云理论及其在空间数据发掘和知识发现中的应用*

邸凯昌[1),2)] 李德毅[3)] 李德仁[2)]

[1)](国土资源部航空物探遥感中心，北京 100083)

[2)](武汉测绘科技大学信息工程学院，武汉 430079)

[3)](中国电子系统工程研究所，北京 100036)

摘 要 云理论是以研究定性定量间的不确定性转换为基础的系统处理不确定性问题的一种新理论，包括云模型、虚云、云运算、云变换、不确定性推理等内容。云理论为数据发掘和知识发现中的许多基础性关键问题提供了新的解决方法，如概念和知识表达、定性定量转换、概念的综合与分解、从数据中生成概念和概念层次结构等。云模型是定性定量间的不确定性转换模型，它用期望值 Ex、熵 En 和超熵 He 表征定性概念，将概念的模糊性和随机性集成在一起。该文介绍了云理论的基本原理和方法及其在空间数据发掘和知识发现中的应用，重点阐述了云模型及其算法。

关键词 云理论 云模型 虚云 云变换 空间数据发掘和知识发现 概念层次结构

0 引 言

近年来，数据发掘和知识发现(Data Mining and Knowledge Discovery, DMKD)成为计算机领域的研究热点，空间数据发掘和知识发现(Spatial Data Mining and Knowledge Discovery, SDMKD)也开始引起重视[1—3]。在数据发掘和知识发现研究中，人们主要集中于对数据发掘算法的研究，而对知识表达、定量定性转换、不确定性推理等一些基础性关键问题则研究得较少，基本上是沿用人工智能以往的研究成果。

在人工智能领域，对知识和推理的不确定性主要分成模糊性和随机性两种展开研究。作为处理模糊性问题的主要工具，模糊集理论用隶属度来刻画模糊事物的亦此亦彼性。然而，一旦用一个精确的隶属函数来描述模糊集，模糊概念就被强行纳入到精确数学的王国，从此以后，在概念的定义、定理的叙述及证明等数学思维环节中，就不再有丝毫的模糊性了。这正是传统模糊集理论的不彻底性。

针对这一问题，李德毅教授在传统模糊集理论和概率统计的基础上提出了定性定量不确定性转换模型——云模型[4,5]。以云模型为基础经过系统研究和发展形成了云理论，包括云模型、虚云、云运算、云变换、不确定性推理等内容，在 DMKD 和 SDMKD 中有着广泛的应用[5—8]。

1 云模型(Cloud model)

1.1 云的基本概念

云是用语言值描述的某个定性概念与其数值表示之间的不确定性转换模型，或者简单地说云模型是定性定量间转换的不确定性模型。设 U 是一个论域 $U=\{x\}$，T 是与 U 相联系的语言值。U 中的元素 x 对于 T 所表达的定性概念的隶属度 $C_T(x)$(或称 x 与 T 的相容度)是一个具有稳定倾向的随机数，隶属度在论域上的分布称为隶属云，简称为云。

$C_T(x)$ 在 $[0,1]$ 中取值，云是从论域 U 到区间 $[0,1]$ 的映射，即：

$$C_T(x): U \to [0,1]$$

$$\forall x \in U \quad x \to C_T(x)$$

图 1 显示了语言值“约 20 公里”的隶属云。云的几何形状对理解定性与定量间转换的不确定性有很好的帮助。首先，所有 $x \in U$ 到区间 $[0,1]$ 的映射是一对多的转换，x 对于 T 的隶属度是一个概率分布而非固定值，从而产生了云，而不是一条明晰的隶属曲线。第二，云由许许多多的云滴组成，一个云滴是

* 本文研究受国家自然科学基金及测绘遥感信息工程国家重点实验室基金项目(No. 49631050 及 No. WKL(97)0302)资助
收稿日期：1999-10-27

