ARTIFICIAL INTELLIGENCE

人工智能

177

177

Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework

量化归纳偏差:人工智能学习算法和瓦兰特学习框架

David Haussler\*

戴维·豪斯勒\*

Department of Computer Science, University of California, Santa Cruz, CA 95064, U.S.A.

美国加州大学圣克鲁斯分校计算机科学系，邮编:95064

Recommended by Tom Mitchell

由汤姆·米切尔推荐

ABSTRACT

摘要

We show that the notion of inductive bias in concept learning can be quantified in a way that directl\_v relates to learning performance in the framework recently introduced by Valiant.Our measure of bias is based on the growth function introduced by Vapnik and Chervonenkis, and on the Vapnik-Chervonenkis dimension.We measure some common language biases, including restriction to conjunctive concepts, conjunctive concepts with internal disjunction, k-DNF and k-CNF concepts.We also measure certain types of bias that result from a preference for simpler hypotheses.Using these bias measurements we analyze the performance of the classical learning algorithm ]or ~onjunctive concepts from the perspective of Valiant's learning framework.We then augment this algorithm with a hypothesis simplification routine that uses a greed~v heuristic and show how this improves learning performance on simpler target concepts.Improved learning algorithms are also developed [~r conjunctive concepts with internal disjunction, k-DNF and k-CNF concepts.We show that all our algorithms are within a logarithmic ]hctor of optimal in terms of the namber of examples th O' require to achieve a given level of learning performance in the Valiant .[?amework.Our results hold .[?~r arbitrary attribute-based instance spaces defined by either tree-structured or linear attributes.

我们证明了概念学习中归纳偏向的概念可以用一种方式来量化，即在Valiant最近引入的框架中，directl\_v与学习绩效相关。我们对偏差的衡量是基于瓦普尼克和切弗嫩基斯引入的增长函数，以及瓦普尼克-切弗嫩基斯维度。我们测量了一些常见的语言偏差，包括对合取概念的限制、带内部析取的合取概念、k-DNF和k-CNF概念。我们也测量某些类型的偏差，这些偏差是由对更简单假设的偏好所导致的。利用这些偏差测量，我们从瓦兰特学习框架的角度分析了经典学习算法或邻接概念的性能。然后，我们用一个假设简化例程来扩充这个算法，这个例程使用一个greed启发式算法，并展示这是如何提高更简单的目标概念的学习性能的。改进的学习算法也发展了具有内部析取、k-DNF和k-CNF概念的[-r合取概念。我们证明了我们所有的算法都在一个最佳的对数函数范围之内，这是为了达到一个给定的学习水平。[？amework。我们的结果成立。[？由树结构或线性属性定义的任意基于属性的实例空间。

Introduction

介绍

The most extensively investigated learning task in artificial intelligence is that of learning a single concept from examples.For example, one might consider the task of learning to distinguish edible mushrooms from non-edible mush-rooms by looking at preclassified examples of actual mushrooms (see e.g. [35]).In this task we select a set of mushroom attributes (e.g. color, shape and size) and attempt to find a rule that distinguishes between edible and non-edible mushrooms expressed in terms of these attributes (e.g. edibleC~((color = red or orange) and (size = small)) or...).The set of attributes selected determines

人工智能中研究最广泛的学习任务是从例子中学习一个单一的概念。例如，人们可以考虑通过观察实际蘑菇的预分类实例来学习区分可食用蘑菇和非可食用蘑菇室的任务(例如参见[35)。在这个任务中，我们选择一组蘑菇属性(例如颜色、形状和大小)，并试图找到一个规则来区分可食用和不可食用的蘑菇，用这些属性来表示(例如，edibleC~((颜色=红色或橙色)和(大小=小))或...)。选定的属性集决定了

\* The author gratefully acknowledges the support of ONR grant N000t4-86-K-0454.

\*作者感谢ONR赠款N000t4-86-K-0454的支持。

Artificial Intelligence 36 (1988) 177 22 I

人工智能36 (1988) 177 22一

0004-3702/88/$3.50 1988, Elsevier Science Publishers B.V. (North-Holland)

0004-3702/88/$3.50 1988，埃尔塞维尔科学出版社(北荷兰)

178

178

D.HAUSSLER

D.HAUSSLER

the instance space used by the learning algorithm, and the type of expression allowed in specifying the rule determines the hypothesis space used by the algorithm.

学习算法使用的实例空间以及指定规则时允许的表达式类型决定了算法使用的假设空间。

In any realistic learning application, the entire instance space will be so large that any learning algorithm can expect to see only a small fraction of it during training.From this small fraction, a hypothesis must be formed that classifies all the unseen instances.If the learning algorithm performs well then most of these unseen instances should be classified correctly.However, if no restric-tions are placed on the hypothesis space and no "preference criterion" 1124] is supplied for comparing competing hypotheses, then all possible classifications of the unseen instances are equally possible and no inductive method can do better on average than random guessing [261.Hence all learning algorithms employ some mechanism whereby the space of hypotheses is restricted or whereby some hypotheses are preferred a priori over others.This is known as inductive bias.

在任何现实的学习应用程序中，整个实例空间都很大，任何学习算法在训练过程中都只能看到一小部分。从这一小部分，必须形成一个假设，将所有看不见的实例分类。如果学习算法表现良好，那么大多数这些看不见的实例应该被正确分类。然而，如果没有对假设空间进行限制，并且没有提供用于比较竞争假设的“偏好标准”，那么所有看不见的实例的可能分类都是同样可能的，并且没有归纳方法能够比随机猜测[261做得更好。因此，所有的学习算法都采用某种机制，从而限制假设的空间，或者某些假设优先于其他假设。这就是所谓的感应偏置。

The most prevalent form of inductive bias is the restriction of the hypothesis space to only concepts that can be expressed in some limited concept descrip-tion language, e.g. concepts described by logical expressions involving only conjunction (see e.g. [6, 10]).A still stronger bias can be obtained by also introducing an a priori preference ordering on hypotheses, e.g. by preferring hypotheses that have shorter descriptions in the given description language (see e.g. [24, 33]).While many forms of bias have been used, up to this point there has been no generally agreed upon language-independent measure of the strength of a bias, in particular, a measure that relates the strength of a bias to the performance of learning algorithms that use it, so that it will be useful in analyzing and comparing learning algorithms.This paper proposes such a measure, and demonstrates how it can be used to compare and prove perform-ance results for learning algorithms.

归纳偏见最普遍的形式是将假设空间限制为只能用某种有限的概念描述语言表达的概念，例如，仅涉及连接的逻辑表达式所描述的概念(例如，见[6，10)。还可以通过引入假设的先验偏好排序来获得更强的偏差，例如，通过偏好在给定描述语言中具有更短描述的假设(参见例如[24，33))。虽然已经使用了许多形式的偏差，但是到目前为止，对于偏差强度的语言无关的测量还没有达成一致，特别是，一种将偏差强度与使用它的学习算法的性能相关联的测量，因此它在分析和比较学习算法中是有用的。本文提出了这样一个度量，并演示了如何使用它来比较和证明学习算法的性能结果。

We measure bias with a combinatorial parameter defined on classes of concepts known as the growth function [43].A theory and methodology of pattern recognition based on this function has been developed by Vapnik [421.Applications of the theory to linear separators and Boolean circuits, and its relation to the preference for simpler hypotheses are discussed in 130].The present work can be viewed as an extension of this methodology to concept learning problems in artificial intelligence.

我们用一个组合参数来衡量偏差，这个组合参数定义在被称为增长函数[43]的概念类上。瓦普尼克·[421提出了一种基于这一功能的模式识别理论和方法。130]中讨论了该理论在线性分离器和布尔电路中的应用，以及它与更简单假设偏好的关系。目前的工作可以被看作是这种方法在人工智能的概念学习问题上的延伸。

The growth function of a hypothesis space can be used to define its

假设空间的增长函数可以用来定义它的

Vapnik-Cherv6nenkis dirnension, a combinatorial parameter closely related to the notion of capacity introduced in [9].Extending the results of [42], in [5, 12] it is shown that this parameter is strongly correlated with learning performance as defined in the learning framework introduced by Valiant ]21, 39-4l].We adopt this framework here as well.

Vapnik-Cherv6nenkis方向，一个与[引入的容量概念密切相关的组合参数9]。扩展[42]的结果，在[5，12]中显示，该参数与学习性能密切相关，如由瓦兰特]21，39-4l]引入的学习框架中所定义的。我们在这里也采用这个框架。

The salient feature of the Valiant framework is that it only requires that the learning algorithm produce a hypothesis that with high probability is a good

瓦兰特框架的显著特征是，它只要求学习算法产生一个假设，即高概率是好的

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

179

179

approximation to the target concept.It does not demand that the learning algorithm identify the target concept exactly.Angluin has called this framework "probably approximately correct" identification [1].By adopting this weaker performance criterion, we are able to show that a number of simple learning algorithms actually perform near optimally in terms of the number of training examples they need.These algorithms include the "classi-cal" algorithm for conjunctive concepts, and variants of this algorithm for related target classes.

接近目标概念。它不要求学习算法精确地识别目标概念。安格伦称这个框架“可能大致正确”地识别了[1]。通过采用这种较弱的性能标准，我们能够证明许多简单的学习算法实际上在它们所需的训练例子的数量方面表现得接近最优。这些算法包括用于连接概念的“类-cal”算法，以及用于相关目标类的该算法的变体。

In the Valiant framework, a training sample is created by drawing instances from the instance space independently at random according to some fixed probability distribution, and labeling them "+" or "-" according to whether or not they are instances of the target concept.Each such labeled instance is called a (random) example of the target concept.The error of a hypothesis is defined as the probability that it will disagree with a random example of the target concept drawn according to the same fixed probability distribution used to generate the training sample.Thus, if we are trained to recognize edible mushrooms on the west coast of the United States, we expect the rule we learn to work well in west coast forests, but not necessarily in east coast forests.

在Valiant框架中，通过根据某个固定的概率分布从实例空间中随机独立地绘制实例，并根据它们是否是目标概念的实例来标记它们“+”或“-”来创建训练样本。每个这样标记的实例被称为目标概念的(随机)例子。假设的误差被定义为与根据用于生成训练样本的相同固定概率分布绘制的目标概念的随机示例不一致的概率。因此，如果我们被训练识别美国西海岸的可食用蘑菇，我们期望我们学习的规则在西海岸的森林中有效，但不一定在东海岸的森林中有效。

A good approximation to the target concept is a hypothesis with small error.Thus formally, the Valiant criterion demands that a learning algorithm produce a hypothesis that with high probability has small error with respect to a given probability distribution and target concept.A class of target concepts is considered learnable by the algorithm only if this happens for any target concept in the class and any probability distribution on the instance space.Thus, while the framework is probabilistic, it is not tied to any particular probability distribution or even to any type of distribution, and hence it provides an extremely robust performance guarantee.

对目标概念的一个很好的近似是一个误差很小的假设。因此，在形式上，瓦兰特准则要求学习算法产生一个假设，即对于给定的概率分布和目标概念，高概率具有小误差。只有当类中的任何目标概念和实例空间上的任何概率分布发生这种情况时，算法才认为一类目标概念是可学习的。因此，尽管该框架是概率性的，但它并不依赖于任何特定的概率分布，甚至也不依赖于任何类型的分布，因此它提供了一个极其可靠的性能保证。

Two measures of learning complexity are relevant in this framework.The first is sample complexity.This is the number of random examples needed to produce a hypothesis that with high probability has small error.As above, the sample complexity of a learning algorithm on a given target class is defined by taking the number of random examples needed in the worst case over all target concepts in the class and all probability distributions on the instance space.For each of the learning algorithms we present, we show that the sample complex-ity is within a poly-logarithmic factor of optimal.

在这个框架中，学习复杂性的两个衡量标准是相关的。首先是样本的复杂性。这是产生假设所需的随机例子的数量，假设概率高，误差小。如上所述，给定目标类上的学习算法的样本复杂度是通过对该类中的所有目标概念和实例空间上的所有概率分布采用最坏情况下所需的随机示例的数量来定义的。对于我们提出的每一种学习算法，我们证明了样本复数在最优的多对数因子内。

The second performance measure is computational complexity, which we take as the worst-case computation time required to produce a hypothesis from a sample of a given size.We show that each of the learning algorithms we present has computational complexity polynomial in the sample size and in the number of attributes that define the instance space.

第二个性能指标是计算复杂度，我们将它视为从给定大小的样本中产生假设所需的最坏情况计算时间。我们证明了我们提出的每一种学习算法在样本大小和定义实例空间的属性数量上都有计算复杂度多项式。

The paper is organized as follows.In Section 1 we define instance spaces on tree-structured and linear attributes, and we define various hypothesis spaces on such instance spaces.In Section 2 we take a new look at Mitchell's version

论文组织如下。在第1节中，我们在树结构和线性属性上定义了实例空间，并在这些实例空间上定义了各种假设空间。在第二部分，我们对米切尔的版本进行了新的审视

18(I

18(一

D.HAUSSLER

D.HAUSSLER

space framework for learning concepts from examples [28], here from a probabilistic point of view.Mitchell defines the version space as the set of all hypothesis in the hypothesis space that are consistent with a given set of examples.We show (Lemma 2.2) that by using a hypothesis space that is strongly biased and by drawing independent random examples, the version space will shrink very rapidly, with high probability, to a set of hypotheses that cluster around the target concept in the sense that their errors are small relative to the target concept.For this initial result, bias is measured in terms of the size of the hypothesis space (see also [28, 32]).The result is then refined in Section 3 (Theorem 3.3) when we introduce the growth function as a measure of bias.

从概率的观点来看，从例子[28]学习概念的空间框架。米切尔将版本空间定义为假设空间中与给定示例集一致的所有假设的集合。我们证明(引理2.2)，通过使用一个有很大偏差的假设空间，并通过画出独立的随机例子，版本空间将非常迅速地缩小，很有可能成为一组围绕目标概念的假设，因为它们相对于目标概念的误差很小。对于这个初始结果，偏差是根据假设空间的大小来衡量的(参见[28，32)。当我们引入增长函数作为偏差的度量时，结果将在第3节(定理3.3)中得到改进。

In Sections 4 to 7 we use these results to analyze the performance of several learning algorithms.We first consider what we call the classical algorithm for learning conjunctive concepts (Algorithm 4.1).This algorithm produces thc unique maximally specific conjunctive hypothesis consistent with the training sample.Corollaries 4.5 and 4.8 provide bounds on the learning performance of this algorithm.The latter results show that its sample complexity is within a logarithmic factor of optimal (see also [12]).

在第4节到第7节中，我们使用这些结果来分析几种学习算法的性能。我们首先考虑学习合取概念的经典算法(算法4.1)。该算法产生与训练样本一致的唯一最大特定合取假设。推论4.5和4.8为该算法的学习性能提供了界限。后一个结果表明，其样本复杂度在最佳对数因子之内(另见[12)。

In Section 5 we consider the problem of learning simple (i.e. syntactically short) conjunctive concepts on instance spaces with many attributes.We adapt the greedy heuristic for set cover [18] to simplify the hypothesis produced by the classical algorithm.The result is a learning algorithm (Algorithm 5.2) that has sample complexity within a poly-logarithmic factor of optimal for simple conjunctive target concepts (Corollary 5.7).Sections 6 and 7 extend these results to k-DNF, k-CNF and internal disjunctive target concepts (see Section 1 ).The main results are given in Corollaries 6.1 and 7.2 respectively.Finally, a number of remaining open problems are outlined in the conclusion,

在第5节中，我们考虑在具有许多属性的实例空间上学习简单(即，语法上短)的连接概念的问题。我们对集合覆盖[18]采用贪婪启发式算法来简化经典算法产生的假设。结果是一个学习算法(算法5.2)，对于简单的连接目标概念(推论5.7)，其样本复杂度在最优的多对数因子内。第6节和第7节将这些结果扩展到k-DNF、k-CNF和内部析取目标概念(见第1节)。主要结果分别在推论6.1和7.2中给出。最后，结论中概述了一些尚未解决的问题，

Notation

注释

We use "log" to denote the logarithm base 2 and "In" to denote the natural logarithm.For any set S, [S I denotes the cardinality of S.

我们用“对数”表示以2为底的对数，用“英寸”表示自然对数。对于任何集合，[表示集合的基数

1.Instances and Concepts

1.实例和概念

In the simplest type of inductive concept learning, each instance of a concept is defined by the values of a fixed set of attributes, not all of which are necessarily relevant.For example, an instance of the concept "red triangle" might be characterized by the fact that its color is red, its shape is triangular and its size is 5.Following [24[, we consider three type of attributes.A nominal attribute is one that takes on a finite, unordered set of mutually exclusive values, e.g. the attribute color, restricted to the six primary and secondary colors, or a Boolean attribute, taking only the values true and false.A linear attribute is one with a linearly ordered set of mutually exclusive values, e.g. a real-valued or integer-

在最简单的归纳概念学习中，一个概念的每个实例都是由一组固定属性的值定义的，并不是所有的属性都是相关的。例如，“红色三角形”概念的一个实例的特征可能是它的颜色是红色，它的形状是三角形，它的大小是5。继[24[之后，我们考虑三种类型的属性。名义属性是一种采用有限的、无序的互斥值集的属性，例如属性颜色，限于六种原色和二次色，或者布尔属性，仅采用真值和假值。线性属性是一个线性有序的互斥值集合，例如实值或整数-

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

181

181

valued attribute.A tree-structured attribute is one with a finite set of hierarchi-cally ordered values, e.g. the attribute shape shown in Fig. 1.Only the leaf values of a tree-structured attribute (e.g. the values triangle, square hexagon, proper\_ellipse, circle, crescent and channel of Fig. 1) are directly observable.

有值属性。树形结构的属性是具有有限组分层有序值的属性，例如图1所示的属性形状。只有树形结构属性的叶值(例如图1中的三角形、正方形、椭圆、圆、新月和通道)是可直接观察到的。

Since a nominal attribute can be converted to a tree-structured attribute by

因为名义属性可以通过

addition of the special value any\_value, we will restrict our discussion to

加上特殊值any\_value，我们将把讨论限制在

tree-structured and linear attributes.Throughout the paper we will assume that each attribute has at least two distinct observable values.

树形结构和线性属性。在整篇论文中，我们将假设每个属性至少有两个不同的可观察值。

Let At,..., A, be attributes with observable value sets V~ ...., V,, respec-

让在，...是具有可观察值集的属性....、五、、、等

tively, i.e. if A i is a linear attribute then V,.contains all values of Ai and if A i is

换句话说，如果A i是一个线性属性，那么V。包含人工智能的所有值，如果人工智能为

tree-structured then V,.contains only the leaf values.The instance space defined by A ~ .....A,, is the cross-product of the value sets V~,..., V,,.Each instance

树形结构，然后V，。仅包含叶值。由A～定义的实例空间.....，是值集V~，的叉积，...，五，，.每个实例

in this space is characterized by an n-tuple giving an observable value for each attribute.The instance space can be thought of as consisting of a large set of simple objects, each object characterized by its properties as given by the

在这个空间中，n元组为每个属性提供了一个可观察的值。实例空间可以被认为是由一大组简单对象组成的，每个对象都由它的属性来表征

values of attributes A~ .....A,,.Such an instance space is called attribute-

属性值A~.....一、二、三。这样的实例空间称为属性-

basedJ

basedJ

Concepts can be specified on an instance space using a concept description language as described in [24].Equations relating attributes to values will be

可以使用[24]中描述的概念描述语言在实例空间上指定概念。将属性与值相关联的等式将是

called atoms, which are either elementary or compound.The possible forms of

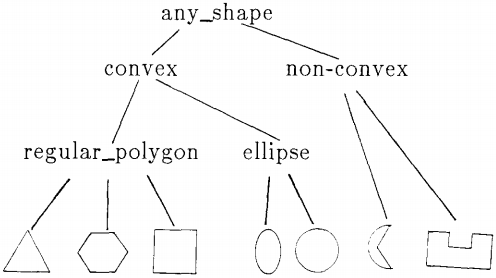
叫做原子，或者是元素或者是化合物。的可能形式

elementary atoms are as follows.

基本原子如下。

shape:

形状:



triangle

三角

hexagon square proper\_ellipse circle crescent channel

六边形正方形椭圆圆形月牙形通道

Fig. 1.The tree-structured attribute shape.

图1。树状结构属性形状。

~ A richer class of instance spaces, called structured instance spaces, can be defined by allowing each instance to include several objects, each with its own attributes, and allowing binary relations that define a structure between objects (see e.g. [10,24]).The technique defined below for quantifying inductive bias and evaluating learning performance is extended to such spaces in [15].

~一类更丰富的实例空间，称为结构化实例空间，可以通过允许每个实例包含多个对象(每个对象都有自己的属性)，并允许定义对象之间结构的二元关系来定义(例如，参见[10，24)。下面定义的量化归纳偏差和评估学习成绩的技术在[扩展到了这样的空间15]。

182

182

- For tree-structured attributes:

-对于树形结构属性:

attribute = value,

属性=值，

e.g. color = red, shape = regular\_polygon.

例如，颜色=红色，形状=正多边形。

-For linear attributes:

-对于线性属性:

value~ <~ attribute <~ value 2 ,

值~ <~属性< ~值2，

D.ttAUSSLER

D.泰特斯勒

e.g. 5 ~< size <~ 12.Strict inequalities are also permitted, as well as intervals unbounded on one side.Atoms such as 5 ~< size <~ 5 are abbreviated as size = 5.

例如5-12号。严格的不等式也是允许的，而且区间在一边是无界的。原子如~ < size < ~ 5缩写为size = 5。

Compound atoms can take the following forms.-For tree-structured attributes:

复合原子可以有以下几种形式。-对于树形结构属性:

attribute = value~ or value 2 or ...or value k ,

属性=值~或值2或...或者k值，

e.g. shape = square or circle.

例如，形状=正方形或圆形。

- For linear attributes: any disjunction of intervals e.g. 0 ~< age ~ 21 or

-对于线性属性:间隔的任何分离，例如0～小于21岁或

age >~65.Disjunctive operators within compound atoms are called internal

年龄> 65岁。复合原子中的析取算符称为内部算符

disjunctions.

分离。

We consider the following types of concepts: (1) Pure conjunctive.Expressions are of the form

我们考虑以下类型的概念:(1)纯合取。表达式的形式如下

atorn~ and atom 2 and ...and atom~.

atorn~和原子2和...和原子。

where each atorni is an elementary atom, 1 ~< i ~ s. For example, color = red

其中每个原子是一个基本原子，例如，颜色=红色

and 5 ~ size <~ 12 is a pure conjunctive concept.

5 ~ size <~ 12是一个纯合取概念。

(2) Pure disjunctive.The same as pure conjunctive, but the atoms are

(2)纯析取。与纯合取相同，但原子是

connected by "or.'"

通过“或”连接。"

(3) lnternal disjunctive.The same as pure conjunctive, but compound atoms are allowed.For example,

(3)内部析取。与纯合取相同，但允许复合原子。例如，

(color = red or blue or yellow) and ((5 ~< size <~ 12) or (size > I(X)))

(颜色=红色、蓝色或黄色)和((5 <尺寸< ~ 12)或(尺寸> 1(X))

is an internal disjunctive concept.

