基于近邻传播算法的最佳聚类数确定方法比较研究

周世兵[1] 徐振源[1,2] 唐旭清[2]
(江南大学信息工程学院 无锡 214122)[1] (江南大学理学院 无锡 214122)[2]

摘 要 在聚类分析中,决定聚类质量的关键是确定最佳聚类数。提出采用聚类效果较好的近邻传播聚类算法对样本进行聚类,运用6种聚类有效性指标分别对聚类结果进行有效性分析,以确定最佳聚类数。具体分析了这些有效性指标,并改进了IGP指标确定最佳聚类数的方法。针对8个数据集,通过实验比较这些指标的性能。分析和实验结果表明,基于近邻传播聚类算法,IGP指标确定最佳聚类数的性能最好。

关键词 近邻传播,聚类数,聚类有效性指标,聚类分析
中图法分类号 TP18 **文献标识码** A

Comparative Study on Method for Determining Optimal Number of Clusters Based on Affinity Propagation Clustering

ZHOU Shi-bing[1] XU Zhen-yuan[1,2] TANG Xu-qing[2]
(School of Information Technology, Jiangnan University, Wuxi 214122, China)[1]
(School of Science, Jiangnan University, Wuxi 214122, China)[2]

Abstract It is crucial to determine optimal number of clusters for the quality of clustering in cluster analysis. Based on Affinity Propagation clustering algorithm, a method for determining optimal number of clusters was proposed to analyze the clustering validity and determine optimal number of clusters by using six clustering validity index. These clustering validity indexes were analyzed concretely and the method of using IGP index to determine optimal number of clusters was improved. In connection with eight datasets, the performances of these indexes were compared by simulation experiments. The results of analysis and experiments show that IGP index is the best to determine optimal number of clusters based on Affinity Propagation clustering.

Keywords Affinity propagation, Number of clusters, Clustering validity index, Cluster analysis

聚类算法是一种有效的数据分析方法。聚类算法是在没有任何样本的先验信息条件下对样本进行聚类分析的,这类算法又称为无监督学习方法。在聚类分析中,决定聚类质量的关键是确定最佳聚类数。目前,大部分聚类算法需要预先给定聚类数,才能对样本进行聚类分析。而如何得到正确的最佳聚类数,一直是聚类有效性研究的重要课题。近年来在《Science》中由Frey等人提出的一种新的聚类算法,称为近邻传播聚类算法(Affinity Propagation clustering,简称AP算法)[1]。与以往的聚类方法相比,此方法可以更快地处理大规模数据,得到较好的聚类结果。该算法通过数据点之间的消息传递产生高质量的聚类中心,避免聚类中心的初始选择。文献[1]中将近邻传播聚类算法应用在人脸图像聚类、基因表达数据的基因识别、手写体字符识别、最优航空路线确定等问题上。实验结果表明,近邻传播聚类算法在很短的时间内就能得到K中心算法花费很长时间才能达到的聚类结果[1,2]。近邻传播聚类算法不能直接将指定类数K作为算法的输入参数,以使算法产生K个聚类的聚类结果。对于类内紧密、类间远离的聚类结构,近邻传播聚类算法通过设定偏向参数p,可以得到比较准确的聚类结果;但对于比较松散的聚类结构,算法倾向于产生较多的局部聚类,这使得算法产生的聚类数往往偏多,而不能给出准确的聚类结果。要获得指定类数(例如K个聚类)的聚类结果,一般采用搜索的方法。文献[1]给出了一种通过对分法搜索近邻传播聚类算法,实现指定类数的聚类分析。目前,现有的有效性评价指标中,有很多可以用来分析聚类结果并确定最佳聚类数。然而在使用近邻传播聚类算法进行聚类分析方面,尚未有文献专门讨论哪些评价指标或方法更适用并具有更好的性能。因此,对近邻传播聚类算法的聚类结果进行有效性分析,并确定样本的最佳聚类数具有重要的意义。

1 近邻传播聚类算法

近邻传播聚类(AP)[1,3]算法是一种基于近邻信息传播的聚类算法,其目的是找到最优的类代表的集合,使得所有样本到最近的类代表的相似度之和最大。AP算法首先将数据集

