基于HMM的Web信息抽取算法的研究与应用

祝伟华 卢 熠 刘斌斌
(重庆大学软件学院 重庆 400044)

摘 要 随着因特网技术的迅速发展,网上信息成几何级数增长,如何从这些海量联机非结构化文本中自动抽取出结构化信息成为目前重要的研究课题。研究了基于隐马尔可夫模型的Web信息抽取算法,着重探讨了隐马尔可夫模型在文本信息抽取中应该如何应用,数据应该如何标记,并对隐马尔可夫模型在文本信息抽取中的应用提出了几个改进的方法,建立了基于HMM的Web信息抽取模型,并对信息抽取后的数据进行了分析对比,验证了改进算法的有效性。

关键词 隐马尔可夫模型,信息抽取,机器学习
中图法分类号 TP311.56 **文献标识码** A

Improvement of Web Information Extraction Algorithm Based on HMM

ZHU Wei-hua LU Yi LIU Bin-bin
(School of Software Engineering, Chongqing University, Chongqing 400044, China)

Abstract With the development of the Internet technologies, the information on the Internet increases exponentially. One important research focuses on how to extract structured data from these great capacities of online documents in unstructured texts. This thesis mainly studied relative algorithms on Web information extraction based on hidden Markov model(HMM), discussed how to use HMM and how to mark data in text information extraction, offered several methods to improve the hidden Markov model in information extraction, introduced the establishment of Web information extraction model based on HMM, Comparatively analysed the output data of information extraction, verified the validity of the algorithm through experiments.

Keywords HMM, Information extraction, Machine learning

基于隐马尔可夫模型(简称HMM)的文本信息抽取还是相对较新的技术,应用HMM模型进行文本信息抽取算法研究主要集中在两个主要方面。首先,从数据中研究学习模型结构。HMM绝大多数的应用都是假定一个固定的模型结构(状态数和状态间的转移链),该模型结构事先根据所适用的范围手工选定,或者从训练数据中学习模型结构。其次,根据给定的或者训练获得的模型对未标记的数据进行识别。已标记的数据被用来训练HMM模型,很多时候,未标记的数据也可以结合到训练数据中来增大模型训练量,以学习到更好的模型参数。

1 HMM模型

HMM模型是一个二重马尔可夫随机过程,它包括具有状态转移概率的马尔可夫链和输出观测值的随机过程。其状态是不确定或不可见的,只有通过观测序列的随机过程才能表现出来。现有的一些模型学习、训练和评估算法,都具有很高的计算效率。

一个HMM包含两层:一个可观察层和一个隐藏层。可观察层是待识别的观察序列,隐藏层是一个马尔可夫过程,即一个有限状态机,其中每个状态转移都带有转移概率。

一个HMM可以看成一个五元组$\{S,V,A,B,\Pi\}$:

$S=\{S_1,S_2,\cdots,S_N\}$

$V=\{V_1,V_2,\cdots,V_M\}$

$A=\{a_{ij}=P(q_{i+1}=S_j|q_i=S_i),1\leqslant i,j\leqslant N\}$

$B=\{b_j(V_k)=P(V_k \text{ at } t|q_t=S_j),1\leqslant i\leqslant j,1\leqslant k\leqslant M\}$

$\Pi=\{\pi_i=P(q_1=S_i,1\leqslant i\leqslant j)\}$

其中,S是状态集,模型共有N个状态;V是词汇集,模型共有M个可能输出的词汇;A是$N*N$的状态转移矩阵,a_{ij}是从状态S_i转换到状态s_j的概率;B是$N*M$的释放概率矩阵,$b_j(V_k)$是在状态与时释放单词V_k的概率;Π是初始状态概率集合,π_i是第i个状态作为初始状态的概率。

1.1 HMM的主要算法

应用HMM主要解决以下3个方面的问题:①评估问题:对于给定的模型λ,求某个观察值序列O的概率$P(O|\lambda)$,本文采用前向算法;②学习问题:对于给定的观察值序列O,调整参数λ,使得观察值出现的概率$P(O|\lambda)$最大,本文采用用于部分标记训练样本的Baum-Welch算法;③解码问题:对于给定的模型λ和观察值序列,求可能性最大的状态序列,本文采用Viterbi算法。对于这3个主要问题,文本信息抽取需要解决其中的学习问题与解码问题。

