从空间数据库发现聚类：一种基于数学形态学的算法*

邸凯昌　李德仁　　李德毅

（武汉测绘科技大学，武汉　430070）（总参谋部第六十一研究所，北京　100036）

摘　要　聚类算法能从空间数据库中直接发现一些有意义的聚类结构而不需要背景知识，是空间数据发掘和知识发现的重要手段。在分析已有聚类算法的基础上，提出了一种基于数学形态学的聚类算法，该算法能够处理任意形状的聚类，采用启发式方法自动确定最优聚类数。同时，该算法也可以在矢量型空间数据库中得到实现。试验表明算法是可行和有效的，且能处理存在噪音的数据。

关键词　聚类算法，数学形态学，数据发掘与知识发现，空间数据库

1　引　言

数据库技术是计算机信息处理技术中应用最广泛、发展最活跃的领域之一，数据和数据库急剧膨胀使数据库处理与分析技术面临挑战。现今数据库的处理主要地还是停留在查询、检索的层次，难以发掘和充分利用数据库中隐藏的丰富的知识。数据发掘（Data Mining，简称 DM ），或称从数据库中发现知识（Knowledge Discovery from Databases，简称 KDD），定义为“从数据库中发现隐含的、先前不知道的、潜在有用的信息”[1,2]，是数据库技术与机器学习、统计、可视化等技术相结合产生的新的技术领域，旨在提高数据库处理与分析的自动化和智能化水平，解决或缓解数据库激增与数据库分析相对困难之间的矛盾。

空间数据库（Spatial Database）是一类重要的、特殊的数据库，除了常规的属性数据外，空间数据库中还含有大量的图形数据，其数据结构和存取方法比一般关系数据库和事务数据库更加复杂，从而也有着更加丰富和复杂的语义信息，隐藏着丰富的知识。从空间数据库中发现知识（Knowledge Discovery from Spatial Databases，简称 KDSD），或称空间数据发掘（Spatial Data Mining，简称 SDM），是指从空间数据库中提取用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其它一些隐含在数据库中的普遍的数据特征[3~6]。从空间数据库中可以发现多种知识，使我们更好地理解和应用空间数据库。

聚类分析是统计学的一个分支，聚类算法能从空间数据库中直接发现一些有意义的聚类结构，而不象归纳学习等方法需要背景知识，因而作为一种重要的数据发掘方法应用于空间数据库中。本文在分析已有聚类算法的基础上，提出了一种基于数学形态学的聚类算法。该算法能处理任意形状的聚类，采用启发式方法自动确定最优聚类数，适用于点状、线状和面状目标。

2　已有聚类算法的分析

已有的聚类算法多数是为模式识别而设计的，将目标用其特征来表示，一个目标表达为多维特征空间的一个点，在特征空间中聚类。空间数据库中的聚类是对目标的图形直接聚类，空间目标有点状、线

* 本文得到国家自然科学基金重点项目“3S 集成理论与关键技术研究”（NO. 49631050）的支持，
收稿日期：1997-06-17；收到修改稿日期：1997-08-06

状、面状等多种类型，有时聚类形状复杂，同时数据量庞大，这就使空间数据发掘对聚类算法提出了更高的要求：能处理任意形状的聚类；适用于点状、线状、面状等多种目标类型；处理大型空间数据库时效率较高；算法需要的参数能自动确定或用户容易确定。下面对有代表性的算法根据上述要求作一简单分析和评述。

聚类算法主要地可以分为两类：分割算法（partitioning algorithm）和层次算法（hierarchical algorithm）。分割算法将 n 个目标划分到 k 个聚类中去，k 为输入的参数。首先选择 k 个代表点，其余目标根据到各类代表点的距离划分到 k 个聚类中；然后用每个类的重心（k-mean 算法）或离重心最近的点（k-medoid 算法）代表这个聚类，将目标重新分割；这一过程迭代进行，直到收敛。分割算法适用于聚类为凸形状、各类相距较远且直径相差不悬殊的情况，否则就可能产生错误的分割[7]。Ng 和 Han[8] 提出了一个称为 CLARANS 的改进的 k-medoid 算法，用随机搜索的方法提高了聚类的效率，从而适用于大型空间数据库中的聚类，然而仍然具有分割算法的上述缺陷。Ester 等提出用基于 R*-tree 的数据聚焦方法进一步提高 CLARANS 算法的效率[9]。

