一种基于先验信息的混合数据聚类个数确定算法

庞天杰[1] 赵兴旺[2]
(太原师范学院计算机系 晋中 030619)[1]
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)[2]

摘 要 聚类个数的确定是聚类分析中一个富有挑战性的难题。现有的聚类个数确定方法主要采用随机选取初始聚类中心的策略,导致聚类过程中迭代次数的稳定性不强。基于此,在利用含有类标签的先验信息优化初始类中心的基础上,提出了一种基于先验信息的混合数据聚类个数确定算法。实验证明,该算法是有效的。

关键词 聚类分析,聚类个数,混合数据,先验信息,最大最小距离
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.023

Algorithm to Determine Number of Clusters for Mixed Data Based on Prior Information

PANG Tian-jie[1] ZHAO Xing-wang[2]
(Department of Computer Science, Taiyuan Normal University, Jinzhong 030619, China)[1]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)[2]

Abstract In cluster analysis, one of the most challenging and difficult problem is the determination of the number of clusters. The strategies for choosing initial prototypes randomly are used to determine the number of clusters in most of the existing methods, resulting in weak stability of iterations in clustering process. So we proposed an prior information based algorithm to determine the number of cluster for mixed data by using priori information which includes class labels to optimize initial prototype. Experiments show that the algorithm is effective.

Keywords Clustering analysis, Number of cluster, Mixed data, Prior information, Max-min distance

1 引言

聚类分析是机器学习中一个重要的研究领域。目前,研究者已经针对不同的需求提出了许多聚类算法[1-5],并且在经济、网络安全、生物信息学等领域得到了广泛应用。然而,已有的这些算法都需要事先直接或间接地指定聚类个数。因此,聚类个数的确定成为了聚类分析中面临的一个难题。

为了解决这一难题,许多研究者也开展了一些探索性的研究。例如,针对数值型数据,Sun HJ 等提出了一种基于模糊 k-means 的聚类个数自动确定方法[6];针对符号型数据,Bai L 等提出了一种初始化的方法[7];针对包含数值型和符号型数据的混合数据,Liang J Y 等提出了一种基于信息熵的混合数据聚类个数确定方法[8],这些方法在不同程度上推进了这一问题的解决。然而,聚类的目标是使类内相似度尽可能大,而类间相似度尽可能小,因此初始聚类中心的选取将直接影响聚类过程中的迭代次数和聚类算法的效率。但是,上述算法在初始聚类中心的选取上都采用了随机的方法,这样就会给算法的执行效率带来很大的不确定性。目前,已有研究者就如何优化初始聚类中心提出了相应的方法[9,10],但是使用这些方法又会提高计算的复杂度,因此,如何简单高效地优化初始类中心成为聚类个数确定过程中需要解决的一个问题。

在实际应用中,我们发现所采集到的样本会存在少量与样本类别相关的先验信息,利用这些先验信息,可以简单高效地优化初始聚类中心。然而,通常情况下,研究者认为先验信息所反映出的类别个数即为真实的个数[11-14],但在一些情况下,由于数据量较大并且先验信息不易获得,因此先验信息所标记的类别个数与数据集蕴含的真实个数并不一致。为此,本文针对实际应用中广泛存在的混合数据,利用最大最小距离提出了一种基于先验信息的混合数据聚类个数确定算法。该算法首先利用已有的先验信息确定部分初始聚类中心;然后基于最大最小距离原则,逐渐在无标签样本中选取新的类中心,在每添加一个类中心后,按照最近距离原则,对无标签样本进行划分获得聚类结果;最后利用评价指标对聚类结果进行评价来确定数据集真实的类个数。在真实数据集上的实验结果表明,本文提出的算法是有效的。

2 基础知识

2.1 混合数据信息系统

在现实世界中,很多数据都是既有数值数据又有符号数

到稿日期:2015-03-23 返修日期:2015-05-14 本文受国家自然科学基金项目:"用户行为数据"稀疏表示的理论与方法研究(61273294),山西省回国留学人员科研资助项目:基于多粒度与变粒度的群决策方法研究(2013-101)资助。
庞天杰(1980—),男,硕士,讲师,主要研究方向为数据挖掘与机器学习,E-mail:pangtj@tynu.edu.cn;赵兴旺(1984—),男,博士生,主要研究方向为数据挖掘与机器学习。

