文章编号：1007-4619 (2004) 01-0045-06

MSCMO：基于数学形态学算子的尺度空间聚类方法

汪 闽，周成虎，裴 韬，骆剑承
（中国科学院 地理科学与资源研究所资源与环境信息系统国家重点实验室，北京 100101）

摘 要： 提出了一种基于数学形态学算子的多尺度聚类方法：首先将数据进行二值图像化处理，利用一次闭开运算去除噪声干扰后再利用逐步增大结构元素的闭运算构建尺度空间，图像内的连通单元集随着尺度上升不断融合，最终全部归并。将连通集覆盖下的点集归为一类，以上步骤就对应了一个多尺度层次聚类过程。本算法的一个最大优点是聚类个数事先无需设定，而被确定为跨越最多尺度（具有最长尺度生存期）的类别个数。此外，参数少、能够提取任意形状的类别、具有较强的抗噪声能力也是算法的优点。对自构建数据与地震数据的聚类实验验证了方法的有效性和实用性。

关键词： 数学形态学，尺度空间，聚类
中图分类号： TP751.1 **文献标识码**： A

1 引 言

聚类被广泛地应用于模式识别、数据分析、图像处理等领域，近年来，它逐渐被看作是从空间数据库中发现知识的一种主要的挖掘方法[1]。针对空间数据往往呈现非球状聚集的特点，面向此类数据的聚类方法应该能够处理不规则的聚集形状[2]。此外，聚类任务中确定类似类别个数这样的参数往往是一个棘手的问题，而空间数据通常数据量庞大、结构复杂，参数选取尤为不易，因此，无参数或少参数聚类算法也是面向空间数据挖掘任务的一个迫切要求[2]。

直观理解，我们感知的类别个数和观察数据的尺度有关：在一个非常粗糙的尺度下（如距数据集非常远的距离），整个数据集是一个类别，而在一个非常精细的尺度下，每个数据点可看作是一个类别。这为最合理类别个数的确定带来了新的思路：将类别个数选定为在一个最长的尺度变化范围内固定不变的个数，换句话说，也就是此类别个数具有最长的“尺度生存期”[3]。

尺度空间理论（Scale Space Theory）最早出现于计算机视觉领域，其目的是模拟图像数据的多尺度特征[4]。Witken[5]等人利用高斯卷积构造一幅尺度空间图像：

$$F(x,\sigma)=f(x)^{*}g(x,\sigma)=\int_{-\infty}^{+\infty}f(u)\frac{1}{\sigma^{2}\sqrt{2\pi}}e^{-\frac{(x-u)^{2}}{2\sigma^{2}}}du \quad (1)$$

其中，F 是信号 f 通过卷积变换后得到的尺度空间图像。参数 σ 对应着尺度的概念：随着 σ 的增大，信号 f 将逐步被平滑，代表着其由精细尺度逐步变换到粗糙尺度的滤波过程。Leung 等人[3]依据上述原理设计了基于尺度空间滤波的多尺度聚类算法，并根据类别个数的生存期、类别的生存期等概念确定合理的类别个数和类别。类似算法尚见 Wong 等人[6]提出的基于热力学融合理论的的多尺度聚类方法等等。上述方法的共同特点是需要求取类别中心点并据其进行迭代融合，因此对类别形状有所限制。

数学形态学（Mathematical Morphology）是分析几何形状和结构的数学方法[7]。将数学形态学与尺度概念、聚类分类结合的研究，曾见 Postaire 等人[8]利用形态学开闭算子寻找簇核（CORE），再用最近邻法归属剩余数据点的聚类思路；Maragos[9]利用形态学滤波器构建一个尺度空间进行形态表示；Acton 等人[10]利用区域形态学（Area Morphology）的开闭运算构建尺度空间，利用其尺度空间矢量（scale space

收稿日期：2002-05-24；**修订日期**：2002-08-29
基金项目：863 计划（2002AA135230）国家自然科学基金（40101021），中科院知识创新项目（CX10G-D00-06），（KZCX1-Y-02）资助。

