KDD中规则提取的收敛网络方法及其应用*

熊范纶, 邓 超

(中国科学院 合肥智能机械研究所,安徽 合肥 230027)

(中国科学技术大学 计算机系,安徽 合肥 230027)

E-mail: flxiong@163.net

摘要: 提出一种新的基于神经网络的规则提取方法.提出的网络由一个主网络及其映射网络组成,具有二次收敛过程.通过主网络的学习(第1次收敛)完成知识学习和网络构造,在此基础上构造了其网络映射,通过该映射网络的收敛过程实现规则的提取.该方法在规则提取时无须遍历解空间,从而很好地提高了搜索效率,降低了计算复杂度.同时,还提出估计规则数下限的信度差方法.模拟实验和应用实验也验证了所提出方法的有效性和正确性.

关键词: KDD(knowledge discovery and data mining);规则提取;神经网络;收敛网络;信度差

中图法分类号: TP18 **文献标识码**: A

从数据库中发现知识(knowledge discovery and data mining,简称KDD)是当今AI和计算机科学研究领域中倍受关注的研究热点.其中,从数据库数据中挖掘模式和提取规则是知识发现过程的主要部分之一.目前,在KDD中主要的规则提取方法有决策树推理、神经网络等,其中神经网络方法是最近在KDD中兴起的一个重要方法[1~3],具有很大潜力.

纵观KDD中的规则提取方法,决策树规则提取方法不能实现多变量搜索[4],因为它在建树时每一个节点只含有一个特征,故属于一种单变元算法,特征间的相关性强调不够.虽然它将多个特征用一棵树连在一起,但这种联系是一种松散形式.另外,决策树推理方法对数据中的噪声较为敏感.尽管神经网络从数据中提取规则的最大困难在于其所表示的知识都隐含在网络的联接中,但是神经网络具有一些独特的优点(如非线性映射能力、高度的容错能力以及对噪声的鲁棒性等),因而相对于其他方法,神经网络在数据挖掘中具有更大的前景,值得深入研究.

KDD作为一个独立的研究领域,有区别于其他领域的特点,如所处理的数据具有庞大性、噪声性、不确定性和稀疏性.因此,在对所挖掘规则的评价准则中,除了前述准则以外,算法的计算复杂度也是一个不可忽视的指标.文献[2]是KDD领域中用神经网络技术进行数据挖掘的较为著名的文献.该文针对数据挖掘中的分类问题提出了一种规则提取的神经网络方法,分析和实验均显示了该方法的有效性.但在搜索空间维数较大的情况下或者隐层离散度太大的情况下,该算法的规则搜索效率很低,尤其是在输入空间或输出空间非常大的情况下,要进行遍历是不可能的.

本文依据网络收敛的思想,提出了一种基于神经网络及其映射网络二次收敛的规则提取方法.该方法可以在不进行输入空间和输出空间遍历的情况下进行规则提取.这样就大大缩短了规则提取时间,从而提高了计算效率.

为了用显式方式表示神经网络提取的规则,必须对输入进行编码.好的编码方式有助于提高算法的有效性.本文采用完全编码和温度计式编码两种编码方式.

* 收稿日期: 1999-05-18; 修改日期: 1999-09-15

基金项目: 国家自然科学基金资助项目(69835001)

作者简介: 熊范纶(1940—),男,江苏靖江人,教授,博士生导师,主要研究领域为人工智能,模式识别,机器学习,农业信息处理;邓超(1966—),女,安徽合肥人,博士,主要研究领域为数据挖掘与知识发现,神经网络,人工智能.

1 收敛网络的规则提取方法

1.1 主网络一次收敛

定义 1(训练网络 training network,简称 TrN). 它是一个 3 层前馈神经网络.其中第 1 层为输入层,输入向量$\subset R^{D_{\mathrm{ITr}}}$;第 2 层为隐层,其向量空间$\subset R^{D_{\mathrm{HTr}}}$;最后一层为输出层,属于 $R^{D_{\mathrm{OTr}}}$空间.其中 ITr,HTr 和 OTr 分别表示网络输入层、隐层和输出层的神经元数目.在 TrN 网络中,X^{Tr}为输入向量;O^{Tr}为输出向量;W^{Tr}为输入层到隐层的连接权矩阵;V^{Tr}为隐层到输出层的连接权矩阵;H^{Tr}为隐层输出.因此,有网络输出:

$$O^{\mathrm{Tr}}=\varphi(V^{\mathrm{Tr}}H^{\mathrm{Tr}})=\varphi[V^{\mathrm{Tr}}\sigma(W^{\mathrm{Tr}}X^{\mathrm{Tr}})]. \tag{1}$$

在上式中,激励函数 σ 和 φ 可以根据需要选择 sigmoid 函数、硬限幅函数、饱和线性函数或双曲正切函数或 Gaussian 函数等.对于网络训练可以采用多种方式,如 BP(back propagation)算法、拟牛顿算法等.

