异构数据的结构熵聚类算法

李志华[1] 顾 言[1] 陈孟涛[1] 王士同[1] 陈秀宏[2]
(江南大学信息工程学院 无锡 214122)[1] (江南大学数字传媒学院 无锡 214122)[2]

摘 要 研究了语义数据的聚类问题,提出了一种基于样本内在结构的结构熵聚类SEC算法。通过给出语义属性相异性度量测度的新定义,挖掘蕴含于数据样本中的结构信息,提出了一种根据结构信息计算样本信息熵的优化方法,即通过熵来确定样本的聚类中心,从而完成样本的聚类,并把此方法向异构数据进行了拓展。SEC算法能实现不平衡数据的聚类,能自动确定初始类中心和聚类数目,具有无需迭代、效率高和相当的鲁棒性优势。实验表明,算法是有效的,与文献中的已有方法相比,聚类准确率得到显著提高,具有一定的实用价值。

关键词 异构数据,相异性度量,聚类线索,结构熵,聚类算法
中图法分类号 TP393 **文献标识码** A

Structure-based Entropy Clustering Algorithm for Heterogeneous Data

LI Zhi-hua[1] GU yan[1] CHEN Meng-tao[1] WANG Shi-tong[1] CHEN Xiu-hong[2]
(Shool of Information and Engineering, Jiangnan University, Wuxi 214122, China)[1]
(Digital Media College, Jiangnan University, Wuxi 214122, China)[2]

Abstract The dissimilarity measure and clustering approach about the heterogeneous dataset were studied, and a structure-based entropy clustering SEC algorithm was presented in this paper. Data often do appear in homogeneous groups, the SEC utilizes these structural information to improve the clustering accuracy. Unlike the distribution of numeric data, nominal data are often unbalancedly distributed, whose distribution are often unrelated with their distance measure, due to the above, a new structural information-based entropy computing technology was proposed. By mining the clues in structural information, constructing the weight implying the different distribution information of nominal and numeric attributes, the SEC can automatically identifies the initial locations and number of cluster centriods, and exhibits its robustness to initialization and no iteration in algorithm. Experimental results comparing with other references demonstrate that the proposed method has promising performance.

Keywords Heterogeneous data, Dissimilarity measure, Clustering clue, Structural entropy, Clustering algorithm

1 引言

聚类分析的主要任务是通过挖掘样本中的各种聚类线索,从而有效地选择聚类算法中的合适参数,以达到确定最佳聚类数目的目的。当前,很多数据样本,如网络安全事件、网页分类和疾病诊断数据等,这些数据集的组成复杂,存在着大量"构造相似"的同质样本[1],同时其属性组成呈异构性特点,既有语义属性又有连续属性,而语义属性的数据类型又可能是字符、二值属性和一些具有枚举数据特征的数值等,表现出鲜明的结构特征。

研究语义属性样本的聚类,必须首先能有效地确定聚类中心,并实现对异构属性相异性的同时度量。由于语义属性数据分布固有的无序性,样本的分布不平衡、分布与空间距离无关[2],因此对其相异性度量比较困难,造成现今大多数聚类算法不能实现语义属性数据、异构属性数据的聚类。如著名的FCM算法就不能实现这两种数据集的聚类。只有少数几个从k-均值算法演变而来的算法[3-5]能实现语义属性数据的聚类,只有基于混合属性数据(k-prototype)的聚类算法[6]等能实现异构属性数据的聚类。但算法或多或少存在一些不足,如需要处理大量的二进制数据、对初始类中心敏感和需要选取特殊的模式[3-5]等。

虽然基于熵的聚类算法[7]已经存在,但同样不适宜异构属性数据的聚类,更没有充分挖掘样本中有利于聚类的结构线索。本文通过给出一种语义属性数据相异性度量测度的新定义,深入挖掘样本中的结构信息,根据信息理论构造了一种体现异构属性数据分布结构的计算熵的新方法,在充分挖掘了样本中的聚类线索后提出了结构熵聚类(Structure-based Entropy Clustering, SEC)算法,算法能实现语义属性数据和

到稿日期:2010-03-24 返修日期:2010-07-09 本文受国家自然科学基金青年科学基金项目(60704047),2007年度教育部高等学校创新工程重大培育项目资助。

