基于 RBF 神经网络的可疑交易监测模型

吕林涛,姬 娜,张九龙
Lv Lin-tao,JI Na,ZHANG Jiu-long

西安理工大学 计算机科学与工程学院,西安 710048
School of Computer Science and Engineering,Xi'an University of Technology,Xi'an 710048,China
E-mail:lvlintao@xaut.edu.cn

Lv Lin-tao,JI Na,ZHANG Jiu-long.Suspicious transaction detection model based on Radial Basis Function Neural Network.Computer Engineering and Applications,2010,46(3):207-210.

Abstract: Aiming at the low detection rate of suspicious transaction at home and abroad in financial field,and with the analysis of Radial Basis Function(RBF) Neural Network,a RBF Neural Network model based on APC-III clustering algorithm and Recursive Least Square(RLS) algorithm for anti-money laundering is proposed.APC-III clustering algorithm is used for determining the parameters of RBF in hidden layer,and RLS algorithm is adopted to update weights of connections between hidden layer and output layer.The proposed method is compared against Support Vector Machine(SVM) and outlier detection methods,which show that the proposed method has the highest detection rate and the lowest false positive rate.Thus the model is proved to have both theoretical and practical value.

Key words: anti-money laundering;Neural Network;Radial Basis Function(RBF);APC-III clustering algorithm;Recursive Least Square(RLS) algorithm

摘 要: 针对国内外金融领域可疑交易的低检测率问题,通过对 RBF(Radial Basis Function)神经网络技术的分析与研究,提出了一种基于 APC-III 聚类算法和 RLS(Recursive Least Square)算法的面向反洗钱的 RBF 神经网络模型并加以实现。APC-III 聚类算法用于确定 RBF 神经网络隐含层的中心向量,RLS 算法用来调整隐含层与输出层之间的连接权值。RBF 神经网络与支持向量机(SVM)和孤立点检测相比,有更高的检测率和较低的误检率,因此,提出的模型具有重要的理论和实用价值。

关键词: 反洗钱;神经网络;径向基函数;APC-III 聚类算法;RLS 算法
DOI: 10.3778/j.issn.1002-8331.2010.03.064 **文章编号:** 1002-8331(2010)03-0207-04 **文献标识码:** A **中图分类号:** TP183

1 引言

洗钱是指隐瞒或掩饰犯罪收益并将该收益伪装起来使之看起来合法的一种活动和过程。当今洗钱犯罪活动日益猖獗,严重威胁全球经济发展和国家安全,是全世界面临的棘手问题之一。由文献[1]可见,选择适当的洗钱交易识别策略与监测方法是当前反洗钱的迫切需要。

目前应用于反洗钱监测领域的技术有[2]:孤立点检测、智能代理,支持向量机、粗糙集和神经网络。相对于洗钱方式的隐蔽性、专业性和创新性,目前被动静止、基于规则监测大额和可疑支付交易的反洗钱监测方法很难适应反洗钱形势的需要,金融领域的海量数据和过高的误检测率迫切需要反洗钱监测技术创新。其中 RBF 神经网络可通过自身无数个神经元持续地对报告数据进行反复计算,从而提高反洗钱的情报收集和分析过程的智能化水平。因此,该文提出基于 RBF 神经网络的可疑交易监测模型,为大额和可疑交易数据分析工作提供了崭新的思路,为相关反洗钱执法部门提供了快捷、准确的预警。

2 RBF 神经网络结构

文献[3-4]中提出的三层前馈 RBF 神经网络能够在很大程度上克服了 BP(Back Propagation)神经网络训练时间较长和隐节点个数难确定的问题。RBF 神经网络描述如下:

$$\{x\in R_p\}\xrightarrow{\varphi(\cdot)}\{h\in R_m\}\xrightarrow{w_j}\{y\in R_n\} \tag{1}$$

$$y_j=f_j(x)=w_0+\sum_{i=1}^{m}w_j\phi(\|x-c_i\|),i=1,2,\cdots,m,j=1,2,\cdots,n \tag{2}$$