层次算法将数据集分解成树状图，即循环地将数据集分裂成子集，直到每一个子集只包含一个目标。树状图可采用分裂或合并的方法构建。层次算法不象分割算法那样需要聚类数 k 这个参数，但需要定义停止条件。层次算法的难点在于确定最优停止条件，同时也难以处理聚类形状复杂的情况。

Ester[7] 等提出了一种基于密度的聚类算法，称为 DBSCAN。其基本思想是认为在一类中的任一点一定半径的范围内有足够的相邻点，即邻域密度超过某一阈值。该算法需要 2 个参数：邻域半径 Eps 和邻域内点数阈值 MinPts，参数 Eps 采用启发式方法交互地确定，在二维数据的试验中将 MinPts 设为 4。算法采用 R*-tree 方法检索相邻点，对全部数据只需搜索一次即可得到最终结果，因而速度很快，同时它能处理任意形状的聚类，根据阈值 MinPts 还可以去除噪音，是一种较好的空间数据库聚类算法。然而，在实际的空间数据库中，并不总是存在 R*-tree 索引结构，R*-tree 的构建是十分耗时的，若将该算法用于没有索引结构的空间数据库中，其效率就会大大降低。

通过上面的分析可知，多数聚类算法不能处理非凸的、复杂的聚类形状，速度较慢，只能处理点状目标，因此应用于大型空间数据库时受到限制。DBSCAN 算法也只能处理点状目标，同时要求具有 R*-tree 索引结构。另外，在 GIS 系统中常见的栅格数据结构，上述算法都无法处理。由此，我们提出了下面的基于数学形态学的聚类算法。

3 基于数学形态学的聚类算法

数学形态学（Mathematical Morphology）是研究数字影像形态结构特征与快速并行处理方法的理论，它是通过对目标影像的形态变换实现结构分析与特征提取的目的。形态变换是通过选择称为结构元的较小特征影像集合与目标影像相互作用来实现。最基本的形态运算是膨胀和侵蚀，它们在二值图像中的定义如下：设 X 为目标影像，B 为结构元，膨胀 $X\oplus b=\bigcup_{b\in B}X_b$，其中 X_b 表示 X 相对于原点沿向量 b 的平移；侵蚀 $X\Theta B=\bigcap_{b\in B}X_{-b}$。根据膨胀和侵蚀运算可以定义多种多样的运算，如开运算 $X\circ B=(X\Theta B)\oplus B$，闭运算 $X\cdot B=(X\oplus B)\Theta B$ 等等[10,11]。

图 1 空间数据库的模拟例子，数据库 a 中聚类的直径相差悬殊且彼此间距离较近，数据库 b 和 c 为凹形聚类，数据库 d 为在 c 上加上了噪音，均为比较复杂的情况，用常规的分割算法以及 CLARANS

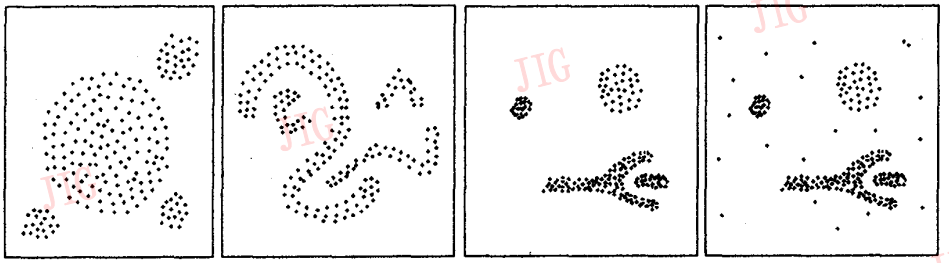

图 1 空间数据库例子

Fig. 1 Examples of spatial databases

算法均不能得到正确的结果[7]。我们可以容易地看出聚类情况,这是由于属于一类的目标在空间分布上比较集中,即任一点到同一类中相邻点的距离大于它到其它类中任意点的距离。需要注意的是,任一点到本类重心的距离未必会小于它到另一类重心的距离,在聚类的大小相差悬殊或聚类形状非凸时就可能出现这种情况,这也就是 k-mean 和 k-medoid 等分割算法缺陷之根源。基于这种认识,我们提出一种基于数学形态学的聚类算法(Mathematical Morphology based Clustering algorithm) MMC,该算法循环地用由小到大的圆形结构元对空间目标进行闭运算连接相邻的目标,连通区即为聚类,开始每个目标为一类,随着结构元由小到大变化,聚类数逐渐减少,到最后用足够大的结构元进行闭运算时,所有的目标连接成一类。每一次闭运算后,用连通区着色方法统计连通区数即聚类数,根据结构元由小到大得到聚类数由多到少的变化规律,提出了一个启发式方法确定最优聚类数。

