文章编号:1000 - 582x(2000)05 - 0049 - 04

信度网分类器

邢永康, 沈一栋
(重庆大学 计算机学院,重庆 400044)

摘 要: 分类问题是人工智能中机器学习研究的一个重要问题,它在模式识别、故障诊断以及数据挖掘等领域有着广泛的应用。利用信度网可以构造出分类性能更好的分类器。文章着重探讨了 Naive Bayes 分类器,增强的 Naive Bayes 分类器和通用信度网分类器的构造方法,并分析比较了这三类分类器的优缺点。

关键词: 分类器; 信度网; 机器学习
中图分类号: TP 181 文献标识码: A

1 分类器(Classifier)简介

分类问题可以描述为:给定一个实例数据集合 D,该集合中的每一个实例是变量 $A_1, A_2, \cdots, A_n$ 和 C 的一个取值组合,即:$\langle a_1, a_2, \cdots, a_n; c\rangle$。寻找一个函数 $f(\mathbf{A})$,对于任意实例$\langle a_1, a_2, \cdots, a_n\rangle$,该函数可以正确输出变量 C 的值 c。习惯上将变量 $A_1, A_2, \cdots, A_n$ 称为属性变量,变量 C 称为类变量,实例数据集合称为训练数据库,并将学习所得的函数 $f(\mathbf{A})$ 称为分类器。作为分类器的函数 $f(\mathbf{A})$ 一般都是非线性函数,表示从属性变量 $A_1, A_2, \cdots, A_n$ 到类变量 C 的一个映射。常用的分类器模型有决策树、决策列表、神经网络、决策图等。Quinlan 的 C4.5 被认为是经典的分类器[1]。它采用 ID3 算法,应用信息论中熵的概念分析实例数据,从而构造出一棵熵值下降最快的判定树作为分类器。

人工智能中的许多问题如模式识别、故障诊断以及数据挖掘等都可以看作是一个分类问题。下面是一个选择隐形眼镜的分类问题。

例:配戴隐形眼镜一般可以分为 3 类。即 C = {配戴硬性隐形眼镜,配戴软性隐形眼镜,不适于配戴隐形眼镜}。为了给希望配戴隐形眼镜的人进行分类,一般需要考虑 4 个属性:

1. 配镜者的年龄 A_1 = {年轻,年老,早期老视}
2. 配镜者的视力缺陷 A_2 = {近视,远视}
3. 配镜者是否有散光 A_3 = {是,否}
4. 配镜者泪液分泌情况 A_4 = {减少,正常}

利用 C4.5 算法构造的分类器如图 1 所示。对于不同的配镜者,根据以上的 4 个属性,可以快速地判定出适合他的隐形眼镜类型。

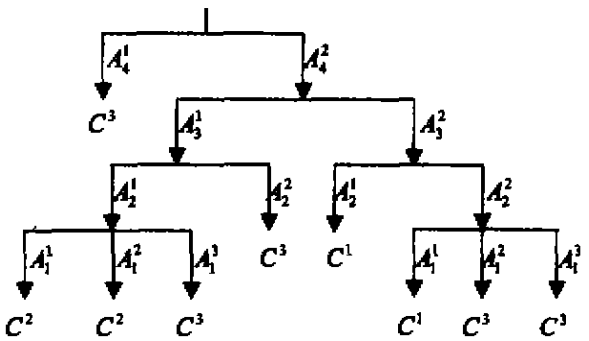

图 1 C4.5 算法学习所得的判定树

2 最简单的分类器 —Naive Bayes

Duda 和 Hart 提出了一种基于贝叶斯公式的分类器 ——Naive Bayes[2] 分类器。在该分类器中,假设各个属性变量之间相互独立。当训练数据库 D 的所有实例都完整时,通过对实例出现频率的统计,求出给定类变量 C 条件下各个属性变量 A_i 的条件概率 $P(A_i \mid C)$ 以及类变量的概率 $P(C)$,就可以完成分类器的学习。

收稿日期:2000-03-27
基金项目:国家自然科学基金(69883009)及教育部跨世纪优秀人才基金
作者简介:邢永康(1971-),男,陕西人,重庆大学博士生。主要从事人工智能研究。

