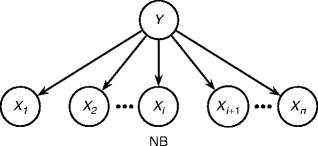
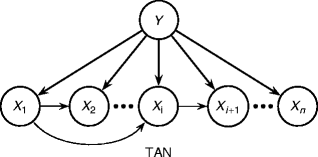
TAN算法

71117230 柳沿河

【算法简介】

树增广朴素贝叶斯是一种半朴素贝叶斯学习方法。它通过使用一个树结构来放宽朴素贝叶斯属性独立性的假设，其中每个属性只依赖于类和另一个属性。使用最大加权生成树，使训练数据的似然性达到最大。

属性之间的相互依赖关系可以通过允许属性依赖于其他非类属性来直接解决。树增广朴素贝叶斯(Tree Augmented Naïve Bayes, TAN) (Friedman et al.，1997)采用了树结构，允许每个属性依赖于类和最多一个其他属性。下图显示了NB和TAN分别创建的模型类型的贝叶斯网络表示。

由于放宽了属性独立性的假设，TAN大大降低了朴素贝叶斯的偏差，代价是增加了方差。实证结果(Friedman et al.， 1997)表明，在许多数据集上，该方法显著降低了朴素贝叶斯的0 - 1损失，而在所有被检验的数据集上，该方法比朴素贝叶斯更容易获得较低的0 - 1损失。

【算法实现方式】

该算法使用python语言结合现有代码自行实现。

代码运行方法：python tan.py 数据集(如vote.arff，需在同一目录下) 学习方法(n:NB, t:TAN) 训练集比例(如0.9)

【算法流程】

算法伪代码见<TAN.pdf>.

【实验结果】

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | breast-w | Kr-vs-kp | labor | mushroom | vote |
| 正确分类数 | 69 | 300 | 6 | 813 | 42 |
| 错误分类数 | 1 | 20 | 0 | 0 | 2 |
| 正确率 | 98.57% | 93.75% | 100.00% | 100.00% | 95.45% |

\*注：采用十折交叉验证，即训练集占比90%

【结果分析】

和朴素贝叶斯的对比结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | breast-w | Kr-vs-kp | labor | mushroom | vote |
| TAN | 98.57% | 93.75% | 100.00% | 100.00% | 95.45% |
| NB | 98.57% | 85.94% | 100.00% | 95.08% | 93.18% |

可以看到，TAN考虑了属性之间的依赖关系，所以相比朴素贝叶斯正确率有了明显的提升；但在较大的数据集上的训练时间比朴素贝叶斯要长。

【参考文献】

[1] Chow, C. K., & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 14, 462–467.

[2] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29(2), 131–163.