到稿日期:2009-10-30 返修日期:2009-12-10 本文受国家自然科学基金项目(No. 101022820080079)资助。
祝伟华(1970—),副教授,主要研究方向为软件工程及应用、人工智能、数据挖掘、搜索引擎等,E-mail:swzhu@yahoo.com;**卢 熠**(1985—),硕士生,主要研究方向为数据挖掘等;**刘斌斌**(1985—),硕士生,主要研究方向为数据挖掘等。

1.1.1 评估问题的解决算法

定义向前概率 $a_t(i)=P(O_1O_2\cdots O_t,q_t=S_i|\lambda)$。它是部分观察序列 $O_1O_2\cdots O_t$（直到时间 t）的概率，通过给定观察序列 $O_1O_2\cdots O_t$ 和模型 λ，可以归纳计算 $a_t(i)$，过程如表 1 所列。

表 1 前向算法

初始化：

$a_1(i)=\pi_i b_1(O_1)$

其中，$1\leqslant i\leqslant N$

迭代：$a_{t+1}(j)=\lfloor\sum_{i=1}^{N}a_t(i)a_{ij}\rfloor b_j(O_{t+1})$

其中 $1\leqslant t\leqslant T-1,1\leqslant j\leqslant N$

终止：

$P(O|\lambda)=\sum_{i=1}^{N}a_T(i)$

前向算法通过点阵结构解决问题，运算复杂性为 $N*N*T$，比直接计算要快得多。

1.1.2 学习问题的解决算法

Baum-Welch 训练算法是一种迭代的最大期望值算法：给定初始的参数配置，调整模型参数到局部最大化未标记数据的可能性。Baum-Welch 训练算法能够到达局部最大并且对初始参数设置敏感。算法可简单描述如表 2 所列。

表 2 Baum-Welch 算法

初始化：

$y_1(i)=\bar{\pi}_i$，时间 $t=1$ 时处于状态 S_i 的期望值

$\lambda=M\langle A_0,B_0,\bar{\pi}\rangle$

迭代计算：

令 $\lambda_0=\lambda$

$\bar{\pi}_i=\gamma_1(i),1\leqslant i\leqslant N$

$$\bar{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_i(i,j)}{\sum_{i=1}^{T-1}\gamma_i(i)}=\frac{\sum_{t=1}^{T-1}a_i(i)a_{ij}b(o_{i+1})\beta_{i+1}(j)}{\sum_{t=1}^{T-1}a_i(i)\beta_i(j)}$$

$$\bar{b}_j(k)=\frac{\sum_{t=1,o_t=v_k}^{T}\gamma_i(i)}{\sum_{t=1}^{T}\gamma_i(i)}=\frac{\sum_{t=1}^{T}a_i(i)\beta_i(i)\delta(o_i,v_k)}{\sum_{t=1}^{T}a_i(i)\beta_i(j)}$$

其中，

$$\delta(o_t,v_k)=\begin{cases}1, & o_t=v_k\\ 0, & \text{otherwise}\end{cases}$$

$\lambda=M\langle\bar{A},\bar{B},\bar{\pi}\rangle$

终止

$|\log P(O|\lambda)-\log P(O|\lambda_0)|<\varepsilon$，其中 ε 是预先设定的阈值。

1.1.3 解码问题的解决算法

寻找单一的最佳状态序列 $Q=q_1q_2\cdots q_r$，对于给定的观察序列 $O_1O_2\cdots O_t$，需要定义量：

$\delta_t(j)=\max q_1q_2\cdots q_{t-1}P[q_1q_2\cdots q_t=i,O_1O_2\cdots O_t|\lambda]$

它是沿着一条单一路径，在 t 时刻最高的概率，它说明了处于第一位置的 t 观察，并且结束于状态 S_j。通过推导，可以得到：

$\delta_{t+1}(j)=[\max i\,\delta_t(i)a_{ij}]b_j(O_{t+1})$

在实际寻找状态序列的过程中，需要记录对于每个 t 和每个 j，使得 $\delta_{t+1}(j)$ 最大的参数。通过数组 $\Psi_t(j)$ 来完成这个工作。完整的寻找最佳状态序列的 Viterbi 算法描述如表 3 所列。