李志华(1969—),男,博士,副教授,硕士生导师,主要研究方向为优化理论、智能信息处理、模式识别和网络技术,E-mail:ezhli@yahoo.com.cn;**顾 言**(1974—),男,硕士生;**陈孟涛**(1986—),男,硕士生;**王士同**(1964—),男,教授,博士生导师,主要研究方向为人工智能、模式识别、模糊系统、生物信息学;**陈秀宏**(1964—),男,博士后,教授,主要研究方向为模式识别、模糊系统、图像处理。

异构属性数据的聚类。

2 相异性度量

假设有样本集 $x_i \in X$，$x_i=(x_{i1},x_{i2},\cdots,x_{il},x_{i(l+1)},\cdots,x_{i(l+m)})^T$，$1\leqslant i\leqslant n$，且前 l 维为语义属性，后 m 维为连续属性，即 X 为异构数据集。

2.1 语义属性相异性度量的计算

对于语义属性数据的聚类算法的研究，最为关键的就是聚类对象（如样本、聚类中心等）间相异性的定义，又称距离的定义。已有算法中已有不同的定义方法[2-6]。以下给出一种新的相异性度量测度的定义，对语义属性样本进行相异性度量。为了讨论方便，假设 X 中仅包含前 l 维语义属性。则语义属性的相异性度量测度定义如下。

定义 1 语义属性的相异性度量测度：

$$s(x_i,x_j)=\sum_{p=1}^{l}\theta(x_{ip},x_{jp})$$

$$\theta(x_{ip},x_{jp})=\begin{cases}1, & x_{ip}=x_{jp}\\ 0, & x_{ip}\neq x_{jp}\end{cases}$$

$$(i\neq j,1\leqslant i\leqslant n,1\leqslant j\leqslant n,1\leqslant p\leqslant l)$$

$$d(x_{ip},x_{jp})=[\dim s-s(x_i,x_j)]/\dim s \tag{1}$$

式中，$\dim s$ 是语义属性数据的维数。则语义属性样本间的差为 $x_i-x_j=D_S(x_i,x_j)$，其中 $D_S(x_i,x_j)$ 形式化地定义为

$$D_S(x_i,x_j)=(d(x_{i1},x_{j1}),d(x_{i2},x_{j2}),\cdots,d(x_{il},x_{jl}))^T \tag{2}$$

为了进一步计算传统意义上的"距离"，即相异性度量测度，对于语义属性而言，可以采用 Euclidean 距离或 Mahalanobis 距离等具体形式计算。

根据距离度量的性质不难证明，式(1)满足距离的非负性、自反性和对称性，但不能总是满足三角不等式原则，即式(1)实质上是一个度量异构样本集的广义(Generalized)距离测度。

2.2 异构距离的计算

本文选用 Mahalanobis 距离的基本形式，用加权的 Mahalanobis 距离计算异构距离，因为 Mahalanobis 距离考虑了样本的协方差，协方差越大，Mahalanobis 模越小，体现了样本的结构[2]。同时，为了计算方便，给出异构样本之间差的形式化表示：

$$x_i-x_j=(D_S(x_i,x_j),x_{i(l+1)}-x_{j(l+1)},\cdots,x_{i(l+m)}-x_{j(l+m)})^T \tag{3}$$

式中，$D_S(x_i,x_j)$ 见式(2)。

另一方面，从样本的内在结构出发构造一个权重，给 Mahalanobis 距离加权，这个权值体现了样本分布中的不平衡信息。算法区分两种情况来分别计算异构样本之间的距离：当两个样本属于同一类时，其间的 Mahalanobis 距离用式(4)计算：

$$D_m(x_i,x_j,\Sigma)=\sqrt{(x_i-x_j)^T\Sigma^{-1}(x_i-x_j)} \tag{4}$$

当两个样本属于两个不同的类 S_i，S_j 时，其间的距离用加权的 Mahalanobis 距离如式(5)进行计算[1]：

$$D_{wm}(x_i,x_j)=\sqrt{\frac{|s_i|}{|s_i|+|s_j|}D_m^2(x_i,x_j,\Sigma_i)+\frac{|s_j|}{|s_i|+|s_j|}D_m^2(x_i,x_j,\Sigma_j)} \tag{5}$$