式(1)和(2)中,输入层实现 $x\to\phi_i(x)$ 的非线性映射,输出层实现 $\phi_i(x)\to y_j$ 的线性映射,w_j 为隐含层与输出层间的权值,$\|\cdot\|$ 为欧氏范数,$\phi(\cdot)$ 通常取高斯函数。

$$\phi(\|x-c_i\|)=\begin{cases}\exp\left(-\dfrac{\|x-c_i\|^2}{\sigma_i^2}\right),i=1,2,\cdots,m\\ 1,i=0\end{cases} \tag{3}$$

基金项目: 陕西省教育厅自然科学研究项目(the Education Department Natural Science Research Project of Shaanxi Province,China under Grant No.07JK339)。

作者简介: 吕林涛(1954-),男,教授,硕士生导师,主要研究方向:计算机网络,网络信息安全和数据挖掘研究;姬娜(1983-),女,硕士研究生,主要研究方向:计算机网络,网络信息安全和数据挖掘研究;张九龙(1974-),男,博士,副教授,主要研究方向:机器学习,图像模式识别。

收稿日期: 2008-07-29 **修回日期:** 2008-10-23

式(3)中,c_i 为第 i 个隐节点中心,σ_i 是控制高斯函数衰减快慢的宽度参数,m 为隐含层节点数。

可见,RBF 网络涉及 3 个参数:中心向量 c_i、宽度参数 σ_i 及连接权值 w_j。RBF 神经网络结构如图 1 所示。

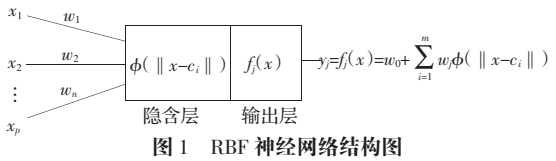

图 1 RBF 神经网络结构图

3 基于 RBF 网络的可疑交易监测模型

通过对金融机构客户信息和交易数据的分析,发现要想判别金融交易行为是否正常,应从两个维度上来判断,一是与自身过往历史行为模式比较;二是与其他账户交易行为做比较。在洗钱交易行为中,资金流动特征可用流量、流向、流动时间来表示,已知 $i,j\in\{$银行账号集合$\}$,$i\neq j$。假定 $Q_j^i(t)$表示在 t 时刻账号 i 与账号 j 之间的资金流动数量,$D_i(t)$表示在 t 时刻账号 i 的存款数量。则 $D_i(t)$可描述为:

$$D_i(t)=D_i(t_0)+\sum_{j=j_0}^{j_0}\int_{t_0}^{t}Q_j^i(t)\mathrm{d}t$$

其中 $i\in\{i_0,i_1,\cdots,i_n\}=I$,$j\in\{j_0,j_1,\cdots,j_n\}=J$,$I$ 为 i 账号在 $t-t_0$ 时间内发生所有交易的账号集合,J 为 j 账号在 $t-t_0$ 时间内发生所有交易的账号集合。当 $Q_j^i(t)>0$ 时,表示 t 时刻资金从账号 i 流到账号 j。当 $Q_j^i(t)<0$ 时,表示 t 时刻资金从账号 j 流到账号 i。在 t 时刻,$Q_j^i(t)=0$,这表明账号 i 与账号 j 之间未发生任何交易行为。

训练集由一组数据库记录构成,每一条记录是一个由有关属性组成的特征向量,另外训练集的每条记录中有一个特定的类标签与之对应。一个训练样本数据集可表示

$$E=(v_1,v_2,\cdots,v_p;s) \tag{4}$$

式(4)中,$v_1,v_2,\cdots,v_p$ 表示要输入的反洗钱属性参数,这些属性参数必须能够反映出企业账户的基本特征以及洗钱活动的不同阶段账户资金流动的特征。s 表示分类标签属性值。可用的属性包括客户号、客户姓名、资金账号、客户证件号码、交易时间、交易账号所属行业、交易地区代号、交易金额、交易次数、交易币种、交易类型、交易频度等。

