混合马尔科夫预测模型及其在反洗钱中的应用研究

李玉华 李栋才 毕 威 李瑞轩
(华中科技大学计算机学院 武汉 430074)

摘 要 反洗钱中的一个重要问题是预测可疑账户未来可能发生的交易。马尔科夫模型在股票、商品价格、市场占有率等经济领域的预测中具有广泛的应用,但单一的马尔科夫模型的预测准确性有待提高。提出一种结合数据挖掘中聚类、关联规则和低序马尔科夫模型的混合马尔科夫模型,并在模型的建立过程中基于置信度进行剪枝以降低时间复杂度,最后将该模型用于预测反洗钱领域中账户之间的交易。实验表明,该模型具有较高的预测准确性,并在预测准确性和时间复杂度两者之间取得了较好的平衡。

关键词 混合马尔科夫模型,预测,聚类,关联规则,反洗钱
中图法分类号 TP311 **文献标识码** A

Hybrid Markov Prediction Model and Research of its Applications in Anti-money Laundering

LI Yu-hua LI Dong-cai BI Wei LI Rui-xuan
(College of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract An important problem in anti-money laundering is to predict the possible transactions conducted by suspicious accounts. Markov model has a wide range of applications in economic predictions such as stock, commodity prices, market share and so on. But the prediction accuracy of the single markov model remains to be improved. A hybrid Markov model jointing with clustering, association rule and low order Markov model was proposed. In the process of constructing the model, the confidence-based pruning was conducted to reduce the time complexity. Finally, the model was used to predict the transactions among accounts in anti-money laundering. The experimental results show that this model has high prediction accuracy and is a good tradeoff between the prediction accuracy and the time complexity.

Keywords Hybrid Markov model, Prediction, Clustering, Association rule, Anti-money laundering

1 引言

金融犯罪是当今国际社会面临的重大问题,尤其是其中的洗钱活动日益猖獗,对世界各国政治和经济秩序的危害越发严重而深远,对国际金融体系安全造成的威胁不容忽视[1]。洗钱等金融犯罪是一种群体犯罪,其交易是随时间变化的序列。异常交易行为预测是反洗钱的关键,它可帮助发现洗钱线索,帮助执法部门有效地预防和打击洗钱等各种金融犯罪。

马尔科夫模型在时间序列预测中应用非常广泛。2002年,Zhu Jianhan[2],Shantha[3]将马尔科夫链用于Web网页的导航预测中。2004年,Mukund Deshpaned和George Karypis[4]在混合序马尔科夫模型和K序马尔科夫模型的基础上提出了选择性马尔科夫模型,它具有较高的预测准确性和覆盖率,但是当K较大时,算法的时间复杂度很高。2006年,Faten Khalil和Li Jiuyong[5]将低序马尔科夫模型和关联规则结合起来,提出了一种新的混合马尔科夫模型,此算法相对选择性马尔科夫模型而言,既降低了时间复杂度,又将预测的准确性保持在一个较高的水平。但此模型对所有的链接序列采用统一的统计方式,忽视了不同类型时间序列之间的差异性。郭景峰[6]提出了一种改进的针对合著关系网络的链接预测方法。

近年来,在经济领域的应用中,高金余[7]利用马尔科夫切换模型来分析中国股市,该模型对股市波动的研究有一定指导意义。张冬青[8]在考虑影响因素的基础上,提出了基于观测向量序列的隐马尔科夫模型预测方法。该方法同时考虑变量自身序列结构以及相关因素的影响。张冬青[9]提出了一种基于小波域隐马尔可夫模型的时间序列分析方法,它可应用在经济领域时间序列分析中。现有单纯基于马尔科夫模型预测的准确性有待提高,而且马尔科夫模型在反洗钱预测中的相关研究还比较少。