式中，Σ 为相应的样本的协方差：$\Sigma=\frac{1}{S}X^TX-\frac{1}{|S|^2}X^T\overline{1}_{|S|}\overline{1}_{|S|}^TX$。若 Σ 为奇异矩阵，则计算它的伪逆：$\Sigma^+=A^TG^{-1}A$，其中 G 是一个仅包含 Σ 正的特征值的对角矩阵，A 是一个包含 Σ 的特征向量的单位正交矩阵；$|S_i|$，$|S_j|$ 是相应类的一阶范数。式(5)更好地体现了样本的结构特征，充分挖掘了样本集中的结构信息。

3 结构熵聚类算法

3.1 熵及 EFC 算法简介

据信息论可知，熵是信息度量的一个有力工具，通常作为某信源所蕴含的信息量的度量，即不确定性的度量。不确定性越小，即概率越大，则熵就越小，相应的信息量也就越小。而对数据样本集而言，熵可以度量一个随机变量的不确定性，度量数据样本集中数据元素依赖关系强弱的程度；从平均意义上而言，熵是随机变量总体信息测度的一个表征[7]。所以文献[7]提出了一种基于样本间相似性度量的熵计算方法，如式(6)所示，并提出了一种基于熵的模糊聚类(Entropy-based fuzzy clustering，EFC)算法[7]。算法逐个计算每一个样本相对于其它样本的熵，并通过标准化处理，使熵值在区间[0.0，1.0]之内。若熵值趋近于区间的下限值 0 时，说明样本离得很近或很远（如离群样本）；若熵值趋近于区间的上限值 1 时，样本以接近该样本集的平均分布距离分布。可见，在不考虑离群样本的情况下，对于熵值趋近于 0 的样本，说明其周围样本分布的密度比较大，可以将其选作为聚类中心，这就是 EFC 算法的精髓。

$$E_i=-\sum_{j\in X}^{j\neq i}(S_{ij}\log_2 S_{ij}+(1-S_{ij})\log_2(1-S_{ij})) \tag{6}$$

式中，S_{ij} 用样本间的欧氏距离计算 $S_{ij}=\exp(-\alpha D_{ij})$，是样本 x_i 和样本 x_j 之间的相似性度量，α 表示指数函数的曲率，估算成 $\alpha=-\ln 0.5/\overline{D}$，$\overline{D}$ 表示数据集的平均距离。EFC 算法的主要思想：用熵值最小的样本充当聚类中心，用相似性来度量聚类。EFC 算法的最大优点是算法中无需迭代，只需遍历样本数据集一遍，算法的效率高。

3.2 异构数据结构熵的计算

异构数据具有明显的结构特征。不仅有样本组成方面的特殊结构，而且有"维"组成方面的特殊性。由于不同的属性具有不同的分布特征，根据信息论可知，它们所包含的信息量不同。在本节的算法中，通过分别计算不同类型属性的熵，再通过加权计算成样本的熵，式(7)从样本"维"的角度同样挖掘了样本中的组成结构信息。

$$E_i=\omega E_{ij'}^{n}+(1-\omega)E_{ij''}^{s},(1\leqslant i\leqslant n,1\leqslant j'\leqslant m,1\leqslant j''\leqslant l) \tag{7}$$

式中，ω 称为属性权重，用来调整不同类型属性对样本 x_i 熵的贡献，$E_{ij'}^{n}$，$E_{ij''}^{s}$ 分别为样本的连续属性和语义属性的熵。此时，为了充分地挖掘样本结构中的聚类线索，式(6)中的相似性系数 S_{ij} 中的欧氏距离用式(4)或式(5)所示的异构距离代替。

显然，式(7)的一个关键就是属性权重 ω 的选择。由于 j' 表示的是样本 x_i 中连续属性的维数，j'' 表示的是 x_i 中语义属性的维数，$j'+j''$ 才是样本 x_i 的维数，因为 ω 的取值在[0，1]之间，理论上讲有无穷多个取值。为了缩短算法的训练时间，提高属性权重系数 ω 的准确度，我们假设 ω 是 j 的函数。这

样 ω 的选择就变成一个多目标优化问题，描述如下：

$$\begin{cases}\min\{E_i=(E_{ij}^n,E_{ij}^c)\}\\0\leqslant\omega\leqslant1\end{cases}\tag{8}$$