即:

$$P(A_i = A_i^k \mid C = C^k) = \frac{N(A_i = A_i^k \& C = C^k)}{N(C = C^k)} \quad (1)$$

$$P(C = C^l) = \frac{N(C = C^l)}{N} \quad (2)$$

其中,$N(A_i = A_i^k \& C = C^k)$ 表示训练数据库中,A_i 取第 k 个值、C 取第 k 个值的实例的数目;N 表示训练数据库中实例的总个数。

对于一个具体的测试实例 $A_1, A_2, \cdots, A_n$,根据贝叶斯公式:

$$P(C \mid A_1, A_2, \cdots, A_n) = \frac{P(C, A_1, A_2, \cdots, A_n)}{\Sigma_C P(C, A_1, A_2, \cdots, A_n)} = \frac{P(C)\prod_{i=1}^{n} P(A_i \mid C)}{\Sigma_C P(C)\prod_{i=1}^{n} P(A_i \mid C)} \quad (3)$$

计算类结点 C 的后验概率。从中选取概率最大的类变量的值 C^l,作为该实例的分类值。

Naive Bayes 分类器具有计算简单、分能性能良好等特点,但它的属性变量之间相互独立假设在许多情况下并不成立。如在关于配戴隐形眼镜的实例中,配镜者的年龄与配镜者的视力缺陷之间明显存在着依赖关系:年轻者常患有近视;而年老者常患有老视。因此武断地假设这两个属性之间相互独立,与实际情况并不相符。如放松甚至放弃 Naive Bayes 分类器中的属性变量之间相互独立假设的限制,则能构造出分类性能更好的分类器。利用 Pearl 提出的信度网模型[4]就能做到这一点。信度网是一种不确定知识的表达及推理模型,它根据变量之间的独立性关系,将随机变量构成的联合概率分布表示为直观的图形方式。如对配戴隐形眼镜的分类问题,它的 Naive Bayes 分类器可以用信度网简单明了地表示为图 2。

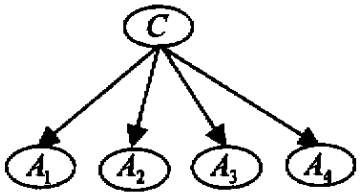

图 2 Naive Bayes 分类器

在图 2 中,根据信度网的结构语义,各个属性变量之间相互独立。同时,Naive Bayes 分类器中需要学习的参数在该图中表现为信度网的的条件概率表 因此,利用信度网的学习、推理算法,可以很容易地完成 Naive Bayes 分类器的学习与分类任务。信度网能够表示更复杂的独立性关系,所以利用信度网,可以彻底地放弃属性独立性假设,在各个属性之间建立复杂的依赖关系,从而构造出分类性能更好的分类器。

3 通用信度网分类器

直接以建立在学习数据库上的信度网作为分类器,称为通用信度网分类器—GBNC(General Belief Network Classifier)。通用信度网分类器的建立过程,就是利用信度网的学习算法,从实例数据建立所有属性变量和类变量构成的信度网的过程;利用这类分类器的分类过程,就是采用信度网的推理算法计算给定属性变量的值时类变量的后验分布的过程。

信度网的学习是当前信度网研究的一个热点,研究者已经提出了许多切实可行的学习算法。这些算法依据不同的基本思想可以分为两类:一类是基于测度的模型选择法(Modal Selection)。这类方法首先根据学习的具体要求,选用一种测度,来衡量一个信度网模型对实例数据的适合程度。再根据该测度从模型空间中找出测度最优的模型作为信度网模型。由于模型空间太大,一般都采用启发式搜索算法,以测度为指导,逐步构造出该测度最优的模型。常用的测度有:贝叶斯测度(BDe)[5]、最小描述长度测度(MDL)[6]、贝叶斯信息测度(BIC)等。第二类是基于独立性测试的方法。这类方法主要着眼于信度网的结构语义,即信度网的结构表达了变量之间的条件独立性关系。通过对实例数据的分析,使用不同的条件独立性测试方法(如互信息测试等)可以获得变量之间的依赖关系,再根据这种依赖关系来构造信度网。这类方法有:CL 算法(Chow 1968)、三阶段学习算法[7]等。