定性概念在数量上的一次实现，单个云滴可能无足轻重，在不同的时刻产生的云的细节可能不尽相同，但云的整体形状反映了定性概念的基本特征。云滴的分布类似天上的云，远看有明确的形状，近看没有确定的边界，这就是我们用“云”来命名它的原因。第三，云的数学期望曲线（Mathematical Expected Curve，MEC）从模糊集理论的观点来看是其隶属曲线。第四，云的“厚度”是不均匀的，腰部最分散，“厚度”最大，而顶部和底部汇聚性好，“厚度”小。云的“厚度”反映了隶属度的随机性的大小，靠近概念中心或远离概念中心处隶属度的随机性较小，而离概念中心不近不远的位置隶属度的随机性大，这与人的主观感受相一致。

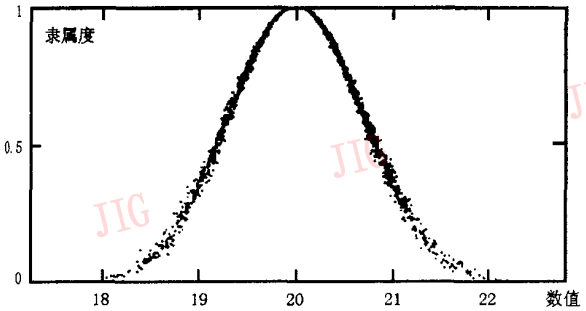

图 1　语言值“20 公里左右”的隶属云

1.2　云的数字特征

云的数字特征用期望值 Ex(Expected Value)、熵 En(Entropy)、超熵 He(Hyper Entropy)3 个数值来表征(如图 2)。

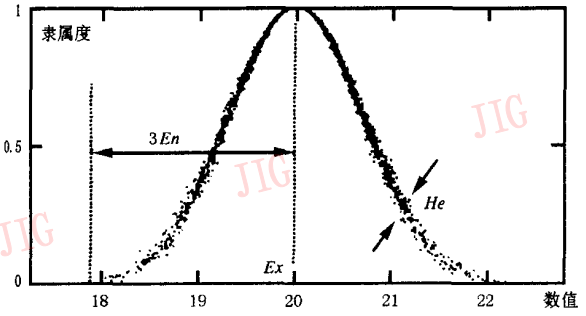

图 2　云的数字特征值的示意图

期望值 Ex：是概念在论域中的中心值，是最能代表这个定性概念的值，换句话说，它 100%地隶属于这个定性概念。

熵 En：是定性概念模糊度的度量，反映了在论域中可被这个概念所接受的数值范围，体现了定性概念亦此亦彼性的裕度。熵越大，概念所接受的数值范围也越大，概念越模糊。

超熵 He：可谓熵 En 的熵，反映了云滴的离散程度。超熵越大，云滴离散度越大，隶属度的随机性越大，云的“厚度”也越大。

可见，云模型的 3 个数字特征值把模糊性(定性概念的亦此亦彼性)和随机性(隶属度的随机性)完全集成到一起，构成定性和定量相互间的映射，作为知识表示的基础。

1.3　正态云模型

正态云模型在表达语言值时最常用，其数学期望曲线 MEC 为：

$$MEC_A(x) = \exp[-(x - Ex)^2/(2En^2)]$$

正态云的生成算法如下：

(1) $x_i = G(Ex, En)$。生成以 Ex 为期望值、En 为标准差的正态随机数 x_i；

(2) $En'_i = G(En, He)$。生成以 En 为期望值、He 为标准差的正态随机数 En'_i；

(3) 计算 $\mu_i = \exp\left[-\frac{(x_i - Ex)^2}{2En'^2_i}\right]$，令 (x_i, μ_i) 为云滴。

给定正态云的 3 个数字特征值 (Ex, En, He)，可以用上面的算法生成任意个数云滴组成的正态云。该算法生成的云自然地具有不均匀厚度的特性，云的腰部、顶部、底部等并不需要精确地定义，3 个数字特征值足以很好地描述整个云的形态。