式(8)是该优化问题的一个妥协模型。为了求得模型的妥协解，把此多目标优化问题转换成单目标优化问题。根据非线性规划中的 α 法求解[8]方法，给出一种近似 α 法求解的 ω 参数的估算方法，过程描述如下：

假设

$$E_{ij_1}^{n*}=\min_{1\leqslant j\leqslant n}E_{ij}^n,$$

$$E_{ij_2}^{c*}=\min_{1\leqslant j\leqslant n}E_{ij}^c,$$

根据 α 法求解法的性质[8]，有

$$E_{ij_2}^{n*}\geqslant E_{ij_1}^{n*}$$

$$E_{ij_1}^{c*}\geqslant E_{ij_2}^{c*}$$

属性权重 ω 最后估算成

$$\omega=\frac{E_{ij_1}^{c*}-E_{ij_2}^{c*}}{(E_{ij_2}^{n*}-E_{ij_1}^{n*})+(E_{ij_1}^{c*}-E_{ij_2}^{c*})}\tag{9}$$

式中，$E_{ij_1}^{n*}$，$E_{ij_2}^{c*}$ 分别为对应样本的连续属性和语义属性的熵值的最小值，$E_{ij_1}^{c*}$，$E_{ij_2}^{n*}$ 分别为对应样本的语义属性和连续属性的熵值的平均值。这一方法可以为异构样本估算一个参考的属性权重 ω，是一个近似寻优的方法，起到了明显缩短算法训练时间的作用。

3.3 结构熵聚类算法

对于样本的聚类，可以把极小熵的样本作为聚类中心[7]，用本文第 2.1 节定义的异构距离作为度量尺度完成样本聚类。于是，可以定义这样一个过程为聚类算法，样本集中的连续属性用文献[9]的方法进行离散化处理，算法具体描述如下。

算法 1 结构熵聚类(Structure-based Entropy Clustering，SEC)方法

Step 1 初始化 $c=0$，β，mat，c 是聚类个数、β 是度量阈值、mat 是距离权重矩阵；

Step 2 对异构样本，根据式(9)估算参数 ω；

Step 3 根据式(7)计算样本的熵 E；

Step 4 $c=c+1$；

Step 5 在未被聚类的样本中找 $E_{\min}=\min_n\{E_n\}$，且令 $E(x_k)=E_{\min}$，$v_c=x_k$，为第 c 类聚类中心；

Step 6 把满足条件 $D(x_i,v_c)<\beta$ 的所有样本 x_i，假设 v_c 表示存在的聚类中与 x_i 最相近的那个聚类，则把 x_i 聚成第 v_c 类，并从样本集 X 中删除 x_i；

Step 7 根据新的聚类结果刷新并修改距离权重矩阵 mat；

Step 8 如果 X 为空，算法结束，否则转步骤 4。

算法中距离权重矩阵 mat 的初始化方法是：首先产生一个随机值矩阵，并根据训练样本对矩阵元素进行第一轮“权重值”的替代，这样首先得到一个真实值和随机值混合的权重矩阵 mat；距离权重矩阵 mat 的刷新方法是，当算法执行到步骤 7 时，在测试样本中，对于那些训练样本中已出现过的类，矩阵元素保持不变。对于新出现的样本，若无法成功聚类，就把它们当成一个新类，并重新计算相关的距离权重值，并刷新矩阵 mat。

以上算法产生 c 个聚类，聚类中心用 $v_i(1\leqslant i\leqslant c)$ 表示。不难看出，该算法从相当程度上把有关熵聚类的算法在语义属性数据、异构数据上进行了有效的推广。并且，从式(7)可知，当属性权重值 $\omega=1$ 时，算法演变为一个纯连续属性样本的聚类算法，即基于加权 Mahalanobis 距离度量的熵聚类算法；当属性权重值 $\omega=0$ 时，算法演变为一个纯语义属性样本的聚类算法。该特性大大增强了有关熵聚类算法的应用范围。

