蚁群聚类组合算法在证券行业客户细分中的应用

Ant Clustering Combination Algorithm Application in Security

邵良杉　王 鹤　（辽宁工程技术大学系统工程研究所 123000）

摘　要： 本文提出了基于信息素的蚁群聚类组合算法，将此种算法应用于证券行业中客户的细分。这种方法既避免了人为事前设定簇的数目，又改善了传统算法中易于陷入局部最优的缺陷。

关键词： 数据挖掘 聚类分析 蚁群算法 信息素 证券行业

1 引言

数据挖掘[1]技术近些年成为人们研究的热点，作为数据挖掘技术之一的聚类分析也越来越受到研究者的关注。随着蚁群算法研究的兴起，人们发现在某些方面采用蚁群模型进行聚类更加接近实际的聚类问题。蚁群聚类算法最大的特定是：不需要设定最终产生的簇的数目，簇的中心是动态变化且可以发现任意形状的簇。

本文提出的蚁群聚类组合算法是对传统算法的改进，避免了 LF 算法中蚂蚁随机的移动，并且利用了蚁群分布式搜索的特性，改善了传统的 K－means 算法易于陷入局部最优的缺陷[2]，并将此聚类组合算法应用于证券公司的客户关系管理中[3]，通过对大量的客户信息进行深层次的挖掘和综合[4]，对不同价值类型的客户进行区分，对客户的信用风险加以预测，使证券公司更好地为客户提供产品组合。

2 蚁群聚类组合算法在客户细分中的应用

2.1 蚁群聚类组合算法原理

蚁群聚类组合算法首先应用 HCBP 算法（Hierarchical Clustering Based on Pheromone）进行初步聚类。这种算法是在 LF 算法[5]基础上的改进。主要思想是尽可能模仿蚂蚁的真实行为，将蚂蚁的空间转换与周围的环境（信息素）紧密地联系在一起，避免蚂蚁随机的移动。将待测对象随机的分布在一个环境中，令空载蚂蚁个体在环境中移动，在运动过程中如果遇到数据对象，则测量当前对象在局部环境的局部相似度，并通过概率转换函数把这个局部相似度转换成拾起或放下对象的概率，以这个概率和标准概率比较，考虑是否拾起该对象，同时逐渐调整局部相似系数，如果是负载的蚂蚁在移动中遇到一个空格，要测量该位置周围的对象和本身携带的对象之间的相似程度，然后判断是否放下该对象。像这样经过大量个体的相互作用，采用简单的递归算法在环境空间中得到聚类结果。

然后在此基础上进一步运用 PCBP 算法（Partioning Clustering Based on Pheromone）。这种算法是在 k－means 基础上的改进。将蚂蚁从食物源 i 到食物源 j 的转移概率引入到 K－means 算法中，数据对象的归属根据转移概率的大小来决定。在下一轮循环中，引入聚类偏差的衡量标准，更新聚类中心，计算偏差，再次判断，直到偏差没有变化或在一定误差范围内，算法结束。

2.2 证券行业客户数据的采集与处理

根据证券公司的要求确定相关的知识源 。例如客户的资产（资金额＋股票市值），佣金贡献，现金（支票）存取频率及差额盈亏情况和交易操作频率。

实例的数据采集对象为阜新市某营业部柜面系统 2004 年的历史交易。数据来源于数据库中的多个表（客户基本资料，年初客户资金情况，年初客户股票库存明细，年末客户资金情况，年末客户股票库存明细，2004 年客户资金变动明细，2004 年客户股票交易明细，2004 年客户股票交易所对应的清算费用明细，

2004 年客户委托交易明细等)。

数据集成就是将来自多文件或多数据库运行环境中异构数据进行合并处理,解决语义的模型性。实例中经过数据集成得到所有客户的年初资产(资金额 + 股票市值),佣金贡献,现金(支票)存取频率及差额,盈亏情况和交易操作频率等内容的数据库。

建立一张表便于以后挖掘与统计,表名 khxfzl(客户细分资料)包含十个字段,是一个客户一条记录。

实例中经过数据集成得到含有八个属性(zczz1 期初资产总值,jycs 交易次数,jygx 佣金贡献,cqcs 存取次数,cqce 存取差额,ndyk 年度盈亏,khzl 客户种类,jyfs 交易主要方式)的数据源表客户细分资料 khxfzl 。