是一个内部分离的概念。

(4) k-DNF.Expressions are of the form

(4)DNF。表达式的形式如下

t I or l, or ...or t s ,

我还是我，还是...或者t-s，

where each t i is a pure conjunctive concept with at most k atoms for some fixed k. For example,

其中每个t i是一个纯合取概念，对于某个固定k，它最多有k个原子。例如，

(color = red and shape = regular\_polygon)

(颜色=红色，形状=正多边形)

or (5 ~< size ~ 12 and shape = circle)

or (5 ~ <尺寸~ 12，形状=圆形)

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

183

183

is a 2-DNF concept.Within k-DNF concepts the pure conjunctive ti are called terms.

是一个双DNF概念。在k-DNF概念中，纯合取ti称为术语。

(5) k-CNF.Expressions are of the form

(5)CNF。表达式的形式如下

c~ and c 2 and ...and c.~ ,

c～和c 2和...和c.~,

where each c i is a pure disjunctive concept with at most k atoms for some fixed k. The c i are called clauses.

其中每个c i是一个纯析取概念，对于某些固定的k，它最多有k个原子。

A concept of any of the above types represents a set of instances in the

任何上述类型的概念都表示

instance space in the usual way, i.e. the concept

实例空间，即概念

shape = regular\_polygon and 5 <~ size <~ 12

形状=正多边形和5 <~尺寸< ~ 12

represents the set of all instances that have a value between 5 and 12 for the

表示所有实例的集合，这些实例的值在5和12之间

attribute size and a value for the attribute shape that is a leaf in the hierarchy

属性大小和属性形状的值，属性形状是层次结构中的一个叶

below regular\_polygon, i.e. triangle, hexagon or square.In what follows, we will not distinguish between the syntactic form of a concept (its intension) and the set of instances it represents (its extension), unless this distinction is required for clarity.We use the notation x ~ h, the phrase "x is included in h" and the phrase "h covers x" interchangeably to denote that the instance x is an instance of the concept h.

低于正多边形，即三角形、六边形或正方形。在接下来的内容中，我们将不会区分一个概念的句法形式(其内涵)和它所代表的一组实例(其外延)，除非为了清晰起见需要这种区分。我们使用符号x ~ h，短语“x包含在h中”和短语“h覆盖x”可互换地表示实例x是概念h的实例。

2.Exhausting a Version Space

2.耗尽版本空间

Let X be an instance space determined by a fixed set of attributes (each

设X是由一组固定属性(每个属性

tree-structured or linear) and let H be a hypothesis space defined on X, i.e. a

树形结构或线性结构)，并且让H是在X上定义的假设空间，即

class of concepts defined using the attributes of X. For example, H might be

用x的属性定义的一类概念。例如，H可能是

the class of pure conjunctive concepts over X. Let Q be a finite set of examples of a target concept c defined on X. The version space ofQ (w.r.t. H) is defined by Mitchell [27] as the set of all hypotheses in H that are consistent with all examples in Q.

设Q是在x上定义的目标概念c的有限个例子集。Q的版本空间(w.r.t. H)由米切尔·[27]定义为与Q中所有例子一致的H中所有假设的集合

Assume the instance space is finite.Then if the target concept c is a member of H, as new examples of c are added to Q, the version space of Q w.r.t. H shrinks until it eventually contains only the target concept c. If the target concept c is not a member of H, then as new examples of c are added to Q, the version space of Q w.r.t. H shrinks until it eventually becomes empty.We denote the fact that the version space has reached one of these two limiting states, i.e. is either empty or reduced to just the target concept, by saying that the version space is exhausted (w.r.t. c).

假设实例空间是有限的。然后，如果目标概念c是H的成员，当新的c的例子被添加到Q时，Q的版本空间缩小，直到它最终只包含目标概念c。如果目标概念c不是H的成员，那么当新的c的例子被添加到Q时，Q的版本空间缩小，直到它最终变成空的。我们表示版本空间已经达到这两种限制状态之一的事实，即要么是空的，要么减少到仅仅是目标概念，通过说版本空间被耗尽(w.r.t. c)。

Note that if the version space is reduced to one hypothesis h, but h is not the

请注意，如果版本空间缩减为一个假设h，但h不是

target concept, then the version space is not yet exhausted.This case can occur when the target concept is not a member of the hypothesis space H. In this case it is always possible to add a new example that distinguishes the

目标概念，那么版本空间还没有耗尽。当目标概念不是假设空间H的成员时，就会出现这种情况。在这种情况下，总是可以添加一个新的示例来区分

184

184

D.HAUSSLER

D.HAUSSLER

hypothesis h from the target concept.This eliminates h from the version space, leaving it empty.

来自目标概念的假设h。这从版本空间中消除了h，使其为空。

We say that the hypothesis h is more specific than the hypothesis h' if h is

我们说假设h比假设h更具体

contained in h', and that h is more general than h' if h' is contained in h. A

包含在h '中，如果h '包含在h. A中，则h比h '更一般

hypothesis h in the version space of Q is maximally specific if there is no other hypothesis h' in the version space of Q that is strictly more specific than h. Maximally general is defined similarly.Mitchell observes that by keeping track

如果在Q的版本空间中没有严格比h更具体的其他假设h’，则Q的版本空间中的假设h是最大具体的。最大一般的定义类似。米切尔观察到，通过跟踪

of only the maximally specific and the maximally general hypotheses in the

只有最具体和最普遍的假设

version space of Q (the sets S and G respectively of [27]), we can monitor this

Q的版本空间(分别是[27的集合S和G)，我们可以监视它

version space as more examples are added to Q, and, while we cannot in

版本空间作为更多的例子被添加到Q，和，虽然我们不能在

general determine when it is exhausted, we can detect when it is either empt~ or has been reduced to just one hypothesis.: If it becomes empty, then we know that the target concept is not in the hypothesis space H. If it is reduced to just one hypothesis h, then we know that if the target concept is in the hypothesis space at all, then it must be h. This is sufficient for most learning applications.

一般确定当它被用尽时，我们可以检测到它是被抢先还是被简化为仅仅一个假设。:如果它变成空的，那么我们知道目标概念不在假设空间h中。如果它被简化为一个假设h，那么我们知道如果目标概念在假设空间中，那么它一定是h。这对于大多数学习应用来说是足够的。

There are two problems with this approach in practice.The first is that it

这种方法在实践中有两个问题。首先是它

may require too many examples to reduce the version space to at most one

可能需要太多的示例来将版本空间减少到最多一个

hypothesis.Consider the simple case when X is the instance space defined by the Boolean attributes A ~ ......4,,, H is the class of pure conjunctive concepts

假设。考虑一个简单的情况，当X是由布尔属性定义的实例空间时......4、、、H是纯合取概念的类

over X and the target c H is the concept A ~ = true.It is possible lo observe

在X上，目标c H是概念A ~ =真。有可能观察到

all the 2" e positive examples of this concept in which A~ = true and (by

这一概念的所有2”e个正面例子，其中A~ =真且(由

coincidence) A~\_ = true as well, and all the 2" :;negative examples of this

巧合)A~\_ =真以及所有的2 ":；这方面的负面例子

concept in which A t = false and A ~ - false, and still not be able to distinguish

其中A t =假和A ~ -假的概念，仍然不能区分

between the target concept A~= true and the other consistent hypotheses

在目标概念A~=真和其他一致的假设之间

A,=true, and (A~ =true) and (A3=true).While it seems "unlikely" that

A，=真，以及(A ~ =真)和(A3 =真)。虽然看起来“不太可能”

such a sequence of examples will be given, if indeed the real target concept is

如果真正的目标概念是

A~= true, this intuition has not yet been quantified.Worse yet.if we use

的确，这种直觉还没有被量化。更糟。如果我们使用

real-valued attributes and atoms that denote intervals of values ~m these

表示这些值的区间的实值属性和原子

attributes, then the version space can never be reduced to at most one

属性，则版本空间永远不会减少到最多一个

hypothesis by any linitc set of examples of any target concept.

任何目标概念的一组例子的假设。

Thc other problem with the version space approach (in Mitchell's m~del) is that even ii wc monitor only the sets S and (~;.the storage needed can still become exponentially large as we build up examples before it starts to drop as

版本空间方法(在米切尔的模型中)的另一个问题是，即使是二级wc也只监视集合S和(~；。当我们在开始下降之前建立示例时，所需的存储空间仍然会呈指数级增长

the \~crsion space approaches its limit.Bundy el al.[71 have noted that if X is

空间接近极限。邦迪·艾尔。[71注意到如果X是

defined by t finite set of tree-structured attributes aud H is the class of pure

由树结构属性的有限集合定义，H是纯的类

conjunctive concepts over X, then the set S of maximally specific hypotheses in H that arc consistent with a sample Q never contains more than one hypoth-esis.This holds for our more general notion of pure conjunctive concepts as well, as is demonstrated in Section 4 below.However, Bundy et al. fail to note

X上的合取概念，那么H中与样本Q一致的最大特定假设的集合S从不包含一个以上的假设。这也适用于我们更一般的纯合取概念，如下面第4节所示。然而，Bundy等人没有注意到

that the set (;<~t" maximally general consistent hypotheses can grow exponen-

那套(；< ~t "最大程度上一般一致的假设可以发展成指数-

tially large in the number of examples.This is demonstrated as follows.

例子的数量相当多。如下所示。

:Other scarcln techniques for version spaces arc given in i22] in a more general ~cttmg.

:版本空间的其他scarcln技术在i22]中给出。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

185

185

As above, let X be defined by Boolean attributes A1,..., An and H be the

如上所述，让X由布尔属性A1定义，...安和H是

class of pure conjunctive concepts.Assume the number n of attributes is even.Let Q be the positive example

一类纯合取概念。假设属性的数量n是偶数。让Q成为正面的例子

(true, true,..., true)

(对，对，...，真)

(i.e. A 1, - • • , An all have the value true) followed by the n negative examples

(即A 1，-，An均为真值)后跟n个负例

(false, false, true, true, true,..., true, true, true),

(假，假，真，真，真，...真的，真的，真的)，

(true, true, false, false, true .....true, true, true),

(真、真、假、假、真.....真的，真的，真的)，

(irue, true, true, true, true .....true, false, false).

(真，真，真，真，真.....真、假、假)。

Assume h is a pure conjunctive hypothesis consistent with Q. In order to

假设h是与q一致的纯合取假设。为了

contain the positive example,

包含正面的例子，

(1) h must be of the form

(1) h必须为以下形式

(Ai~ = true) and (Ai2 = true) and ...and (A ik ~ - tFUe)

(Ai~ =真)和(Ai2 =真)和...和(一千英尺)

for some {Ai,,..., Ai~} C\_ {A 1 ...., An).In other words, h cannot include

对某些人来说{艾，，...，艾~ } C1....，安)。换句话说，h不能包括

an atom of the form (A~ = false).Given this restriction, to avoid containing a

形式为(A~ = false)的原子。给定此限制，为避免包含

negative example,

反面例子，

(2) h must contain the following atoms:

(2) h必须包含下列原子:

either the atom (A~ = true) or the atom (A 2 = true) and

要么是原子(A 2 =真)，要么是原子(a2 =真)

either the atom (A 3 = true) or the atom (m 4 = true) and

原子(A 3 =真)或原子(m 4 =真)和

~ither the atom (An\_ 1 = true) or the atom (A, = true).

~有原子(An\_ 1 =真)或原子(A，=真)。

The maximally general concepts that meet criteria (1) and (2) are those with the fewest atoms.It is easy to see that there are 2 ~/2 of these, all incomparable, each created by choosing one atom from each pair according to criterion (2).Hence, G has size exponential in the size of Q in this case.

满足标准(1)和(2)的最一般的概念是那些原子最少的概念。很容易看出，有2 ~ 2个这样的原子，都是无与伦比的，每一个都是根据标准(2)从每一对原子中选择一个来创造的。因此，在这种情况下，G的大小与Q的大小成指数关系。

These problems with the version space approach are overcome by incor-porating into it the probabilistic ideas of the Valiant framework [39].To overcome the first problem, we will abandon the goal of completely exhausting the version space and settle for a version space that is "probably almost exhausted" (cf. Dana Angluin's characterization of the Valiant framework as a "probably approximately correct" identification of a concept [I]).We will see below how this reduces the number of examples needed.

版本空间方法的这些问题可以通过在其中引入[39的随机思想来解决。为了克服第一个问题，我们将放弃完全耗尽版本空间的目标，而满足于一个“可能几乎耗尽”的版本空间(参见达纳·安格伦将“勇敢的框架”描述为“可能近似正确的”对概念[一号的识别)。我们将在下面看到这是如何减少所需示例数量的。

To overcome the second problem, we will simply avoid keeping track of the exact version space in any form.Instead, we will set things up so that any hypothesis from an "almost exhausted" version space will accurately approxi-mate the target concept.Thus, the strategy of keeping track of all consistent

为了克服第二个问题，我们将简单地避免跟踪任何形式的确切版本空间。相反，我们将设置一些东西，以便来自“几乎耗尽”版本空间的任何假设都能准确地逼近目标概念。因此，保持跟踪所有一致的策略

186

186

D, HAUSSLER

d，HAUSSLER

hypotheses is replaced by the simpler strategy of drawing enough examples to probably almost exhaust the version space and then finding at least one hypothesis consistent with these examples.

假设被更简单的策略所取代，即画出足够多的例子，可能几乎耗尽版本空间，然后找到至少一个与这些例子一致的假设。

We will assume that there is a fixed but arbitrary probability distribution

我们将假设有一个固定但任意的概率分布

defined on the instance space, unknown to the learner.This distribution can be as complex as it needs to be to adequately represent the real-world prob-abilities in any application domain.The complexity of the distribution will not affect the sample size bounds obtained below.

在实例空间中定义，学习者不知道。这种分布可能非常复杂，因为它需要充分地表示任何应用领域中的真实世界的可能性。分布的复杂性不会影响下面获得的样本大小范围。

As outlined in the introduction, the notion of the error of a hypothesis with

正如引言中所概述的，假设的错误概念

respect to the target concept is defined relative to this distribution: it is the

关于目标概念的定义是相对于这个分布的:它是

combined probability of all instances that are either in the hypothesis and not in the target concept or in the target concept and not in the hypothesis, i.e. the probability of drawing a random example on which the hypothesis and target concept disagree.When the error is slnall, the hypothesis and the target concept differ only by a set of instances that rarely occur, i.e. the hypothesis is a good approximation to the target concept relative to the fixed "real-world" distribution on instances.

在假设中而不在目标概念中或在目标概念中而不在假设中的所有实例的组合概率，即得出假设和目标概念不一致的随机例子的概率。当误差很小时，假设和目标概念的区别仅在于一组很少出现的实例，即假设是相对于实例上固定的“真实世界”分布的目标概念的良好近似。

The idea of "almost exhausted" can now be formalized as follows.

“几乎筋疲力尽”的概念现在可以形式化如下。

Definition 2.1.Given a hypothesis space H, a target concept c, a sequence of

定义2.1。给定一个假设空间，一个目标概念，一系列

examples Q of c, and an error tolerance e, where 0 <~ e ~< 1, the version space of Q (w.r.t. H) is e-exhausted (w.r.t. c) if it does not contain any hypothesis that has error more than e with respect to c.

c的示例Q和误差容限e，其中0 <~ e ~< 1，Q的版本空间(w.r.t. H)是e-穷尽的(w.r.t. c)，如果它不包含任何相对于c的误差大于e的假设

Note that if the instance space is finite and no instance has zero probability, then setting e = 0 in the above definition is equivalent to demanding that no hypothesis in the version space differ at all from the target concept, and thus this reduces to our original definition of an exhausted version space.Note also that since every hypothesis in an e-exhausted version space has error at most e with respect to the target concept, then any two hypotheses can have error at most 2e with respect to each other, i.e. they will agree with each other except on a set of instances that has combined probability at most 2e.Hence, when e is small and the target concept is in H, although the version space may not be reduced to a single hypothesis, it is at least reduced to a set of hypotheses that are all almost identical to each other and to the target concept (with respect to the fixed probability distribution on the instance space).

请注意，如果实例空间是有限的，并且没有实例具有零概率，那么在上面的定义中设置e = 0相当于要求版本空间中的假设与目标概念没有任何不同，因此这简化为我们对耗尽版本空间的原始定义。还要注意的是，由于e-excluded版本空间中的每个假设相对于目标概念最多具有e个错误，因此任何两个假设相对于彼此最多具有2e个错误，即它们将彼此一致，除了在具有最多2e个组合概率的一组实例上。因此，当e小而目标概念在H中时，尽管版本空间可能不会被简化为单个假设，但它至少被简化为一组假设，这些假设彼此几乎相同，并且与目标概念几乎相同(相对于实例空间上的固定概率分布)。

How may examples are required to e-exhaust a version space?As above, if we take the worst case over all possible sequences of distinct examples, then this number can be exponential or even infinite.The situation is considerably improved if we assume that the examples are drawn independently at random,

如何要求示例来耗尽版本空间？如上所述，如果我们对所有可能的不同例子的序列采取最坏的情况，那么这个数可以是指数甚至是无穷大。如果我们假设这些例子是随机独立地画出来的，情况就大大改善了，

'The idea of e-exhausting a version space is a special case of the more general idea of finding an e-net for a set of regions, introduced in [17].

电子版耗尽版本空间的想法是在[17中引入的为一组区域寻找电子网的更一般的想法的一个特例。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

187

187

and insist only that the version space be e-exhausted with high probability (hence the term "probably almost exhausted").

并且只坚持版本空间以高概率耗尽(因此术语“可能几乎耗尽”)。

Lemma 2.2 [4, 42].If the hypothesis space H is finiw, its cardinality denoted by ]HI, and Q is a sequence of m >~ 1 independent random examples (chosen according to any fixed probability distribution on the instance space) of any target concept c, then for any 0 < e < 1, the probability that the version space of Q (w.r.t. H) is not e-exhausted (w.r.t. c) is less than

引理2.2 [4，42]。如果假设空间H是有限的，其基数由]HI表示，并且Q是任何目标概念c的m >~ 1个独立随机示例(根据实例空间上的任何固定概率分布选择)的序列，则对于任何0 < e < 1，Q的版本空间未用尽(w.r.t. c)的概率小于

IHle ~m.

IHle ~m

Proof.Let h a .....h k be the hypotheses in H that have error greater than e

证据。让h a.....h k是H中误差大于e的假设

with respect to c. We fail to e-exhaust the version space w.r.t. H if and only if there is a hypothesis in this set that is consistent with all m independent random examples.Since each hypothesis has error greater than e, an individual example of c is consistent with a given h i with probability at most (1- e).Thus, m independent random examples are consistent with h i with probability at most (1 - e) m. Since the probability of a union of events is at most the sum of their individual probabilities, the probability that all m examples of c are consistent with any of the hypotheses in h~,..., hk is at most k(1 - e) m. The result follows from the fact that k~ < ]HI and (1 - e) m ~e -~m for O~ < e~ < 1 and m~>0.[]

关于c .我们不能排除版本空间w.r.t. H当且仅当在这个集合中有一个假设与所有m个独立的随机例子一致。因为每个假设的误差都大于e，所以c的单个例子与给定的h i是一致的，概率至多为(1- e)。因此，m个独立的随机例子与h . I一致，概率至多为(1 - e) m .由于事件并合的概率至多是它们各自概率的和，c的所有m个例子与h .中的任何假设一致的概率，...这一结果是由O~ < e~ < 1和m~>0时k~ < ]HI和(1-e)m ~ e ~-m这一事实得出的。[]

As a corollary of the above result, for any 8, 0 < 6 < 1, if Q has size

作为上述结果的推论，对于任何8，0 < 6 < 1，如果Q有大小

m ~ (In(1/8) + lnlHl)/e,

m ~ (In(1/8) + lnlHl)/e，

(1)

(1)

then the version space of Q is e-exhausted with probability at least 1 - 8.This follows from setting IHle -Era = 6 and solving for m. What is significant about this formula is that the number of random examples needed to e-exhaust a version space is logarithmic in the size of the underlying hypothesis space, independent of the target concept and independent of the distribution on the instance space.

那么Q的版本空间以至少1 - 8的概率耗尽。这个公式的意义在于，对一个版本空间进行e-exclude所需的随机样本的数量在基本假设空间的大小上是对数的，与目标概念无关，也与实例空间的分布无关。

Compare this to the number of queries needed to (completely) exhaust the version space using the standard (nonprobabilistic) model.By a query we mean a question of the form "is x an instance of the target concept?," where x is any instance chosen by the learning algorithm.The minimum number of queries needed in the worst case to reduce a version space over a finite hypothesis space H to at most one hypothesis is log IH[.This is achieved if there is a strategy that always cuts the version space in half with each new query [27].For fixed e and 6 this is of the same order of magnitude as the bound given in equation (1) above.

将此与使用标准(非概率)模型(完全)耗尽版本空间所需的查询数量进行比较。我们所说的查询是指“x是目标概念的一个实例吗？，，其中x是学习算法选择的任何实例。在最坏的情况下，将有限假设空间H上的版本空间减少到最多一个假设所需的最小查询数是对数IH·[。如果有一种策略总是用每个新的查询将版本空间减半，这是可以实现的[27]。对于固定的e和6，这与上面等式(1)中给出的界限数量级相同。

Returning to our example of the instance space defined by n Boolean

回到由n个布尔定义的实例空间的例子

188

188

1).HAUSSt.ER

1)。豪斯斯特。呃

attributes A~ .....A,,, if H is the hypothesis space of all pure conjnnctive

属性A~.....如果H是所有纯连通的假设空间

concepts over this instance space, then [HI = 3" (for each attribute A we can

这个实例空间上的概念，那么[HI = 3”(对于每个属性A，我们可以

include the atom A = true, the atom A - false or neither)~ hence the version

包括原子A =真，原子A -假或两者都没有)~因此版本

space wilt be e-exhausted with probability 1 - ~ after

空间将耗尽，概率为1 - ~之后

(ln(l/fi) + n In 3)/e

(ln(l/fi) + n In 3)/e

independent random examples, regardless of the underlying distribution gov-erning the generation of these examples.Note that the number of examples required grows only linearly in the number n of attributes, instead of exponen-tially in n as it does for completely exhausting the version space.

独立的随机例子，不管潜在的分布——宁儿这些例子的产生。请注意，所需的示例数量仅在属性数量n中线性增长，而不是像完全耗尽版本空间那样在n中指数增长。

Upper estimates on the number of examples needed to e-exhaust a version space that are derived by the above method are still very crude, and for the case of infinite hypothesis spaces, such as the set of intervals on the real line, the method does not even apply.We remedy this in the next section.

对通过上述方法导出的版本空间进行e-exclude所需的示例数量的上限估计仍然非常粗略，并且对于无限假设空间的情况，例如实线上的间隔集，该方法甚至不适用。我们将在下一节对此进行补救。

3.The Growth Function and the Vapnik-Chervonenkis Dimension

3.增长函数与瓦普尼克-切尔沃涅斯基维数

We use the following notions from [42, 43] (see also [9]).