本文根据反洗钱应用的特点,将数据挖掘中的聚类、关联规则的相关理论和低序马尔科夫模型结合起来,给出一种基于混合马尔科夫模型的预测算法并应用于反洗钱实践。采用k均值聚类算法对所有账户的交易序列进行聚类,将具有相似交易特点的交易序列划为一类,更有针对性。然后对每个聚类分别建立基于置信度剪枝的低序马尔科夫模型,用以预测用户的交易路径。当预测出现模糊时,引入关联规则给出更准确的预测结果。

到稿日期:2010-08-04 返修日期:2010-11-12 本文受国家自然科学基金项目(70771043),国家自然科学基金项目(60873225),国家863计划项目(2007AA01Z403)资助。

李玉华(1968-),女,博士,副教授,CCF会员,主要研究方向为数据挖掘、金融信息化,E-mail:yuhua_yy@163.com。

2 K序马尔科夫模型

定义1 设有随机过程$\{X_n, n\in T\}$，若对于任意的整数$n\in T$和任意的$i_0, i_1, \cdots, i_{n+1}\in I$，条件概率满足

$$P\{X_{n+1}=i_{n+1}\mid X_0=i_0, X_1=i_1, \cdots, X_n=i_n\}$$
$$=P\{X_{n+1}=i_{n+1}\mid X_n=i_n\} \quad (1)$$

则称$\{X_n, n\in T\}$为马尔科夫链。但是在实际应用中，并不是所用情况都满足式(1)，即

$$P\{X_{n+1}=i_{n+1}\mid X_n=i_n, \cdots, X_0=i_0\}\neq P\{X_{n+1}=i_{n+1}\mid X_n=i_n\}$$

在这种情况下，必须将之前所有的状态都考虑进来，毫无疑问这增加了问题的复杂程度。为了简化问题，可以只考虑当前时间点之前的K个状态($K<n$)，并假设

$$P\{X_{n+m}=i_{n+m}\mid X_n=i_n, \cdots, X_0=i_0\}=P\{X_{n+m}=i_{n+m}\mid X_n=i_n, \cdots, X_{n-(K-1)}=i_{n-(K-1)}\} \quad (2)$$

满足式(2)的马尔科夫模型为K序马尔科夫模型，其中K称为马尔科夫模型的序。本文中提到的低序马尔科夫模型是指K取值比较小的情况，比如$K=2,3$等。至于具体K取多大才是低序，则取决于具体问题的需要。

$S_t^K=\langle i_{t-(K-1)}, i_{t-(K-2)}, \cdots, i_t\rangle$称为$K$序马尔科夫模型的状态[4]。$K$序马尔科夫模型有如下评估参数[4]：

(1)准确率(accuracy)：评价模型预测能力的参数。

(2)状态数量(number of states)：评价模型时间和空间复杂度的参数。

(3)覆盖率(coverage)：训练集的数量是一定的，当K越大时，模型的状态数量越大，有限的训练集不能保证涵盖所有的状态，其比值就是模型的覆盖率。

已有的研究表明，增大K值能提高预测的准确性。但是随着K的增大，模型的状态数量有了更大幅度的提高。

3 基于混合马尔科夫模型的金融交易预测

在一般的马尔科夫模型应用中，往往对整个数据集建立统一的马尔科夫模型，忽视了个体的差异。比如在Web浏览预测中，不同的用户有着不同的浏览偏好，其浏览路径是有区别的。在反洗钱中，考虑到不同的交易账户有不同的交易特点，将所有账户对应的交易序列划为彼此不同的若干类别，可提高解决问题的针对性。基于这种考虑，可以先对所有账户交易序列进行聚类，将具有相似特点的交易序列划为一类，然后对每个聚类分别建立独立的K序马尔科夫模型。这样将更具有针对性，从而提高预测的准确性，同时减少马尔科夫模型状态数量，进而降低时间复杂度。

