第 36 卷 第 3 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.3 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)03—0044—03 文献标识码：A 中图分类号：TP311.12

基于变精度粗糙集的 Web 用户聚类方法

纪洲鹏，周 军，何 明
(辽宁工业大学电子与信息工程学院，锦州 121000)

摘 要：针对 Web 使用挖掘中的用户聚类问题，提出一种基于变精度粗糙集理论的粗糙聚类方法，该方法放宽经典粗糙集中不可区分关系的传递性将其扩展为相容关系，使用变精度粗糙集的相对错误分类率 β 来形成新的相似 β 上近似，从而将一个用户划分到多个聚类，该方法不需要区分用户会话，降低了数据预处理的难度，通过理论推导和实例证明了其有效性。

关键词：粗糙聚类；变精度粗糙聚类；相似 β 上近似；Web 使用挖掘

Web User Clustering Method Based on Variable Precision Rough Set

JI Zhou-peng, ZHOU Jun, HE Ming
(College of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121000)

【Abstract】Focus on solving the user clustering issues, a rough clustering method based on the variable precision rough set theory is proposed. The indiscernibility relation in classical rough set is extended to a tolerance relation with the transitivity property being relaxed. The new proposed similarity β upper approximations are formed using the relative degree of misclassification β, so a user can be assigned to more than one cluster, and this approach does not need to identify the users' sessions, therefore, the complexity of data preprocessing decreases. Experimental example shows the effectiveness of the proposed algorithm.

【Key words】rough clustering; variable precision rough clustering; similarity β upper approximation; Web usage mining

1 概述

聚类是数据分析中最基本的步骤，分为硬计算方法和软计算方法。在硬聚类方法中，根据相似度将对象划分到不同的聚类中，不同的聚类之间没有交集。在软聚类方法中，一个对象可以被分配到 2 个或者 2 个以上的聚类中。软聚类可能具有模糊或粗糙的边界域。在模糊聚类方法中，通过隶属函数刻画对象的特征，需要专门的领域知识来定义模糊隶属函数并且需要先验统计信息。粗糙聚类无需提供数据集合之外的任何先验信，处理结果比较客观。

近些年来，粗糙集理论已被广泛地应用于聚类。文献[1]提出一种利用粗糙近似的概念对登录日志中的 Web 事务进行聚类的算法。文献[2]提出一种基于受限的上近似，在上近似的迭代过程中引入相对相似度作为聚类合并标准的对序列化数据进行聚类的方法。但已提出的方法大多需要区分用户会话，粗糙聚类所基于的模型大多是经典粗糙集模型。

本文使用兴趣度将不可区分关系扩展为相容关系，引入相对错误分类率 β 作为聚类的合并条件形成相似 β 上近似，提出一种基于变精度粗糙集模型的 Web 用户聚类的新方法。

2 Web用户的变精度粗糙聚类算法

数据预处理包括数据清洗、用户识别、会话识别、路径补全和事务识别，由于代理服务器和客户端缓存的使用，给用户会话的识别带来极大的困难，更不用说事务识别带来的额外的算法开销[3-4]。本文提出的方法避免了用户会话的识别，实现了用户的有效聚类。

2.1 登录日志预处理

服务器登录日志是 Web 挖掘最为丰富的数据来源，其格式一般分为通用日志格式和扩展日志格式，对登录日志进行预处理，可以统计得到用户浏览行为的原始信息，包括：用户浏览的网页，点击网页的次数，浏览网页的时间，网页的内容长度等，进而得到用户原始的登录信息矩阵：

$$\boldsymbol{AccessR}_{m\times n}=\begin{pmatrix} Infor(1,1) & \cdots & Infor(1,n) \\ \vdots & \ddots & \vdots \\ Infor(m,1) & \cdots & Infor(m,n) \end{pmatrix}$$

其中，矩阵的行数 m 代表用户数；矩阵的列数 n 代表网页数；$Infor(i,j)$ 表示第 i 个用户 $user_i$ 浏览第 j 个网页 $page_j$ 时的统计信息。

2.2 用户兴趣矩阵的建立

聚类的目的是为了将兴趣相似的用户分到一个聚类中，本文使用点击率和相对浏览时间来定义用户的兴趣度[5]。既考虑了用户的点击率、浏览时间，又考虑了网页内容的长度。

定义 1 点击率：

$$Hits(user,page)=\frac{NumberOfVisits(page)}{\Sigma_{page\in VisitedPage}(NumberOfVisits(page))}$$