当隐层节点到输出节点之间的连接权值幅度很小时,它对网络输出的贡献相对于权值幅度大的连接就很小.为了提高网络的表示效率,我们对它施加一个惩罚衰减因子 $\zeta(0<\zeta<1)$继续参加网络训练.当权值小于预先设定的门限时,我们就可将权值小于一定门限的连接从网络中删除.

(1) 权惩罚阶段.每当新观察样本输入时,

$$\text{if } \alpha_j<\alpha_{pu} \text{ then } \alpha_j=\zeta\alpha_j,\quad j=1,\ldots,D_{\mathrm{HTr}}. \tag{2}$$

这里,α_{pu}表示惩罚衰减门限,α_j 表示网络的连接权.

(2) 连接权裁减阶段.

$$\text{if } \alpha_j<\alpha_{pr} \text{ then } \alpha_j=0,\quad j=1,\ldots,D_{\mathrm{HTr}}. \tag{3}$$

其中 α_{pr}称为隐节点裁减门限.

如果与 j 所对应的隐节点相连的所有连接权值均为 0,则将该隐节点从网络中删除,同时隐层节点数减 1.

定义 2(学习网络 learned network,简称 LN). 它是 TrN 训练后生成的网络.它仍是一个 3 层前馈神经网络,输入向量空间$\subset R^{D_{\mathrm{IL}}}$、隐层向量空间$\subset R^{D_{\mathrm{HL}}}$、输出层属于$\subset R^{D_{\mathrm{OL}}}$空间,则在 LN 网络中,$X^L$ 为输入向量;O^L 为输出向量;W^L 为输入层到隐层的连接权矩阵;V^L 为隐层到输出层的连接权矩阵;H^L 为隐层输出向量.

网络输出向量为

$$O^L=\varphi(V^LH^L)=\varphi[V^L\sigma(W^{XL})]. \tag{4}$$

1.2 网络二次收敛

当网络训练完成后,就开始进行规则提取.在 LN 的基础上,我们建立一个规则提取网络(rule extract network,简称 REN).在 REN 中同样定义,输入向量为

$$X^R=(x_1^R,x_2^R,\ldots,x_{D_{\mathrm{IR}}}^R)^T, \tag{5}$$

输出向量为

$$O^R=(o_1^R,o_2^R,\ldots,o_{D_{\mathrm{OR}}}^R)^T, \tag{6}$$

输入层到隐层的连接权矩阵为

$$W^R=[w_{ij}^R], \tag{7}$$

其中 w_{ij}^R表示矩阵 W^R 中的第 i 行和第 j 列的元素;$i=1,\ldots,D_{\mathrm{HR}}$;$j=1,\ldots,D_{\mathrm{IR}}$.

隐层到输出层的连接权矩阵为

$$V^R=[v_{ij}^R], \tag{8}$$

其中 v_{ij}^R表示矩阵 V^R 中的第 i 行和第 j 列的元素;$i=1,\ldots,D_{\mathrm{OR}}$;$j=1,\ldots,D_{\mathrm{HR}}$.

网络的传输特性:

$$H^R=\sigma(W^RX^R), \tag{9}$$

$$O^R=\varphi(V^RH^R)=\varphi[V^R\sigma(W^RX^R)], \tag{10}$$

其中 $H^R\in R^{D_{\mathrm{HR}}}$.

定义 3. REN 与 LN 的映射关系为 Ψ:LN→REN,即

$$\Psi:\begin{cases}\boldsymbol{X}^R=f(\boldsymbol{W}^L)\\ \boldsymbol{W}^R\leftarrow g(\boldsymbol{X}^L)\\ \boldsymbol{V}^R=\boldsymbol{V}^L\\ D_{\mathrm{HR}}=D_{\mathrm{HL}}\end{cases},\tag{11}$$

其中函数 f 为多维映射,$f:R^{M\times N}\to R^{1\times MN}$. 具体地,$\forall\ \boldsymbol{X}\in R^{M\times N}$,则$\exists\ \boldsymbol{Y}\in R^{1\times MN}$,有 $\boldsymbol{Y}=f(\boldsymbol{X})$.