表 3 Viterbi 算法

初始化

$\delta_1(i)=\pi_1b_1(O_1)$

$\Psi_1(i)=0,1\leqslant i\leqslant N$

归纳

$\delta_t(j)=\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]b_j(O_t)$

$\Psi_t(j)=\operatorname{argmax}_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]$

其中 $2\leqslant t\leqslant T$ and $1\leqslant j\leqslant N$

终止

$p^*=\max_{1\leqslant i\leqslant N}[\delta_t(i)]$

$q_T^*=\operatorname{argmax}_{1\leqslant i\leqslant N}[\delta_T(i)]$

返回

$q_t^*=\Phi_{t+1}(q_{t+1}^*)$

$t=T-1,T-2,\cdots,1$

2 HMM 模型在信息抽取中的改进

针对 HMM 模型在 Web 信息抽取中的实际应用，对其做了如下改进。

2.1 平滑处理

在隐马尔可夫模型实际应用中，会遇到未知观测值的情况，而在隐马尔可夫模型中对观测值概率的定义为：令 $T_j(k)$ 为从状态 j 释放观察值 k 的次数，则 $\sum_{k=1}^{M}T_j(k)$ 为状态 j 释放所有观察值的次数，有如下公式：

$$b_j(k)=\frac{T_j(k)+1}{\sum_{k=1}^{M}T_j(k)}$$

其中，$1\leqslant j\leqslant N,1\leqslant k\leqslant M$。

这会导致观察值概率为零，平滑处理就是给未知观察值赋予正确的概率。平滑处理的方法很多，基本思路是减小已知观察值的概率，剩余部分分配给未知观察值，传统的 Laplace 方法是将分子加 1，分母加观察值总数 M，即：

$$b_j(k)=\frac{T_j(k)+1}{\sum_{k=1}^{M}T_j(k)+M}$$

于是对于一个未知观察值在状态 j 的释放概率为：

$$b_j(k)=\frac{1}{\sum_{k=1}^{M}T_j(k)+M}$$

Laplace 方法简单易行，对所有抽取域都是平等对待的，即不区分各个域中可能出现未知观测值情况的多少，事实上各个域中出现未知观测值的概率相差很大，由此本系统采用改进的 Laplace 方法，即令已知观察值释放概率为：

$$b_j(k)=b_j(k)-\delta$$

其中，剩余的概率分配给未知观测值即可。

$$\delta=\frac{1-\frac{\sum_{k=1}^{M}T_j(k)}{M}}{\sum_{k=1}^{M}T_j(k)+M}$$

2.2 符号串聚类

如上介绍在域中出现的观测值包括“数字”，对于数据抽取来说释放的数据一定是待抽取文本中的数据，而文本中是“2”还是“3”对抽取过程本身并没有影响，可以统一作为“数字”来进行处理，而“数字”则是该域中的重点观测值，可将其附带后面的量词等抽取出来成为有价值的数据。这种将符号串归类统一进行考虑的方法称为字符串聚类。

给定两个符号串序列 $A=\{a_1,a_2,\cdots,a_3\}$ 和 $B=\{b_1,b_2,\cdots,b_3\}$，若存在单调增加的整数序列：$i_1<i_2<\cdots<i_x$ 和 $j_1<j_2<\cdots<j_x$，满足 $a_{ik}=b_{ik}=c_k,k=1,2,\cdots,x$，则记符号串 $C=\{c_1,c_2,\cdots,c_3\}$ 为 A 和 B 的公共符号串，用符号 $CS(A,B)$ 表示。若公共符合串 C 中的两个相临元 c_s 与 c_{s+1} 满足条件 i_s+

$1=j_s$ 与 $i_{s+1}+1=j_{s-1}$，则称 c_s 与 c_{s-1} 为公共符号串 C 中的一对连续符号。定义公共符号串分值公式为：

$$sore(C)=|C|+p\times\delta$$

其中，$|C|$ 为公共符号串 C 的长度，p 是 C 中连续符号的对数，δ 为每对连续符号的奖惩值。

利用公共符号串分值可定义符号串相似度为：

$$\mathrm{sim}(A,B)=\frac{score(CS(A,B))}{f(|A|,|B|)}$$

$$f(|A|,|B|)=\min(|A|,|B|)\times(1+\delta)-\delta$$

其中，分子为 A 和 B 的公共符号串的分值，$|A|$ 和 $|B|$ 分别表示 A 和 B 的长度。定义两个符号串完全相同或者其中一个符号串是另一个符号串的连续子串时相似度最大，等于 1。

