基于覆盖的粗糙聚类算法

王慎超　苗夺谦　陈　敏　王睿智
(同济大学计算机科学与工程系　上海　201804)

摘　要：传统的聚类算法大都得到了样本集的一个划分，类之间是严格的互斥关系，而现实世界中类与类之间往往没有明确的边界。该文将粗糙集理论引入到聚类分析中，提出了一种基于覆盖的粗糙聚类算法 KMMRSC，它用多个中心点代表一个类，并用上、下近似来刻画样本的归属，类与类之间是一种覆盖关系。实验结果表明，该算法聚类质量优于 k-均值算法，且能发现非球状簇。

关键词：粗糙聚类；覆盖；多中心点
中图分类号：TP181　　**文献标识码**：A　　**文章编号**：1009-5896(2008)07-1713-04

Overlapping-Based Rough Clustering Algorithm

Wang Shen-chao　Miao Duo-qian　Chen Min　Wang Rui-zhi
(*Department of Computer Science and Engineering*, *Tongji University*, *Shanghai* 201804, *China*)

Abstract: Most of traditional clustering algorithms get a partition of sample set with mutually exclusive classes, while there is no explicit boundary between classes mostly in the real world. Introducing rough set theory into clustering analysis, this paper proposes a kind of overlapping-based rough clustering algorithm called KMMRSC which represents a class with multiple centroids and describes the belongingness of samples with the concepts of upper approximation and lower approximation, thus there is overlapping relationship between classes. Experiments show that the algorithm KMMRSC, which can find non-spherical clusters, outperforms classic k-means.

Key words: Rough clustering; Overlapping; Multi-centroid

1　引言

聚类[1]是指把一组个体按照相似性归成若干类别，即“物以类聚”，它的目的是使属于同一类别的个体之间的相似度尽可能大，而不同类别的个体之间的相似度尽可能小。从机器学习的角度来看，聚类属于无指导学习，和分类不同，它不依赖于预先定义的类和带类标号的训练。聚类分析是知识发现的重要方法，在图像识别、信息检索、基因分析、客户关系管理等领域有着广泛的应用[2−4]。

经典的基于划分的聚类算法主要有 k-means 和 k-medoids 两种。前者是以簇中所有样本的均值为簇的代表，后者以簇中某个具有代表性的点作为簇的代表，通过反复迭代获得稳定的簇。两者均只能发现球状簇，且都是将每个待处理的对象严格地划分到某个类中，类之间是严格的互斥关系。而大量的事实说明，类与类之间不一定存在明确的边界，对象的类属存在着一定的过渡，即对象对于聚类簇并非只有属于和不属于两种状态。近年来，Lingras 等人[5, 6]将粗糙集理论引入到数据聚类中，取得了一些阶段性的成果，其主要思想是将聚类簇看成一种不确定集合，通过粗糙集中的上近似和下近似概念来描述聚类簇。

本文以能反映簇轮廓的多个中心点代表一个簇，从而能够发现非球状簇。同时引入上、下近似的概念，将聚类过程中缺乏明确归属的样本点归入距离相当的所有簇的上近似中，使得簇与簇之间是一种覆盖关系，聚类结果更加自然。实验过程中用簇间距离均值与簇内距离均值之比来检验聚类结果，使聚类质量的评价更有说服力。

2　粗糙集理论与粗糙聚类

2.1 粗糙集的概念

设非空对象集合 U 为论域。如果要在 U 上定义一个概念，那么它能由 U 的子集 X 表示。粗糙集理论的中心观点是集合的近似表示：任何在 U 上的集合概念 X 都能用它的下近似集合和上近似集合表示。

定义 1　R 是集合 U 上的二元关系，如果它是自反、对称和传递的，则它是 U 上的等价关系。

定义 2　设 R 为集合 U 上的等价关系，对任何 $x \in U$，下列集合

$$[x]_R = \{y \mid y \in U, xRy\} \tag{1}$$

称为元素 x 形成的 R 等价类。

定义 3　对知识库 $A=(U,R)$，$X \subseteq U$，X 的下近似和上近似集合如下：

2007-03-26 收到，2007-08-20 改回
国家自然科学基金(60475019)和教育部博士点基金(20060247039)资助课题

$$L_R(X)=\{x\in U\mid [x]_R\subseteq X\} \tag{2}$$

$$U_R(X)=\{x\in U\mid [x]_R\bigcap X\neq\varnothing\} \tag{3}$$

从定义 3 可以知道 $L_R\subseteq X\subseteq U_R$，于是 $\mathrm{BN}_R=U_R-L_R$，我们称为边界。直观上看，下近似集合中的对象肯定属于概念 X，上近似集合中的对象则可能属于概念 X，而 (L_R,U_R) 表示了概念 X 的一种粗糙近似关系，即粗糙集。