其中

$$\boldsymbol{Y}=\{y_k=f(x_{ij})\mid y_k=x_{[\frac{k}{N}]\,k\%N}\}.\tag{12}$$

定义 4. 设 $X\in R^N$,I 为单位矩阵,则定义 X 与 I 的替换乘为

$$\boldsymbol{X}\otimes\boldsymbol{I}\begin{bmatrix}\boldsymbol{X}&0&0&0\\0&\boldsymbol{X}&0&0\\0&0&\cdots&0\\0&0&0&\boldsymbol{X}\end{bmatrix}.\tag{13}$$

式(11)中函数 g 为映射,$g:R^{D_{\mathrm{HL}}}\to R^{D_{\mathrm{HL}}\times(D_{\mathrm{IL}}D_{\mathrm{HL}})}$. 由式(11)的映射可知,

$$\begin{aligned}D_{\mathrm{HR}}&=D_{\mathrm{HL}},\\ D_{\mathrm{OR}}&=D_{\mathrm{OL}}.\end{aligned}\tag{14}$$

对于$\forall$ 矩阵 $X\in R^{D_{\mathrm{IL}}}$,则$\exists\ \boldsymbol{Y}\in R^{D_{\mathrm{HL}}\times(D_{\mathrm{IL}}D_{\mathrm{HL}})}$,有 $\boldsymbol{Y}=f(\boldsymbol{X})$. 具体地,$\boldsymbol{Y}=\boldsymbol{X}\otimes\boldsymbol{I}$.

断言 REN 的收敛解对应于 LN 的有效解,经寻优得出的极值点给出系统的一个具体规则. 因为由 $\boldsymbol{O}^R=\varphi[\boldsymbol{V}^R\cdot\sigma(\boldsymbol{W}^R\boldsymbol{X}^R)]$,根据映射关系式(11)可得

$$\boldsymbol{O}^R=\varphi[\boldsymbol{V}^L\sigma(g(\boldsymbol{X}^L)f(\boldsymbol{W}^L))]=\varphi[\boldsymbol{V}^L\sigma((\boldsymbol{X}^L\otimes\boldsymbol{I})f(\boldsymbol{W}^L))].\tag{15}$$

由 f 映射的单调性和替换乘$\otimes$的定义,有以下对应解关系:

$$\boldsymbol{O}^R\to\varphi[\boldsymbol{V}^L\sigma(\boldsymbol{W}^L\boldsymbol{X}^L)]=\boldsymbol{O}^L.\tag{16}$$

下面给出规则提取算法.

Step 1. 随机初始化 $\boldsymbol{W}^R$ 中的非零元素.

Step 2. 取与 LN 网络同样的目标函数和同等量级的误差准则. 用 LMS(least mean square)算法搜索最优 $\boldsymbol{W}^R$,并限制 $\boldsymbol{W}^R$ 元素的取值范围为 $w_{ij}^R\in[0,1]$,但不调整 $\boldsymbol{V}^R$ 和 $\boldsymbol{X}^R$.

Step 3. 当 REN 训练达到收敛时,即得假设空间中的一个具体规则.

Step 4. 若达到目标,则停止规则提取过程. 否则,返回 Step 1.

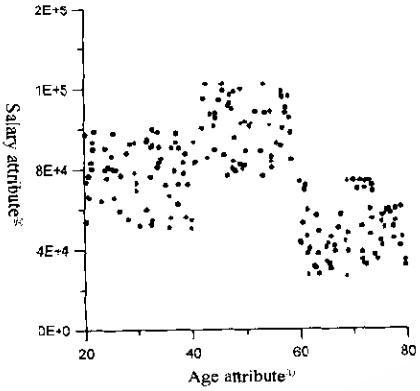
1.3 最少规则数的估计

为了获得上述规则产生算法的停止准则,应获得对规则最少数量的估计,为此本文提出了一种规则下限确定方法——信度差值(DBD)法. 建立在模式发现上的规则是对模式的高级描述,因此,最少模式数的估计也反映了最少规则数的估计. 信度(belief)是 KDD 中衡量模式意义的一个重要指标,而信度的差值则反映了模式未预见性[5]. 信度的计算有 Bayesian 方法、频率方法、Cyc 方法和统计方法[6]等. 其中 Bayesian 估计方法能够适应任意问题类型的信度计算. 这里,命题 α 的信度定义为在先验事件 ζ 下的先验概率 $P(\alpha|\zeta)$. 当新的事件 E 到来时,根据 Bayesian 规则,信度 $P(\alpha|E,\zeta)$这样计算:

$$P(\alpha|E,\zeta)=\frac{P(E|\alpha,\zeta)P(\alpha|\zeta)}{P(E|\alpha,\zeta)P(\alpha|\zeta)+P(E|\neg\alpha,\zeta)P(\neg\alpha|\zeta)}.\tag{17}$$

计算上述后验概率的困难在于有时很难获得条件概率 $P(E|\alpha,\zeta)$. 这里,用 $P(\hat{x}|\boldsymbol{x},\mathrm{cov}(\bar{\boldsymbol{x}}))$来估计 $P(E|\alpha,\zeta)$. 其中 $\boldsymbol{x}=(x_1,\ldots,x_n)$为数据特征参数集,$\bar{\boldsymbol{x}}=(\bar{x}_1,\ldots,\bar{x}_n)$为基于事件 ζ 对 x 均值的估计,而 $\hat{\boldsymbol{x}}=(\hat{x}_1,\ldots,\hat{x}_n)$为基于事件 E 对 x 的估计,$\mathrm{cov}(\boldsymbol{x})$表示 x 的协方差.

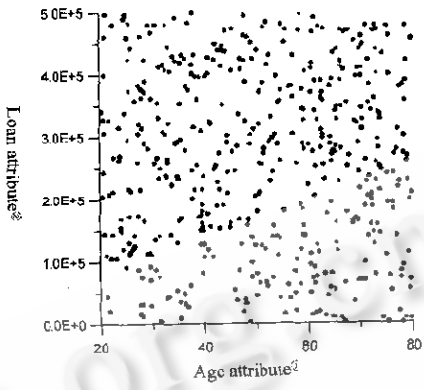
在下面第 2 节的实验 2 中随机产生了 500 个训练样本. 将其中属于 A 组的 283 个样本取出,对每两个属性绘出二维样本分布图于图 1 中. 从图 1 可以直观地看出该属性对之间至少与 3 个模式相关. 而其他属性对的数据分布图均与 age-car 相同(如图 2 所示). 从图 2 可见,该类属性对之间的关系为随机关系.



①年龄属性,②工资属性.

Fig. 1　Distribution of attribute pair age-salary

图 1　关于属性 age-salary 对的分布

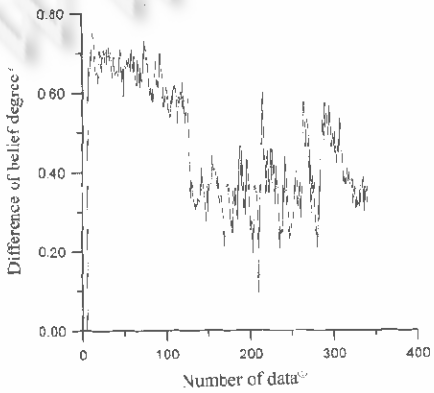


①年龄属性,②贷款属性.

Fig. 2　Distribution of attribute pair age-loan

图 2　关于属性 age-loan 对的分布

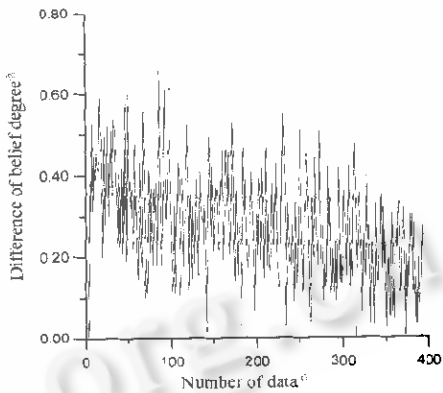
对上述数据计算它们所对应的相邻项信度差 DBD,分别如图 3 和图 4 所示.这两类模式的信度差的区别是十分明显的,因而采用这种信度估计最少模式(规则)数是合理的.



①数据个数,②信度差.

Fig. 3　Belief degree difference of attribute pair age-salary

图 3　关于属性 age-salary 对的信度差



①数据个数,②信度差.