另一方面，K序马尔科夫模型中K值的选取是一个重要的问题。如前所述，K越大，预测准确性越高，但同时将导致时间复杂度的提高以及覆盖率的降低。为了解决这个问题，可以选择K较小的低序马尔科夫模型。但较低的序很难保证预测的准确性，这主要是因为低序马尔科夫模型的规则容易出现模糊的情况。即最大的条件概率和第二大的条件概率差值很小，此时引入关联规则对预测结果进行修正，可得到更准确的预测结果。这样在提高预测准确性的同时，保持了较低时间复杂度，实现了预测准确性和时间复杂度之间的均衡。

下面介绍混合马尔科夫模型在可疑交易预测中的应用。

3.1 建立交易序列

先从数据库中读取所有的交易账户，然后按照时间顺序建立所有账户的交易序列。这个过程并不复杂，关键问题是将交易方向考虑进来，而不仅仅是交易对象的账号。可借助一种扩展编码的思想。假设在账号为60384的账户的交易序列中，该账户按照时间顺序有如表1所列的交易记录，其中主体账户表示交易资金的流出方，客体账户表示交易资金的流入方。根据这些记录可以得到如下交易序列：

$\langle 692750, 687401, 687400, 690540, \cdots\rangle$

即通过在交易对象的账号后面添加一位方向编码0或者1来区分资金的流向。这样，即使交易对象是同一个交易账户，只要资金流向不同，仍被视为不同的交易对象。

表1 账户60384的交易记录

序号	主体账户	客体账户	交易时间
1	60384	69275	2003-01-12
2	68740	60384	2003-01-31
3	60384	68740	2003-02-08
4	60384	69054	2003-02-25
5	……	……	……

3.2 交易序列的聚类

本节讨论如何对交易序列进行聚类。假设$C=\{c_1, c_2, \cdots, c_n\}$是所有账户的集合，$n$表示账户的总数量，$W_i=\langle c_1^i, c_2^i, \cdots, c_{l_i}^i\rangle$是账户$c_i$的交易序列，$l_i$表示交易序列$W_i$的长度，$D=\{W_1, \cdots, W_n\}$是所有交易序列的集合(每个账户都对应一个属于自己的交易序列)。不同的账户分别属于不同的类型P_i，$P_i=\{c_1^i, c_2^i, \cdots, c_{k_i}^i\}(i=1,2,\cdots,m)$，其中$k_i$表示属于类型$P_i$的账户数量，显然$P_i\subset C$。对于每一个交易序列$W_i$，分别计算属于每个账户类型的账户有多少个，可以得到$S_i=\{(p_1^i, w_1^i), \cdots, (p_m^i, w_m^i)\}$，其中$(p_1^i, w_1^i)$表示属于类型$p_1^i$的账户有$w_1^i$个。$D_S$是所有$S_i$的集合，$D_S=\{S_1, S_2, \cdots, S_n\}$。于是对$D$进行聚类就转化为对$D_S$聚类。得到$D_S$的流程如下：

输入：所有账户的集合D和交易序列集合W
输出：$D_S=\{S_1, S_2, \cdots, S_n\}$

```
(1) For each W_i in W
(2)     For each c_i in W_i
(3)         If c_i ∈ P_i
(4)             w_i ← w_i + 1
(5)         Else
(6)             w_i ← 0
(7)         EndIf
(8)     EndFor
(9) EndFor
```

接下来，对D_S中的元素进行聚类。这里采用一种无监督不分层的快速k均值聚类算法[10]对交易序列进行聚类。k均值算法以k为输入参数，把n个对象的集合分成k个簇，使得结果簇内的相似度高，而簇间的相似度低。

3.3 建立基于置信度剪枝的低序马尔科夫模型

在聚类之后，接下来的工作是对每个聚类分别建立低序马尔科夫模型。回到之前的表述：假设$C=\langle c_1, c_2, \cdots, c_n\rangle$表示所有账户的集合，$W=\langle c_1, c_2, \cdots, c_n\rangle$表示当前某账户的交易序列，序列长度为$l$，概率$P(c_i\mid W)$表示当前账户接下来与账户$c_i$发生交易的可能性。于是交易序列的预测问题可以