1.4　云发生器

云的生成算法可以用软件的方式实现，也可以固化成硬件实现，称为云发生器（Cloud Generator）。由云的数字特征产生云滴，即实现从定性到定量的转换，称为正向云发生器(如图 3)，上面的云生成算法即为正向云发生器算法。

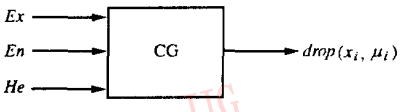

图 3　正向云发生器

云可以根据不同的条件来生成，在给定论域中特定的数值 x 的条件下的云发生器称为 X 条件云发生器，给定特定的隶属度值 μ 的条件下的云发生器称为 Y 条件云发生器。X 条件云发生器生成的云滴位于同一条竖直线上，横坐标数值均为 x，纵坐标隶属度值呈概率分布。Y 条件云发生器生成的云滴

位于同一条水平线上,被期望值 Ex 分成左右两组,纵坐标隶属度值均为 μ,两组横坐标数值分别呈概率分布。两种条件云发生器是运用云模型进行不确定性推理的基础。

X 条件云发生器算法如下:

(1) $En'_i=G(En,He)$。生成以 En 为期望值、He 为标准差的正态随机数 En'_i;

(2) 计算 $\mu_i=\exp\left[-\frac{(x-Ex)^2}{2En'^2_i}\right]$,令 (x,μ_i) 为云滴。

Y 条件云发生器算法如下:

(1) $En'_i=G(En,He)$。生成以 En 为期望值、He 为标准差的正态随机数 En'_i;

(2) 计算 $x_i=Ex\pm\sqrt{-2\ln(\mu)}En'_i$,令 (x_i,μ) 为云滴。

给定符合某一正态云分布规律的一组云滴作为样本 (x_i,μ_i),产生云所描述的定性概念的 3 个数字特征值 (Ex,En,He),即从定量到定性的转换,其软件或硬件实现称为逆向云发生器,如图 4 所示。正向云发生器和逆向云发生器相结合,实现定性与定量的随时转换。

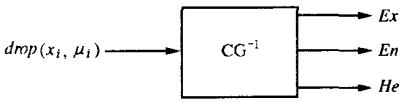

图 4 逆向云发生器

逆向云发生器算法如下:

(1) $Ex=mean(x_i)$

(2) $En=stdev(x_i)$

(3) $En'_i=\sqrt{\frac{-(x_i-Ex)^2}{2\ln(\mu_i)}}$, $He=\mathrm{stdev}(En'_i)$

其中,$mean()$、$stdev()$ 分别为求均值和标准差的函数。

以上的逆向云发生器算法是一种统计方法,求出的数字特征值是一种估计值,当云滴数较少时,误差可能比较大,随着云滴数的增加,误差将减小。当云滴数很少时,采用最小二乘法为宜。最小二乘法拟合精度高,但算法复杂。

1.5 二维和多维正态云模型

在空间数据库中,很多概念是由多个密不可分的因素决定的,例如,地理位置由经度和纬度两个值确定,彩色由红、绿、蓝 3 个分量确定,它们对应的论域为二维和三维论域。沿着上面一维云模型的的思路,我们把云模型扩展至二维和多维,使之适用于描述二维和多维语言值表示的定性概念。

二维云模型是用语言值描述的某个定性概念与其二维数值表示之间的不确定性转换模型。二维云的概念可以绘成三维图形,图 5 是语言值“中心”对应的二维云的表面图。从图中我们可以看到它象一个坟头或山包,在山顶和山脚较光滑、变化缓慢,而在山腰表面粗糙、变化剧烈,这说明二维云的“厚度”是不均匀的,在山腰分散,在山顶和山脚更汇聚。因此,二维云是一维云的自然扩展。