在参考文献[10]中,作者在理论上研究了数学形态学用于聚类的问题,归纳提出了八种基本运算,其基本思想是根据样本分布的形状来区分不同的样本集,比如区分团状分布与线状分布的样本,这与我们所指的空间数据库聚类的要求并不一致,其所列方法是对理论上可能性的探讨,无法直接应用于空间数据库聚类中。而本文提出的 MMC 算法是一个面向实际应用的算法。

数学形态学算法是针对栅格图像的,若空间数据库为栅格型的,可直接应用 MMC 算法,若空间数据库为矢量型的,可将矢量转换成栅格,转换时可采用较大的采样间距,只要能保持目标间邻接的拓扑关系不变即可。由于是在栅格图像中处理,对于点状、线状和面状目标的处理是一样的,一个算法能同时适用于所有的目标类型,这是 MMC 算法的优越性之一。

MMC 算法采用一系列由小到大的圆形结构元,在离散空间中用多边形来逼近,下图中分别为半径为 1、2、3、4、5 的圆形结构元,结构元的原点在中心,像元值均为 1。

图 2　圆形结构元

Fig. 2　Circular structure elements

MMC 算法描述如下:

输入:空间数据库中感兴趣的数据 X(背景为 0,目标为 1)。

输出:聚类结果 Y(同一类中值相同,不同类值不同)。

过程:

初始化 $i = 1$;

(1) 建立圆形结构元 B_i,其半径为 i;

(2) 闭运算 $Y_i = X \cdot B_i$;

(3) 统计 Y_i 的连通区数即聚类数 n_i,如果 $n_i > 1$,则 $i = i+1$,转(1);否则继续(4);

(4)根据 n_i 计算最优聚类数 n_k,得到对应的结构元半径 k;

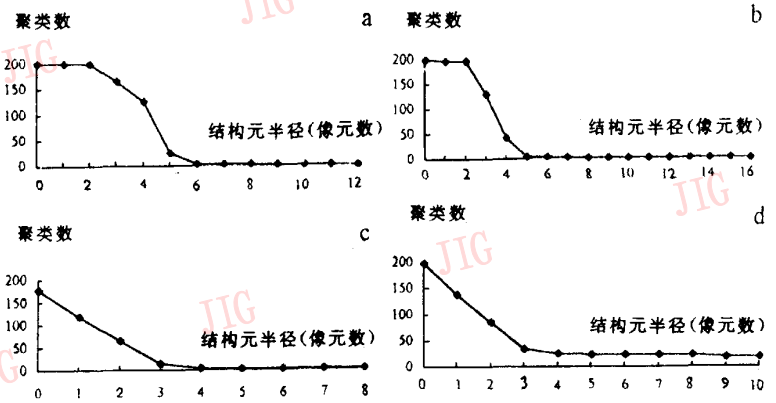

图 3　聚类状态图

Fig. 3　Clustering states(Clusters VS. iteration times)

(5)$Y = X \cdot B_k$;

(6)对 Y 的连通区着色,不同聚类着不同色。

连通区着色方法可采用条件序贯膨胀的形态学方法,或常规的种子蔓延填充方法。由于采用闭运算,按最优聚类数得到的结果为聚类的凹包,可方便地应用于后续的形状分析与识别中。

最优聚类数的确定是 MMC 算法的关键。在对图 1 的试验中,我们将聚类数随结构元变化情况绘于图 3,称为聚类状态图,从图中我们观察到与最优聚类数对应的点位于图形中由多到少变化的腰部,其左边变化显著,右边变化平缓或不变。在聚类状态一阶导数图 4 上,最优聚类点在局部极大值右部相邻点,在聚类状态二阶导数图 5 上对应极大值点。根据这一启发式信息,搜索聚类状态二阶导数的极大值就可以得到最优聚类数。值得注意的是,对于数据库 d 中有噪音的情况,MMC 算法也能得到正确的结果,可以根据预先设定的聚类点数阈值滤除噪音点。