表示为

$$c_{l+1}=\underset{c\in C}{\operatorname{argmax}}\{P(c_{l+1}=c|W)\}=\underset{c\in C}{\operatorname{argmax}}\{P(c_{l+1}=c|c_l,c_{l-1},\cdots,c_1)\} \quad (3)$$

其关键步骤是计算所有的条件概率 $P(c_{l+1}=c|c_l,c_{l-1},\cdots,c_1)$，并选择其中条件概率最大的那个所对应的交易账户作为预测结果。事实上，这几乎是不可行的。因为一方面，实际中的交易序列可以是任意长的，但是基于历史数据计算条件概率其交易序列长度总是有限的，无法满足实际需要；另一方面，过长的交易序列需要分析的马尔科夫状态数量是巨大的，导致很高的时间和空间复杂度。基于这两个原因，可以假设交易序列满足 K 序马尔科夫模型的条件，即未来发生交易情况不是由整个交易序列决定，而是由其中与当前时间最接近的 K 个账户组成的序列决定，其中 $K<<l$。这种假设也比较符合实际情况，即未来交易情况与最近的交易最相关，而与发生时间离现在较远的交易相关性小。

于是可以得到

$$\begin{aligned}c_{l+1}&=\underset{c\in C}{\operatorname{argmax}}\{P(c_{l+1}=c|W)\}\\&=\underset{c\in C}{\operatorname{argmax}}\{P(c_{l+1}=c|c_l,c_{l-1},\cdots,c_1)\}\\&=\underset{c\in C}{\operatorname{argmax}}\{P(c_{l+1}=c|c_l,c_{l-1},\cdots,c_{l-(K-1)})\}\end{aligned} \quad (4)$$

运用 K 序马尔科夫模型进行预测，其关键问题是计算 $P(c_{l+1}=c|c_l,c_{l-1},\cdots,c_{l-(K-1)})$，即 $S_j^K=\langle c_{l-(K-1)},c_{l-(K-2)},\cdots,c_l\rangle$。对于含有 n 个账户的金融网络，其状态数量为 P_n^K，所要计算的条件概率数量为 P_n^{K+1}。根据最大似然法则，可以通过在训练集中进行统计的方法得到所有的条件概率：

$$P(c_i|S_j^K)=\frac{Frequency(\langle S_j^K,c_i\rangle)}{Frequency(S_j^K)} \quad (5)$$

式中，$Frequency(S_j^K)$ 表示 S_j^K 出现的频率(次数)，$Frequency(\langle S_j^K,c_i\rangle)$ 表示交易序列中 c_i 紧跟在 S_j^K 后出现的频率(次数)。如前所述，当 K 增加时，状态数量急剧增长，因此考虑到预测准确性和时间复杂度之间的平衡，这里选取的 K 值较小，即低序。

在计算每个状态的概率值时，可能会遇到概率值很小的情况。这种状态在最后的预测结果中意义不大，因此在实际应用中可以采用基于置信度剪枝马尔科夫模型[4]。置信度剪枝模型保留那些预测状态概率差值高于某个阈值(ϕ_c)的状态，这个阈值称为置信度阈值。如果给定的条件状态得到的预测状态之间的概率差值低于置信度阈值(预测是模糊的)，那么这个状态将被排除。置信度阈值可以由式(6)得到：

$$\phi_c=\max-\lambda\cdot\sqrt{\max\cdot(n-\max)} \quad (6)$$

式中，max 是后续状态中最大的频率，n 是初始状态出现的频率，λ 为置信度系数。随着 n 的增加，置信度阈值减小，剪枝状态越多。一个状态出现频率越高，置信度阈值越小。因此，如果一个状态出现频率很高，即使其预测结果中两个最大的概率差值很小，这个状态仍然很可能被保留。同理，如果两个预测结果间概率差值很大，但是这个状态出现的频率太低，那么该状态仍然可能被剪枝掉。

