粗糙几何及其在脱机手写数字识别中的应用*)

岳晓冬　苗夺谦　张超琼　王睿智

(同济大学嵌入式系统与服务计算教育部重点实验室　上海 201804)

(同济大学计算机科学与技术系　上海 201804)

摘　要　粗糙几何学将粗糙集理论应用于几何学之中，利用图形的上近似变换，在更粗糙的粒度上构造并分析几何图形。本文着重介绍了粗糙几何学的研究动机和理论基础，同时将其应用于脱机手写数字识别，并对粗糙几何未来的研究方向进行了展望。

关键词　粗糙集理论，粗糙几何，上近似变换，几何不变性，手写数字识别

Research on Rough Geometry and its Applications to Off-line Handwritten Digit Recognition

YUE Xiao-Dong　MIAO Duo-Qian　ZHANG Chao-Qiong　WANG Rui-Zhi

(The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 201804)

(Department of Computer Science and Technology , Tongji University , Shanghai 201804)

Abstract　Rough geometry is the result from applying the rough set theory to the traditional geometry. In the new theory, the geometric configuration can be constructed by its upper approximation at different levels of granularity. This paper focuses on the introduction of the motivation and the foundation of the rough geometric theory and then we apply it to the off-line handwritten digit recognition. Finally, we present the future research perspectives about the rough geometry.

Keywords　Rough set theory, Rough geometry, Upper approximation, Geometric fixity, Handwritten digit recognition

1　粗糙几何研究的动机

欧几里德几何统治了世界两千年，但欧几里德几何中“没有长度，没有宽度，没有厚度的点”，“没有宽度，没有厚度的线”在现实世界中是无法做出的，例如在屏幕上显示的线都是有宽度的。而屏幕中的点和线具有尺寸的原因在于对给定的分辨率，一个像素中的欧几里德点是不可区分的。

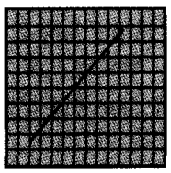


图1　屏幕中的欧几里德线

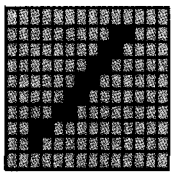


图2　覆盖线的像素

如图1、2所示，屏幕上显示的欧几里德线实际上是由覆盖这条线的像素组成，即是由与它相交不空的等价类组成，也即是它的上近似(显然这种情况下近似是空)。因此，相关研究将粗糙集理论应用于几何学之中，提出了粗糙几何学[1]。

2　粗糙几何的理论基础

2.1　粗糙集方法

简单介绍与粗糙几何相关的粗糙集理论[2,3]的基本概念。

设信息系统 $S=(U,A)$。其中 U 是一个非空的有限样本集合，A 是一个非空的有限属性集合。对于每一个属性子集 $B\subseteq A$ 决定了如下一个二元的不可区分关系 $IND_S(B)$：$IND_S(B)=\{(x,y)\in U\times U\mid \forall a\in B, a(x)=a(y)\}$。由此形成的不可区分关系 $IND_S(B)$ 具有自反性，对称性及传递性，因此，$IND_S(B)$ 是一个等价关系，将在样本集 U 上形成一个划分。

对于一个等价关系 R，任意样本 $x\in U$ 在其划分下所对应的等价类定义为如下集合：$[x]_R=\{y\mid y\in U\wedge xRy\}$，显然一个等价类中的样本是不可区分的。简便起见，将用 $[x]_B$ 来表示属性子集 $B\subseteq A$ 对应的 x 在等价关系 $IND_S(B)$ 下所形成的等价类。

对于信息系统 $S=(U,A)$，设样本子集 $X\subseteq U$，属性子集 $B\subseteq A$，可以用 B 所含的信息来描述 X，即构造 X 的上，下近似。定义 $\underline{B}X=\{x\mid [x]_B\subseteq X\}$ 为 X 的下近似，$\overline{B}X=\{x\mid [x]_B\cap X\neq\phi\}$ 为 X 的上近似。