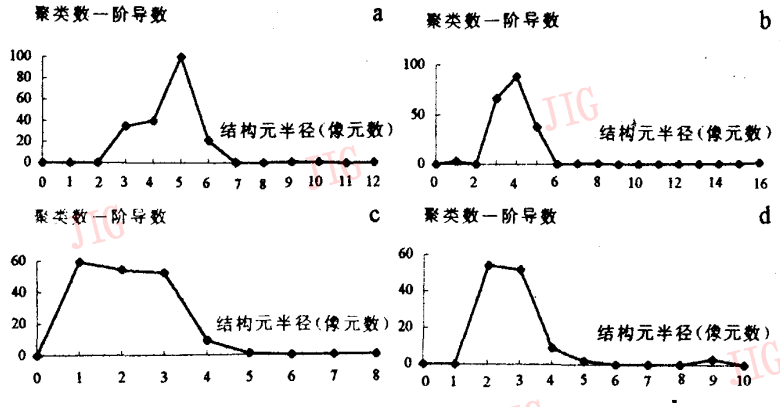

图 4 聚类状态一阶导数图

Fig. 4 First order derivatives of clustering states

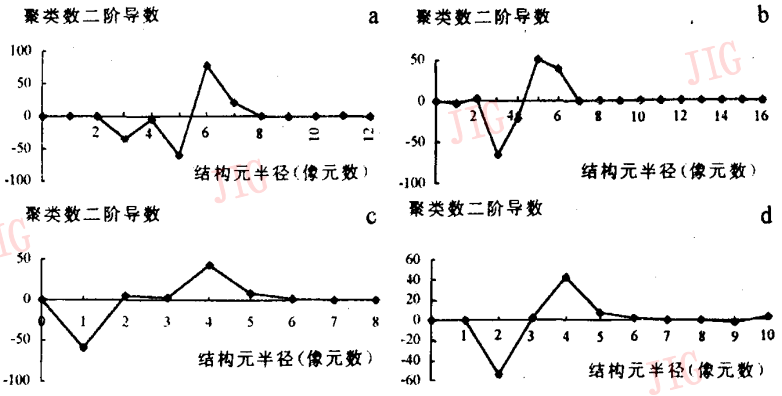

图 5 聚类状态二阶导数图

Fig. 5 Second order derivatives of clustering states

MMC 算法是根据距离来聚类的,其基本假设是任一点到同一类中相邻点的距离大于它到其它类中任意点的距离,即属于一类的目标在空间分布上比较聚集。这一假设在空间数据库中是普遍满足的。至于聚类中不同形态的聚类的区分问题,可以在聚类完成以后根据形态特征加以区分和识别。

上面的算法是针对二维空间的,绝大多数空间数据库中的数据是分层以二维形式存贮的,因此算法有广泛的适用性,也可以将其扩展成三维形式,应用于三维空间数据库的聚类中。

4　MMC 算法在矢量型空间数据库中实现的探讨

上述的 MMC 算法也是针对栅格型空间数据库的,对于大量的矢量型空间数据的聚类分析,可以将感兴趣的矢量数据转换成栅格形式处理,但这毕竟增加了工作量,这里我们探讨 MMC 算法在矢量型空间数据库中实现的可能性与方法。

缓冲区(buffer)分析是地理信息系统中常用的空间分析方法,它与数学形态学的膨胀与侵蚀运算是一致的,缓冲区的大小对应于圆形结构元的半径,当缓冲区为正时相当于膨胀,缓冲区为负时相当于侵蚀。数学形态学的平移和反射在矢量 GIS 系统中更易实现。这几种基本运算都能在矢量形式下实现,所以,在矢量型 GIS 中实现数学形态学中圆形结构元的算法是完全可行的,其它形状结构元的算法原则上也可以实现,我们将在另文中展开论述。

