Computer Science and Application 计算机科学与应用, 2012, 2, 271-275
http://dx.doi.org/10.12677/csa.2012.25048 Published Online December 2012 (http://www.hanspub.org/journal/csa.html)

An Optimal Algorithm of Selection Books in Library Based on Rough Set and Fuzzy Set

Liping Zhao[1], Wenjun Liu[2]

[1]Library of Changsha University of Science and Technology, Changsha
[2]Department of Mathematics and Computing Science, Changsha University of Science and Technology, Changsha
Email: zhlp731118@126.com, liuwjzhlp@126.com

Received: Oct. 14th, 2012; revised: Oct. 24th, 2012; accepted: Nov. 11th, 2012

Abstract: Combining rough set and fuzzy set theory, an optimal decision algorithm of library to select books is put forward. During this algorithm, firstly, we construct the similarity matrix from the original information system; secondly, we classify all the programs according to fuzzy clustering; thirdly, an algorithm to choose the optimal program is put forward according to the minimum distance of weighted relative deviation.

Keywords: Rough Set; Similarity Matrix; Fuzzy Set; Optimal Program

基于粗糙集与模糊集理论的图书馆最优选书算法

赵利萍[1]，刘文军[2]

[1]长沙理工大学图书馆，长沙
[2]长沙理工大学数学与计算科学学院，长沙
Email: zhlp731118@126.com, liuwjzhlp@126.com

收稿日期：2012 年 10 月 14 日；修回日期：2012 年 10 月 24 日；录用日期：2012 年 11 月 11 日

摘　要：本文将粗糙集理论与模糊集理论结合起来，给出一种图书馆最优选书算法。该算法首先从已知数据的初始信息系统出发，计算各选书方案之间的相似度，从而构造相似矩阵，然后根据相似矩阵的传递闭包对各方案进行聚类，并根据粗糙集理论求各属性重要性，最后利用加权综合的思想及最小距离方法选择最优买书方案。

关键词：粗糙集；相似矩阵；模糊集；最优方案

1. 引言

图书馆是社会公众文化领域的主阵地，是社会知识信息的存储、咨询中心，也是弘扬社会主义精神文明主旋律的重要载体。随着科技的发展，图书馆不仅在数量上需要增加，而且图书种类也须向多样化发展，图书馆的价值不再仅仅以其所拥有的馆藏图书的数量来衡量，而是以它为用户提供各种形式的信息的能力和质量来衡量。在这种新形式下，图书馆在选书决策时，如何利用目前有限的人力、经费资源，而又使所做决策符合读者阅读或参考，从而为广大读者提供高质量的服务，是目前图书工作者需要认真研究和解决的一个重要课题。

在实际过程中，由于影响选书决策的因素很多，且大多数具有模糊性与不确定性，所以在处理这类问题时，可以结合不确定性理论。本文就是基于这种想法，结合粗糙集与模糊集这两种不确定性理论，给出一种图书馆最优选书算法。

2. 预备

粗糙集理论[1]是由波兰科学家 Z. Pawlak 在 1982 年提出的一种处理含糊和不确性问题的新型数学工具。经过近三十年的发展，该理论已渗透到人工智能

的各个分支，在机器学习、决策分析、过程控制、模式识别与数据挖掘等领域取得了成功的应用[2-6]。该理论的一个最大优点是它无须提供问题所需处理的数据集合之外的任何先验信息，能客观有效地分析和处理不精确、不确定与不完全数据，并从中发现隐含的知识，揭示潜在的规律。

为了处理智能数据，粗糙集理论将知识进行符号化，将所要研究的数据用一个信息系统的形式给出，信息系统的基本成分是研究对象的集合，关于这些对象的知识是通过指定对象的基本特征(属性)和它们的特征值(属性值)来描述。信息系统的数据以关系表的形式表示，关系表的行对应要研究的对象，列对应对象的属性，对象的信息是通过指定对象的各属性值来表达。

形式上，$S=(U,A,V,f)$是一类信息系统，其中U是有限论域；A为所有属性的集合，$V=\bigcup_{a\in A}V_a$，V_a是属性a的值域；$f:U\times A\to V$是信息函数，即对于任意的$u\in U$，$a\in A$，有$f(u,a)\in V_a$。

对U上任意属性集B，定义B上的不可区分关系$ind(B)$如下：

$$ind(B)=\left\{(u_i,u_j)\middle|\forall a\in B, f(u_i,a)=f(u_j,a)\right\},$$
$$(u_i,u_j)\in ind(B),$$

