doi:10.16104/j.issn.1673-1891.2017.03.006

粗糙集三支决策理论及其应用

安世勇[a],肖成英[a],王珑滔[b]
(四川工商学院 a.云计算与智能信息处理重点实验室;b.建筑工程学院,成都 611745)

摘要:利用粗糙集中的三支决策思想,将类用正域、负域和边界域刻画,得到初始聚类结果。然后通过定义重叠度和类与类的合并策略,将初始聚类结果进行合并,得到最终聚类结果。之后应用2个关系网络数据展示了具体的聚类步骤,并通过比较2个例子的聚类结果,分析了影响聚类结果的一个主要因素:阈值的选取。实验表明:阈值的选取对简单的网络结构数据集的聚类结果的影响并不明显,然而对复杂的网络结构数据集的聚类结果的影响则较为显著。

关键词:粗糙集;三支决策;重叠度;阈值
中图分类号:TP18 **文献标志码**:A **文章编号**:1673-1891(2017)03-0018-04

Three-way Decision Theory in Rough Sets and Its Application

AN Shi-yong[a], XIAO Cheng-ying[a], WANG Long-tao[b]
(a. Key Laboratory of Cloud Computing and Intelligent Information Processing;
b. Construction Engineering School, Sichuan Technology and Business University, Chengdu 611745,China)

Abstract: By using the three-way decision ideas in rough sets, the class is characterized by positive domain, negative domain and boundary domain to get the initial clustering results. The final clustering result was merged by defining the degree of overlap and the merging strategy of class. The two given relational network data have shown the specific clustering steps, and by comparing the clustering results of two examples, a major factor influencing the clustering results is arisen: the selection of threshold value. Experiments show that: the effect of threshold value selection on clustering results of simple network structure datasets is not obvious, but the impact of clustering results on complex network structure datasets is more significant.

Keywords: rough sets; three-way decision; overlap degree; threshold.

0 引言

粗糙集理论(Rough Sets Theory)是一种新的处理模糊和不确定性知识的数学工具。其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。目前粗糙集理论已被成功地应用于机器学习、决策分析、过程控制、模式识别与数据挖掘等领域。粗糙集理论利用2个精确集(下近似和上近似)去逼近一个粗糙集,而上近似下近似又将论域分为正域、边界域和负域3个部分。这一理论赋予了三支决策一种数学的结构表示。

三支决策理论[1]是加拿大里贾纳大学姚一豫教授等人在粗糙集[2]和决策粗糙集[3]基础上于2009年提出的新的决策理论。自此学术界开始了对三支决策方法的系统理论研究。三支决策理论在国内外已经受到了许多大学和研究机构的关注,在国内如南京大学、西南交通大学、闽南师范大学、上海师范大学等,国际上姚一豫教授所在的加拿大里贾纳大学是研究主力。2013年7月国内出版了《三支决策与粒计算》[4]一书,全面地介绍了三支决策方面的国内外研究动向。马云的成功就是三支决策中的一个典型案例[5],他改变了传统交易的一手交钱一手交货的支付决策的二支模式,即支付和不支付。但电子商务存在空间和时间的距离,不能实现一手交钱一手交货。马云便通过支付宝实现了支付的三支决策,即增加了“延迟支付”,从而减少了交易的风险。

三支决策是一种基于人类认知过程的决策方法。它以决策粗糙集为研究背景,利用粗糙集理论

收稿日期:2017-03-07
基金项目:四川工商学院科研立项:基于粗糙集的三支决策方法与应用研究(2016ZYB22X)。
作者简介:安世勇(1989—),女,河南驻马店人,助教,硕士,研究方向:粗糙集与三支决策和数学教育。

