

## 基于用户兴趣分类优化的聚类模型仿真

刘建粉, 史永昌

(平顶山学院 计算机科学与技术学院, 河南 平顶山 467000)

**摘 要:** 对用户兴趣进行聚类分析对研究消费心理有着重要的意义. 提出一种考虑用户兴趣分类优化的聚类模型, 采用 ID3 决策树算法提高用户兴趣分类计算速度, 将最高信息增益的属性当成前节点的检测属性, 确保结果分解中的用户兴趣样本分类所需的信息量最小, 构建用户兴趣分类优化的自适应模糊聚类目标函数, 更新聚类原型矩阵, 在自适应模糊聚类模型下, 直接给出聚类原型的迭代等式, 保证分类准确. 实验结果说明, 所提模型相对于传统聚类模型不容易陷入局部最优解, 具有较高的查全率和查准率, 对进一步用户行为研究有着较大的意义.

**关键词:** 用户兴趣; 分类; 优化; 聚类模型; 仿真

中图分类号: F127

文献标识码: A

文章编号: 1000-7180(2014)05-0171-04

## Classification Optimization Clustering Model Simulation Based on User Interest

LIU Jian-fen, SHI Yong-chang

(Department of Computer Science & Technology, Pingdingshan University, Pingdingshan 467000, China)

**Abstract:** Clustering analysis was carried out on the user's interests is of great significance to the study of consumer psychology. Considering user's interests is a kind of classification optimization clustering model, improve the user's interests using the algorithm of ID3 decision tree classification calculation speed, the attribute of the highest information gain as the test attributes of nodes before, to ensure the result of decomposition users interested in samples required minimum amount of information, building user interest classification optimization of adaptive fuzzy clustering objective function, the update matrix clustering prototype, under adaptive fuzzy clustering model, clustering prototype iterative equation is given directly, guarantee the accuracy of the classification. Experiment result shows that the proposed model is compared with traditional clustering model is not easy to fall into local optimal solution, has higher recall ratio and precision, and has great significance for further user behavior research.

**Key words:** user interest; classification; optimization; clustering model; simulation

### 1 引言

当前在海量数据中进行用户归类划分, 类中的数据相似性运算获取决策方案, 差异数据间的相似性可忽略, 进而降低决策方案的运算量, 提高决策方案的实时性具有重要意义<sup>[1-2]</sup>. 该种过程的应用基础是用户间的聚类分析问题. 聚类分析是一种无指导的机器学习方法, 其依据“物以类聚”的原理, 将一组个体依据相似性划分成若干类别, 确保属于相

同种类的个体差异最小化, 不同种类上的个体差异最大化. 不同用户的评分规范不同, 不同规范要求下不同用户对同一商品的评分值存在差异, 而传统聚类方法<sup>[3-5]</sup>大都依据用户对商品评分数值大小进行聚类, 无法有效描述用户对该商品的兴趣度, 会形成不可避免的人为误差, 存在收敛速度慢、精度不高的缺陷<sup>[6-9]</sup>.

文中提出了一种考虑用户兴趣分类优化的聚类模型, 具有较高的查全率和查准率<sup>[10-12]</sup>.

收稿日期: 2013-07-13; 修回日期: 2013-09-07

基金项目: 河南省重点科技攻关计划项目(132102210442)

## 2 基于用户兴趣分类优化的聚类模型

### 2.1 融入用户兴趣分类的 ID3 决策树算法的快速分类优化

本文采用 ID3 决策树算法实现用户兴趣的分类优化,为构建基于用户兴趣分类优化的聚类模型提供可靠的依据. ID3 算法是由 Quinlan 为了从数据中归纳分类模型而构造的算法,也称为判定树. 在树的每个节点上采用信息增益(information gain)度量选取检测属性. 寻求具有最高信息增益的属性当成前节点的检测属性,该属性导致对结果分解中的用户兴趣样本分类所需的信息量最小,同时可描述分解的最小随机性.