Fig. 4　Belief degree difference of attribute pair age-loan

图 4　关于属性 age-loan 对的信度差

当设置适当的门限时,由图 3 可检测出有 3 个模式.而对于图 4 则几乎是每输入数据该检测系统都出现信度差的大幅度变化,这是随机模式的一种表现形式.对于图 3 的门限设置需要经过适当的试验,以便准确地提取正确的模式数.

2　实验及分析

实验 1. 异或问题

TrN 网络输入为两个节点,隐层为 3 个节点,输出为 1 个节点.经过训练(学习后的总误差 $E+P=0.0020$)得到 LN,其中隐层节点经裁减变为两个.LN 映射得到 REN.对 REN 进行训练,并限制所有 W 元素在[0,1]之间调整.

对于给定导师信号 $O=0$,当 REN 收敛时,所得权值矩阵分别为

$$\begin{bmatrix}1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\end{bmatrix}$$ (总误差 $E+P=0.0032$),

$$\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$ (总误差 $E+P=0.0027$).

由此可得:

规则 1. $(x_1=1\wedge x_2=1)\vee(x_1=0\wedge x_2=0)\Rightarrow o=0$.

对于给定导师信号 $O=1$,当 REN 收敛时所得权值矩阵分别为

$$\begin{bmatrix}0.7176 & 0.0841 & 0 & 0\\ 0 & 0 & 0.7176 & 0.0841\end{bmatrix}$$ (总误差 $E+P=2.7135e-004$),

$$\begin{bmatrix}0.0129 & 0.7791 & 0 & 0\\ 0 & 0 & 0.0129 & 0.0779\end{bmatrix}$$ (总误差 $E+P=2.5929e-004$).

由此可得:

规则 2. $(x_1=1\wedge x_2=0)\vee(x_1=0\wedge x_2=1)\Rightarrow o=1$.

规则 1 和规则 2 正是异或规则.

实验 2. 典型 DM(data mining)函数问题

本实验采用文献[7]中所定义的数据挖掘函数.它包括 9 个属性.见表 1.

Table 1 The attributes of function 2

表 1 函数 2 的属性

Attributes①
Elevel (education level)
Car
Zipcode (zip code of the town)
Salary
Commission
Age
Hyear (value of the house)
Loan (total amount of loan)
Hvalue (value of house)

①属性.

定义以下函数:

如果((age<40)∧(50000≤salary≤100000))∨((40≤age<60)∧(75000≤salary≤125000))∨((age≥60)∧(25000≤salary≤75000)),则属于 A 组;否则属于 B 组.

随机选择 500 个样本,作为训练样本用于 TrN 的训练.经过网络训练,其训练后网络的识别精度为 96%.经过网络映射,在 REN 收敛时,所得规则见表 2.

Table 2 The mined rules of function 2

表 2 函数 2 所挖掘的规则

Rule① 1:	((20≤age<30)∧(65000≤salary≤107000))⇒A group
Rule 2:	((50≤age<60)∧(42000≤salary≤85000))⇒A group
Rule 3:	((70≤age<80)∧(20000≤salary≤100000))⇒A group
Rule 4:	((80<age)∧(20000≤salary≤65000))⇒A group
Rule 5:	((30≤age<60)∧(65000≤salary≤107000))⇒A group
Rule 6:	((20≤age<80)∧(65000≤salary≤107000))⇒A group
Default rule②	⇒B group

①规则,②缺省规则.

从这些提取出的规则中可以看出,规则 1~4 是原来函数表示规则的一个子集,由于数据中噪声和无关属性干扰边界稍有出入.规则 5 和规则 6 是对应该年龄段的 salary 的平均而得出的规则.根据第 1.3 节的最少规则

数的估计方法可知,该实验的规则至少应为 3 个.本实验结果满足该估计方法.

实验 3. 土壤肥力评估规则提取

研究表明,以下因素(U)是决定土壤肥力的主要因素集:

$$U=(U_1,U_2,\ldots,U_{10})$$

=(有机质,全氮,碱解氮,速效磷,速效钾,土壤自然生产力,PH 值,土壤 CEC,容重,物理性粘粒).

