一种高维聚类算法及在洗钱侦测中的应用*)

陈云开　卢正鼎　刘　芳　郭　洁
(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　从技术的角度看,洗钱侦测问题实际上是一个数据分析问题。本文首先给出了一个可疑交易判定模型,并提出了一个基于超图模型的高维聚类算法,运用该算法从案例库中形成可疑交易模式,最后给出了可疑交易的判定方法。该基于超图的高维聚类算法具有以下特点:1)能处理大数据集;2)能适应高维数据;3)聚类结果是可理解、可解释和可用的。

关键词　高维,超图模型,聚类,洗钱

A High Dimension Clustering Algorithm and its Application in Detecting Money Laundering

CHEN Yun-Kai　LU Zheng-Ding　LIU Fang　GUO Jie
(School of Computer Science and Technology, Huazhong University of Science&Technology, Wuhan 430074)

Abstract　From the point of view of technique, detectiing money laundering is process of data analysis. This paper first proposes a hypergragh-based clustering algorithm to find out suspicious business mode about money laundering. And then it discusses way of determination suspicious business. The algorithm could solve the problems of 1) large volume of data set; 2) data set of high dimension; 3) easy understanding of result. The algorithm has been employed in the money laundering detection system, which is the first anti-money laundering system of our nation.

Keywords　High dimension, Hypergragh model, Clustering, Money laundering

1　引言

从技术的角度看,洗钱侦测问题实际上是一个数据分析问题。数据聚类是进行数据分析的一个重要技术。在数据挖掘中常用的数据聚类方法是基于距离的分割聚类算法,该类算法能有效地应用于空间维数较小的情况,对于高维空间则不能产生有意义的聚类结果。因此,一种常用的高维空间数据聚类的处理方法是:首先在保留原始数据之间关系的前提下,减少数据项的维数,例如:主要成份分析法(PCA)、多维缩放法(MDS)都是常用的减少数据维数的方法;其次,将传统聚类算法例如:K-means 算法、AutoClass 算法应用于转换后的低维数据空间。这里,由于维数的缩小使得在低维空间中噪声点可以非常接近有效数据点进而直接影响到数据聚类的质量[1]。

在过去的几十年里,图论已被证明是解决几何、数论、运筹学和优化领域中重要问题非常有用的工具。为了解决更多的组合问题,把图的概念进行推广是非常自然的事情[2]。图概念是 C. Berge 于 1970 年提出的。由于超图理论比较抽象,研究者很不容易入门,超图理论的发展一直比较缓慢[3]。近年来,超图理论及其应用的研究越来越为人们所重视[4]。

基于超图模型的分割可能适用于高维数据的聚类[5~7],高维空间的关系转化成超图,用超边的权重来描述空间点间的关系[8,9]。对超图的分割实际上就是聚类的过程,将权重大的超边中包含的数据点尽量放在一个类中,同时使被切割的超边权重之和最小[10]。

2　可疑交易判定模型

在对洗钱行为进行监测的过程中将对金融交易划分为四个层次进行分析:第一个层次是交易层(Transaction Level)。包括一系列的单独行为,如现金的存取,电汇和支票;第二个层次是个人/账户层(individual or account level)。多笔交易都是由特殊的个人或银行账户构成的;第三个层次是公司/组织层(business or organizational level)。洗钱的主体可能是一个公司,其中包含多个账户和个人;第四个层次是"环状"层("ring" level)。这一层包括与一项洗钱方案有关的多个公司,账户和个人。

在洗钱侦测中的金融交易形式大致为:汇款人、汇款金额、汇款时间、汇款地、收款人、收款地,根据这些信息分析交易,可看出,从单独的一笔汇款无法判断该交易是否属于可疑的洗钱交易,需要扩充单笔交易的信息,增加频度、交易范围、累积金额等反映交易主体的可用于决策的信息,形成以交易主体的交易模式为中心的判别模型[1]。

以交易主体的交易模式为中心的判别模型的原理如下:单笔交易信息无法支持是否可疑的判断,需要扩充单笔交易的信息,增加频度、交易范围、累积金额等可用于决策的信息,与案例库中的可疑交易信息匹配,并返回决策结论。可分为如下几个步骤:

1. *构建案例库*

根据检查出的可疑交易主体,搜索该交易主体的历史记录,从中筛选可疑交易,形成案例库。

给定包含 n 个交易对象的案例库 $D=\{P_1, P_2, \cdots, P_n\}$,对象 $P_i(1\leqslant i\leqslant n)$的形式为向量:(交易主体 *ID*、时间范围、频度、平均金额〈包括收款和付款〉、累积金额〈包括收款和付款〉、交易数量、对方交易主体数目、跨地区数、跨银行数、使用帐户数……)。

*)基金资助:国家自然科学基金(60403027)。陈云开　博士研究生,研究方向为金融数据挖掘;卢正鼎　博士生导师,研究方向为分布式系统和数据库系统。

2. 从案例库中形成可疑交易模式

给定包含 n 个交易对象的案例库 $D=\{P_1, P_2, \cdots, P_n\}$，对象 $P_i(1\leqslant i\leqslant n)$ 的形式为向量：(交易主体 ID、时间范围、频度、平均金额(包括收款和付款)、累积金额(包括收款和付款)、交易数量、对方交易主体数目、跨地区数、跨银行数、使用帐户数……)。

通过对案例库 D 聚类，获取 k 个类，类 j 的交易集合用 C_j 表示$(1\leqslant j\leqslant k)$，$|C_j|$ 表示类 j 包含的对象个数。

类 j 的模式用 M_j 表示，M_j 以类中所有对象的向量和取均值得到，即 $M_j=\vec{C}_j/|C_j|$。

由于洗钱行为的多样化，形成的可疑交易模式也存在多种。

3. 可疑交易的判定

对于需要判断的交易，根据其中的两个交易主体(即汇款人、收款人)的历史交易记录，获取频度、累积金额、交易范围等信息，再将这些信息与案例库中已标记为可疑的交易模式匹配，如存在可匹配模式，则判定为可疑，如不存在则判定为正常。

根据给定交易 t 的两个交易主体，搜索一定时间范围内的信息，形成两个交易对象，分别为 P_x 和 P_y，对象形式用向量表示，具体同上。

对于给定的阈值 τ，将 P_x 与 k 个 $M_j(1\leqslant j\leqslant k)$ 逐个匹配，计算两者的相似度 $\mathrm{sim}(P_x, M_j)$，其中 $\mathrm{sim}(x, y)$ 表示对象 x 和对象 y 的相似度；以同样方式对 P_y 操作，这样得到两个相似度值，其中一个大于阈值 τ，则认为交易 t 为可疑交易；否则，交易 t 为正常交易。

3 基于超图的聚类

模式发现过程涉及到聚类过程，本文采用基于超图的方法对案例库 D 聚类，本节详细描述了聚类过程，并根据聚类结果得到可疑交易模式。

3.1 频繁项集和关联规则

由于基于距离的传统描述方法只能描述两个点之间的相关性，而用一个关联规则算法开采出的频繁项目集能够描述多个高维数据点之间的相关性，因此，可以考虑用频繁项目集对超图进行描述。超图模型要用到关联规则和频繁项集，因此首先介绍这两个概念的定义。

令 I 为 m 个不同项目的集合，即 $I=\{i_1, i_2, \cdots, i_m\}$，$D$ 为交易数据集合，每条交易 T 是一组项目的集合，$T\subseteq I$。一条关联规则具有以下形式：$X\Rightarrow Y$，其中 X，$Y\subset I$，且 $X\cap Y=\varnothing$。项目的集合称作项目集(itemset)，项目集的长度为项目集中所包含项目的数量，用 k-itemsets 表示长度为 k 的项目集，k 也称作维数，每个项目集都有一个支持度(support)，对于项目集 X，它的支持度记作 $\sup(X)$。给定项目集 $X\subset I$，如果在 D 中包含 X 的交易占整个集合中交易数量的百分比为 s，则 $\sup(X)=s$。规则 $X\Rightarrow Y$ 的支持度定义为 $\sup(X\cup Y)$，置信度定义为 $\sup(X\cup Y)/\sup(X)$，那么关联规则开采问题可以描述为从数据集合中生成满足下列两个条件的规则：① 规则的支持度大于给定的阈值最小支持度 minsup；② 规则的置信度大于给定的阈值最小置信度 minconf。满足条件①的项目集称作频繁项目集，否则称作非频繁项目集。