符号串聚类即是将同一抽取域中的数据两两进行相似度计算，将相似度大于临界值的数据进行合并。由于无法预先确定符号串的类别，因此采用无指导的分类，对于数字和特殊符号则需要进行特征符号识别。

2.3 状态合并

状态合并可分为横向合并和纵向合并，横向合并是指具有相同类型的多个连续状态可以合并为一个状态，并引入自循环。纵向合并是指具有相同类型的多个状态如果具有相同的起止状态则可合并为一个状态，纵向合并减少了模型的分支。

2.4 隐马尔可夫模型中结合规则

这里的规则是指经过人工定义可作为必然事件执行特定处理的特殊过程。在一个实际应用的信息抽取系统中，除了存在不确定状态及其不确定释放观察值的情况外，还存在确定状态释放观察值的情况。定义确定状态，在隐马尔可夫模型中作为已知状态处理，其观察值释放概率设为 1，这主要适用于固定特征词和一些结构化数据的处理，例如对抽取域名称字符的提取以及简单表结构中数据的抽取，既可以减少计算观察值释放概率的时间，又可减少由于数据稀疏等问题带来的对计算模型参数的负面影响。同时由于部分状态已确定，运用 Viterbi 算法可减少循环次数，从而减少运行时间，提高效率。规则的合理使用也可以在更大范围内减轻人工标注的工作量，但是制定规则本身也需要人工进行分析和提取，制定统一的接口则可以方便规则的定义和添加，降低对系统进行升级的工作量。

3 基于 HMM 的信息抽取模型建立及实验结果

本文通过二手汽车垂直搜索项目中的信息抽取进行实验研究。该搜索引擎项目的主要目的是为二手汽车的购买人群提供全方位的二手汽车销售信息，以及为二手汽车的销售人群提供发布信息的平台，并为其提供有效的广告服务。待建的信息抽取模型的具体功能有：输入数据为原始网页、转换为带有简单标记信息的树状结构半格式化数据，由经过训练的隐马尔可夫模型处理后提取出精确信息存入数据库的对应表中，成为格式化数据以便进行查取。其中的重点和难点是对隐马尔可夫模型进行训练和利用其进行信息的精确抽取。

该模型主要处理过程如下。

① 网页预处理：对抓取来的相关网页进行初步处理，生成信息节点树，降低信息抽取的输入状态序列，并对列表、图片等信息进行标记，这对模型训练和数据聚类起着关键作用，且可大大提高信息抽取的准确率。

② 模型训练：在利用模型进行信息抽取之前必须先通过大量的精确标注数据源对模型进行训练，实际应用中可利用已有模型直接进行信息抽取，同时也要根据信息抽取过程中遇到的新情况调整模型参数，这体现了隐马尔可夫模型自动调整的特性。

③ 抽取信息：将预处理生成的信息节点树作为输入信息，在扫描各节点的同时按照当前模型参数计算节点信息作为观察值序列输出的概率，当概率高于临界值时实施对该域内数据的提取。