我们使用[42，43中的以下概念(另见[9)。

Definition 3.1.Let X be an instance space and H be a hypothesis space defined on X. Let 1 be a finite set of instances in X. For a given hypothesis h ~ H, label I so that it becomes a sample of h, i.e. label all the instances of I included in h with "+'" and the others with .... ".This labeling partitions I into a set of positive instances and a set of negative instances.This partition is called the

定义3.1。设X是一个实例空间，H是在X上定义的一个假设空间。设1是在X上的一个有限的实例集。对于给定的假设h ~ H，标记1，使其成为H的一个样本，即，用“+”标记包含在H中的所有实例，用“+”标记其他实例…。”。这种标记将I划分为一组正实例和一组负实例。这个分区被称为

dichotomy of I induced by h. I1H(I ) denotes the set of all dichotomies of 1

由1H(I)诱导的I的二分法表示1的所有二分法的集合

induced by hypotheses in H, i.e. the set of all ways the instances in 1 can be

由H中的假设诱导，即1中的实例可以被

labeled with "+'" and "-" so as to be consistent with at least one hypothesis in

用“+”和“-”标记，以便与中的至少一个假设一致

H.For any integer m, l<~m~lX[, HH(m)=maxllll4(I)t over all sets of

H.对于任意整数m，l<~m~lX[，HH(m)=最大11l 4(I)t

instances I C\_ X such that 111 = m. Hence, lI.(m) is the maximum number of

实例1C \_ X使得111 = m。因此，lI。(m)是最大数量

dichotomies induced by hypotheses in H on any set of m instances.As in [42 I,

在任意m个实例的集合上，由H中的假设引起的二分法。就像在[42 I，

we call II.(m) the growth function of H.

我们称之为二。(1)人的生长函数

As an example, let X be the instance space defined by the tree-structured

作为一个例子，让X是由树形结构定义的实例空间

attribute shape given in Fig. 1 and let H be the hypothesis space of all pure

图1中给出的属性形状，假设H是所有纯的假设空间

conjunctive concepts on X. Since X is defined by a single tree-structured

因为X是由一个单一的树形结构定义的

attribute, any conjunction in H can be reduced to a single atom, and hence the

属性，H中的任何连接都可以简化为单个原子，因此

hypotheses in H are given by the nodes in the hierarchy depicted in Fig. 1.

H中的假设由图1中描述的层级中的节点给出。

Let 1 = {tri, sq, cir) be a set of three instances in X, where tri is an instance

设1 = {tri，sq，cir)是X中三个实例的集合，其中tri是一个实例

with shape = triangle, sq an instance with shape = square, and cir an instance with shape = circle.Then the hypothesis shape = regular\_polygon induces the dichotomy {(tri, +), (sq, +), (cir,-)} of I. The hypothesis shape = ellipse induces the complementary dichotomy {(tri,-), (sq,-), (cir, +)} of I. The

对于shape =三角形，sq是shape =正方形的实例，cir是shape =圆形的实例。然后假设形状=正多边形导致了I的二分法{(三，+)，(平方，+)，(圆，-)}。假设形状=椭圆导致了I的互补二分法{(三，-)，(平方，-)，(圆，+)}

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

189

189

hypotheses shape = triangle, shape = square, shape = convex and shape =

假设形状=三角形，形状=正方形，形状=凸形，形状=

non\_convex induce four more distinct dichotomies on I, for a total of six

非凸在I上诱导了四个更明显二分性，总共六个

dichotomies.However, there is no hypothesis that induces the dichotomy

二分法。然而，没有一个假设会导致这种二分法

{(tri, +), (sq, -), (cir, +)} of I. Because the least common ancestor of triangle and circle in the shape hierarchy is convex, which already includes square, the

因为形状层次中三角形和圆的最少公共祖先是凸的，它已经包括了正方形，所以

concept description language of H cannot represent a hypothesis that includes triangles and circles but not squares.The same is true for the dichotomy

H的概念描述语言不能代表包括三角形和圆形但不包括正方形的假设。二分法也是如此

{(tri, -), (sq, +), (cir, +)}.Hence, in this case II1~t(I)1 = 6.This implies that

{(tri，-)，(sq，+)，(cir，+)}。因此，在这种情况下，II1~t(I)1 = 6。这意味着

H~t(3)~>6, since H~(m) is the maximum of ]H~(I)] over all sets I of m

H~t(3)~>6，因为在m的所有集合I中，H(m)是[H(I)]的最大值

instances.

实例。

In fact H~(3)= 6 in this case, since it is easily verified that I/7,,(I)1 ~< 6 for

事实上，在这种情况下，H3 = 6，因为很容易验证I/7，(I)1 ~< 6

any set I of 3 distinct instances whenever X is defined by a single tree-

每当X由一棵树定义时，任何一组3个不同的实例

structured attribute and the hypothesis space//is pure conjunctive.Ultimate-ly, this follows from the fact that for any 3 leaves of a tree, 2 of them always have a least common ancestor that is either equal to or a descendant of the least common ancestor of all 3 leaves.Since not all of the 8 possible dichotomies of 3 instances can be expressed, this represents a kind of bias inherent in the hypothesis space H, which may be attributed to its restricted concept description language.This bias is not evident when we consider sets I

结构化属性和假设空间//是纯合取的。最终，这是基于这样一个事实:对于一棵树的任何3片叶子，它们中的2片总是有一个最不共同的祖先，它等于或者是所有3片叶子的最不共同的祖先的后代。由于3个实例的8个可能的二分法不能全部被表达，这代表了假设空间H中固有的一种偏见，这可能归因于其受限的概念描述语言。当我们考虑集合I时，这种偏见并不明显

containing only 2 instances.Even in the shape example, for any such set all 4

仅包含两个实例。即使在形状示例中，对于任何这样的集合，所有4

dichotomies are induced by hypotheses in H. Hence, //,(2)=4=2 I~1, its

二分法是由h中的假设引起的。因此，//，(2)=4=2 I~1，其

maximum possible value.

最大可能值。

Definition 3.2.Let I be a set of instances in X. If H induces all possible 2 Izl

定义3.2。设我是x中的一组实例。如果H诱导所有可能的2个元素

dichotomies of I, then we say that H shatters I. The Vapnik-Chervonenkis

我的二分法，那么我们说他粉碎了我

dimension of H, denoted VCdim(H), is the cardinality of the largest finite

H的维数，表示为VCdim(H)，是最大有限的基数

subset I of X that is shattered by H, or equivalently, the largest m such

被H粉碎的X的子集I，或者相当于最大的m

that [I~(m)= 2 m. If arbitrarily large subsets of X can be shattered, then

[I(m)= 2m。如果任意大的X子集可以被粉碎，那么

VCdim(H) = :c.

VCdim(H)= c

Continuing with the example given above, since Hn(2) = 4 and Hn(3) = 6 <

继续上面给出的例子，因为Hn(2) = 4且Hn(3) = 6 <

8, VCdim(H)= 2.(Note that by the definition of II n, if HH(m,~ ) < 2 m'' then

8，VCdim(H)= 2。(注意，根据II n的定义，如果HH(m，~ ) < 2 m”，则

II~(m) < 2 m for all m ~> m0, ) In general, VCdim(H) ~< 2 whenever the instance

II(m)< 2m(对于所有m ~> m0)，一般来说，只要实例

space X is defined by a single tree-structured attribute and the hypothesis space H is pure conjunctive.In a similar manner, it is easily verified that whenever X

空间X由单个树结构属性定义，假设空间H是纯合取的。以类似的方式，很容易验证每当X

is defined by a single linear attribute, say size, and the hypothesis space H is

由单个线性属性定义，比如大小，假设空间H是

pure conjunctive, then VCdim(H)~<2 as well, In this case, the hypothesis space H can be represented by all possible size intervals.For any 3 instances with distinct sizes x < y < z, there is no size interval that includes the instances with sizes x and z without also including the instance with size y. Thus no set

纯合取，然后VCdim(H)~<2，在这种情况下，假设空间H可以用所有可能的大小区间来表示。对于具有不同大小x < y < z的任何3个实例，不存在包含大小为x和z的实例而不包含大小为y的实例的大小间隔。因此，没有设置

I C\_ X of cardinality 3 is shattered by H. Note that this holds even when size is

基数为3的元素被h打破了。请注意，即使大小为

real-valued, and hence the cardinalities of X and H are infinite.Note also that in the linear case H~(3) is 7 instead of 6.

实值的，因此X和H的基数是无限的。还要注意，在线性情况下，H3是7而不是6。

The following result, derived from the pioneering work in [42, 43], is a

下面的结果，来源于[的开创性工作42，43]，是一个

190

190

D.HAUSSLER

D.HAUSSLER

natural analogue to Lemma 2.2 of the previous section.It relates the growth function lift(m) to the number of examples required to e-exhaust a version space with respect to H.

与上一节的引理2.2自然相似。它将增长函数提升(m)与相对于h对版本空间进行e-exhaust所需的示例数相关联

Theorem 3.3 ~ ([5, Theorem A2.4].See also [17]).If H is a hypothesis ,space and Q is a sequence of m >~ 1 independent random examples (chosen according to any fixed probability distribution on the instance space) of any target concept c, then for any 0 < e < l, the probability that the version space of Q (w.r.t. tt) is not e-exhausted (w.r.t. c) is less than

定理3.3([5，定理2.4)。另见[17)。如果H是一个假设，空间和Q是一个序列的m >~ 1个独立的随机例子(根据实例空间上的任何固定概率分布选择)的任何目标概念c，那么对于任何0 < e < l，Q的版本空间(w.r.t. tt)不是e-穷尽(w.r.t. c)的概率小于

21I,(2m)2 ~;2.

21I，(2m)2 ~；2.

The following bounds on the growth function in terms of VCdim(H) are

以下是关于增长函数的极限

given in [5, Proposition A2.5] (also derived from [42]).

在[5中给出，命题A2.5](也来自[42)。

Lemma 3.4.If VCdim(H) = d and m ~ d >~ 1, then Ht;(m ) <~ (era~d) ~, where e is the base of the natural logarithm.

引理3.4。如果VCdim(H) = d且m ~ d > ~，则Ht；(m ) <~ (era~d) ~，其中e是自然对数的基数。

As in the previous section, using Lemma 3.4 we can set the value given in

如前一节所述，使用引理3.4，我们可以设置中给出的值

Theorem 3.3 to 6 and "solve" for m. From the calculation given in [5, Lemma A2.6] we have the following:

定理3.3到6和m的“求解”。从[5，引理A2.6]中给出的计算，我们有如下:

Corollary 3.5.ff the sample Q has size at least

推论3.5。如果样本Q至少有大小

(41og(2/6) + 8 VCdim(H) log(13/e))/e,

(410g(2/6)+8 VCDim(H)对数(13/e))/e，

then the version space of Q (w.r.t. H) is ~-exhausted with probability at least 1-6.

那么Q的版本空间以至少1-6的概率被耗尽。

Let us compare these results to the analogous results from the previous

让我们将这些结果与之前的类似结果进行比较

section.Assume the hypothesis space H is finite and VCdim(H) = d. Hence, there exists a set I of d distinct instances that is shattered by H. Since this requires 2 d distinct hypotheses, IHI ~> 2 d. Therefore d =VCdim(H) ~< loglH I whenever H is finite.In the above example of pure conjunctive hypotheses on a single tree-structured or linear attribute VCdim(H)~<2, but loglH I can be arbitrarily large.This shows that in many cases VCdim(H) is much less than loglH 1.This often happens when the hypothesis space has some special structure that weakens its "power of expression" and thereby holds its growth function down.In these cases Corollary 3.5 can be significantly better than the corollary to Lemma 2.2 given in equation (1), despite the larger constants and the additional log(13/e) factor.

部分。假设假设空间H是有限的，并且VCdim(H) = d。因此，存在一组由d个不同的实例组成的集合，这些实例被H破坏。由于这需要2 d个不同的假设，IHI ~> 2 d。因此，只要H是有限的，d = VCDim(H)~ < logLH 1。在上面关于单个树结构或线性属性的纯合取假设的例子中，VCdim(H)~<2，但loglH I可以任意大。这表明，在许多情况下，VCdim(H)比loglH 1小得多。当假设空间有某种特殊结构削弱其“表达能力”，从而抑制其增长功能时，这种情况经常发生。在这些情况下，推论3.5可以明显好于方程(1)中给定的引理2.2的推论，尽管有较大的常数和附加的对数(13/e)因子。

~ Here and in subsequent results we are suppressing some additional measurability assumptions required in the general form of the theorem since they will not be relevant in our intended applications (see [5, Appendix]).

~在这里和随后的结果中，我们抑制了定理一般形式中要求的一些额外的可测性假设，因为它们与我们的预期应用无关(见[5，附录)。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

191

191

On the other hand, if H is finite and VCdim(H) is not significantly smaller than log[HI, then instead of using the bound on HH(m ) given in Lemma 3.4, we can simply use the bound Iltt(m ) <~ [H I. This follows from the fact that each dichotomy must be induced by a distinct hypothesis in H. Using this bound, setting the value given in Theorem 3.3 to 6 and solving for m gives results similar to those given in equation (1), but with slightly higher constants.

另一方面，如果H是有限的，并且VCdim(H)不明显小于对数[高，那么我们可以不使用引理3.4中给出的关于HH(m)的界限，而简单地使用界限Iltt(m)< ~[高。这是因为每个二分法必须由H中的不同假设引起。使用该界限，将定理3.3中给出的值设置为6，并且求解m给出的结果类似于等式(1)中给出的结果，但是具有稍高的常数。

Because it reflects limitations on the power of discrimination and expression inherent in the hypothesis space H, the growth function Flu(m ) is a natural way to quantify the bias of learning algorithms that use H. It is also a useful measure of bias.Theorem 3.3 provides a direct way to use this measure of bias to determine how fast a version space with respect to H shrinks, in a probabilistic sense.In subsequent sections we will see how this result translates into performance bounds on learning algorithms that use the hypothesis space H.

因为生长函数Flu(m)反映了假设空间H中固有的辨别和表达能力的限制，所以它是量化使用H的学习算法的偏差的自然方法。它也是偏差的有用度量。定理3.3提供了一种直接的方法来使用这种偏差的度量来确定版本空间相对于H在概率意义上收缩的速度。在接下来的章节中，我们将看到这个结果如何转化为使用假设空间的学习算法的性能界限

Lemma 3.4 shows that FlH(m) grows as 2 m until m reaches a critical value d = VCdim(H), and thereafter grows polynomially in m, with exponent at most d. Beyond this critical value, the polynomial growth function llH(m ) is rapidly dominated by the negative exponential 2 ..../2 in the formula of Theorem 3.3.Because of this, many useful learning performance bounds can be obtained directly from VCdim(H), without considering other details of the growth function.In some cases this is also true of loglH [, which we have seen is an upper bound on VCdim(H).Hence, these values are also useful measures of bias.

引理3.4表明，FlH(m)以2 m增长，直到m达到临界值d = VCdim(H)，然后以m多项式增长，指数至多为d。超过该临界值，多项式增长函数llH(m)迅速由负指数2支配..../2在定理3.3的公式中。正因为如此，许多有用的学习性能界限可以直接从VCdim(H)中获得，而无需考虑生长函数的其他细节。在某些情况下，对数[也是如此，我们已经看到它是VCdim(H)的一个上界。因此，这些值也是衡量偏差的有用指标。

We now give bounds on the growth function and VC dimension of each of the more general concept classes introduced in Section 1.These results are derived in part from results in [23, 44].The reader anxious to forge ahead to learning applications can safely skip the proof of the following theorem without loss of continuity.

我们现在给出第1节中介绍的每个更一般的概念类的增长函数和虚函数维数的界限。这些结果部分来自[23，44]的结果。急于学习应用程序的读者可以安全地跳过下面定理的证明，而不会丧失连续性。

Theorem 3.6.Let X be an instance space defined by n >~ 1 attributes, each

定理3.6。假设X是由n > 1个属性定义的实例空间，每个属性

tree-structured or linear.

树形结构或线形。

(i) If H is the hypothesis space of all pure conjunctive concepts on X, then

如果H是X上所有纯合取概念的假设空间，那么

and

和

n ~< VCdim(H) ~< 2n

n ~< VCdim(H) ~< 2n

Fl,(m) <~ ( em/n) 2" for all m >~ 2n .

Fl，(m) <~ ( em/n) 2 "适用于所有m >~ 2n。

(ii) If H is the hypothesis space of

如果H是的假设空间

(a) all pure conjunctive concepts on X that contain at most s atoms, (b) all pure disjunctive concepts on X that contain at most s atoms, or (c) all internal disjunctive concepts with at most s occurrences of tree-

(a)X上最多包含s个原子的所有纯合取概念，(b)X上最多包含s个原子的所有纯析取概念，或(c)最多出现s个树的所有内部析取概念-

structured attribute values or linear attribute value ranges in all compound atoms combined, then

所有复合原子的结构化属性值或线性属性值范围的总和，然后

192

192

and

和

ll.(m) <~ n'm ~ for all m >~ 2,

ll。(m) <~ n'm ~对于所有m >~ 2，

VCdim(H) ~< 4s log(4s x/~) ,

VCdim(H) ~< 4s log(4s x/~)，

s[Iog(n/s)] ~<VCdim(H) for s <~ n .

s[日志(n/s)] ~<VCdim(H)代表s <~ n。

D.HAUSSLER

D.HAUSSLER

(iii) If H is the hypothesis space of all k-DNF concepts on X with at most s

(3)如果H是X上所有k-DNF概念的假设空间，至多为s

terms (or k-CNF concepts with at most s clauses), then

术语(或最多含s个子句的k-CNF概念)，然后

and

和

lltt(m ) <~ nkSm 21c" jbr all m >~ 2,

lltt(m ) <~ nkSm 21c" jbr所有m >~ 2，

VCdim(H) ~< 4ks log(4ks x/-~),

VCdim(H) ~< 4ks日志(4ks x/-~)，

ks log ~<VCdim(H) for k<~n and s<-

k < n和s<-的ks log ~<VCdim(H)

Proof.The proof of this result is given in a series of lemmas.

证据。这个结果的证明在一系列引理中给出。

Lemma 3.7 [44].If X is an instance space defined by n linear attributes

引理3.7 [44]。如果X是由n个线性属性定义的实例空间

A 1 .....A,, and H is the set of pure conjunctive concepts over X, then

A 1.....，和H是X上的纯合取概念的集合，那么

VCdim(H) ~< 2n.

VCdim(H) ~< 2n。

Proof.Recall that instances in X are represented as n-tuples of values over

证据。回想一下，X中的实例被表示为

A i ...., A,,.Let I be a subset of X of cardinality 2n + 1.For each i. 1 ~< i <~ n.

人工智能....，A，，.让我是基数2n + 1的X的子集。对于每个i. 1 ~< i <~ n

choose a member of I that has the largest value for the attribute A i among all

在所有成员中，选择一个对属性A具有最大值的成员

members of I, and a member of I that has the smallest value for the attribute

I的成员，以及对该属性具有最小值的I的成员

A i among all members of I. Let S be the set of all members of I that are

我的所有成员中的一个。让我们成为我的所有成员的集合

chosen.Elements of S will be called extreme members of I (see Fig. 2).Clearly

被选中。S的元素将被称为I的极端成员(见图2)。显然地

I can have at most 2n extreme members, and thus I has at least one element

我最多可以有2n个极端成员，因此我至少有一个元素

that is not extreme.Furthermore.since the hypotheses of H are cross-products

这并不极端。此外。因为H的假设是交叉积

of intervals of values of the attributes A l .....A,,, it is easily verified that any

属性的值的区间.....很容易验证任何

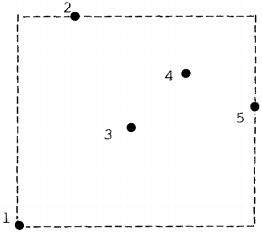


Fig. 2.Case n = 2, cardinality of 1 is 5.Extreme points are 1~ 2 and 5.Any pure conjunctive hypothesis that contains all extreme points must contain the dashed region, hence the points 3 and 4.

图2。情况n = 2，基数1是5。极值点是1~ 2和5。任何包含所有极值点的纯合取假设都必须包含虚线区域，即点3和点4。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

193

193

hypothesis that includes all extreme members of 1 includes all members of I. Hence, if we form a dichotomy of I by labeling all extreme members of I as positive instances and all other members of I as negative instances, then no hypothesis in H is consistent with this labeling.Thus I is not shattered by H. It follows that no subset of cardinality 2n + 1 can be shattered by H, showing that the VC dimension of H is at most 2n.[]

包含1的所有极端成员的假设包含了1的所有成员。因此，如果我们通过将1的所有极端成员标记为正实例并将1的所有其他成员标记为负实例来形成1的二分法，那么在H中没有假设与这种标记一致。由此可见，基数2n + 1的子集都不能被H粉碎，表明H的VC维数至多为2n。[]

It is shown in [44] that VCdim(H)=2n when H is the space of pure

在[44]中表明，当H是纯空间时，VCdim(H)=2n

conjunctive concepts on n real-valued linear attributes.Hence, this upper bound cannot be improved.

n个实值线性属性上的合取概念。因此，这一上限无法提高。

Corollary 3.8.If X & an instance space defined by n attributes, each linear or

推论3.8。如果X &一个实例空间由n个属性定义，每个属性为线性或

tree-structured, and H is the set of pure conjunctive concepts over X, then

树型结构，H是X上的纯合取概念的集合，那么

VCdim(H) ~< 2n.

VCdim(H) ~< 2n。

Proof.The observed values (leaves) of any tree-structured attribute can be ordered in such a way that any higher-level value (internal node) represents an interval of observed values.Hence, H is a subset of some H', where H' is the class of pure conjunctive concepts over some set of n linear attributes.Since the VC dimension of a subclass of concepts is never more than the VC dimension of the class itself, by the previous lemma, this implies that the VC dimension of H is at most 2n.[]

证据。任何树形结构属性的观察值(叶子)都可以以这样的方式排序，即任何更高级别的值(内部节点)代表观察值的间隔。因此，H是一些H’的子集，其中H’是在一组n个线性属性上的纯合取概念的类。由于概念子类的虚函数维数永远不会超过类本身的虚函数维数，根据前面的引理，这意味着H的虚函数维数至多为2n。[]

This establishes the upper bound on VCdim(H) in this case.For the lower bound we will assume that the instance space X is defined by n Boolean attributes A ~,..., A n. Since we are assuming throughout the paper that each attribute has at least two observable values, we can make this assumption without loss of generality.Let I be the set of instances

在这种情况下，这就建立了VCdim(H)的上限。对于下界，我们假设实例空间X由n个布尔属性A ~，...因为我们在整篇论文中都假设每个属性至少有两个可观察值，所以我们可以在不失一般性的情况下做出这个假设。让我成为一组实例

X 1 ~ - (false, true, true,..., true) ,

X 1 ~ -(假，真，真，...，真的)，

x 2 = (true, false, true ...., true) ,

x 2 =(真、假、真....，真的)，

x 3 = (true, true, false .....true),

x 3 =(真、真、假.....真的)，

~,, = (true, true, true .....false).

~，，=(真，真，真.....false)。

It is easily verified that for any {i~, i 2 ...., ik} C\_ {1 .....n}, the dichotomy of

很容易验证对于任何{i~，I ^ 2....，ik} C\_ {1.....n}的二分法

1 in which all instances xi,, xi2,..., X~k are labeled "-" and all others are

1，其中所有的实例xi，xi2，...，X~k标记为“-”，其他所有标记为

labeled "+" is induced by the pure conjunctive concept

标有“+”的词是由纯合取概念引起的

(Ai~ = true) and (Ai2 = true) and ...and (Aik = true) .

(Ai~ =真)和(Ai2 =真)和...和(Aik =真)。

Hence, I is shattered by H and thus VCdim(H) ~> n.

因此，我被氢击碎了，因此，我被氮击碎了

Now note that by Lemma 3.4, if the VC dimension of H is 2n, then

现在注意，根据引理3.4，如果H的VC维数是2n，那么

194

194

I3.HAUSSLER

I3。HAUSSLER

Hn(m ) <~ ( era~n) ~'~ for all rn ~> 2n.This also holds for smaller VC dimensions, since for all k, any hypothesis space with VC dimension less than k is contained in a hypothesis space with VC dimension equal to k. This, combined with the above results, establishes Theorem 3.6(i).