据的混合数据。通常情况下，混合数据存储在一个表中，每一行代表一个样本，混合数据信息系统描述如下。

四元组 $MDT=(U,A,V,f)$ 是一个信息系统，其中 $U=\{x_1,x_2,\cdots x_n\}$ 为非空有限集，称为论域，$x_i=(x_{i1},x_{i2},\cdots,x_{im})$ 表示由 m 个属性描述的对象；$A=A^r\cup A^c$ 为属性集合，A^r 为数值条件属性集合，A^c 为符号条件属性集合；$V=\bigcup_{a\in A}V_a$，V_a 为属性 a 的值域；$f:U\times A\to V$ 是一个信息函数，对于任意 $x\in U,a\in V$，有 $f(x,a)\in V_a$。

为了讨论方便，MDT 也可以表示为 $NDT=(U,A^r,V,f)$ 和 $CDT=(U,A^c,V,f)$ 的组合，NDT 称为数值数据系统，CDT 称为符号数据系统。

2.2 最大最小距离算法

最大最小距离法是一种基于欧氏距离的试探性算法，它可以有效避免 k-means 算法中聚类中心过近的情况，同时可以提高聚类的效率。算法的基本思想是在样本集中以最大距离原则选取聚类中心，以最小距离原则对样本进行划分，算法主要步骤如下。

Step 1：在样本集 U 中随机选取一个样本作为初始聚类中心 Z_1，加入类中心集合 Z 中；

Step 2：在剩余的样本中找到一个与 Z_1 距离最远的样本作为第二个聚类中心 Z_2，加入集合 Z 中；

Step 3：将剩余数据集中的每一个样本 x_i 与现有的聚类中心的最小距离记为 D_i，将最大的 D_i 所对应的 x_i 作为新的聚类中心加入集合 Z 中；

Step 4：重复 Step 3，直到找到 k 个聚类中心。

最大最小距离算法是一种简单高效的算法。由于聚类结果依赖于初始聚类中心和聚类个数 k 的选取，而该算法中采用了随机选取初始聚类中心的方法，聚类个数 k 也需要根据经验事先给出，在没有先验知识指导的情况下，聚类结果并不理想。同时，由于最大最小距离算法是利用欧氏距离来度量样本的相似性，因此无法对包含数值属性和符号属性的混合数据进行聚类。

3 一种基于先验信息的混合数据聚类个数确定算法

基于上述分析，本节首先提出了利用先验信息确定初始类中心的方法，同时介绍了一种混合数据相似性度量方法以及聚类有效性评价指标，并在此基础上提出了一种基于先验信息的混合数据聚类个数确定算法。

3.1 初始聚类中心选取

假设在样本集 U 中，有类标签信息的样本构成集合 U'，没有类标签信息的样本构成集合 U''，那么 $U=U'\cup U''$。在 U' 中，根据类标签信息进行分类，得到对 U' 的一个划分：

$$U'=\{C_1,C_2,\cdots,C_{k_{\min}}\}$$

在 U' 的每一类中计算聚类中心 $Z=\{Z_1,Z_2,\cdots,Z_{k_{\min}}\}$，其中第 l 类 c_l 的聚类中心为 $Z_l=(z_{l1},z_{l2},\cdots,z_{lm})$，$1\leqslant l\leqslant k_{\min}$，如果 $a_j\in A^r$，则

$$z_{lj}=\frac{1}{|c_l|}\sum_{x_i\in c_l}x_{ij}$$

如果 $a_j\in A^c$，则

$$z_{lj}=a_j^{(q)}\in V_{a_j}$$

其中，$|\{w_{li}\mid x_{ij}=a_j^{(q)},w_{li}=1\}|\geqslant|\{w_{li}\mid x_{ij}=a_j^{(t)},w_{li}=1\}|$，$1\leqslant t\leqslant n_j,t\neq q,V_{a_j}=\{a_j^{(1)},a_j^{(2)},\cdots,a_j^{(n_j)}\}$ 为 a_j 的值域，n_j 为 a_j 属性值的个数，如果 x_i 属于第 l 类，$w_{li}=1$，否则等于 0。