2.2 粗糙聚类

由上、下近似的概念不难想到，在聚类过程中，可用下近似和上近似分别描述聚类簇的最小轮廓和最大轮廓。这样，簇的下近似集成为簇的核心部分，反映了簇的位置和形状，而上近似中去掉下近似得到的边界部分由于可能同时属于多个簇，因此对于簇而言是相对次要的，属于补充。

粗糙聚类[5−8]算法主要有两类，一类是遗传粗糙聚类，另一类是粗糙 k-均值聚类。前者将粗糙集的上、下近似概念用到了遗传算法中，通过反复的复制、交叉、变异，寻求理想的聚类结果，在小样本集上取得了很好的效果，但由于时间复杂度相对较高，不适于处理大样本集。后者将对象划分到聚类簇的下近似和上近似中，通过下近似和上近似计算新的聚类中心，然后重新划分对象，此过程反复进行，直到形成稳定的聚类结果。但此类算法只能发现球状簇，且对输入参数敏感，可伸缩性较差。

3 基于覆盖的粗糙聚类算法

鉴于现有粗糙聚类算法的不足，本文提出一种基于覆盖的粗糙聚类算法。算法中用上、下近似描述样本归属，类之间是覆盖关系，聚类结果更加自然。通过随机抽样过程，避免了在整个大样本集上的反复迭代，降低了时间复杂度。同时以能反映簇轮廓的多个中心点代表一个簇，从而能够发现非球状簇。以下部分是这样安排的：3.1 节给出了与算法相关的一些定义，3.2 节展开介绍算法本身，3.3 节简单分析了算法的时间复杂度。

3.1 相关定义

在簇集 $C=\{C_1,C_2,\cdots,C_K\}$ 中，每个簇 C_j 由 M 个样本组成的代表点集合 $\mathrm{Rps}_j=\{\boldsymbol{R}_{j1},\boldsymbol{R}_{j2},\cdots,\boldsymbol{R}_{jM}\}$ 代表，其代表点由实际样本点组成，簇 C_j 的上、下近似分别表示为 $\overline{A}_j$，$\underline{A}_j$。样本 $\boldsymbol{X}_i$，$\boldsymbol{X}_j$ 间的相异度用欧氏距离 $d(\boldsymbol{X}_i,\boldsymbol{X}_j)=||\boldsymbol{X}_i-\boldsymbol{X}_j||^2$ 衡量。

定义 4 质心 簇的下近似集合中所有点的几何中心，即 $\mathbf{CTD}_j=\sum_{X_i\in\underline{A}_j}\boldsymbol{X}_i\Big/\left|\underline{A}_j\right|$，$\left|\underline{A}_j\right|$ 表示下近似集合的基数。

定义 5 收缩因子 将代表点集合 Rps_j 按因子 α 向簇心收缩，得到的集合为中心点集合 $\mathrm{Cen}_j=\{\boldsymbol{C}_{j1},\boldsymbol{C}_{j2},\cdots,\boldsymbol{C}_{jM}\}$。由此可见，该集合并不是由实际样本点组成。收缩公式为 $\boldsymbol{C}_{jt}=\boldsymbol{R}_{jt}+\alpha(\mathrm{CTD}_j-\boldsymbol{R}_{jt}),\alpha\in(0,1),t=1,2,\cdots,M$。

定义 6 点簇距 样本点与某簇的距离 $d_{i,j}(\boldsymbol{X},C)$ 等于点到该簇所有中心点中最近的一个距离，即 $d_{i,j}(\boldsymbol{X},C)=\min d(\boldsymbol{X}_i,\boldsymbol{C}_{jt}),t=1,2,\cdots,M$，称离点 $\boldsymbol{X}_i$ 最近的簇 C_m 为最近距离簇，该最近距离 $d_{i,m}=d(\boldsymbol{X}_i,C_m)=\min d_{i,j}(\boldsymbol{X},C),j=1,2,\cdots,K$ 称为最近簇距离。

定义 7 点簇距比值 样本 $\boldsymbol{X}_i$ 到簇 C_j，C_z 的距离之比，计算公式为 $Q(i,j,z)=d_{i,j}(\boldsymbol{X},C)/d_{i,z}(\boldsymbol{X},C)$。

定义 8 点心距 样本点到簇心的距离，用 $d(\boldsymbol{X}_i,\mathbf{CTD}_j)$ 表示。

定义 9 簇间距 簇之间的距离，用两簇几何中心间的欧氏距离衡量，即 $d_C(i,j)=d(\mathbf{CTD}_i,\mathbf{CTD}_j)$。

定义 10 凝聚度 用 AD 表示，它等于簇间距离均值除以簇内距离均值，用于评价聚类结果，值越大聚类结果越好，即

$$\mathrm{AD}=\frac{2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}d_C(i,j)\sum_{t=1}^{K}\left|\underline{A}_t\right|}{K(K+1)\sum_{t=1}^{K}\sum_{n\in\underline{A}_t}d(\boldsymbol{X}_n,\mathbf{CTD}_t)} \tag{4}$$