Hn(m ) <~ ( era~n) ~'~所有rn ~> 2n。这也适用于较小的虚函数维数，因为对于所有k，虚函数维数小于k的任何假设空间都包含在虚函数维数等于k的假设空间中。

Lemma 3.9.If H is the hypothesis space of

引理3.9。如果H是的假设空间

(a) all pure conjunctive concepts on X that contain at most s atoms,

(a)X上最多包含s个原子的所有纯合取概念，

(b) all pure disjunctive concepts on X that contain at rnost s atorns, or

(b)在X上的所有纯析取概念，包含在最后面，或

(c) all internal disjunctive concepts with at rnost s occurrences o[ tree-

(c)所有[树出现次数最多的内部析取概念-

structured attribute values or linear attribute value ranges in all compound atorns cornbined, then

组合的所有复合属性中的结构化属性值或线性属性值范围，然后

• 2s Hr~(m ) <~ n'rn for all rn >~ 2.

所有rn的2s Hr(m)< ~ n ' rn > ~ 2。

Proof.Let I be a set of rn >~2 instances in X. We first claim that for any

证据。让我成为x中的一组rn > 2个实例。我们首先声明对于任何

(elementary) atom involving a linear attribute A, there are at most ( ~' ) + rn +

(初等)原子涉及线性属性A，最多有(~' ) + rn +

1 ways this atom can induce a dichotomy on the set 1 by partitioning it into

这个原子可以通过把集合1分成几个部分来引发一个二分法

positive instances whose values on A satisfy the atom and negative instances whose values do not.To see this, order the elements of I as x~ .....x,,, such that for each i, 1 ~< i < m, the value of A on Xg is less than or equal to the value of A on xi+ 1.Since each atom involving the attribute A specifies an interval of values of A, each such atom induces a dichotomy on 1 by making positive some interval of instances xi,.....r i, where 1 <~i<~j<~ m, and making the rest negative, or by making all instances negative.This gives at most ('~') + rn + 1 dichotomies.

其值在A上满足原子的正实例和其值不满足原子的负实例。要看到这一点，将I的元素排序为x~.....这样，对于每个I，1 ~< i < m，Xg上的值A小于或等于xi+ 1上的值A。由于涉及属性A的每个原子指定了值为A的区间，因此每个这样的原子通过使实例xi的某个区间为正而在1上诱导二分法。....r i，其中1 <~i<~j<~ m，并使其余部分为负，或使所有实例为负。这至多给出(' ~') + rn + 1个二分法。

As in the previous lemma, since the leaves of any tree can be ordered so that the set of leaves of the subtree defined by any internal node forms an interval of this ordering, this result also holds for tree-structured attributes.(A tighter bound of at most 2rn dichotomies can also be derived for the tree-structured

如同前面的引理一样，由于任何树的叶子都可以被排序，使得由任何内部节点定义的子树的叶子集合形成这种排序的间隔，这一结果也适用于树结构的属性。(对于树形结构，也可以导出至多2rn个二分法的更紧的界限

case.)

凯斯。)

It is easily verified that (!])+ rn + 1 ~< rn 2 for all rn ~>2.Hence~ we have

很容易验证(！])+ rn + 1 ~< rn 2所有rn ~>2。因此~我们有

shown that for each attribute A, the atoms involving A are capable of inducing at most rn dichotomies on a set I of rn instances.The dichotomy induced by a hypothesis formed by the conjunction or disjunction of a set of atoms is entirely determined by the dichotomies induced by the individual atoms.Since it does not change the hypothesis to include the same atom more than once, we can assume without loss of generality that each hypothesis h ~ H contains exactly s atoms.Since for each of the s atoms in the hypothesis h there are n ways to assign it an attribute and at most rn 2 ways to choose the dichotomy induced by its value range given its assigned attribute, this gives a bound of (nrn2) ~ = n~rn z~ on the total number of distinct dichotomies induced by H on I. Hence, H~(m)<~ n2m z" in cases (a) and (b).

表明对于每个属性A，包含A的原子最多能够在一组rn实例上诱导rn二分法。由一组原子的结合或分离形成的假设所引发的二分法完全由单个原子所引发的二分法决定。因为它不会改变假设，把同一个原子包含一次以上，我们可以不失一般性地假设，每个假设h ~ H恰好包含s个原子。因为对于假设H中的每个s原子，有n种方法给它分配一个属性，并且在给定给定给定属性的情况下，最多有2种方法选择由它的值域引起的二分法，这给出了由I上的H引起的不同二分法的总数的(nrn2) ~ = n~rn z~的一个界限。因此，在情况(a)和(b)中，H~(m)<~ n2m z”。

Clearly the same argument works in case (c) for internal disjunctive con-

显然，同样的论点适用于情况(c)的内部析取条件-

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

195

195

cepts.Once we have assigned attributes to elementary atoms, we can collect all the elementary atoms that share a common attribute together to form com-pound atoms and then form the conjunction of these.Every internal disjunc-tive concept can be formed in this way.The dichotomy it induces is determined by the dichotomies of the elementary atoms and the way attributes are assigned to them.[]

cepts。一旦我们给基本原子赋予了属性，我们就可以把所有具有相同属性的基本原子集合在一起，形成复合原子，然后形成这些原子的结合。每一个内部矛盾的概念都可以这样形成。它引起的二分法是由基本原子的二分法和属性分配给它们的方式决定的。[]

Lemma 3.10.If H is the hypothesis space of all k-DNF concepts on X with at most s terms (or k-CNF concepts with at most s clauses), then

引理3.10。如果H是X上最多含s个项的所有k-DNF概念(或最多含s个子句的k-CNF概念)的假设空间，那么

IIH(m ) <~ n~Sm 2~s for all m >~ 2.

IIH(m ) <~ n~Sm 2~s所有m >~ 2。

Proof.By Lemma 3.9, the number of dichotomies induced by a single term of a k-DNF is at most nkm 2k.As above, we can assume that the k-DNF expression contains exactly s terms.Since the dichotomies induced by a k-DNF expression are determined by the dichotomies induced by each of its terms, there are at most (nkm-~k) s= n~'m 2~ dichotomies induced by k-DNF expres-sions with s terms.Clearly the same argument works for k-CNF.[]

证据。根据引理3.9，由k-DNF的单个项诱导的二分法的数目至多为nkm 2k。如上所述，我们可以假设k-DNF表达式正好包含s项。由于由k-DNF表达式引起的二分性是由其每个项引起的二分性决定的，所以至多有(nkm-~ k)s = n ~ m ^ 2个二分性是由带有s项的k-DNF表达式引起的。显然，同样的论点也适用于k-CNF。[]

Lemma 3.11.Assume n,s >~ 1.Then for any m > 4s log(4sv'~), nSm 2s < 2 m.

引理3.11。假设n，s > 1。然后，对于任何m > 4s日志(4sv'~)，nSm 2s < 2 m。

Proof.This is easily verified.[]

证据。这很容易验证。[]

The upper bounds in Theorem 3.6(ii) and (iii) follow directly from Lemmas 3.9-3.11.The lower bounds follow from [23, Lemma 4.6] (see also [23, Example 4 in Section 5]), which uses an example on an instance space of Boolean attributes remotely related to that given above for the lower bound in part (i).This completes the proof of Theorem 3.6.[]

定理3.6(ii)和(iii)的上限直接来自引理3.9-3.11。下限来自[23，引理4.6](另见[23，第5节中的例子4)，它使用了一个布尔属性实例空间的例子，该实例空间与上面第(I)部分中给出的下限远程相关。这就完成了定理3.6的证明。[]

As an example application of Theorem 3.6, we can now extend the result obtained in the previous section for the hypothesis space of pure conjunctive concepts over an instance space of n Boolean attributes to pure conjunctive concepts over n arbitrary tree-structured and linear attributes.Since by Theorem 3.6(i) the Vapnik-Chervonenkis dimension of this hypothesis space is at most 2n, using Corollary 3,5, after

作为定理3.6的一个示例应用，我们现在可以将上一节中获得的关于n个布尔属性的实例空间上的纯合取概念的假设空间的结果扩展到n个任意树结构和线性属性上的纯合取概念。因为根据定理3.6(i)，这个假设空间的Vapnik-Chervonenkis维数至多是2n，使用推论3，5，之后

(4 1og(2/6) + 16n log(13/e))/e,

(4 10g(2/6)+16n log(13/e))/e，

independent random examples of any target concept c, the version space w.r.t. this hypothesis space will be e-exhausted (w.r.t. c) with probability at least

任何目标概念c的独立随机示例，版本空间w.r.t .该假设空间将至少以概率e-穷尽(w.r.t. c)

1 - 6, independent of the distribution governing the generation of the ex-

1 - 6，独立于支配ex-的产生的分布

amples.

样本。

Note that this bound is not much higher than that given in Section 2 for the

请注意，该界限并不比第2节中给出的高多少

196

196

D, HAUSSLER

d，HAUSSLER

case of Boolean attributes.In particular, this bound does not depend on the

布尔属性的情况。特别是，这个界限不依赖于

size or complexity of the hierarchies of values defined for the tree-structured attributes, nor on the number of values of the linear attributes.In fact, the linear attributes can be real-valued.This is because increasing the number of values of the attributes does not increase the Vapnik-Chervonenkis ditnension of the hypothesis space beyond 2n, no matter how much it increases the size of the hypothesis space.

为树形结构属性定义的值层次的大小或复杂性，也不取决于线性属性的值的数量。事实上，线性属性可以是实值的。这是因为增加属性值的数量不会使假设空间的Vapnik-Chervonenkis ditnension增加到2n以上，无论假设空间的大小增加了多少。

Similar bounds hold for the other kinds of hypothesis spaces treated in

类似的界限适用于中处理的其他类型的假设空间

Theorem 3.6.

定理3.6。

4.The Performance of the Classical Learning Algorithm for Conjunctive Concepts

4.合取概念的经典学习算法的性能

The fact that the hypothesis space of pure conjunctive concepts is rapidly

事实上，纯合取概念的假设空间是迅速的

e-exhausted as independent random examples of any target concept are drawn tells us a good deal about the performance of learning algorithms that use this hypothesis space.Here we apply this result to analyze the performance of one of the simplest learning algorithms for pure conjunctive concepts, which wc

当任何目标概念的独立随机例子被绘制出来时，e-excluded告诉我们使用这个假设空间的学习算法的性能。这里我们应用这个结果来分析一个最简单的纯合取概念学习算法的性能

will call the classical algorithm.To analyze learning performance we will adopt

会调用经典算法。为了分析学习成绩，我们将采用

the viewpoint of Valiant [39] and ask how many random examples and how much computational effort is required for the algorithm to, with high probabili-ty, find a hypothesis that is a good approximation of the target concept.

瓦兰特·[的观点39]并询问算法需要多少随机例子和多少计算努力来以高概率找到一个与目标概念很接近的假设。

Let X be a fixed instance space defined by n attributes, each tree-structured or linear.Let Q be a sample of any concept defined on X. For simplicity, we will assume here and in what follows that the sample Q contains at least one

设X是由n个属性定义的固定实例空间，每个属性都是树形结构或线性的。让Q是定义在x上的任何概念的一个样本。为了简单起见，我们将在这里和下面假设样本Q包含至少一个

positive example.Under this assumption, for any attribute A the minimal dominating atom for A (w.r.t. Q) is defined as the most specific elementary

正面的例子。在这种假设下，对于任何属性，最小支配原子被定义为最具体的元素

atom involving the attribute A that includes all the positive examples of Q.

包含所有正例的属性A的原子

It is easily verified that this atom is always uniquely defined for tree-

很容易证明这个原子总是为树唯一定义的-

structured and linear attributes.If A is a linear attribute, the minimal

结构化和线性属性。如果A是线性属性，则最小

dominating atom for A is the atom v~<~A<~v 2, where v~ and v: are the

A的主要原子是原子v < ~ A < ~ v2，其中v~和v:是

smallest and largest values of A that occur among the positive examples.This atom is the result of applying the "closing interval rule" of [24].If A is a tree-structured attribute, the minimal dominating atom is A = v, where v is the value of the node that is the least common ancestor of all the leaf values of A that occur among the positive examples (see Fig. 3(b) for an example).This atom is the result of using the climbing tree heuristic of [24].It also corres-ponds to the "lower mark" in the attribute trees of [7].

出现在正例中的A的最小和最大值。这个原子是应用[24的“闭区间规则”的结果。如果A是树形结构属性，则最小支配原子是A = v，其中v是节点的值，该节点是在正示例中出现的所有A的叶值的最不共同的祖先(参见图3(b)的示例)。这个原子是使用[24]的爬树启发法的结果。它也对应于[7的属性树中的“较低标记”。

We can use the minimal dominating atoms to find the unique most specific pure conjunctive concept consistent with a given sample.This learning method can be traced back in various forms at least to [6].It leads to the following: 5

我们可以用最小支配原子来找到与给定样本一致的唯一的最具体的纯合取概念。这种学习方法可以以各种形式追溯到至少[6]。它导致以下结果:5

s This algorithm is typically presented as an incremental algorithm, but this causes problems with the negative examples [7, 27].Therefore we give it in a non-incremental form.

这种算法通常以增量算法的形式出现，但这会给反面例子[7，27带来问题。因此，我们以非增量的形式给出它。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

197

197

Algorithm 4.1 (Classical algorithm for learning conjunctive concepts).

算法4.1(学习合取概念的经典算法)。

Step 1.Find the minimal dominating atom for each attribute with respect to the given sample.Let the conjunction of these atoms be the hypothesis h.

第一步。相对于给定的样本，找出每个属性的最小支配原子。假设这些原子的结合是h。

Step 2.If no negative examples are included in h then return h, else report that the sample is not consistent with any pure conjunctive concept.

第二步。如果h中没有包含否定的例子，则返回h，否则报告样本与任何纯合取概念不一致。

To illustrate this algorithm, consider an instance space with attributes shape, size and shade, where shape is the tree-structured attribute given in Fig. 1, size is a real-valued linear attribute, and shade is Boolean.Let the sample Q consist of the positive examples

为了说明该算法，考虑具有形状、大小和阴影属性的实例空间，其中形状是图1中给出的树形结构属性，大小是实值线性属性，阴影是布尔型。让样本Q由正例组成

(square, 5.2, true), (triangle, 3.4, true), (square, 2.9, true),

(正方形，5.2，真)，(三角形，3.4，真)，(正方形，2.9，真)，

and the negative examples

负面的例子

(circle, 4.3, true), (channel, 5.1, true), (square, 3.7, false).

(圆圈，4.3，真)，(通道，5.1，真)，(正方形，3.7，假)。

Then the minimal dominating atoms are

那么最小的支配原子是

shape = regular\_polygon, 2.9 ~< size <~ 5.2, shade = true.

shape =正多边形，2.9 ~ < size < ~ 5.2，shade = true。

Hence, Algorithm 4.1 forms the conjunction of these as its hypothesis.No negative examples are included in this hypothesis, hence it is returned.

因此，算法4.1把这些结合起来作为它的假设。这个假设中不包含否定的例子，因此它被返回。

Lemma 4.2.If there exists a pure conjunctive concept consistent with the

引理4.2。如果存在一个与

sample, Algorithm 4.1 will find the unique maximally specific such concept, otherwise it correctly reports that the sample is not consistent with any pure conjunctive concept.

示例，算法4.1将找到唯一的最大特定的此类概念，否则它会正确地报告该示例与任何纯合取概念不一致。

Proof.Let h = a~ and a 2 and ...and a,,, where a i is the minimal dominating atom for the attribute A i w.r.t.Q. For each i, let V,.denote the set of values for A i included in the atom a i. The hypothesis h represents the set of all instances in the cross-product of V~ .....V,,.By the definition of a minimal dominating atom, for any positive example (v~,..., v,,) we must have vi ~ ~ for all i and hence this example is included in h. Thus if h does not include any negative examples, then it is consistent with Q. On the other hand, since each a~ is the unique minimal dominating atom for A~, any other atom that includes all

证据。设h = a~和a 2和。。。，其中I是属性的最小支配原子。对于每个I，让V，.表示原子A i中包含的a i的一组值。假设h表示V~的叉积中所有实例的集合.....v、、、。根据最小支配原子的定义，对于任何正例(v~，...因此，如果h不包括任何负的例子，那么它与q是一致的。另一方面，因为每个a~是A~的唯一的最小支配原子，任何其他原子包括所有

198

198

D.HAUSSLER

D.HAUSSLER

values of A i that occur in positive examples must include all values in V,..Therefore any conjunction of such atoms must represent a hypothesis that includes all examples in the cross product of V~ .....V,,.Therefore any pure conjunctive hypothesis that is consistent with the sample must contain h. It follows that if h does not include any negative examples, then h is the unique maximally specific pure conjunctive hypothesis that is consistent with Q, otherwise no pure conjunctive hypothesis is consistent with Q. (2

出现在正例中的A i值必须包括V中的所有值，..因此，这些原子的任何结合都必须代表一个假设，这个假设包含了V～的叉积中的所有例子.....v、、、。因此，任何与样本一致的纯合取假设都必须包含h。因此，如果h不包含任何否定的例子，那么h就是与Q一致的唯一的最大特定纯合取假设，否则，没有任何纯合取假设与Q一致

In order to analyze the performance of this algorithm, let us first make the following general definition.

为了分析这种算法的性能，让我们首先作出以下一般定义。

Definition 4.3.We say that a learning algorithm uses the hypothesis space H consistently if for any sequence of examples Q:

定义4.3。我们说学习算法一致地使用假设空间H，如果对于任何例子序列Q:

(1) if the version space of Q (w.r.t. H) is not empty, then the algorithm produces a hypothesis in this version space,

(1)如果Q (w.r.t. H)的版本空间不是空的，则算法在该版本空间中产生一个假设，

(2) else it indicates that no hypothesis in H is consistent with the given examples.

(2)否则，它表明在H中没有假设与给定的例子一致。

Lemma 4.2 shows that Algorithm 4.1 uses the hypothesis space of pure conjunctive concepts consistently.More sophisticated learning algorithms may handle case (2) more intelligently by "shifting the bias" when the version space becomes empty, as described in [38].However, it is still likely that they will use procedures that act as described in (1) and (2) to detect the need to shift bias, so in general, the performance of such procedures still warrants investigation.In this regard, we have the following result.

引理4.2表明算法4.1一致地使用纯合取概念的假设空间。如[38所述，当版本空间变空时，更复杂的学习算法可以通过“偏移偏差”更智能地处理情况(2)。然而，他们仍有可能使用(1)和(2)中所述的程序来检测是否需要转移偏差，因此，总体而言，这些程序的性能仍然值得调查。在这方面，我们有以下结果。

Theorem 4.4.Let H be a hypothesis space and L be a learning algorithm that uses H consistently.For any 0 < e,6 < 1, given

定理4.4。假设H是一个假设空间，L是一个始终使用H的学习算法。对于任何0 < e，6 < 1，给定

(4 log(2/6 ) + 8 Vfdim(H) log(13/e))/e

(4 log(2/6)+8 Vfdim(H)log(13/e))/e

independent random examples of any target concept c, with probability at least 1 - 6, algorithm L will either

任何目标概念c的独立随机示例，概率至少为1 - 6，算法L将

(1) produce a hypothesis in H that has error at most e with respect to c, or

(1)在H中提出一个假设，该假设相对于c至多有误差e，或

(2) indicate correctly that the target concept c is not in H. Moreover, this result holds regardless of the particular probabili O, distribution on the instance space that governs the generation of examples.

(2)正确地指出目标概念c不在h中。此外，不管控制实例生成的实例空间上的特定概率分布如何，该结果都成立。

(Note: we do not claim that whenever c ~E'H the algorithm detects this with high probability.It may instead find a good approximation to c in H.)

(注意:我们并不声称每当c ~ E’h算法以高概率检测到这一点。相反，它可以找到一个很好的近似

Proof.By Corollary 3.5, after this many examples the version space with respect to H is e-exhausted with probability 1- 6.When the version space is e-exhausted then either it is empty, in which case, since L uses H consistently,

证据。根据推论3.5，在这许多例子之后，关于H的版本空间以概率1- 6耗尽。当版本空间耗尽时，要么它是空的，在这种情况下，由于L一直使用H，

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

199

199

L halts, indicating correctly that no hypothesis in H is consistent with the given examples of c and hence c~H, or it is not empty, in which case L produces a hypothesis from this space and, because the space is e-exhausted, this hypoth-esis has error at most e. []

L停止，正确地表明在H中没有假设与c和c~H的给定例子一致，或者它不是空的，在这种情况下，L从这个空间产生一个假设，并且，因为这个空间是耗尽的，这个假设最多有错误e[]

This gives the following result on the performance of the classical learning algorithm for pure conjunctive concepts.

这给出了以下关于纯合取概念的经典学习算法的性能的结果。

Corollary 4.5.Let X be an instance space defined by n attributes, each

推论4.5。假设X是由n个属性定义的实例空间，每个属性

tree-structured or linear.For any 0 < e,6 < 1, given

树形结构或线形。对于任何0 < e，6 < 1，给定

(4 log(2/6) + 16n log(13 /e)) /e

(4 log(2/6) + 16n log(13 /e)) /e

independent random examples of any target concept c defined on X, with

在X上定义的任何目标概念的独立随机示例，带有

probability at least 1 - 6, Algorithm 4.1 will either

概率至少为1 - 6，算法4.1将

(1) produce a pure conjunctive hypothesis that has error at most e with

(1)提出一个纯粹的合取假设，该假设至多有e个错误

respect to c, or

关于c，或

(2) indicate correctly that the target concept c is not a pure conjunctive

(2)正确指出目标概念c不是一个纯粹的连词

concept.

概念。

This holds for any probability distribution on X governing the generation of examples.

这适用于X上任何控制示例生成的概率分布。

Proof.Lemma 4.2 shows that Algorithm 4.1 uses the hypothesis space H of pure conjunctive concepts on X consistently and Theorem 3.6(i) shows that VCdim(H) ~< 2n.The result then follows directly from Theorem 4.4.[]

证据。引理4.2表明算法4.1一致地使用X上纯合取概念的假设空间H，定理3.6(i)表明VCdim(H) ~< 2n。结果直接来自定理4.4。[]

This result shows that whenever the target concept is pure conjunctive, the classical learning algorithm will find a good approximation to it with high probability using relatively few random examples.The number of examples required is at most linear in the number of attributes in the instance space, almost linear in the inverse of the error parameter e, and logarithmic in the inverse of the confidence parameter & One remarkable aspect of this result is that this bound on the number of examples required does not depend on the number of values that each attribute in the instance space has.As mentioned in the previous section, this is because all pure conjunctive hypothesis spaces on n tree-structured or linear attributes have VC dimension at most 2n, regardless of the number of values per attribute.

这一结果表明，只要目标概念是纯合取的，经典学习算法就可以用相对较少的随机例子找到一个很好的近似。所需的示例数在实例空间中的属性数中最多是线性的，在误差参数e的倒数中几乎是线性的，在置信度参数的倒数中是对数的&这个结果的一个显著方面是，所需的示例数的界限不取决于实例空间中每个属性的值的数量。如前一节所述，这是因为n个树形结构或线性属性上的所有纯合取假设空间最多有2n个虚函数维数，而不考虑每个属性的值数量。

How close does this upper estimate come to the actual number of examples needed for probably approximately correct learning?How does this number of examples compare to the number of examples needed by other algorithms?In order to answer these questions, we make the following definition.