中的正域、边界域和负域,提出了一种三支决策理论:从正域里获取的正规则用来接受某事物,从负域里获取的负规则用来拒绝某事物,落在边界域上的规则表示延迟决策。这种将论域分为三部分的决策方式,很好地描述了人类在解决实际决策问题时的思维模式,为粗糙集方法应用于数据驱动的决策分类问题提供了可靠的理论依据,是信息处理的一种新的概念和计算方式,主要用于描述和处理不确定的、模糊的、不完整的和海量的信息,以及提供一种基于粒和粒间关系的问题求解方法。2016年8月,于洪、毛传凯将K均值思想融入到三支决策聚类方法中,提出了基于K-means的自动三支决策聚类方法 。

三支决策理论方法在解决不确定性度量及其应用、海量信息处理、计算机领域的数据决策、机器学习、模式识别、聚类分析[8-9]等领域有重要的应用。然而目前决策粗糙集的三支决策理论研究还处于起步环节,故研究决策粗糙集的三支决策方法与应用有着重要的理论意义和实际意义。

1 三支决策理论知识

定义1[2](粗糙集)假设用U表示非空的论域集合,R表示U上的等价关系,即$R\subseteq U\times U$,根据等价关系可将论域划分为若干个等价类,划分用U/R表示。设集合X是U的子域,即子集$X\subseteq U$表示一个概念所包含的对相集。X不一定可以准确地用R的等价类描述,即X不一定是一组等价类的并集。因此用一对上近似和下近似集刻画X,在等价关系上的下近似集和上近似集分别为

$$\underline{apr}(X)=\{x\in U|[x]\subseteq X\} \tag{1}$$

$$\overline{apr}(X)=\{x\in U|[x]\cap X\neq\varnothing\} \tag{2}$$

式中,$[x]$表示元素$x\in U$在等价关系R上的等价类。

粗糙集是一种刻画边界非空集合的理论,它采用间接刻画边界的方法,用2个分明的上下近似集合,界定了边界的范围。基于上下近似的概念,可以得到关于论域集合的一个划分,将其划分为,3个没有交叉集的子集,即正域、负域、边界域:

$$POS(X)=\underline{apr}(X) \tag{3}$$

$$BND(X)=\overline{apr}(X)-\underline{apr}(X) \tag{4}$$

$$NEG(X)=U-\overline{apr}(X) \tag{5}$$

在传统的二支决策中,往往只有接受或拒绝、是或非两种选择。但在有些实际问题中,强制做出接受或拒绝的决策,若决策失误则会导致不必要代价或后果。在这种情况下,考虑到所掌握信息的不精确或不完全,可以通过不承诺或延迟决策的第三种选择来减小或规避风险。

定义2[4](三支决策)设U为有限非空实体集合,C为有限条件集。基于条件集,三支决策的主要任务是将实体集U分为3个两两不相交的域,分别记为正域、负域、边界域,对于落在正域、负域和边界域中的实体,分别使用接受、拒绝和不承诺规则。

为实现三支决策,需引入评价函数对有限非空实体集合U中的元素x进行评估,依据评价函数的值对实体x进行决策。下面介绍基于全序集的评价函数。

设$(L,\preceq)$是一个全序关系集,L表示所有可能的决策状态值,$\preceq$表示集合L上的一个全序关系,即$\preceq$具有自反、反对称和传递关系,且集合L中任意2个元素是可比较的,$\preceq$是一个小于等于关系。

以全序关系$([0,1],\leqslant)$为例,其中L是单位区间$[0,1]$,全序关系$\preceq$是小于等于关系。实体评价函数可以看成从实体集合U到全序集L的映射,即$v:U\to L$。

采用评价函数,通过引入一对阈值(α,β),将实体集合U分为如下三个部分:

正域:$POS_{(\alpha,\beta)}(v)=\{x\in U\mid v(x)\succeq\alpha\}$; (6)

负域:$NEG_{(\alpha,\beta)}(v)=\{x\in U\mid v(x)\preceq\beta\}$; (7)

边界域:$BND_{(\alpha,\beta)}(v)=\{x\in U\mid\alpha\prec v(x)\prec\beta\}$。 (8)