3.2 混合数据相似性度量

由于最大最小距离算法是基于欧氏距离的一种聚类方法，无法对混合数据进行聚类，因此在生成新的聚类中心时，需要一种新的针对混合数据样本进行相似性度量的方法。本文采用了如下度量方法：

$$D_A(x,y)=\frac{|A^r|}{|A|}D_{A^r}+\frac{|A^c|}{|A|}D_{A^c}\tag{1}$$

其中，

$$D_{A^c}(x,y)=\sum_{a\in A^c}d_a(x,y)$$

$$d_a(x,y)=\begin{cases}0, & f(x,a)=f(y,a)\\ 1, & f(x,a)\neq f(y,a)\end{cases}$$

$$D_{A^r}(x,y)=\sum_{a\in A^r}(f(x,a)-f(y,a))^2$$

$|\cdot|$ 为属性集中属性的个数。

3.3 基于先验信息的混合数据聚类个数确定算法

基于上述分析，本文提出了一种基于先验信息的混合数据聚类个数确定算法。该算法的主要思想是首先利用已有的先验信息确定部分初始聚类中心及最小聚类个数；然后基于最大最小距离原则，逐渐在无标签样本中生成新的类中心，直到达到最大聚类个数，在每添加一个类中心后，按照最近距离原则，对无标签样本进行划分获得聚类结果；最后利用评价指标对聚类结果进行评价来确定数据集真实的类个数。在本算法中，使用如下评价指标：

$$CUM(C^k)=\frac{|A^r|}{|A|}CUN(C^k)+\frac{|A^c|}{|A|}CUC(C^k)\tag{2}$$

$$CUN(C^k)=\frac{1}{k}\sum_{l=1}^{|A^r|}(\delta_l^2-\sum_{j=1}^{k}p_j\delta_{jl}^2)$$

$$CUC(C^k)=\frac{1}{k}\sum_{a\in A^c}Q_a$$

其中，

$$\delta_l=\sum_{x\in U}(f(x,a_l)-m_l)^2/|U|$$

$$m_l=\frac{1}{|U|}\sum_{x\in U}f(x,a_l)$$

$$\delta_{jl}=\sum_{x\in C_l}(f(x,a_l)-m_{jl})^2/|C_j|$$

$$m_{jl}=\frac{1}{|C_j|}\sum_{x\in C_j}f(x,a_l)$$

$$p_j=\frac{|C_j|}{|U|}$$

$$Q_a=\sum_{X\in U/IND(\{a\})}\sum_{j=1}^{k}\frac{|C_j|}{|U|}\left(\frac{|X\cap C_j|^2}{|C_j|^2}-\frac{|X|^2}{|U|^2}\right)$$

利用这一指标，对聚类结果进行评价，找出最佳聚类个数。

具体算法步骤如下。

Step 1：输入混合数据信息系统 MDT=(U,A,V,f)，$k_{max}=\sqrt{|U|}$；

Step 2：在集合 U′中，利用标签信息计算初始聚类中心 $Z=\{Z_1,Z_2,\cdots,Z_{k_{min}}\}$，并得到 k_{min}；

Step 3：将 U″中每一个元素划分到与其距离最小的聚类中心所代表的类中，并计算 $CUM(C^{k\_min})$；

Step 4：For i=k_{min} to k_{max}

Step 4.1：根据式(1)，找出 U″中每一个样本与集合 Z 中所有聚类中心的最小距离，记为 D_i；

Step 4.2：将值最大的 D_i 所对应的样本作为新的聚类中心，加入集合 Z 中；计算 U″中每一个样本与 Z 中所有聚类中心的距离，将该样本划分到与其距离最近的一类中；

Step 4.3：计算 $CUM(C^i)$；

i=i+1;

Step 5:$k=\underset{i=k_{min}}{\overset{k_{max}}{\operatorname{argmax}}}\ CUM(C^i)$;

Step 6:输出 k。

在算法中耗时最多的是 Step 3,每一次循环所耗时间最长为计算 D_j,需要计算 $|U''|\times\frac{k(k+1)}{2}-\frac{k(k+1)(2k+1)}{2}$ 次,由于 $k\leqslant\sqrt{|U|}$,假设 U 中只有少数样本具有标签信息,则 $|U''|\approx|U|$,那么每次循环最坏情况下的时间复杂度为 $O(|U|^2|A|)$,整体算法的时间复杂度为 $O((k\_max-k\_min+1)|U|^2|A|)$。