这个较高的估计值与可能近似正确的学习所需的实际示例数有多接近？与其他算法所需的示例数量相比，这些示例数量如何？为了回答这些问题，我们做出如下定义。

Definition 4.6.Let L be a learning algorithm and C be a class of target

定义4.6。假设L是一个学习算法，C是一类目标

200

200

D.HAUSSLER

D.HAUSSLER

concepts on the instance space X. For any 0< ~',6 < 1, 5c(~, 6) denotes the

对于任何0< ~ '，6 < 1，5c(~，6)表示

minimum sample size m such that for any target concept c~ C and any

最小样本量m，使得对于任何目标概念c~ C和任何

distribution on X, given m random examples of c, L produces a hypothesis L that, with probability at least 1 - 6, has error at most e. So(e, 6) is called the sample complexity of L for the target class C.

在X上的分布，给定m个c的随机例子，L产生一个假设L，其概率至少为1 - 6，误差至多为e。所以(e，6)被称为目标类c的L的样本复杂度

Theorem 4.7 [12].If C is a class of concepts with VCdim(C)~> 2, then there exists a positive constant c o such that for all learning algorithms L,

定理4.7 [12]。如果C是一类VCdim(C)~> 2的概念，则存在一个正常数C 0，使得对于所有的学习算法L，

Slci(e, 6) >>- c0(log(1/6 ) + VCdim(C))/e

Slci(e，6) >>- c0(log(1/6 ) + VCdim(C))/e

for all sufficiently small ~ positive e and 6.

对于所有足够小的~正e和6。

Corollary 4.8.There are positive constants c o and c~ such that for any instance space X defined on n attributes, each tree-structured or linear

推论4.8。有正常数c 0和c~使得对于在n个属性上定义的任何实例空间X，每个都是树形结构或线性的

c0(log(1/6) + n)/e <~ So(e, ~ 6) <~ c~(log( l /6 ) + n log(1/e))/e

c0(log(1/6) + n)/e <~ So(e，~ 6)< ~ c(log(1/6)+n log(1/e))/e

for all sufficiently small e and 6, where L & Algorithm 4.1 and C is the class of pure conjunctive concepts on X. Moreover, this lower bound holds .for any learning algorithm L.

对于所有足够小的e和6，其中L &算法4.1和C是x上的纯合取概念类。此外，这个下界成立。对于任何学习算法

Proof.Using the fact that n ~<VCdim(C) from Theorem 3.6(i), the first

证据。利用定理3.6(i)中n < VCdim(C)的事实，第一

inequality follows from Theorem 4.7.The second inequality follows from Corollary 4.5.[]

不等式源于定理4.7。第二个不等式来自推论4.5。[]

Corollary 4.8 shows that we have overestimated the sample complexity of Algorithm 4.1 by at most an O(log(1/e)) factor.More importantly, it shows that the actual sample complexity of Algorithm 4.1, whatever it is, is within an O(log(1/e)) factor of optimal for any learning algorithm for pure conjunctive concepts.

推论4.8表明，我们高估了算法4.1的样本复杂度，最多高估了0(对数(1/e))因子。更重要的是，它表明算法4.1的实际样本复杂度，无论它是什么，对于任何纯合取概念的学习算法来说，都在最优的0(log(1/e))因子之内。

Algorithm 4.1 is also extremely efficient computationally.In order to analyze the time complexity of this algorithm, for simplicity we assume that for a linear attribute the time required to compare two values is constant, and for a tree-structured attribute the time required to determine if one value is in the subtree below another value or to compute the least common ancestor of two values is constant.This will not be an unreasonable approximation in most applications.

算法4.1在计算上也非常高效。为了分析该算法的时间复杂度，为了简单起见，我们假设对于线性属性，比较两个值所需的时间是恒定的，而对于树形结构属性，确定一个值是否在另一个值之下的子树中或者计算两个值的最少公共祖先所需的时间是恒定的。在大多数应用中，这不会是不合理的近似。

Under these assumptions the time required to find the minimal dominating atom for a single attribute with respect to a sample of size m is O(m).Hence, the time for Step 1 of Algorithm 4.1 is O(nm) on an instance space with n

在这些假设下，对于大小为m的样本，找到单个属性的最小支配原子所需的时间是O(m)。因此，算法4.1的步骤1的时间是在具有n的实例空间上的0(nm)

'The result in [12] shows that this holds for all e< 1/8 and 6 ~< 1/100.

[12]的结果表明，这适用于所有e< 1/8和6 ~< 1/100。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

201

201

attributes.Step 2 takes no longer, hence the overall time for Algorithm 4.1 is

属性。步骤2不再需要，因此算法4.1的总时间是

O(nm).This is essentially optimal, since for the standard encoding of instances as n-tuples, the size of the sample itself is proportional to nm, and hence ~(nm) time is required merely to read the sample.

O(nm)。这基本上是最佳的，因为对于作为n元组的实例的标准编码，样本本身的大小与nm成比例，因此仅读取样本就需要~ nm时间。

5.Using a Greedy Heuristic to Improve Performance on Simpler Target Concepts

5.使用贪婪启发式提高简单目标概念的性能

In many AI learning situations where conjunctive concepts are used, the task is to learn relatively simple conjuncts from examples over instance spaces with many attributes.This is because without a fairly strong domain theory, it is hard to anticipate in advance which attributes each individual target concept will depend on, and so a large number of possible attributes are considered for all target concepts.This problem becomes particularly acute in large scale systems in which each new learned concept is allowed to depend on previously learned concepts (viewed as Boolean attributes), and in systems where a large "library" of attributes is derived from simple combinations of primitive attri-butes [23, 35].

在许多使用连词概念的人工智能学习环境中，任务是从具有许多属性的实例空间的例子中学习相对简单的连词。这是因为如果没有一个相当强的领域理论，很难预先预测每个单独的目标概念将依赖于哪些属性，因此对于所有的目标概念都要考虑大量可能的属性。这个问题在大规模系统中变得尤其尖锐，在大规模系统中，每个新学习的概念被允许依赖于先前学习的概念(被视为布尔属性)，在大规模系统中，属性的“库”是从原始属性的简单组合中导出的([23，35)。

It is therefore of some interest to consider the problem of learning target

因此，考虑学习目标的问题是有意义的

concepts on an instance space defined by n attributes, where each target

由n个属性定义的实例空间上的概念，其中每个目标

concept is represented by a pure conjunctive expression with at most s atoms, with s much smaller than n. If C is the class of all such target concepts, then Theorem 3.6(ii) shows that VCdim(C)~<4s log(4s~/'~).Since this bound is logarithmic in n, when s is small relative to n it is considerably better than n, which is a lower bound on the VC dimension of the class of all pure conjunctive concepts on an n-attribute instance space.In view of Theorems 4.4 and 4.7, this indicates that it may be possible to learn concepts in C with considerably fewer random examples than are required to learn arbitrary pure conjunctive concepts on an n-attribute instance space.

概念由一个最多有s个原子的纯合取表达式表示，s比n小得多。如果C是所有这类目标概念的类，那么定理3.6(ii)表明VCdim(C)~<4s log(4s~/'~)。因为这个界在n中是对数的，当s相对于n小时，它比n好得多，n是n-属性实例空间上所有纯合取概念类的VC维的下界。根据定理4.4和4.7，这表明在C语言中学习概念可能比在n属性实例空间中学习任意纯合取概念所需的随机例子少得多。

This is indeed the case.Instead of using the classical algorithm, which finds the most specific conjunct that is consistent with the sample, consider an

事实的确如此。不要使用经典算法来寻找与样本一致的最具体的结合点，而是考虑

algorithm that finds the simplest conjunct, i.e. the conjunct with the least

找到最简单的合取，即最少的合取的算法

number of atoms, that is consistent with the sample.For now, let us assume that this is accomplished by an exhaustive search.

原子数，这与样品一致。现在，让我们假设这是通过彻底的搜索完成的。

Given a sample of any target concept c in C, this algorithm always produces

给定C语言中任何目标概念C的样本，该算法总是产生

a conjunct that is consistent with the sample, and contains no more atoms than c itself.Hence, given any sample of a target concept in C, this algorithm will

一种与样品一致的混合物，其原子含量不超过c本身。因此，给定C语言中目标概念的任何样本，该算法将

find a consistent hypothesis in C. If it cannot find a consistent hypothesis in C,

在C中找到一致的假设。如果在C中找不到一致的假设，

then the target concept cannot be in C. Thus for any particular C, the

那么目标概念不能在C中。因此，对于任何特定的C

algorithm can easily be adapted to use the hypothesis space C consistently.If L is the resulting algorithm, then by Theorem 4.3, using the bound from Theorem 3.6(ii) on VCdim(C), we can show that

算法可以很容易地适应一致地使用假设空间C。如果L是最终的算法，那么根据定理4.3，利用VCdim(C)上定理3.6(ii)的界，我们可以证明

S~.(e, 6 ) ~< (4 log(2/6 ) + 32s log(4s~,/~) 1o8(13/e))/~.

S~。(e，6 ) ~< (4 log(2/6 ) + 32s log(4s~，/~) 1o8(13/e))/~。

(2)

(2)

202

202

D.HAUSSLER

D.HAUSSLER

When s is very small and n very large, this sample complexity is considerably smaller than that given in Corollary 4.8 (with constants from Corollary 4.5) for the target class of all pure conjunctive concepts using the classical learning algorithm.

当s很小时，n很大时，这个样本的复杂度比推论4.8(用推论4.5中的常数)中使用经典学习算法的所有纯合取概念的目标类给出的要小得多。

Of course this result is of limited value since exhaustively searching for the simplest consistent conjunct requires exponential time, and thus this learning algorithm is entirely impractical as it stands.Can this algorithm be efficiently implemented using a different method?The following shows that it probably cannot.

当然，这个结果的价值是有限的，因为彻底搜索最简单的一致合取需要指数时间，因此这个学习算法是完全不切实际的。使用不同的方法可以有效地实现这个算法吗？下面的例子表明它可能不能。

Theorem 5.1.Given a sample on n attributes that is consistent with some pure

定理5.1。给定一个与一些纯

conjunctive hypothesis, it is NP-hard to find a pure conjunctive hypothesis that is both consistent with this sample and has the minimum number of atoms.

合取假设，很难找到一个既符合这个样本又具有最小原子数目的纯合取假设。

Proof.We will reduce the following problem, known to be NP-hard [13], to the above problem.

证据。我们将把下面的问题，也就是众所周知的NP-hard [13]，简化为上面的问题。

Minimum set cover problem.Given a collection of sets with union T (i.e~

最小集合覆盖问题。给定一组具有并T(即~)的集合

that cover T), find a subcollection whose union is T that has the minimum

覆盖T)，找到其并集为T的子集合，该子集合具有最小值

number of sets.This is called a minimum cover of T.

器械包数量。这被称为t的最小覆盖

Given an instance of the minimum set cover problem defined by the

给定一个由

collection of sets S~ .....S,, with union T = {x~ .....x~}, let A~ ......4,, bca

集合S~.....s，，与联合T = {x~.....x~}，让A~......4、bca

set of Boolean attributes.Let the sample Q consist of one positive example

布尔属性集。让样本Q由一个正例组成

(true, true .....true)

(对，对.....真)

followed by k negative examples

接着是k个否定的例子

.Ut.I, U1.2, - .., Ul.n) ,

. Ut.I，U1.2，-。。，Ul.n)，

(b~,l, v~,~, ..., v~.,,) •

(b~，l，v~，~，。。。，v~。，，)

where for all i, 1 ~< i ~< k and all j, 1 <~j<~ n, vi./ =false ifx~ ~ S~ and v~., = true

其中对于所有I，1 ~ I ~ k和所有j，1 ~ j ~ n，vi。/=假ifx~ ~ S~和v~。，=真

otherwise.

否则。

Suppose that Si,,..., S~ is a subcollection of S ,..., S, that covers T. Then

假设斯，...S~是S的一个子集合，...，S，包括t。那么

we claim that the hypothesis Ai~ = true and ...and A~f = true is consistent

我们声称假设Ai~ =真且。。。A~f =真是一致的

with Q. To verify this, note that it clearly includes the positive example of (2 and furthermore, because every x~, 1 <~ i ~< k appears in some S~,, l ~/~< p, every one of the negative examples has some attribute in Ai, .....Ai, that is set to false, and thus is not included in this hypothesis.

为了验证这一点，请注意它清楚地包括(2)的正例，而且，因为每个x~，1 <~ i ~< k出现在一些S~，l ~/~< p中，所以每个负例在Ai中都有一些属性，.....人工智能，即设置为假，因此不包括在这个假设。

On the other hand, if h is any pure conjunctive hypothesis that is consistent with (2, then h must have the form A~, = true and ....and A~ = true for some {i~ .....i,} G {1,..., n}, for otherwise it would not include the positive example of (2.Furthermore• each of the negative examples of (2 must have the

另一方面，如果h是符合(2)的任何纯合取假设，那么h必须具有形式A~，=真和....对某些人来说，A~ =真.....我，}全球{1，...，否则它将不包括(2)的正例。此外，(2)的每个否定示例都必须有

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

203

203

value false for some attribute in A.q, ..., A ip , otherwise it would be included

在问答中，某些属性的值为false。。。知识产权，否则将被包括在内

in h. Because of the way the negative examples are defined, this implies that Si~ , ..., Sip cover T.

在h .中，因为定义负的例子的方式，这意味着Si~，...，啜盖t

It follows that finding the minimum cover of T from the given collection of sets reduces to finding the pure conjunctive hypothesis that is consistent with Q and has the minimum number of atoms.Hence, since the minimim set cover problem is NP-hard, so is the problem of finding the smallest consistent pure conjunctive hypothesis.[]

由此可见，从给定的集合集合中寻找T的最小覆盖简化为寻找与Q一致且具有最小原子数的纯合取假设。因此，因为最小集合覆盖问题是NP难的，所以寻找最小一致纯合取假设的问题也是如此。[]

The above argument shows how the difficulty of finding the smallest consis-tent pure conjunctive hypothesis is related to the problem of finding the minimum cover of a set T among a collection of sets whose union is T. There is, however, an obvious heuristic for approximating the minimum cover of a set T: First choose a largest set in the collection.Then remove the elements of this set from T and choose another set that includes the maximum number of the remaining elements, continuing in this manner until T is exhausted.This is called the greedy method.

上面的论点说明了寻找最小相容纯合取假设的困难与在并集为T的集合中寻找集合T的最小覆盖的问题是如何联系在一起的。然而，有一个明显的启发式方法来近似集合T的最小覆盖:首先在集合中选择一个最大的集合。然后从T中移除该集合的元素，并选择另一个包含最大数量的剩余元素的集合，以这种方式继续，直到T用尽。这被称为贪婪方法。

To apply this method to the problem of finding pure conjunctive concepts, we first make the following definition.Given an atom a involving an attribute A and a negative example, we say that a eliminates that negative example if it has a value for A that is not included in the set of values for A specified in a. For example, if a is the atom 2 ~ size ~< 5, then a eliminates all negative examples that have sizes outside the range from 2 to 5.We now define the following algorithm.

为了将这种方法应用到寻找纯合取概念的问题上，我们首先做如下定义。给定一个包含属性A的原子A和一个反例，如果它的值A不包含在A中指定的值A的集合中，我们说A消除了那个反例。例如，如果A是原子2 ~大小< 5，那么A消除了所有大小在2到5范围之外的反例。我们现在定义以下算法。

Algorithm 5.2 (Greedy algorithm for learning pure conjunctive concepts).

算法5.2(学习纯合取概念的贪婪算法)。

Step 1.Find the minimal dominating atom for each attribute with respect to the given sample.

第一步。相对于给定的样本，找出每个属性的最小支配原子。

Step 2.Starting with the empty pure conjunctive hypothesis h, while there are negative examples in the sample do:

第二步。从空的纯合取假设h开始，虽然在样本中有否定的例子做:

(a) Among all attributes, find the minimal dominating atom that eliminates

(a)在所有属性中，找到消除的最小支配原子

the most negative examples and add it to h, breaking out of the loop if no minimal dominating atom eliminates any negative examples.(b) Remove from the sample the negative examples that are eliminated.Step 3.If there are no negative examples left return h, else report that the

最负的例子并把它加到h上，如果没有最小支配原子消除任何负的例子，就跳出循环。从样品中去除被消除的阴性样品。第三步。如果没有反例，返回h，否则报告

sample is not consistent with any pure conjunctive concept.

样本不符合任何纯合取概念。

To see how this algorithm differs from Algorithm 4.1, consider again the same instance space and sample used in the previous section to illustrate Algorithm 4.1.The positive examples were

要了解此算法与算法4.1有何不同，请再次考虑上一节中说明算法4.1时使用的相同实例空间和示例。积极的例子有

(square, 5.2, true), (triangle, 3.4, true), (square, 2.9, true)

(正方形，5.2，真)，(三角形，3.4，真)，(正方形，2.9，真)

204

204

and the negative examples were

负面的例子是

(circle, 4.3, true) ,

(圆圈，4.3，真)，

(channel, 5.1, true),

(通道，5.1，真)，

(square, 3.7, false) .

(平方，3.7，假)。

D.HAUSSLER

D.HAUSSLER

As in Algorithm 4.1, Step 1 of Algorithm 5.2 produces the set of minimal

与算法4.1一样，算法5.2的第1步产生一组极小值

dominating atoms

主导原子

shape = regular\_polygon ,

shape =正多边形，

2.9 ~< size <~ 5.2,

2.9 <尺寸< ~ 5.2，

shade = true.

阴影=真。

Initially the hypothesis h is empty.The atom shape = regular\_pol,vgon elimi-nates the most negative examples (two), so it is chosen first and conjoined to h. The examples that it eliminates are removed, leaving only one negative example (square, 3.7, false).The atom shade = true eliminates this example, whereas the size atom does not, so it is now conjoined to h. All negative examples are now eliminated, so the hypothesis

最初假设h是空的。原子形状=正则pol，vgon消除了最负的例子(两个)，因此它被首先选择并连接到h。它消除的例子被去除，只留下一个负的例子(正方形，3.7，假)。原子的阴影=真消除了这个例子，而原子的大小却没有，所以它现在和h连在一起。所有的反例现在都被消除了，所以这个假设

shape = regular\_polygon and shade = true

shape =正多边形，shade =真

is returned.Because it omits the atom 2.9 <~ size ~< 5.2, this hypothesis is

返回。因为它省略了原子2.9 <~尺寸< 5.2，这个假设是

simpler than that produced by Algorithm 4.l.

比算法4产生的更简单。长度

It can readily be verified that, like Algorithm 4.1, Algorithm 5.2 uses the hypothesis space of all pure conjunctive concepts consistently.The proof is similar to that given in Lemma 4.2 and so is omitted.This means that the overall performance of Algorithm 5.2 is at least as good as that established for Algorithm 4.1 in Corollary 4.5.

可以容易地验证，像算法4.1一样，算法5.2一致地使用所有纯合取概念的假设空间。该证明类似于引理4.2中给出的证明，因此被省略。这意味着算法5.2的整体性能至少与推论4.5中的算法4.1一样好。

However, Algorithm 5.2 has the additional property that, while it does not always find the simplest consistent conjunct, it does tend to find simpler conjuncts.This is guaranteed by the following bound on the approximation given by the greedy set cover heuristic.

然而，算法5.2有一个额外的属性，虽然它不总是找到最简单的一致合取，但它确实倾向于找到更简单的合取。这由贪婪集覆盖启发式给出的近似的以下界来保证。

Theorem 5.3 [18, 29].If the set T to be covered has m elements and s is the size

定理5.3 [18，29]。如果要覆盖的集合T有m个元素，并且s是大小

of the minimum cover, then the greedy method is guaranteed to find a cover of

的最小覆盖，那么贪婪的方法保证找到

size at most s(ln m + 1).

大小最多为s(ln m + 1)。

From this theorem it follows that given m examples of an s-atom pure

从这个定理可以得出，给定m个纯s原子的例子

conjunctive concept, Algorithm 5.2 is guaranteed to find a consistent pure conjunctive hypothesis with at most s(ln m + 1) atoms.One way to look at this is as follows.Given a class of target concepts C to be learned, where in this case C is the class of all pure conjunctive concepts with at most s atoms, this

合取概念，算法5.2保证找到一个最多有s(ln m + 1)个原子的一致的纯合取假设。看待这个问题的一种方式如下。给定一类要学习的目标概念，在这种情况下，C是最多有个原子的所有纯合取概念的类，这个

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

205

205

algorithm learns concepts in C using a larger hypothesis space H, namely the class of all pure conjunctive concepts with at most son m + 1) atoms, where m is the sample size.Algorithm 4.1 does this as well when applied to target concepts in C, except that in that case H is the class of all pure conjunctive concepts.

算法使用一个更大的假设空间H学习C中的概念，即最多有m + 1个原子的所有纯合取概念的类，其中m是样本大小。算法4.1在应用于C中的目标概念时也是这样做的，除了在这种情况下H是所有纯合取概念的类。

By using a larger hypothesis space than is strictly needed, it may be computationally easier to find a consistent hypothesis.This is certainly the case here.On the other hand, by using a larger hypothesis space, or more accurately, a hypothesis space with a larger growth function, more random examples will be required in general before we will have confidence that the hypothesis produced is a good approximation to the target concept.Thus it is important to strike a balance between the size or growth function of the hypothesis space and the computational difficulty of finding a consistent hypothesis in this space.Algorithm 5.2 does this in a particularly interesting way by, in effect, dynamically adjusting the size of its hypothesis space to the size of the sample and the complexity of the underlying target concept that generates the sample.This general technique leads to the following.

通过使用比严格需要的更大的假设空间，在计算上可能更容易找到一致的假设。这里肯定是这种情况。另一方面，通过使用一个更大的假设空间，或者更准确地说，一个具有更大增长函数的假设空间，在我们确信所产生的假设是对目标概念的良好近似之前，通常需要更多的随机例子。因此，在假设空间的大小或增长函数与在该空间中寻找一致假设的计算难度之间取得平衡是很重要的。算法5.2以一种特别有趣的方式做到了这一点，实际上是根据样本的大小和生成样本的潜在目标概念的复杂性动态调整其假设空间的大小。这种通用技术导致以下结果。

Definition 5.4.Let L be a learning algorithm, C be a class of target concepts and m be a sample size.By H~(m) we denote the set of all hypotheses produced by L from samples of size m of target concepts in C. We call HZ~:(m) the effective hypothesis space of L for target concepts in C and sample size m.

定义5.4。假设L是一个学习算法，C是一类目标概念，m是样本量。我们用H(m)来表示由L从C中目标概念的m个样本产生的所有假设的集合，我们称之为HZ ~:(m)C中目标概念的L的有效假设空间和样本大小m

The following corollary of Theorem 3.3 can now be used to obtain bounds on the learning performance of algorithms that dynamically adiust their hypothesis space according to sample size.

定理3.3的以下推论现在可以用来获得算法学习性能的界限，这些算法根据样本大小动态地限定它们的假设空间。

Theorem 5.5.Let C be a class of target concepts and let L be a learning

定理5.5。让C成为一类目标概念，让L成为一种学习

algorithm that always produces a consistent hypothesis (not necessarily in C) when given a sample of a target concept in C. Then given a sequence of m >~ 1 independent random examples (chosen according to any fixed probability distri-bution on the instance space) of any target concept c ~ C, for any 0 < e < 1, the probability that L returns a hypothesis with error greater than e is less than

当给定C中的目标概念的样本时，总是产生一致假设(不一定是C中的)的算法。然后给定任何目标概念c ~ C的m > 1个独立随机示例序列(根据实例空间上的任何固定概率分布选择)，对于任何0 < e < 1，L返回误差大于e的假设的概率小于

2//~,/o,,}(2m)2 -~'/e

2//~，/o，，}(2m)2 -~'/e

.