3.2 超图模型定义

根据前面对问题的描述，案例库 D 形成向量空间，其中每个对象对应为空间中的一个点。这里，超图 $H=(V,E)$，V 对应案例库 D，每个顶点对应 D 中的一个对象，超边 $E_k(k=1,2,\cdots,t)$ 连接包含一个频繁项集的所有顶点，顶点集记为 $V(E_k)$，它对应的频繁项集记为 $I(E_k)$。

3.3 建立超图模型

超边 $E_k(k=1,2,\cdots,t)$ 的支持数为：

$$\mathrm{se}(E_k)=|V(E_k)|=|\{v|\in V, I(E_k)\subseteq I(v)\}| \tag{1}$$

超边 $E_k(k=1,2,\cdots,t)$ 的支持度为：

$$\mathrm{support}(E_k)=\mathrm{se}(E_k)/n \tag{2}$$

设超边 E_k 中发现的关联规则 r 为 $X\Rightarrow Y$，其中 $X\subseteq I(E_k)$，$Y\subseteq I(E_k)$，则该关联规则的置信度为：

$$\mathrm{confi}(r)=\mathrm{se}(X\cup Y)/\mathrm{se}(X) \tag{3}$$

设超边 E_k 中发现 r_k 个关联规则，则超边 E_k 的平均置信度为：

$$\mathrm{confi}(E_k)=\sum_r \mathrm{confi}(r)/r_k \tag{4}$$

当然，需要设定最小支持度和最小置信度，只有满足最小支持度和最小置信度的超边才是聚类时要考虑的对象。在文[9]中，用了关联规则来定义超边的权重，这里给出一种加权的超边权重定义方法。它用超边的平均置信度以及与其它超边的公共顶点在其它超边中所占比例来描述超边的权重，有助于更好地描述顶点之间联系的紧密程度。于是按式(5)定义这些满足条件的超边的权重。

$$w(E_k)=\mathrm{confi}(E_k)/\left(1+\sum_{j\neq k, j=1}^{t}\left(\frac{|V(E_k)\cap V(E_j)|}{\mathrm{se}(E_j)}\cdot \mathrm{confi}(E)\right)\right) \tag{5}$$

其中，t 表示与 E_k 顶点的超边数目。

通常，每个(属性名，属性值)称为项，项的集合称为项集 I。选择这些感兴趣的属性及对应属性值为项的项集作候选项集，然后采用 Apriori 算法，它是目前最有影响的挖掘布尔关联规则频繁项集的算法之一。找到频繁项集后，频繁项集 $I_j(j=1,2,\cdots,s, s$ 为频繁项集个数$)$ 与记录 $v_i(i=1,2,\cdots,n)$ 是多对多的映射关系，如图 1 所示。然后根据频繁集，建立超图模型，依式(5)计算超边权重。

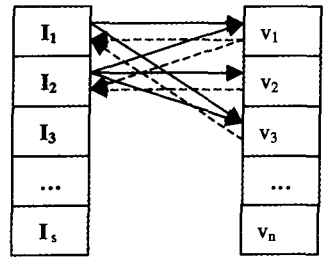

图 1 记录与频繁集的对应关系

3.4 基于超图的聚类算法

得到超图的基础上，对数据集进行聚类。设将数据分成 k 类，基于超图的聚类算法就是完成对超图的 k 类分割，它的基本思想是使用超边分割方法，保持每次打断的超边权重和最小，直到每个分割中的数据都密切相关为止，最终得到的分割就是聚类的结果及其模式。聚类及模式发现过程详细描述如下：

(1)超图粗化

定义形成的超图模型为 $H_0=(V_0,E_0)$，H_0 不断粗化形成一个超图序列 $H_1, H_2, \cdots, H_m$，其中 $H_i=(V_i,E_i)(i=1,2,\cdots,m)$ 且 $|V_0|>|V_1|>|V_2|>\cdots>|V_m|$。每步粗化都选择

当前权重最大的边，该边包含的点合并成一个点后映射到新超图中，当所有的超边都被粗化后停止，原超图模型 H_0 已经被初步地分割成 k'。通常用最小支持度和最小置信度来控制超边数目的大小，也就控制了 k' 值。如果 $k' \leqslant k$，则算法停止，否则进行到下一步骤中。