4 实验结果及分析

本文从 UCI 数据集中选取了 4 个数据集,从以下 3 方面进行了实验。

首先为了验证算法的有效性,选取了数据集 Zoo。Zoo 数据集有 16 个属性,其中有 1 个数值属性,15 个符号属性,包含 101 个样本,分为 7 类。在该数据集上做了 3 次实验,每次随机保留 10%样本的类标签,这些有类标签的样本分别可以分为 4 类、5 类、6 类,实验结果如图 1—图 3 所示。

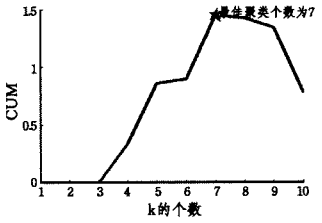

图 1 数据集 Zoo 先验类标签信息为 4 类

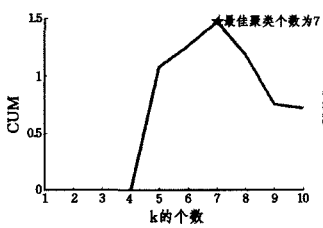

图 2 数据集 Zoo 先验类标签信息为 5 类

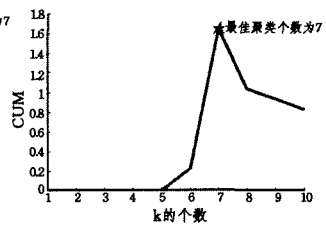

图 3 数据集 Zoo 先验类标签信息为 6 类

实验证明,在先验类标签信息没有覆盖到所有类时,使用该算法可以有效地找到正确的聚类个数。

其次,为了验证算法在先验类标签信息极少时的有效性,选取了数据集 CMC (*Contraceptive Method Choice*),该数据集有 8 个属性,其中有 2 个数值属性和 6 个符号属性,包含 1473 个样本,分为 3 类。在该数据集上,进行了 4 次实验,分别随机保留 5%、1%、0.5%、0.1%样本的类标签,这些有标签的样本根据类标签可以分为 2 类,实验结果如图 4—7 所示。

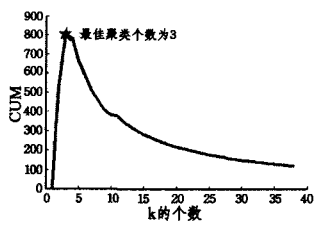

图 4 CMC 保留 5%样本的类标签

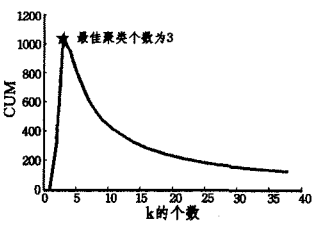

图 5 CMC 保留 1%样本的类标签

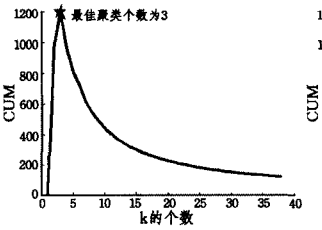

图 6 CMC 保留 0.5%样本的类标签

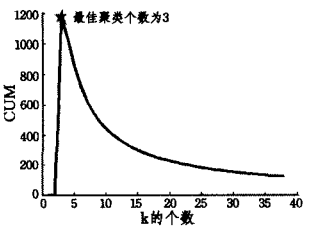

图 7 CMC 保留 0.1%样本的类标签

实验证明,当先验信息极少时该算法也可以有效地找到正确的聚类个数。

最后,为了验证该算法对于只包含数值属性或者符号属性样本集的有效性,选取了数据集 Wine 和 Balance。数据集 Wine 有 13 个数值属性,包含 178 个样本,分为 3 类。实验中保留 1%样本的类标签,实验结果如图 8 所示。