本实验中把各肥力因素划分为 4 个等级作为肥力评估集(V):

$$V=(C_1\quad C_2\quad C_3\quad C_4)=(\text{高}\quad \text{中}\quad \text{低}\quad \text{极低}).$$

从该数据库中随机选择 500 个元组作为训练样本.采用完全编码方式.网络一次收敛,其精度为 92%.经映射网络二次收敛共得 36 个规则,表 3 列出了其中的 12 个规则.

Table 3 The mined rules of soil fertility on the network's second times convergence (part)

表 3 网络二次收敛所得的土壤肥力评估规则(部分)

Organic matter①	Total nitrogen②	Soluble nitrogen③	Fast available phosphorus④	Fast available potassium⑤	Soil natural productivity⑥	PH value⑦	Soil CEC⑧	Bulk weight⑨	Physical thin granule⑩	Soil fertility level⑪
low⑫	low	low				medium⑬	medium	medium		very low⑭
	medium	low		high⑮		high		medium		very low
		low				high		low	high	very low
	medium			low		low	low	medium	low	low
high	medium	medium		low		medium	high	medium		low
		low	high	medium				high	medium	low
high	low	medium			medium	low	medium		medium	medium
low	low	low	medium	low	medium			low	medium	medium
high			medium		medium	low	high	high	medium	medium
low	medium	high	medium	low	high	low				high
medium	medium	high	high	medium	high	high	high	low	medium	high
high	medium	high	medium		high	high	medium			high

①有机质,②全氮,③碱解氮,④速效磷,⑤速效钾,⑥土壤自然生产力,⑦PH 值,⑧土壤 CEC,⑨容重,⑩物理性粘粒,⑪土壤肥力等级,⑫低,⑬中,⑭极低,⑮高.

表 3 中所空的栏缺省为极低.经专业人员评定,所得这些规则的有效率为 90%.

3 结 论

本文提出了一种基于神经网络二次收敛的规则提取方法.该方法能够高效地提取规则,尤其对有较高输入空间维数的数据能避免计算复杂度带来的困难.同时提出了估计最少规则数的信度差方法.通过在标准数据和真实数据库的实验都表明该方法的有效性,同时显示该方法在 KDD 中具有很大的前景.进一步的研究,如规则完备性的研究,将在后续工作中进行.

References:

[1] Omlin, C. W., Giles, C. L. Rule revision with recurrent neural networks. IEEE Transactions on Knowledge and Data Engineering, 1996,8(1):183~188.

[2] Lu H. J., Setiono, R., Liu, H. Effective data mining using neural networks. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):957~961.

[3] Fu, L. M. A neural-network model for learning domain rules based on its activation function characteristics. IEEE Transactions on Neural Networks, 1998,9(5):787~795.

[4] Quinlan, J. R. Comparing connectionist and symbolic learning methods. In: Hanson, S. J., Drastall, G. A., Rivest, R. L. eds. Computational Learning Theory and Natural Learning Systems. MIT Press, Vol 1. 1994. 445~456.

[5] Silberschatz, A., Tuzhilin, A. What makes patterns interesting in knowledge discovery system. IEEE Transactions on

Knowledge and Data Engineering, 1996,8(6):970~974.

[6] Kachigan, S. K. Statistical Analysis. Radius Press, 1986.

[7] Agrawal, R., Imielinski, T., Swami, A. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, 1993,5(6):914~925.

Convergent Network Approach for Rule Extraction in KDD and Its Applications

XIONG Fan-lun, DENG Chao

(*Institute of Intelligent Machines, The Chinese Academy of Sciences, Hefei* 230027, *China*);
(*Department of Computer, University of Science and Technology of China, Hefei* 230027, *China*)
E-mail: fanlxiong@163.net

Received May 18, 1999; accepted September 15, 1999

Abstract: A novel neural network based rule extraction method is proposed in this paper. This method consists of a primary network and its corresponding mapping network, which includes twice convergent processes. The knowledge acquisition and network construction of the method are fulfilled by the first convergence of the primary network. Here by a mapping network corresponding to the converged primary network is created whose convergence is capable of realizing the rule extraction. Since there is no need of enumerating the overall space of solutions for this method to extract rules, therefore the searching efficiency is greatly increased and the computation complexity is dramatically reduced. Meanwhile, a stop criterion of rule extraction in terms of difference of belief degree is also proposed in this paper. A lot of simulation experiments and practical applications illustrate and verify the validity and correctness of the proposed method.

Key words: KDD (knowledge discovery and data mining); rule extraction; neural network; convergent network; difference of belief degree