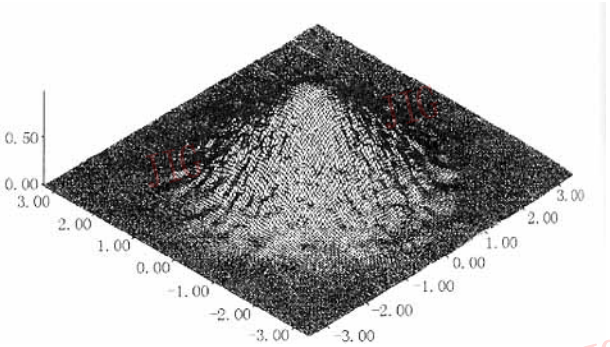

图 5 语言值“中心”对应的二维云的表面图

设二维论域的两维之间不相关,二维正态云可以用 6 个数字特征量来描述:(Ex,Enx,Hx,Ey,Eny,Hy)。其中 Ex 和 Ey 为期望值,Enx 和 Eny 为熵,Hx 和 Hy 为超熵。

(Ex,Ey) 是二维云表示的定性概念在论域中的中心值,是最能代表这个二维概念的值。Enx 和 Eny 是概念的模糊测度,它们反映了该语言值对二维数值的可覆盖程度。Hx 和 Hy 分别是 Enx 和 Eny 的熵,反映了云滴的离散程度。

二维正态云的数学期望曲面(Mathematical Expected Surface, MES)方程为:

$$MES_A(x,y)=\exp\left[-\frac{1}{2}\left[\frac{(x-Ex)^2}{Enx^2}+\frac{(y-Ey)^2}{Eny^2}\right]\right]$$

MES 向 x-y 平面的投影为一个椭圆面(如果 Enx 与 Eny 相等则是圆面)。当椭圆的两个轴与 x 和 y 轴不平行时,需要增加一个数字特征量 θ 来描述二维云,称为旋转云。同理,可以把正态云模型进一步推广至多维正态云模型。

1.6 二维正态云发生器

给定一组数字特征量 (Ex,Enx,Hx,Ey,Eny,Hy),二维正向正态云发生器(见图 6)可以产生任意多个云滴 (x_i,y_i,μ_i),其中 (x_i,y_i) 服从二维正态分布,μ_i 服从一维概率分布。

二维正向正态云发生器的算法如下：

(1) $(x_i, y_i)=G(Ex, Enx, Ey, Eny)$。生成以$(Ex, Ey)$为期望值、$(Enx, Eny)$为标准差的二维正态随机数$(x_i, y_i)$，具体实现方法是先产生两个一维标准正态随机数 t_0 和 t_1，计算 $x_i=Enx \cdot t_0+Ex$，$y_i=Eny \cdot t_1+Ey$，则(x_i, y_i)为符合要求的二维正态随机数；

(2) $(Enx'_i, Eny'_i)=G(Enx, Hx, Eny, Hy)$。生成以$(Enx, Eny)$为期望值、$(Hx, Hy)$为标准差的二维正态随机数$(Enx'_i, Eny'_i)$；

(3)计算 $\mu_i=\exp\left[-\frac{1}{2}\left[\frac{(x_i-Ex)^2}{Enx'^2_i}+\frac{(y_i-Ey)^2}{Eny'^2_i}\right]\right]$，令$(x_i, y_i, \mu_i)$为云滴。

同理得到二维 X 条件云发生器和 Y 条件云发生器的算法，两种二维条件云发生器是运用二维云模型进行二因素和多因素不确定性推理的基础。

二维逆向云发生器从给定符合二维正态云分布规律的一组云滴样本(x_i, y_i, μ_i)，产生云所描述的二维定性概念的 6 个数字特征值$(Ex, Enx, Hx, Ey, Eny, Hy)$，如图 7 所示。二维逆向云发生器算法如下：

(1) $Ex=mean(x_i)$，$Ey=mean(y_i)$；

(2) $Enx=stdev(x_i)$，$Eny=stdev(y_i)$；

(3) 从样本中选取 $y=Ey$ 的云滴(x_j, y_j, μ_j)，计算 $Enx'_j=\sqrt{\frac{-(x_j-Ex)^2}{2\ln(\mu_j)}}$，$Hx=stdev(Enx'_i)$；

