DOI:10.13860/j.cnki.sltj.2006.06.008
2006年 11月 第25卷 第6期 数理统计与管理 Application of Statistics and Management Nov., 2006 Vol.25 No.6

文章编号:1002-1566(2006)06-0667-08

特征选择方法在信用评估指标选取中的应用

刘 扬 刘伟江

(吉林大学商学院,吉林长春,130012)

摘要:在信用评分模型中所运用的指标变量对模型的表现有重要的影响,指标选取方法的科学化规范化水平有待于进一步提高。本文研究了机器学习领域的特征选择方法在定量确定信用评分模型指标体系上的应用。以实际信用评估问题为例,对四种特征选择方法(ReliefF方法、基于相关性的方法、基于一致性的方法和包裹性)进行了比较试验,验证了特征选择方法可以在精简性、速度和准确率三个方面提高信用评分模型的表现。其中基于一致性的方法和包裹法表现优于ReliefF方法和基于相关性的方法。

关键词:信用评分模型;特征选择;分类算法

中图分类号:F830 文献标识码:A

The Application of Feature Selection Methods in Variables Selections For Credit Scoring Models

LIU Yang, LIU Wei-jiang

(Business School opf Jilin University, Jilin Changchun, 130012, China)

Abstract: The variables included in credit scoring models have important impacts on models' performance, but the variable selevtion methods nee4d to be more standardized and systemized. This paper presents an empirical study on the application of feature selection methods in variables selections compare four feature selection methods—"ReliefF", "Correlation-based", "Consistency-based" and "Wrapper" methods. The study illustrates how these methods help to improve the performance of scoring models in three aspects: model simplicity, model speed and model accuracy. The results show that "Consistency-based" and "Wrapper" methods perform better than "ReliefF" and "Correlation-based" methods.

Key words: credit scoring model; feature selection; classification algorithm

0 引言

信用评分模型是一种建立在历史数据基础上,对信用指标进行量化分析的信用风险评估方法。线性判别分析和逻辑回归等线性统计方法是传统的信用评分建模技术。近几十年来,包括分类树、回归树、神经网络、遗传算法和K-最近邻法在内的诸多人工智能和非线性统计等领域的分类算法开始成功地应用于信用评分模型的建立[1-6]。无论是传统的还是新的建模算法都必须面对一个至关重要的问题——评估指标的选取。在实际应用中,信用评估的数据可能来源不同,而且是为其他用途而收集,因此源数据很可能包含大量指标。信用评分模型除要求达到一定的判别准确率之外,模型的速度、简洁性、可解释性等性能在实际应用时也非常

收稿日期:2005年6月17日
基金项目:教育部留学回国人员科研启动基金项目,项目名称:基于数据挖掘方法的信用风险评估。

重要[7]。无关的和冗余的指标进入模型会导致模型运算的时间增加、模型结果的可解释度下降,也可能引起模型判别准确率的降低。因此,信用评估指标的确定关系到评分模型的有效性和实用性。

过去人们为了从大量的待选指标中选取其中最具判别能力的和减少冗余的指标,采用了各种方法。一些方法根据单个指标的统计属性选取:例如,Fritz 和 Hosemann 采用单变量线性判别分析法估测每个指标与信用类别的相关程度[1],Joos 等根据 Kolmogorov－Smirnov 检验和 x^2 检验的结果排列每个指标的判别能力[2]。但是,单个指标的判别能力高并不能保证它们的组合也一定具有好的判别能力。在另外一些方法中,指标的选取隐含在建模算法内部:例如,在回归分析中利用显著性逐级检验的方法逐步筛选指标;在分类树算法中选择合适的指标作为分类树的判断分支结点。但当待选指标中存在较多的不相关或冗余的指标时,由建模算法本身筛选指标会增加算法运行时间,使生成的模型过于复杂而难以解释,甚至影响模型的准确率。

Hand 和 Henley 在其文章中总结了确定信用评估指标的各种方法,并就此问题指出:在实际应用中,指标体系的确定常常具有随意性,各种方法都可能被先后反复采用,也包括专家依据主观经验对指标的增减[8]。Fritz 和 Hosemann 的研究也得出结论,认为信用评分模型选择指标的过程只是经验的、主观的、无系统的实验过程[1]。这样的方式在待选指标很多的情况下显然不能科学有效地确定信用评估指标体系。