使用基于置信度剪枝的 K 序马尔科夫模型，在聚类完成以后，预测流程如下所示：

(1)对所有的交易序列进行聚类；

(2)对每个聚类中的交易序列分别建立 K 序马尔科夫模型，预测特定账户的下一个交易对象；

(3) 如果预测结果是模糊的，则引入关联规则进行修正，给出最终的预测结果。

前面提到，基于置信度剪枝的马尔科夫模型，其预测结果可能是模糊的，此时需要引入关联规则，以便给出更准确的预测结果。下面讨论如何运用关联规则。

3.4 基于关联规则挖掘的预测结果修正

低序马尔科夫模型在 K 取值较小时，可能会出现预测结果模糊的情况，此时我们引入关联规则，修正预测结果。1994 年，R. Agrawal 和 R. Srikant 提出了布尔关联规则挖掘频繁项集的 Apriori 算法。Apriori 算法通常用于挖掘大量的无序数据中的频繁项集，但是交易序列显然是有序的。2004 年，Yang[11] 提出了 5 种序列关联规则挖掘算法。针对本文的研究内容，将采用子序列关联规则[11]。

下面结合交易序列的特点介绍子序列关联规则挖掘的思想：

假定我们有表 2 所列的属于同一个聚类的 3 个交易序列，这里为描述简便，仅以 2 序($K=2$)马尔科夫模型为例，在该聚类上建立 2 序马尔科夫模型。

表 2 聚类后的交易序列

W_1	B,G,I,J,F,D,E,K,A,I,J,H,E,K,A,D,I,J,H
W_2	G,E,K,A,I,J,F,B,D,E,K,A,I,J,H
W_3	G,E,K,I,J,F,K,G,I,J,F,E,K,N

现在考虑这样的交易序列$\langle I,J,?\rangle$。若以 0.2745 作为置信度阈值，可以得到如下的统计结果：

$c_{l+1}=\operatorname{argmax}\{P(F|J,I)\}=\operatorname{argmax}\{F\rightarrow 0.57\}$

$c_{l+1}=\operatorname{argmax}\{P(H|J,I)\}=\operatorname{argmax}\{H\rightarrow 0.43\}$

$0.57-0.43<0.2745$，作为后续状态的 F 和 H 之间出现的概率差值小于置信度阈值，这个预测结果是模糊的。为了得到准确的预测结果，考虑$\langle I,J\rangle$之前的状态。运用子序列关联规则挖掘算法可以得到表 3 所列的结果。

表 3 子序列关联规则

B,G	⟨I,J⟩	F
D,E,K,A	⟨I,J⟩	H
E,K,A,D	⟨I,J⟩	H
G,E,K,A	⟨I,J⟩	F
B,D,E,K,A	⟨I,J⟩	H
G,E,K	⟨I,J⟩	F
K,G	⟨I,J⟩	F

其中预测为 F 的关联规则的置信度如表 4 所列。

表 4 预测为 F 的关联规则的置信度

B->F	BF/B	1/2	50%
G->F	GF/G	4/4	100%
E->F	EF/E	2/5	40%
K->F	KF/K	3/6	50%
A->F	AF/A	1/4	25%

预测为 H 的关联规则的置信度如表 5 所列。

表 5 预测为 H 的关联规则的置信度

D->H	DH/D	3/3	100%
E->H	EH/E	3/5	60%
K->H	KH/K	3/6	50%
A->H	AH/A	3/4	75%
B->H	BH/B	1/2	50%

从表4和表5中可以看到$G->F$和$D->H$有最高的置信度，于是得到如下结论：如果I,J之前的状态中有G，则预测结果为F；若有D，则预测结果为H。若G,D同时出现或者同时不出现，则考虑第二高的置信度，即B,K和A，此时A具有较高的置信度，即预测结果为H。若置信度仍然相同，则重复以上步骤，直到给出最终的预测结果。