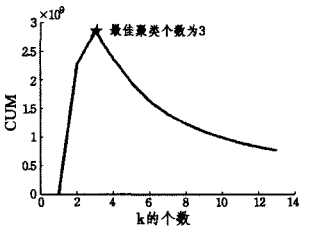

图 8 Wine 保留 1%样本的类标签

数据集 Balance 有 4 个符号属性,包含 625 个样本,分为 3 类。实验中保留 1%样本的类标签,实验结果如图 9 所示。

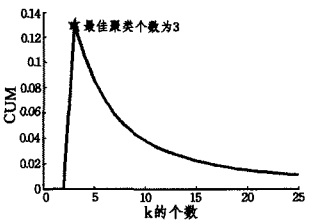

图 9 Balance 保留 1%样本的类标签

实验证明,对于只包含数值属性或者符号属性的样本集,该算法同样可以有效找到正确的聚类个数。

结束语 本文针对现有聚类个数确定方法中,由于随机选取初始聚类中心产生的算法执行效率不稳定的情况,利用先验信息初始化聚类中心,提出了一种基于先验信息的混合数据聚类个数确定算法,较好地解决了这一问题。并在 UCI 数据集中选取了 Zoo、CMC、Wine 和 Balance 等 4 个数据集,从不同方面对算法进行了有效性验证,实验证明该算法是有效的。

参 考 文 献

[1] MacQueen J B. Some methods for classification and analysis of multivariate observations [C] // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297

[2] Ruspini E R. A new approach to clustering [J]. Information and Control, 1969, 15(1): 22-32

[3] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38

[4] Camastra F, Verri A. A novel kernel method for clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 801-805

[5] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm

for discovering clusters in large spatial databases with noise [C]// Proceedings of the 2th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1996,226-231

[6] Sun Hao-jun, Wang Sheng-rui, Jiang Qing-shan. FCM-based model selection algorithms for determining the number of clusters[J]. Pattern Recognition,2004,37(10):2027-2037

[7] Bai Liang, Liang Ji-ye, Dang Chuang-yin. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data[J]. Knowledge-Based Systems,2011,24(6):785-795

[8] Liang Ji-ye, Zhao Xing-wang, Li De-yu, et al. Determining the number of clusters using information entropy for mixed data [J]. Pattern Recognition,2012,45:2251-2265

[9] Tou J, Gonzales R. Pattern Recognition Principles[M]. MA: Addison-Wesley. Reading,1974

[10] Pal N R, Bezdek J C. On clustering validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy Systems,1995,3(3):370-379

[11] Xiao Yu, Yu Jian. Semi-Supervised Clustering Based on Affinity Propagation Algorithm[J]. Journal of Software,2008,19(11):2803-2813(in Chinese)
肖宇,于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报,2008,19(11):2803-2813

[12] Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering [C]// Russ G, Dale S, eds. Proc. of the 21st Int'l Conf. on Machine Learning (ICML 2004). Banff: ACM Press,2004:81-88

[13] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding[C]// Claude S, Achim GH, eds. Proc. of 19th Int'l Conf. on Machine Learning (ICML 2002). Sydney: Morgan Kaufmann Publishers,2002:27-34

[14] Kamvar S D, Klein D, Manning C D. Spectral learning[C]// Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2003). Acapulco, Mexico: Morgan Kaufmann Publishers, 2003:561-566

(上接第 85 页)

$\delta_A^\beta(U,D)$,$G_\beta(B|D)=G_\beta(A|D)$,所以 $B=\{A_2,A_3\}$ 为 $\beta=1/2$ 时粒度集 A 的一个粒度约简;同理可计算出当 $\beta=1$ 时,$\delta_A^\beta(U,D)=\{\{x_6\},\emptyset\}$,当 $B=\{A_3\}$ 时,$\delta_B^\beta(U,D)=\{\{x_6\},\emptyset\}=\delta_A^\beta(U,D)$,$G_\beta(B|D)=G_\beta(A|D)$,所以 $B=\{A_3\}$ 为 $\beta=1$ 时粒度集 A 的一个粒度约简。

实例表明,在计算变精度悲观多粒度粗糙集的粒度约简时,针对给定的精度 β,首先计算粒度集的核心粒度,然后以下近似分布粒度熵的变化作为启发式信息,逐个选择粒度重要度最大的粒度加入到核中,最终求得的粒度约简与原始多粒度空间在相同精度下具有同样的决策能力。在本例中,当 $\beta=1/2$ 时,$B=\{A_2,A_3\}$ 为粒度集 A 的一个约简,评价指标 A_1 可忽略,当 $\beta=1$ 时,$B=\{A_3\}$ 为粒度集 A 的一个约简,评价指标 A_1、A_2 可忽略。

