外汇领域的洗钱侦测系统及关键算法研究*)

陈云开　孙小林　马君华
(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　目前,反洗钱成为世界各政府机构关注的热点领域。本文从技术角度探讨了外汇领域的洗钱侦测系统及其关键算法的实现。首先,描述了我国第一个外汇反洗钱侦测系统的架构;然后提出了一个以语义核心树 SCT(Semantic Core Tree)为基础的增量概念聚类算法。该算法能解决以下问题:1)能处理海量数据集;2)能处理分类和数值型混合的数据集;3)能够清楚地解释聚类结果,使得结果易于理解。该算法已在反洗钱框架下实现并投入使用。

关键词　反洗钱系统,数据挖掘,增量概念聚类,语义核心树,分类属性

Study on Money-Laundering Detection System and a Key Algorithm

CHEN Yun-Kai　SUN Xiao-Lin　MA Jun-Hua
(School of Computer Science and Technology, Huazhong University of Science&Technology, Wuhan 430074)

Abstract　At present, anti-money laundering has become a hot topic of the entire world. From the point of view of technique, this paper discusses a money-laundering detection system for administration of foreign exchange and one of its key algorithms. It first gives the framework of the system, which is the first anti-money laundering system of our nation. And then it proposes an SCT(Semantic Core Tree)-based incremental conceptual clustering algorithm. The algorithm could solve the problems of 1) large volume of data set; 2) mixture of categorical and numerical data; 3) easy understanding of result. The algorithm has been employed in the money laundering detection system.

Keywords　Anti-money laundering system, Data mining, Incremental conceptual clustering, Semantic core tree, Categorical

1　引言

洗钱活动与走私、贩毒、恐怖活动、贪污腐败等刑事犯罪相联系,对国家的政治稳定、社会安定、经济安全及国际政治经济体系的安全构成严重威胁。因此,各国政府及有关国际组织除了对反洗钱进行国家立法和国际立法,开展国际合作外,还利用各种信息技术,建立高效的反洗钱信息系统,对洗钱活动进行有效的分析、监测和跟踪。美国金融犯罪执法网络(FinCEN)建立的基于人工智能的反洗钱系统 FAIS,通过10多年的不断完善,已经建立了300多条洗钱规则,能够通过对上报的金融交易报告进行自动识别,来发现未知的、潜在的可疑金融交易行为。澳大利亚交易报告中心(AUSTRAC)的交易报告分析与查询系统(TRAQ)的交易报告筛选系统(ScreenIT)能根据专家知识和样本数据自动推断某些交易、账户,或公司是否具有洗钱特征。

我国的洗钱侦测电子系统建设刚刚起步,在外汇领域履行反洗钱职能的国家外汇管理局2005年初开始建立我国第一个洗钱侦测系统——外汇领域洗钱侦测系统。该系统通过基础平台完成洗钱侦测业务流程,实现了数据收集、洗钱侦测分析、查询反馈、辅助管理等功能。本文首先描述了洗钱侦测系统的业务流程和系统架构,然后就该框架下的关键算法——一个改进的增量概念聚类算法——进行了详细描述。

聚类算法是一种重要的数据挖掘方法,按照聚类方法的不同大体可分为五类:1)基于划分的方法(K-means, PAM, CLARANS…);2)基于层次的方法(BIRCH, CURE, Chamelone…);3)基于密度的方法(DBSCAN, DENCLUE…);4)基于网格的方法(STING, WaveCluster);5)以及基于模型(COBWEB…)的方法。这些算法的研究使得聚类效率不断提高,并在多个领域得到广泛应用。采用传统的聚类算法,外汇洗钱侦测系统面临的问题是:

(1)交易量巨大;

(2)交易类型为分类和数值混合类型;

(3)聚类结果的难以理解。

针对以上问题,本文提出一个改进的增量概念聚类算法,它以语义核心树 SCT(Semantic Core Tree)为基础,利用它处理分类和数值型混合的数据集,并且能够清楚地解释聚类结果。该算法已在反洗钱框架下实现并投入使用。

2　洗钱侦测的业务流程和系统结构

2.1　洗钱侦测的业务流程

洗钱是公司或者个人,采用非法手段将非法所得合理化。洗钱侦测系统主要通过以下过程来进行洗钱行为侦测(见图1)。

(1)洗钱线索监控:洗钱在客户帐户上一定有一些线索可以反应。洗钱线索监控就是根据反洗钱监管人员指定的线索监控点对客户帐户上的所有异常操作进行监控。找到各种线索。不过洗钱线索具有多变的特点,也就是说根据不同的时期犯罪分子会采用不同的方法来洗钱,因此,线索监控方式也随着改变,所以,要求线索监控有灵活定制的功能。