(2)合并

这个步骤实际上是一个合并的过程，考虑现有的每个类与其它类的合并，找到最匹配的两个类加以合并。最匹配的两个类选择的标准就是使得切断的超边权重之和最小，重复这一过程直到出现 k 个分割为止。

(3)模式计算

聚类结束后，得到 m 个类，求得均值，获取可疑交易模式。

3.5 算法描述

基于超图的聚类算法的伪代码如下：

算法 1 HCA(Hypergraph-based Clustering Algorithm)

输入：数据集 X，超边列表 E(其中每个超边为 $E_t(t=1,2,\cdots,k')$)，每个超边 $E_t(t=1,2,\cdots,k')$ 中发现的关联规则集合 R_t(其中每个元素表示为(关联规则，对应的置信度)值对)，每个超边 $E_t(t=1,2,\cdots,k')$ 的顶点集 $V(E_t)$，目标聚类数 k

输出：$M[1,2,\cdots,m]$，m // $M[1,2,\cdots,m]$为聚类得到的类所形成的模式，m 为类的个数

```
(1){
(2)while(t in {1,2,…,k'})按式(5.4)计算超边 E_t 的平均置信度;
(3)while(t in {1,2,…,k'})按式(5.5)计算超边 E_t 的权重;
(4)按超边权重降序排列超边列表 E;
(5)for(t=1;t<=k';t++)
(6){
(7)E_t 中未分类的点合成类C[t];
(8)C[t]=t;
(9)}//超图粗化结束，得到 k' 个初始类，形成类列表 C[1,2,…,k']
(10)int n=k';
(11)while(getCount(C)>k)// getCount(C)得到类列表 C 的元素个数
(12){
(13)int p=1,q=2;
(14)for(int i=1;i<=getCount(C);i++)
(15)for(int j=i+1;j<=getCount(C);j++)
(16){
(17)if(! (w(C[i], C[j])=hash(C[i], C[j])))
   //查 hash 表
(18)计算 C[i]与 C[j]合并时所切割的超边权重之和 w(C[i], C[j])，加入 hash 表中;//hash 表中，hash(C[i], C[j])=w(C[i], C[j])
(19)if(w(C[i], C[j])< w(C[p], C[q])){p=i;q=j;}
(20)}
(21)合并 C[p]和 C[q]，在类列表 C 中删除 C[p]和 C[q]，同时加入新类 C[s]
(22)n++;C[s]=n;
(23)}
(24)for C 中的每一个类C[i]
(25)M[i]=means(C[i])//将每个类 C[i]中的向量和平均值赋给 M[i]，获取模式
(26)}
```

3.6 算法分析

选取合适的最小支持度值是很重要的，当最小支持度值选取合适值时，它能快速发现大数据库中的频繁集。阈值过小，造成算法耗时过长，而且可能发现一些关联度很小的频繁集，这将进一步影响到聚类结果。

超边的数目只跟发现的关联规则数目有关，而关联规则的数目不会随着数据量的增加而变多，同时在聚类的粗化阶段采用超边粗化能够很快降低数据的规模，因此此算法适用于大数据集的聚类。

高维空间中数据分布较稀疏，用超边的权重来描述数据间的相似度，能够克服常用聚类算法中用数据之间的距离尺度来衡量数据间的相似度造成聚类结果不佳的情况。

该算法耗时主要花费在计算两类合并时切割超边权重之和上，在实现过程中使用 hash 函数，使得计算切割超边权重之和的次数大大减少，从而减少了时间花费，计算复杂度为 $O(E^2)$，其中 E 为超边条数，即满足最小支持度和最小置信度的频繁集个数。

4 可疑交易判定算法

步骤大致为：获取给定交易 t 的详细信息，并形成两个向量 P_x，P_y，这两个向量分别以交易 t 的汇款人和收款人的信息构成。接着将向量与交易模式 M 中的模式匹配，一旦发现相似度超过阈值 τ 的，即为可疑交易，否则为正常交易。其算法的伪代码如下：

算法 2 Suspicious($t, T, M^{[1,,m]}, \tau$)

