# 模糊关联规则挖掘和推理系统①

方 睿，黄方胜

(武汉大学 计算机学院，武汉 430072)

**摘 要：**为了准确对用户的消费提供个性化建议，智能推荐系统应运而生。对智能推荐系统体系结构及实现方法进行了有益的探索，将模糊聚类、模糊关联规则挖掘与模糊推理相结合，设计并实现了一个原型智能推荐系统。该系统在经过实际数据运行后，经过模糊聚类、模糊关联规则挖掘和规则筛选，并经过模糊推理，系统可以给出一些符合实际背景的结论。

**关键词：**智能推荐；模糊规则；模糊推理；FuzzyJess；FCM

## Recommendation System Based on Fuzzy Association Rules Mining and Inference

FANG Rui, HUANG Fang-Sheng

(Computer School, Wuhan University, Wuhan 430072, China)

**Abstract:** In order to provide personalized suggestions for consumers in a more accurate way, Intelligent Recommendation System (IRS) has emerged. This study made a meaningful exploration on its structure and its implementation. By combining the fuzzy clustering, fuzzy association rules mining and fuzzy reasoning, this study designed and implemented an original IRS. It could give more matching suggestions based on the authentic data and the operation of fuzzy clustering, fuzzy associations rules mining, rules filtering, and fuzzy reasoning.

**Keywords:** intelligent recommendation; fuzzy rules; fuzzy reasoning; fuzzy Jess; FCM

## 1 引言

为了方便用户在网络环境下更好的购物，满足不同用户的不同偏好和不同需求，智能推荐系统(Intelligent Recommendation System)应运而生[1]。虽然传统的数据挖掘方法广泛应用于智能推荐系统，但是这些方法在处理模糊性方面还显得不足。因此，本文的研究内容主要是将模糊关联规则挖掘、模糊聚类和模糊推理应用于智能推荐系统实现中，对智能推荐系统体系结构及实现方法进行了有益的探索和一定程度上的丰富。

本文设计并实现的智能推荐系统的应用场景在超市，系统能够根据超市的历史销售数据所挖掘出的模糊规则进行推理，并对用户当前的消费给予合适的建议。

## 2 基本算法

本论文所实现的智能推荐系统首先要对数量型的销售数据进行模糊聚类，将具体的销售数据映射到相应的类别中。然后，根据转化后的销售数据挖掘模糊关联规则。最后，依据模糊关联规则对当前的订单进行推理，提供有效的消费建议。因此，为了完成系统功能，需要实现模糊聚类算法和模糊关联规则挖掘算法。

### 2.1 模糊 C 均值聚类算法

模糊 C 均值聚类算法，是由硬 C-均值聚类算法改进得来，它也是用隶属度矩阵和聚类中心来确定样本数据的类别属性的一种聚类算法，其包含了模糊集合的概念[2]。样本数据不明确属于任何类别，其属于每一个类别的相似程度用隶属度来表示，隶属度为[0,1]

① 收稿时间:2010-09-08;收到修改稿时间:2010-11-15

之间的实数。

模糊 C 均值聚类算法在计算样本中所有数据与划分的聚类簇的非相似度的公式与硬 C 均值聚类算法不一样，其计算公式如下：

$$J=\sum_{i=1}^{m}J_i=\sum_{i=1}^{m}\sum_{j=1}^{n}u_{ij}^{a}\|x_k-P_i\|^2 \tag{1}$$

此方法计算出来的值越大，说明样本数据与簇聚类中心越不相似；值越小，样本数据与簇聚类中心越相似。

另外，模糊 C-均值聚类算法要根据聚类簇中心，生成模糊隶属度矩阵 U，其计算公式如下：

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{2}$$

模糊 C-均值算法的简要描述如下。

输入：参数分为基本聚类样本信息、算法的参数信息、算法控制信息、输出信息。并设：pattern 为样本点向量；dimension 为样本的维数；numpattern 为样本的个数；cata 分类的数目；maxcycle 为最大循环次数；limit 为算法结束迭代的阈值。

输出：rescenter 为输出的聚类中心向量；umatrix 为输出的划分矩阵。

算法的基本步骤为：

1) 初始化聚类中心{$P_1$，$P_2$，…，$P_m$}。典型的做法是从所有数据点中随机选取初始迭代点作为初始的聚类中心；

2) 根据初始化的聚类中心，计算与初始化聚类中心相对应的隶属度矩阵 umatrix。其方法是，如果||patternk - catai||=0，则将 umatrix[k][i]置为 1，并对第 k 行的其它隶属度设为 0；