(2)线索预警:通过洗钱线索监控发现了洗钱线索以后,快速传递给反洗钱监管人员,是线索预警的主要功能。

(3)线索分析:通过洗钱线索监控发现了洗钱线索,但并不是所有的洗钱线索都是真正的洗钱犯罪。怎样区分是否真正洗钱,需要监管人员根据自己的理解和经验进一步来判断。

*)基金资助:国家自然科学基金(60403027)。陈云开　博士研究生,研究方向为数据挖掘;孙小林　博士研究生,研究方向为数据挖掘,语义网。

线索分析功能为监管人员提供了这样的手段来处理这些任务。监管人员可以根据发现的线索，将与线索相关的所有信息(包括详细交易信息、客户信息等)提取出来，进一步进行判断。

(4)立项侦察：根据线索分析得到的资料，如果能够确认洗钱线索的真实性，则可以移交给相关的部门进行立项侦察(也许需要公安部门协助)。

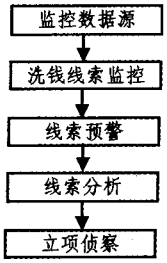

图1　洗钱侦测业务结构

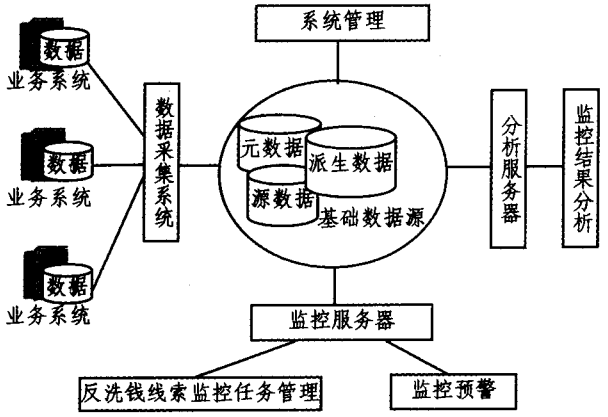

图2　洗钱侦测系统结构

2.2　洗钱侦测系统结构

洗钱侦测系统系统由数据采集子系统、洗钱线索监控子系统、查询分析和管理子系统等构成。其整体结构如图2。

数据采集子系统：从外管局各分局分布于12个业务系统的数据中，采集细粒度的交易数据到基础数据库中。

洗钱线索监控子系统：运用机器学习或数据挖掘算法和模型从交易数据库中侦测可疑交易，并从中发现洗钱交易模式，模式的表达必须容易理解。本文将要描述的改进的增量聚类算法属于该功能模块。

查询分析和管理子系统：为最终用户提供交互界面，以便发送请求及浏览结果。

3　改进的增量概念聚类算法

K-means[5]是划分方法聚类的代表算法，顾名思义，该算法可以基于某种准则将数据划分为 K 个簇，典型的准则有K-平均距离算法以及K-中心点算法。基于划分方法的K-means中心点算法是基于最小化所有对象与其参照点之间的相异度之和的原则来执行的，当存在“噪声”和孤立点数据时，K-中心点算法比K-平均方法更健壮，这是因为中心点不像平均值那么容易被极端数据影响[6]。但是K-中心点算法执行代价比较高，不太适应大型数据库，而且它和K-平均值算法一样需要用户提供 K 值。

Tian Zhang 等人提出的聚类算法 BIRCH[7](Balanced Iterative Reducing and Clustering)通过逐个读入对象构造CF(Clustering Feature)树完成聚类过程，因此是一个增量方法，BIRCH另一个特点是可以通过控制CF树的大小使存储需求符合实际内存大小，适合大型数据库。BIRCH算法只用扫描数据库一次，只要聚类特征树建造完毕，以后的速度会非常快，所以对于海量数据效率很高；但是由于使用汇总信息，因此当簇不是球型的时候，该算法会出现比较大的误差，所以BIRCH很难处理“噪音”或者孤立点数据，而且当数据是分类属性的时候，汇总信息就变得很难表达，BIRCH算法的应用就被限制。

