Hans Journal of Data Mining 数据挖掘, 2015, 5, 32-38
Published Online April 2015 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2015.52005

Hans汉斯

# Health Rapid Assessment Based on Fuzzy Rough Set Reduction Algorithm

**Linghua Wu**

Chongqing Bureau of Navy Equipment, Chengdu Sichuan
Email: hql101@126.com

Received: Apr. 7th, 2015; accepted: Apr. 23rd, 2015; published: Apr. 30th, 2015



## Abstract

**Health is a good important indicator of people's living standard, but how to quickly evaluate health level based on physiological index is the core issue of concern for medical wisdom. Due to redundancy characteristics between various physiological indicators to determine, it's important to analyze the properties for physical test indexes. Rough Set has a strong processing capacity, in terms of knowledge discovery. This paper based on Rough Set of Fuzzy data processing method, puts forward a kind of Fuzzy Rough Set reduction algorithm; this algorithm can implement rapid assessment to the health using physical testing data. At the same time due to the use of Fuzzy theory, the result is more suitable for people to accept, and can reflect the probability characteristics in nature of the data. The simulation analysis of actual data can be found that the algorithm has a high recognition accuracy in the rapid assessment of health.**

## Keywords

**Rough Set, Pattern Recognition, Attribute Reduction, Fuzzy Theory**

# 基于Fuzzy Rough Set约简的健康快速评估算法

**吴凌华**

海装重庆局，四川 成都
Email: hql101@126.com

收稿日期：2015年4月7日；录用日期：2015年4月23日；发布日期：2015年4月30日

## 摘　要

**健康是人们生活水平优良的重要指标，但是如何快速基于生理检测指标来评估其健康水平是智慧医学所关注的核心问题。由于各种生理检测指标之间具有冗余特性，因此对所分析的生理检测指标进行属性分析具有重要作用。Rough Set在知识挖掘方面具有很强的处理能力，本文依据Rough Set对Fuzzy数据的处理方法，提出一种Fuzzy Rough Set约简算法，该算法可以实现利用生理检测数据对健康快速评估。同时由于使用来Fuzzy理论，所以其结果更加适合人们的接受方式，更能体现数据本质的概率特性。对实际数据的仿真分析可以发现该算法在健康快速评估上具有高的识别准确度。**



## 关键词

**Rough Set，模式识别，属性约简，Fuzzy 理论**


## 1. 概述

医疗卫生行业是事关百姓民生的行业，人们期望应用规范化的、科学化的、现代化的、信息化的手段来加强医院的管理，提高医院的工作效率。目前新型医疗系统是一门由医学、信息学、管理科学、计算机科学等多种学科交叉为一体的边缘科学，其在发达国家已经得到了广泛的应用，并且创造了良好的社会效益和经济效益。

目前医疗信息系统及医保信息管理系统已广泛应用在大中型医院及相关机构中，在这些医院和机构中普遍使用关系数据库来存储医院里的档案数据，从这些海量的繁杂的档案数据中运用各种数据挖掘技术和方法来寻找和挖掘疾病本身的内在规律和各种疾病之间的联系和关联规律，以及疾病和患者病理外在特征之间的关系，发现隐藏在大量的数据中存在的关系和规则，都具有重要应用价值和深远的影响。它可以有效地促进医疗信息系统的纵向完善和横向的发展，可以方便病人及病人家属对病人病情的客观了解；帮助医务人员进行分析、决策和管理，开展大规模、高水平的医学研究及建立完善的治疗方案，向医疗卫生事业相对落后的医院和社区医疗的医生提供临床诊断和决策支持；通过对已有的医疗信息系统的数据进行有效地挖掘和分析，构建病历档案知识库，为相应的疾病患者寻求最为合适的治疗方案和技术，以最小的风险和较低的花费来获得治疗效益的最大化[1] [2]。

数据挖掘技术在医疗系统上的应用为以上问题提供了有力的技术支持。20 世纪 90 年代以来，数据挖掘技术在商业领域得到了成功的应用和发展，而医学技术与人类社会紧密相关，具有很强的实验性、实践性和统计性，与商业领域有一定的共性，因此，探索数据挖掘技术在医疗系统方面的应用尤为迫切。