到稿日期:2010-03-26 返修日期:2010-06-25 本文受国家863计划项目(2007AA1Z158),国家自然科学基金(60703106)资助。
周世兵(1972-),男,博士生,讲师,主要研究方向为人工智能、模式识别、生物信息学,E-mail:worldguard@sina.com;徐振源(1946-),男,教授,博士生导师,主要研究方向为混沌、同步控制、人工智能、生物信息学;唐旭清(1963-),男,博士,副教授,硕士生导师,主要研究方向为计算智能、生物信息学。

的所有 N 个样本都视为候选的类代表，为每个样本建立与其它样本的吸引程度的信息，即任意两个样本 x_i 和 x_k 之间的相似度(采用欧式距离为测度时，$s(i,k)=-||x_i-x_k||^2$)被存储在 $N\times N$ 的相似度矩阵中。AP 算法用 $s(i,k)$ 表示样本 x_k 在多大程度上适合作为样本 x_i 的类代表。AP 算法初始假设所有样本被选中成为类代表的可能性相同，即设定所有 $s(k,k)$ 为相同值 p。AP 算法为选出合适的类代表而不断从样本中搜集有关证据，为此，AP 算法引入了两个重要的信息量参数，即可信度 r 和可用度 a，两个信息量代表了不同的竞争目的。$r(i,k)$ 是从 x_i 指向 x_k，它代表 x_k 积累的证据，用来表示 x_k 适合作为 x_i 的类代表的代表程度；$a(i,k)$ 是从 x_k 指向 x_i，它代表 x_i 积累的证据，用来表示 x_i 选择 x_k 作为类代表的合适程度。对于任意样本 x_i，计算所有样本的可信度 $r(i,k)$ 和可用度 $a(i,k)$ 之和，则两者之和最大的样本 x_k 为类代表。AP 算法的迭代过程就是这两个信息量交替更新的过程。其中，可信度 $r(i,k)$ 定义为：

$$r(i,k)\leftarrow s(i,k)-\max_{k'\ s.t.\ k'\neq k}\{a(i,k')+s(i,k')\}$$

可用度 $a(i,k)$ 定义为：

$$a(i,k)\leftarrow \min\{0,r(k,k)+\sum_{i'\ s.t.\ i'\neq\{i,k\}}\max\{0,r(i',k)\}\}$$

另外，自可用度 $a(k,k)$ 定义为：

$$a(k,k)\leftarrow \sum_{i'\ s.t.\ i'\neq k}\max\{0,r(i',k)\}$$

AP 算法的突出优点是，它能在较短的时间里得到很好的聚类结果，且不需要初始化聚类中心，这使得 AP 算法比较稳定。AP 算法的另一个优点在于它对相似度矩阵的对称性没有要求，这扩大了 AP 算法的应用范围。

2 聚类有效性指标与最佳聚类数的确定

运用 AP 算法搜索聚类空间能够输出一系列具有不同聚类数目的聚类结果，对聚类结果进行分析并确定最佳聚类数属于聚类有效性问题。聚类有效性是指评价聚类结果的质量并确定最适合特定数据集的划分。通常采用聚类有效性指标来评价聚类算法产生的哪个聚类结果是最优的，并将最优的聚类结果所对应的聚类数目作为最佳聚类数。现有文献中可用于确定最佳聚类数的有效性指标主要包括：Calinski-Harabasz 指标、Davies-Bouldin 指标、Weighted inter-intra 指标、Krzanowski-Lai 指标、Hartigan 指标和 In-Group Proportion 指标等。

2.1 Calinski-Harabasz(CH)指标

CH 指标是基于全部样本的类内离差矩阵和类间离差矩阵的测度，其最大值对应的类数作为最佳聚类数。该指标不适用于聚类数为 1 的情况。设 k 表示聚类数，$trB(k)$ 与 $trW(k)$ 分别表示类间离差矩阵的迹和类内离差矩阵的迹。CH 指标定义为：

$$CH(k)=\frac{trB(k)/(k-1)}{trW(k)/(n-k)}$$

2.2 Davies-Bouldin(DB)指标

DB 指标是基于样本的类内散度与各聚类中心间距的测度，进行类数估计时其最小值对应的类数作为最佳聚类数。该指标不适用于聚类数为 1 的情况。设 W_i 表示聚类 C_i 的所有样本到其聚类中心的平均距离，C_{ij} 表示聚类 C_i 和聚类 C_j 中心之间的距离，则 DB 指标定义为：

$$DB(k)=\frac{1}{k}\sum_{i=1}^{k}\max_{j=1\sim k,j\neq i}\left(\frac{W_i+W_j}{C_{ij}}\right)$$