定义 2 相对浏览时间：

$$Duration(user,page)=\frac{TotalDuration(page)/Length(page)}{\max_{Page\in VisitedPage}(TotalDuration(page))/Length(page)}$$

鉴于 HTTP 协议的无状态性，用户最后访问网页的浏览时间无法断定，以 30 min 为限，使用所有用户的平均浏览时

基金项目：国家自然科学基金资助项目(60674056)；辽宁省教育厅基金资助项目(20031066)
作者简介：纪洲鹏(1982－)，男，硕士研究生，主研方向：数据挖掘，粗糙集理论；周 军，教授、博士；何 明，硕士研究生
收稿日期：2009-09-16 **E-mail：**horizon_jzp@hotmail.com

间对定义 2 进行适当调整：

$$Duration(user,page)=\frac{TotalDuration(page)/Length(page)}{Average_{Page\in VisitedPage}(TotalDuration(page))/Length(page)}$$

定义 3 兴趣度：

$$Interest(user,page)=\partial Hits(user,page)+\beta Duration(user,page)$$

其中，$\alpha+\beta=1$ ($\alpha>0,\beta>0$)，α,β 各取 0.5。

定义 4 用户兴趣矩阵：

兴趣矩阵用来描述网页与浏览这些网页的用户之间的关系。矩阵的行数 m 代表用户数，矩阵的列数 n 代表网页数：

$$\boldsymbol{InterestR}_{m\times n}=\begin{pmatrix} Interest(1,1) & \cdots & Interest(1,n) \\ \vdots & \ddots & \vdots \\ Interest(m,1) & \cdots & Interest(m,n) \end{pmatrix}$$

其中，$Interest(i,j)$ 代表在一段时间内 $user_i$ 对第 j 个网页 $page_j$ 的兴趣度。

假定 U 表示一段时间内访问网站的用户集，记作 $U=\{u_1,u_2,\cdots,u_i,\cdots,u_m\}$，其中，$1\leqslant i\leqslant m$，$m$ 是用户的数目。URL 表示网站的网页的集合，记作 $URL=\{url_1,url_2,\cdots,url_i,\cdots,url_n\}$，其中，$1\leqslant i\leqslant n$，$n$ 是网页的数目。

定义 5 对每一个用户 $u_j\in U(1\leqslant j\leqslant m)$，使用用户对每个网页 $url\in URL$ 的兴趣度信息来描述用户的浏览行为，用户的模糊子集 μ_{u_j} 反映了第 j 个用户的浏览行为，记作：

$$\mu_{u_j}=\{(url_i,f_{\mu_{u_j}}(url_i))\,|\,url_i\in URL\wedge 1\leqslant i\leqslant n\}$$

其中，$f_{\mu_{u_j}}(url_i)$，$f_{\mu_{u_j}}(url_i)\to[0,1]$，是隶属函数，并且 $f_{\mu_{u_j}}(url_i)=Interest(u_j,url_i)$。

定义 6 模糊相似度：

设 X 是一个模糊子集，$X=\{x_1,x_2,\cdots,x_i,\cdots,x_m\}(1\leqslant i\leqslant m)$，每个模糊子集 $x_i\in X$ 又可以由一个模糊子集元素描述，记作 $x_i=\{x_{i,1},x_{i,2},\cdots,x_{i,n}\}$。2 个模糊子集 x_i 和 x_j 的模糊相似度采用经典的最大最小法求得：

$$sim_f(x_i,x_j)=\frac{\sum_{k=1}^{n}\min(x_{ik},x_{jk})}{\sum_{k=1}^{n}\max(x_{ik},x_{jk})}$$

其中，$sim_f(x_i,x_j)\in[0,1]$。在实际生活中，2 个用户对网站中感兴趣的网页或多或少存在差异。

2.3 模糊相似矩阵

模糊相似矩阵如下：

$$\boldsymbol{SIM}^{f}_{m\times m}=\begin{pmatrix} 1 & sim_f(x_2,x_1) & \cdots & sim_f(x_m,x_1) \\ sim_f(x_1,x_2) & 1 & \cdots & sim_f(x_m,x_2) \\ \cdots & \cdots & \cdots & \cdots \\ sim_f(x_1,x_m) & sim_f(x_2,x_m) & \cdots & 1 \end{pmatrix}$$