作者简介：汪 闽（1975 年— ），男，浙江衢州人。中科院资源与环境信息系统国家重点实验室博士生。主要研究方向为空间数据挖掘与知识发现，空间数据聚类、分类算法。已发表论文 6 篇。

vectors)进行图像分类的方法等等。

考虑到高斯尺度空间聚类方法的不足之处与聚类空间数据库中任意形状类别的需求，本文提出了一种基于数学形态学算子的多尺度聚类方法 MSCMO(Multi-Scale Clustering algorithm based on mathematical Morphology Operators)。方法的基本过程是：将数据空间离散变换为图像空间，利用一次闭开运算去除噪声影响，再利用闭运算构造图像的尺度空间，闭运算的结构元素大小对应尺度大小，随着结构元素的不断增大，图像内连通集相互间不断融合，最终全部归并为一类，而最终类别个数则确定为跨越尺度最多的个数。MSCMO 的优点在于事先无需确定类别个数、所需参数少且设置简易、能够提取任意形状的簇，算法效率较高且具有较强的抗噪声能力，适合于空间数据的挖掘。

2 数据的图像化处理

为将数学形态学方法应用于空间聚类研究，须对原数据进行图像化处理。为此用指定精度划分一个覆盖整个数据区的格网，并对落入数据点的格网单元进行标识。格网划分的关键在于避免噪声对后继运算的影响。通过试验，可根据如下方法指导精度选取：在点聚集区、噪声区进行数据块采样，为采样点集的每点(考察点)计算其到最近邻的距离，将距离按大到小排序，以点序为横坐标，距离为纵坐标作一曲线(此曲线称为排序 N-DIST 图[11]，N 指距考察点最近的第 N 个点，计算最近邻则 $N=1$)，寻找

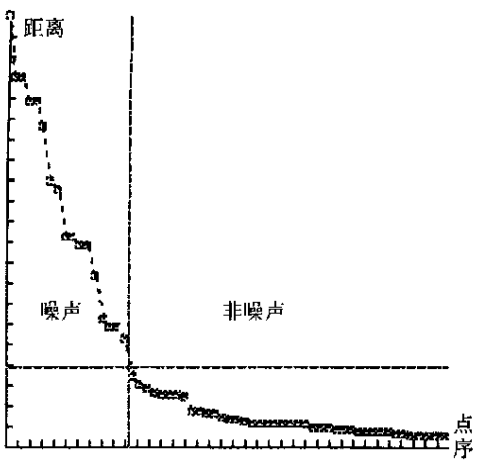

图 1 排序 1-DIST 图

Fig.1 Sorted 1-DIST graph

噪声与非噪声的分界距离，以此距离的一半或更少作为格网跨度，划分单元格。图 1 是试验 2 采样区的排序 1-DIST 图，发现在交叉线左部点集距离变化剧烈而其右迅速平缓，此外可认为是噪声与非噪声的分界，格网单元跨度应小于其一半距离。如此处理是由于闭开运算去除噪声中使用模板的最小尺度为 3×3，因此应尽量保证噪声单元间隔在一格以上使其不致在膨胀运算中相互粘连而被保留下来。

显然同一格网单元可能落入多个数据点，我们将其简便处理为保留一个。这样就将原连续的数据空间变换为一个离散的二值图像空间，有值单元为 1，无值单元为 0，可应用二值形态学算子对其进行一系列的操作。

将图像的连通量覆盖下的点集归为一类可将图像的连通量和聚类类别相对应。关于搜索图像连通成分的经典算法很多，如传统的种子填充算法、扫描行填充算法[12]等等，这里不再赘述。