将参加聚类的六个属性进行数据变换,即 zczz1 期初资产总值,jycs 交易次数,jygx 佣金贡献,cqcs 存取次数,cqce 存取差额,ndyk 年度盈亏。为表达方便,假设表中共有 n 条记录。例表示第 i 条记录的第三个属性。

对每条记录的每个属性进行标准差标准化变换:

$$X_{ij}=\frac{X_{ij}-X_j}{S_j} \quad i=1,2\cdots\cdots,n \quad j=1,2,3,4,5,6 \tag{1}$$

其中:

$$X_j=1/n\sum X_{ij},\ S_j=[1/(n-1)\sum(X_{ij}-X_j)^2]^{1/2} \tag{2}$$

经过变换后各属性的均值为 0,标准差均为 1。

2.3 应用组合聚类算法进行数据聚类

对表 khxfzl 先执行 10 次 GCBP 算法,每次的初始代表对象为随机产生,得到 50 聚类代表对象。然后再对所产生的 50 个聚类代表对象执行 PCBP 算法,由此产生 5 个聚类中心。

具体步骤如下:

(1) 初始化。

把 9800 个客户资料作为数据对象随机的分布在 200×200 的二维网格上。用 400 只蚂蚁在网格上独立的移动并完成聚类。任意时刻蚂蚁都处在某个网格单元中,蚂蚁的速度范围 η 为[2~10]。整个过程采用 3×3 的网格区域作为局部面积,该区域上数据对象的多少决定该区域的密度,并隐含了对象的相似度。每次实验分为外循环 Cycles 和内循环 CycleSize,所以每次迭代次数 M 为(Cycles × CycleSize)。在实验中 Cycles = 100, CycleSize = 100。

蚂蚁初始化为空载状态,且随机的放置在网格上。

(2) 对每只蚂蚁,若空载且位置 r 有对象 o_i,则计算 $f(o_i)$ 和 $p_p(o_i)$ 在[0,1]之间随机取一个实数 R 来判断蚂蚁是否捡起对象;如果蚂蚁负载且位置 r 为空,则计算 $f(o_i)$ 和 $p_p(o_i)$ 并根据 R 判断蚂蚁是否放下对象。

计算对象间距离:

$$d(o_i,o_l)=\sqrt{\sum_{k=1}^{m}(o_{ik}-o_{lk})^2} \tag{3}$$

计算局部相似度:

$$f(o_i)=\begin{cases}\dfrac{1}{S^2}\displaystyle\sum_{l\in Neight_{s\times s}}\left[1-\frac{d(o_i,o_l)}{\alpha}\right] & \text{if } f(o_i)>0\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

负载的移动蚂蚁拾起数据对象的概率 $p_p(o_i)$:

$$p_p(o_i)=\begin{cases}1-\varepsilon & \text{if } f(o_i)<\tau\\ \varepsilon & \text{otherwise}\end{cases} \tag{5}$$

一个负载的移动蚂蚁放下数据对象的概率 $p_d(o_i)$:

$$p_d(o_i)=\begin{cases}1-\varepsilon & \text{if } f(o_i)<\tau\\ \varepsilon & \text{otherwise}\end{cases} \tag{6}$$

本实验中以上各式参数的取值如下:噪音常量 ε: 0.001;阈值恒量 z: 0.3;发射信息素量 τ:0.2;信息素挥发率 x: 0.01;属性个数 m:6;局部面积: 3×3。

(3) 计算 $W(\tau_l)$ 和 p_{lk}。

计算信息素相应函数:其中:发射信息素量 τ:0.2,信息素挥发率 k:0.01。

$$W(\tau)=(1+\frac{\tau}{1+\delta\tau})^{\beta} \tag{7}$$

$$p_{lk}=\frac{W(\tau_l)w(\Delta_l)}{\sum_{l/k}W(\tau_l)w(\Delta_l)} \tag{8}$$

(4) 蒸发所有网格的信息素,输出对象的位置。得到初始聚类个数 num 为 50,计算聚类中心 $\bar{c}_l$。

(5) 输入聚类个数 num = 50, 聚类中心 $\bar{c}_l$。

(6) 取不同与 $\bar{c}_l$ 且未被标识过的 X_l,只要 $p_{ll}<p_o$,就计算 $p_{ll}(t)$。令:k = 3; R = 0.35; p_o = 0.4;

$$p_{ll}(t)=\frac{\tau_{ll}^{\alpha}(t)\eta_{ll}^{\beta}(t)}{\sum_{s\in S}\tau_{sl}^{\alpha}(t)\eta_{sl}^{\beta}(t)} \tag{9}$$