。

Proof.By Theorem 3.3, for any target concept c and distribution on the instance space, this is an upper bound on the probability that any hypothesis in H~(m) with error greater than e is consistent with all m random examples of c. Since L always produces a consistent hypothesis in H~(tn) for any sample of a target concept in C, when the target concept is in C this is therefore an upper bound on the probability that the hypothesis returned by L has error greater than e. [~

证据。根据定理3.3，对于任何目标概念C和实例空间上的分布，这是误差大于e的H~(m)中的任何假设与C的所有m个随机例子一致的概率的上界。由于L总是对C中的目标概念的任何样本产生一致的假设，当目标概念在C中时，因此这是由L返回的假设的误差大于e的概率的上界[~

206

206

D.HAUSSLER

D.HAUSSLER

In order to apply this result to obtain bounds on the sample complexity of

为了应用这个结果来获得样本复杂度的界限

Algorithm 5.2 we will use the following

算法5.2我们将使用以下内容

Lemma ~.6.ff~,,( is a real-valued function and there exist a,b,d >~ 1 such that

引理。ff~，，(是实值函数，并且存在a，b，d >~ 1，使得

tim) <~ a(bm) ~ g m for all m >~ 2, then there exists a constant c~ such that

tim) <~ a(bm) ~ g m对于所有m >~ 2，则存在常数c~使得

f(m)2 -~m/2 ~< 6

f(m)2 -~m/2 ~< 6

~br all 0 < ~,6 < 1

~br全部0 < ~，6 < 1

and 7

和7

m ~ Cl(log(a/6 ) + d(log(bd/e))~)/,: .

m ~ Cl(log(a/6 ) + d(log(bd/e))~)/，: .

Proof.This follows from [16, Lemma l(iii)].The calculations are outlined in

证据。这源于[16，引理l(iii)]。计算概述于

[16, Appendix].[]

[16，附录]。[]

Corollary 5.7.There are positive constants c o and c ~ such that for any instance space X defined on n attributes, each tree-structured or linear,

推论5.7。存在正常数c0和c0 ~使得对于在n个属性上定义的任何实例空间X，每个都是树形结构或线性的，

c0(log(1/6 ) + s log(n/s))/e

c0(log(1/6 ) + s log(n/s))/e

~l, ~ ,

~l，~，

~< 5 ,:,(e, 6 ) ~< c~ (log( 1/6 ) + s(log(sn/~:)) ~)/~"

~< 5，:(e，6)~ < c(log(1/6)+s(log(sn/~:)~)/~ "

for all sufficiently small e and 6, where L is Algorithm 5.2 and C is the class of

对于所有足够小的e和6，其中L是算法5.2，C是

pure conjunctive concepts on X with at most s atoms, s <~ n. Moreover, this

至多有s个原子的X上的纯合取概念，s <~ n。此外，这

lower bound holds for any learning algorithm L.

任何学习算法的下限成立

Proof.As in Corollary 4.8, the lower bound follows from Theorem 4.7, using the lower bound on VCdim(C) given in Theorem 3.6(ii).For the upper bound, note that from Theorem 5.3 it follows that H~(m) is contained in the class of pure conjunctive hypotheses with at most s(ln m + 1) atoms.Thus by Theorem 3.6(ii)

证据。如同在推论4.8中，下限遵循定理4.7，使用定理3.6(ii)中给出的VCdim(C)的下限。关于上限，请注意，从定理5.3可以得出，H~(m)包含在最多有s(ln m + 1)个原子的纯合取假设类中。因此，根据定理3.6(ii)

2Ht4~,im~(2m) ~< 2(2x/Bm) 2~ ....+ ~) ~ 2(2x/-gm) 4~' ~'g '" for ~n ~ 2.

2Ht4~，im~(2m) ~< 2(2x/Bm) 2~....对于~ N2 ,+ ~)~ 2(2x/-GM)4 ~ ' ~ ' g'。

Now let a = 2, b = 2x/~ and d = 4s.Then by Lemma 5.6 there exists a constant

现在让a = 2，b = 2x/~和d = 4s。根据引理5.6，存在一个常数

c~ such that

这样

2H,~,~,,,)(2m)2 .......: ~ 6

2H，~，~，，，)(2m)2.......:~ 6

fbr all 0< e,6 < 1 and m >~ cl(log(1/6) + s(log(sn/e)):)/e .

fbr均为0< e，6 < 1和m > ~ cl(log(1/6)+s(log(sn/e)):)/e。

Hence, by Theorem 5.5, for any distribution on the instance space, given a

因此，根据定理5.5，对于实例空间上的任何分布，给定一个

random sample of this size of any target concept in C, the probability that L

任意目标概念的随机样本，概率为L

produces a hypothesis with error greater than e is at most 6.Thus, this is an

产生误差大于e的假设最多为6。因此，这是一个

upper bound on the sample complexity of L for targets concepts in C. [~

[中目标概念的样本复杂度的上界

7 The (log(bd/e)): factor in this equation can be improved to (log(d/~)) 2 + log b log((d/e) log b),

7(log(BD/e)):该等式中的因子可改进为(log(d/~)) 2 + log b log((d/e) log b)。

which replaces the (log b) ~ term with a log b log log b term (see [16]).

它将(log b) ~项替换为log b log b项(参见[16)。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

207

207

Note that in spite of the fact that the greedy heuristic comes with only a fairly weak guarantee as to the simplicity of the hypothesis it produces, it still performs nearly as well as the algorithm that exhaustively searches for the simplest consistent conjunct (equation (2)), and, more importantly, comes within a poly-logarithmic factor of the optimal sample complexity.Thus it successfully trades off only a small increase in the number of examples needed for a very significant decrease in computational complexity.We can estimate the computation time required by Algorithm 5.2 as follows.

请注意，尽管贪婪启发式算法对其产生的假设的简单性只有相当弱的保证，但它的性能仍然与穷尽搜索最简单的一致合取(等式(2))的算法相当，更重要的是，它在最佳样本复杂度的多对数因子内。因此，它成功地在计算复杂度显著降低所需的示例数量的少量增加之间进行了权衡。我们可以如下估计算法5.2所需的计算时间。

As in the previous section, for simplicity, we assume that for a linear attribute the time required to compare two values is constant and for a tree-structured attribute the time required to determine if one value is in the subtree below another value or to compute the least common ancestor of two values is constant.Thus the time required for finding the minimal dominating atom for a single attribute with respect to a sample of size m and determining how many negative examples it eliminates is O(m).

如前一节所述，为了简单起见，我们假设对于线性属性，比较两个值所需的时间是恒定的，对于树形结构属性，确定一个值是否在另一个值之下的子树中或者计算两个值的最少公共祖先所需的时间是恒定的。因此，对于大小为m的样本，找到单个属性的最小支配原子并确定它消除了多少个负样本所需的时间是O(m)。

Under these assumptions, a simple implementation of Algorithm 5.2 would take time O(nm) for Step 1, O(n) for each execution of Step 2(a) and O(nz) for each execution of Step 2(b), assuming that in Step 2(b) the number of negative examples removed from the sample is z and we also update an array that maintains the number of negative examples eliminated by each minimal dominating atom in light of this new, smaller sample.(Initialization of this array can be done in Step I at no additional cost.) Since the total number of negative examples removed from the sample during the course of the algorithm is less than m, the total time spent in Step 2(b) is O(nm).By the performance bound on the greedy method given Theorem 5.3 above, the total number of iterations of the loop of Step 2 is bounded by O(slogm), where s is the number of atoms in the target concept, so the total time spent in Step 2(a) is O(ns log m).Thus, the overall time is bounded by O(n(m + s log m)).Often we will have s log m ~< m, in which case this algorithm is optimal to within a constant factor.

在这些假设下，算法5.2的简单实现对于步骤1将花费时间0(nm)，对于步骤2(a)的每次执行花费时间0(n)，并且对于步骤2(b)的每次执行花费时间0(NZ)，假设在步骤2(b)中，从样本中移除的负样本的数量是z，并且我们还根据这个新的更小的样本更新了一个数组，该数组保持由每个最小支配原子消除的负样本的数量。(该阵列的初始化可以在步骤1中完成，无需额外成本。)由于在算法过程中从样品中去除的阴性样品的总数小于m，所以步骤2(b)中花费的总时间为0(nm)。根据上面定理5.3给出的贪婪方法的性能界限，步骤2的循环的总迭代次数由0(slogm)限定，其中s是目标概念中的原子数，因此步骤2(a)中花费的总时间是0(ns log m)。因此，总时间由0(n(m+s log m))限定。通常我们会有s log m ~< m，在这种情况下，该算法在一个常数因子内是最优的。

6.Learning Pure Disjunctive, k-DNF and k-CNF Concepts

6.学习纯析取、k-DNF和k-CNF概念

The complements of pure conjunctive concepts can be represented as pure disjunctive concepts.Hence this is the dual form of pure conjunctive concepts.A variant of Algorithm 5.2 can be used to learn pure disjunctive concepts.For a pure disjunction to be consistent with a sample, each atom must eliminate all negative examples and need only include some subset of positive examples, and all atoms together must include (cover) all positive examples.To achieve this, in place of minimal dominating atoms we use their dual counterparts, which we call maximal subordinate atoms.For each attribute A, these are the most general elementary atoms involving A that include at least one positive example and no negative examples.For tree-structured attributes, they are

纯合取概念的补语可以表示为纯析取概念。因此，这是纯粹合取概念的双重形式。算法5.2的变体可以用来学习纯析取概念。为了使一个纯粹的析取与一个样本相一致，每个原子必须排除所有的反例，并且只需要包括一些正例的子集，并且所有的原子一起必须包括(覆盖)所有的正例。为了实现这一点，我们用它们的对偶对应物来代替最小支配原子，我们称之为最大从属原子。对于每个属性A，这些是涉及到A的最一般的基本原子，包括至少一个正的例子而没有负的例子。对于树形结构的属性，它们是

208

208

D.HAtJSSLER

D.HAtJSSLER

nodes closest to the root that define subtrees whose leaves contain only values from positive examples.For linear attributes, they are maximal intervals that contain only values from positive examples.Note that unlike minimal dominat-ing atoms, each attribute can have more than one maximal subordinate atom.

定义子树的最接近根的节点，子树的叶子只包含正例中的值。对于线性属性，它们是仅包含正例中的值的最大间隔。请注意，与最小支配原子不同，每个属性可以有多个最大从属原子。

The dual greedy method is to repeatedly choose the maximal subordinate atom that covers the most positive examples and add it to the hypothesis, removing any new positive examples that are covered, until either all positive examples are accounted for, or no maximal subordinate atom covers any of the remaining positive examples.

对偶贪婪方法是重复选择覆盖最积极例子的最大从属原子，并将其添加到假设中，删除任何新的被覆盖的积极例子，直到所有积极例子都被考虑，或者没有最大从属原子覆盖任何剩余的积极例子。

As in the previous section, this method produces a consistent pure disjunc-tive hypothesis if any exist, and this hypothesis has at most s(ln m ~ 1 ) atoms for any sample of size rn of a pure disjunctive target concept with at most .s atoms.The VC dimension and the growth function for the hypothesis space of pure disjunctive concepts with at most s(In m + 1) atoms are bounded in the same way that the corresponding VC dimension and growth function for pure conjunctive concepts are bounded (Theorem 3.6(ii)).Hence the results given in Corollary 5.7 also hold when L is the learning algorithm defined by the dual greedy method and C is the target class of pure disjunctive concepts with at most s atoms.

如前一节所述，这种方法产生了一个一致的纯分离假设(如果存在的话)，对于一个纯分离目标概念的rn大小的任何样本，该假设至多有s(ln m ~ 1)个原子。s原子。至多有s(In m + 1)个原子的纯析取概念的假设空间的虚函数维数和增长函数的有界方式与纯合取概念的相应虚函数维数和增长函数的有界方式相同(定理3.6(ii))。因此，当L是由对偶贪婪方法定义的学习算法，而C是具有最多s个原子的纯析取概念的目标类时，推论5.7中给出的结果也成立。

The dual greedy method is a variant of the "star" methodology of Michalski

双重贪婪方法是迈克尔斯基的“明星”方法论的变体

[24].However, in Michalski's method, you repeatedly pick a "seed" positive example at random and then add the (in this case unique) maximal subordinate atom that includes it to the hypothesis, removing any newly covered positive examples.The important difference here is that since we are using the greedy heuristic to select our maximal subordinate atoms rather than random draw.we arc able to give quantitative performance bounds using the known bounds on the greedy method for set cover (Theorem 5.3).Michalski also suggests using the number of positive examples covered as a criterion for selecting between competing ~naximal subordinate atoms in more complicated learning domains where there can be more than one such atom for a given seed.However, this filtering comes later in his mcthod, after the seed has already been randomly selected.This does not allow lor the possibility thai some seeds may be better than others for producing atoms that cover man\ positive examples.

[24]。然而，在迈克尔斯基的方法中，你反复随机选择一个“种子”正例，然后将包含它的(在这个例子中是唯一的)最大从属原子添加到假设中，删除任何新覆盖的正例。这里的重要区别在于，因为我们使用贪婪的启发式来选择我们最大的从属原子，而不是随机抽取。我们能够使用集合覆盖的贪婪方法的已知界限给出定量的性能界限(定理5.3)。Michalski还建议，在更复杂的学习领域中，使用所涵盖的正面例子的数量作为在竞争的~naximal从属原子之间进行选择的标准，在这些领域中，对于给定的种子可以有多个这样的原子。然而，在他的方法中，这种过滤是在种子已经被随机选择之后进行的。这不允许某些种子可能比其他种子更好地产生覆盖人类正面例子的原子。

The dual greedy method can be extended to learn k-DNF concepts for fixed k (see definition in Section 1).Apply the tnethod as above, except at each step choose the k-atom pure conjunctive concept (term) that includes the most positive examples without including any negative examples.This choice can be made as follows.

对偶贪婪方法可以扩展到学习固定k的k-DNF概念(见第1节的定义)。应用上述方法，除了在每一步选择k原子纯合取概念(术语)，包括最积极的例子，但不包括任何消极的例子。这个选择可以如下进行。

For every attribute A and every pair of w~lues ~ and ~, of A that occur

对于出现的每一个属性A和每一对值A和A

among the examples of the sample, calculate either the least common ancestor v of v~ and v, and form the atom A = v (if A is a tree-structured attribute) or form the atom v~ ~< A ~< v2 (if A is linear and v~ ~ t,\_~).This creates a pool of at most nrn -~ atoms, where n is the number of attributes and rn is the sample size.

在示例中，计算v~和v的最少公共祖先v，并形成原子A = v(如果A是树形结构属性)或形成原子v~ ~< A ~< v2(如果A是线性的且v~ ~ t，\_~)。这将创建一个最多包含nrn个原子的库，其中n是属性的数量，rn是样本大小。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

209

209

For every conjunction of k atoms from this pool, determine the number of positive and negative examples it includes, and select the conjunction that includes the largest number of positive examples without including any nega-tive examples.This can be done in time O((nm2)~m)= O(n~m2k+l).

对于这个库中k个原子的每个连词，确定它包含的正例和反例的数量，并选择包含最多正例但不包含任何反例的连词。这可以在时间0((nm2)~ m)= 0(n ~ m2k+1)内完成。

The overall time analysis of the algorithm is now easy.By the basic

算法的总体时间分析现在很容易。基本上

performance bound on the greedy method (Theorem 5.3), given m examples of a k-DNF target concept with s terms, this method produces a consistent k-DNF concept with at most s(ln m + 1) terms.Hence, the main loop is executed at most O(s log m) times, giving an overall time bound of O(snkm ~'~+~ log m).(Since each iteration of the loop takes so long, we dispense with the more refined approach to the analysis taken in the previous section.)

贪婪方法的性能界限(定理5.3)，给定带有s项的k-DNF目标概念的m个例子，该方法产生具有最多s(ln m + 1)项的一致k-DNF概念。因此，主循环最多执行0(s log m)次，总的时间范围为0(snkm ~ ' ~+~ log m)。(由于循环的每次迭代都需要很长时间，所以我们省去了上一节中更精细的分析方法。)

In analogy with Corollary 5.7, we have the following bounds on the sample complexity of this algorithm.

与推论5.7类似，我们对该算法的样本复杂度有以下限制。

Corollary 6.1.There are positive constants c o and cl such that for any instance space X defined on n attributes, each tree-structured or linear,

推论6.1。存在正常数c0和C1，使得对于在n个属性上定义的任何实例空间X，每个都是树形结构或线性的，

c0(log(1/6) + ks log(n/ksl/k)) /e

c0(log(1/6) + ks log(n/ksl/k)) /e

L

L

<~ So(e,6 ) <~ c 1 (log(1/6 ) + ks(log(ksn/e)) 2)/e

< ~ So(e，6)< ~ C1(log(1/6)+ks(log(ksn/e))2)/e

for all sujficiently small e and 6, where L is the above algorithm and C & the

对于所有足够小的e和6，其中L是上述算法，C和

class of all k-DNF concepts on X with at most s terms, k <~ n and s ~ ( ~ ).

最多含s项，k <~ n和s < ~的X上的所有k-DNF概念的类。

Moreover, this lower bound holds for any learning algorithm L.

此外，这个下界适用于任何学习算法

Proof.Similar to that of Corollary 5.7, but using Theorem 3.6(iii).[]

证据。类似于推论5.7，但使用定理3.6(iii)。[]

Again, this shows that the sample complexity of the algorithm is within a poly-logarithmic factor of optimal.This improves on Valiant's result [40] for learning k-DNF by reducing the required sample size from O(n k) to a size logarithmic in n.

这再次表明算法的样本复杂度在最佳的多对数因子内。这改进了瓦兰特的结果[40]学习k-DNF通过减少所需的样本大小从0(n k)到一个对数的大小在n

By duality, these results also extend to the class of k-CNF concepts.

通过对偶性，这些结果也扩展到k-CNF概念类。

Theorem 3.6(iii) shows that the same bounds hold for the growth function and the VC dimension.Clearly the algorithm outlined above can be dualized again so that, as in Algorithm 5.2, in each step we choose the k-atom clause that includes all the positive examples and eliminates the most negative examples.If L is the resulting algorithm and C is the class of k-CNF concepts on n attributes with at most s clauses, then the same computational and sample complexity bounds derived above still hold.

定理3.6(iii)表明，增长函数和可变成本维数有相同的界限。显然，上面概述的算法可以再次双重化，因此，如在算法5.2中一样，在每个步骤中，我们选择包含所有正面示例的k-atom子句，并删除最负面的示例。如果L是结果算法，C是n个属性上的k-CNF概念类，最多有s个子句，那么上面导出的相同的计算和样本复杂度界限仍然成立。

This greedy method, like Valiant's method, is clearly computationally im-practical for large k. Thus, in practice, the exhaustive search part of the algorithm should be replaced by a limited heuristic search (e.g. as in [24]).However, we have not found any heuristic techniques that lead to provably good learning performance for arbitrary distributions.

这种贪婪的方法，就像瓦兰特的方法一样，显然在计算上对大k不实用。因此，在实践中，算法的穷举搜索部分应该被有限的启发式搜索所代替(例如，在[24中)。然而，我们还没有发现任何启发式技术能够为任意分布带来良好的学习性能。

210

210

7.Internal Disjunctive Concepts

7.内部析取概念

D.HAUSSLER

D.HAUSSLER

We now tackle the problem of learning internal disjunctive concepts.There are several ways to go about simplifying internal disjunctive hypotheses to improve the performance of a learning algorithm.One extreme is to try to get rid of as many compound atoms as possible, similarly to what we did with pure conjunctive hypotheses.The other is to try to reduce the number of internal disjunctions within one or more of the compound atoms of the hypothesis.A good compromise is to try to minimize the total number of atoms plus internal disjunctions in hypothesis, which we call the (syntactic) size of the hypothesis.For an internal disjunctive hypothesis h, the size of h is equal to the total number of occurrences of tree-structured attribute values and linear attribute value ranges in all compound atoms combined.

我们现在解决学习内部析取概念的问题。有几种方法可以简化内部析取假设，以提高学习算法的性能。一个极端是试图尽可能多地去除复合原子，就像我们对纯合取假设所做的那样。另一种是试图减少假设中一个或多个复合原子内部分离的数量。一个很好的折衷办法是尽量减少假设中的原子总数加上内部析取，我们称之为假设的(句法)大小。对于内部析取假设h，h的大小等于所有复合原子中树形结构属性值和线性属性值范围出现的总数之和。

Let h be an internal disjunctive hypothesis that is consistent with a given sample.As with pure conjunctive hypotheses, each atom in h includes all positive examples and eliminates some set (possibly empty) of negative exam-ples.A compound atom with this property will be called a dominating compound atom.We would like to eliminate all the negative examples using a conjunction of dominating compound atoms with the smallest total size.This leads to the following.

假设h是与给定样本一致的内部析取假设。和纯合取假设一样，h中的每个原子都包含所有的正例，并排除了一些(可能是空的)负例。具有这种性质的化合物原子称为支配化合物原子。我们想用具有最小总尺寸的主要化合物原子的结合来消除所有的否定例子。这导致以下结果。

Minimum set cover problem with positive integer costs.Given a collection of sets with union T, where each set has associated with it a positive integer cost, find a subcollection whose union is T that has the minimum total cost.

正整数成本的最小集合覆盖问题。给定一组具有并集T的集合，其中每个集合都有一个与之相关联的正整数代价，找到一个其并集为T的子集合，该子集合具有最小的总代价。

Since it generalizes the minimum set cover problem, this problem is clearly NP-hard as well.However, approximate solutions can be found by a general-ized greedy method.Let T' be a set of elements remaining to be covered.For each set in the collection, define the gain~cost ratio of this set as the number of elements of T' it covers divided by its cost.The generalized greedy method is to always choose the set with the highest gain/cost ratio and add it to the cover.As with the basic minimum set cover problem, it can be shown that if the original set T to be covered has m elements and s is the minimum cost of any cover, then the generalized greedy method is guaranteed to find a cover of cost at most s(ln m + 1) [8].

因为它推广了最小集合覆盖问题，这个问题显然也是NP难的。然而，近似解可以通过一般的贪婪方法找到。让“T”是一组有待覆盖的元素。对于集合中的每个集合，将该集合的收益成本比定义为它覆盖的元素数量除以它的成本。广义贪婪方法总是选择增益/成本比最高的集合，并将其添加到封面。与基本最小集合覆盖问题一样，可以证明，如果要覆盖的原始集合T有m个元素，并且s是任何覆盖的最小代价，那么广义贪婪方法保证找到一个代价至多为s(ln m + 1) [8]的覆盖。

To apply this method to learning internal disjunctions, let the gain/cost ratio of a dominating compound atom be the number of negative examples it eliminates divided by its size.

为了将这种方法应用于学习内部析取，让占支配地位的化合物原子的增益/成本比等于它消除的负例数除以它的大小。

Algorithm 7.1 (Greedy algorithm for learning internal disjunctive concepts).

算法7.1(学习内部析取概念的贪婪算法)。

Step 1.Starting with the empty internal disjunctive hypothesis h, while there are negative examples in the sample do:

第一步。从空的内部析取假设h开始，虽然在样本中有否定的例子做:

(a) Among all attributes, find the dominating compound atom a with the

在所有的属性中，用

highest gain/cost ratio and add it to h, breaking out of the loop if none have positive gains.