3 几个主要的数学形态学算子及其在 MSCMO 中的应用

设图像 S 与结构元素 B 均为二维欧氏空间中的集合，欧氏空间中 B 平移距离 x 的平移运算表示为：

$$B+x=\{b+x \mid b\in B\} \tag{2}$$

则数学形态学的基本运算如下[13]：

B 对 S 的腐蚀运算为：

$$S\Theta B=\{x \mid B+x\in S\} \tag{3}$$

B 对 S 的膨胀运算为：

$$S\oplus B=\{x \mid (-B+x)\cap S\neq\phi\} \tag{4}$$

B 对 S 的开运算为：

$$S\circ B=(S\Theta B)\oplus B \tag{5}$$

B 对 S 的闭运算为：

$$S\bullet B=[S\oplus(-B)]\Theta(-B) \tag{6}$$

开闭运算互为对偶运算，具有平移不变性、单调递增性、非扩展性和幂等性，处理图像时既能消除细节，又能保持整体形态不变，因此在去除图像的胡椒状噪声和砂眼噪声中有着广泛应用[13]。闭运算实际上是利用相同结构元素对原图像进行一次膨胀后再作一次腐蚀操作，它较膨胀操作更易保存原连通集的基本形态，并能起到跨接连通集间间隙的作用。而由于其幂等性质，一次闭滤波后再用相同的结构元素作重复运算则不再有新的效果，这是和经典方法(如中值滤波、线性卷积)不同的性质[7]，然而增大结构元素，由于其膨胀操作部分，此前因间隔较大而不能跨接的像元就有可能相互融合。闭运算的上述

特性启发我们用逐渐增大的结构元素对图像进行迭代闭操作以构造一个尺度空间,结构元素的增大对应着尺度的上升,图像内的连通集随着尺度的上升相互间不断融合,最终全部融合为一个连通集,此处理也可看作是一个图像"基底空间"与"细节空间"不断剥离的滤波过程,其形式化描述如下:

设图像尺度空间为$\{I\}$,I_s 为尺度 s 下的图像。则$\{I\}$是定义在图像位置域 Ω 和尺度域 Ω_s 上的 I_s 集。尺度域中尺度 $s(t)$(结构元素)以 t 为参量,$s(0)$代表最精细尺度,而 $s(|\Omega_s|-1)$代表最粗糙尺度。则应用闭运算构建$\{I\}$可表示为:

$$I_{s(t)}=I_{s(t-1)}\bullet s(t) \tag{7}$$

本方法中,$s(0)$为去噪后图像对应尺度,$S(|\Omega_s|-1)$则被定义为所有点集初次融合为一类的对应尺度。将某尺度下单个连通集所覆盖的点集归为一类,$\{I\}$的构建过程就可看作为一个考察点集多尺度从聚特征的层次聚类过程。而由于闭操作的形态学性质,提取的类别可以是较随意的形状,这是 MSCMO 的一个重要优点。

实验中发现如果数据集中存在噪声点则对连通集的形态构建有较大干扰,为此,在生成图像之后,利用一次闭、开运算去除图像噪声,并使之不再参与后继形态学运算,但可归入最终聚类结果。由于开运算是先对图像进行腐蚀之后再进行膨胀,而原数据离散化后连通集中难免存在相对的有值单元"稀疏区",仅使用开运算可能会造成合理丛聚点集的丢失。为此,在开运算之前先进行一次闭运算,这样的操作流程在填补了丛聚内的较小"砂眼"的同时也有效地避免了合理点集的丢失。

4　MSCMO 算法步骤

输入:数据点集、格网精度、去除噪声的结构元素 B_0。

输出:类别

1. 以指定精度划分覆盖特征空间的格网,将数据栅格化为图像 S;

2. $S=S\bullet B_0\circ B_0$(去除噪声点,使之不再参与后继计算);

3. 设当前图像的初始尺度 $c=1$,对应的结构元素为 B_c;

4. 统计类别数,并在类别数大于 1 时,do{

$$S=S\bullet B_c$$
$$c=c+1$$

}

5. 选择具有最长生存期的连通子集个数作为最终类别数;

6. 输出类别。

需指出的是,采用不同形状的结构元素对图像进行形态学运算可能会产生不同结果。图 2 是对原始矩形利用 3 * 3 方形、圆形结构元素进行 3 次膨胀操作后的结果(从左到右依次为原矩形、采用方形结构元素、圆形结构元素的膨胀结果)。可见,两者间存在边缘细节上的差异。为保证图像总体位置不因形态学处理而发生"偏移"应采用原点对称模板(因此模板最小尺度选择 3 * 3 较为合适,如 1 * 1 则不起作用)。圆形结构元素因具有旋转不变性而最常被采用,但在某些情况下,根据应用目的的不同,使用其他类型的结构元素效果可能会更好一些。在地震数据中的聚类实验中,发现采用圆形结构元素造成提取的地震带边界过于曲折,使用方形结构元素则较平直,效果较好。