本文运用特征选择方法解决信用评分模型指标的确定问题。特征选择方法(feature selection)是机器学习领域的一种自动的特征选取方法,通过使用某种评价标准和搜索策略将已知数据集合中的特征数目减少,其目的在于优化分类模型。特征选择方法在模式识别和数据挖掘领域应用广泛,却未见其在信用评分模型中的应用。作为一种优化分类模型的方法,特征选择方法能自动从数据中选择判别性好、冗余低的特征集,如果能有效地运用,将为信用评分模型的指标选取提供一种定量的科学的方法。本文首先介绍使用的建模算法和特征选择方法,然后以一个实际信用评估问题为例论述信用指标的选择过程,最后比较分析不同特征选择方法的有效性和不同建模算法的适用性。

1 建模算法和特征选择方法

1.1 本文的研究包含了下面四种曾运用于信用评分模型的分类算法:

回归树 M5:M5 算法采用了一种和分类树类似的非返回跟踪的分割方法,将样本集递归分割成不相交的子集,分割准则旨在极大化分割子集包含样本的相似性。和一般分类树算法的不同之处在于,用该方法建立的树的叶结点不是离散的类别值,而是线性回归模型模拟的连续值[9]。

误差反向传播神经网络(Multi-layer perceptron with back-propagation,简称 MLP):MLP 由输入层、隐含层和输出层的相互连结的神经元组成,是应用于信用评估模型的神经网络中最有效、最活跃的方法。通过样本数据对模型的多次反复训练,沿输出结果与实际结果的误差的负梯度方向修正各层神经元的权值和阈值,如此反复,直至网络全局误差平方和达到预期精度。本文采用了三层结构(输入层、一个隐含层和输出层),以交叉验证(Cross-validation)确定隐含层包含的神经元的最佳个数。激活函数为 sigmoid 函数,学习率初始设为 0.3,训练中若网络不收敛,则自动调低学习率重新训练。

k-最近邻法(k-nearest-neighbors 简称 k-NN): 运用该方法解决分类问题, 首先存储一个由已知类别的训练样本组成的实例集合, 当判别一个新实例的类别时, 与其最相近的 k 个实例中多数属于哪个类别, 新实例就被分到哪一个类别。为此, 要给定一个计算方法用以衡量实例之间的相近程度。本文运用欧几里德距离函数定义两个实例的相近程度, 以交叉验证确定最佳的 k 值。

逻辑回归(Logistic Regression, 简称 LR): 逻辑回归应用于两个类别的分类问题, 是在西方国家的信用行业中普遍采用的一种建立评分模型的方法。该方法先将样本属于两个类别的概率之比取对数(也称为 Logit 变换), 再用样本的指标变量对其进行线性回归, 以此估计某个样本实例属于两个类别中的一个类别的概率。

2.2 本文的研究运用了下面四种有代表性的特征选择方法: ReliefF 方法评价单个特征的判别能力[10,11], 可以对特征集合中的特征进行排序; 基于相关性的方法(correlation-based method)[12]、基于一致性的方法(consitency-based method)[13]和包裹法(wrapper method)[14]分别按照各自的评价标准评价特征子集的判别能力。对于后三种评价特征子集的方法, 采用不同的搜索策略可能得出不同的最优特征子集, 为了便于统一比较, 本文采用 Hall 和 Holmes 提出的爬山法搜索策略[15], 使评价特征子集的方法也可以对特征进行排序。该搜索方法起始于一个空集合, 然后逐个加入特征。首先按照某个评价标准评价每一个特征并选出其中最好的, 然后将它与其它每一个特征组合, 评价并选择其中最好的一对特征, 这个过程一直继续直到所有特征都被组合进来, 如果加入任何一个特征都会减少评价标准的取值, 则选取使其取值减少最小的特征。这样, 特征被选择加入的顺序形成一个特征排序, 这种排序体现了每个特征对特征集合的贡献大小(按照某个评价标准计算得出的贡献)。下面介绍每种特征选择方法评价特征或特征子集的标准。

- ReliefF 方法:

ReliefF 方法(简称 REF)评价单个特征判别能力的依据是: 特征能否辨别相互邻近的样本。有判别能力的特征具有的特点是: 对两个相邻的来自不同类别的样本, 特征取值差别大; 而对每个相邻的来自相同类别的样本, 特征取值相同或相近。用 $M_{REF}(A)$ 表示特征 A 的判别能力, 可用如下公式定量计算:

$$M_{REF}(A) = \frac{\sum_{i=1}^{m} diff(A, R_i, D_i) - diff(A, R_i, S_i)}{m}$$

m 次随机从数据集合中抽取一个样本(m 是一个待设参数), R_i 是第 i 次从数据集合中随机抽取的一个样本, D_i 是 R_i 最近邻的来自不同类别的样本, S_i 的 R_i 最近邻的来自相同类别的样本。$diff(A, *, *)$ 表示两上样本对特征 A 的取值之差(经过正规化处理以保证不同特征之间的可比性)。对于离散的特征, $diff$ 取"1"(两样本对 A 取值相同时)或"0"(两样本对 A 取值不同时)。在此基础上进一步改进, 采用的办法是对 k 个最邻近的样本的贡献值取平均, 以平滑数据中的噪音。本文将 k 设为 30(k 取 20 和 40 所作的实验表明, 对本研究数据, k 的取值对结果的影响很小), m 设为整个训练数据集合的样本数(m 越大, 结果的可靠性越大)。

- 基于相关性的方法

基于相关性的方法(简称 CFS)评价特征子集的标准是: 好的特征子集包含的每个特征与类别高度相关, 同时这些特征相互之间不相关或弱相关。设特征子集 S 包含 k 个特征, 评价 S

好坏的标准如下计算：

$$M_{CFS}(S)=\frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

其中$\overline{r_{ff}}$是 S 中每对特征的相关程度的均值，$\overline{r_{cf}}$是 S 中每个特征与类别的相关程度的均值。相关程度由“对称不确定性”(Symmetrical Uncertainty)度量计算：

$$r_{XY}=2^{*}\frac{H(X)-H(X|Y)}{H(X)+H(Y)}$$

其中 $H(X)$是 X 的熵，$H(Y)$是 Y 的熵，$H(X|Y)$是给定 Y 时的X 的条件熵。对称不确定性的分子称为互信息量，具有对称性：

$$H(X)-H(X|Y)=H(Y)-H(Y|X)$$

设 X 和 Y 的取值范围分别为R_X 和R_Y，则有：

$$H(X)=-\sum_{x\in R_x}p(x)\log(p(x)),$$

$$H(Y)=-\sum_{x\in R_y}p(y)\log(p(y))$$

$$H(X|Y)=-\sum_{x\in R_Y}p(y)\sum_{x\in R_X}p(x|y)\log(p(x|y))$$

对于连续的特征变量需要先用 Fayyad 和 Irani 的方法做离散化处理[16]，再用上面公式计算。

- 基于一致性的方法

基于一致性的方法(简称 CON)认为好的特征子集S 具有的特点是：如果某些样本对 S 的取值相同，则这些样本的类别也应该趋于一致。假设给定数据集合 T 中的样本属于两个类别(C_1 和 C_2)，衡量特征子集 S 的标准采用下式计算的一致性比率：

$$M_{CON}(S)=\frac{\sum_{i=1}^{J}Max(n_i^1,n_i^2)}{N},$$

其中，N 是 T 的样本总数，J 是 S 所有取值的数目(包括 S 中特征的所有取值组合，对于连续的特征变量需要用 Fayyad 和 Irani 的方法做离散化处理[16])。T 中具有 S 第 i 个取值的样本有n_i 个，它们的集合设为 D_i，其中有 n_i^1 个样本属于C_1 类，n_i^2 个样本属于 C_2 类($n_i^1+n_i^2=n_i$)。$Max(n_i^1,n_i^2)$越大(最大为 n_i)，则表示 D_i 中样本的类别越趋于一致，故称 $Max(n_i^2,n_i^2)$为 D_i 的一致数。所有 $D_i(i=1,2,\cdots,J)$的一致性之和占所有样本总数的比率就是特征子集 S 的一致性比率。

在两个特征子集的一致性比率相同时，选择较小的特征子集。事实上，原特征集合的一致性比率最大，如果进行穷尽搜索，可以找到一致性比率与原特征集合相同的最小的特征子集。运用这种方法一方面能保留具有判别能力的特征，另一方面又可以有效地减少冗余的特征。