3.5 建模和预测算法

结合前面的讨论，本节将给出完整的建模和预测算法。这里考虑到描述的简便，同样仅以2序($K=2$)马尔科夫模型为例对算法进行描述，$K>2$的情况类似。

首先是建模算法。

算法1 根据已有的账户交易记录，建立混合2序马尔科夫模型，得到规则矩阵Rule(即预测结果)。步骤如下：

1)根据所有账户的交易序列集合，运用k均值聚类算法[10]对交易序列集合进行聚类。

2)对于每一个聚类，执行以下步骤：

对于当前聚类中的所有交易序列$S_1,S_2,\cdots,S_n$，遍历所有的子序列S_i,S_{i+1},S_{i+2}，统计账户序列S_i,S_{i+1}下账户S_{i+2}出现的频率，可得到初始的规则矩阵Rule。Rule[p][i][j]记录了第p个聚类中在账户序列i,j出现的情况下，接下来最有可能出现的两个账户以及出现的频率。

3)对于规则矩阵Rule[k][i][j]，如果接下来最有可能出现的两个账户的频率差值小于置信度阈值，则利用3.4节中的方法更新Rule[k][i][j]，也即更新账户i,j的下一个可能发生交易的账户。

该算法包含3个部分：聚类、建立2序马尔科夫模型、挖掘关联规则。聚类部分时间复杂度为$O(nkt)$，其中n为账户总数量，t为迭代次数，k为聚类个数。建立2序马尔科夫模型需要遍历一遍所有的交易序列，假设交易序列平均长度为l，则时间复杂度为$O(ln)$。挖掘关联规则需要遍历三维规则矩阵，在遍历过程中判断马尔科夫模型规则是否有效；若无效，则遍历一遍所有交易序列，挖掘相应的关联规则，时间复杂度为$O(kln^3)$，其中k为聚类个数。综上，建模部分总的时间复杂度最差的情况为$O(kln^3)$，最好的情况下为$O((n+l)n)$。空间方面，需要一个三维矩阵和kn^2个含n个键值对的哈希表，复杂度为$O(kn^3)$。

下面介绍预测算法。

算法2 预测账户的未来交易发生情况。根据账户，由算法1可得账户所属的聚类，记为p，该账户交易序列的最后两个账户记为i,j，则当前账户的未来交易发生情况可直接由Rule[p][i][j]得到。

该算法只需要在马尔科夫规则矩阵中去查找相应的规则即可，时间复杂度为O(1)，空间方面只需存储预测结果，复杂度为O(1)。

最后讨论一下算法的动态性体现。建模算法的核心是建立马尔科夫规则矩阵，当有新的交易数据存入数据库时，只需分析新增加的数据更新规则矩阵即可。

4 实验和结果分析

实验环境为Intel Core2 Duo 2.26GHz，2048MB RAM，测试平台为Window XP sp3。实验数据来自某金融监管单位的经过预处理消除了用户敏感信息的大额可疑交易数据，实验中一共使用了来自2003年的115个交易账户、1472条大额可疑交易数据。实验数据基本结构如表6所列。

表6 实验数据基本结构

交易主体账号	交易客体账号	交易时间	交易地点	交易金额	交易类别

在试验过程中，将数据分为两个数据集，每个数据集又分为测试集和训练集。第一个数据集为2003年7月1日到2003年12月31日期间的数据，共734条交易记录，其中取2003年7月1日到2003年10月31日期间的493条交易记录作为训练集，2003年11月1日到2003年12月31日的241条交易记录为测试集；第二个数据集为2003年1月1日到2003年12月31日期间的数据，共1472条交易记录，其中取2003年1月1日到2003年10月31日期间的1231条交易记录作为训练集，2003年11月1日到2003年12月31日的241条交易记录为测试集。

实验分为两个部分：第一部分分别测试了基本马尔科夫模型、2序马尔科夫模型、结合关联规则的2序马尔科夫模型、结合聚类的2序马尔科夫模型(聚类个数为2，迭代次数100次)、3序马尔科夫模型以及结合聚类、关联规则的2序马尔科夫模型(聚类个数为2，迭代次数100次)。分别对比了它们在不同数据集上的状态数量和预测的准确性。