易知，此矩阵满足自反性、对称性。

2.4 经典粗糙集模型的扩展

定义 7 给定任意一个非负的阈值 $\delta\in(0,1]$ 和论域 U 的一个子集 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，论域 U 中任何一个对象 x_i 的第一个上近似定义如下：

$$\overline{R}(x_i)=\{x_j\,|\,sim_f(x_i,x_j)\geqslant\delta\}$$

相容关系 R 为：xRy iff $sim_f(x_i,x_j)\geqslant\delta$，满足自反性、对称性。$U/R=\{\overline{R}(x_1),\overline{R}(x_2),\cdots,\overline{R}(x_i),\cdots,\overline{R}(x_m)\}$，是论域 U 的一个覆盖。

本文采用凝聚的算法对 Web 用户进行粗糙聚类。开始将每个用户看成一个聚类，将第 i 个聚类记为 $C_i=\{x_i\}$。显然，C_i 是 U 的一个子集，C_i 的上近似 $\overline{R}(C_i)$ 是与 x_i 相似的用户的集合，即一个对 x_i 浏览的页面感兴趣的用户可能还会对属于 $\overline{R}(C_i)$ 中的用户浏览的页面感兴趣。

由第一个上近似产生的聚类之间可能共有某些元素(即这些聚类之间存在粗糙边界域)，由此可以通过相对错误分类率计算不同聚类中共有元素的强度来指导聚类的迭代进程。

2.5 相对错误分类率

定义 8 设 X 和 Y 表示有限论域 U 的非空子集。令

$$c(\overline{R}(x_i),\overline{R}(x_j))=1-\frac{|\overline{R}(x_i)\cap\overline{R}(x_j)|}{|\overline{R}(x_i)|},\quad |\overline{R}(x_i)|>0$$

其中，$|\overline{R}(x_i)|$ 表示集合 $\overline{R}(x_i)$ 的基数；$c(\overline{R}(x_i),\overline{R}(x_j))$ 表示集合 $\overline{R}(x_i)$ 对于集合 $\overline{R}(x_j)$ 的相对错误分类率。$\overline{R}(x_i)$ 至少包含其自身，故 $|\overline{R}(x_i)|$ 大于 0。

令 $0\leqslant\beta<0.5$，若 $c(\overline{R}(x_i),\overline{R}(x_j))\leqslant\beta$，则称 $\overline{R}(x_j)$ 以误差 β 多数包含 $\overline{R}(x_i)$，记为 $\overline{R}(x_j)\overset{\beta}{\supseteq}\overline{R}(x_i)$。多数包含意味着 $\overline{R}(x_i)$ 与 $\overline{R}(x_j)$ 的公共元素的数目大于 $\overline{R}(x_i)$ 中元素数目的 50%。

如图 1 所示，易知，随着分类误差 β 的增大，X 的正域与负域将扩大，边界域将缩小，而 X 的上近似域将变小。反之，随着分类误差 β 的减小，X 的正域与负域将缩小，边界域将扩大，而 X 的上近似域将变大。

图 1 变精度粗糙集模型的 β 变化情况

调节分类误差 β 的大小，可以将对象从下近似移到上近似中，即减小下近似域，增大上近似域，意味着一个物体被划分到多个聚类中的机会大大增加。

2.6 相似β上近似

定义 9 设 $A=(U,R)$ 为近似空间，对于 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_m\}(1\leqslant i\leqslant m)$，$X\subseteq U$，$U/R=\{\overline{R}(x_1),\overline{R}(x_2),\cdots,\overline{R}(x_i),\cdots,\overline{R}(x_m)\}$。

x_i 的相似 β 上近似：

$$\overline{RR}_\beta(x_i)=\bigcup\{x\in U\,|\,x_j\in U,c(\overline{R}_\beta(x_j),\overline{R}_\beta(x_i))<1-\beta\}$$

$$\overline{RRR}_\beta(x_i)=\bigcup\{x\in U\,|\,x_j\in U,c(\overline{R}_\beta(x_j),\overline{RR}_\beta(x_i))<1-\beta\}$$

论域 U 中的对象 x_j 只有在满足条件：$\overline{R}_\beta(x_j)$ 与 $\overline{R}_\beta(x_i)$ 的相对错误率小于 $1-\beta$ 时才能被合并到下一个相似 β 上近似 $\overline{RR}_\beta(x_i)$ 中。