④ 数据精化：对数据进行粗略抽取后需去掉数据中的标注等辅助信息，调用数据库管理模块将数据存入表中，至此完成了从 HTML 半结构化信息向数据库结构化信息的转换。

就 IE 而言，召回率可粗略地看成是测量被正确抽取的信息的比例，而抽准率用来测量抽出的信息中有多少是正确的。计算公式如下：

P=抽出的正确信息点数/所有抽出的信息点数

R=抽出的正确信息点数/所有正确的信息点数

两者取值在 0 和 1 之间，通常存在反比的关系，即 P 增大会导致 R 减小，反之亦然。

评价一个系统时，应同时考虑 P 和 R，但同时要比较两个数值，毕竟不能做到一目了然。许多研究人员提出合并两个值的办法。其中包括 F 值评价方法：

$$F=\frac{(\beta^2+1)PR}{\beta P+R}$$

其中，β 是一个预设值，决定对 P 侧重还是对 R 侧重。通常设定为 1。这样用 F 一个数值就可以看出系统的好坏。以下就采用上述公式来表示信息抽取成功率指数。

通过对 100 个汽车信息页面的抽取结果进行对比，使用传统 HMM 模型和改进后的 HMM 模型进行抽取的部分结果如表 4 所列。

表 4 信息抽取结果与对比

抽取域	传统 HMM 模型的 PR 合并值	改进 HMM 模型的 PR 合并值
Engine	0.683	0.914
Price	0.738	0.925
Mileage	0.789	0.965
Type	0.621	0.896
Fule	0.725	0.917
Transmission	0.744	0.909
……	……	……

从表 4 显示的实验结果可以看出，经过改进的 HMM 比简单的 HMM 抽取效果要好得多，这是因为当测试数据中存在大量的未知观测值时，对它们赋予零概率将大大影响数据标记结果，而改进后的平滑算法和符号串聚类等方法可以有效地解决这些问题。可见通过构建改进的隐马尔可夫模型并配合适当的规则和大规模的词典后，进行信息抽取的精确度提升非常明显。

结束语 本文提出了基于 HMM 模型的信息抽取技术，并针对 HMM 模型在应用中需要注意的 4 个方面，包括平滑处理、字符串聚类、状态合并和与规则相结合，提出了改进的方法。根据改进后的 HMM 模型，对二手汽车信息搜索引擎中的应用进行信息抽取，并将 PR 结合后的实验结果加以对比，证明对模型的改进是极其有效的，但同时还有需要改进的地方，如需要更详细的状态集和规则处理，信息抽取的抽准率和召回率还有提高的空间，这些都是今后改进的研究方向。

参 考 文 献

[1] 钟敏娟，郝谦，刘云中. 基于多模板隐马尔可夫模型的文本信息

抽取算法[J]. 计算机工程,2006,32(2):203-205
[2] 王雷,陈治平,李志成. 基于文本分块的多模板隐马尔可夫模型的文本信息抽取[J]. 山东大学学报:理学版,2006,41(3):21-24
[3] 杜世平,李海. 二阶隐马尔可夫模型及其在计算语言学中的应用[J]. 四川大学学报:自然科学版,2004,41(2):284-289
[4] 林亚平,刘云中,陈治平. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报,2005,33(2):236-241
[5] 胡宇舟,王雷,顾学道. 基于多模板隐马尔可夫模型的文本信息抽取算法[J]. 计算机应用,2008,28(3):609-702
[6] 邢永康,马少平. 一种基于 Markov 链模型的动态聚类方法[J]. 计算机研究与发展,2003,40(2):129-135
[7] Frietag D, McCallum A. Information extraction with HMMs and shrinkage[A] //Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction[C]. Orlando: AAAI Press, 1999:31-36
[8] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization[A] //Proceedings of the Eighteenth Conference on Artificial Intelligence[C]. Edmonton: AAAI Press, 2002:584-589
[9] Ray S, Craven M. Representing sentence structure in hidden Markov models for information extraction[A] //Proceedings of the Sev2 enteenth International Joint Conference On Artificial Intelligence[C]. Washington: Morgan Kaufmann, 2001: 1273-1279
[10] Scheffer T, Decomain C, Wrobel S. Active hidden Markov models for information extraction[A] //Proceedings of the Fourth International Sym2 posium on Intelligent Data Analysis[C]. Lisbon: Springer, 2001:301-309
[10] 朱征宇,周智,罗颖,等. 基于浏览行为量化分析的兴趣网页提取[J]. 重庆工学院学报:自然科学版,2009,23(7):79-84

(上接第 166 页)

集,样本数据分布如图 1 所示。核函数为高斯核函数:取 $\sigma^2 = 0.02$, $K(x_i, x_j) = \exp(-\sigma^2 \| x_i - x_j \|^2)$。