2.3 Weighted inter-intra(Wint)指标

Wint 指标[4]的目标是最大化类内相似度和最小化类间相似度。通常采用带罚项的 $(1-2k/n)$ Wint 指标进行类数估计，其最大值对应的类数作为最佳聚类数。该指标不适用于聚类数为 1 的情况。设 $inter(i,j)$ 表示类内相似度，$intra(i)$ 表示类间相似度，Wint 指标定义为：

$$Wint(k)=1-\frac{1}{\sum_{i=1}^{k}n_i\cdot intra(i)}\sum_{i=1}^{k}\frac{n_i}{n-n_i}\sum_{j=1,j\neq i}^{k}n_j\cdot inter(i,j)$$

2.4 Krzanowski-Lai(KL)指标

KL 指标是基于全部样本的类内离差矩阵的测度，其最大值对应的类数作为最佳聚类数。该指标不适用于聚类数为 1 的情况。设 $trW(k)$ 表示类内离差矩阵的迹，KL 指标定义为：

$$KL(k)=\frac{|Diff(k)|}{|Diff(k+1)|}$$

$$Diff(k)=(k-1)^{2/p}trW(k+1)-k^{2/p}trW(k)$$

2.5 Hartigan(Hart)指标

Hart 指标可以用于聚类数为 1 的情况，其满足 $Ha\leqslant 10$ 的最小类数作为最佳聚类数。设 $trW(k)$ 表示类内离差矩阵的迹，Hart 指标定义为：

$$Hart(k)=\left(\frac{trW(k)}{trW(k+1)}-1\right)(n-k-1)$$

2.6 In-Group Proportion(IGP)指标

IGP 指标[5]用来衡量在某一类中距离每个样本最近的样本是否在同一类中。所有聚类的平均 IGP 指标越大表示聚类的质量越好，其最大值对应的类数为最佳聚类数。该指标不适用于聚类数为 1 的情况。设 j^N 表示距离样本 j 最近的样本，$Class(j)$ 表示样本 j 的类标，$\#$ 表示满足条件的个数。对于类标为 u 的聚类，IGP 指标定义为：

$$IGP(u)=\frac{\#\{j|Class(j)=Class(j^N)=u\}}{\#\{j|Class(j)=u\}}$$

2.7 有效性指标确定最佳聚类数分析

这里结合有效性指标的定义和我们使用有效性指标的经验，对以上 6 种有效性指标确定最佳聚类数的过程进行具体分析。

2.7.1 IGP 指标确定最佳聚类数

在这些有效性指标中，IGP 指标是最近提出的指标，它使用类内数据点的 in-group 比例来衡量聚类结果的质量，具有较优的性能，特别适合对层次聚类算法和 K 近邻聚类算法的聚类结果进行评估和最佳聚类数的确定。由于 IGP 指标是基于概率统计思想提出的有效性指标，从粒度计算观点来看，是对聚类结果的粗粒度估计。在研究过程中我们发现，对有些数据集进行聚类搜索的过程中，会出现多个不同类数的平均 IGP 指标值都是 1 的情况。根据 IGP 指标的定义，IGP 指标的最大值为 1，若出现多个平均 IGP 指标值都为 1 的情况，说明指定的这些不同类数的聚类效果很好。在聚类效果相同的情况下，类数越多说明划分越细，聚类质量越高。因此，当出现多个平均 IGP 指标值都为 1 的情况时，我们取相同指标值中的类数最大值作为最佳聚类数。而其它有效性指标都是基于几何的指标，针对相同数据集，不同类数出现相同指标值的概率几乎为 0。

2.7.2 其它指标确定最佳聚类数

KL 指标适合聚类结构比较容易判别的情况，如相距远

的分离的聚类结构。Wint 指标的目标是最大化类内相似度和最小化类间相似度，对于一般数据集，其评估能力欠佳；对于基因表达数据集，由于该指标和 Pearson 距离有一定的关联度，其评估能力和确定最佳聚类数的能力较好。通常，Hart 指标的评估能力较差，但可以用 Hart 指标来判断聚类数是否为 1。DB 指标的评估能力欠佳，仅适合完全分离的聚类结构。CH 指标的评估能力不稳定，随着最佳聚类数搜索范围的变化，CH 指标得到的最佳聚类数会发生变化，并且随着搜索范围增大，CH 指标得到的最佳聚类数有逐渐增大的趋势。