3) 根据公式 2，对 umatrix 的其他未赋值的项进行计算并赋值；

4) 根据公式 1，计算所有样本数据与每个簇聚类中心非相似程度之和。如果该值小于程序最初定义的阈值 limit 或者算法迭代次数大于 maxcycle，则算法停止，返回样本的聚类中心向量 rescenter 和划分矩阵 umatrix；否则，转下一步。

5) 对原定的聚类中心进行修改，并初始化聚类中心相对应的隶属度矩阵 umatrix，返回第 3 步。

### 2.2 模糊关联规则挖掘算法

模糊关联规则挖掘，是由关联规则挖掘算法 Apriori[3]改进得来。该过程主要包含连接与剪枝两个步骤。连接步骤是将频繁模糊 k-项集连接成候选模糊(k+1)-项集。剪枝操作是对所有的候选模糊项集进行模糊支持度计算，如果计算得到的模糊支持度小于设定的最小模糊支持量或者候选模糊项集的子项不是频繁模糊项集，则将此候选模糊项集从中删除。

模糊关联规则挖掘是基于存放历史数据的数据仓库，分两个步骤完成。首先要对数据仓库中的数量型字段进行模糊化，然后，根据模糊属性和隶属度矩阵挖掘模糊关联规则。在规则挖掘过程中要计算模糊支持度和模糊置信度，定义如下：

定义 1. 模糊支持度。假设任意模糊属性集合 L={$L_1$，$L_2$，…，$L_n$}，转化后的数据库为 Q，则对于模糊属性集合 L 的模糊支持度为 FSupport(L)，具体公式如下：

$$\text{FSupport(L)}=\frac{\text{FSupport\_num(L)}}{\text{number}}=\frac{\sum_{i=1}^{number}\prod_{m=1}^{n}q_i(L_m)}{number} \tag{3}$$

其中，number 为数据库 Q 中的元组数量，Fsupport_num(L)为数据库对模糊属性集合 L 的模糊支持数，计算展开为 $\sum_{i=1}^{number}\prod_{m=1}^{n}q_i(L_m)$。因此，如果 Fsupport(L)大于或等于模糊最小支持度，这些模糊属性集合 L 可称为模糊频繁属性集，如果此模糊频繁属性集的长度为 n，则称其为 n-项频繁模糊属性集。

定义 2. 模糊置信度。任何模糊关联规则“X=>Y”，其中模糊属性集合 X 的长度为 m，模糊属性集合 Y 的长度为 n，则此模糊关联规则为 FConfidence(Y|X)，其具体计算公式如下：

$$\text{FConfidence(Y|X)}=\frac{\text{FSupport}(X\cup Y)}{\text{FSupport(X)}}=\frac{\sum_{i=1}^{number}\prod_{j=1}^{m+n}q_i(X_j)}{\sum_{i=1}^{number}\prod_{j=1}^{m}q_i(X_j)} \tag{4}$$

模糊关联规则的挖掘算法的简要描述如下[4]。

输入：最小支持度阈值；最小置信度；事务数据库

输出：模糊关联规则，频繁模糊项集

具体步骤为：

1) 对数据库中的每一个数量型数据采用 FCM 算

法进行离散化操作，对每个数量型数据划分成若干个模糊集等级。

2) 将 FCM 算法得到的数量型数据的模糊集等级作为模糊属性，创建一个新的数据库，原数据库中的每一个数量型属性在新数据库中由几个模糊属性来代替，并将共同代替原数量型属性的几个模糊属性标记为同类模糊属性。

3) 在新数据库中统计所有 1-项候选项集的模糊支持度，当模糊支持度大于最小支持度阈值时，此 1-项候选项集为 1-项频繁项集。

4) 连接 k-项频繁项集，对于同类模糊属性无需进行连接操作，得到(k+1)项候选项集。考察(k+1)项候选项集的支持度，如果此(k+1)项候选项集的模糊支持度大于最小支持度阈值时，此(k+1)项候选项集为频繁项集。

5) 重复第 4 步操作，直至发现所有的模糊频繁属性项集。

6) 考察所有模糊频繁属性项集，生成候选模糊关联规则，当候选规则的置信度大于最小置信度阈值时，此候选规则为模糊关联规则。

## 3 智能推荐系统的设计

本文所设计的智能推荐系统能够从存放交易数据的数量型数据库挖掘得到模糊关联规则；然后将有效的模糊关联规则输入模糊推理模块；最后，根据当前用户的消费信息推理并给予个性化的建议。