初次迭代的时候 $\overline{R}_\beta(x_i)=\overline{R}(x_i)$，并且去除长度较小的用户，即 $\forall\overline{R}_\beta(x_j)\in U/R$，$\left|\overline{R}_\beta(x_j)\right|\geqslant 2$。重复迭代相似 β 上近似的计算过程，直到前后 2 次迭代的结果相同。其中，相对错误分类率 β 是给定的阈值。

2.7 VPRSCluster算法

VPRSCluster 算法(变精度粗糙聚类算法)如下：

Input

***Access*$R_{m\times n}$**：用户登录信息矩阵

δ：相似度阈值

β：自定义相对错误分类率阈值

Output

CoU：聚类方案

Begin

Step 1：将用户登录信息矩阵转变为用户兴趣矩阵

Step 2：在给定的阈值δ条件下将用户兴趣矩阵转变为用户模糊相似矩阵

Step 3：对于每个用户 $x_i \in U$，利用公式 $\overline{R}(x_i)=\{x_j \mid sim_i(x_i,x_j) \geqslant \delta\}$ 得到每个用户的第一个上近似 $S_i=\overline{R}(x_i)$

Step 4：假定 $US=\{S_1,S_2,\cdots,S_i,\cdots,S_m\}$，$CoU=\varnothing$

Step 5：对于所有的 $S_i \in US$，在给定的阈值 β 条件下计算它们的下一个相似 β 上近似 S_i^{β}

If $S_i^{\beta}=S_i$

$CoU=CoU\cup S_i^{\beta}$

$US=US\backslash\{S_i\}$

Endif

Step 6：重复 Step 5 直到 $US=\varnothing$

Step 7：Return CoU

2.8 时间复杂度

设论域 U 中的用户数目为 N，用户平均浏览路径长度为 L，用户两两之间计算相似度的时间复杂度为 $O(N^2 \mathrm{lb}L)$，计算一次相似 β 上近似的时间复杂度为 $O(N/|R|)$，假设每次上近似迭代中涉及到合并操作的聚类的平均数目为 K，则合并 K 个聚类的时间复杂度为 $k\mathrm{lb}k$，假设迭代次数的最大值为 N/k，迭代过程中所有用户合并相同聚类的时间复杂度为 $O((N/k)k\mathrm{lb}k)$ [2]，算法的总的时间复杂度为

$$O(N^2\mathrm{lb}L)+O(N/|R|)+O((N/k)k\mathrm{lb}k)$$

即：$O(N^2\mathrm{lb}L)+O(N/|R|)+O(N\mathrm{lb}k)$

3 举例

数据源来自 GHS 系统的 IIS 服务器登录日志，采用 w3c 扩展日志文件格式，经数据预处理得到用户的模糊相似矩阵，如下所示：

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
u_1	1	0	0	0.44	0.45	0.48	0	0	0	0
u_2	*	1	0	0.46	0.37	0.37	0.47	0	0.55	0.35
u_3	*	*	1	0	0	0.43	0.54	0.28	0.54	0.41
u_4	*	*	*	1	0.35	0	0.65	0.47	0.26	0.7
u_5	*	*	*	*	1	0.38	0	0	0	0
u_6	*	*	*	*	*	1	0.34	0.42	0	0.37
u_7	*	*	*	*	*	*	1	0.78	0.36	0.82
u_8	*	*	*	*	*	*	*	1	0	0.7
u_9	*	*	*	*	*	*	*	*	1	0.64
u_{10}	*	*	*	*	*	*	*	*	*	1

假定 $\partial=0.45$，论域 U 的覆盖：

$$U/R=\{\overline{R}(x_1),\overline{R}(x_2),\cdots,\overline{R}(x_i),\cdots,\overline{R}(x_m)\}$$

如下所示：

$$\overline{R}(u_1)=\{u_1,u_5,u_6\}$$

$$\overline{R}(u_2)=\{u_2,u_4,u_7,u_9\}$$

$$\overline{R}(u_3)=\{u_3,u_7,u_9\}$$

$$\overline{R}(u_4)=\{u_2,u_4,u_7,u_8,u_{10}\}$$

$$\overline{R}(u_5)=\{u_1,u_5\}$$

$$\overline{R}(u_6)=\{u_1,u_6\}$$

$$\overline{R}(u_7)=\{u_2,u_3,u_4,u_7,u_8,u_{10}\}$$

$$\overline{R}(u_8)=\{u_4,u_7,u_8,u_{10}\}$$

$$\overline{R}(u_9)=\{u_2,u_3,u_9,u_{10}\}$$

$$\overline{R}(u_{10})=\{u_4,u_7,u_8,u_9,u_{10}\}$$

取 $\beta=0.4$，即论域 U 中的对象（$u_j \in U$）只有满足它的等价类（$\overline{R}(u_j)$）与 $\overline{R}(u_i)$ 的公共元素的数目大于等于 $\overline{R}(u_j)$ 中元素数目的 60%，才能被合并到下一个相似 β 上近似 $\overline{RR}_{0.4}(u_i)$ 中。