最高的增益/成本比，并将其加到h上，如果没有正增益，就打破循环。

211

211

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

(b) Remove from the sample the negative examples a eliminates.Step 2.If there are no negative examples left return h, 8 else report that the sample is not consistent with any internal disjunctive concept.

(b)从样品中去除a消除的阴性样品。第二步。如果没有反面例子，则返回h，否则报告样本与任何内部析取概念不一致。

As an example, we trace the development of the hypothesis h in Algorithm 7.1, given the examples of Fig. 3(a).At the start of the algorithm h is empty.During the first iteration of the loop of Step 1 the dominating compound atom shape = convex is found to have the highest gain/cost ratio: it eliminates all the nonconvex negative examples in the bottom row of Fig. 3(a) and at a cost of 1, since only one value for shape is specified.Its gain/cost ratio is thus 4.This atom is thus added to h, giving

作为一个例子，我们跟踪算法7.1中假设h的发展，给出了图3(a)的例子。算法开始时，h为空。在步骤1的循环的第一次迭代期间，发现占优势的复合原子形状=凸具有最高的增益/成本比:它消除了图3(a)的底行中的所有非凸负示例，并且成本为1，因为只指定了一个形状值。因此，它的收益/成本比为4。这个原子因此被加到h上，给出

h = (shape = convex)

h =(形状=凸形)

and these four negative examples are removed from the sample.On the next iteration, the dominating compound atom 1.7 <~ size ~ 3.0 is selected.It elimi-nates the large yellow square and the small red square for a gain of 2, at a cost of 1, because only one interval is specified.After this iteration

这四个负面的例子从样本中移除。在下一次迭代中，选择主要的化合物原子1.7 <~尺寸3.0。它消除了大的黄色方块和小的红色方块，增益为2，代价为1，因为只指定了一个间隔。在这个迭代之后

h = (shape = convex) and (1.7 ~< size <~ 3.0).

h =(形状=凸起)和(1.7 ~ <尺寸< ~ 3.0)。

On the next iteration, we find the atom shape = regular\_polygon or circle has the highest gain/cost ratio ( ), eliminating the ellipse at a cost of 2.Now

在下一次迭代中，我们发现原子形状=正多边形或圆具有最高的增益/成本比( )，以2的成本消除了椭圆。现在

h = (shape = convex) and (1.7 <~ size <~ 3.0)

h =(形状=凸形)和(1.7 <~尺寸< ~ 3.0)

and (shape = regular\_polygon or circle),

和(形状=正多边形或圆形)，

which can be reduced to

这可以简化为

h = (shape = regular\_polygon or circle) and (1.7 ~< size <~ 3.0).

h =(形状=正多边形或圆形)和(1.7 <尺寸< ~ 3.0)。

Finally, the last iteration eliminates the green triangle by adding the atom color = red or yellow or blue, giving the final hypothesis (in reduced form)

最后，最后一次迭代通过添加原子颜色=红色或黄色或蓝色来消除绿色三角形，给出最终假设(以简化形式)

h = (shape = regular\_polygon or circ'&)

h =(形状=正多边形或circ'&)

and (1.7 ~< size <~ 3.0) and (color = red or yellow or blue).

和(1.7-3.0)和(颜色=红色或黄色或蓝色)。

All negative examples have been eliminated, so this hypothesis is consistent with the sample, and is returned.

所有否定的例子都被排除了，所以这个假设与样本是一致的，并且被返回。

It is clear that to implement this algorithm, we need an efficient procedure to find a dominating compound atom with the highest gain/cost ratio for a given

很明显，为了实现这个算法，我们需要一个有效的过程来找到给定的增益/成本比最高的支配化合物原子

~h may include several compound atoms for the same attribute.In practice these would be combined into one logically equivalent compound atom so that the final hypothesis is given in the simplest form.

~h可能包含几个具有相同属性的化合物原子。在实践中，这些原子将被组合成一个逻辑上等价的化合物原子，因此最终的假设以最简单的形式给出。

212

212

1.7

1.7

+

+

2.5

2.5

+

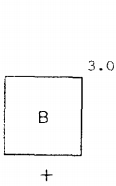
+

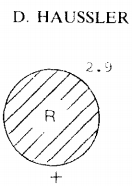
1.8 O

1.8欧

+

+





--

-

\_

\_

1.5

1.5

P 'I

皮伊

3.0

3.0

2.3

2.3

C

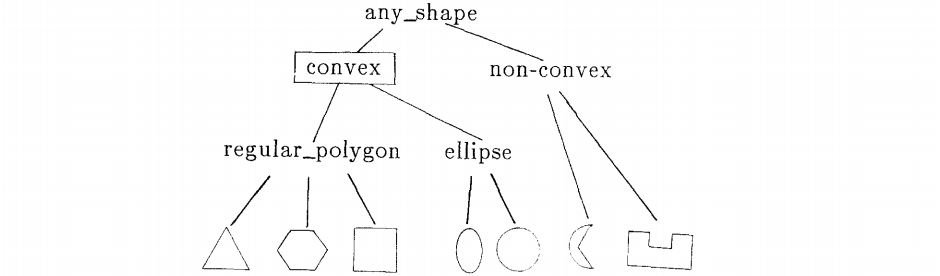
C

(a)

(a)

4.0

4.0



(1)

(1)

•

﹍)

(o)

(o)

(,)

(，)

~

~

(1)

(1)

(b)

(b)

(o)

(o)

•

﹍)

(~)

(~)

(e)

(e)

Fig. 3.(a) Sample Q on an instance space defined by attributes: shape: as in Fig. 2;color: {red (R), yellow (Y), blue (B), green (G), purple (P), orange (0), any\_color};size: real-valued (values

图3。(a)在由属性定义的实例空间上的样本Q:形状:如图2所示；颜色:{红色(R)、黄色(Y)、蓝色(B)、绿色(G)、紫色(P)、橙色(0)、任意\_颜色}；大小:实值(值

indicated next to example);shade: {true (111), false (), a.y\_shade/.{b) The minimal dominating

紧挨着例子指示)；阴影:{true (111)，false()，a.y\_shade/。最小支配

atom for the attribute shape with respect to the sample Q of Fig. 3(a) is shap~ = convex.Values

相对于图3(a)的样本Q的属性形状的原子是shap~ =凸的。价值观念

that appear in one or more positive example are marked with a star.The number in parentheses is the number of negative examples that the value appears in (used in Section 7).

出现在一个或多个正面例子中的标记有星号。括号中的数字是该值出现在其中的反例数(在第7节中使用)。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

213

213

attribute.Since there are in general exponentially many distinct dominating compound atoms with respect to the number of leaves of a tree-structured attribute or the number of values of a linear attribute, this cannot be done by an exhaustive search.However, there is a reasonably efficient dynamic pro-gramming procedure that does this for tree-structured attributes, and a simple iterative procedure for linear attributes.The reader that is not interested in the implementation details of these procedures can safely skip ahead to Corollary 7.2, where the learning performance of Algorithm 7.1 is evaluated.

属性。由于相对于树形结构属性的叶子数目或线性属性的值数目，通常有指数形式的许多不同的支配化合物原子，这不能通过穷举搜索来完成。然而，对于树形结构的属性，有一个相当有效的动态编程过程，对于线性属性，有一个简单的迭代过程。对这些过程的实现细节不感兴趣的读者可以安全地跳到推论7.2，在那里评估算法7.1的学习性能。

The procedures we use to find a dominating compound atom with the highest

我们用来寻找具有最高

gain/cost ratio for a given attribute actually produce what we call a candidate list, which is a list of dominating compound atoms with the highest gain, with

给定属性的增益/成本比实际上产生了我们所说的候选列表，这是一个具有最高增益的主要化合物原子的列表，具有

one for each possible cost.We discuss the procedure for tree-structured

每种可能的成本一个。我们讨论树形结构的过程

attributes first.

属性优先。

Assume we are given a sample Q and a tree-structured attribute A. We first

假设给我们一个样本Q和一个树形结构的属性a

derive from A a tree T, called the projection of Q onto A, and two numbers, called the base\_gain and base\_cost.These objects are defined as follows.The

从A导出一棵树T，称为Q到A的投影，和两个数，称为基本增益和基本成本。这些对象定义如下。这

leaves of T include only the leaves of A whose values occur among the positive examples of Q. The internal nodes of T are all the least common ancestors in A of sets of leaves of T. The descendant relationship among the nodes in T is the same as it was in A. Hence, the root of T is the least common ancestor in A of the set of all leaves of T. Each internal node of T is labeled with the name of the value it represents, taken from A, and two nonnegative integers called the

T的叶子仅包括其值出现在Q的正例中的A的叶子。T的内部节点都是T的叶子集合中的最不常见的祖先。T中的节点之间的后代关系与A中的相同。因此，T的根是T的所有叶子集合中的最不常见的祖先。T的每个内部节点都用它所代表的值的名称来标记，该值取自A和两个称为的非负整数

gain and cost.The gain of an internal node ~r is the number of additional

收益和成本。内部节点r的增益是附加的

negative examples eliminated when o- is expanded in a dominating compound

当o-在主要化合物中膨胀时，消除了负例

atom, i.e. when the value represented by o- is replaced by the disjunction of the values represented by its immediate successors in T. Assume the immediate successors of o- in T are o-l,..., o- k. The gain of cr is calculated by determining the total number of negative examples with values associated with leaves that are in the subtree of o- in A, but not in any of the subtrees of ~rl,..., o- k in A. The cost of ~ is the increase in the size of a dominating compound atom containing o- when o" is expanded.The cost is simply the number of immediate

原子，即当由o-表示的值被由它在T中的直接后继者表示的值的析取所代替时。假设o-在T中的直接后继者是o-1，...cr的增益是通过确定负样本的总数来计算的，这些负样本的值与在A的子树中的叶子相关联，但是不在~ rl的任何子树中。...~的成本是当o”膨胀时，含o-的主要化合物原子尺寸的增加。成本只是眼前的数字

successors of o- in T minus one.base\_gain(T) is the number of negative

o- in T减1的后继者。基数增益是负数

examples eliminated by the dominating compound atom A = v, where v is the

被主要化合物原子A = v消去的例子，其中v是

value represented by the root of T. base\_cost(T) is the cost of this atom, i.e. one.The projection of the sample Q given in Fig. 3(a) onto the attribute shape

由成本的根表示的值是这个原子的成本，即1。图3(a)中给出的样本Q在属性形状上的投影

given in Fig. 3(b) is illustrated in Fig. 4.

图3(b)中给出的内容在图4中示出。

By a predecessor-closed subtree of a tree T, we mean a subtree T' such that

对于树T的前闭子树，我们指的是子树T’

whenever a node cr of T is in T', then all predecessors of o- in T are also in T'.

每当T的一个节点cr在T’中时，那么o- in T的所有前置节点也在T’中。

With each predecessor-closed subtree T' of the internal nodes of T, there is

对于T的内部节点的每个前置封闭子树T ’,有

associated a cut of T, denoted cut(T'), defined as the set of all immediate

关联一个T的割，表示为割(T’)，定义为所有立即的集合

successors of the leaves of T'.When T' is empty, cut(T') is the root of T.

‘T’叶的继承者。当“T”为空时，切(T’)是T的根

These definitions are illustrated in Fig. 5.

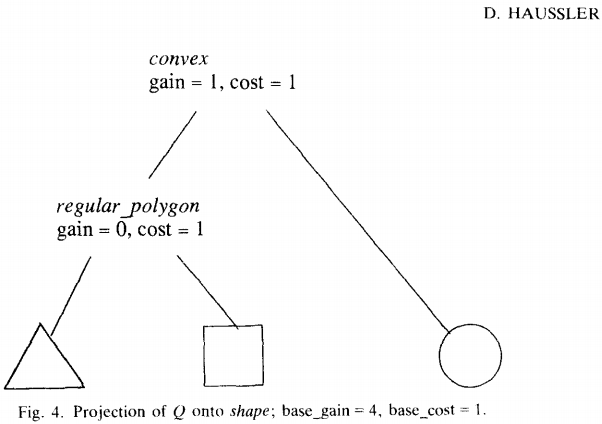
这些定义如图5所示。

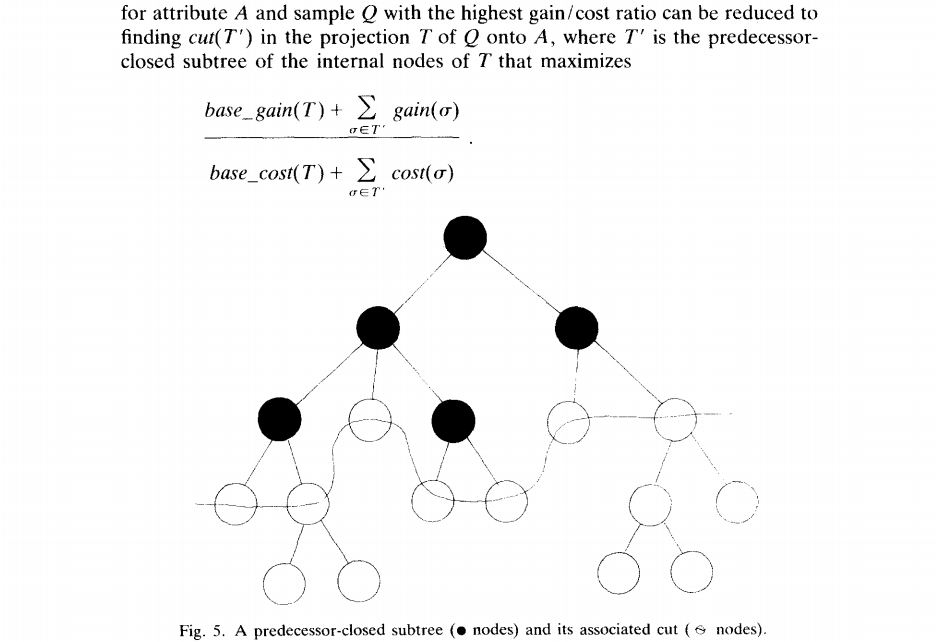
It is easily verified that the problem of finding a dominating compound atom

很容易证明找到一个占优势的化合物原子的问题

214

214





QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

215

215

By adding a "dummy root" to T that has gain base\_gain(T) and cost

通过将“虚拟根”添加到具有增益基数增益和成本的T

base\_cost(T), and deleting the leaves of T (see Fig. 6), the latter problem

基本成本(T)，并删除T的叶子(见图6)，后一个问题

reduces to the following:

简化为以下内容:

Investrnent problem.Given a set of investments I, each of which has a

投资问题。给定一组投资，每个投资都有一个

nonnegative gain and a nonnegative integer cost, and a rooted tree T with node set I specifying which investments in I must be made prior to other investments (investment cr must be made before investment/3 if ~r is an ancestor of/3 in the tree), find a feasible investment scheme with the highest gain/cost ratio, i.e. a nonempty predecessor-closed subtree T' of T that maximizes

非负收益和非负整数成本，以及带有节点集I的根树T，该树指定在I中的哪些投资必须在其他投资之前进行(如果~r是树中/3的祖先，则投资cr必须在投资/3之前进行)，找到具有最高收益/成本比的可行投资方案，即T的非空的前体封闭子树T ′,该子树最大化

Z gain(~r)/ Z cost(~r).

Z增益(~r)/ Z成本(~r)。

~r~T' ~o, GT'

~r~T' ~o，GT '

This investment problem is a variant of the similar investment problem solved in [19] by dynamic programming.There we are given a bound/3 on the maximum total cost of the investments we can make and seek to maximize our gain subject to this constraint.The dynamic programming technique given in

这个投资问题是在[通过动态规划解决的类似投资问题的变体。在这里，我们得到了我们所能进行的投资的最大总成本的1/3的界限，并在这一限制下寻求我们收益的最大化。中给出的动态编程技术

[19] solves this problem by (essentially) calculating for each possible total cost the predecessor-closed subtree with the maximal total gain that has at most that cost.Not only does this solve our investment problem as well, but, under the above reduction, the cuts for these subtrees form the candidate list used for selecting the dominating compound atom with the highest gain/cost ratio.The

[19]解决了这个问题，通过(本质上)计算每个可能的总成本，具有最大总收益的前一个封闭子树最多有那个成本。这不仅解决了我们的投资问题，而且，在上述缩减下，这些子树的削减形成了用于选择具有最高增益/成本比的主要化合物原子的候选列表。这

combined time required for these calculations is O(tq), where t is the number

这些计算所需的总时间是O(tq)，其中t是数字

nodes in the tree and q is the number of distinct possible total costs.9 When we

树中的节点数，q是不同的可能总成本数。9当我们

dummy root

虚拟根

gain = 4, cost = 1

增益= 4，成本= 1

COFIVeX

COFIVeX

gain = cost = /

增益=成本= /

regular\_polygon

正多边形

gain = 0, cost ;1

增益= 0，成本；1

Fig. 6.Investment problem derived from Fig. 4.

图6。投资问题源自图4。

~Because the algorithm runs in time proportional to the sum of the costs of the nodes, rather than the total number of bits required to represent these costs, it is only a pseudo-polynomial time algorithm [13].In this ease this is likely to be the best one can hope for, since the investment problem with bound B is NP-hard [191.We do not know if the investment problem we have given above is also NP-hard.

~因为该算法的运行时间与节点成本的总和成正比，而不是代表这些成本所需的总位数，所以它只是一个伪多项式时间算法[13]。在这种情况下，这可能是一个最好的希望，因为投资问题的边界B是NP-hard [191。我们不知道我们上面给出的投资问题是否也是NP难的。

216

216

D.HAUSSLER

D.HAUSSLER

are producing a candidate list from a projection, the size t of the tree is

根据投影生成候选列表，树的大小为

proportional to the number of distinct values for the attribute A that appear in positive examples in the sample, which is bounded by the sample size m. Since each possible total cost corresponds to the size of some subtree of this tree, the number of distinct possible total costs is also bounded by m, giving an O(m 2) procedure.Of course this is a considerable overestimate if the trees for the attributes are small and the sample size is large.

与样本中出现在正例中的属性A的不同值的数量成比例，该数量由样本大小m来限定。由于每个可能的总成本对应于该树的某个子树的大小，所以不同的可能总成本的数量也由m来限定，从而给出一个O(m 2)过程。当然，如果属性的树很小，样本量很大，这是相当大的高估。

We can calculate the candidate list for a linear attribute by a much simpler procedure.To do this, we partition the sequence of ordered values of the attribute using the maximal subordinate atoms, i.e. the maximal intervals that contain at least one positive example and no negative examples.The intervals between two consecutive intervals for maximal subordinate atoms will be called gaps.We rank the gaps in increasing order according to the number of negative examples that have their value in the gap.By selecting i gaps of highest rank for any i >~ 0, we can find a dominating compound atom of cost i + 1 with the maximum gain as follows.

我们可以通过一个简单得多的过程来计算线性属性的候选列表。为此，我们使用最大从属原子来划分属性的有序值序列，即包含至少一个正实例而不包含负实例的最大间隔。两个最大从属原子的连续间隔之间的间隔称为间隙。我们根据在差距中有其价值的负面例子的数量，按升序排列差距。通过为任何i >~ 0选择最高等级的I间隙，我们可以找到具有最大增益的成本i + 1的主要化合物原子，如下所示。

First remove all other gaps by (temporarily) throwing away all negative

首先通过(暂时)扔掉所有消极的东西来消除所有其他的差距

examples with values that lie in these other gaps.Then form the dominating compound atom consisting of the disjunction of all the maximal subordinate atoms for the resulting sample.Since there will be only i gaps between consecutive intervals of maximal subordinate atoms, there will be only i + 1 atoms, hence the resulting compound atom will have cost i + 1.By construc-tion, it will cover all positive examples (hence be dominating) and have maximal gain among compound atoms of the same size, since it eliminates all the negative examples with values in the i highest rank gaps, plus all the negative examples that do not lie in a gap because they have values that are either smaller than the smallest positive value or larger than the largest positive value.To make this procedure more compatible with the one for the tree-structured attributes, we then shrink each of the intervals in the compound atom as far as possible without uncovering any positive examples.This makes each interval the most specific that covers its positive examples, just as each atom formed by computing the least common ancestor in a tree-structured attribute of a set of positive examples is the most specific that covers these examples.Finally, to form the entire candidate list we do this for each cost i from 0 to the total number of gaps, hence this procedure, like the one for tree-structured attributes, also takes time quadratic in the number of distinct values of the attribute that appear in positive examples, and thus is O(m2).

值位于这些其他间隙中的示例。然后形成由最终样品的所有最大从属原子的分离组成的主要化合物原子。因为在最大从属原子的连续间隔之间只有I个间隙，所以只有i + 1个原子，因此所得到的复合原子的成本为i + 1。通过构造，它将覆盖所有的正例(因此是支配性的)，并且在相同大小的化合物原子中具有最大增益，因为它消除了具有最高等级间隙值的所有负例，加上不在间隙中的所有负例，因为它们具有小于最小正值或大于最大正值的值。为了使这个过程与树形结构的属性更加兼容，我们尽可能地缩小复合原子中的每个间隔，而不暴露任何正面的例子。这使得每个区间覆盖其正例的最具体，正如通过计算一组正例的树结构属性中的最少公共祖先而形成的每个原子覆盖这些示例的最具体一样。最后，为了形成整个候选列表，我们对从0到间隙总数的每一个成本I都这样做，因此这个过程，就像对树结构属性的过程一样，也需要时间，该时间与出现在正例中的属性的不同值的数量成二次方，因此是O(m2)。

The overall analysis of Algorithm 7.1 can now be given.Let s denote the size of the internal disjunctive target concept.The bound on the generalized greedy method guarantees that the loop in Step 1 of Algorithm 7.1 will be executed at most O(s log m) times, where m is the sample size.The cost of each iteration is dominated by the time it takes to produce the candidate lists for each attribute, which is O(nm2), giving an overall time bound of O(snm 2 log rn).Again, this is

现在可以给出算法7.1的整体分析。让我们表示内部分离目标概念的大小。广义贪婪方法的界限保证算法7.1的步骤1中的循环最多执行0(s log m)次，其中m是样本大小。每次迭代的成本主要取决于为每个属性生成候选列表所需的时间，该时间为0(nm2)，总时间范围为0(snm 2 log rn)。同样，这是

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

217

217

a considerable overestimate if the number of distinct values for any attribute that appear in positive examples is small.

如果出现在正面例子中的任何属性的不同值的数量很少，这是相当大的高估。

Finally, we can also give fairly tight bounds on the learning performance of Algorithm 7.1.

最后，我们还可以对算法7.1的学习性能给出相当严格的限制。

Corollary 7.2.There are positive constants c o and c 1 such that for any instance space X defined on n attributes, each tree-structured or linear,

推论7.2。存在正常数c0和c 1，使得对于在n个属性上定义的任何实例空间X，每个都是树形结构或线性的，

c0(log(1/6) + s log(n/s))/e

c0(log(1/6) + s log(n/s))/e

<- S~( e,6 ) <~ cl(log(1/6 ) + s(log(sn/e)) ~) /e

<- S~( e，6)< ~ cl(log(1/6)+S(log(sn/e))~)/e

for all sufficiently small e and 6, where L is Algorithm 7.1 and C is the class of internal disjunctive concepts on X with size at most s, s <~ n. Moreover, this lower bound holds for any learning algorithm L.

对于所有足够小的e和6，其中L是算法7.1，而C是X上的内部析取概念类，其大小至多为s，s <~ n。此外，该下界适用于任何学习算法L

Proof.Similar to that of Corollary 5.7.[]

证据。类似于推论5.7。[]

This shows that the sample complexity of Algorithm 7.1 is also within a poly-logarithmic factor of optimal.