则称u_i与u_j是B不可区分的。容易证明不可区分关系$ind(B)$是U上的等价关系，符号$U/ind(B)$(简记为U/B)表示不可区分关系$ind(B)$在U上导出的划分，$ind(B)$中等价类称为B基本集。

3. 聚类分析

对数据进行模糊聚类分析，一般有数据规格化、建立模糊相似矩阵、聚类三大步。

第一步：数据规格化。在实际应用中，不同的数据可能有不同的量纲和数量级，故在运算过程中可能突出某数量级特别大的特性指标对分类的作用，而降低甚至排除了某些数量级很小的指标的作用，致使对各特性指标的分类缺乏一个统一的尺度，为了清除特性指标单位的差别和特性指标数量级不同的影响，必须对各指标值施行数据规格化的处理，从而使每一个指标值统一于某种共同的数值特性范围。

设$U=\{u_1,u_2,\cdots,u_n\}$为被分类的对象，每个对象有m个指标描述，即对第i个对象有$u_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ $(i=1,2,\cdots,n)$，对应的信息系统如表1所示。

数据规格化方法一般有[7,8]：

1) 数据标准化

$x'_{ij}=\dfrac{x_{ij}-\overline{x_j}}{\sigma_j}$ $(i=1,2,\cdots,n;\ \ j=1,2,\cdots,m)$，其中$\overline{x_j}=\dfrac{1}{n}\sum_{i=1}^{n}x_{ij}$，$\sigma_j=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)^2}$；

2) 均值规格化

$x'_{ij}=\dfrac{x_{ij}}{\sigma_j}$ $(i=1,2,\cdots,n;\ \ j=1,2,\cdots,m)$;

3) 中心规格化

$x'_{ij}=x_{ij}-\overline{x_j}$ $(i=1,2,\cdots,n;\ \ j=1,2,\cdots,m)$;

4) 最大值规格化

$x'_{ij}=\dfrac{x_{ij}}{M_j}$ $(i=1,2,\cdots,n;\ \ j=1,2,\cdots,m)$，其中$M_j=\max\{x_{1j},x_{2j},\cdots,x_{nj}\}$;

5) 极差规格化

$x'_{ij}=\dfrac{x_{ij}-m_j}{M_j-m_j}$ $(i=1,2,\cdots,n;\ \ j=1,2,\cdots,m)$，其中$m_j=\min\{x_{1j},x_{2j},\cdots,x_{nj}\}$，$M_j=\max\{x_{1j},x_{2j},\cdots,x_{nj}\}$;

6) 对数规格化

$x'_{ij}=\log x_{ij}$ $(i=1,2,\cdots,n;\ \ j=1,2,\cdots,m)$。

第二步：构造模糊相似矩阵。建立模糊相似矩阵的方法主要有相似系数法、距离法、贴近度法和主观评定法。在此我们采用 Hamming 距离法来构造模糊相似矩阵，即对象u_i与u_j的相似度

$r_{ij}=1-\dfrac{1}{m}\sum_{k=1}^{m}\left|x'_{ik}-x'_{ik}\right|$。

第三步：建立模糊等价矩阵。通过求传递闭包将n阶模糊相似矩阵改造成为n阶模糊等价矩阵，我们一般采用平方法求R的传递闭包。

Table 1. Information system
表 1. 信息系统

U	a_1	a_2	…	a_m
u_1	x_{11}	x_{12}	…	r_{1m}
u_2	x_{21}	x_{22}	…	x_{2m}
⋮	⋮	⋮	⋮	⋮
u_n	x_{n1}	x_{n2}	…	r_{nm}

第四步：选取最佳分类阈值进行聚类。

在分类过程中，如何确定分类阈值是一个重要的问题，在此，我们用最优分类变化率来确定分类阈值[8]。

在等价矩阵 $\boldsymbol{R}^*$ 中，将 $\boldsymbol{R}^*$ 中元素 λ_i 从大到小排列，即 $1>\lambda_1>\lambda_2>\cdots>\lambda_k>0$，定义 λ_i 的分类变化率 C_i 为：

$$C_i=\frac{\lambda_{i-1}-\lambda_i}{n_i-n_{i-1}},$$

其中 n_i 与 n_{i-1} 分别为第 i 次和第 $i-1$ 次聚类的对象个数。

若 $C_j=\max_i\{C_i\}$，则认为第 j 次聚类的置信水平 λ_j 为最佳值。

4. 最佳阈值 λ 下的各属性权重

下面，结合模糊集与粗糙集理论，我们给出一种求连续信息表的属性权重的方法。

输入连续信息系统 $S=(U,A,V,f)$，$U=\{u_1,u_2,\cdots,u_n\}$，$A=\{a_1,a_2,\cdots,a_m\}$。

输出各属性的权重。

步骤 1　根据上述方法找出最佳分类阈值 λ_i，即 $C_i=\max_j\{C_j\}$，其中 $C_j=\frac{\lambda_{j-1}-\lambda_j}{n_j-n_{j-1}}$；