输入：需要判定是否可疑的交易 t，原始交易库 T，以及从案例库中获取的模式 $M[1..m]$，阈值 τ

输出：布尔值 suspicious // 1 表示交易 t 可疑，0 表示 t 为正常交易

```
(1){
(2)GetInfo(t, T, P_x, P_y); //从交易库 T 中获取 t 的两个交易主体信息，形成向量 P_x, P_y
(3)for M 中的每个模式 M[i]
(4)s_x=Sim(P_x, M[i]); // 将交易信息与可疑交易模式匹配
(5)s_y=Sim(Py, M[i]);
(6)if s_x>=τ or s_y>=τ
(7)return(1)
(8)return(0)
(9)}
```

5 实验及结论

5.1 实验

以 2003 年上半年某沿海地区的外汇交易数据为数据源，验证上面的基于超图聚类算法的可疑交易模式发现方法和可疑交易判别方法。对外汇交易原始表进行预处理，最终形成交易主体表(交易主体 ID、时间范围、频度、平均金额〈包括收款和付款〉、累积金额〈包括收款和付款〉、交易数量、对方交易主体数目、跨交易地区数、跨交易银行数、使用帐户数、累计交易金额/注册资本金、公司注册时间)。

(下转第 213 页)

容忽视。以名词和动词两者作为候选词条要明显好于所有词性的词，这表明通过只选择名词和动词作为词条，足以反映网页内容的实际意义，并且还可以消除代词、形容词、数量词、介词、副词等对分类产生的噪音，改善分类效果。

3.4 最终实验结果

经过前面的实验分析和比较，我们确定最终的网页分类方案，即，抽取网页中的普通文字，提高标题的权重，使标题与正文的权重之比为9，选择名词和动词作为候选特征，并采用文档频率的特征选择方法选择4000个特征词，以朴素贝叶斯算法作为分类器模型，对我们的中文网页数据集进行训练和测试，各个类别的分类效果(Precision和Recall)如图5所示。采用上述一系列预处理和特征处理方法后，Micro-F1可达到94.9%，相比之下，仅抽取全文且不做特殊特征处理时的Micro-F1仅有81.5%(注：图2中"全文"曲线取特征词为4000时，Micro-F1=81.5%)。

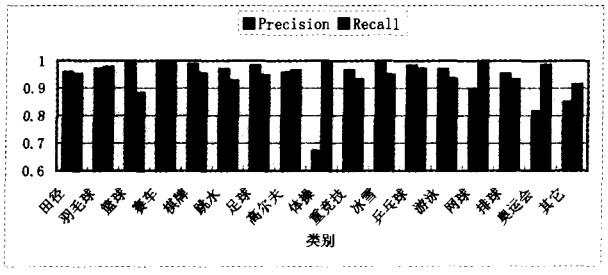

图5 各类别的分类结果

总结 本文介绍了我们设计实现的一个中文网页分类系统。通过实验分析，得出如下结论：通过抽取网页中的"普通文字"可以有效地去除噪音，适当提高网页标题在特征空间中的权重，只选择名词和动词作为候选词条，都可以改善分类效果。在我们采集的中文网页数据集上，与不做特殊处理相比，我们的方法可以使分类效果指标Micro-F1由81.5%提高到94.9%。今后我们将扩大中文网页数据集的数量和主题范围，进行更大规模的研究，同时将我们的系统应用于医学网页的分类应用中。

参考文献

1 Shih L K, Karger D R. Using URLs and Table Layout for Web Classification Tasks. In: Proceedings of WWW'04, New York, New York, USA, 2004

2 孙承杰，关毅. 基于统计的网页正文信息抽取方法的研究. 中文信息学报，2004，18(5)：17～22

3 Yang Yinming, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of ICML, Nashville, Tennessee, USA, 1997

4 单松巍，冯是聪，李晓明. 几种典型特征选取方法在中文网页分类上的效果比较. 计算机工程与应用，2003，39(22)：146～148

5 Chinese Segmenter. http://www.mandarintools.com/segmenter.html

6 ICTCLAS. http://www.nlp.org.cn/project/project.php?proj_id=6

7 Porter Stemming Algorithm. http://www.tartarus.org/martin/PorterStemmer/

8 黄科，马少平. 基于统计分词的中文网页分类. 中文信息学报，2002，16(6)：25～31

9 周水庚，关佶红，胡运发，等. 一个无需词典支持和切词处理的中文文档分类系统. 计算机研究与发展，2001，38(7)：839～844

(上接第193页)