计算 u_7 的相似 β 上近似：

$$c(\overline{R}(u_9),\overline{R}(u_7))=1-\frac{|\overline{R}(u_9)\cap\overline{R}(u_7)|}{|\overline{R}(u_9)|}=1-\frac{|u_2,u_3,u_{10}|}{|u_2,u_3,u_9,u_{10}|}=0.25<0.4$$

只有 u_9 满足相对错误率的条件，才可以并到 $\overline{RR}_{0.4}(u_7)$ 中，所以：

$$\overline{RR}_{0.4}(u_7)=\{u_2,u_3,u_4,u_7,u_8,u_9,u_{10}\}$$

继续求 u_7 的相似 3 次 β 上近似 $\overline{RRR}_{0.4}(u_7)$，知：$\overline{RRR}_{0.4}(u_7)=\overline{RR}_{0.4}(u_7)$，$u_7$ 的相似 β 上近似为

$$\overline{RRR}_{0.4}(u_7)=\{u_2,u_3,u_4,u_7,u_8,u_9,u_{10}\}$$

由此得到 u_7 的聚类结果：$CoU_7=\{u_2,u_3,u_4,u_7,u_8,u_9,u_{10}\}$。

同理可求论域中所有对象的相似 β 上近似：

$$CoU_1=\{u_1,u_5,u_6\};\quad CoU_2=\{u_2,u_3,u_4,u_7,u_8,u_9,u_{10}\}$$

$$CoU_3=\{u_3,u_7,u_9\};\ CoU_4=\{u_2,u_4,u_7,u_8,u_{10}\}$$

$$CoU_5=\{u_1,u_5\};\ CoU_6=\{u_1,u_6\}$$

$$CoU_7=\{u_2,u_3,u_4,u_7,u_8,u_9,u_{10}\}$$

$$CoU_8=\{u_4,u_7,u_8,u_{10}\};\quad CoU_9=\{u_2,u_3,u_9,u_{10}\}$$

$$CoU_{10}=\{u_2,u_3,u_4,u_7,u_8,u_9,u_{10}\}$$

最后合并相同的聚类，得到在阈值 $\beta=0.4$ 条件下的用户聚类方案：

$$CoU=\{CoU_1,CoU_2,CoU_3,CoU_4,CoU_5,CoU_6,CoU_8,CoU_9\}$$

4 结束语

已提出的粗糙聚类算法大多建立在经典粗糙集模型的基础上，本文使用的聚类算法建立在变精度粗糙集模型的基础之上，通过引入相对分类错误率相对有效的控制上近似的迭代过程，不仅可以将一个对象分类到多个聚类中，而且可以得到较多的不同的聚类。通过调节相对错误分类率 β 的大小，可以得到不同精度的聚类方案，当相对错误分类率 $\beta=0$ 时，变精度粗糙集模型就蜕化成了经典粗糙集模型，故变精度粗糙聚类算法是经典粗糙聚类算法的推广。

下一步工作将完成 Web 页面的粗糙聚类，在用户聚类和页面聚类的基础上挖掘用户的频繁访问模式。

参考文献

[1] De S K, Krishna P R. Clustering Web Transactions Using Rough Approximation[J]. Fuzzy Sets and Systems, 2004, 148(1): 134-138.

[2] Kumar P, Krishna P R, Bapi R S, et al. Rough Clustering of Sequential Data[J]. Data & Knowledge Engineering, 2007, 63(2): 183-199.

[3] 陈子军, 王鑫昱, 李 伟. 一种 Web 日志会话识别的优化方法[J]. 计算机工程, 2007, 33(1): 95-97.

[4] 刘立军, 周 军, 梅红岩. Web 使用挖掘的数据预处理[J]. 计算机科学, 2007, 34(5): 200-204.

[5] Liu Haibin, Keselj V. Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and

Predicting Users' Future Requests[J]. Data & Knowledge Engineering, 2007, 61(2): 304-330.

编辑　任吉慧