3.4 SEC 算法的时空复杂度分析及特点

SEC 方法的空间开销主要来自两部分：样本存储的空间复杂度 $O(n)$、在求每个样本熵时的相似性度量存储空间开销 $n(n-1)/2$，所以最坏情况下的空间复杂度为 $O(n^2)$；时间复杂度方面，SEC 方法的本质是一个聚类过程，根据文献[10]可知，从总体上而言，SEC 方法的时间复杂度就是其期望复杂度 $O(nkm)$，其中 n 是样本个数、k 是聚类个数、m 是属性的维数，这样的复杂度通常达不到平方阶 $O(n^2)$。但是，因为算法中的第 4 步涉及到样本间的相似性计算，所以当以样本间“相似性计算”为基本操作时，算法在最坏情况下的时间复杂度还是为 $O(n^2)$。

通过分析研究，SEC 方法具有如下特点：

(1) 算法无需迭代，能一次性确定聚类中心；

(2) 算法充分地挖掘了样本集中的各种聚类线索，如样本的组成结构和维组成信息；

(3) 类中心的选择方法决定算法对初始类中心的选择不敏感和不容易收敛到局部极小值；

(4) 算法拓展了基于熵聚类算法的应用范围。

4 仿真实验及分析

算法的调试环境是 MATLAB 7.0，在具有 P IV 2.4GHz CPU、1GB 内存、Windows XP 操作系统的机器上运行。

本节引进一个聚类准确率 γ 的概念，是指各个类中所有被正确聚类的样本数总和与样本集样本总数的比值，如式(10)所示：

$$\gamma=\frac{\sum_{i=1}^{c}number_i}{n}\tag{10}$$

式中，$number_i$ 为第 i 类中被正确聚类的样本个数，n 为样本的总数。γ 值越大，就说明聚类准确度越高，聚类效果越好；越小，则刚好相反。

为了评价 SEC 算法对语义属性数据、异构属性数据的聚类能力，仿真实验使用了一个纯语义属性数据样本集和两个异构属性数据样本集。样本集的组成如表 1 所列[11]，它们充当仿真实验的测试样本。训练样本从原始样本中随机抽取，如表 1 所列。SEC 算法均采用有监督学习的样本训练方式，并把 SEC 算法的聚类结果与文献[3-6]的结果进行比较。

表 1 实验样本集的组成

样本名称	属性类型	样本数	训练样本数	维数	类
Soybean disease	categorical	47	25	21	4
Flag dataset	heterogeneous	194	100	30	4
Kddcup99 dataset	heterogeneous	181	100	41	5

第 1 个 soybean 样本集共分为 4 类，即 4 种疾病：Phytophthora Rot，Diaporthe Stem Canker，Charcoal Rot，Rhizoctonian Root，样本个数分别为 17，10，10，10，仿真实验仅选用其中的 21 维语义属性；第 2 个 flag 样本集，30 维中有 10 维是

连续属性，其余20维是语义属性，是一个典型的异构属性数据集[6]，可以根据该样本预测某个国家所在的地区、宗教信仰等，本文算法在去掉第1维国家名后，按照zone属性进行聚类，预测样本所属的象限为NE，SE，SW和NW，即4类，每类样本的个数分别为91，29，16和58；第3个是从kddcup99的10%标准子集中随机抽取的样本集，共有5类：normal，smurf，land，pod，ipsweep，相应样本的数量分别为32，100，47，1，1。

Soybean数据集是国际公认[3,5]的语义属性数据样本集，所以它是进行算法比较的主要样本数据集。通过多次实验，当参数值$\beta=9.6$时，SEC算法聚类效果最佳，聚类准确率$\gamma=93.62\%$，聚类结果如表2所列。并把该算法同文献[3]的硬k-modes算法、文献[4]的模糊k-modes算法从聚类准确率、聚类稳定性的角度进行比较，硬k-modes算法和模糊k-modes算法的参数$\alpha=1.1$，初始模式(initial modes)选用文献[3]中提供的，比较结果如表3所列。由于文献[3,4]两种经典的语义属性数据聚类算法，在聚类的过程中，每次迭代均需要重新计算新的聚类中心，造成算法对初始类中心敏感，因此每两次间聚类的效果不一定相同。进一步在表3中给出了上述两种算法、概念k均值(Conceptual K-means，CKM)聚类算法[5]和本文的SEC算法的100次聚类运算的聚类准确度的统计平均值的比较。这一指标体现了算法的聚类稳定性，SEC算法明显高于参与比较的算法，但SEC算法的聚类准确率低于文献[4]的模糊k-modes算法的最佳聚类准确率$\gamma=95.7\%$。