### 3.1 系统体系结构

本论文所要实现的智能推荐系统是基于模糊关联规则的挖掘和模糊推理系统。智能推理系统的设计结构图如图 1 所示。

该系统中，系统数据层包含数据仓库和实时数据库。以下简要介绍其关键部分。

1) 数据仓库：存放历史交易数据，主要用于模糊关联规则挖掘。

2) 实时数据库：存放实时交易数据。智能推荐系统以实时交易数据为基础，根据挖掘出的模糊规则进行推理，对用户的实时交易给予个性化建议。

3) 模糊聚类：以历史交易数据为基础，采用模糊 C-均值聚类算法对数据仓库中的数量型属性进行模糊聚类划分。对数量型属性划分的若干模糊属性也会传给模糊推理系统，以帮助其对实时数据进行模糊化。

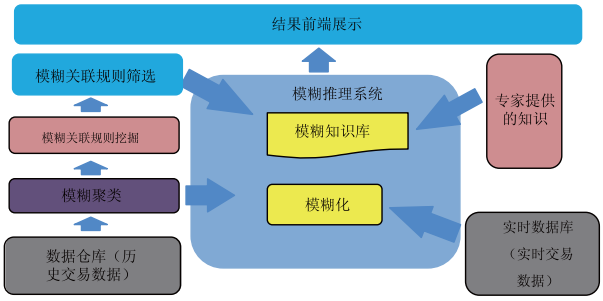


图 1 智能推荐系统设计结构图

4) 模糊关联规则挖掘：采用模糊关联规则的挖掘算法进行挖掘。

智能推荐系统的知识来源有两处：一为专家提供的知识；二为基于数据仓库所挖掘出的模糊规则。

5) 专家知识：此为业务知识，如超市的节日打折活动规则，以及长年累月积累的销售经验都可以转化为专家提供的知识。

6) 模糊关联规则筛选：通过模糊关联规则挖掘算法挖掘出的部分模糊规则可能存在没有意义或者不合逻辑的情况，因此，有必要对挖掘出的模糊关联规则进行过滤，删除其中不合理的规则。规则筛选过程可以分为系统过滤和人工过滤。在挖掘出的规则数量不多的情况下，人工筛选是一种方便且有效率的方法。

7) 模糊推理系统：其包含模糊知识库和模糊化的事实。本文使用由加拿大国家研究委员会信息技术研究所开发出的模糊推理工具 FuzzyJ Toolkit[5]，加上 Jess (JAVA Expert System Shell)[6]专家系统推理框架相结合进行模糊推理的实现。

FuzzyJ Toolkit 是一组能够处理模糊概念和进行模糊推理的 Java 类，可以有效地探索模糊逻辑并进行模糊推理。此工具将 FuzzyCLIPS 扩展到 CLIPS 专家系统外壳[7]。其中 FuzzyJ Toolkit 的应用程序端口（API）可以单独使用来创造模糊规则并进行推理。FuzzyJ Toolkit 与 Jess 结合能够发挥出和 FuzzyCLIPS 相类似甚至更灵活的功能。

8) 用户交互的界面：其作用是辅助人工过滤无效规则；显示模糊推理结果，对用户给予个性化消费建议。

### 3.2 系统工作流程

智能推荐系统可以抽象为两条工作流，一条为挖掘模糊关联规则，另一条为模糊推理。

模糊关联规则挖掘的流程如图 2，步骤如下：

1) 将数据仓库中的数量型属性模糊聚类划分为多个模糊属性，并将原数据库转化为以模糊属性为字段的新数据库，其数据为相应的隶属度。

2) 以转化后的新数据库为基础，挖掘出频繁模糊项集，从中发现模糊关联规则。

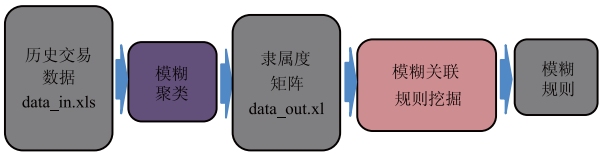


图 2　模糊关联规则挖掘工作流程图

模糊推理的流程如图 3，步骤如下：

1) 将系统挖掘出的模糊关联规则进行人工筛选，只将有效的模糊规则输入模糊推理系统中。

2) 将专家提供的知识输入模糊推理系统中。

3) 根据模糊聚类得到的模糊属性，对实时数据库中的事务数据进行模糊化。

4) 根据模糊推理系统中的知识库，对模糊化的事实进行模糊推理，最后将推理结果反馈给用户。

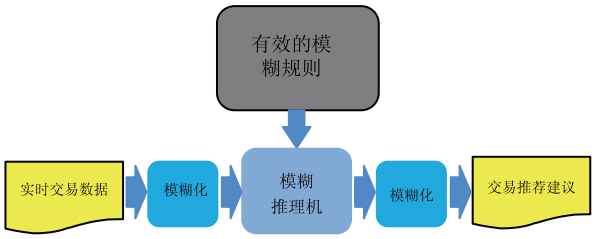


图 3　模糊推理工作流程图

## 4　系统功能展示

为了从功能上检测所实现原型智能推荐系统的效果，从某超市的销售数据中过滤得到只包含啤酒与尿布两种商品的交易数据，共包含 12406 条，实验以此交易数据作为模糊关联规则挖掘的原始数据库数据，从中挖掘规则指导消费。