- 包裹法

包裹法(简称 WRP)采用的衡量特征子集的标准是分类模型的分类准确率。本文研究的实际数据样本容量足够大，没有采用 Kohavi 和 John 针对小样本提出的以交叉验证估计错误率的算法[14]，而是用检验样本直接估计模型的错误率。衡量特征子集 S 的标准为

$$M_{WRP}(S)=1-E_{test} \tag{9}$$

其中 E_{test}为用检验样本估计的分类错误率。运用包裹算法选择出的特征子集通常能提高用目标分类算法建立的分类模型的准确率。然而，这种方法在评价特征子集的过程中要启用目标分类算法建立模型并检验其结果，因此花费时间巨大，尤其是对某些大的数据集或某些耗

时的分类算法, 包裹法可能由于耗时巨大而无法实际运用。正是由于这个原因, 本文的研究无法对神经网络算法 MLP 进行包裹法的实验。

2 指标选择及模型建立过程

本文以德国某信用保险公司的信用风险评估问题为例。该保险公司为进行信用评估从不同渠道搜集客户企业的数据。其中一个数据源是这些客户企业的结算银行提供的银行评估报告。原报告以自由文本格式描述的信息转换为 72 个指标(见表 1 给出的例子), 其中包括银行对企业运行情况、在银行建立的帐户、企业的基本财务状况及其历史信用状况等方面的评价。

该数据集中包含 38283 个样本数据。根据信用分析专家的经验将样本分为两大类, 34562 个"信用好" 样本和 3721 个"信用差" 样本。为了保证模型参数的准确估计, 使用的训练样本中"好" 和"差" 样本数相同: 选取全部的"差" 样本, 再随机选取相同数目的"好" 样本。选出的样本分为三个(见表 2)。在指标选择过程中运用样本 A 进行指标评价和训练模型, 并运用样本 B 检验模型并选择指标: 运用 M_{REF}、M_{CFS}、M_{CON}和 M_{WRP}四个评价标准以及"爬山法" 搜索策略得出四个指标排序, 按照顺序依次使用前 1、2、……、72 个指标建立模型, 每次增加一个指标, 从而进行 72 次建模。然后, 观察 72 次建模的检验错误率的变化, 按顺序选择前面的尽可能少的指标, 并且达到尽可能低的或同水平的错误率。

表 1 银行报告提供的企业信息例子

指标	指标的含义
A1	企业发展总体水平(六个等级)
A2	企业是否有良好的声誉
A3	企业是否有固定资产
A4	企业业务往来帐户运行情况(四个等级)
A5	帐户透支情况(四个等级)
A6	债务偿付情况(四个等级)
……	……

指标选择确定后, 采用大训练样本(样本 A 与样本 B 的并集)重新建立信用评分模型(包括指标选择前运用 72 个指标的模型以及指标选择后运用减少了的指标的模型), 并用另外一个样本数据(样本 C)对模型在指标选取前后的准确率进行检验和比较。用新样本重新建立和检验模型是为了避免因过适应(over-fitting)导致模型准确率的估计及其比较的失真。

表 2 研究中运用的三个样本的容量

样本	"好"	"差"	总数
A: 训练样本	1721	1721	3442
B1: 检验样本 1	1000	1000	2000
B2: 检验样本 2	1000	1000	2000

3 指标选择及模型结果比较

由图 1—图 4 用曲线显示出各种建模算法分别按照四个指标排序进行 72 次建模的检验错误率的变化, 并例示了运用 CON 方法选择出的指标数目。从图 1—图 4 的曲线上观察到, CON 和 WRP 方法的曲线在指标排序的前面部分比 REF 和 CFS 方法的两个曲线的位置低, 相对而言, CON 和 WRP 方法的曲线更早地达到了较低的错误率, 换句话说, 运用更少数目的指标达到了较低的错误率。因此, 运用 CON 和 WRP 方法进行指标排序显然优于另外两种方法。这点可从下面的模型性能的比较得到证实。

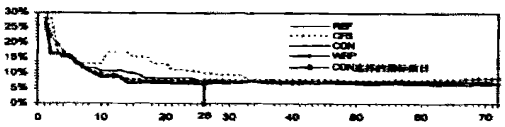
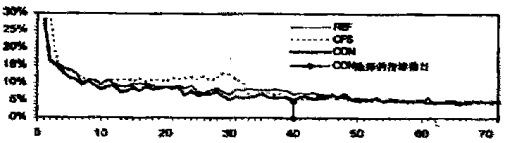