上述关于配戴隐形眼镜的例子,通过学习得到的通用信度网分类器如图 3 所示。该分类器真实地表达了属性之间的相互依赖关系。

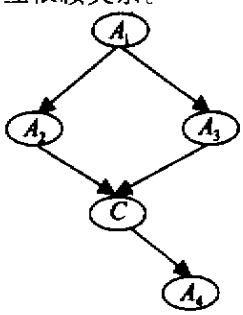

图 3 通用信度网分类器

利用信度网分类器进行分类时，将要求分类的实例的各个属性值作为证据输入到信度网中，利用任意一种信度网的推理算法（如消息传递算法、关联树算法等），求出在给定证据下，类变量的各种取值概率。其中概率值最大的类变量取值就是该实例的分类值。

Friedman 用 28 个实例数据库，对通用信度网分类器和 Naive Bayes 分类器进行了试验比较，结果表明[8]：在一般情况下，前者的分类能力优于后者，但在某些情况下其分类准确性弱于后者。尤其当属性变量的数目超过 15 时，这种情况更为明显。这可以从 Friedman 在试验中所采用的测度——MDL 来分析。MDL 测度是基于编码理论提出来的，其定义如下：

$$\mathrm{MDL}(B^S) = \frac{1}{2}\log N^* \mid B^S \mid - \log P(D \mid B^S) \quad (4)$$

等式右端第一项表示保存一个信度网所需的编码长度，第二项表示采用该模型对实例数据编码压缩后的编码长度。学习过程中，要求该模型的 MDL 测度值最小，即要求式中第二项的值最大。如果假设实例数据库完整且各个实例数据之间相互独立，则第二项可以转化为：

$$\log P(D \mid B^S) = \sum_{i=1}^{N} \log P(C^i \mid A_1^i, A_2^i, \cdots, A_n^i) + \sum_{i=1}^{N} \log P(A_1^i, A_2^i, \cdots, A_n^i) \quad (5)$$

上式右端第一项是分类测度，从分类器的角度看，这一项应该达到最大值。但在信度网学习中要求的 MDL 测度最小并不能保证分类测度最大。随着属性变量数目的增加，属性变量的取值组合的总数目将以 n 的指数增加，所以每一种取值组合的概率 $P(A_1^i, A_2^i, \cdots, A_n^i)$ 将变得很小，根据对数的性质，(5)式右端第二项将变成一个非常小的负数。与此相比，(5)式右端第一项的变化却要小得多，因此，第二项将对结果起主导作用。也就是说，第一项中的微小变动都将被第二项所掩盖。可见，一个 MDL 测度最优的信度网结构并不一定是分类测度最优的结构，这就是在某些情况下，通用信度网分类器的分类性能弱于 Naive Bayes 分类器的一个原因。由于当实例数据数目很大时，BDe 测度、BIC 测度与 MDL 测度是相互等价的，所以，用其他两个测度进行学习也存在与 MDL 测度同样的问题。

基于独立性测试的学习方法，是从信度网结构本身所包含的独立性来入手的，学习的目的是尽可能地表现出实例数据中所包含的条件独立性关系，所以在这种方法中，不会存在基于测度的学习所遇到的问题。Chen 通过实验比较也证实了这一点[9]。因此，在通用信度网分类器中，一般采用基于独立性测试方法来学习建立信度网分类器。

4　增强的 Naive Bayes

在信度网中，将一个结点的所有父结点、所有子结点以及所有子结点的父结点，三者构成的集合称为该结点的马尔科夫覆盖。根据有向马尔科夫属性，一个结点状态的变化，只受它的马尔科夫覆盖中的节点状态变化的影响，与其它结点的状态无关。因此，在通用信度网分类器中，那些不属于类结点的马尔科夫覆盖的属性，将对分类结果不产生作用。而在实际应用中，这些属性对某些分类有可能是一个关键属性，这就会造成在某些情况下，通用信度网分类器分类错误。由于在 Naive Bayes 分类器中，所有的属性结点都属于类结点的马尔科夫覆盖，所以可以对 Naive Bayes 分类器直接进行增强，保留其结构特点，放松它的独立性假设，使属性变量之间存在复杂的依赖关系，这类分类器称为增强的 Naive Bayes 分类器——ANBC(Augmented Naive Bayes Classifier)。