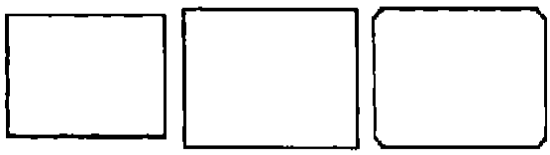

图 2　采用不同形状的结构元素对矩形作膨胀运算

Fig.2　Dilation operation with different structure elements

为保证算法通用性,MSCMO 提供用户选取或自定义不同结构元素的能力,算法缺省值设置选择 3 * 3 方形对称模板。算法也提供用户定义不同尺度下相应结构元素的能力,缺省大小设置为$(2*c+1)*(2*c+1)$。

作为一种层次聚类法,MSCMO 提供输出多个不同类别个数的合理聚类结果的能力。通过试验,最佳聚类结果则按以下方法确定:将类别个数的生存期定义为该类别个数所跨越的尺度个数,并选取首次出现该类别个数的图像尺度所对应的聚类结果作为最合适的输出结果。

5　实　验

5.1　实验一

实验一采用自构建数据,包含点数 814 个,目的是说明 MSCMO 的特点与优点。实验参数为:如前述确定格网精度 100 * 100 格,选取 3 * 3 方形结构元素去除噪声,且 3 * 3 的结构元素对应于第 1 尺度,第 2 尺度选取 5 * 5 结构元素,并依次类推。

观察图 3 原始数据，首先的问题就是：有几个类？似乎划分 3 个或 5 个类别都是合理的，但人们似乎更倾向于将图右角的 3 个点集归为一类，因为归并后其“尺度”更接近于其它两个类别。通过 DBSCAN[11] 方法、ISODATA 方法、基于高斯卷积的尺度空间聚类方法对该图数据聚类结果的对比，发现：基于密度概念的 DBSCAN 方法虽具有一定的抗噪声能力且可以聚集任意的形状，通过调整算法输入参数，可以将图中右上角 3 个点集归为一类，但造成下方“方框”类别包含了噪声；反之如果去除了噪声，又造成 3 个点集各自成类；这说明单一尺度的 DBSCAN 方法对本数据的聚类效果并不理想；ISODATA 则受噪声影响大，且类别形状趋于球状，也不适合；基于高斯卷积的尺度空间方法类别形状也趋于球状，并受噪声较大干扰，具最长尺度生存期的类别个数为 4，更不能满足需求。而 MSCMO 则可以很好地提取出图中三类。利用 MSCMO 方法计算各个类别数目对应的尺度生存期，见表 1：