从表7中可以明显看到，基本马尔科夫模型的状态是很少的，状态数量两倍于预测对象的数量(考虑了方向性)，2序马尔科夫模型在同样的数据集上其状态数量有了大幅度的提升，3序马尔科夫模型的状态数量最多。应用关联规则无法减少状态数量(可以提高预测的准确性，后面将会提到)，聚类只能在一定程度上减少状态数量，而采用置信度剪枝方法可以较好地减少状态数量。当序大于3时，马尔科夫模型的状态数已经急剧增加，由此带来的准确性较之时间复杂度显得得不偿失，故在试验中并未显示出来。

表7 6种马尔科夫模型在两个数据集上的状态数量

模型	数据集	
	数据集1	数据集2
基本马尔科夫模型	230	230
2序马尔科夫模型	863	1724
结合关联规则的2序马尔科夫模型	863	1724
结合聚类的2序马尔科夫模型	782	1607
3序马尔科夫模型	1149	2037
结合聚类、关联规则的2序马尔科夫模型(置信度剪枝)	591	1307

图1是6种模型在两个数据集上预测的准确性比较。

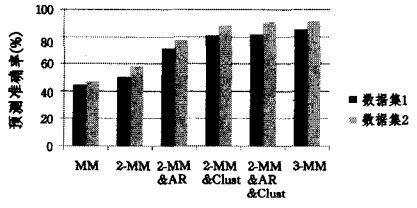

图1 6种马尔科夫模型在两个数据集上预测的准确度

该实验表明：

(1)基本马尔科夫模型的状态数量最少，等于预测对象的

数量;2 序马尔科夫模型的状态有了质的提升;应用关联规则无法降低状态数量;聚类能在一定程度上降低状态数量;基于置信度的剪枝可以较好地降低状态数量。

(2)基本马尔科夫模型的预测准确性比较低;采用 2 序马尔科夫模型可以提高预测准确性,但是时间复杂度也有很大幅度的增加;结合关联规则可以进一步提高预测准确性,但无法降低时间复杂度;结合聚类比关联规则有更好的效果,不仅能提高预测的准确性,还能降低时间复杂度;将三者结合起来能达到最好的效果。

(3)3 序马尔科夫模型的预测准确性比结合聚类、关联规则的 2 序马尔科夫模型更高,但是其状态数量也更多,这意味着时间和空间复杂度提高,而预测准确性的提高幅度与时间空间开销的增加相比是非常有限的。当序大于 3 时,时间空间开销急剧增加,由此带来的准确性提高显得得不偿失,因此结合聚类和关联规则的 2 序马尔科夫模型更有效。

第二个实验针对结合聚类和关联规则的 2 序马尔科夫模型,分别设置不同的聚类个数,看聚类个数对预测准确性的影响。具体过程如下:数据库中的账户分别属于 3 种不同的类型,这样计算每个交易序列中分别属于每个类型的账户的数量,就可以将一个交易序列转化为一个三维向量。然后对所有三维向量进行聚类,迭代次数为 100。若设置聚类个数为 2,则两个聚类包含的交易序列个数分别为 79,38;若设置聚类个数为 3,则 3 个聚类包含的交易序列个数分别为 46,15,54;若设置聚类个数为 4,则 4 个聚类包含的交易序列个数分别为 30,8,51,26。3 种初始聚类个数造成的预测准确性如图 2 所示。