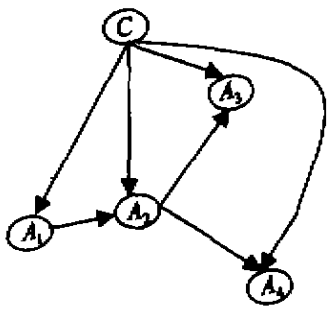

图 4　树形加强的 Naive Bayes 分类器

最简单的增强的 Naive Bayes 分类器是 Geiger 提出的树形增强 Naive Bayes 分类器——TANBC(Tree Augmented Naive Bayes Classifier)。其基本思想是假设类结点是所有属性结点的父结点，但各个属性结点之间可以形成树形关系（即在该结构中，一个结点只能有一个父结点）。如配戴隐形眼镜的例子，其树形增强的 Naive Bayes 分类器如图 4 所示，在该图中，所有的属性结点都是类结点的子结点，各个属性结点之间构成了树形依赖关系。

通过对 Chow 等的 CL 算法[10]的修改，Friedman 提出了树形增强 Naive Bayes 分类器的学习算法[8]，算法时间复杂度为 $O(n^2 \cdot N)$。

TAN 算法：

1) 利用下式，计算出每一对属性变量的条件互信息：

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j, c} P(x_i, x_j \mid c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c) P(x_j \mid c)}$$

2) 以所有的变量作为结点，建立一个无向完全图。并对每一对结点 (X_i, X_j) 的连接边赋予权值 $I(X_i, X_j \mid C)$。

3) 利用求最大权生成树算法，求出该无向图的最大权生成树。

4) 任意选定一个结点作为根结点，由此出发，将所有的无向边转化为指向其邻居结点的有向边，形成树 B_T。

5) 将类结点 C 作为所有属性结点的父结点，加入树 B_T 中，形成图形 B_{TAN}。

6) 学习以图形 B_{TAN} 作为结构的信度网的条件概率表。

树形增强 Naive Bayes 分类器的优点在于，其学习算法的时间复杂度要比通用信度网分类器低。同时，它假设各个属性变量之间存在树形依赖关系，这要比 Nvive Bayes 分类器中各个属性变量之间独立假设更符合实际情况。对 Naive Bayes 分类器的进一步增强是使各个属性变量之间不再局限于树形网络关系，而是可以形成任意的信度网结构，这样的分类器称为信度网增强的 Naive Bayes 分类器——BNAN (Belief Network Augmented Naive - Bayes)。信度网增强的 Naive Bayes 分类器的表达能力要优于树形增强 Naive Bayes 分类器，其分类性能也更好，但其学习的复杂性也明显地增加了。

5 总结与进一步的研究

基于贝叶斯公式的简单分类器——Naive Bayes 分类器在试验中表现出良好的分类性能，但其属性变量的独立性假设在实际中往往不成立。以信度网为工具，可以从两个方向对其进行改进。一个是利用信度网对独立性的表达能力，直接以属性变量和类变量构成的信度网作为分类器，称为通用信度网分类器。一个是直接对 Naive Bayes 分类器进行增强，即放松它的属性独立性假设，使属性之间可以形成树形结构或任意的信度网结构，分别称为树形增强的 Naive Bayes 分类器和信度网增强的 Naive Bayes 分类器。实验证明，改进后的分类器分类能力明显提高。

以上的研究都假设用于分类的学习数据库是完整的，这种假设在实际应用中往往不成立。如在一个医学疾病的分类系统中，类变量代表各种疾病，属性变量代表各种疾病所表现出的症状。对于一种疾病，某些症状对确诊该疾病很关键，而有些症状则无足轻重，这些对该疾病无关紧要的症状往往不会被观测，所以由此产生的数据库中的实例就不是完整的。对不完整的实例数据如何进行分类，仍然是一个有待解决的问题。

参考文献：

[1] QUINLAN J. C4.5: Programs for Machine Learning[M] San Francisco, CA: Morgan Kaufmann, 1993

[2] DUDA R O, HART P E. Pattern Classification and Scene analysis[M]. New York: John Wiley & Sons, 1973

[3] LANGLEY P, IBA W, THOMPSONM K. An analysis of bayesian classifiers[A] Proceedings of Tenth National Conference on Artificial Intelligence[C]. Menlo Park, CA: AAAI Press, 1992 223-228.