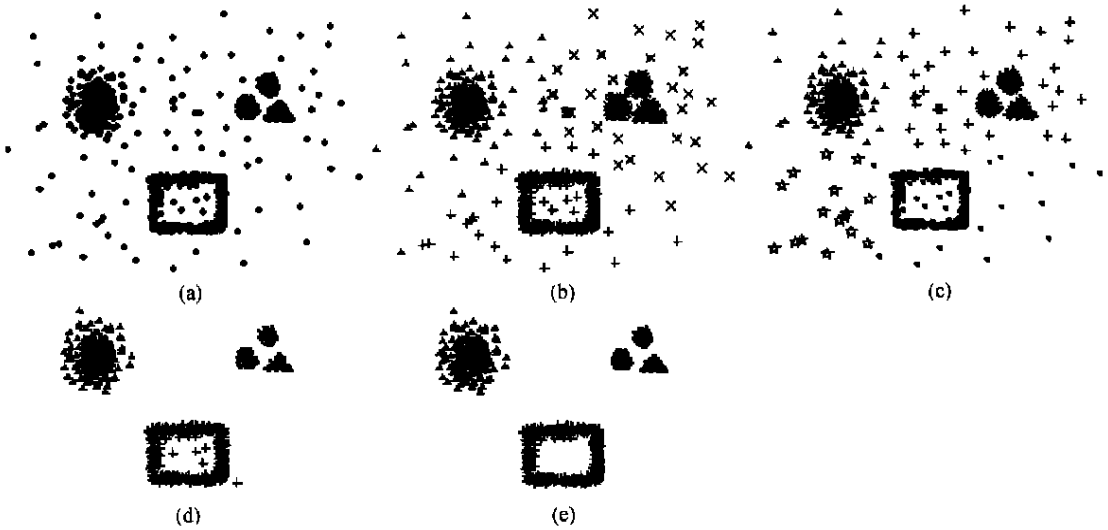

图 3 MSCMO 与几种聚类方法聚类结果对比 (a) 原数据；(b) DBSCAN；(c) ISODATA；(d) 高期尺度空间；(e) MSCMO

Fig. 3 Comparison between MSCMO with some of the other clustering algorithms (a) the raw data; (b) results of DBSCAN; (c) ISODATA; (d) scale space with Ganssian filtering; (e) MSCMO

表 1 图 3 中类别数的生存尺度

Table 1 Survival scales of the clusters in Fig. 3

类别数	6	5	3	1
生存尺度	0—1	2	3—13	14

分析表 1，明显可见，5 个类别的生存期只有 1 个尺度，而 3 个类别的生存期则有 11 个尺度，远远高于其它类别数，可见将图 3 划分为 3 个类别是最为合理的，它较好地反映了人眼在不同尺度下观察图 3 感知最久的丛聚个数，参照图 3 中的聚类结果，可见效果是最理想的。

5.2 实验二

实验数据来自中国及邻区地震数据库，共收集地震条目 510255 条[44]。我们从中抽取了大致范围 [34°—42°N, 106°—115°E]，震级大于 2.2 级的地震条目共 3201 条，目标是考察 MSCMO 方法在挖掘地震带（地震集中成带分布，并受活动构造带或地壳结构变异控制的地带[15]）任务中的能力。如图 4，一眼望去，图中明显存在两条接近南北走向的大地震带（图中虚线位置）。实际中它们分别对应了南北地震带的北段（左）和山西地震带（右）[16]。然而困难在于带内地震的密集区并不连续，相互间存在一定的间隔，利用单一尺度的密度聚类方法如 DBSCAN 很难将地震带作为一个类别整体提取出来，而 MSCMO 则能够完成这一任务。

参数设置为：如前述确定格网精度 150 * 150 格，其它参数和试验一完全相同。表 2 列举了不同类别数的尺度生存期，图 4 选取了几个不同尺度下连通集的分布状况。分析表 2 和图 4，明显可见：从第 6 个尺度开始类别个数进入一个相对较长的稳定期，选取 7 个类别作为最终的聚类个数，对比图 4 的第 6 尺度聚类结果图，可以发现两个地震带作为整体被较好地提取了出来。