综上所述,提出基于 RBF 神经网络的可疑交易监测模型如图 2 所示。

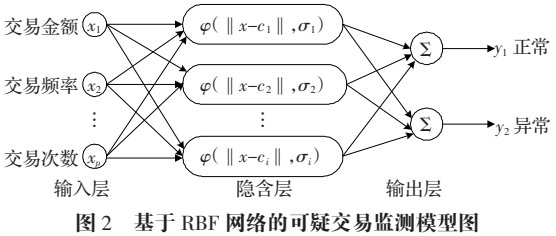

图 2 基于 RBF 网络的可疑交易监测模型图

4 基于 RBF 神经网络的可疑交易监测模型实现

反洗钱中需要进行可疑交易识别的数据是由金融机构上报的大额和可疑交易数据。大额交易是指规定金额以上的货币支付交易,可疑交易是指金融交易的金额、频率、来源、流向和用途等有异常特征的交易行为。可以把提供的可疑交易数据和洗钱案例所归纳出来的案例数据作为训练样本,从而进行参数学习。对经过预处理的金融交易数据,通过调用训练好的 RBF 神经网络来判断是否可疑,当可疑交易数据经过证实的确为洗钱交易数据后,把它添加到案例数据中再次进行参数学习。模型实现流程如图 3 所示。

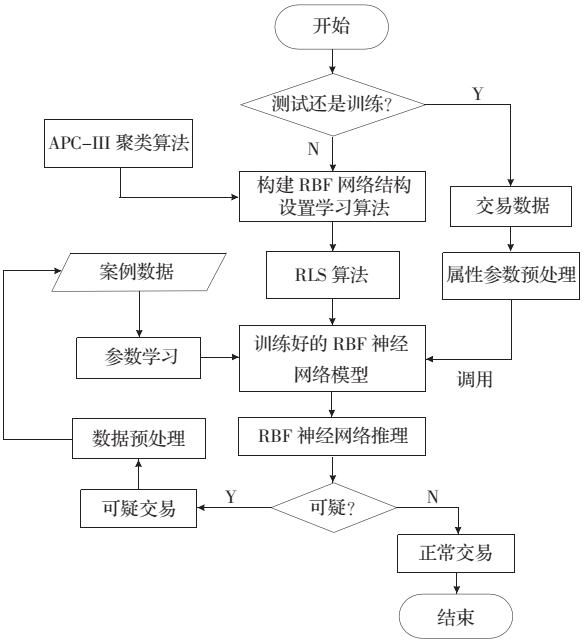

图 3 模型实现流程图

4.1 属性参数预处理

银行交易记录是典型的高维和异构数据。已知数据集 R,设 R 中每个数据 e 有 w 个属性,则 e 的第 i 个属性值表示为 e_i $(i=1,2,\cdots,w)$,其中 $e_i(i=1,2,\cdots,r)$取值为连续值,$e_i(i=r+1,\cdots,w)$取值为离散值,这样的数据叫异构数据。反洗钱交易数据中并非每个属性都与洗钱活动有关。反洗钱部门在监测、分析、甄别洗钱交易时,要结合反洗钱领域知识,对一些无关属性进行过滤,将银行交易源数据转换成反应洗钱特征的属性集合。也有一些无关的属性需要相关性分析技术来判断是否将该属性过滤掉,设数据集 $D=\{d_1,d_2,\cdots,d_n\}$,数据的属性集 A_S。现在检测 $x\in A_S$和目标属性 $y\in A_S$是否相关。对于所有 x 的值排序后可以得到相应 y 属性值的分布。设 x_i 表示第 i 条记录的 x 属性值,$0<i<n$,定义距离函数:

$$F=\sum_{2}^{n}\sqrt{(x_i-x_{i-1})^2+(y_i-y_{i-1})^2}$$

可以描述该分布的离散型,若超过设定的阈值,则认为这两个属性没有相关性,可以过滤属性 x。

由于 RBF 神经网络只能处理数值型数据,因此首先需要对反洗钱属性参数进行预处理。对定性属性参数,在输入时按程度统一分类,然后将定性属性参数予以量化。假定数据集 U,Q 为 U 的一个子集,则对于 Q 中的每一个数据,存在一个数学函数 C,使产生的数据值在[0,1]之间,即:$\forall Q\subseteq U$,$C:Q\rightarrow[0,1]$。该文采用了极差标准化方法:

$$\xi_i=\max_i(x_{ij})-\min_i(x_{ij}) \tag{5}$$

式(5)中，ξ_i代表极差(训练样本中每个特征向量的最大值与最小值之差)，x_{ij}表示第 i 个样本数据第 j 个属性的值。

通过把 x_{ij}转换成 $x'_{ij}=(x'_{ij}-\min(x_{ij}))/\xi_i$来达到数据标准化，经过这种数据变换后，新得到的样本数据值均属于区间[0,1]。这样就使得变量间具有一定的可比性和可计算性。同时，在使用这些标准化后的数据值时，相对于原来的数据减少了计算量。