图 1 运用 M5 算法 72 次建模的检验错误率变化曲线



注: 因耗时巨大而无法得到 MLP 的 WRP 试验结果。

图 2 运用 MLP 算法 72 次建模的检验错误率变化曲线

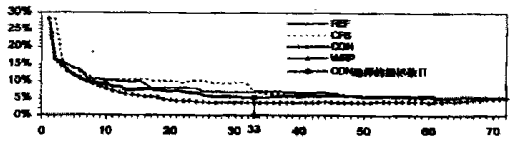
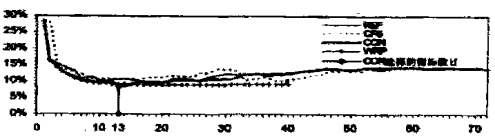


图 3 运用 LR 算法 72 次建模的检验错误率变化曲线



注: k—NN 的 WRP 试验中, 40 个指标以后的点因耗时过长而省略, 但指标的选择结果不受影响。

图 4 运用 k—NN 算法 72 次建模的检验错误率变化曲线

从精简性、速度和准确率三个方面比较模型的性能。信用评分模型在选择指标之前(运用 72 个指标)的性能表现如表 3 所示。运用四种特征选择方法进行指标选择之后模型性能的比较见表 4。

表 3 指标选择前的模型性能

模型算法	模型的错误率	模型速度(秒)
M5	7.15%	655
MLP	4.2%	7244
LR	4.95%	325
k—NN	12.75%	807

注: 模型速度用训练模型所花费的时间表示。由于 k— 最邻近法模型几乎不需要时间进行训练, 它的速度指用样本 C 检验模型的时间。

表 4 指标选择后模型性能比较

	REF	CFS	CON	WRP
M5	1)58% 2)52% 3)—13%	1)53% 2)50% 3)0	1)**36%** 2)45% 3)0	1)67% 2)**35%** 3)0
MLP	1)86% 2)122% 3)—1.5%	1)89% 2)92% 3)0	1)**56%** 2)**25%** 3)0	
LR	1)68% 2)34% 3)0	1)56% 2)30% 3)0	1)**46%** 2)21% 3)0	1)47% 2)**16%** 3)0
k-NN	1)25% 2)28% 3)+1.8%	1)53% 2)59% 3)+3.0%	1)**18%** 2)**22%** 3)+3.3%	1)23% 2)30% 3)**+3.4%**

1) 模型精简性变化(选取的指标数目占原数目的百分比)

2) 模型速度变化(模型花费时间占原时间的百分比)

3) 模型准确率变化(+xx%: 提高百分点— xx%: 降低百分点 0: 无显著变化)

注: 粗体下划线的数字为横向相比的最佳值。对模型在选择指标前后判别准确率的变化进行了显著性检验(Z 值取 2), 表中所给的准确率差异是显著的。

从表4的比较结果可见,特征选择方法能在不同程度上减少模型所用指标的数目,提高模型速度并且同时保持或提高模型原有的准确率(除了REF选择的指标使M5和MLP算法的模型准确率、MLP算法的模型速度有些微下降)。相比而言,运用CON和WRP方法,不管是在减少指标数目、提高模型速度、还是在提高模型准确率方面,都好于REF和CFS方法。

另外,指标选择之后,只有k-最近邻法的模型准确率有明显提高(最多可提高3.4%),其他模型的准确率没有明显改善。这个结果是由分类算法本身的特点决定的:回归树方法M5可以在建立树结构的过程中选择有用的指标作为树结点,神经网络MLP方法在同样本数据进行训练过程中给不相关的指标赋予小的权重,逻辑回归算法根据指标的不同预测能力估计不同的权重系数。这就是说,这三种算法本身具有一定的选择指标的能力。而k-最近邻法只能同样地对待每个指标,不管其判别能力如何,因而进行事先的指标选择可以去掉冗余和不相关的指标,从而明显提高k-最近邻法模型的准确率。