设置  $S$  表示  $s$  个用户兴趣度样本的集合,如果类标号属性存在  $m$  个不同值,定义  $m$  个不同类  $G_i (i = 1, \dots, m)$ , 采用属性  $A$  将  $S$  分解成  $v$  个子集  $\{S_1, S_2, \dots, S_v\}$ , 其中  $S_j$  包含  $S$  中这样一些样本,在  $A$  上具有值  $a_j$ . 若将  $A$  当成检测属性,则这些子集对应于由包含集合  $S$  的节点生长出来的分枝. 设  $s_{ij}$  是子集  $S_j$  中类  $C_i$  的用户兴趣度样本数. 由  $A$  划分成子集的期望信息如式(1)所示:

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) \quad (1)$$

项  $s_{1j} + \dots + s_{mj}$  充当第  $j$  个子集的权,并且等于子集中的用户兴趣度样本数量除以  $S$  中的样本中核,该熵值越小,子集分解的纯度越高. 在  $A$  上分枝得到的编码信息是

$$Cain(A) = I(s_1, s_2, \dots, s_m) - E(A) \quad (2)$$

式中,  $Cain(A)$  是由于知道属性  $A$  的值而导致的熵的期望压缩.

针对规则生成方法即属性选择标准算法进行了改进. 通过加权和增加兴趣参数,加强了重要的属性标注,降低了非重要属性的标注,把加权和转换为加权和与兴趣度参数之和,使生成决策树时数量少的数据元组不会被覆盖,最终使决策树减少了对取值教多的属性的依赖. 把式(1)修改为

$$E_1(A) = - \sum_{j=1}^v \sum_{i=1}^m \left( \sum_{i=1}^m s_{ij} / s + Q \right) \frac{s_{ij}}{|S_j|} \log_2 \left( \frac{s_{ij}}{|S_j|} \right) \quad (3)$$

相应的式(2)就转化为

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E_1(A) \quad (4)$$

本文优化算法的主要思想是降低式(5)的运算

复杂度,则有

$$E_1(A) = - \frac{1}{S \ln 2} \sum_{j=1}^v \sum_{i=1}^m \left( \sum_{i=1}^m s_{ij} + SQ \right) \frac{s_{ij}}{|S_j|} \ln \left( |S_j| - \sum_{k=1}^m s_{kj} / |S_j| \right) \quad (5)$$

式中,  $s_{ij}$  是子集  $S_j$  中类  $C_i$  的样本数,  $|S_j|$  是  $S_j$  中所有类别的样本总数,  $\sum_{k=1}^m s_{kj}$  是  $S_j$  中非类  $C_i$  的类别的样本数和.

令  $\alpha = - \sum_{k=1}^m s_{kj} / |S_j|$ , 则存在一个实数  $\beta$ , 当  $\ln(1 + \alpha) < \ln(1 + \beta)$ , 且  $\alpha$  和  $\beta$  很小时, 有  $\ln(1 + \alpha) \approx \alpha < \ln(1 + \beta) \approx \beta$ , 所以可知由  $B$  划分成子集的熵或期望信息有

$$E_1(B) = - \frac{1}{S \ln 2} \sum_{j=1}^v \sum_{i=1}^m \left( \sum_{i=1}^m s_{ij} + SQ \right) \frac{s_{ij}}{|S_j|} (\beta) \quad (6)$$

当  $E_1(A) < E_1(B)$  时, 必然有  $E_1(A) < E_1(B)$  式(4)转化为

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E_1(A) \quad (7)$$

对于引入用户兴趣度参数的 ID3 算法的快速分类的优化极大节约了时间, 因为经过优化, 用  $-\sum_{k=1}^m s_{kj} / |S_j|$  代替了  $\ln(s_{ij} / |S_j|)$ , 使得用户兴趣分类的复杂程度得到极大的简化, 为后续塑造基于用户兴趣分类优化的聚类模型提供可靠的依据.

### 2.2 考虑用户兴趣分类优化的自适应模糊聚类目标函数

考虑上述分析的用户兴趣分类优化结果, 融合合理的优化条件, 可设置考虑用户兴趣分类优化的自适应模糊聚类目标函数见式(8), 并给出两个优化的约束条件, 分别见式(9)和(10).

$$J_e(X, U, V, W) = \sum_{j=1}^n \omega_j^p \sum_{i=1}^c u_{ij}^m d_{ij}^2 \quad (8)$$

$$\text{s. t. } \sum_{j=1}^n u_{ij} > 0, \sum_{i=1}^c u_{ij} = 1 \quad (9)$$

$$\prod_{j=1}^n \omega_j = 1 \quad (10)$$

式中,  $J_e$  称为聚类模型的目标函数. 一组新的参数  $\omega_j$  引入到  $J_e$  中, 称为用户兴趣  $x_j$  的自适应度, 所有  $\omega_j$  构成一个向量  $\mathbf{W}$ , 称为自适应度向量.  $p$  是自适指数, 是一个预先给定的常数, 用来调节自适应量的取值. 式(10)表明全体自适应度  $\omega_j$  施加了“乘为1”这样的约束条件, 这个约束条件描述了用户兴趣分类