医疗数据挖掘与其他商业领域的数据挖掘有着很大的不同，它有着自己的特点。医疗数据的特点是数据量大，种类繁多，具有复杂性和多样性；具有数据冗余性：具有数据不完整性；具有数据的隐私性；具有数据的时间序列性。医疗数据既包括临床信息系统数据，管理信息系统数据。临床信息系统数据又包括患者检查、入院、住院、治疗、出院等一些与患者有关的信息，也包括在治疗过程中的产生的图像和信号等数据。管理信息系统数据又包括财务、人事和设备的管理等信息，所以医疗数据数据量大，种类比较多。同时几乎每天都有大量的种类繁多的类似和相同的数据信息保存到医疗数据库中，形成庞大的医疗数据的数据源，会导致了医疗数据的冗余。医疗数据的数据不完整性体现在病历档案的登记不够不完整；再加上疾病患者的个体上的差异以及主治医师的不同，更增加了病历档案记录和医学信息的表达的不确定性。大量的医疗数据涉及到患者的个人隐私，患者的隐私如果受到不法的侵害势必会对其生活和工作造成不良的影响，所以从事医疗数据挖掘的研究者在对医疗数据合理挖掘分析利用的同时，也

要保护好患者的隐私。患者的发病是具有过程性的，医疗检测的图像和波形往往都具有时间序列性，甚至患者的诊断治疗也具有一定的时间序列性，所以医疗数据具有时间序列性[3]。

针对上述问题，本文提出一种基于 Fuzzy Rough Set 约简算法，该算法可以实现利用生理检测数据对健康快速评估。

## 2. Fuzzy Rough Set 理论

知识就是人们在认识和改造客观世界的过程中得到的一些认知和经验。所以，知识可以被用来描述客观世界里的任何事物或对象，而且还可以根据这些知识如对象的不同特征来对它们进行有效的分类。

定义 2.1：给定一个非空的有限集合$U$，如果它由需要被研究的对象构成，那么就称其为论域。对于集合$U$ 中的任意一子集 $X$，称其是论域$U$ 中的一个概念或范畴。这些概念簇就被称作为$U$ 的知识[4]。

定义 2.1：设 R 是论域$U$ 上的模糊等价关系也即模糊属性，对 $x \in U$，定义

$$\mu_{[x]_R} = \mu_R(x, y) \tag{1}$$

为对象 $x$ 的模糊等价类，它表示论域 U 中和对象 $x$ 邻近的全部元素的聚集，是一个模糊集。

定义 2.2：对于模糊决策表 $\mathrm{DT} = (U, \underset{\sim}{A} = C \cup D, V, f)$，$R$ 是论域 $U$ 上的模糊属性，且 $X \subseteq U$，那么 $X$ 关于模糊属性 $R$ 的下近似 $\underline{RX}$ 隶属度函数和关于 $R$ 的下近似 $\overline{RX}$ 的隶属度函数分别为：

$$\mu_{\underline{RX}}(F_i) = \inf_{x \in U} \max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{2}$$

$$\mu_{\overline{RX}}(F_i) = \sup_{x \in U} \max\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{3}$$

其中，$\underline{RX}$ 和 $\overline{RX}$ 均是 $U$ 上的模糊集合。模糊上、下近似还可定义为

$$\mu_{\underline{RX}}(x) = \sup_{F_i \in U/R} \min\left(\mu_{F_i}(x), \inf_{y \in U} \max\{1 - \mu_{F_i}(y), \mu_X(y)\}\right) \tag{4}$$

$$\mu_{\overline{RX}}(x) = \sup_{F_i \in U/R} \min\left(\mu_{F_i}(x), \inf_{y \in U} \min\{\mu_{F_i}(y), \mu_X(y)\}\right) \tag{5}$$

由模糊下近似和上近似构成的二元对 $\langle \underline{RX}, \overline{RX} \rangle$ 就被称为 Fuzzy Rough Set。

定义 2.3：在粗糙集理论中，属性集合 $A$ 对论域 $U$ 的划分可表示为

$$U / A = \otimes\{U / \alpha \mid \alpha \in A\} \tag{6}$$

其中 $\otimes$ 表示：

$$S_1 \otimes S_2 = \{X \cap Y \mid X \in S_1, Y \in S_2, X \cap Y \neq \varnothing\} \tag{7}$$

如果 $A = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$，那么

$$U / A = \{X_{1_{i1}} \cap X_{2_{i2}} \cap \cdots \cap X_{1_{im}} \mid X_{1_{i1}} \in U / \alpha_1, X_{1_{i2}} \in U / \alpha_2, \cdots, X_{1_{im}} \in U / \alpha_m\} \tag{8}$$