步骤 2　在最佳阈值 λ_i 下将对象进行聚类，所得分类看做是在对象在等价关系 $ind(A)$ 下的等价类；

步骤 3　从 A 中删除属性 a_k $(k=1,2,\cdots,m)$，类似地，在阈值 λ_i 下将对象分类，此分类看做是在等价关系 $ind(A-\{a_k\})$ 下的分类，若 $U/ind(A-\{a_k\})\neq U/ind(A)$，那么属性 a_k 在 A 中是不可省的，属性 a_k 的重要性

$$\sigma(a_k)=1-\frac{\left|U/ind(A)\cap U/ind(A-\{a_k\})\right|}{|U|};$$

步骤 4　归一化属性重要性，得到在 λ_i 阈值下，属性 a_k $(k=1,2,\cdots,m)$ 的权重，即 $w_i=\frac{\sigma(a_i)}{\sum_{k=1}^{m}\sigma(a_k)}$。

5. 图书馆最优选书方案

设 $S=(U,A,V,f)$，$U=\{u_1,u_2,\cdots,u_n\}$ 是 n 个选书方案，$A=\{a_1,a_2,\cdots,a_m\}$ 为对决策起重要作用的个属性所构成的集合，则各方案可由其相应的 m 个属性值所确定，设 $u_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ $(i=1,2,\cdots,n)$，式中 x_{ij} 表示第 i 个方案的第 j 个属性值，称为方案 u_i 的属性指标向量。把这 n 个方案的属性指标向量作为行构成如上述表 1 所示的信息系统。

在信息系统 $S=(U,A,V,f)$ 中，令

$$\Delta_{ij}=\frac{\left|x_j^0-x_{ij}\right|}{x_{j\max}-x_{j\min}}\ (i=1,2,\cdots,n,\ \ j=1,2,\cdots,m),$$

其中 $x_{j\max}=\max\{x_{1j},x_{2j},\cdots,x_{nj}\}$，

$x_{j\min}=\min\{x_{1j},x_{2j},\cdots,x_{nj}\}$，

$$x_j^0=\begin{cases}x_{j\max}, & \text{当属性指标}a_j\text{是正指标},\\ x_{j\min}, & \text{当属性指标}a_j\text{是负负指}。\end{cases}$$

这里正指标是指因素指标值越大方案越优的因素指标，负指标是指因素指标值越小方案越优的因素指标。称 Δ_{ij} 为相对偏差值。称 x_j^0 为属性 a_j 的标准值，而称 $u_0=(x_1^0,x_2^0,\cdots,x_m^0)$ 为标准值向量。

由上述 $n\times m$ 个相对偏差值 Δ_{ij} 作为元素构成一个模糊矩阵

$$\Delta=\begin{pmatrix}\Delta_{11} & \Delta_{12} & \cdots & \Delta_{1m}\\ \Delta_{21} & \Delta_{22} & \cdots & \Delta_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ \Delta_{n1} & \Delta_{n2} & \cdots & \Delta_{nm}\end{pmatrix}$$

叫做相对偏差矩阵。

对于给定的 n 种选书方案，若这个方案的加权相对偏差距离与标准值向量的距离最小，则我们选择这个方案为最优选择方案。根据这种思想，我们可得如下最优购书算法：

步骤 1　根据前面的算法确定每个属性 a_j 的权重 w_j；

步骤 2　根据公式

$$d_i=d(u_i,u_0)=\sqrt{\sum_{j=1}^{m}\left(w_j\cdot\Delta_{ij}\right)^2}\ \ (i=1,2,\cdots,n),$$