设空间数据集为 X,缓冲区半径为 r,$r>0$,buffer(X, r) 为对 X 进行缓冲区操作的结果,另设 B 为半径为 r 的圆形结构元,则 buffer(X, r)和 buffer(X, $-r$)分别等同于膨胀 $X\oplus B$ 和侵蚀 $X\Theta B$,开 $X\circ B$ 和闭 $X\cdot B$ 分别用 buffer (buffer (X, $-r$), r)和 buffer (buffer(X, r), $-r$)来实现。用由小到大变化的 r 对空间目标进行正负缓冲区操作(即先做正缓冲区紧接着做负缓冲区),统计包含原始目标的多边形数即聚类数。根据聚类数随缓冲区半径由小到大而由多变少的规律,用同样的启发式方法自动确定最优聚类数。用聚类点数阈值去除噪音。不包含原始空间目标的小多边形为空洞,对聚类而言是无效的,聚类的结果为有效的聚类多边形,每个多边形有着不同的标识,这些多边形即为聚类的凹包。同样地,矢量 MMC 算法同时适用于点状、线状和面状目标。

MMC 的矢量算法与栅格算法完全对应。在栅格算法中圆形结构元的半径大小即为迭代次数 i;在矢量算法中可定义初始缓冲区半径 r_1 及半径随迭代次数的递增量 Δr,在第 i 次迭代时缓冲区半径 $r_i=r_1+(i-1)*\Delta r$ 。Δr 应根据空间数据库的比例尺和精度来确定,它相当于栅格数据的像元大小。Δr 太小则算法迭代次数增加,速度降低;Δr 太大则可能会得不到合适的聚类数。

MMC 的矢量算法可在现今流行的矢量型 GIS 平台上进行二次开发实现,例如用 MapInfo 的 MapBasic 语言,在开发语言中有进行缓冲区操作和判断多边形包含其它目标的命令或语句,可以比较方便地实现 MMC 算法,在其它商用的 GIS 平台上均可通过二次开发实现 MMC 算法。

5　试验及结论

我们用 Microfost C 语言在 MS-DOS 下编程,用串行的方法实现 MMC 算法,用图 1 的数据进行了试验,对于数据库 a、b、c、d,MMC 算法自动确定的最优聚类数均为 4 类 (对于 d 用 3、4 或 5 作为聚类点数阈值将噪音滤掉),对应的闭运算结构元半径分别为 6、5、4、4,聚类结果见图 6。灰色的连通区即为聚类,用不同浓淡的灰色表示不同的聚类。与人眼的观察完全一致。另外,按照第 4 节中的设想,我们用 MapInfo 4.0,对 MMC 算法在矢量型 GIS 中的实现方法进行了简单的试验,得到了完全相同的聚类结果,说明设想是正确的,也证明了数学形态学算法可以在矢量系统中实现。

由算法分析和试验结果可知,MMC 算法在理

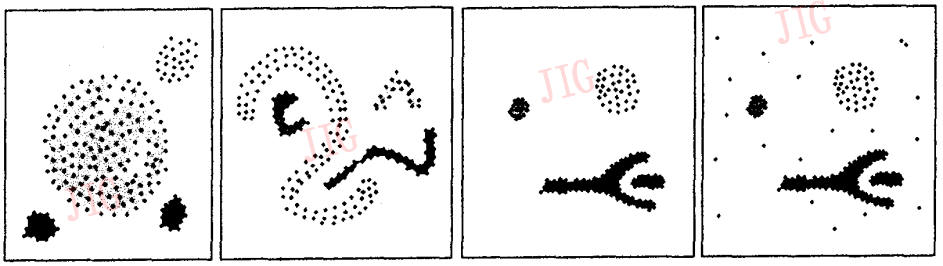

图 6　聚类结果图

Fig. 6　Clustering results

论上是合理的，在实践上是可行的。该算法可以处理任意形状的聚类；用启发式方法自动确定最优聚类结果，在存在噪音的情况下仍能得到最优结果，具有稳健性；适用于点状、线状和面状目标；并且可以在矢量系统中实现，在空间数据发掘和知识发现中具有实用价值。同时数学形态学算法本质上是并行的，便于并行高速处理，应用于未来的并行空间数据库时更具有优越性。

由于 MMC 算法通过迭代计算确定最优聚类数，本身比较耗时，同时由于我们是用串行方式实现该算法，当感兴趣的空间数据规模不是很大时效率尚可，而当算法用于大型空间数据库时整体效率还不够高。因此，MMC 的并行算法、能够减少迭代次数的快速算法、在串行条件下 MMC 的快速算法、在实际的大型空间数据库中的应用以及与其它算法的深入比较等，是值得进一步研究的内容。

参考文献

1 Frawley W, Piatesky-Shapiro G, Matheus C. Knowledge Discovery in Databases: An Overview. In Piatesky-Shapiro G. and Frawley W (Ed.), Knowledge Discovery in Databases, AAAI/MIT Press, 1991.