通过上面对属性集合 A 的模糊等价类的定义，一个对象属于这样的一个模糊等价类的隶属函数可以定义为：

$$\mu_{F_1 \cap F_2 \cap \cdots \cap F_m}(x) = \mu_{F_1}(x) \wedge \mu_{F_2}(x) \wedge \cdots \wedge \mu_{F_m}(x) = \min\left(\mu_{F_1}(x), \mu_{F_2}(x), \cdots, \mu_{F_m}(x)\right) \tag{9}$$

其中，$F_1 \cap F_2 \cap \cdots \cap F_m$ 为 U/A 的模糊等价类。

在经典的粗糙集理论中，正域被定义为下近似的并集。根据这一规则，就可以定义 Fuzzy Rough Set 中一个对象属于模糊正域的隶属函数。

定义 2.4：给定模糊决策表 $\mathrm{DT}=\left(U, \underset{\sim}{A}=C \cup D, V, f\right)$， $P \in C$， $Q \in D$，定义

$$\mu_{pos_P(Q)}(x)=\sup_{X \in U/Q} \mu_{\underline{P}(X)}(x) \tag{10}$$

为对象 $x$ 属于模糊正域的隶属度。利用模糊正域的定义，Fuzzy Rough Set 中新的依赖函数可被定义为[6]

$$\gamma'_P(Q)=\left|\mu_{pos_P(Q)}(x)\right| \Big/ |U|=\sum_{x \in U} \mu_{pos_P(Q)}(x) \Big/ |U| \tag{11}$$

它表示属性 Q 依赖于属性 P 的程度。

## 3. 基于 Fuzzy Rough Set 的属性约简

基于 Fuzzy Rough Set 理论，对于其属性约简，设计其过程为[7]：

如图 1 所示，采用模糊粗糙集对目标进行识别主要有四个步骤：模糊化预处理、约简、提取决策规则、规则匹配。

本文拟采用模糊 C 均值(Fuzzy C-means: FCM)聚类的来对属性进行模糊化。

模糊 C 均值(Fuzzy C-means: FCM)通过不断地优化目标函数来实现对对象集合的分类。FCM 目标函数的一般化形式是[6]。

$$J=\sum_{j=1}^{c}\sum_{i=1}^{n}\left(u_{ij}\right)^m d_{ij}^2 \qquad 1 \le m < \infty \tag{12}$$

其中，$d_{ij}=\left\|x_i-v_j\right\|$ 为第 $i$ 个对象与第 $j$ 类中心之间的欧氏距离；$\{x_i, i=1,\cdots,n\}$ 是 $n$ 个对象组成的样本集合；$c$ 是事先指定的类别数目；$v_j, j=1,2,\cdots,c$ 为每个类别的聚类中心值；$\mu_{ij}$ 表示第 $i$ 个对象对于第 $j$ 类的隶属程度，它的范围区间为[0,1]；m 为模糊加权指数，它可以控制聚类结果中隶属度值。

对于模糊 C 均值，它要求每个对象关于各个类别的隶属度的总和等于 1 [7]，即

$$\sum_{i=1}^{c} u_{ij}=1, \quad i=1,\cdots,n \tag{13}$$

在条件(13)式的约束下，对(12)式求极小值，令 $J$ 对 $v_j$ 和 $\mu_{ij}$ 的偏导数为 0，得聚类中心为：

$$v_j=\sum_{i=1}^{n} u_{ij}^m x_i \Big/ \sum_{i=1}^{n} u_{ij}^m, \quad j=1,2,\cdots,c \tag{14}$$

和隶属度为：

$$u_{ij}=1 \Big/ \sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{2/(m-1)}, \quad i=1,2,\cdots,n \;\; j=1,2,\cdots,c \tag{15}$$

用迭代的方法，求解(14)式和(15)式，当算法收敛的时候，获得各个类别的中心值，还有每个对象属于各个类别的隶属度值，模糊划分过程便结束。这就是模糊 C 均值算法，该算法快速且简单。图 2 给出了基于 FCM 的单个属性模糊化的流程[8]。