尽管除k-最近邻法的模型,其他模型的准确率没有明显改善,但这些模型的指标数目显著减少(使用CON方法选择指标,M5、MLP、LR和k-NN的模型分别只用了26、40、33和13个指标),大大提高了模型的精简性和速度。这些性能的提高在实际应用时非常重要,尤其在待选指标很多时,以牺牲少量模型准确率为代价获得更加快速更加易于解释的精简的评分模型甚至是相当必要的,因为否则无法得到在实际中可行的评分模型。

4 结论

本文运用实际的信用数据建立了信用评分模型,试验了四种不同的特征选择方法(ReliefF方法、基于相关性的方法、基于一致性的方法和包裹法)在信用评分模型指标选择上的表现。实验研究结果表明,基于一致性的方法和包裹法的表现优于另外两种方法。经过指标选择之后,信用评分模型的性能在精简性、速度和准确率方面得到了改善。其中应用k-最近邻法,模型性能提高的幅度最大。本文的研究结果可见,机器学习中的特征选择方法为信用评估指标的选取提供了有效的途径,当存在大量待选指标时,这种方法作为一种客观定量的指标评价工具,可以优化和提高信用评分模型的有效性和实用性。

[参考文献]

[1] Fritz S and Hosemann D. Restructuring the Credit Process: Behavior Scoring for Deutsche Bank's German Corporates[J]. International Journal of Intelligent Systems in Accounting, Finance & management, 2000, 9: 9-21.

[2] Joos P, Banhoof L, Ooghe H, and Sierens N. Credit classification: A comparison of logit models and decision trees[A]. 10th European Conference on Machine Learning, Workshop notes: Application of machine learning and data mining in finance[C]. TU Chemnitz, Germany: 1998: 59-70.

[3] Desai VS, Conway DG, Crook IN, and Overstreet GA: Credit scoring models in the credit-union environment using neural networks and genetic algorithms[J]. IMA Journal of Mathematics Applied in Business and Industry, 1997, 8: 323-346.

[4] Galindo J and Tamayo P. Credit Risk Assessment Using Statistical and Machine Learning Basic Methodology and Risk Modeling Application[J]. Computational Economics, 2000, 15(1-2): 107-143.

[5] Yobas MB, Crook IN and Ross P. Credit scoring using neural and evilutionary techniques[J]. IMA Journal of

Mathematics Applied in BVusiness and Industry, 2000, 11: 111—125.

[6] Henley WE and Hand DJ. Construction of a k-neares-neighbor credit-scoring system[J]. IMA Journal of Mathematics Applied in Business and Industry, 1997, 8: 305—321.

[7] Liu Yang. A theoretical and empirical study on the sata mining process for credit scoring[M], Goettingen, Germany: Cuvillier Verlag, 2003: 96—97.

[8] Hand DJ and henley WE. Statistical Classification Methods in Consumer Credit Scoring: A Review[J]. Journal of the Royal Statistical Society, 1997, Series A 160(3): 523—541.

[9] Quinlan JR. Learning with continuous classes[A]. Proceedings Australian Joint Conference on Artificial Intelligence[C], Singapore: World Scientific, 1992: 343—348.

[10] Kira K and Rendell LA. A practical approach to feature selection[A]. Proceedings of the International Conference on machine Learning [C], Berlin: Morgan Kaufmann, 1992: 249—256.

[11] Kononenko I. Estimating Att6ributes: Analysis and Extensions of RELIEF[A]. Proceedings of the European Conference on Machine Learning[C], Berlin: Springer, 1994: 171—182.

[12] Hall MA. Correlation-based Feature Selection for Discrete and Numeric Class machine Learning[A]. Proceedings of the Seventeenth International Conference on Machine Learning[C], San Francisco, CA: Morgan Kaufmann, 2000: 359—366.

[13] Dash M, Liu H and Motoda H. Consistency Based Feature Selection[A]. The 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining[C], Berlin: Springer, 2000: 98—109.

[14] Kohavi R and John GH. Wrappers for the feature subset selection[J]. Artificial Intelligence, 1997, 97(1—2): 273—324.

[15] Hall M and Holmes G. Benchmarking attribute selection techniques for discrete class data mining[J]. IEEE Transactions on Knowledge and Data Engineering. 2003, 15(3): 1437—1447.

[16] Fayyad UM. and Irani KB. Multi-Interval discretization of continuousvalued attributes for classification learning[A]. Proceedings of the thirteenth International Joint Conference on Artificial Intelligence[C], Morgan Kaufmann, 1993: 1022—1027.