以下从数据流角度来展示原型系统中模糊聚类、模糊关联规则挖掘与模糊推理的处理过程，并分析实验效果。

1) 通过模糊 C-均值聚类算法对原始的交易数据进行模糊分类。在聚类操作前设置啤酒购买数量分为 4 类，分别为少量，一般量，较多量，大量；设置尿布购买数量分为 3 类，分别为少量，一般量，大量。表 1 为对啤酒购买数量模糊分类得到的聚类中心，表 2 为对尿布购买数量模糊分类得到的聚类中心：

表 1　啤酒交易数据模糊分类得到的聚类中心

| 啤酒模糊分类 | 少量 | 一般量 | 较多量 | 大量 |
|---|---|---|---|---|
| 聚类中心 | 2.47622788 | 5.479904517 | 8.578261205 | 17.035256486 |

表 2　尿布交易数据模糊分类得到的聚类中心

| 尿布模糊分类 | 少量 | 一般量 | 大量 |
|---|---|---|---|
| 聚类中心 | 1.4139118612 | 3.7377158297 | 6.394021572 |

模糊聚类后的结果除聚类中心外，还包括相应的隶属度矩阵，其形式为以模糊概念——商品数量类别为字段的交易数据。表 3 为模糊聚类得到的以模糊属性为字段的隶属度矩阵，只显示部分数据。

表 3　以模糊属性为字段的隶属度矩阵

| 啤酒：少量 | 啤酒：一般量 | 啤酒：较多量 | 啤酒：大量 | 尿布：少量 | 尿布：一般量 | 尿布：大量 |
|---|---|---|---|---|---|---|
| 0.01028921 | 0.04965134 | 0.93625577 | 0.003803681 | 0.16057108 | 0.80381162 | 0.03561730 |
| 0.95411308 | 0.03722796 | 0.00747865 | 0.001180323 | 0.97334422 | 0.02126989 | 0.00538588 |
| 0.03067866 | 0.08504244 | 0.84947415 | 0.034804727 | 0.01175690 | 0.97439523 | 0.01384785 |
| 0.01938088 | 0.94302554 | 0.03565110 | 0.001942464 | 0.97334422 | 0.02126989 | 0.00538588 |
| 0.97276058 | 0.02036881 | 0.00576739 | 0.001103208 | 0.97334422 | 0.02126989 | 0.00538588 |
| …… | …… | …… | …… | …… | …… | …… |

由于得到的隶属度矩阵中含有大量的浮点数，会对模糊关联规则挖掘阶段的计算速度产生负面影响。因此，为了提高挖掘的速度，系统可设置一个隶属度过滤阈值，将数值很小并且对候选项集的支持度计算影响很小的(如小于 0.01)值，将其设为 0。所以，对表 3 的数据进行优化得到表 4。

表 4　以模糊属性为字段的隶属度矩阵优化

| 啤酒：少量 | 啤酒：一般量 | 啤酒：较多量 | 啤酒：大量 | 尿布：少量 | 尿布：一般量 | 尿布：大量 |
|---|---|---|---|---|---|---|
| 0.01028921 | 0.04965134 | 0.93625577 | 0 | 0.16057108 | 0.80381162 | 0 |
| 0.95411308 | 0.03722796 | 0 | 0 | 0.97334422 | 0.02126989 | 0 |
| 0.03067866 | 0.08504244 | 0.84947415 | 0.034804727 | 0.01175690 | 0.97439523 | 0.01384785 |
| 0.01938088 | 0.94302554 | 0 | 0 | 0.97334422 | 0.02126989 | 0 |
| 0.97276058 | 0.02036881 | 0 | 0 | 0.97334422 | 0.02126989 | 0 |
| …… | …… | …… | …… | …… | …… | …… |