2 三支决策理论在重叠聚类中的应用

在这个信息日渐复杂化和多样化的时代,对信息的分类显得尤为重要。一般的聚类方法大都要求对象只能属于一个类,但现实生活中的许多应用,如社交网络结构分析、基因数据以及生物信息处理等。普遍存在类类重叠的现象,故可用三支决策的方法对具有类类重叠现象的数据进行聚类分析。

考虑任意2个类C_i,C_j,它们的正域为$POS(C_i)$,$POS(C_j)$边界域分别为$BND(C_i)$,$BND(C_j)$。当类的正域与其它类的正域和边界域有重叠时,就只考虑正域部分;当类的正域与其它类不存在重叠部分时,则考虑该类的边界域与其它类的重叠情况。因此,类类之间重叠情况有三种:正域与正域重叠、正域与边界域重叠、边界域与边界域重叠(图1)。

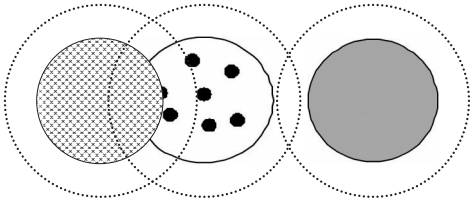

图1　类类重叠示意图

针对上面的3种重叠情况,分别定义不同的重

叠度[11]：

（1）正域与正域重叠

$$DopPP(C_i,C_j)=\frac{|POS(C_i)\cap POS(C_j)|}{|POS(C_i)\cup POS(C_j)|} \quad (9)$$

（2）正域与边界域重叠

$$DopPPB(C_i,C_j)=\frac{|POS(C_i)\cap BND(C_j)|}{|POS(C_i)\cup POS(C_j)\cup BND(C_j)|} \quad (10)$$

（3）边界域与边界域重叠

$$DopPPB(C_i,C_j)=\frac{|BND(C_i)\cap BND(C_j)|}{|BND(C_i)\cup POS(C_j)\cup BND(C_j)|} \quad (11)$$

注：|C|表示集合C中元素的个数，也称集合C的模[12]。

依据重叠类型和重叠度的定义，结合初始聚类结果，针对不同的重叠类型和不同程度的重叠度定义类与类的合并策略。定义合并策略需引入一对阈值(α,β)，这对阈值可由贝叶斯最小风险原则[8]确定出。

C_i的正域与C_j的正域重叠。

当$DopPP(C_i,C_j)\geq\alpha$时，将两类C_i，C_j的正域合并为一个新类C_k的正域，两类C_i，C_j的边界域合并为一个新类C_k的边界域。

当$\beta\leq DopPP(C_i,C_j)<\alpha$时，将两类$C_i$，$C_j$的正域分别划分到对方的边界域中。

C_i的正域与C_j的边界域重叠。

当$DopPPB(C_i,C_j)\geq\alpha$时，将类$C_j$的正域加边界域添加到$C_i$的正域。

当$\beta\leq DopPPB(C_i,C_j)<\alpha$时，将类$C_j$的正域加边界域添加到$C_i$的边界域。

C_i的边界域与C_j的边界域重叠。

当$DopBPB(C_i,C_j)\geq\alpha$时，将类$C_j$的正域加边界域添加到$C_i$的边界域。

当$\beta\leq DopPPB(C_i,C_j)<\alpha$时，考虑到2个类的边界域重叠度不大，为避免边界域过大，此时不采取任何行动。

此处的三支决策聚类方法是一种两步聚类方法：首先，针对不同类型的数据集给出初始聚类方法；然后再根据初始聚类结果不同的重叠情况给出合并策略[13]，从而得到最终的聚类结果。下面以网络型数据[14]为例，进行数据集聚类的初始化。