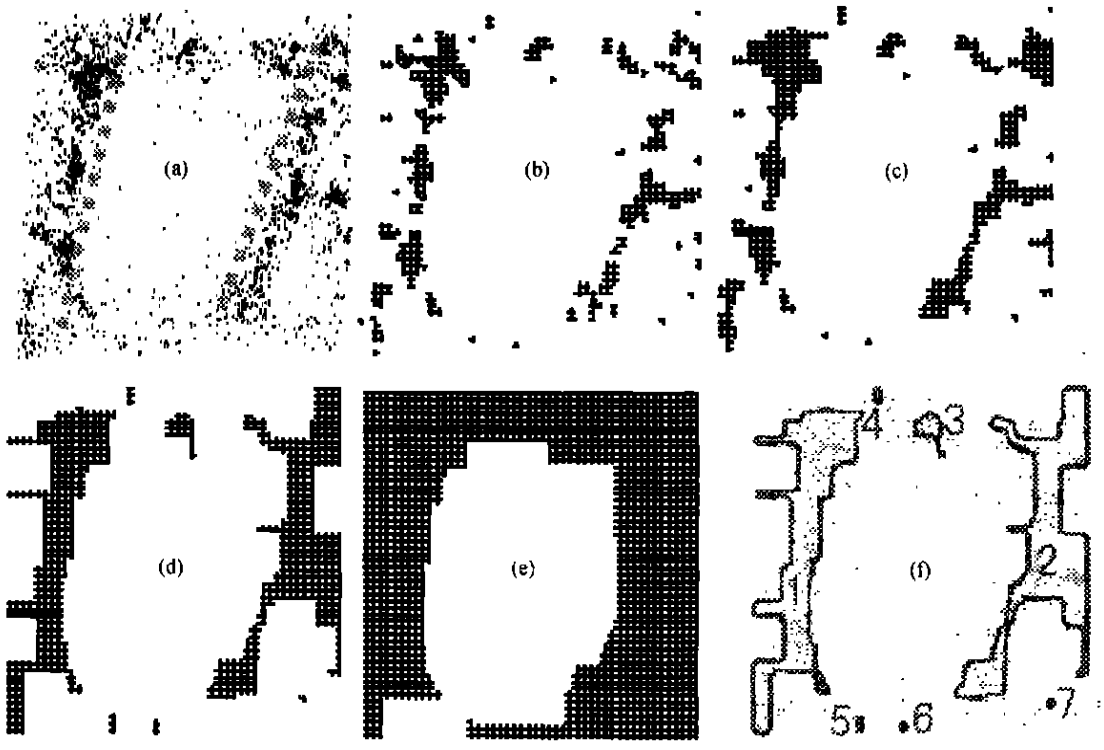

图 4　MSCMO 在地震数据中的应用实例 (a)原始数据集 (b)去噪声后图像 (c)第 3 (d)6 (e) 15 尺度 (f)第 6 尺度对应聚类结果图

Fig. 4　Application of MSCMO on earthquake data (a)the raw data (b)image removed noises. (c)the 3st (d)6th (e)15th scale (f)clustering result of the 6th scale

表 2　图 4 中类别数的生存尺度

Table 2　Survival scales of the Clusters in Fig. 4

类别数	…	14	10	7	7	7	6	5	3	…
生存尺度	…	4	5	6	7	8	9	10	11	…

6　总　结

本文提出了一种基于形态学算子的尺度空间聚类方法 MSCMO,并用实验证明了方法的有效性和实用性。MSCMO 本质上是一种适合于大数据集的二值图像分割方法,通过本文提出的矢栅转换规则,方法可实现矢量数据的聚类。应该指出,本文使用的实验数据仅为离散点数据,而稍加改进,MSCMO 应可适合于线状数据、面状数据的多尺度聚类工作:如间断道路的提取与归类、城镇聚落的丛聚特征分析等等。本算法已集成入课题组编制的时空数据库挖掘系统 STDBMiner 的聚类挖掘子系统 CFinder 中,目前我们正在如何在尺度空间构造过程中融入地学知识,在类聚搜索和生成过程中加入辅助信息和领域专家知识进行决策,使得整个非监督过程具有一定的知识监督能力、方法在线状、面状矢栅数据、高维特征空间中的扩充应用等方面做进一步的研究、改进,以更好地满足海量空间数据挖掘的需要。

参 考 文 献 (References)

[1] Martin Ester, Hans-Peter Kriegel, Jorg Snader, Xiaowei Xu. Clustering for Mining in Large Spatial Databases[J]. *Special Issue on Data Mining, KI-Journal*, 1998, **12**(1):18—24.

[2] Erica Kolatch. Clustering Algorithms for Spatial Databases: A Survey. [EB/OL]. URL:http://citeseer.nj.nec.com/436843.html, 2002-5-1.

[3] Yee Leung, Jiang-She Zhang, Zong-Ben Xu. Clustering by Scale-Space Filtering[J]. *IEEE transaction on pattern analysis and machine intelligence*, 2000, **22**(12):1396—1409.