3.2 KMMRSC 算法

算法首先通过随机抽样和一步 k-均值迭代获得初始的样本集和簇。然后开始一个迭代过程：首先计算当前的簇心。然后根据簇心确定代表点集 Rps，同时根据收缩公式算出中心点集 Cen。接着计算凝聚度 AD，如果 AD 已收敛，算法结束；否则继续对所有的样本点重新确定其归属。待所有对象分配完毕后，如果已达到预定迭代次数，算法终止；否则开始新的迭代过程。在迭代过程中使用上、下近似分配样本点可以得到更自然的聚类结果。对于聚类算法来说，如何选取每个簇的代表对聚类质量的好坏有着十分重要的影响。因此，接下来首先详细阐述本文的中心点选取原则及步骤，然后给出算法的描述。

3.2.1 中心点的选取 为了能够得到更自然的聚类结果，发现任意形状的簇，本文采用多个中心点描述一个簇。用 M 表示每簇最少需要选取的中心点数，如果 M 选得过小，则不能有效地描述一个簇的轮廓；M 选得过大，又会大大提高算法的时间复杂度。对于给定的样本集，每个属性均有一个取值区间，在该属性上具有最大或最小值的样本携带了样本集在该维度上的最值，将其作为代表便能够获得该属性值的跨度区间。因此，本文给出如下结论：至少需要 $2N$ 个向量才能近似地描述 N 维空间中一个封闭连通区域的轮廓或范围。于是取 $M=2|A|$，其中 $|A|$ 表示样本空间的维数，即描述样本的属性的个数。

每次迭代过程选择代表点时，对每个簇 C_j，求各点到其质心的距离，取最远的一点作为 C_j 的第 1 个代表点 C_{j1}，然后求离 C_{j1} 最远的一点作为该簇第 2 个代表点 C_{j2}，接着求离前两个已有代表点距离之和最远的点作为第 3 个代表点，依次类推直至取满簇 C_j 的 M 个代表点。这样做的目的是期望取出的代表点能够覆盖样本所有属性的值域宽度，从而使这有限个代表点尽可能地捕获其所代表的簇的轮廓。在样本点分配时，点与簇的距离就可以用点与簇代表点间的距

离来表示。然而实验过程中发现，这样选取代表点会使簇的范围逐步扩张，簇与簇趋向重叠，从而导致聚类失败。为了既能发现任意形状的簇，又能获得良好的聚类效果，本文引入一个介于 0 与 1 之间的收缩因子 α，使选出的代表点以该比例向簇心收缩，从而得到中心点集 Cen，并以此来代表簇。

3.2.2 算法描述

输入　数据集 D，期望簇个数 K，阈值 ε，收缩因子 α。

输出　K 个聚类结果簇。

步骤 1　从 D 中随机抽取样本集 S，随机选取 K 个样本点作为簇心，并通过一步 k-均值迭代获得初始的簇，迭代计数器 $\text{iter}=0$；

步骤 2　计算簇心集 CTD；

步骤 3　计算代表点集 Rps 和中心点集 Cen；

步骤 4　计算凝聚度 AD。如果 AD 已收敛，转到步骤 7，否则转到步骤 5；

步骤 5　将各样本点重新分配到其对应的最近距离簇 C_m 中。对于每个样本点 $\boldsymbol{X}_i \in S$，计算点簇距 $d_{i,j}(\boldsymbol{X},C)$, $j=1,2,\cdots,K$，如果存在 $C_j \in C$ 且 $j \neq m$ 使得 $Q(i,j,m)<\varepsilon$ (即点 $\boldsymbol{X}_i$ 到簇 C_j 的距离与最近簇距离的比值不超过阈值 ε)，则将点 $\boldsymbol{X}_i$ 分配到最近距离簇和满足以上条件的所有簇的上近似中，否则将点只分配到最近距离簇 C_m 的下近似 $\underline{A}_m$ 中；

步骤 6　$\text{iter}=\text{iter}+1$，如果 $\text{iter}<\text{MAX\_RUN\_TIME}$，转到步骤 3，否则转到步骤 7；

步骤 7　算法终止。

3.3 算法复杂度分析

由于该算法不需要存储相异度矩阵，故空间复杂度为 $O(N)$，N 为数据集大小。设抽取的样本集大小为 $|S|$，迭代次数为 t，簇个数为 K，每个簇的代表点个数为 M，则算法的时间复杂度为 $O(t(K+M+1)M|S|)$，复杂度与 k-means 接近，低于 k-medoids。