图2是网络型数据集基本的存储结构。网络型数据集可以用图论中网络的二元组表示$V=(N,E)$，其中N为顶点集合，E为网络中边的集合。

定义3[4]（好友）在网络数据集中，顶点与顶点之间存在直接相连时，称这2个顶点分别为对方的好友。

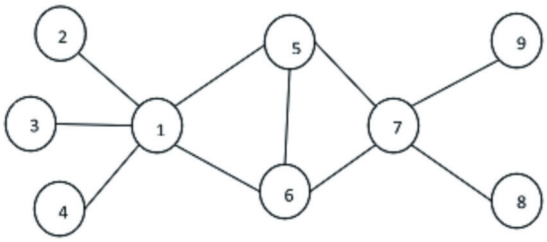

图2 关系网络示意图Ⅰ

定义4（亲密好友）在网络型数据集中，顶点与顶点是好友关系，而且2个顶点有超过一定数量的共同好友时，称这2个顶点分别为对方的亲密好友（文中设定亲密好友的阈值为1，即2个人的共同好友数目大于等于1，则称他们为彼此的亲密好友）。

定义5（一般好友）在网络型数据集中，顶点与顶点是好友关系，但2个顶点之间没有超过一定数量的共同好友时，称这2个顶点分别为对方的一般好友。

在网络型数据集中，先分别求出每个顶点的一般好友与亲密好友，其中亲密好友为该顶点的正域，一般好友为该顶点的边界域。

由图2中的网络关系示意图可看出共有9个顶点，9条边，并且好友关系为对称关系。对图2中的数据集进行聚类分析，算法步骤如下：

步骤一：求初始聚类结果。初始聚类中类的正域和边界域是根据顶点的亲密好友和一般好友组成的，每个顶点的亲密好友为该顶点的正域，一般好友为该顶点的边界域。若每个人为一个类，则图2中的网络关系的初始聚类结果见表1。

表1 图2关系网络的初始聚类结果

顶点	亲密好友	一般好友
1	5, 6	2, 3, 4
2	无	1
3	无	1
4	无	1
5	6, 7	1, 9
6	5, 7	1, 8
7	5, 6, 8, 9	无
8	7	6, 9
9	7	5, 8

以顶点1为例。顶点1的好友集合为{2,3,4,5,6}，计算顶点1的边界域和正域（共同好友的阈值为1）。根据定义4和定义5可得顶点1的正域为{5,6}，边界域为{2,3,4}。

步骤2：求解聚类结果。依据步骤α，β中的初始聚类结果，给定阈值 的值，以及合并策略，求出最终

聚类结果。根据不同的阈值可以得到不同的聚类结果。

取$\alpha=0.7$,$\beta=0.3$,得到的聚类结果为一个以1、7为核心的一个类,正域为{1,5,6,7,8,9},边界域为{2,3,4}。从聚类结果可以看出这个关系网络是一个群体,在这个群体中正域中的个体较为活跃,边界域中的个体相对低调,而1、7为群体的组织者或领导者。若更换阈值,取$\alpha=0.7$,$\beta=0.4$或$\alpha=0.8$,$\beta=0.3$,得到的聚类结果是一样的。可见图3中的网络结构相对简单,阈值的改变并不影响聚类结果。下面来分析一个复杂点的网络数据结构[15]。