计算每个方案 u_i 与标准值向量 u_0 的加权相对偏差距离；

步骤 3　根据最小加权相对偏差距离确定最优方案，即若 $d_k=\min\{d_1,d_2,\cdots,d_n\}$，则 u_k 选择为最优方案。

6. 实例分析

下面我们用一个例子说明上述算法的有效性与

可行性。某高校图书馆计划新购一批计算机图书。计算机科学系王老师根据多年的教学经验，结合学生的兴趣爱好，计划从 A,B,C,D,E,F,G,H 8 类图书中选择一类最优的图书，即选择方案是：$U=\{A,B,C,D,E,F,G,H\}$。按照图书价格(a_1)、教学效果(a_2)、科研价值(a_3)、售后服务(a_4)、学生爱好(a_5)等因素来进行选取。由于影响图书的几个指标有很强的专业性，因此采用专家主观评分法，对 8 类图书的基本情况进行打分，结果如表 2 所示。

由于此数据都在$[0,1]$之间，且量纲相同，我们选择极差规格化，然后，根据公式 $r_{ij}=1-\frac{1}{m}\sum_{k=1}^{m}\left|x'_{ik}-x'_{jk}\right|$，计算 u_i 与 u_j 之间的相似度，构造相似度矩阵 $\boldsymbol{R}=\left(r_{ij}\right)_{n\times n}$ 及求得它的传递闭包 $\boldsymbol{R}^*$ 分别如下：

Table 2. Expert scoring information system
表 2. 专家打分情况信息表

U	a_1	a_2	a_3	a_4	a_5
u_1	0.1	0.2	0.3	0.3	0.2
u_2	0.3	0.2	0.4	0.2	0.1
u_3	0.4	0.5	0.3	0.5	0.3
u_4	0.5	0.3	0.2	0.1	0.5
u_5	0.2	0.1	0.5	0.4	0.2
u_6	0.3	0.3	0.5	0.5	0.3
u_7	0.5	0.4	0.4	0.3	0.4
u_8	0.2	0.2	0.3	0.2	0.3

$$\boldsymbol{R}=\begin{pmatrix}1&0.73&0.55&0.43&0.72&0.57&0.53&0.85\\0.73&1&0.48&0.47&0.68&0.63&0.6&0.78\\0.55&0.48&1&0.48&0.47&0.72&0.68&0.6\\0.43&0.47&0.48&1&0.25&0.4&0.67&0.58\\0.72&0.68&0.47&0.25&1&0.75&0.48&0.67\\0.57&0.63&0.72&0.4&0.75&1&0.63&0.62\\0.53&0.6&0.68&0.67&0.48&0.63&1&0.58\\0.85&0.78&0.6&0.58&0.67&0.62&0.58&1\end{pmatrix}$$

$$\boldsymbol{R}^*=\begin{pmatrix}1&0.78&0.72&0.67&0.72&0.72&0.68&0.85\\0.78&1&0.72&0.67&0.72&0.72&0.68&0.78\\0.72&0.72&1&0.67&0.72&0.72&0.68&0.72\\0.67&0.67&0.67&1&0.67&0.67&0.67&0.67\\0.72&0.72&0.72&0.67&1&0.75&0.68&0.72\\0.72&0.72&0.72&0.67&0.75&1&0.68&0.72\\0.68&0.68&0.68&0.67&0.68&0.68&1&0.68\\0.85&0.78&0.72&0.67&0.72&0.72&0.68&1\end{pmatrix}$$