[4] Tony Lindeberg. Scale-space: A framework for handling image structures at multiple scales[A]. Proc. CERN school of Computering, Egmond aan Zee[C], The Netherlands, 1996.

[5] A P Witkin. Scale-space filtering[A]. Proc. 8th Int. Joint Conf.

Art. Intell[C]. 1983:1019－1022.

[6] Wong Y. Clustering data by melting[J]. *Neural Computation*, 1993, **5**:89－104.

[7] He Bin, Ma Tianyu, Wang Yunjian, Zhu Honglian. Digital Image Processing with Visual C＋＋[M]. Beijing: People's Posts and Telecommunications Press, 2001,335－371. [何斌,马天予,王运坚,朱红莲. Visual C＋＋数字图像处理[M]. 北京:人民邮电出版社,2001,335－371.]

[8] J G Postaire, R D Zhang, and C Lecocq-Botte. Cluster Analysis by Binary Morphology[J]. *IEEE transactions on pattern analysis and machine intelligence*, 1993, **15**(2):170－180.

[9] P Maragos. Pattern spectrum and multiscale shape representation[J]. *IEEE transaction on pattern analysis and machine intelligence*, 1989, **11**(7):701－716.

[10] Scott T Acton, Dipti Prasad Mukherjee. Scale Space Classification Using Area Morphology[J]. *IEEE transactions on image processing*, 2000, **9**(4):623－635.

[11] M Ester, H-P Kriegel, J Sander and X Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [A]. Proc. of the Second International Conference on Knowledge Discovery and Data Mining[C], Portland, Oregon, 1996, 324－331.

[12] Sun Jiaguang, Yang Changgui. Computer Graphics[M]. Beijing: Tsinghua University Press, 1995, 185－186. [孙家广,杨长贵. 计算机图形学[M]. 北京:清华大学出版社,1995, 185－186.]

[13] Cui Yi. Image Processing and Analysis: Mathematical Morphology and its Applications[M]. Beijing: Science Press, 2000,67－76. [崔屹. 图像处理与分析——数学形态学方法及应用[M]. 北京:科学出版社,2000,67－76.]

[14] Pei Tao. Spatio-Temporal Characteristic Analysis and its Methods Research into Large Scale Seismic Database of China and its Adjacent Areas[R]. Beijing: Geography Institute, Acta, 2000,23. [裴韬. 中国及邻区大型地震数据库时空特征分析及其方法研究[博士后出站报告][R]. 北京:中科院地理所,2000,23.]

[15] National Department of Earthquake. Conspectus of the Layout Map on China's Earthquake Intensity (1990)[M]. Beijing: Earthquake Press, 1996, 64. [国家地震局. 中国地震烈度区划图(1990)概论[M]. 北京:地震出版社,1996,64.]

[16] Fu Zhenxiang. Research on the Earthquake Activity Mechanics in China's Mainland[M]. Beijing: Earthquake Press, 1997, 124－128. [傅征祥. 中国大陆地震活动性力学研究[M]. 北京:地震出版社,1997,124－128.]

MSCMO: A Scale Space Clustering Algorithm Based on Mathematical Morphology Operators

WANG Min, ZHOU Cheng-hu, PEI Tao, LUO Jian-cheng

(*State Key Laboratory of Resources and Environment Information System, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract: In this paper, a scale space clustering algorithm based on mathematical morphology operators (MSCMO) is proposed. The data are firstly converted into a binary image, the noises are then deleted with close-open operators. A scale space is constructed with the close operator and structure elements as well as increased size. The connected cells merge with each other with the increasing scale until all of them combine into one. We suggest this is just a multi-scale hierarchy clustering process considering the data under the connected cells into one class. One of the biggest advantages is that we do not need to set the cluster number before hand, it is fixed in the end on the cluster number which spans the longest scale range (with the longest 'scale survival time'). Besides, less arguments the ability to extract clusters with arbitrary shapes, and the robustness against noises are also the advantages of MSCMO. The validity and practicality of the algorithm are validated with constructed data and earthquake data.

Key words: mathematical morphology; scale space; clustering