[4] JUDEA PEARL Fusion, Propagation, and Structuring in Belief Networks[J]. Artificial Intelligence, 1986, 29: 241-288

[5] HECKERMAN D, GEIGER D, CHICKERING M Learning Bayesian Networks: The combination of knowledge and statistical data[J]. Machine Learning, 1995, 20: 197-243

[6] LAM W, BACHUS F Learning Bayesian Networks: An Approach based on the MDL principle[J] Computation intelligence, 1994, 10: 269-293

[7] CHEN J, BELL D A, LIU W An algorithm for Bayesian Belief network construction from data[A] Proceedings of AI &&STAT'97[C]. Lauderdale, Florida, 1997 83-90

[8] FRIEDMAN N, GOLDSZMIDT M. Building Classifier using Bayesian networks[A]. Proceedings of National Conference on Artificial Intelligence[C] Menlo Park, CA: AAAI Press, 1996. 1 277-1 284.

[9] CHENG J, GREINER R, COMPARING Bayesian network classifiers[A]. Henri P. Proceedings of the fifteenth conference on uncertainty in artificial intelligence[C]. San Francisco: Morgan Kaufmann Publishers, 1999.

[10] CHOW C K, LIU C N. Approximation discrete probability distributions with dependence trees[J]. IEEE Trans on Information Thoery, 1968, 14: 462-467

(下转 77 页)

色物质产生总量，催化剂的使用可提高单位时间的 PET 分解率。

参考文献：

[1] 李绍英．聚脂废料的再生利用[J]．河北轻化工学报，1996，17(4)：21-23.

[2] 雅克·贝扎里亚．制备对苯二甲酸盐或对苯二甲酸的方法[P] 1066056.1992-11-11

[3] ROSEN B I, GROVE M. Preparation of Purified Terephthalic Acid From Wast Polyethylene Terephthalate [P] USP5,095,145.10,1992-03-10

[4] 杨治伟．PET 在 EG 中的醇解反应[J] 北京服装学院学报，1994，14(1)：29-32.

Research on Mechanism of Depolymerizing Waster PET by Uniting the Ethandiol and the Luis Alkali

YANG Guang, JIANG Tao, WU Li-mei

Research Ctr.-Resource Comp. Util. Engr., Chongqing University, Chongqing 400044, China

Abstract: The research shows that the depolymerizing reaction of waste PET is a synergetic and accelerated process that alcoholysis and alkaline hydrolysis act each other under the condition of excessive ethandiol and Luis alkali existing together. The enhancement of depolymerisation rate and depolymerizing speed are very remarkable. The process conditions of depolymerizing waste PET and recovery TPA and EG are oversimplified. The depolymerisation rate exceed 99% at the normal pressure and 180 ℃ in 15 min. The product of colored substance in depolymerizing system is affected by alkaline degree, reacting temperature and reacting time. It can be controlled effectively by choolsing different Luis alkali and adjusting the process conditions of depolymerisation.

Key words: waste PET; depolymerizing; mechanism

（责任编辑　钟学恒）

（上接 52 页）

Belief Network Classifier

XING Yong-kang, SHEN Yi-dong

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: Classification is an important research area in Artificial Intelligence, which has a broad-range of applications such as pattern recognition, diagnosis, data mining, and so on. A best classifier can be built by using belief networks. This paper mainly discusses how to build the Naive Bayes classifiers, the Augmented Naive Bayes classifiers, and the General Belief Network classifiers. Their respective advantages and shortcomings also be shown by a detailed comparison.

Key words: classifier; belief networks; machine learning

（责任编辑　吕寨英）