表2 SEC对soybean数据集的聚类结果

No. of cluster $\beta=9.6$				
	1	2	3	4
Actual 1				10
Class 2		10		
Label 3	7	3		
4			17	

表3 不同算法对soybean数据集的聚类准确率比较(100次统计平均)

	Conceptual K-means	Hard k-modes	Fuzzy k-modes	SEC
准确率(%)	70.40	78.20	79	93.62

对flag样本集的聚类，当属性权重参数$\omega\in[0.6,0.7]$时，算法的鲁棒性最好。度量阈值$\beta=45.5$时，样本被正确地聚成4类，聚类准确率$\gamma=80.928\%$。聚类效果如表4所列，虽然同文献[6]的参照属性不同，但单纯从聚类准确率γ而言，明显高于文献[6]的66%。并且从表4可以看出，被错分的样本仅19个。

表4 SEC对flag数据集的聚类结果

No. of cluster $\beta=45.5$				
	1	2	3	4
Actual 1		1		90
Class 2			29	
Label 3				16
4	38	18	2	

kddcup99样本集，其结构复杂，是一个分布不平衡的样本数据集[1]。当把样本聚成6类，属性权重参数$\omega\in[0.6,0.7]$时，聚类结果如表5所列，其中第3，4类可以认为是噪音数据，此时的聚类准确率$\gamma=98.34\%$，并仅有1个样本被错分。

表5 SEC对kddcup99数据集的聚类结果

No. of cluster $\beta=18.9$						
	1	2	3	4	5	6
Actual 1			1	1	30	
Class 2						100
Label 3		47				
4	1					
5	1					

通过实验发现，SEC算法具有比较高的聚类准确率、更高效的算法运行效能、更佳的聚类稳定性和更好的鲁棒性，能实现高维的语义属性数据和异构属性数据的聚类。

结束语 异构属性数据的分布特殊，样本分布不平衡，样本集中往往蕴含有各种有利于聚类的线索，充分挖掘它们，能显著提高聚类性能。本文通过对语义属性、异构属性相异度量的研究和挖掘样本中的结构线索，提出了一种结构熵聚类SEC算法。SEC算法继承了EFC算法的优点，是一种非监督学习的聚类算法，文中总结了此算法的优点。通过仿真实验及同其它算法的比较，说明此算法具有比较好的算法聚类性能。由于聚类结果对距离权重矩阵的依赖性比较大，所以更高效地挖掘样本中的结构信息是下一步的研究重点。

参考文献

[1] Wang De-feng, Yeung D S, Tsang E C C. Weighted Mahalanobis Distance Kernels for Support Vector Machines[J]. IEEE Transaction on Neural Networks, 2007, 18(5): 1453-1462

[2] Kim M, Ramakrishna R S. Projected Clustering for Categorical Datasets[J]. Pattern Recognition Letters, 2006, 27(12): 1405-1417

[3] Huang Zhexue. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

[4] Huang Zhe-xue, Ng M K. A Fuzzy k-modes Algorithm for Clustering Categorical Data[J]. IEEE Transactions on Fuzzy System, 1999, 7(4): 446-452

[5] Ralambondrainy H. A Conceptual version of the k-means Algorithm[J]. Pattern Recognition Letter, 1995, 16(11): 1147-1157

[6] Chen Ning, Chen An, Zhou Long-xiang. Fuzzy k-prototypes algorithm for clustering mixed numeric and categorical valued data [J]. Journal of Software, 2001, 12(8): 1107-1119

[7] Yao J, Dash M, Tan S T, et al. Entropy-based fuzzy clustering and fuzzy modeling[J]. FUZZY Sets and Systems, 2001, 113 (3): 381-388

[8] 陈宝林. 最优化理论与算法[M]. 北京：清华大学出版社，2005

[9] 郭山清，谢立，曾英佩，入侵检测在线规则生成模型[J]. 计算机学报，2006，29(9)：1523-1531

[10] Jiang Sheng-yi, Song Xiao-yu, Wang Hui, et al, A clustering-based method for unsupervised intrusion detections[J]. Pattern Recognition Letters, 2007, 28(13): 989-1003

[11] UCI repository of machine learning database[EB/OL]. http://www.ics.UCI.edu/~mlearn/MLRepository.html, 1998