(4) 从样本中选取 $x=Ex$ 的云滴(x_k, y_k, μ_k)，计算 $Eny'_k=\sqrt{\frac{-(y_k-Ey)^2}{2\ln(\mu_k)}}$，$Hy=stdev(Eny'_k)$。

算法中第(3)、(4)两步，分别从样本中选取坐标值等于期望值的样本子集，目的是为了将 Hx、Hy 分开计算，具体编程实现时，可以取坐标值与期望值相差很小的子集。由此也可以看出，二维逆向云发生器需要很多的云滴样本。

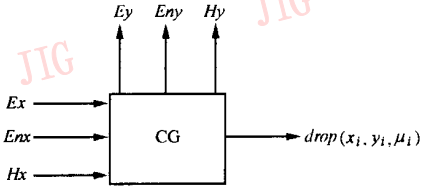

图 6　二维正向正态云发生器

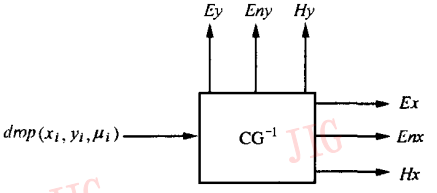

图 7　二维逆向云发生器

我们不仅将一维正态云模型扩展至二维和多维云模型，而且还将正态云扩展至多种分布的云，如将Γ型、三角形、梯形的隶属函数扩展至Γ云(Gamma云)、三角形云、梯形云，同样也完全可以将模糊集理论中的任何隶属函数扩展为隶属云。

2　虚云、云运算、云变换及不确定性推理

虚云(Virtual Cloud)，又称虚拟云，是指将给定的云的数字特征进行某种运算，得到新的数字特征所构造的云[5-7]。这些给定的、用来计算虚云的云称为基云。而虚云所表达的语言值称为虚拟语言值。虚云主要有浮动云、综合云、分解云、几何云等几类。

浮动云的熵和超熵值是由基云的熵和超熵值线性内插得到。浮动云方法可以在论域中给定的语言值没有覆盖的空白区自动生成虚拟语言值。

综合云用于将两个或多个同类型(对应同一个语言变量)的语言值综合为一个更广义的语言值。综合云的熵大于基云的熵，它覆盖了更大范围的论域空间。如果我们从低的概念层次到高的概念层次迭代地应用综合云构建方法，就可以得到语言变量的概念层次结构。

沿着与构建综合云相反的思路，我们也可以把一个基云进行分解，形成若干个虚拟云，称为分解云。分解云熵的和等于基云的熵，因此在论域中覆盖的范围是一致的，单个分解云的熵小于基云的熵，对概念有更细致的描述。分解云用于概念树层次间的概念细化操作，给定一个高层次概念，可以分解为若干个低层次概念。分解云与综合云相结合可以实现概念层次结构上的自由下降与提升。

已知云的某些局部特征，例如若干云滴，我们可以用数学拟合方法生成一个完整的虚云，称为几何云。

云运算包括代数运算、逻辑运算、语气运算等，运算过程中能够保持和传播不确定性。

云变换(Cloud Transform)将任意函数(如直方图)分解为基云的叠加。我们提出了一种启发式云变换算法，能够在限定的误差范围内将任意数据分布分解为多个云的叠加。

基于云理论的不确定性推理，包括单条件单规

则推理、多条件单规则推理、单条件多规则推理、多条件多规则推理等。由于云模型的特性,使得基于云的不确定性推理能较好地解决不确定性的表达和传播问题。

3 云理论在空间数据发掘和知识发现中的应用

云理论在空间数据发掘和知识发现中有着广泛的应用,如用于概念和知识表达、定量定性转换、概念的综合与分解、从数据中生成概念和概念层次结构、不确定性推理和预测、与其它数据发掘算法的结合等。限于篇幅,这里仅展开介绍云理论用于 SDMKD 中的概念和知识表达、概念和概念层次结构的生成,云理论在 DMKD 和 SDMKD 中的应用实例见文献[5—8]。