增量概念聚类算法是指在聚类结束后对新增的数据仍然可以对其进行分类并且可以根据新数据的特征调整聚类结果。洗钱侦测系统提供了两个数据集[8]：D_1 代表所有金融电汇数据源，其数据量庞大；D_2 代表案例数据源，其数据量较小且信息完备。基于以上情况，本文提出的增量概念聚类算法结合BIRCH算法和K-means算法构建核心树Core-Tree，并以此为基础进行反洗钱分析。具体构建Core-Tree的方法为：在提供完备案例数据的基础上，使用BIRCH的思想构建CF Tree，并使用K-means中心点算法计算出其叶子节点的语义中心，完成Core-Tree的构造。然后确定条件属性和决策属性，利用规则生成算法求出条件属性类和决策属性类的概念描述并引入确信因子 $\partial(Y)$ 以及包容因子 $k(Y)$ 来对生成的规则进行评价。这样在一定程度上提高了聚类结果的输出质量。接下来我们来讨论在混合类型变量相异度计算标准下增量概念聚类算法的原理和改进。

3.1　混合类型数据的距离计算

作为聚类算法的基础，我们首先讨论一下聚类分析中经常出现的数据类型以及如何计算不同类型数据之间的距离即相异度。

在许多真实的数据库中(比如金融电汇数据)，对象是被混合类型变量描述的，如果对每种类型的变量进行单独的聚类分析，在实际应用中由于其代价太高，是几乎不可能实现的，所以本文采用将所有的变量规格化后一起处理的方法。所谓规格化，即将所有的变量的值域控制在[0，1]这同一范围内，那么变量之间就会具有一定的可比性和可计算性。那么对于对象 i 中的 h 变量，将常用数据类型的相异度(距离)的计算公式分别规格化以后得到以下相异度计算公式：

假设一个包含混合型变量的数据集包括 p 个不同类型的变量，那么对于对象 X_i 和 X_j 有混合类型变量的相异度公式：

$$\mathrm{dis}(X_1,X_2)=\frac{\sum_{h=1}^{p}\delta_{ij}^{(h)}w(h)d_{ij}^{(h)}}{\sum_{h=1}^{p}\delta_{ij}^{(h)}w(h)} \tag{1}$$

其中 $\delta_{ij}^{(h)}$ 是指示值，仅当 X_i 或 X_j 的 h 变量不存在或都为0的时候改值为0，其余的时候为1；$w(h)$ 是变量 h 在所有变量中的权重；$d_{ij}^{(h)}$ 是 X_i 或 X_j 根据 h 变量类型计算的距离或者相异度[3]。

数据开采中的聚类算法在提高效率方面做了大量研究，这些算法都利用空间距离来衡量两个对象的相似性，空间距离较小的对象分到同一个类，而空间距离较大的对象被分到不同的类中。针对洗钱的数据，我们提出以下思想。

3.2　确信因子 $\partial(Y)$ 和包容因子 $k(Y)$

决策属性指的是对知识的产生有决策意义的属性，例如公司的盈亏，客户的消费等等，决策属性概念集表示为 $T(D)$。条件属性就是指可能对这些结果造成影响的属性，例

如产品的销售、客户的种类等等，条件属性概念集表示为 $T(C)$。设 $X_i \in T(C), Y_j \in T(D)$ 分别是条件概念集分块和决策概念集分块，当 $|X_i \cap Y_j| \neq \Phi$，定义 $\mathrm{Des}(X_i)$，$\mathrm{Des}(Y_j)$ 分别是 X_i, Y_j 的描述，则 $\mathrm{Des}(X_i) \Rightarrow \mathrm{Des}(Y_j)$ 是信息系统 IS 上的判定规则。指定允许误差 $\beta(0<\beta\leqslant 0.5)$，引入确信因子 $\partial(Y)$ 以及包容因子 $k(Y)$

$$\partial(Y)=\frac{|X_i \cap Y_j|}{|X_i|} \tag{2}$$

$$k(Y)=\frac{|X_i \cap Y_j|}{|Y_j|} \tag{3}$$

当 $\partial(Y), k(Y) \geqslant \beta$ 时，规则成立。

3.3 判决函数

传统的数据类型一般分数值属性型和分类属性型两种，现实世界中分类属性占的比重更大，比如姓名、地址等，金融领域的数据的特征之一就是分类属性的数据比重大。对于这种数据我们采用混合类型距离公式。