2.2　粗糙几何中的基本概念

作为一种新的非欧几何，粗糙几何学将粗糙集理论中的等价类及近似概念应用于几何学之中，对于用等价类构造的全新的几何对象，很多概念和性质都需要严格的定义。下面介绍一些粗糙几何学中的基本概念。

定义 2.1　粗糙空间，粗糙点，粗糙图形

设 φ 是从 $\mathcal{R}$ 到 $\mathcal{R}$ 的一个满射，$\forall x,y\in\mathcal{R}$，若有 $x\leqslant y\Rightarrow\varphi(x)\leqslant\varphi(y)$，则称 φ 是从 $\mathcal{R}$ 到 $\mathcal{R}$ 的一个同态，同时定义 φ 的同态核为 $Ker\varphi=\{(x,y)\in\mathcal{R}\times\mathcal{R}\mid \varphi(x)=\varphi(y)\}$，显然是 $\mathcal{R}$ 上的

*)基金项目：国家自然科学基金项目(No. 60475019)，2006年博士学科点专项科研基金(No. 20060247039)。**岳晓冬**　博士研究生，主要研究方向为粗糙集理论，粒度计算；**苗夺谦**　教授，博士生导师，主要研究方向为人工智能、模式识别、数据挖掘、粗糙集理论、主曲线；**张超琼**　硕士研究生，主要研究方向为粗糙集理论，粒度计算；**王睿智**　博士研究生，主要研究方向为模式识别，数据挖掘，粗糙集理论。

一个等价关系。$\mathscr{R}^n$ 是一个 n 维欧几里德空间。在 $\mathscr{R}^n$ 中由 $Ker\varphi$ 定义的不可区分关系"$\approx_\varphi$"如下：点$(x_1,\cdots,x_n)$与$(y_1,\cdots,y_n)$是不可区分的，当且仅当$(x_1,y_1)\in Ker\varphi,(x_2,y_2)\in Ker\varphi,\cdots,(x_n,y_n)\in Ker\varphi$。显然$\approx_\varphi$是自反的，对称的，传递的，因此$\approx_\varphi$也是$\mathscr{R}^n$上的一个等价关系（当$n=1$时，$\approx_\varphi$就是$Ker\varphi$）。

定义 2.1.1 对于$\mathscr{R}^n$上的一个等价关系$\approx_\varphi$，称$\approx_\varphi$划分的$\mathscr{R}^n$所有等价类集合$\mathscr{R}^n/\approx_\varphi$为一个粗糙空间。

定义 2.1.2 对于$\mathscr{R}^n$上的一个等价关系$\approx_\varphi$形成的粗糙空间$\mathscr{R}^n/\approx_\varphi$，称粗糙空间中的一个元素为一个粗糙点。

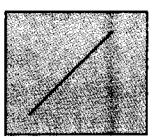
定义 2.1.3 对于$\mathscr{R}^n$上的一个等价关系$\approx_\varphi$形成的粗糙空间$\mathscr{R}^n/\approx_\varphi$，称粗糙空间中的粗糙点子集为一个粗糙图形。

定义 2.2 粗糙子空间，粗糙父空间

设$\mathscr{R}^n/\approx_1$及$\mathscr{R}^n/\approx_2$是两个粗糙空间，若$\mathscr{R}^n/\approx_1$的每个粗糙点都是$\mathscr{R}^n/\approx_2$的某个粗糙点的子集，则称$\mathscr{R}^n/\approx_1$是$\mathscr{R}^n/\approx_2$的子空间，$\mathscr{R}^n/\approx_2$为$\mathscr{R}^n/\approx_1$的父空间，并且记为：$\mathscr{R}^n/\approx_1\leqslant\mathscr{R}^n/\approx_2$。