2.7.3 聚类算法与有效性指标确定最佳聚类数

有效性指标方法是基于数据的类内与类间的离差矩阵或相似度的明显差别对聚类结果进行评估的。有效性指标能否得到正确的最佳聚类数，除了和有效性指标本身有关，还与所采用的聚类算法有很大关系。比如对本文实验中的人工数据集 Kes3 采用 AP 算法进行聚类分析，再使用 IGP 指标可以得到正确的最佳聚类数；而采用 K-均值聚类算法，IGP 指标却无法得到正确的最佳聚类数。由于传统的 K-均值聚类算法采用随机的方法确定初始聚类中心，使得聚类结果不稳定，采用有效性指标进行评估，得到的最佳聚类数也是不确定的，因此传统的 K-均值聚类算法不适合作为确定最佳聚类数的有效算法。而 AP 算法不需要初始化聚类中心，算法比较稳定，并且 AP 算法的聚类效果较好，因此采用 AP 算法作为确定最佳聚类数的聚类分析算法比较合适。

3 实验与分析

本文使用 AP 算法对 8 个已知正确类数的数据集进行聚类分析，并运用以上 6 种有效性评价指标对聚类结果进行评估和确定最佳聚类数。这 8 个数据集包括一般数据集和基因表达数据集。其中，有 3 个 UCI 标准数据集，分别是 Ionosphere，BUPA，Breast-cancer-wisconsin(本文简称 Bcw)，来源于 UCI Machine Learning Repository。Model2 数据集来源于文献[6]，Y14c 数据集来源于文献[7]。Kes2(其分布结构见图 3)是二维两类的人工数据集，Kes3(其分布结构见图 4)是二维三类的人工数据集，这两个数据集的共同特征是有一个类中的样本较多，并且部分类内样本之间的距离大于类间样本之间的距离。Leukemia 数据集是基因表达数据集，来源于文献[8]，表示白血病的基因表达数据。实验中聚类数的搜索范围为$[2,k_{max}]$，根据普遍使用的经验规则 $k_{max}\leqslant\sqrt{n}$，取 $k_{max}=\mathrm{Int}(\sqrt{n})$，其中 n 为样本数。

对于一般数据集，两个样本 x_i 和 x_k 之间的相似度采用欧式距离为测度，即 $s(i,k)=-||x_i-x_k||^2$。对于基因表达数据集，采用普遍使用的 Pearson 相关系数作为相似性测度，即两个样本之间的线性相关系数为：

$$R(i,k)=\frac{\sum_{m=1}^{d}(x_{im}-\overline{x_i})(x_{km}-\overline{x_k})}{\sqrt{\sum_{m=1}^{d}(x_{im}-\overline{x_i})^2}\sqrt{\sum_{m=1}^{d}(x_{km}-\overline{x_k})^2}}$$

为避免负数引起计算混乱，将 $R(i,k)\in[-1,1]$进行转换，使 $R(i,k)=1-(1+R(i,k))/2$，从而使 Pearson 相关系数转换为正的 Pearson 距离。$R(i,k)$越大表示两个样本相距越远，这样，基因表达数据的相似度表示为 $s(i,k)=-R(i,k)$。

基于 AP 算法，对标准数据集 BUPA 采用以上 6 种有效性指标得到的聚类结果如表 1 所列。其中带下划线的指标值所对应的聚类数为该列指标得到的最佳聚类数；Hart 指标所对应的列无下划线，表示该指标无法得到最佳聚类数。由于数据集 BUPA 的正确类数为 2，因此 KL 指标和 IGP 指标得到的最佳聚类数是正确的。