设 A_j 是数据表 T 的第 j 个概念属性的值域的集合，$p(A_j=a_{jk}|l)$ 是第 l 个聚类划分中第 j 个属性取值 a_{jk} 的概率。调整聚类中心的判决函数 J_l 用下式表达：

$$J_l=\{(\max\ p(A_j=a_{jk}|l))|_{j=l}^{m}\}$$

$$(l=1,2,\cdots,p) \tag{4}$$

求得 J_l，便可以确定第 l 个聚类的各属性字段（$k=1,2,\cdots,m$）上出现频繁程度最大的概念值集合构成聚类的中心 Y_l，其中 $l=1,2,\cdots,p$。

3.4 改进的增量概念聚类算法流程

3.4.1 基于中心点思想的 BIRCH 算法

聚类算法 BIRCH(Balanced Iterative Reducing and Clustering)通过逐个读入对象构造 CF(Clustering Feature)树完成聚类过程，因此是一个增量方法，该算法提出类汇总信息的思想[1]，CF 树是一棵高度平衡 B 叉树（B 为每个结点能容纳的最大孩子数）。

由于电汇数据库中混合了几个变量类型，因此我们拟用一种新的方法来计算中心点汇总信息：找到每个叶子的中心点对象 X_i，上级非叶子节点中包括所属所有叶子中心点的信息，CF={N, means [N]}，其中 N 为该子树所含所有叶子个数（即所含所有叶子中心点的个数），means [N]是一个结构体数组的首地址，means [i]表示第 i 个叶子节点的中心点，并包括节点完整信息，每个叶子节点只有一个中心点，所以 CF={1, means [1]}。

当读入对象时，计算到左右子树所含叶子节点的平均距离来判断聚类路径；当对象达到叶子时，判断该对象到该叶子的语义中心的距离是否大于阈值来决定是否聚为一类。

3.4.2 语义 CF 树的构建

先介绍 T, B 的概念：T 用来判断对象是否属于类的一个阈值，当对象与类的距离小于 T 的时候，对象归并到该类中去；B 是项目数，即为每个结点所能容纳孩子结点的最大个数。

根据 T 与 B 的大小构造一棵高度平衡树，对每一个对象，从根结点开始计算该对象与根节点下的所有节点进行比较，选择距离最小的节点继续以上操作。当对象从跟节点搜索到叶子节点那一层的时候，如果叶子汇总 CF 与对象的相异度小于阈值 T，则将它加入到某个类中，否则增加一个叶子节点。如果项目已满，则需要进行分裂操作。

该算法的算法复杂度为 $O(\log_B(n))$，和所有的增量聚类算法一样，BIRCH 算法也具有一个大多数增量聚类所不能避免的明显缺点：依赖顺序。如果一个算法对数据集的任何排序都能生成相同的分区，它就是不依赖顺序的。

聚类很重要的一个特征就是不依赖顺序，而增量算法对样本的顺序非常敏感。对于不同的顺序会产生不同的分区。接下来我们来讨论一下如何结合 K-means 聚类算法来改进这一缺点，以提高算法质量以及如何将构造好的 CF-Tree 转化为 SCT。

3.4.3 使用改进的 K-means 算法来提高增量聚类的质量

在 BIRCH 算法第一步执行完时，内存中会存有一棵 CF-Tree，为了方便以下工作，我们将 CF-Tree 所有的叶子如图 3 连接起来：

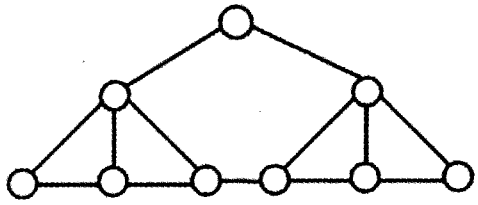

图 3 语义核心树 SCT

则每个叶子都有左右邻居（头叶子和尾叶子除外）。

由于增量算法第一次构造树生成的叶子节点有可能是依赖顺序的，因此算法第二步就是运用对所有叶子节点进行适当调整，使其达到高质量得聚类结果。

K-centers 算法分别以随机选取的 k 个对象作为中心点，计算其它对象与各个中心点的距离，选取距离最小的中心点聚为一类。然后反复迭代使用非中心对象来替代中心点对象（迭代重定位）来提高聚类质量[2]。K-centers 算法不管存不存在孤立点都是比较健壮的，但是由于需要用户给出类的个数 K 并且由于需要不停的测量各个对象之间的相异度，该算法的代价还是比较高的。如何打破这些限制并降低其运行代价？