3.1 用于概念和知识表达

SDMKD 从空间数据中发现概念和模式,空间概念的表达是一个基础性的关键问题。我们用多种云模型表达多种空间概念,提出不确定点、不确定线、不确定面、不确定方向、不确定距离等的云模型,同时定义某概念所有隶属度为 1 的元素集合为概念的核。

不确定点用二维正态云模型表示,用来表达诸如"人民英雄纪念碑周围"、"山顶附近"等空间概念,不确定点的核为一个点。不确定线用二维云模型来表示,它的核是一条线(直线或曲线),沿线的方向为均匀分布,线的垂直方向为一维正态云。不确定线云模型用来表达诸如"草地与林地的边界线"、"京广线沿线"等空间概念。不确定面用二维云模型来表示,它的核是一个面(多边形),在核多边形的内部隶属度均为 1,在核多边形的外部依点到多边形的距离呈半正态云。不确定面云模型用来表达诸如"北京中心地区"、"北京郊区"、"大兴安岭林区"等空间概念。不确定方向用二维云模型来表示,其核为一条直线。不确定方向用来表达"东"、"西"、"南"、"北"、"东北"、"东南"、"北偏东"等语言值。不确定距离用一维云模型表示,一般用半降正态云表达"近"、"很近"等语言值,用半升正态云表达"远"、"很远"等语言值,用完整正态云表达"20 公里左右"、"距离适中"等语言值,而用梯形云表达"不远不近"等覆盖度较宽的语言值。

在用云模型表达空间和非空间概念的基础上,我们就可以利用云发生器表达知识。一条单条件定性规则"If A,then B。"可以用 A 的 X 条件云发生器与 B 的 Y 条件云发生器连接起来表达,一条多条件定性规则"If $A_1,A_2,\cdots,A_n$ then B。"可以用 A 的多维 X 条件云与 B 的 Y 条件云发生器连接起来表达。

3.2 用于从数据中生成概念和概念层次结构

从数据中产生概念是 SDMKD 中属性归纳和概念提升的关键问题,我们提出 3 种基于云模型的概念生成方法:基于云变换的数据驱动法、基于黄金分割率的模型驱动法以及基于虚云的方法。

云变换方法是一种数据驱动的"客观"方法,当数据量充分大时,数据直方图能很好地表示数据的分布特点,这种情况下用云变换方法从数据中生成概念比较有利。若从大到小给定一系列误差容限实施云变换,则可以得到由少到多多个层次的一系列的云模型,从而形成概念层次结构。

当数据量较小时,数据直方图不能充分地表示数据的分布特点,因而云变换方法的结果也就不理想。在这种情况下,我们提出一种"主观"方法,称为基于黄金分割率的模型驱动法。其基本思想是,给定的属性(即论域)看成语言变量,每个语言变量有几个语言值,语言值用云模型来表达,越接近论域的中心,云的熵和超熵越小,越远离论域的中心,云的熵和超熵越大,相邻云的熵和超熵的较小者是较大者的 0.618 倍。一般取奇数个云,如 3 个或 5 个。基于黄金分割率的方法可以生成概念层次结构,依次分别生成 1 个、3 个、5 个、…云模型来表达定性概念,则这些云模型就形成了一个概念层次结构。

若用户直接给出了几个关键的云模型,则可以用虚云方法生成概念层次结构,即用浮动云方法给定的概念层次上论域的空白区生成云模型,用综合云方法生成高层次概念,用分解云方法生成低层次概念,三者相结合生成概念层次结构。

基于云模型的概念层次结构是对论域的一种软划分,相邻云之间允许有重叠,高概念层次上的一朵云包含低概念层次上几朵云的信息,低概念层次上的一朵云可能被高概念层次上的多朵云包含,由于云的特性,云模型重叠区范围内的数值在不同的机会可能属于不同的云,在非重叠区范围内的数值在不同的机会属于同一朵云的隶属度也不同,是随机值。因此基于云模型的概念层次结构与传统的概念层次结构有显著的区别,在表达和处理概念层次的不确定性方面