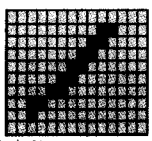
我们还定义一个特殊的粗糙空间$\mathscr{R}^n/=$为$\{\{(x_1,\cdots,x_n)\}|(x_1,\cdots,x_n)\in\mathscr{R}^n\}$，即$\mathscr{R}^n/=$是每个等价类都只含一个欧几里德点的粗糙空间。不区分一个元素与只含一个元素的集合，则$\mathscr{R}^n/=$就是$\mathscr{R}^n$，就是普通n维欧几里德空间。规定$\mathscr{R}^n/=$是所有粗糙空间$\mathscr{R}^n/\approx_\varphi$的子空间。

定义 2.3 图形的上近似变换。设S是某个粗糙空间$\mathscr{R}^n/\approx_1$中的图形，则S在另一个粗糙空间中$\mathscr{R}^n/\approx_2$的上近似及下近似分别记作$U_{\approx_2}(S)$及$L_{\approx_2}(S)$，如图3、4、5所示，当$\mathscr{R}^n/\approx_1\leqslant\mathscr{R}^n/\approx_2$，称$U_{\approx_2}(S)$为$S$的"上近似变换"。



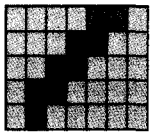
$\mathscr{R}^n/=$空间中的线

图3



线在$\mathscr{R}^n/\approx_1$中的变换

图4



线在$\mathscr{R}^n/\approx_2$中的变换

图5

2.3 上近似变换的几何不变性

英国数学家 Felix Klein 将不同的几何学定义为研究不同变换下的几何不变性的科学。例如投影几何就是研究投影变换下图形不变性的科学。在粗糙几何学中将研究图形的"上近似变换的几何不变性"。它具有如下性质：

性质 2.1 由粗糙子空间到其对应的粗糙父空间的上近似变换中，粗糙点不变。

性质 2.2 由粗糙子空间到其对应的粗糙父空间的上近似变换中，两粗糙点间相对位置不变。

性质 2.3 由粗糙子空间到其对应的粗糙父空间的上近似变换中，粗糙线段不变。

性质 2.4 由粗糙子空间到其对应的粗糙父空间的上近似变换中，粗糙图形凹凸性不变。

性质 2.5 由粗糙子空间到其对应的粗糙父空间的上近似变换中，粗糙图形相交性，通过性不变。

性质 2.6 由粗糙子空间到其对应的粗糙父空间的上近似变换中，粗糙图形对称性不变。

性质 2.7 由粗糙子空间到其对应的粗糙父空间的上近似变换中，粗糙图形两点间距离相等性不变。

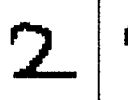
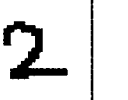
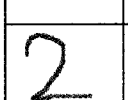
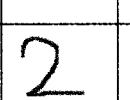
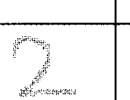
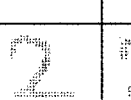
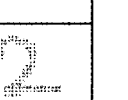
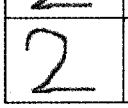
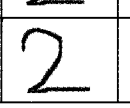
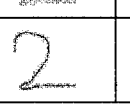
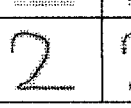
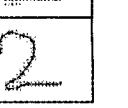
3 粗糙几何在主曲线中的应用

主曲线概念是 Hastie 和 Stuetzle 于 1984 年提出的[4]，主曲线是通过数据分布"中间"并满足"自相合"的光滑曲线，其目的是根据给定的数据集合寻找通过数据分布"中间"的曲线和曲面，使其能够在低维空间真实地反映数据的形态。主曲线对数据信息保持性好，其理论基础是寻找嵌入高维空间的非欧氏低维流形，是线性主成分的非线性推广[5]。基于以上优良性质，主曲线研究自 20 世纪 90 年代以来取得了较快的发展，且研究成果已经广泛应用于计算机处理的实际问题之中，如线性对撞机中电子束运行轨迹的控制，图像处理中辨识冰原轮廓，脱机手写体数字及字符的主曲线识别[6,7]等。