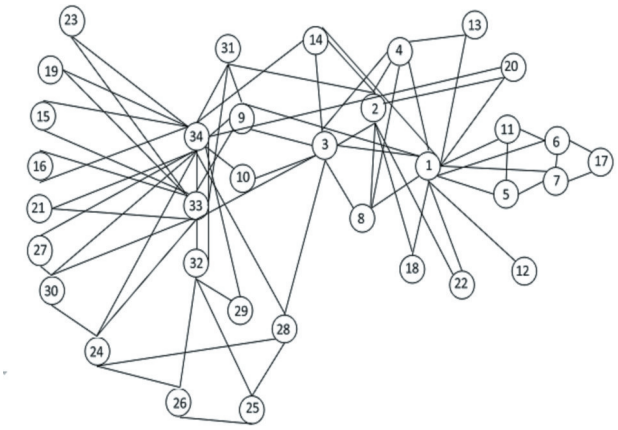

图3　关系网络示意图Ⅱ

第一组实验:$\alpha=0.7$,$\beta=0.4$。求出聚类结果为2个类,分别是以1为核心和以33、44为核心的两个群体。类1的正域为{1,2,3,4,5,6,7,8,9,11,12,13,14,17,18,20,22,32,29,28},类1的正域为{3,9,10,14,15,16,19,20,21,23,24,25,26,27,28,29,30,31,32,33,34}。可见集合{3,9,14,20,28,29,32}是类1和类2的重叠部分,{10,31,33,34}是类 的边界域,{1,2,8}是类2的边界域。

第二组实验:$\alpha=0.81$,$\beta=0.4$。在这对阈值下聚类的结果为4个类,分别为:第一个类的正域为{1,2,3,4,5,6,7,8,9,11,12,13,14,17,18,20,22,28},边界域为{31,33,34};第二个类的正域为{3,9,10,14,15,16,19,20,21,23,24,27,28,29,30,31,32,33,34},边界域为{1,2,8,24,26};第三个类的正域为{24,25,26,28,29,30,32,33,34},边界域为{3,9,10,14,15,16,19,20,21,23,27,31};第四个类的正域为{3,24,25,28,34},边界域为{9,10,14,26,29,30,31,32,33}。

3　结语

综上所述,三支决策理论在重叠聚类有着重要的应用,尤其是重叠度的定义和类与类的合并策略,更是为让人头痛的重叠聚类提供了良好的解决方案。然而聚类的结果与阈值的选取有着密切的关系,通过对比图2、图3中2个网络数据集的聚类结果,可知阈值的选取对复杂的网络结构数据集的聚类结果更为灵敏。阈值的选取对聚类结果的影响是以后需要深入研究的内容。

参考文献:

[1] YAO Y Y,WONG S K M , LINGRAS P. A Decision-theoretic Rough Set Model, Methodologies for Intelligent Systems[C]// RAS Z W, ZEMANKOVA M,EMRICHM M L (Eds.). The 5 - Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems. Tennessee: North-Holland,1990,17-25.

[2] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.

[3] YAO Y Y.Interval-set Algebra for Qualitative Knowledge Representation[C]// ABOU-RABIA O, CHANG C K, KOCZKODAJ W W (Eds.). Proceedings of the 5th International Conference on Computing and Information. Ontario: IEEE Computer Society Press,1993,370-375.

[4] 刘盾,李天瑞,苗夺谦,等.三支决策与粒计算[M].北京:科学出版社,2013:237-299.

[5] 殷业,柯德营,刘传勇.三支决策理论及应用[J].上海师范大学学报(自然科学版),2015,44(1): 95-104.

[6] 刘盾,姚一豫,李天瑞.三支决策粗糙集[J].计算机科学,2011,38(1):246-250.

[7] 于洪,毛传凯.基于k-means的自动三支决策聚类方法[J].计算机应用,2016(8):2061-2065.

[8] 于洪.三支聚类分析[J].数码设计,2016 (1):31-35.

[9] 焦鹏,于洪.软聚类中的重叠类型[J].昆明理工大学学报(自然科学版),2015(3):64-69.

[10] 魏贵莹.基于决策粗糙集的代价敏感多类分类模型与多目标决策[D].安徽:安徽大学,2016.

[11] 张聪.一种基于树结构的三支增量聚类算法研究[D].重庆:重庆邮电大学,2015.

[12] 杜丽娜.三支决策理论与应用研究[D].新乡:河南师范大学,2015.

[13] 张聪,于洪.一种三支决策软增量聚类算法[J].山东大学学报(理学版),2014(8):40-47.

[14] 于洪,王国胤,姚一豫.决策粗糙集理论研究现状与展望[J].计算机学报,2015(8):1628-1639.

[15] 刘保相,李言,孙杰.三支决策及其相关理论研究综述[J].微型机与应用,2014(12):1-3.