2 Piatesky-Shapiro G. An Overview of Knowledge Discovery in Databases: Recent Progress and Challenges. In W. P. Ziarko (Ed.), Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer-Verlag, 1994.

3 Han J. Data Mining Techniques. ACM-SIGMOD'96 Conf. Tutorial, June, 1996.

4 Koperski K., Adhihary J., Han J. Spatial Data Mining: Progress and Challenges. SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96), Montreal, Canada, June, 1996.

5 邸凯昌，李德仁. KDD 技术及其在 GIS 中的应用与扩展. 中国 GIS 协会第二届年会，北京，1996.

6 邸凯昌，李德仁，李德毅. 空间数据发掘和知识发现的框架. 第十届全国遥感技术学术交流会，青岛，1997.

7 Ester M. et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, USA, 1996.

8 Ng R T., Han J. Efficient and Effective Clustering Methods for Spatial Data Mining. Proc. of the 20th VLDB Conf., Santiago, Chile, 1994.

9 Ester M. et al. A Database Interface for Clustering in Large Spatial Databases, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, Montreal, Canada, 1995.

10 唐常青等. 数学形态学方法及其应用. 北京：科学出版社，1990.

11 陈晓勇. 数学形态学与影象分析. 北京：测绘出版社，1991.

邸凯昌，1992 年毕业于武汉测绘科技大学，获硕士学位。现任地矿部地质遥感中心工程师，1995 年开始攻读武汉测绘科技大学摄影测量与遥感专业博士生，主要从事数据发掘与知识发现、遥感与地理信息系统的集成理论等研究。

李德仁，博士，中国科学院、中国工程院院士，武汉测绘科技大学校长、教授，现从事 3S 集成理论研究。

李德毅，博士，总参六十一所副所长、高级工程师，主要从事 C^3I、知识发现、模糊控制等领域的研究。

（下转 183 页）

钱昱明,男 26 岁,自动控制工学硕士主要研究方向为计算机控制系统、网络通讯、图象处理等,目前在中兴新通讯研究所从事通讯研究。

Dynamic Images' Motion Vector Multi-Tracing Method and its Realization

Huangfu Zhengxian, Qian Yuming

(*Southeast university, Nanjing* 210096)

Abstract In the process of dynamic images' motion vector estimation, normal logarithm vector searching method has high probality of not able to find the true motion vector. The reason that causes such a problem is analysised in the paper and a new motion vector estimation method called multi-tracing estimation method is introduced. In the process of image matching, a sub-sample stencil is used by this multi-tracing vector estimation method in order to reduce the quantity of calculation. Many frames of image are tested by this method. The result shows that this method can adapt the multi-pole matching image situation better than logarithm vector searching one and can find the motion vector more accurately.

Keywords Image compression, Image matching, Motion vector, Multi-tracing, Subsample-stencil

(上接 178 页)

A Mathematical Morphology Based Algorithm for Discovering Clusters in Spatial Databases

Di Kaichang, Li Deren

(*Wuhan Technical University of Surveying and Mapping, Wuhan* 430070)

Li Deyi

(*The Institute of China Electronic System Engineering, Bejing* 100036)

Abstract Cluster analysis is an important technique for data mining and knowledge discovery in spatial databases. Its main advantage is the ability to find interesting structures or clusters directly form the spatial data without using any background knowledge. Some available algorithms are reviewed and a mathematical morphology based clustering algorithm (MMC) is presented in this paper. Clusters with arbitrary shape can be discovered by using MMC algorithm, and the optimal cluster number is automatically determined by a heuristic method. The algorithm can be implemented in vector databases as well as in raster databases. The experiments show that the new algorithm is feasible and effective for discovering clusters in spatial databases and is robust when clustering in databases with noise.

Keywords Clustering algorithm, Mathematical morphology, Data mining and knowledge discovery, Spatial database