$$\bar{c}_l=\frac{1}{l}\sum_{\Omega}^{l}X_l \tag{10}$$

(7) 标识 X_i，并归到 $\bar{c}_i$

(8) 计算第 j 个聚类的偏离误差 ζ_j 和所有聚类的总偏离误差 ζ，其中：$\zeta_0 = 0.6$

$$\zeta_j = \frac{1}{j}\sum_{i=1}^{j}(X_i - \bar{c}_j) \quad (11)$$

$$\zeta = \sum_{j=1}^{k}\zeta_j \quad (12)$$

(9) 更新聚类中心 $\bar{c}_j$，更新信息素 τ_{ij}。

(10) 得到 5 个聚类结果。

2.4 实验结果分析

实例中将聚类结果根据属性的特征进行统计，并显示每类客户的数量，占总客户数的比例，佣金贡献，占总佣金贡献比例，资产总值范围，平均资产总值，占总资产总值比例交易次数范围，平均交易次数，平均存取次数，平均存取差额，年度盈亏范围，平均年度盈亏。该实例经过多次挖掘后根据要求应分为五类客户每类具体属性值如表 1。

表 1 聚类后每类属性的统计值(资金单位：元)

类别 sort	数量 number	资产总值(平均) asset (average)	佣金贡献(汇总) commission(total)
1	380	248211.23	479060.45
2	213	1902413.80	1763316.61
3	69	2634166.22	1824489.80
4	8213	12990.66	1258340.56
5	892	1059.34	3300.98

类别 sort	年度盈亏(平均) number	次数(平均) time (average)	存取差额 access balance	存取次数(平均) access time
1	-1.42	29.2	9499.32	4.1
2	-5.2%	679.9	-1198.99	24.9
3	+6.89%	788.9	3089.20	39.2
4	-4.01	89.1	599.98	11.3
5	+0.3%	0.21	-6.01	0.04

解读数据并提供模拟策略：

(1) 客户行为：盈亏不大，现金存取频率低，总存款额大于总取款额，交易操作次数少。客户特征：收入稳定，投资渠道少，对市场不敏感，利润要求不高，是潜在优质客户。服务对策：可以推荐一些信托产品或代客理财，改变其投资理念，将其发展为优质客户。

(2) 客户行为：亏损较大，交易操作频繁，有一定的资金实力。客户特征：专业知识不强，对市场敏感，迫切解套。是公司利润的贡献者，但有潜在流失的危险。服务对策：提供咨询服务，亦可介绍专业力量强的咨询公司提供各种服务。

(3) 客户行为：有较高的盈利能力，交易操作频繁，有较大的资金量。客户特征：有专业知识，对市场敏感，是公司利润的重要贡献者，是优质客户。服务对策：应经常与客户沟通，及时发现客户的真正需求，提供相应的政策照顾，必须提高客户的忠诚度和满意度。

(4) 客户行为：大多数为中小散户，有一定量的买卖交易和资金存取。客户特征：无专业知识，资金量偏小，是公司的主要客户群。服务对策：提供大众性咨询服务，使其行为向对公司有利的方向转变。

(5) 客户行为：基本不进行股票交易也不存取资金客户特征属于睡眠客户。服务对策：应减少这类客户的比例。

3 结论

实验证明，应用本文提出的蚁群聚类组合算法在对大量数据进行聚类时可以得到一个比较好的聚类结果。本算法还可以应用于与此相近的大型数据库的数据聚类。

参考文献

1 尹松、周永权、李陶深，数据聚类方法的研究与分析[J]，航空计算机，，2005，35(1)：63－66.

2 Dorigo M, Maniezzo V. 1996. Ant system: optimization by a colony of cooperating agents. IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics, Feg., 26(1):29－41.

3 Refer 企业核心竞争力的 Web 挖掘研究，计算机系统应用，2005. N(8)，91－94.

4 邵峰晶、于忠清，数据挖掘原理与算法，中国水利水电出版社，2003.

5 李瑞，蚁群聚类算法及其在推荐系统中的应用：[硕士学位论文]. 重庆：西南师范大学 计算机与信息科学学院，2005.

6 邵良杉、那宝贵，基于 Web 挖掘的虚拟企业合作伙伴选择决策支持系统研究：计算机系统应用，2006. N(10)，2－5.