由于主曲线的理论基础是寻找通过数据分布"中间"并满足"自相合"的光滑曲线，这样，多种主曲线算法的复杂度就与初始数据规模紧密相关。但是，在实际应用中，我们也许不必遍历全部初始数据以生成数据分布的"骨架"，只需要选择可以代表数据分布的典型数据来生成主曲线以提取数据分布的近似特征，达到识别目的。例如，在数字或字符的识别过程中，现有的方法主要由图像中待识别对象的像素数据集合生成的主曲线来获取其结构特征，然而，依据粗糙几何学的上近似变换不变性，如凹凸性不变，通过性不变等，我们可以在更粗糙的粒度下看待识别对象，在更粗糙的几何空间中生成对应的主曲线对其进行特征抽取。这样不仅可以极大地缩小主曲线算法在识别任务中处理的数据规模，提高算法效率，使实时处理成为可能，而且可以在一定程度上减小噪声数据在识别过程中的影响，突出识别对象的结构特征。

综上所述，我们将粗糙几何应用于基于主曲线的脱机手写数字识别任务中，做一些简单的探索性工作。实验中采用推广的多边形(PL)算法来提取手写数字的骨架结构。实验采集到的数据见表 1，可以看到主曲线算法在不同粗糙空间中对手写数字的识别效果。实验采取屏幕上的单位像素作为粗糙空间中可度量的最小等价类($\delta=1$)，对应形成的粗糙空间为 $SPACE(1)$，并且选取了五个不同的粒度来形成对应的粗糙空间。表 1 分别展示了不同粗糙空间中手写体数字"2"在屏幕上的显示效果，形成的待识别的二维数据点集合以及对数据点集合采用主曲线算法进行识别得到的数字骨架。

表 1 主曲线算法在粗糙空间的手写数字识别效果

SPACE (1)	*SPACE*(2)	*SPACE*(5)	*SPACE*(8)	*SPACE*(10)
				
				
				

通过实验，我们发现手写体数字的主要结构特征在进行一定范围内的上近似变换过程中损失很小，这与之前所述的图形上近似变换的不变性质是一致的。当以更粗糙的粒度[8~10]视角来看待识别对象时，构成粗糙图形的等价类集合

（下转第 270 页）

集，不与其它任务共享。各个并行执行的任务之间通过显示的消息传递来交换数据、协调步伐、控制执行[3]。

在贸易地图生成软件中应用 MPI 并行编程模型，需要重新设计一个基于 MPI 的调度模块。由于把生成贸易地图作业分解成多任务后，各任务的执行相对独立，不存在通过显示的消息传递来交换数据、协调步伐，而在 MPI 模块中唯一可以控制的消息传递是对任务数量进行广播，这不是问题的关键。因此，在并行处理方案中选择何种编程模型，需要把握原系统的特征，确定系统性能瓶颈。同时也说明不是所有的问题都适合消息传递编程模型。所以贸易地图生成软件采用的是基于参数扫描分析的并行处理方案。

5 并行处理方案的性能评测

贸易地图生成软件并行处理方案性能评测的主要内容是：测加速比和并行效率。加速比和并行效率是最传统的并行算法评价标准，它体现并行机上运行并行算法，求解实际问题所能获得的性能。对求解具有相同规模的统一问题，并行算法加速比可定义为

$$S_P = T_S / T_P$$

其中 T_S 为最佳串行算法在单处理机上的运行时间；T_P 为并行算法在并行机上使用 P 台处理机所需时间。相应并行效率可定义为：

$$E_P = S_P / P$$

式中：P 为处理器数[4]。本方案在测试时，保证了任务规模、机器硬件配置，以及其他环境相同。测试数据如表 1 所示。

由表 1 可知，并行计算随着节点数的增加，完成整个作业的时间明显缩短。但由于网络存在通信和延时的问题，造成相应的加速比值低于理想状态以及并行效率的下降。

结束语 本文通过对贸易地图生成软件主要特征的分析和研究，明确了系统运行性能瓶颈，把基于参数扫描分析的数据并行编程模型应用到系统中，缩短了生成单个贸易地图的时间，提高了批量生成贸易地图的效率，改善了系统的整体性能。通过对并行计算方案的研究，探讨了数据并行中参数扫描分析任务调度策略的使用范围和特征，并且在集群上实现了这种任务调度策略，从性能测试数据证明达到了预期成果。最后还简要分析了 MPI 消息传递编程模型的使用特点。