表 1 数据集 BUPA 的聚类有效性指标值

聚类数	CH	DB	Wint	KL	Hart	IGP
2	0.0309	0.3074	0.2111	3.2089	103.0362	0.9583
3	0.0089	0.2492	0.2665	1.3435	82.0595	0.8801
4	0.1957	0.2174	0.4738	2.1967	48.6593	0.8533
5	0.2369	0.2197	0.5312	0.9147	63.3009	0.8721
6	0.2871	0.2299	0.5211	1.4993	47.3124	0.8641
7	0.3529	0.1879	0.5839	1.6679	31.1627	0.8399
8	0.3941	0.1866	0.5092	1.0225	33.4066	0.8337
9	0.4424	0.1888	0.5096	1.4976	23.8214	0.8321
10	0.4804	0.1870	0.5055	1.4199	21.6422	0.7471
11	0.3104	0.1961	0.4837	1.2732	14.3851	0.7618
12	0.3262	0.1934	0.4833	0.8023	18.8926	0.7848
13	0.3479	0.1953	0.4843	0.9080	18.4936	0.7993
14	0.5944	0.1754	0.4871	1.6463	13.9692	0.8054
15	0.6233	0.1777	0.4848	0.8100	18.1423	0.7709
16	0.6625	0.1826	0.4811	1.6252	11.5197	0.7829
17	0.6889	0.1846	0.4785	0.8451	13.7813	0.7792
18	0.7498	0.1923	0.4584	0.8473	11.3623	0.7624

对数据集 Y14c 和 Kes3 的聚类结果运用 IGP 指标进行评估，出现多个 IGP 指标值为 1 的情况。采用本文方法，即针对同一个数据集，相同指标值中的最大聚类数为最佳聚类数，可以得到正确的结果。对数据集 Y14c 和 Kes3，运用 IGP 指标确定最佳聚类数的实验情况分别如图 1 和图 2 所示。从中可以看出，对于 Y14c 数据集，IGP 指标得到的最佳聚类数为 14；对于 Kes3 数据集，IGP 指标得到的最佳聚类数为 3。对于数据集 Y14c 和 Kes3，IGP 指标能够得到正确的最佳聚类数。

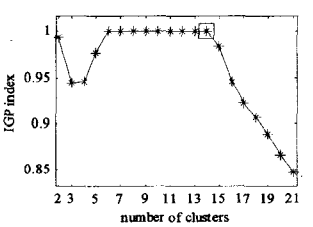

图 1 Y14c 的类数-IGP 指标关系图

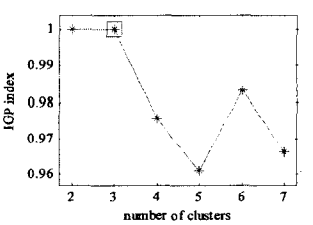

图 2 Kes3 的类数-IGP 指标关系图

针对以上 8 个数据集，6 种有效性评价指标估计出的最佳聚类数实验情况如表 2 所列。其中“—”表示无法得到最佳聚类数，带下划线的最佳聚类数表示正确的最佳聚类数。实验结果表明 IGP 指标的性能最优，对 8 个数据集都能得到正确的最佳聚类数；KL 指标次之，对 4 个数据集能得到正确的结果；效果最差的是 CH 指标，该指标对每个数据集都无法得到正确的最佳聚类数。

表 2 6 种有效性评价指标估计出的最佳聚类数

数据集	正确类数	CH	DB	Wint	KL	Hart	IGP
Ionosphere	2	6	18	4	6	9	2
BUPA	2	18	14	7	2	—	2
Bcw	2	7	10	8	7	—	2
Model2	3	10	9	3	3	—	3
Y14c	14	21	16	10	14	14	14
Kes2	2	4	3	4	4	5	2
Kes3	3	7	3	2	4	5	3
Leukemia	3	6	6	3	3	3	3

基于 AP 算法，设定数据集 Kes2 的类数为 2 类，Kes3 的类数为 3 类，数据集 Kes2 和 Kes3 的聚类结果分别如图 3 和图 4 所示。从中可以看出 AP 算法的聚类效果很好，而在提供正确类数的前提下，采用 K-means 算法或 FCM 算法，对这两个数据集仍然无法正确聚类。

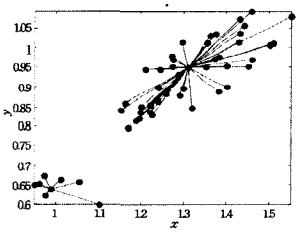

图 3　数据集 Kes2 的聚类结果

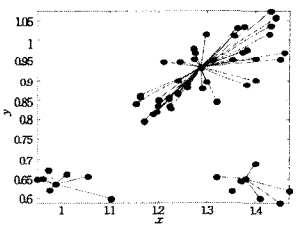

图 4　数据集 Kes3 的聚类结果

结束语　AP 算法能在较短的时间里得到很好的聚类结果，且不需要初始化聚类中心，这使得 AP 算法比较稳定，适合进行聚类分析。本文基于 AP 算法，研究了采用 6 种有效性指标来确定最佳聚类数的方法。通过分析和实验，可以得出这样的结论：与其它有效性指标相比，IGP 指标具有优良的性能，适合对 AP 算法的聚类结果进行评估并可用来确定最佳聚类数。运用聚类有效性指标确定最佳聚类数是一项重要的研究课题，需要我们作深入的研究，对聚类有效性指标的评估机理以及新的具有优良性能的有效性指标的研究，将是我们下一步研究工作的重点。