为考察某个条件属性相对决策属性是否重要，粗糙集理论的方法是：将该条件属性从决策表中删除，然后检查决策表是否有变化。如果这个条件属性对决策表来说是很重要的，那么决策表的分辨能力就会发生很大的改变，但如果这个条件属性对决策表来说没那么重要，甚至一点用都没有，那么这个决策表的分辨能力就不会发生太大改变，甚至没变化。通过这个思想就可以找到决策表的一个相对约简。通常，条件属性的重要度可通过该属性相对决策属性的正域来描述[9]：

$$\mathrm{sig}(\alpha, P; D)=\gamma_P(D)-\gamma_{P-\{\alpha\}}(D) \tag{16}$$

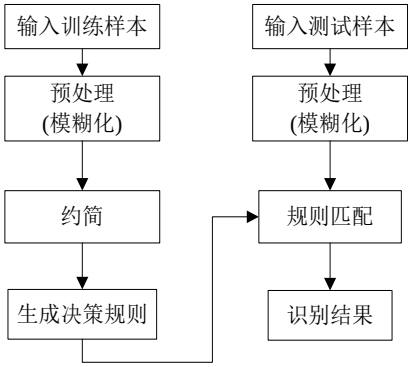


**Figure 1.** The framework of knowledge acquisition based on fuzzy rough set
**图 1.** 基于模糊粗糙集的知识获取框架

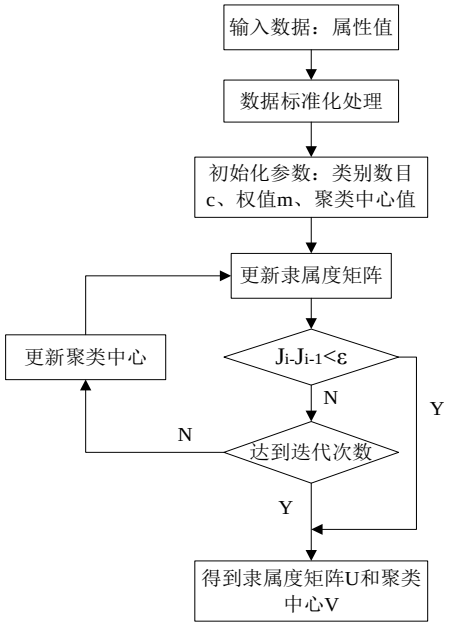


**Figure 2.** The flow chart of a single blurred attribute based on FCM
**图 2.** 基于 FCM 的单个属性模糊化流程图

基于模糊粗糙集条件熵(FRCE)的属性约简算法可以通过判断属性重要度来逐步添加属性到当前的约简集合中。由于条件熵是单调非递增函数，所以 FRCE 属性约简算法的思想是将使 $H(D|R)$ 减少最多的条件属性依次添加到约简集合中，图 3 给出了该算法的流程图，其中 $\eta$ 为误差阈值。

## 4. 基于 Fuzzy Rough Set 的健康快速评估

粗糙集及模糊粗糙集理论在医疗领域最通常的应用是诊断或结果预测。通常通过生成决策规则完成，从而实现基于 Fuzzy Rough Set 的健康快速评估，其系统构建见图 4 [10]。

按照上述流程，本文对威斯康星大学医学院提供的乳腺癌数据库来进行测试。该数据库包含 569 个病例，其中，良性 357 例，恶性 212 例。每个病例的一组数据包括采样组织中各细胞核的 10 个特征量的平均值、标准差和最坏值(各特征的 3 个最大数据的平均值)共 30 个数据。

首先是进行模糊化流程，然后计算不同类别数目 $c$ 下(所有条件属性的类别数目均设为 $c$)划分后的不

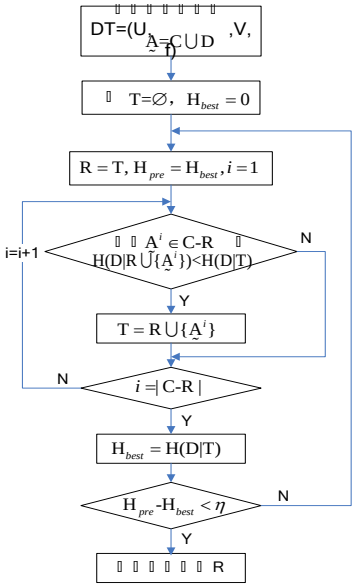


**Figure 3.** The flow chart of conditional entropy attribute reduction algorithm based on the fuzzy rough set
**图 3.** 基于模糊粗糙集条件熵的属性约简算法流程图