表 1 并行方案性能测试表

测试次数	串行/并行(s/p)	CPU 个数	时间(s)	S_P	E_P
1	S	1	5642	1	1
	P	2	3241	1.7408	0.8704
	P	4	1999	2.8224	0.7056
2	S	1	6076	1	1
	P	2	3418	1.7776	0.8888
	P	4	2077	2.9254	0.7314
3	S	1	5680	1	1
	P	2	3192	1.7794	0.8897
	P	4	1987	2.8586	0.7147

参 考 文 献

1 Microsoft Windows Server TechCenter. Using Windows Compute Cluster Server 2003 Job Scheduler. http://technet2.microsoft.com/windowsserver/en/technologies/featured/ccs/default.mspx, 2006,4

2 Chingchit S, Kumar M, Bhuyan L N. A flexible clustering and scheduling scheme for efficient parallel computation. In:13th International and 10th Symposium on Parallel and Distributed Processing,1999 IPPS/SPDP. Proceedings, Suan Juan, April 1999

3 张国晓，袁立强，徐炜民. 基于任务类型的集群调度策略. 计算机工程，2004,30(13)

4 李俊照，罗家融. 基于 linux 集群的并行计算. 计算机测量与控制，2004,12(11)

(上接第 252 页)

就是在更粗粒度上代表数据分布特征的点集，应用主曲线算法对更粗粒度上的点集进行识别，仍然可以得到对象比较完整的结构信息，从而达到理想的识别效果。在更粗糙的空间中运用主曲线算法对识别对象进行特征抽取，待处理的数据规模，主曲线算法的迭代次数以及生成的主曲线数据规模都成倍减少，从而极大改进了算法的执行效率。

总结 粗糙几何学将粗糙集理论应用于几何学之中，利用图形的上近似变换，以更粗糙的粒度构造并分析几何图形，带领我们进入了一个崭新的，更贴近实际的几何空间。本文上述章节着重介绍了粗糙几何学的研究动机和理论基础，并对其在脱机手写数字识别领域中的应用前景做了探索性的研究工作。

粗糙几何是崭新的几何学，其理论及应用的各个方面仍有待不断完善改进，但它开始引导人们在不同的粒度空间中审视传统的几何图形，以更新、更现实的视角来发现并分析几何图形的重要性质。粗糙几何学不仅具有深刻的理论意义，而且在图像分析，图像压缩以及模式识别领域具有广泛而直观的应用前景。

参 考 文 献

1 马垣. 粗糙几何学. 计算机科学，2006,33(11A):8

2 王国胤. Rough 集理论与知识获取. 西安交通大学出版社，2001

3 张文修，吴伟志，梁吉业，李德玉. 粗糙集理论与方法. 科学出版社，2000

4 Hastie T. Principal Curves and Surfaces. Laboratory for Computational Statistics, Stanford University, Department of Statistics: [Technical Report 11]. 1984

5 张军平，王珏. 主曲线综述. 计算机学报，2003,26(2):129～146

6 苗夺谦，张红云，李道国，王真. 基于主曲线的脱机手写数字识别. 电子学报，2005,33(9): 1639～1643

7 张红云，苗夺谦，张东星. 基于主曲线的脱机手写数字结构特征分析及选取. 计算机研究与发展，2005,42 (8):1344～1349

8 Yao Y Y. Granular Computing: basic issues and possible solutions. In:P P Wang, ed. Proceedings of the 5th Joint Conference on Information Sciences, 2000,1:186～ 189

9 Yao Y Y. Information Granulation and Rough Set Approximation. International Journal of Intelligent Systems, 2001,16:87～104

10 Skowron A, Stepaniuk J. Information Granules: Towards Foundations of Granular Computing. International Journal of Intelligent Systems, 2001,16:57～85