ 使得它们的和为1.0。

假定每一个独立的分类器都是有用的，即 $\varepsilon^{(t)}<0.5$，那么 $\beta^{(t)}<1$，且当 $|y_i-h_i^{(t)}(x_i)|$ 增加时 $w_i^{(t+1)}$ 也增加。实验表明，只要增加错误分类实例的权重，改变其他细节对算法的结果没有影响。因此，Freund 等建议最后组合假设取：

$$H(x)=\frac{1}{1+\prod_{t=1}^{(a)}(\beta^{(t)})^{2h(x)-1}}$$

其中单独分类器的线性组合为：

$$l(x)=\frac{\sum_{t=1}^{(a)}(\log 1/\beta^{(t)})H^{(t)}(x)}{\sum_{t=1}^{(a)}\log 1/\beta^{(t)}}$$

提升的朴素贝叶斯分类器的性能一般说优于已经发表的使用其他学习方法的最好的结果，起码与之相当。提升算法的时间复杂度为 $O(Tef)$，其中 T 是提升的次数，e 是训练例的个数，f 是属性的个数。

小结 贝叶斯分类器性能优于或相当于其他分类器，具有语义明确和容易理解的优点。其中朴素贝叶斯分类器虽然做了一个很强的关于属性之间相互条件独立的假设，而且这个假设在实际问题中往往不能满足，但是在实际应用中却取得了引人注目的成功，其性能可以同 C4.5相比，当类变量的属性值较多时，结构无约束的一般的贝叶斯网络作为分类器，分类的准确性下降，寻找更合理的评分函数是一个有待研究的问题。树扩展朴素贝叶斯网络 TAN 对朴素贝叶斯分类器作了改进，允许属性变量以另一个属性变量为父节点，取消了属性之间相互条件独立的假设，其性能优于朴素贝叶斯分类器。提升算法对朴素贝叶斯分类器也有较好的效果。

参 考 文 献

1 Friedman N. Bayesian Network Classifiers. Machine Learning, 1997, 29: 131～163

2 Duda R O, Hart P E. Pattern Classification and Scence Analysis, New York: John Wiley & Sons, 1973

3 Langley P, et al. An analysis of Bayesian classifiers. In: Proc. of the National Conf. on Artificial Intelligence (AAAI'92). Menlo Park, CA: AAAI Press, 1992. 223～228

4 Chow C K, Liu C N. Approximating discrete probability distributions with dependence tree. IEEE Trans. on Information Theory, 1968, 14: 462～467

5 Pearl J. Probabilistic Reasoning in Intelligent Systems. San Francisco, CA: Morgan Kaufmann, 1988. 387～390

6 Elkan C. Boosting and naive Bayesian learning: [Technical Report No. CS97-557]. Department of Computer Science & Engineering, Univ. of California, 1997

（上接第82页）

让 $\dot{x}$="do nothing"，去掉因果机制 $x(Pa(x),\dot{x})$（只是在规范影响图中不标出因果机制），那么得到的就是因果网。我们只考虑对 D 响应的机会变量（对 D 不响应的机会变量，我们认为它不具有我们讨论的因果机制或者说因果关系），我们通过构造规范影响图可以得到因果网络结构，利用因果网络结构进行因果定性分析。如果具有机制独立性假设、组件独立性假设、参数独立性假设、参数模块性假设及似然等价性假设，我们就可以用学习非因果网络参数和概率分布的方法学习因果网的参数和概率分布，利用概率联合分布进行因果定量分析。

结束语 因果关系理论无论在数据分析还是预测中都是非常重要的，因果关系是相互关联的变量中最强的一种关系。贝叶斯因果分析的理论还不够成熟，还需要进一步的理论探讨和实验检验。

参 考 文 献

1 Heckerman D. A Bayesian Approach to Causal Discovery. [Technical Report MSR-TR-97-05]. Microsoft Research, Microsoft Corporation, 1994

2 Heckerman D. A Bayesian Approach to Learning Causal Networks: [Technical Report MSR-TR-95-04]. Microsoft Research, Microsoft Corporation, 1995

3 Heckerman D. Learning Bayesian Networks: [Technical Report MSR-TR-95-02]. Microsoft Research, Microsoft Corporation, 1995

4 Heckerman D. Learaning Bayesian networks: The Combination of Knowledge and Statistical Data. Mcahine Learning, 1995, 20: 197～243