我们从三个方面入手：

第一：我们的数据是经过一次增量聚类后的结果，这些数据已经被大概分成了几个类，所以 K 值是不用用户给出的，K=CF Tree 叶子数目。

第二：中心点的选取不再是随机的，而是根据属性出现的频率，利用判决函数计算其语义中心，这样迭代的次数就会减少。

第三：新的对象根据 CF Tree 寻找到一个叶子节点，如果该对象到该叶子类的语义中心以及到该叶子左右邻居叶子节点的语义中心的距离小于阈值，则并入该类或其左右邻居，然后根据以上提到的替换规则判断语义中心是否变化；反之，如果距离大于阈值，则将其归于异常点并运用孤立点分析。此时使用的替换规则不需要计算整个数据集的方差变化，由于左右邻居具有与到达叶子最大的相似性，因此其他的叶子都不需要做改动，这样就会大大减少代价，只需计算该叶子以及其左右邻居的代价变化即可。

该步骤的复杂度为 $O(Mn)$，当 M 取较小值的时候其代价较小。当 CF-Tree 树调整完毕后，对所有的叶子节点我们只保留该叶子的语义中心数据，使该 CF Tree 成为 Core-Tree，很明显，Core-Tree 作为语义中心包括该叶子类的代表信息，使其所占内存空间大大减少。

3.4.4 算法描述

算法输入：数据表 D，相似度阈值 T

算法输出：表达 h 个聚类中心的列表 Lst[h]

算法：

[1] for ($i<=n$)计算 x_i 与其它 $j(j=n-i)$个对象的 similar 值，并写入 CF 树 $M=\{u_{ij}=\text{similar}(x_i,x_j)\}$；
[2] $h=1$；
[3] for(M 的第 i 个叶子){
[4] if($u_{ij}>=T$){cluster(h)=cluster(h)$\cup u_{ij}$}
[5] 计算类的语义中心 u_i
[6] while($D\neq\Phi$){
[7] if(similar(x_i,u_i)最大){x_i(cluster)=k; cluster(k)=cluster(k)$\cup x_i$}
[8] 按照判决函数调整语义中心；}
[9] repeat；

3.4.5 引入确信因子的规则生成算法

将源数据表的属性分成条件集 $T^*(C)$和决策集 $T^*(D)$并按照上文提到的聚类算法分别进行聚类得到条件概念集 $T(C)$和决策概念集 $T(D)$。设条件概念集有 k 个类 $T(C)=\{X_1,X_2,\cdots,X_k\}$，决策概念集有 h 个类 $T(D)=\{Y_1,Y_2,\cdots,Y_h\}$，算法用决策等价类逐个测试对条件等价类进行包含，依据给定的 β 误差判断规则是否成立，当满足相应的可信度和支持度阈值时，产生的规则有效。

4 实验及结果

以下本文通过研究洗钱案例的电汇数据来分析洗钱行为的规律的实例来简要说明本文提出的算法。由于篇幅限制，经简化(概念化)后的一部分电汇记录表如表 1。

表中客户类型={散户，中户，大户}；电汇金额={无，少，较少，一般，较多，多，很多，特别多，极多}；交易频度={很少，少，不频繁，较频繁，频繁，很频繁，极频繁}，结算方式={电汇，票汇，托收，信用证，其他}，交易类型={结售汇，跨境，境内}。我们选择客户类型，结算方式和交易类型作为条件概念集，电汇金额和交易频度作为决策概念集，取相似度阈值 2 和 1，有：

$T(C)=\{C_1,C_2,C_3,C_4,C_5,C_6\}=\{\{x_1,x_2,x_{12}\},\{x_3,x_4,x_{13},x_{16},x_{17}\},\{x_7,x_9,x_{11},x_{15},x_{19}\},\{x_8,x_{10},x_{18}\},\{x_5,x_6,x_{14},x_{20}\}\}$

$T(D)=\{D_1,D_2,D_3,D_4,D_5,D_6,D_7\}=\{\{x_1,x_2,x_{12}\},\{x_3,x_7,x_{13},x_{16}\},\{x_4,x_6,x_{14}\},\{x_8,x_{18},\{x_9,x_{10},x_{19}\},\{x_{11},x_{15}\},\{x_5,x_{17},x_{20}\}\}$

有关条件属性类的描述有：$Des(C_1)_{1,2,12}$ =[客户类型=大户][结算方式=信用证][交易类型=跨境]，$Des(C_2)_{3,4,13,16,17}$ =[客户类型≥中户][结算方式=电汇][交易类型=境内]等。