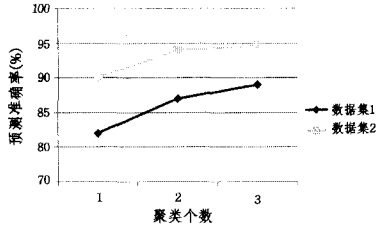

图 2 不同的聚类个数对预测准确率的影响

该实验表明,聚类个数的增加能提高预测的准确性,但当聚类个数达到一定数量后,这种提升是有限的,甚至趋于不变。

对于实验现象做如下分析。

(1)第一个实验说明聚类和关联规则的应用确实提高了预测的准确性,其中聚类造成的效果更好。主要体现在两个方面:一方面是预则准确性提升的程度,这主要是因为聚类使模型更具有针对性,而关联规则的局限性在于它并非总是被使用的,仅当马尔科夫模型生成的规则出现模糊时才会采用关联规则;另一方面,聚类能降低马尔科夫状态数量。前面已经提到,含有 n 个节点的 k 序马尔科夫模型的状态数量是 P_n^k,现在引入聚类,假设分为 m 个聚类,每个聚类含 t_i 个节点,则状态数量为 $\sum_{i=1}^{m} P_{t_i}^k$,其中 $\sum_{i=1}^{m} t_i = n$。很容易证明 $\sum_{i=1}^{m} P_{t_i}^k <$ P_n^k。

(2)第二个实验现象说明聚类个数的增加能提升预测的准确性。因为聚类越多,模型的针对性越强,其预测准确性自然有所提升。但这种提升是有限的,当聚类数量达到一定规模后,其预测准确性会趋于不变。

结束语 这里提出一种结合低序马尔科夫模型和聚类、关联规则的混合马尔科夫模型,并将其用于反洗钱的账户交易预测中。通过对不同聚类中的账户分别建立独立的低序马尔科夫模型,降低了模型的规模,同时在模型的建立过程中基于置信度进行剪枝,降低了时间复杂度。由于序较低,会出现较多预测结果是模糊的情况。为了提高预测准确性,引入了关联规则。在马尔科夫模型规则无效的情况下,给出最终的预测结果。通过对 1 序、2 序和 3 序及大于 3 序的马尔科夫模型的对比发现,2 序马尔科夫模型在预测准确性和时间复杂度方面的效果最好。当然,不同的应用背景可能结果并不一定相同。实验表明,该混合马尔科夫模型具有较高的预测准确性,并在预测准确性和时间复杂度两者之间取得了较好的平衡。

参 考 文 献

[1] 李东荣. 洗钱与反洗钱(第 1 版)[M]. 北京中国财政经济出版社,2004:1-37

[2] Zhu Jian-han, Hong Jun, Hugnes J G. Using Markov Chain for Link Prediction in Adaptive Web Sites[J]. Lecture Notes in Computer Science, 2002, 2311:55-66

[3] Jayalal S. Website link prediction using a Markov chain model based on multiple time periods[J]. International Journal of Web Engineering and Technology, 2007, 3(3):271-287

[4] Deshpande M, Karypis G. Selective Markov Models for Predicting Web Page Accesses[J]. ACM Transactions on Internet Technology, 2004, 4(2):163-184

[5] Khalil F, Li J, Wang H. A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses [C]// AusDM. 2006:177-184

[6] 郭景峰,王春燕,邹晓红,等. 一种改进的针对合著关系网络的链接预测方法[J]. 计算机科学,2008,35(12):126-128

[7] 高金余,陈翔. 马尔科夫切换模型及其在中国股市中的应用[J]. 中国管理科学,2007,15(6):20-25

[8] 张冬青,宁宣熙,刘雪妮. 考虑影响因素的隐马尔科夫模型在经济预测中的应用[J]. 中国管理科学,2007,15(4):105-110

[9] 张冬青,韩玉兵,等. 基于小波域隐马尔科夫模型的时间序列分析-平滑、插值和预测[J]. 中国管理科学,2008,16(2):123-127

[10] Kanungo T. An Efficient k-Means Clustering Algorithm: Analysis and Implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7):881-892

[11] Yang Q, Li I, Wang K. Building Association-Rule Based Sequential Classifiers for Web-Document Prediction[C]// Proceedings of Data Min. Knowl. Discov. 2004:253-273