4.2　RBF 神经网络的学习过程

4.2.1　APC-III 聚类算法

通常用 K-均值聚类算法确定 RBF 神经网络的隐含层，但由于它是随机选取初始聚类中心，因此，学习速率慢，难以获得满意的聚类效果；文献[5]提出的 APC-III 聚类算法对样本集只学习一遍，有较高的学习速率，其聚类半径(样本集中各数据间的最小距离)为：

$$R_0=\alpha\frac{1}{p}\sum_{i=1}^{p}\min_{i\neq j}(\|X_i-X_j\|) \tag{6}$$

式(6)中，p 是样本集中的样本个数，α 为常数。R_0的大小决定了最后的聚类数，即数据中心个数(隐含层神经元个数)。APC-III 聚类算法具体描述如下：

输入：训练样本 $X=\{x_i|x_i\in R^p,i=1,2,\cdots,p\}$

输出：聚类中心 $c_i(i=1,2)$

变量：$\bar{L}$：聚类数；c_i：第 i 个聚类中心；n_i：第 i 个聚类中心拥有的样本数目；d_{ji}：样本数据 x_j到第 i 个聚类中心的距离。

(1)初始化：$\bar{L}=1,c_1=x_1,n_1=1$；/* 对训练集中的每个样本 */

(2)For（$j=2;j\leqslant p;j++$）/* 对每个聚类 */

(3)For（$i=1;i\leqslant L;i++$）

{计算 d_{ji};}

(4)if（$d_{ji}\leqslant R_0$）/* 把数据 x_j加到第 i 个聚类中心范围 */

{$c_i=(c_in_i+x_j)/(n_j+1);n_j=n_j+1;$}

(5)if（$x_j\notin c_i$）/* 建立新聚类 */

{$\bar{L}=\bar{L}+1;c_{\bar{L}}=x_j;n_{\bar{L}}=1;$}

4.2.2　宽度参数的确定

当 RBF 网络的中心向量 $\boldsymbol{c}_i$确定后，高斯函数的宽度参数为：

$$\sigma_i=\frac{d_i}{\sqrt{2m}} \tag{7}$$

式(7)中，d_i是第 i 个中心向量与其他中心向量间的最大距离，m 为选取的中心向量个数。

4.2.3　递推最小二乘(RLS)算法

求取隐含层与输出层间的连接权值是一个线性优化问题，由于传统的梯度下降(Gradient Descent)算法收敛速度较慢，因此，该文采用 RLS(Recursive Least Square)算法。

设在第 k 步时，输出向量为：

$$\boldsymbol{\phi}(k)=[\phi_1(k),\phi_2(k),\cdots,\phi_{nl}(k)]^{\mathrm{T}}=$$

$$[\phi_1(l_1(k)),\sigma,\phi_2(l_2(k)),\sigma,\cdots,\phi_{nl}(l_{nl}(k)),\sigma]^{\mathrm{T}} \tag{8}$$

第 k 步中第 j 个节点的估计输出为：

$$\hat{y}_j(k)=\sum w_{ij}\phi_l[l_i(k),\sigma] \tag{9}$$

若实际输出为 $y_j(k)$，则有误差

$$\varepsilon_j(k)=y_j(k)-\hat{y}_j(k) \tag{10}$$

权值更新为：

$$w_j(k+1)=w_j(k)+\boldsymbol{\mu}(k)\boldsymbol{\phi}(k)\varepsilon_j(k)\frac{1}{\lambda(k)}\cdot$$

$$\left[\boldsymbol{\mu}(k)\frac{\boldsymbol{\mu}(k-1)\boldsymbol{\phi}(k)\boldsymbol{\phi}^{\mathrm{T}}(k)\boldsymbol{\mu}(k-1)}{\lambda(k)+\boldsymbol{\phi}^{\mathrm{T}}(k)\boldsymbol{\mu}(k-1)\boldsymbol{\phi}(k)}\right] \tag{11}$$