有关决策属性类的描述有：$Des(D_1)_{1,2,12}$ =[电汇金额=很多][交易频度≥较频繁]，

$Des(D_2)_{3,7,11,15,17}$=[电汇金额≥较多][交易频度=频繁]等。

设 $\beta=0.5$，分析确信因子和包容因子得出的规则有：$Des(C_1)_{1,2,12}\Rightarrow Des(D_1)_{1,2,12}(\partial=1,k=1)$，$Des(C_4)_{8,10,18}\Rightarrow Des(D_4)_{8,18}(\partial=0.67,k=1)$

$Des(C_1)_{1,2,12}\Rightarrow Des(D_1)_{1,2,12}$

$(\partial=1,k=1)$，$Des(C_4)_{8,10,18}\Rightarrow Des(D_4)_{8,18}(\partial=0.67,k=1)$

以上描述说明了洗钱行为的特征，例如规则一，可以表示为：一旦客户类型是大户，且其交易类型是跨境交易，结算方式是使用信用证，则其电汇金额很大，交易频度频繁。如果对洗钱的案例数据可以作如上分析的话，就比较容易发现洗钱行为的特征，从而提高洗钱侦测的效率。

结论 在我国，反洗钱相关法规的制订刚刚完成，如何将信息技术应用到该领域还处于探索阶段。本文探讨了我国第一个反洗钱系统，该系统用于外汇领域的洗钱侦测。论文还给出了系统中的一个关键算法的描述，针对外汇交易数据的大容量特征，该算法设计为增量式；同时它还解决了传统聚类方法应用于该领域遇到的问题：无法处理混合型数据以及聚类结果难以理解。不过，在利用确信因子和包容因子解决结果易理解问题时，选择条件属性和决策属性必须具有一定的专家知识来指定特定的属性项。

表 1 洗钱案例电汇记录表(部分)

编号	客户类型	电汇金额	交易频度	结算方式	交易类型
1	大户	很多	频繁	信用证	跨境
2	大户	很多	较频繁	信用证	跨境
3	中户	较多	频繁	电汇	境内
4	中户	一般	极频繁	电汇	跨境
5	散户	较少	很频繁	票汇	境内
6	散户	一般	频繁	电汇	跨境
7	中户	较多	频繁	托收	结售汇
8	大户	极多	少	托收	跨境
9	大户	特别多	不频繁	托收	结售汇
10	大户	特别多	少	托收	跨境
11	中户	多	频繁	托收	结售汇
12	大户	很多	较频繁	信用证	跨境
13	中户	较多	频繁	票汇	境内
14	散户	一般	极频繁	电汇	跨境
15	中户	多	频繁	托收	结售汇
16	中户	较多	频繁	电汇	境内
17	大户	多	频繁	电汇	境内
18	大户	极多	很少	托收	跨境
19	大户	特别多	不频繁	托收	结售汇
20	散户	少	很频繁	电汇	跨境

参考文献

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

2 Kaufman L, Rousseeuw P J. Finding Groups in Data - An Introduction on Cluster Analysis// Wiley Series in Probability and Mathematical Statistics [C], New York: John Wiley & Sons Inc, 1990. 32～71

3 Han Hui, Zha Hongyuan. Name Disambiguation in Author Citations Using a K-way Spectral Clustering Method//JCDL' 2005 [C]. New York: IEEE Press, 2005. 334～343

4 Liu Fang, Lu Zhengding, Lu Songfeng. Mining association rules using clustering [J]. Intelligent Data Analysis, 2001, 5(4): 309～326

5 Wang W, Yang J, Muntz R. STING: A Statistical Informaition Grid Approach to Spatial Data Mining//Very Large Data Bases (VLDB'97)[C]. New York: IEEE Press, 1997. 186～195

6 Mohammed J Z, Markus P. CLICKS: Mining Subspace Clusters in Categorical Data via K-partite Maximal Cliques// ICDE' 2005 [C]. New York: IEEE Press, 2005. 355～356

7 Zhang Tian, Ramakrishnan R. BIRCH: A New Data Clustering Algorithm and Its Applications [J]. Data Mining and Knowledge Discovery, 1997, 10(1): 141～182

8 Chen Hsinchun, Chung Wingyan, Jennifer J, et al. Crime Data Mining: A General Framework and Some Examples [J]. IEEE Computer, 2004, 37(4): 50～56