(2) 对表 4 中的数据进行模糊关联规则挖掘，并设置最小模糊支持度为 0.1，最小模糊置信度为 0.5。挖

掘得到 10 条频繁模糊项集和 4 个模糊规则。表 5 为挖掘得到 10 条频繁模糊项集，表 6 为挖掘得到的 4 个模糊规则。

表 5 挖掘得到的 10 条频繁模糊项集

| 频繁模糊项集 | {啤酒：少量} | {啤酒：一般量} | {啤酒：较多量} | {啤酒:大量} | {尿布：少量} |
|---|---|---|---|---|---|
| 支持度 | 0.177711893 | 0.319359059 | 0.261266686 | 0.241662362 | 0.507757071 |
| 频繁模糊项集 | {尿布：一般量} | {尿布：大量} | {啤酒：少量，尿布：少量} | {啤酒：一般量，尿布：少量} | {啤酒：大量，尿布：大量} |
| 支持度 | 0.303362305 | 0.188880625 | 0.156832002 | 0.238840672 | 0.169454429 |

表 6 挖掘出的 4 个模糊关联规则

| 模糊关联规则 | 模糊支持度 | 模糊置信度 |
|---|---|---|
| 啤酒(大量) => 尿布(大量) | 0.169454429 | 0.701203229977618 |
| 尿布(大量) =>啤酒(大量) | 0.169454429 | 0.8971509329582078 |
| 啤酒(一般量)=>尿布(少量) | 0.156832002 | 0.7478750486804031 |
| 啤酒(少量)=>尿布(少量) | 0.238840672 | 0.8825070718775216 |

3) 将表 6 中挖掘出的模糊关联规则以程序形式输入到 FuzzyJess 模糊推理系统中。在此为简化系统测试，只将尿布(大量) => 啤酒(大量)模糊规则输入到推理系统中，应用 FuzzyJess 模糊推理系统进行推理。输入三个不同的啤酒购买数量数据到的推理系统中，比较推理系统结果。

当输入的尿布购买数量为 3 时，系统推理后得出：无规则被激活，对于当前交易无推荐建议。

当输入的尿布购买数量为 5 时，推理结果显示“推荐用户购买 14.9 听啤酒”，为了体现推理后结果的精度和效果，在此就没有对解模糊化后的数量进行四舍五入。模糊推理后的图形解释如图 4 所示。

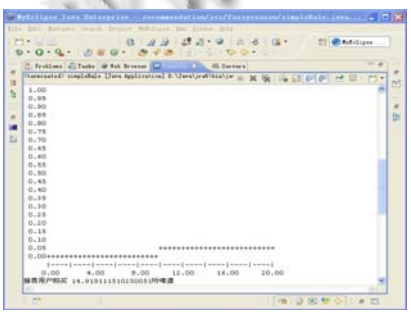

图 4 当购买 6 包尿布时，基于模糊规则推理后的结果

当输入的尿布购买数量为 7 时，推理结果显示“推荐用户购买 15.4 听啤酒”。模糊推理后的图形解释如图 5 所示。

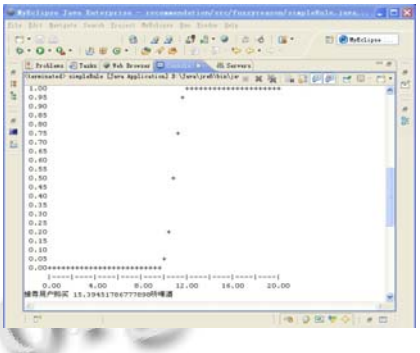

图 5 当购买 7 包尿布时，基于模糊规则推理后的结果

## 5 结语

从以上结构可以看出，原型智能推荐系统的基本功能已经实现。经过模糊聚类、模糊关联规则挖掘和筛选，并经过模糊推理，系统可以给出一些符合实际背景的结论。系统还有很多地方可改进，例如，本论文所使用的模糊关联规则算法，其计算模糊支持度的公式比较简单，没有考虑到事务的加权问题。如果有的数据更加有指导意义，可以适当将其权值加大，使得挖掘出的规则更有意义，并且模糊关联规则挖掘阶段的增量更新也是一个比较值得研究的问题。

## 参考文献

1 Schafer JB, Konstan J, Riedl J. Recommender System in E-Commerce. Proceedings of the first ACM conference on Electronic Commerce. 1999:158−166.

2 Zadeh L A. Fuzzy Sets. Information and Control. 1965, 8:33−35.

3 Han JW, Kamker M. Concepts and Techniques. 2nd Edition. pp:4−6.

4 陆建江,张亚非,宋自林.模糊关联规则的研究与应用.北京:科学出版社,2008.30−34.

5 Friedman-Hill E. Jess in action. Manning Publications, 2003.

6 Orchard RA. FuzzyCLIPS version 6.04 user’s guide. http://ai.iit.nrc.ca/IR-public/fuzzy/,1998-10-25/2000, 07－25.