案例库中的交易主体表中有230条数据。设最小支持度为0.03，最小置信度为0.15，建立的超图模型有212个顶点和223个超边。超图模型中的顶点数比交易主体表中的顶点数少，可以看出由于在建立超图时设置了最小支持度和最小置信度，使部分交易主体排除在外，将它们看作是奇异点，在聚类时不予考虑。运用超图聚类算法最终得到12个类，仔细分析这些类得到部分可疑交易模式。如：

1)单笔交易金额大，单笔交易金额远远以往平均交易金额。

2)交易主体跨多家金融机构、跨多个地区、同时使用多个账户进行资金收付。

3)交易金额大幅超出其注册资本金股本，且公司成立时间较短。

5.2 结论

基于超图模式的聚类方法的主要思想是把一个求解高维空间聚类问题转换为一个超图分隔寻优问题。该算法具有以下特点：1)能处理大数据集；2)能适应高维数据；3)聚类结果是可理解、可解释和可用的。该算法适用于金融可疑交易判定，实验表明：与传统方法相比，该方法能有效祛除噪声点，在高维空间获得优秀的聚类结果。

参考文献

1 Mehammed K. Data Mining Concepts, Models, Methods, and Algorithms [M]. BeiJing: Qinghua university Press, 2002

2 李刚. 知识发现的图模型方法：[博士学位论文]. 中国科学院软件研究所，2001

3 Karypis G, Kumar V. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. SIAM Journal on Scientific Computing, 1998, 20(1): 359～392

4 Karypis G, Aggarwal R, Kumar V, et al. Multilevel Hypergraph Partitioning: Application in VLSI Domain. In: the 34th ACM/IEEE Design Automation Conf., Anaheim, California, United States, Jun. 1997. 526～529

5 Karypis G, Kumar V. Multilevel k-way Hypergraph Partitioning. In: The 36th ACM IEEE Design Automation Conference, New Orleans, LA, Jun. 1999, 11(3): 343～348

6 Karypis G, Kumar V. Multilevel k-way Partitioning Scheme for Irregular Graphs. Journal of Parallel and Distributed Computing, 1998, 48(1): 96～120

7 Karypis G, Kumar V. Multilevel Algorithms for Multi-Constraint Graph Partitioning. In: Proceedings of the 1998 ACM/IEEE conference on Supercomputing, San Jose, CA, Nov. 1998. 28～44

8 Sanchis L A. Multiple-Way Network Partitioning. IEEE transactions on computers, 1989, 38(1): 62～81

9 Han Eui-Hong(Sam), Karypis G, Kumar V, et al. Clustering in A High-Dimensional Space Using Hypergraph Models: [Technical Report TR-97-063]. Department of Computer Science, University of Minnesota, Minneapolis, 1997

10 Han Eui-Hong (Sam), Karypis G, Kum V, et al. Hypergraph Based Clustering in High Dimensional Data Sets: A Summary of Results. IEEE Data Engineering Bulletin, 1998, 21(1): 15～22

11 Karypis G, Kumar V. Multilevel k-way Partitioning Scheme for Irregular Graphs. Journal of Parallel and Distributed Computing, 1998, 48(1): 96～120

12 Karypis G, Kumar V. Multilevel Algorithms for Multi-Constraint Graph Partitioning. In: Proceedings of the 1998 ACM/IEEE conference on Supercomputing, San Jose, CA, Nov. 1998. 28～44

13 Sanchis L A. Multiple-Way Network Partitioning. IEEE transactions on computers, 1989, 38(1): 62～81

14 Han Eui-Hong(Sam), Karypis G, Kumar V, et al. Clustering in A High-Dimensional Space Using Hypergraph Models: [Technical Report TR-97-063]. Department of Computer Science, University of Minnesota, Minneapolis, 1997

15 Han Eui-Hong (Sam), Karypis G, Kum V, et al. Hypergraph Based Clustering in High Dimensional Data Sets: A Summary of Results. IEEE Data Engineering Bulletin, 1998, 21(1): 15～22