4 实验数据与分析

为了评价 KMMRSC 算法的性能，本文在 AMD 1.54GHz 主频、512MB 内存和 Windows XP 操作系统的 PC 机上对该算法进行了测试，并与 k-means 算法进行了比较。实验样本取自于圣玛莉大学 16 周以来计算机科学一年级课程的 web 访问日志[6]，预处理后获取的样本集大小为 21637，属性个数为 5，如表 1 前 5 列属性所示，含义分别为在校内还是校外访问网站(校内为 1/校外为 0)、在白天还是晚上访问(白天为 1/晚上为 0)、在做实验或上课期间还是其它时间访问(做实验或上课期间为 1/其它为 0)、访问期间的点击量和课件下载数量。KMMRSC 算法中 $\alpha \in (0,1), \varepsilon \in (1,2)$。实验过程中，对 α 和 ε 进行了线性调整，并观察这两个参数对算法结果的影响。限于篇幅，表 1 仅给出了 $K=3, \alpha=0.5, \varepsilon=1.5$ 时的实验结果，其中 ClusterSize 列表示聚类结果中各簇集内的样本个数(对于 KMMRSC 算法来说是下近似、上近似集合中分别包含的样本个数)。

从表 1 可以看出，两种算法均得到了良好的聚类结果，KMMRSC 算法得到的簇凝聚度更高一些，各簇之间更加离散。好的聚类结果要求簇间距离越大越好，簇内距离越小越好，因此实验过程中采用了凝聚度 AD 来描述聚类效果，AD 越大，聚类效果越好。整个实验表明，参数 α 和 ε 的选取对聚类结果有一定影响，且 $\text{AD} \propto \alpha, \text{AD} \propto \varepsilon$，即 α 和 ε 越大，AD 越大。但 α 越接近 1，M 个中心点越趋向重叠于簇心，从而便无法发现任意形状的簇。ε 越大，属于每个簇下近似集合(即真正隶属于该簇)的样本点越少，聚类也失去了意义。实验发现，$\alpha \geq 0.5$ 且 $\varepsilon \geq 1.1$ 时，KMMRSC 算法的聚类结果总是优于 k-means。

5 结束语

本文将粗糙集理论应用于知识发现中的聚类分析，提出了一种基于覆盖的粗糙聚类算法KMMRSC，用上、下近似来刻画对象的归属，类与类之间是一种覆盖关系，并且用多个中心点代表一个聚类簇，从而能够得到任意形状的聚类结果，同时应用了采样方法，提高了算法效率，使之能够处理大数据集。实验结果验证了算法的可行性，并且达到了理论分析的效果。但该算法目前尚不能同时处理数据属性和符号属性，且要事先给出聚类簇的个数。因此，如何定义距离的计算方法，使算法能同时处理符号属性和数值属性；如何自适应地寻找聚类簇的个数K以获得更自然的聚类结果将是我们今后研究的重点。

表 1 k-means 和 KMMRSC 算法聚类结果比较

	On campus	Day time	Lab/class day	Hits	Classnotes	ClusterSize(Lower/Upper)	AD
k-means	0.45	0.75	0.34	112.64	9.52	472	4.65
	0.62	0.76	0.44	41.45	4.89	3335	
	0.61	0.73	0.44	8.52	0.82	17830	
KMMRSC	0.46	0.75	0.31	125.10	10.17	331/839	5.44
	0.61	0.75	0.43	46.50	4.96	2572/839	
	0.61	0.73	0.44	8.58	0.84	17895/0	

参 考 文 献

[1] Berkhin Pavel. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[2] Han Jiawei and Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2000, chapter 8.

[3] Grabmeier J and Rudolph A. Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, 2002, 6(4): 303–360.

[4] Jain A K, Murty M N, and Flynn P J. Data clustering: A review. *ACM Computing Surveys*, 1999, 31(3): 264–323.

[5] Lingras P. Unsupervised roughset classification using GAs. *Journal of Intelligent Information Systems*, 2001, 16(3): 215–228.

[6] Lingras P and West C. Interval set clustering of web user with rough K-means. *Journal of Intelligent Information Systems*, 2004, 23(1): 5–16.

[7] Peters Georg. Some refinements of k-means clustering. *Pattern Recognition*, 2006, 39(8): 1481–1491.

[8] Asharaf S, Murty M N, and Shevade S K. Rough set based incremental clustering of interval data. *Pattern Recognition Letters*, 2006, 27(6): 515–519.

王慎超： 男，1983年生，硕士生，研究方向为粗糙集理论、数据挖掘等.

苗夺谦： 男，1964年生，教授，博士生导师，主要研究方向为粗糙集理论、模式识别、人工智能等.

陈 敏： 女，1980年生，博士生，研究方向为web日志挖掘、粗糙集等.

王睿智： 女，1968年生，博士生，研究方向为粒度计算、粗糙集、聚类分析、数据挖掘等.