利用算法 1 得到的分类结果如图 2 所示。实验 1 的结果表明本文提出的下降算法是可行有效的。

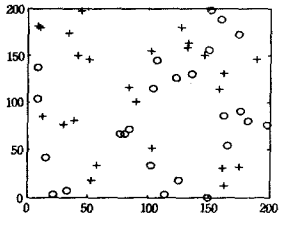

图 1 实验 1 的样本数据

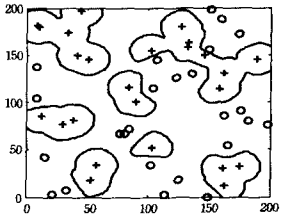

图 2 利用算法 1 的分类结果

实验 2 该实验对来自文献[13,14]的不同规模和维数的实际数据集进行试验。这些数据都进行了零均值和标准化处理。分别采用 LSVM 迭代算法和本文提出的算法 1 对数据集进行分类。为方便起见,均采用高斯核函数。结果如表 1 所列。

表 1 算法 1 与 LSVM 对实际数据的性能比较

Input Dataset		Parameters	LSVM		Descent method	
Datasets	Dimension/ train/test	σ^2	Time (s)	Errors (%)	Time (s)	Errors (%)
Titanic	3/150/2051	1	0.023	21.84	0.014	21.84
Diabetis	8/468/300	10	0.291	23.67	0.031	23.67
Breast cancer	9/200/77	0.2	0.047	22.15	0.015	22.08
Heart	13/170/ 100	0.1	0.031	12.35	0.016	11.93
Image	18/1300/ 1010	0.1	7.241	3.61	0.135	3.18
German	20/700/ 300	1	1.328	23.63	0.048	22.86

实验结果表明,随着数据规模的增大,算法 1 所需时间并没有明显增加。在错误率与 LSVM 算法相当的情况下,其耗时较短,表现出明显的优势。

结束语 基于支持向量机一个修正模型,提出了互补型支持向量机,进一步给出了求解该支持向量机的一个新的下降算法。对实际数据集的实验结果表明,该算法的训练速度要比 LSVM 算法快,且训练正确率相当甚至更好。同时,算法由于不需要求解 Hesse 阵以及任何矩阵的求逆运算,因此适用于求解大规模的非线性分类问题,为求解 SVM 优化问题提供了一个新的途径。

参 考 文 献

[1] Vapnik V N. The Nature of Statistical Learning Theory(econd Edition)[M]. SNew York: Springer, 2000
[2] 王巍,赵国杰,陈作斌. 一种新的混合智能分类模型及其应用[J]. 兰州大学学报,2008,44(1):77-81
[3] 廖东平,魏玺,章黎湘,等. 一种新的支持向量机快速训练算法[J]. 系统工程与电子技术,2007,11:1954-1957
[4] 王玲,薄列峰,刘芳,等. 稀疏隐空间支持向量机[J]. 西安电子科技大学学报,2006,33(6):896-901
[5] Zhou S S, Liu H W, Zhou L H, et al. Semismooth Newton support vector machine[J]. Pattern Recognition Letters, 2007, 28: 2054-2062
[6] Mangasarian O L, Musicant D R. Lagrangian Support Vector Machines[J]. Journal of Machine Learning Research, 2001, 1: 161-177
[7] Fischer A. A Special Newton-type optimization methods[J]. Optimization, 1992, 24: 269-284
[8] Mangasarian O L, Musicant D R. Successive over relaxation for support vector machines[J]. IEEE Trans. on Neural Networks, 1999, 10: 1032-1037
[9] Zhang X Z, Jiang H F, Wang Y J. A smoothing Newton-type method for generalized nonlinear complementarity problem[J]. Journal of Computational and Applied Mathematics, 2008, 212: 75-85
[10] Facchinei F, Pang J S. Finite-Dimensional Variational Inequalities and Complementarity Problems[M]. New York: Springer, 2003
[11] Chen J S, Pan S H. A descent method for a reformulation of the second-order cone complementarity problem[J]. Journal of Computational and Applied Mathematics, 2008, 213: 547-558
[12] Ho T K, Kleinberg E M. Checkerboard dataset[EB/OL]. http://www.cs.wisc.edu//math-prog//mpml.html, 1996
[13] Murphy P M, Aha D W. UCI Repository of machine learning databases[EB/OL]. http://www.ics.ci.edu/~mlearn/ML Repository.html. 1998
[14] Blake C L, Merz C J. UCI Repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/mleam/ML Repository.html. 1998