间的内在关联性。

聚类目标函数求最有是一种非线性优化问题,通常可以用拉格朗日极值法来求解。在考虑两类约束条件的前提下,优化函数可以写为

$$J_{e, \varphi_1, \varphi_2}(X, U, V, W) = \sum_{j=1}^n \omega_j^p \sum_{i=1}^c u_{ij}^m d_{ij}^2 + \varphi_1 \left( \sum_{i=1}^c u_{ij} - 1 \right) + \varphi_2 \left( \prod_{j=1}^n \omega_j - 1 \right) \quad (11)$$

此处,  $\varphi_1$  和  $\varphi_2$  是两个拉格朗日乘法算子,分别对应于模糊隶属度和自适应度的约束条件。

在式(11)的两侧分别  $u_{ij}$  对  $\omega_j$  和求偏导数,可得

$$\begin{cases} \frac{\partial J_{AFCM, \varphi_1, \varphi_2}}{\partial u_{ij}} = m \cdot \omega_j^p \cdot u_{ij}^{m-1} d_{ij}^2 + \varphi_1 \\ \frac{\partial J_{AFCM, \varphi_1, \varphi_2}}{\partial \omega_j} = p \cdot \omega_j^{p-1} \cdot \sum_{i=1}^c u_{ij}^m d_{ij}^2 + \varphi_2 \left( \prod_{k=1, k \neq j}^n \omega_k \right) \end{cases} \quad (12)$$

在求  $J_e$  最优解的过程中,需要更新聚类原型矩阵  $V(c \times q)$ 。在自适应模糊聚类模型下,直接给出聚类原型的迭代等式,见式(13), $v_i$  则是考虑用户兴趣分类优化的自适应模糊聚类模型的第  $i$  个聚类原型。

$$v_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (13)$$

式中,  $u_{ij}$  称为自适应模型下的模糊隶属度,其计算式为

$$u_{ij}^m = \omega_j u_{ij}^m \quad (14)$$

在式(13)和式(14)的基础上,可获取新的用户兴趣分类优化的自适应模糊聚类原型迭代式,如式(15)所示:

$$v_i = \frac{\sum_{j=1}^n \omega_j u_{ij}^m x_j}{\sum_{j=1}^n \omega_j u_{ij}^m} \quad (15)$$

上述分析过程采用 ID3 决策树算法,将最高信息增益的属性当前节点的检测属性,确保结果分解中的用户兴趣样本分类所需的信息量最小,极大简化了用户兴趣分类的复杂程度,构建用户兴趣分类优化的自适应模糊聚类目标函数,更新聚类原型矩阵,在自适应模糊聚类模型下,直接给出聚类原型

的迭代等式,获取考虑用户兴趣分类优化的聚类模型。

### 3 实验分析

为了验证改进本文模型和传统模型的聚类结果,将某电影网页上每个用户标注的电影按照时间的先后顺序分为3份:前70%的电影数据作为用户稳定兴趣模型训练集;中间10%模拟用户浏览行为建立用户的即时兴趣模型;后20%作为测试集。

表1给出5次测试中本文聚类模型以及传统聚类模型的目标函数最大值(max)和整个迭代中全局目标函数均值(mean),以及是否陷入局部最优解情况(LOS)。分析表1可以看出,本文模型的各项指标都优于传统聚类模型,并且本文模型陷入局部最佳解的概率远远低于传统模型。

表1 5次测试中两种模型的目标函数各项指标情况

| 模型   |      | 1     | 2     | 3     | 4     | 5     |
|------|------|-------|-------|-------|-------|-------|
| 本文模型 | max  | 1 528 | 1 449 | 1 545 | 1 530 | 1 356 |
|      | mean | 1 260 | 1 140 | 1 221 | 1 036 | 1 002 |
|      | Los  | 否     | 否     | 否     | 否     | 是     |
| 传统模型 | max  | 1 203 | 883   | 1 546 | 1 259 | 1 459 |
|      | mean | 995   | 658   | 1 209 | 1 002 | 1 224 |
|      | Los  | 是     | 是     | 否     | 是     | 否     |