式(11)中，$\boldsymbol{\mu}(k)$为误差方差阵，$\lambda(k)(0<\lambda(k)<1)$为遗忘因子，它逐步减弱历史样本对当前计算值的影响。该文采用动态调整遗忘因子 $\lambda(k)$的方法

$$\lambda(k)=1-\exp\left(-\frac{k}{\tau_0}\right) \tag{12}$$

式(12)中，τ_0是根据经验进行设定的初始平滑因子。

5　实验

文章将 RBF 神经网络、支持向量机和孤立点检测 3 种技术在反洗钱监测中的应用作了分析与比对。这里仅给出比较检测率(检测出的不正常交易占整个测试集中不正常交易总数的比例)和误检测率(被误认为是不正常交易的正常交易的总数占整个正常交易总数的比例)。

5.1　实验用例

实验采用某商业银行的一个真实金融交易记录数据库集，它包括 6 000 个账号，8 个月中 100 万笔交易记录，表 1 为原始数据集。根据案例数据的特性，通过利用统计理论对反洗钱属性参数处理后，提取出 3 个属性：账户转出频率 d_1(账户出账和入账交易总频数中转出频数的比重)、账户转入频率 d_2(账户出账和入账交易总频数中转入频数的比重)和交易金额 d_3。然后利用 4.1 小节中对反洗钱属性参数进行预处理的方法，对这 3 个提取出的属性进行归一化处理后，得到的训练数据集如表 2 所示。

表 1　原始数据集

客户号	行业类型	交易类型	存款次数	取款次数	交易金额/元
18735	制造业	存款	7	6	109 060
15789	服务业	取款	5	2	8 610
19967	私人	取款	4	3	76 431
…	…	…	…	…	…

表 2　训练数据集

样本	账户转出频率 d_1	账户转入频率 d_2	交易金额 d_3
s_1	0.04	0.21	0.65
s_2	0.18	0.15	0.71
…	…	…	…

由于真实的银行洗钱数据很难获取，将异常交易数据加入到正常交易数据中，把得到的合成数据作为可疑交易数据来检验 RBF 神经网络算法的有效性。假定正常交易数据为 $\psi_1(\kappa)$，则异常交易数据可表示为：

$$\psi_2(k)=\psi_1(k)+\gamma\cdot e(k),k=1,2,\cdots,200 \tag{13}$$

式(13)中,$e(k)=\begin{cases}\sin(\frac{\pi}{2}k) & k\in[80,90]\\ 0 & \text{其他}\end{cases}$为一异常事件,$\gamma$ 为度量交易异常强度的常数。

5.2 实验分析比对

SVM 算法适合高维异构数据集中的分类设计和可疑发现,采用 LIBSVM 软件包提供的 SVM 算法来进行实验。该 SVM 算法使用 HVDM 距离,给出错误分类惩罚参数 G=100,控制因子 η=1;同时在孤立点检测算法中,给定数据集 $L(l\in L)$和 $R(r\in R)$,则有平均值

$$\hat{n}(l,r,\alpha_1)=\frac{\sum_{q\in N_{L(l,r)}} n_R(q,\alpha_1 r)}{n_L(l,r)}$$

其中 $N_{L(l,r)}=\{q\in L|d(l,q)\leqslant r\}$,$d(l,q)$为点 l 和 q 之间的距离,$n_L(l,r)$为 $l\in L$ 的 r 近邻数目;标准偏差

$$\sigma(l,r,\alpha_1)=\sqrt{\frac{\sum_{q\in N_{L(l,r)}}(n_R(q,\alpha_1 r)-\hat{n}(l,r,\alpha_1))^2}{n_L(l,r)}}$$

α_1 的值为 $1/2^{\delta}$(δ 是一个正数)。如果$|\hat{n}(l,r,\alpha_1)-n_R(l,\alpha_1 r)|>\varepsilon\sigma(l,r,\alpha_1)$,则点 l 为孤立点,即为监测出的可疑交易记录,其中的 ε 为常数。

根据文献[5]中的实验结果,发现当决定 APC-III 聚类算法的聚类半径 α 小于 1.04 时,RBF 神经网络的精度相对较高,所以该文取 α 为 1.02, 然后取初始平滑因子 τ_0 为 0.1,RBF 神经网络的初始隐含层节点数在 $\mathrm{lb}p$ 左右摆动,p 为输入层节点数。