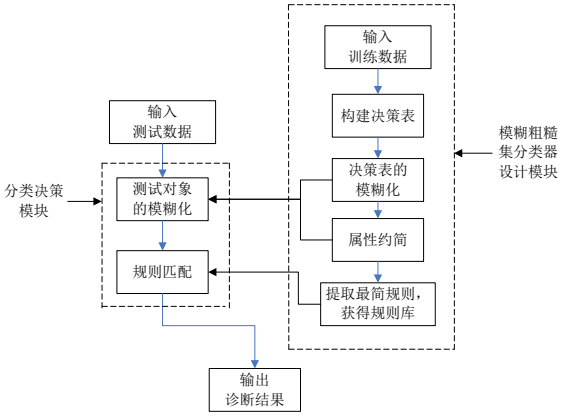


**Figure 4.** SAR recognition process based on the fuzzy rough set
**图 4.** 基于模糊粗糙集的 SAR 识别流程

相容度，其结果如图 5 所示：

通过上述分析，最终从 10 个特征中采用经过约简之后的三个属性来实现对数据的快速识别。按照此约简结果的基础上，提取决策规则，得到了包含 53 条的训练规则库，这相比原来的 569 条件规则来说，可以减少诊断时间，提高系统效率。然后依据图 4 所构建的快速评估系统对测试数据进行模糊化，并提取到测试对象的规则，从而完成与训练规则库的规则匹配，最终得到诊断正确率为 80.29%。从上述的分析可以看出，本文提出的基于 Fuzzy Rough Set 的健康快速评估在满足识别率的情况下实现诊断的高效性，使得病人不需要进行太多的冗余生理检查而实现健康的快速评估。

## 5. 结论

本文在分析健康快速评估的需求上，提出利用 Fuzzy Rough Set 来实现对生理检测数据的属性约简和

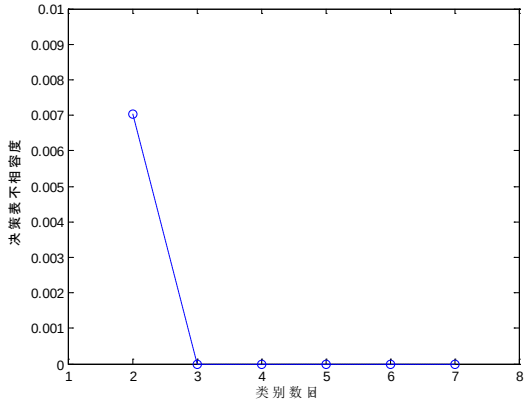


**Figure 5.** The incompatible degrees of breast cancer decision table under different number of category $c$
**图 5.** 乳腺癌决策表在不同类别数目 $c$ 下的不相容度

快速评估算法。该算法利用了 Rough Set 在属性约简上的优势，同时也利用 Fuzzy 理论采用的模糊分类方式，从而在获得更好的分类效果上也同时实现来快速评估的结果。该系统可以有利于实现人们更为合理进行生理检测，用最少的检测结果来实现高效的健康评估，从而有利改善当前的医疗现状。

## 参考文献 (References)

[1] 朱金伟, 鞠时光 (2006) 基于数据挖掘的中医药数据预处理方法. *计算机工程*, **15**, 280-283.

[2] 杨淑莹 (2011) 模式识别与智能计算. 电子工业出版社, 北京, 1-193.

[3] Jensen, R. and Shen, Q. (2009) Are more features better? A response to attributes reduction using fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, **17**, 1456-1458

[4] 苗夺谦, 李道国 (2008) 粗糙集理论、算法与应用. 清华大学出版社, 北京, 24-230.

[5] Chen, D., Zhang, L., Zhao, S.Y., Hu, Q.H. and Zhu, P.F. (2012) A novel algorithm for finding reducts with fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, **20**, 385-389.

[6] Zhao, S.Y., Tsang, E.C.C., Chen, D.G. and Wang, X.Z. (2010) Building a rule-based classifier—A fuzzy-rough set approach. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 624-638.

[7] Huang, H.H. and Kuo, Y.H. (2010) Cross-lingual document representation and semantic similarity measure: A fuzzy set and rough set based approach. *IEEE Transactions on Fuzzy Systems*, **18**, 1098-1111.

[8] Cock, M.D., Cornelis, C. and Kerre, E.E. (2007) Fuzzy rough sets: The forgotten step. *IEEE Transactions on Fuzzy Systems*, **15**, 121-130.

[9] Starczewski, J.T. (2010) General type-2 FLS with uncertainty generated by fuzzy rough sets. *FUZZ Conference Proceeding*, 1-6.

[10] Hu, Q.H., Yu, D.R., Pedrycz, W. and Chen, D.G. (2011) Kernelized fuzzy rough sets and their applications. *IEEE Transactions on Knowledge and Data Engineering*, **23**, 1649-1667.