分析图1和图2可以看出,本文聚类模型相对于传统聚类模型不容易陷入局部最优解,在本文聚类模型没有陷入局部最优解,并且全局最优解不改变时,本文聚类模型能够可快跳出局部最佳解,而在图2中,传统聚类模型的收敛速度慢,而且当全局最优解多次不改变时,其将一直无法得到改进,导致其无法跳出局部最优解。

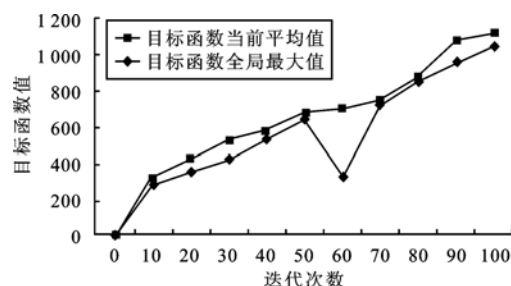


图1 本文模型的最大适应度值变化

实验结果表明,随着用户数量的不断增加,两种模型聚类结果的查准率和查全率都呈现下降趋势,并且本文模型的聚类结果的查准率和查全率都优于

传统模型,本文模型呈现平稳的下降趋势,而传统模型出现了显著的波动问题.

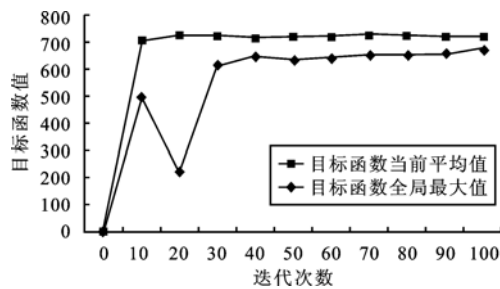


图2 传统模型的最大适应度值变化

#### 4 结束语

本文提出了一种考虑用户兴趣分类优化的聚类模型,根据系统对用户聚类的要求,采用ID3决策树算法提高用户兴趣分类计算速度,将最高信息增益的属性当成前节点的检测属性,确保结果分解中的用户兴趣样本分类所需的信息量最小,极大简化了用户兴趣分类的复杂程度,构建用户兴趣分类优化的自适应模糊聚类目标函数,更新聚类原型矩阵,在自适应模糊聚类模型下,直接给出聚类原型的迭代等式,获取考虑用户兴趣分类优化的聚类模型.实验结果说明所提模型相对于传统聚类模型不容易陷入局部最优解,具有较高的查全率和查准率.

#### 参考文献:

[1] 张海荣,朱信忠,赵建民,等.一种优化的基于用户聚类的过滤推荐策略[J].计算机系统应用,2008(11):95-98.

- [2] 黄茂峰,倪巍伟,王佳俊,等.一种面向聚类的对数螺旋线数据扰动方法[J].计算机学报,2012(11):2275-2282.
- [3] 刘露,彭涛,左万利,等.一种基于聚类的PU主动文本分类方法[J].软件学报,2013(11):2571-2583.
- [4] 傅强.某型涡轴发动机起动过程建模及仿真[J].计算机仿真,2013,30(12):57-60.
- [5] 韩忠明,陈妮,乐嘉锦,等.面向热点话题时间序列的有效聚类算法研究[J].计算机学报,2012(11):2337-2347.
- [6] 赵捷,赵荣彩,丁锐,等.基于嵌套循环分类的并行识别技术[J].软件学报,2012(10):2695-2704.
- [7] 高灿,苗夺谦,张志飞,等.主动协同半监督粗糙集分类模型[J].模式识别与人工智能,2012(5):745-754.
- [8] 杨杰,陈恩红.面向个性化服务的用户兴趣偏移检测及处理方法[J].电子技术,2009(11):72-76.
- [9] 黄世祥,电磁场复杂模型计算过程的仿真分析[J].科技通报,2013,29(11):37-40.
- [10] 陈克寒,韩盼盼,吴健.基于用户聚类的异构社交网络推荐算法[J].计算机学报,2013(2):349-359.
- [11] 张宏莉,鲁刚.分类不平衡协议流的机器学习算法评估与比较[J].软件学报,2012(6):1500-1516.
- [12] 张玉红,周全,胡学钢.面向跨领域情感分类的特征选择方法[J].模式识别与人工智能,2013,11:1068-1072.

#### 作者简介:

刘建粉 女,(1972—),硕士,副教授.研究方向为网络信息处理.

史永昌 男,(1971—),硕士,副教授.研究方向为智能数据处理和数据挖掘、信息安全.