这表明算法7.1的样本复杂度也在最优的多对数因子内。

8.Conclusion

8.结论

This work provides one step toward putting the empirical investigations in concept learning since Winston [45] on a solid theoretical foundation.We have taken the popular theme of inductive bias and formalized it quantitatively, relating this measure directly to learning performance.In so doing we have shown that simple, near-optimal learning algorithms exist for the well-studied classes of conjunctive, internal disjunctive, k-DNF and k-CNF concepts.With the exception of the algorithms for k-DNF and k-CNF concepts when k is large, these learning algorithms are also computationally efficient.Further-more, the method we have developed is also quite general.In principle, it can be applied to any algorithm that learns single concepts from examples.It is required only that the algorithm produce consistent hypotheses, and that the hypothesis space used have a polynomially bounded growth function.

这项工作为将自温斯顿·[以来的概念学习的实证研究建立在坚实的理论基础上迈出了一步。我们采用了归纳偏见这一流行的主题，并将其定量化，将这一衡量标准与学习成绩直接联系起来。在这样做的过程中，我们已经表明，对于研究充分的合取、内部析取、k-DNF和k-CNF概念，存在简单的、近似最优的学习算法。除了当k较大时用于k-DNF和k-CNF概念的算法之外，这些学习算法在计算上也是高效的。此外，我们开发的方法也非常通用。原则上，它可以应用于任何从例子中学习单一概念的算法。只要求算法产生一致的假设，并且所使用的假设空间具有多项式有界增长函数。

Nevertheless, the theoretical framework we have used here for analyzing concept learning algorithms is still inadequate on several accounts.First, we have made no mention of the possibility of misclassifications in the training sample.It is not clear how our algorithms could be modified to tolerate such misclassifications.Since all our general theorems demand that the hypothesis be consistent with the training sample, they would also need to be modified to deal with learning situations that involve misclassifications of the training examples.Clearly any practical theory of concept learning must deal with this possibility.

然而，我们在这里用来分析概念学习算法的理论框架在几个方面仍然是不充分的。首先，我们没有提到训练样本中错误分类的可能性。不清楚我们的算法如何被修改来容忍这样的错误分类。因为我们所有的一般定理都要求假设与训练样本一致，所以它们也需要修改以处理涉及训练样本错误分类的学习情况。显然，任何概念学习的实践理论都必须处理这种可能性。

218

218

D.HAUSSLER

D.HAUSSLER

There are a number of approaches here.The methodology that Vapnik and others have used for pattern recognition starts from the general assumption that there is a fixed probability distribution on the set of all possible examples (i.e. instances and their labels).Hence, each instance may at times be classified as either "+" or "-", and the probability of a "+" classification may vary arbitrarily from instance to instance.Special cases of this general framework can be used to model many common types of "noisy" training data and/or "fuzzy" target concepts, depending on your point of view.A generalization of Theorem 3.3 given in [5, Appendix] (derived from [42]) is sometimes useful in such cases, in [2] a noisy training data viewpoint is adopted in their develop-ment and analysis of a noise resistant learning algorithm for k-CNF concepts in Boolean domains.Valiant has introduced a completely different model in which an adversary to the learning algorithm is allowed to maliciously modify the training examples [40].This model is further developed in [20].It is still not clear which, if any, of these "noise" models will be most appropriate for AI concept learning work.

这里有许多方法。Vapnik和其他人用于模式识别的方法是从一个普遍的假设开始的，即在所有可能的例子(即实例和它们的标签)的集合上有一个固定的概率分布。因此，每个实例有时可以被分类为“+”或“-”，并且“+”分类的概率可以随实例的不同而任意变化。这种通用框架的特殊情况可以用来模拟许多常见类型的“噪声”训练数据和/或“模糊”目标概念，这取决于您的观点。在这种情况下，[5附录中给出的定理3.3的推广(源自[42)有时是有用的，在[2中，在布尔域中的k-CNF概念的抗噪声学习算法的开发和分析中采用了噪声训练数据观点。瓦兰特引入了一个完全不同的模型，在这个模型中，学习算法的对手被允许恶意修改训练示例[40]。这一模式在[得到进一步发展。尚不清楚这些“噪音”模型中的哪一个(如果有的话)最适合人工智能概念学习工作。

Second, the methodology we have proposed may not be the most appropri-ate one for incremental learning algorithms, i.e. algorithms that maintain a working hypothesis and update this hypothesis as new examples are received.In many applications it is desirable to have a learning algorithm that works in this way.It does not appear that the algorithms based on the greedy heuristic that we have given can be used in such applications, short of storing all examples and recomputing the updated hypothesis from scratch when each new example is received.This problem has been addressed by the recent results of Littlestone [23].For Boolean domains, he develops extremely efficient in-cremental algorithms for pure conjunctive and k-DNF concepts with perform-ance very similar to those given here.These algorithms are based on a new variant of the perceptron learning algorithm, and are thus eminently suited for implementation in a "connectionist" architecture as well.However, they do not always maintain a consistent hypothesis, and much of their analysis appears to require mathematical techniques fundamentally different from those used here.The VC dimension is still used to provide lower bounds on the learning performance, however.

第二，我们提出的方法可能不是最适合增量学习算法的方法，增量学习算法是指保持一个有效假设并在收到新的例子时更新该假设的算法。在许多应用中，希望有一种以这种方式工作的学习算法。我们给出的基于贪婪启发式的算法似乎不能用于这样的应用，除非存储所有的例子，并在收到每个新的例子时从头开始重新计算更新的假设。利特尔斯通·[23号最近的研究结果已经解决了这个问题。对于布尔域，他为纯合取和k-DNF概念开发了非常有效的内建算法，其性能与这里给出的非常相似。这些算法基于感知器学习算法的一个新变体，因此非常适合在“连接主义”架构中实现。然而，他们并不总是保持一致的假设，而且他们的大部分分析似乎需要与这里使用的数学技术有根本不同的数学技术。然而，可变成本维度仍然用于提供学习性能的下限。

Third, the techniques here have been applied only to passive learning algorithms, i.e. algorithms that simply receive examples and form hypotheses.Angluin and others have demonstrated the power of learning algorithms that

第三，这里的技术仅应用于被动学习算法，即简单地接收示例并形成假设的算法。Angluin和其他人已经证明了学习算法的力量

can also make queries', e.g. ask questions of the form "is x an instance of the

也可以进行查询，例如，提问的形式是x

target concept?," where x is an instance constructed by the learning algorithm

目标概念？，其中x是由学习算法构造的实例

[1, 36,37] (see also [39]).Any comprehensive theoretical foundation for concept learning should also encompass such algorithms.

[1，36，37](另见[39)。任何概念学习的综合理论基础也应该包括这样的算法。

Fourth and finally, the methodology given here should be extended to richer types of instance spaces, to learning problems for multi-valued functions, to allow kinds of domain-specific background knowledge other than just orders

第四，也是最后一点，这里给出的方法应该扩展到更丰富类型的实例空间，学习多值函数的问题，允许除了顺序之外的各种特定领域的背景知识

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

219

219

and hierarchies on attribute values, and to learning problems that require simultaneously learning sets of related concepts.In [15] one extension to structured instance spaces (e.g. the blocks world [45]) is given.Background knowledge and multi-valued functions are considered in [28].In [42] the basic methodology used here is extended to real-valued functions.Some richer types of background knowledge are considered in [25].In [14] a few speculations on the issue of simultaneously learning sets of related concepts are given, based on ideas from [3, 7].Much more work remains to be done in all these areas.

以及需要同时学习相关概念集的学习问题。在[15]中，给出了对结构化实例空间的一个扩展(例如，块世界[45])。背景知识和多值函数在[28中被考虑。在[，这里使用的基本方法扩展到实值函数。一些更丰富的背景知识在[被认为是25]。在[，根据[的观点，对同时学习相关概念的问题给出了一些推测。在所有这些领域还有许多工作要做。

Apart from extending the analytical methodology presented here to over-come the above mentioned shortcomings, a number of other significant open problems remain within the present framework.We mention only two.First, how does the greedy heuristic we have used relate to Quinlan's information theoretic heuristic for learning decision trees [31]?Can the techniques given here be extended to exhibit efficient and provably effective learning algorithms for small decision trees in the Valiant framework?(See [34] and [11] for one approach here.) Second, is there an efficient and provably effective learning algorithm for simple DNF concepts, i.e. short DNF expressions with no fixed limit on the number of atoms per term?This problem was first posed by Valiant for Boolean domains, and still remains a central question today.

除了扩展这里提出的分析方法以克服上述缺点之外，许多其他重要的未决问题仍然存在于本框架内。我们只提到两个。首先，我们使用的贪婪启发式与昆兰学习决策树的信息论启发式[31有什么关系？此处给出的技术是否可以扩展，以便在Valiant框架中展示针对小决策树的高效且可证明有效的学习算法？(见[34和[11)这里有一种方法。第二，对于简单的DNF概念，即每个项的原子数目没有固定限制的短DNF表达式，是否有一个有效且可证明有效的学习算法？这个问题最初是由瓦兰特针对布尔域提出的，现在仍然是一个中心问题。

ACKNOWLEDGMENT

承认

I would like to thank Larry Rendell for suggesting the relationship between the Vapnik-Chervonenkis dimension and Mitchell and Utgoff's notion of inductive bias, and Ryszard Michalski for suggesting I look at the problem of learning internal disjunctive concepts.I also thank Les Valiant, Leonard Pitt, Manfred Warmuth, Nick Littlestone, Phil Laird, Ivan Bratko and Stephan Muggleton and Andrzej Ehrenfeucht for helpful discussions of these ideas.I thank Ranan Banerji and Anselm Blumer for pointing out errors in an earlier version of this manuscript.

我要感谢拉里·伦德尔，他提出了瓦普尼克-切尔沃内基斯维度与米切尔和乌特戈夫的归纳偏见之间的关系，以及雷沙德·米切斯基，他建议我研究学习内部分离概念的问题。我也感谢莱斯·瓦兰特、伦纳德·皮特、曼弗雷德·沃默斯、尼克·利特尔斯通、菲尔·莱尔德、伊万·布拉特科和斯蒂芬·麻瓜顿以及安杰伊·埃伦弗赫特对这些想法的有益讨论。我感谢安冉·班纳吉和安瑟姆·布鲁姆指出了这份手稿早期版本中的错误。

REFERENCES

参考

1.Angluin, A., Queries and concept learning, Machine Learning 2 (4) (1988) 319-342.2.Angluin, D., and Laird, P.D., Learning from noisy examples, Machine Learning 2 (4) (1988)

1.《查询和概念学习》，机器学习2 (4) (1988) 319-342。2.从嘈杂的例子中学习，机器学习2 (4) (1988)

343-370.

343-370。

3.Banerji, R., The logic of learning: a basis for pattern recognition and improvement of performance, Adv. Comput.24 (1985) 177-216.

3.学习的逻辑:模式识别和性能改进的基础。24 (1985) 177-216。

4.Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M., Occam's razor, Inf.Proc.Lett.24 (1987) 377-380.

4.艾伦富赫特，艾伦富赫特，豪斯勒，和沃茅斯，奥卡姆剃刀，Inf。继续。列特。24 (1987) 377-380。

5.Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M., Learnability and the Vapnik-Chervonenkis dimension, J. ACM, to appear.

5.布鲁默，阿，埃伦弗赫特，阿，豪斯勒，d .和温穆斯，m .，Learnability和瓦普尼克-切弗嫩基斯维度，J. ACM，出现。

6.Bruner, J.S., Goodnow, J. and Austin, G.A., A Study in Thinking (Wiley, New York, 1956).7.Bundy, A., Silver, B. and Plummer, D., An analytical comparison of some rule-learning

6.布鲁纳，J.S .，古德诺，j .和奥斯汀，美国，思考研究(威利，纽约州，1956年)。7.一些规则学习的分析比较

programs, Artificial Intelligence 27 (1985) 137-181.

程序，人工智能27 (1985) 137-181。

8.Chvatal, V., A greedy heuristic for the set covering problem, Math.Oper.Res. 4 (3) (1979) 233-235.

8.集合覆盖问题的贪婪启发式算法，数学。操作。第4 (3) (1979)号决议233-235。

9.Cover, T., Geometrical and statistical properties of systems of linear inequalities with applications to pattern recognition, 1EEE Trans.Elect.Comput.14 (1965) 326-334.10.Dietterich, T.G. and Michalski, R.S., A comparative review of selected methods for learning

9.盖，t。线性不等式系统的几何和统计性质及其在模式识别中的应用，1EE翻译。当选。计算机。14 (1965) 326-334。10.迪特里希，T.G .和迈克尔斯基，R.S .,学习方法选择的比较评论

220

220

D.ttAUSSLER

D.泰特斯勒

from examples, in: R.S. Michalski, J.G. Carbonell and T.M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach (Tioga, Palo Alto, CA, 1983) 41-81.I I. Ehrenfeucht, A. and Haussler, D., Learning decision trees from random examples, Tcch.Rept.UCSC-CRL-87-15, University of California, Santa Cruz, CA (1987).12.Ehrenfeucht, A., Haussler, D., Kearns, M. and Valiant, L.G., A general lower bound on the number of examples needed for learning, Inf.Comput., to appear.

举例来说，在:R.S .迈克尔斯基，J.G .卡本内尔和T.M .米切尔(编辑。)，机器学习:人工智能方法(Tioga，帕洛阿尔托，加利福尼亚州，1983) 41-81。从随机例子中学习决策树。报告。加利福尼亚州圣克鲁斯加州大学UCSC分校(1987年)。12.艾伦富赫特，A .，豪斯勒，d .，卡恩斯，m .和瓦兰特，L.G .，学习所需的例子数量的一般下限，信息。计算机。出现。

13.Garey, M. and Johnson, D., Computers and Intractability: A Guide to the Theor 3 q~ NP-Completeness (Freeman, San Francisco, CA, 1979).

13.计算机与难解性:定理3q ~ NP-完备性指南(弗里曼，旧金山，加利福尼亚，1979)。

14.Haussler, D., Quantifying the inductive bias in concept learning, Tech.Rept.UCSC-CRl.-80- 25, University of California, Santa Cruz, CA (1986).

14.量化概念学习中的归纳偏向。报告。UCSC-CRl。-80- 25岁，加利福尼亚大学圣克鲁斯分校(1986年)。

15.Haussler, D., Learning conjunctive concepts in structural domains, Machine Learning, m appear.

15.学习结构领域中的连词概念，机器学习，m。

16.Haussler, D., Applying Valiant's learning framework to AI concept learning problems, Tcch.Rept.UCSC-CRL-87-11, University of California, Santa Cruz, CA (1987): also in: R.S. Michalski and Kodratoff (Eds.), Machine Learning III (Morgan Kaufmann, Los Altos, CA 1987).

16.将瓦兰特的学习框架应用于人工智能概念学习问题。报告。加利福尼亚州圣克鲁斯市加州大学UCSC分校(1987):1987年。)，机器学习三(摩根考夫曼，洛杉矶，加利福尼亚州1987年)。

17.Haussler, D. and Welzl E., Epsilon nets and simplex range queries, Discrete ('omput.Geometry 2 (1987) 127-151.

17.艾司隆网和单形范围查询，离散(' omput。几何学2 (1987) 127-151。

18.Johnson, D.S., Approximation algorithms for combinatorial problems, J. Comput.Syst.Sci.9 (1974) 256-278.

18.组合问题的近似算法，计算机学报。系统。Sci。9 (1974) 256-278。

19.Johnson, D.S. and Niemi, K.A., On knapsacks, partitions and a new dynamic programming technique for trees, Math.Oper.Res. 8 (1) (1983) 1 - 14.

19.关于背包、分区和树的新动态编程技术，数学。操作。第8 (1) (1983)号决议1 - 14。

20.Kearns, M. and Li, M., Learning in the presence of malicious errors, in: Proceedings 20th ACM Symposium on Theory of Computing, Chicago, IL (1988) 267-28(I. 21.Kearns, M., Li, M., Pitt, L. and Valiant, L., Recent results on Boolean concept learning, in:

20.在恶意错误的存在下学习，收录于:美国计算机学会第20届计算理论研讨会论文集，芝加哥，1988年，第267-28页。卡恩斯，m，李，m，皮特，l和瓦兰特，l，布尔概念学习的最新结果，在:

Proceedings 4th International Workshop on Machine Learning, Irvine, CA (1987) 337-352.22.Laird, P.D., Inductive inference by refinement, Tech.Rept.YALEU/DCS/TR-376, Yale University, New Haven, CT (1986).

第四届机器学习国际研讨会论文集，加州欧文(1987) 337-352。22.通过提炼归纳推理，技术。报告。耶鲁大学，康涅狄格州纽黑文(1986年)。

23.Littlestone, N., Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Machine Learning 2 (4) (1988) 245-318.

23.当不相关的属性大量存在时快速学习:一种新的线性阈值算法，机器学习2 (4) (1988) 245-318。

24.Michalski, R.S., A theory and methodology of inductive learning, in: R.S. Michalski, J.G. Carbonell and T.M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach (Tioga, Palo Alto, CA, 1983) 83-134.

24.归纳学习的理论和方法，收录于:迈克尔斯基，卡本内尔和米切尔。)，机器学习:人工智能方法(Tioga，帕洛阿尔托，加利福尼亚州，1983) 83-134。

25.Milosavljevic, A., Learning in the presence of background knowledge, Tech.Rept.UCSC-CRL-87-27, University of California, Santa Cruz, CA (1987).

25.米洛萨夫列维奇，在背景知识中学习，技术。报告。加利福尼亚州圣克鲁斯加州大学UCSC分校(1987年)。

26.Mitchell, T.M., The need for biases in learning generalizations, Tech.Rept.CBM-TR-117, Department of Computer Science, Rutgers University, New Brunswick, NJ (1980).27.Mitchell, T.M., Generalization as search, Artificial Intelligence 18 (1982) 203-226.28.Natarajan, B.K. and Tadepalli, P., Two new frameworks for learning, in: Proceedings 5th

26.米切尔，麻省理工学院，《学习归纳中偏见的必要性》，科技。报告。新泽西州新不伦瑞克罗格斯大学计算机科学系，CBM-TR-117(1980)。27.米切尔，麻省理工学院，《作为搜索的一般化》，人工智能18 (1982) 203-226。28.Natarajan，B.K .和Tadepalli，p .,两个新的学习框架，载于:第5卷

International Workshop on Machine Learning, Ann Arbor, MI (1988).29.Nigmatullin, R.G., The fastest descent method for covering problems (in Russian), in:

机器学习国际研讨会，密歇根州安阿伯(1988年)。29.覆盖问题的最快下降法(俄语)，见:

Proceedings Symposium on Questions of Precision and Efficiency of Computer Algorithms 5, Kiev, U.S.S.R. (1969) 116-126.

计算机算法的精度和效率问题会议录5，基辅，苏联(1969) 116-126。

30.Pearl, J., On the connection between the complexity and credibility of inferred models, Int.J. General Syst.4 (1978) 255-64.

30.关于推理模型的复杂性和可信性之间的联系，国际。j .通用系统。4 (1978) 255-64。

31.Quinlan, J.R., Induction of decision trees, Machine Learning 1 (1) (1986) 81-106.32.Rendell, L., A general framework for induction and a study of selective induction, Machine

31.决策树归纳，机器学习1 (1) (1986) 81-106。32.归纳的一般框架和选择性归纳的研究，机器

Learning 1 (2) (1986) 177-226.

学习1 (2) (1986) 177-226。

33.Rissanen, J., Modeling by shortest data description, Automatica 14 (1978), 465 - 471.34.Rivest, R., Learning decision-lists, Machine Learning 2 (3) (1987) 229-246.35.Schlimmer, J.C., Incremental adjustment of representations for learning, in: Proceedings 4th

33.用最短数据描述建模，自动化14 (1978)，465 - 471。34.学习决策表，机器学习2 (3) (1987) 229-246。35.施林梅，朱建中，学习表征的增量调整，载于:第4卷

International Workshop on Machine Learning, Irvine, CA (1987) 79-90.

机器学习国际研讨会，加州欧文(1987) 79-90。

QUANTIFYING INDUCTIVE BIAS

量化归纳偏差

221

221

36.Sammut, C. and Banerji, R., Learning concepts by asking questions, in: R. Michalski, J.

36.萨姆特和班纳吉，通过提问来学习概念，见:迈克尔斯基

Carbonell and T. Mitchell (Eds.), Machine Learning II (Morgan Kaufmann, Los Altos, CA,

卡本内尔和米切尔(编辑。)，机器学习二(摩根考夫曼，洛杉矶，加利福尼亚州，

1986).

1986年)。

37.Subramanian, D. and Feigenbaum, J., Factorization in experiment generation, in: Proceedings AAAI-86, Philadelphia, PA (1986) 518-522.

37.实验生成中的因子分解，载于:AAAI会议录-86，费城，宾夕法尼亚州(1986) 518-522。

38.Utgoff, P., Shift of bias for inductive concept learning, in: R. Michalski, J. Carbonell and T.

38.归纳概念学习的偏差转移，收于:迈克尔斯基，卡本内尔和t

Mitchell (Eds.), Machine Learning II (Morgan Kaufmann, Los Altos, CA, 1986).39.Valiant, L.G., A theory of the learnable, Commun.ACM, 27 (11) (1984) 1134-1142.40.Valiant, L.G., Learning disjunctions of coniunctions, in: Proceedings IJCA1-85, Los Angeles,

米切尔(Eds。)，机器学习二(摩根考夫曼，洛杉矶，加利福尼亚州，1986年)。39.瓦兰特，洛杉矶，一个可学习的理论，社区。ACM，27 (11) (1984) 1134-1142。40.瓦兰特，洛杉矶，1985年，《认知功能的分离》，

CA (1985) 560-566.

CA (1985) 560-566。

41.Pitt, L. and Valiant, L.G., Computational limitations on learning from examples, Tech.Rept.TR-05-86, Aiken Computing Lab., Harvard University, Cambridge, MA (1986);also: J.

41.皮特和瓦兰特，《从例子中学习的计算限制》，科技。报告。TR-05-86，爱肯计算实验室。哈佛大学，剑桥，硕士(1986年)；还有:j

ACM, to appear.

ACM，出现。

42.Vapnik, V.N., Estimation of Dependences Based on Empirical Data (Springer, New York,

42.基于经验数据的依赖性估计(斯普林格，纽约，

1982).

1982年)。

43.Vapnik, V.N. and Chervonenkis, A.Ya., On the uniform convergence of relative frequencies of

43.瓦普尼克，维恩和切尔沃内基斯，阿雅，关于相对频率的一致收敛

events to their probabilities, Theor.Probab.Appl. 16 (2) (1971) 264-280.44.Wenocur, R.S. and Dudley, R.M., Some special Vapnik-Chervonenkis classes, in: Discrete Math.33 (1981) 313-318.

事件的概率，定理。可能吧。申请号16 (2) (1971) 264-280。44.《离散数学》中的一些特殊的Vapnik-Chervonenkis类。33 (1981) 313-318。

45.Winston, P., Learning structural descriptions from examples, in: P.H. Winston (Ed.), The Psychology of Computer Vision (McGraw-Hill, New York, 1975).

45.温斯顿，p。从例子中学习结构描述。)，计算机视觉心理学(纽约麦格劳-希尔，1975)。

Received November 1986;revised version received February 1988

1986年11月收到；修订版于1988年2月收到