接下来从银行原始交易数据记录集中选取 200 个数据,然后混入 70 个异常交易数据来训练 RBF 神经网络模型。通过查看训练效果,反复调制网络参数和学习参数的设置,以获得较好的基于 RBF 神经网络的可疑交易监测模型。当误差小于 0.01 时,学习结束,图 4 为误差变化曲线。

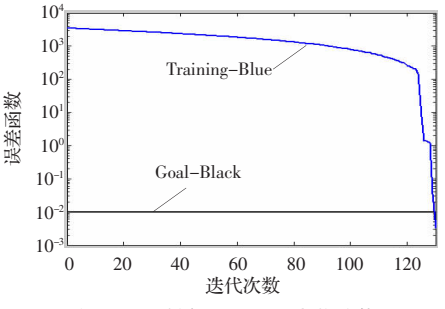

图 4　RBF 神经网络误差变化曲线

对于测试集,从原始数据集中另外选取 50 个数据,然后再混入另外 40 个异常交易数据。设 α_1=1/8,ε=3,γ=0.2,通过对测试集中 90 个数据进行实验分析后,发现支持向量机、孤立点检测和 RBF 神经网络 3 种方法检测出的不正常交易个数分别为 27、38 和 78,测试结果如图 5 所示。

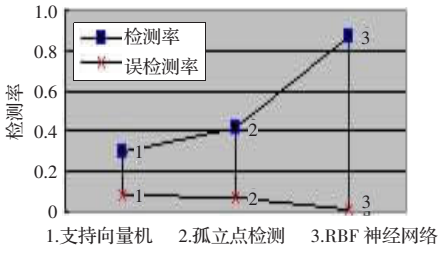

图 5　3 种方法的比较

6 结束语

文章提出基于 RBF 神经网络的可疑交易监测模型, 可有效提高洗钱交易行为的识别效率和准确性,并能适应不断变化的洗钱风险和交易模式。实验证明了该模型在反洗钱监测中可取得满意的效果。在进一步的研究中,将设计一个反洗钱监测系统以及利用真实的洗钱交易数据来进行实验分析,同时对多种判别可疑交易的检测方法进行组合应用,以期达到更好的反洗钱监测效果。

参考文献:

[1] Gao Shijia,Xu Dongming.Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering[J].Expert System with Applications,2007(11).

[2] Liu Xuan,Zhang Pengzhu.An agent based anti-money laundering system architecture for financial supervision[C]//International Conference on Wireless Communications,Networking and Mobile Computing,Sept 2007,2007:5472-5475.

[3] Ahmat Nor N L,Harun S,Mohd Kassim A H.Radial basis function modeling of hourly streamflow hydrograph[J].Journal of Hydrologic Engineering,2007,12(1):113-123.

[4] Yang Zhengrong.A novel radial basis function neural network for discriminant analysis[J].IEEE Transactions on Neural Networks,2006,17(3):604-612.

[5] Idri A,Abran A,Mbarki S.An experiment on the design of radial basis function neural networks for software cost estimation[J].Information and Communication Technologies:From Theory to Applications,2006,1:1612-1617.

(上接 191 页)

[2] Dorigo M,Cambardella L M.A cooperative learning approach to the traveling salesman problem[J].IEEE Transactions on Evolutionary Computation,1997,1(1):53-66.

[3] Colorni A,Dorigo M,Maniezzo V,et al.Ant system for job shop scheduling[J].Belgian Journal of Operations Research Statistics and Computer Science,1994,34(1):39-53.

[4] 段海滨.蚁群算法原理及其应用[M].北京:科学出版社,2005.

[5] 吴斌,史忠.一种基于蚁群算法的 TSP 问题分段求解算法[J].计算机学报,2001(12):1328-1333.

[6] 黄贵玲,高西全,靳松杰,等.基于蚁群算法的最短路径问题的研究和应用[J].计算机工程与应用,2007,43(13):233-235.

[7] 韦绥线,黄胜华.一种仿 Dijkstra 的蚂蚁算法[J].计算机应用,2005,25(12):2908-2910.

[8] 谢民,高利新.蚁群算法在最优路径规划中的应用[J].计算机工程与应用,2008,44(8):245-248.

[9] 朱庆保,杨志军.基于变异和动态信息素更新的蚁群优化算法[J].软件学报,2004,15(2):185-192.