结束语 本文针对变精度悲观多粒度粗糙集模型的粒度约简进行了研究,设计了变精度悲观多粒度粗糙集粒度约简算法,并通过实例验证了该算法的有效性,针对不同的精度,计算粒度集的约简,有利于进一步从变精度悲观多粒度粗糙集中提取更加简洁的决策规则,这为变精度多粒度粗糙集的应用提供了理论基础。

参 考 文 献

[1] Pawlak Z. Rough set[J]. International Journal of Computer and Information Science,1982,11:341-356

[2] Lin T Y. Granular computing on binary relations II: Rough set representations and belief functions [M] // Rough Sets and Knowledge Discovery. 1998:122-140

[3] Qian Yu-hua, Liang Ji-ye. Rough set method based on multi-granulations[C]// Proceeding of the Fifth IEEE International Conference on Cognitive Informatics. Beijing, China, July 2006: 297-304

[4] Qian Yu-hua, Liang Ji-ye, Yao Yi-yu, et al. MGRS: A multigranulation rough set[J]. Information Sciences,2010,180:949-970

[5] Qian Yu-hua, Liang Ji-ye, Wei Wei. Pessimistic rough decision [C]// Second International Workshop on Rough Sets Theory. 2010:440-449

[6] Yang Xi-bei, Dou Hui-li, Yang Jing-yu. Hybird Multigranulation Rough Sets Based on Equivalence Relations[J]. Computer Science,2012,30(11):165-169(in Chinese)
杨习贝,窦慧莉,杨静宇. 基于等价关系的混合多粒度粗糙集[J]. 计算机科学,2012,30(11):165-169

[7] Zhang Ming, Tang Zhen-min, Xu Wei-yan, et al. Variable Multi-granulation Rough Set Model[J]. Pattern Recognition and Artificial Intelligence,2012,25(4):709-720(in Chinese)
张明,唐振民,徐维艳,等. 可变多粒度粗糙集模型[J]. 模式识别与人工智能,2012,25(4):709-720

[8] Sang Yan-li, Qian Yu-hua. A Granular Space Reduction Approach to Pessimistic Multi-Granulation Rough Sets[J]. Pattern Recognition and Artificial Intelligence,2012, 25(3):361-366(in Chinese)
桑妍丽,钱宇华. 一种悲观多粒度粗糙集中的粒度约简算法[J]. 模式识别与人工智能,2012,25(3):361-366

[9] Liu Cai-hui. Covering-based Multigranulation Rough Set Model Based on Maximal Description of Elements[J]. Computer Science,2013,40(12):64-67(in Chinese)
刘财辉. 一种元素最大描述下的多粒度覆盖粗糙集模型[J]. 计算机科学,2013,40(12):64-67

[10] Qian Yu-hua, Zhang Hu, Sang Yan-li, et al. Multigranulation decision-theoretic rough sets[J]. International Journal of Approximate Reasoning,2014,55(1):225-237

[11] Dou Hui-li, Wu Chen, Yang Xi-bei, et al. Variable Precision Multigranulation Rough Sets[J]. Journal of Jiangsu University of Science and Technology,2012,26(1):65-69(in Chinese)
窦慧莉,吴陈,杨习贝,等. 可变精度多粒度粗糙集模型[J]. 江苏科技大学学报,2012,26(1):65-69

[12] Zhai Yong-jian, Zhang Hong. Reduction of Variable Precision Multi-granulation Rough Sets[J]. Journal of Jinling Institute of Technology,2013,29(4):1-8(in Chinese)
翟永健,张宏. 变精度多粒度粗糙集的约简研究[J]. 金陵科技学院学报,2013,29(4):1-8

[13] Ziarko W. Variable precision rough set model [J]. Journal of Computer and System Sciences,1993,46(1):39-59

[14] Zhang Wen-xiu, Mi Ju-sheng, Wu Wei-zhi. Knowledge Reductions in Inconsistent Information Systems[J]. Chinese Journal of Computers,2003,26(1):12-18(in Chinese)
张文修,米据生,吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报,2003,26(1):12-18