参 考 文 献

[1] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315(5814): 972-976

[2] Mézard M. Where Are the Exemplars? [J]. Science, 2007, 315(5814): 949-951

[3] 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

[4] 王开军，李健，张军英，等. 聚类分析中类数估计方法的实验比较[J]. 计算机工程，2008, 34(9): 198-202

[5] Kapp A V, Tibshirani R. Are clusters found in one dataset present in another dataset? [J]. Biostatistics, 2007, 8(1): 9-31

[6] Dudoit S, Fridlyand J. A Prediction-based Resampling Method for Estimating the Number of Clusters in a Dataset[J]. Genome Biology, 2002, 3(7): 1-21

[7] Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data[J]. Bioinformatics, 2003, 19(8): 973-980

[8] Armstrong S A, Staunton J E, Silverman L B, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia[J]. Nature Genetics, 2002, 30: 41-47

(上接第 224 页)

(2)对任意的 $q\in Q$，有 $Y\neq Y_{Q-q}$。

证明：假设 X,Y 是由 P,Q 导出的二进制粒矩阵，Y_{Q-q} 是由等价关系族 $Q-q$ 导出的二进制粒矩阵 $Y\neq Y_{Q-q}$，则

(1)$Q\subseteq P$ 是 P 的一个约简$\Leftrightarrow IND(P)=IND(Q)\Leftrightarrow H(P)=H(Q)\Leftrightarrow m=n, Y=X$;

(2) Q 独立$\Leftrightarrow Y\neq Y_{Q-q}\Leftrightarrow H(q/Q-\{q\})>0$。

从以上各定理证明过程可看到 Rough 集的代数定义、信息熵定义以及粒矩阵定义基本等价。

结束语　同一问题在不同知识表示下的算法难度不同。文献[10]探讨了粗糙性和信息熵的概念，文献[11]讨论了粗糙集理论代数观点与信息观点，分析了二者在相容信息系统中的等价关系，并发现二者在不相容信息系统中的不等价关系。文献[12]在粗糙集理论中提出了信息量和信息粒度的定义。本文通过定义粒矩阵和粒矩阵的运算来表示 Rough 集理论中的知识和知识运算，并证明了知识及知识运算在代数表示、信息表示和粒矩阵表示下的等价性。进一步，这种证明还可以推广到粗糙集的信息量、信息熵、包含度[13]的表示方法以及粗糙集的各种扩展模型的研究和计算中。相关研究还在继续。

参 考 文 献

[1] 王珏，袁小红，石纯一. 关于知识表示的讨论[J]. 计算机学报，1995, 18(3): 212-224

[2] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991

[3] 苗夺谦，王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报，1999(2): 113-116

[4] Zadeh L A. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems [J]. Soft Computing, 1998, 2: 23-25

[5] Lin T Y. Data Mining and Machine Oriented Modeling: A Granular Computing Approach[J]. Appl. Intell., 2000, 13(2): 113-124

[6] 刘澜，刘清. 基于粒的二进制运算的关联规则提取方法[J]. 南昌大学学报，2003, 27(1): 98-101

[7] 刘清. Rough 集及 Rough 推理[M]. 北京：科学出版社，2001

[8] 陈泽华，谢刚，谢珺，等. 基于粒计算的 Rough 集模型[J]. 计算机科学，2009, 36(5): 200-203

[9] 陈泽华，谢刚，谢珺，等. 粒矩阵及其在知识约简中的应用[J]. 计算机科学与探索，2010, 4(3): 283-288

[10] 苗夺谦，王珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论[J]. 人工智能与模式识别，1998, 11(1): 34-40

[11] 王国胤. Rough 集理论代数与信息论观点的关系研究[J]. 世界科技研究与发展，2002, 24(5): 20-26

[12] 梁吉业，钱宇华. 粗糙集理论中的不确定性与知识粒度[M]//张文修，姚一豫，梁怡. 粗糙集与概念格. 西安：西安交通大学出版社，2006

[13] 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