下面，利用最优分类变化率找最佳阈值。

根据模糊等价矩阵 $\boldsymbol{R}^*$，得：

当 $0.85<\lambda\le 1$ 时，

$U/ind(A)=\{\{u_1\},\{u_2\},\{u_3\},\{u_4\},\{u_5\},\{u_6\},\{u_7\},\{u_8\}\}$；

当 $0.78<\lambda\le 0.85$ 时，

$U/ind(A)=\{\{u_1,u_8\},\{u_2\},\{u_3\},\{u_4\},\{u_5\},\{u_6\},\{u_7\}\}$；

当 $0.75<\lambda\le 0.78$ 时，

$U/ind(A)=\{\{u_1,u_2,u_8\},\{u_3\},\{u_4\},\{u_5\},\{u_6\},\{u_7\}\}$；

当 $0.72<\lambda\le 0.75$ 时，

$U/ind(A)=\{\{u_1,u_2,u_8\},\{u_3\},\{u_4\},\{u_5,u_6\},\{u_7\}\}$；

当 $0.68<\lambda\le 0.72$ 时，

$U/ind(A)=\{\{u_1,u_2,u_3,u_5,u_6,u_8\},\{u_4\},\{u_7\}\}$；

当 $0.67<\lambda\le 0.68$ 时，

$U/ind(A)=\{\{u_1,u_2,u_3,u_5,u_6,u_7,u_8\},\{u_4\}\}$；

当 $0<\lambda\le 0.67$ 时，$U/ind(A)=\{U\}$。

由于所在对象各自成类或全部对象并入一类没有实际意义，根据最佳分类阈值的选取方法，可得 $\lambda=0.72$ 为最佳阈值。此时

$$U/ind(A)=\{\{u_1,u_2,u_3,u_5,u_6,u_8\},\{u_4\},\{u_7\}\};$$

从 A 中分别删除 a_1,a_2,a_3,a_4,a_5，在阈值 $\lambda=0.72$ 下，用相同的方法，可得

$U/ind\left(A-\{a_1\}\right)=\left\{\{u_1,u_2,u_8\},\{u_3\},\{u_4\},\{u_5,u_6\},\{u_7\}\right\}$,

$U/ind\left(A-\{a_2\}\right)=\left\{\{u_1,u_2,u_8\},\{u_3,u_5,u_6\},\{u_4\},\{u_7\}\right\}$,

$U/ind\left(A-\{a_3\}\right)=\left\{\{u_1,u_2,u_5,u_8\},\{u_3,u_6\},\{u_4,u_7\}\right\}$,

$U/ind\left(A-\{a_4\}\right)=\left\{\{u_1,u_2,u_5,u_6,u_8\},\{u_3,u_4,u_7\}\right\}$,

$U/ind\left(A-\{a_5\}\right)=\left\{\{u_1,u_2,u_8\},\{u_3\},\{u_4\},\{u_5,u_6\},\{u_7\}\right\}$,

所以$\sigma(a_1)=\sigma(a_2)=\sigma(a_3)=\sigma(a_5)=0.75$，$\sigma(a_4)=1$，可得各个属性的权重为：(0.1875,0.1875,0.1875,0.25,0.1875)。

显然对于考虑的这些因素指标除第一个图书价格外，其余都是正指标，由信息系统得$u_0=(0.1,0.5,0.5,0.5,0.5)$，相对偏差矩阵为：

0	0.75	0.67	0.5	0.75
0.5	0.75	0.33	0.75	1
0.75	0	0.67	0	0.5
1	0.5	1	1	0
0.25	1	0	0.25	0.75
0.5	0.5	0	0	0.5
1	0.25	0.33	0.5	0.25
0.25	0.75	0.67	0.75	0.5

根据公式

$$d_i=d(u_i,u_0)=\sqrt{\sum_{j=1}^{5}\left(w_j\cdot\Delta_{ij}\right)^2}\ \ (i=1,2,\cdots,8),$$

求得每种选择与标准值向量u_0的加权相对偏差距离为：

$$d_1=0.266;d_2=0.321;d_3=0.210;d_4=0.376;$$

$$d_5=0.247;d_6=0.162;d_7=0.243;d_8=0.286.$$

因为d_6最小，所以认为方案 6 最优，即购买第 6 种书籍最合适。

7. 小结

随着科学技术的发展，社会对人才的要求最来越高，而图书馆的建设与发展是提高人们素质的一个重要基础。在新形势下，各图书馆如何针对自身的特色，选择适合读者研究需要和阅读参考的图书，是图书馆面临的一项重要任务。本文结合模糊集与粗糙集理论，对拟选图书根据给定的条件进行计算分析，为馆员提供最佳选择方案，从而让馆员在有限精力的条件下选择确定合适图书。节省了馆员的时间，同时也可以使做出的购书决策更全面地符合实际需要。

参考文献 (References)

[1] Z. Pawlak. Rough set: Theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers, 1991.

[2] Y. Y. Yao, Y. Zhao. Attribute reduction in decision-theoretic rough set models. Information Sciences, 2008, 178(17): 3356-3373.

[3] 胡清华, 谢宗霞, 于达仁. 基于粗糙集加权的文本分类方法研究[J]. 情报学报, 2005, 1: 91-100.

[4] 王国胤, 张清化. 不同知识粒度下粗糙集的不确定性研究[J]. 计算机学报, 2008, 31(9): 1588-1598.

[5] 史忠植. 知识发现(第二版)[M]. 北京: 清华大学出版社, 2011.

[6] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1243.

[7] 张振良. 模糊集理论与方法[M]. 武汉: 武汉大学出版社, 2010.

[8] 梁保松, 曹殿立. 模糊数学及其应用[M]. 北京: 科学出版社, 2007.