优于传统的概念层次结构,是一种"软"的、更广义的结构,我们称之为泛概念层次结构。

4 结束语

云模型用期望值 Ex、熵 En 和超熵 He 表征定性概念,将定性定量转换中的模糊性和随机性集成到一起,克服了模糊集理论中隶属函数的固有缺陷,加上虚云、云变换等新方法,使云理论为空间数据发掘和知识发现中的许多基础性关键问题提供了新的解决方法,有着广泛的应用。同时,云理论在数据处理、模式识别、智能控制等多方面都有着广泛的潜在的应用,值得进一步研究和发展。

参 考 文 献

1 Frawley W, Piatesky-Shapiro G, Matheus C. Knowledge discovery in databases: An overview. In: Piatesky-Shapiro G, Frawley W (eds), Knowledge Discovery in Databases, AAAI/MIT Press, 1991.

2 Li Deren, Cheng Tao. KDG: Knowledge Discovery from GIS——Propositions on the use of KDD in an intelligent GIS. In: Proc ACTES, The Canadian Conf GIS, 1994.

3 邸凯昌,李德仁,李德毅. 空间数据发掘和知识发现的框架. 武汉测绘科技大学学报,1997,22(4):328～332.

4 李德毅,史雪梅,孟海军. 隶属云和隶属云发生器. 计算机研究与发展, 1995,42(8):32～41.

5 Li Deyi, Han J, Chan E, Shi Xuemei. Knowledge representation and discovery based on linguistic atoms. In: Proc the 1st Pacific-Asia Conf KDD & DM, Singapore, Feb. 1997.

6 Li Deyi, Di Kaichang, Li Deren, Shi Xuemei. Mining association rules with linguistic cloud models. In: PAKDD'98, The Second Pacific-Asia Conference on Knowledge Discovery & Data Mining, Melbourne, Australia, April. 1998.

7 Di Kaichang, Li Deyi, Li Deren. Knowledge representation and discovery in spatial databases based on cloud theory. International Archives of Photogrammetry and Remote Sensing, Vol. 32, Part 3/1, Columbus, Ohio, July. 1998.

8 Di Kaichang, Li Deren, Li Deyi. Intelligent query in spatial databases based on cloud model. In: Li Deren *et al.* (eds) Spatial Information Science, Technology and Its Applications, WTUSM Press, SIST'98, Wuhan, Dec 1998.

Cloud Theory and Its Applications in Spatial Data Mining and Knowledge Discovery

Di Kaichang[1),2)], Li Deyi[3)] and Li Deren[2)]

[1)] (*Center for Remote Sensing in Geology, Beijing* 100083)

[2)] (*School of Information Engineering, Wuhan Technical University of Surveying and Mapping, Wuhan* 430079)

[3)] (*Institute of China Electronic System Engineering, Beijing* 100036)

Abstract Cloud theory is a new theory handling uncertainty based on the uncertain transition between qualitatives and quantitatives. The theory includes cloud model, virtual cloud, cloud operation, cloud transform and uncertainty reasoning. It provides new solutions for many basic problems in data mining and knowledge discovery, such as concept and knowledge representation, transition between qualitatives and quantitatives, concept synthesization and resolution, concept and concept hierarchy generation from data, etc. Cloud model is a model of the uncertain transition between a linguistic term of a qualitative concept and its numerical representation. Cloud model represents a qualitative concept with three digital characteristics, expected value Ex, entropy En and hyper entropy He, which integrate the fuzziness and randomness of a linguistic term in a unified way. This paper presents the fundamentals of cloud theory and its applications in spatial data mining and knowledge discovery, focusing on the cloud models and their algorithms.

Keywords Cloud, Cloud model, Virtual clouds, Cloud transform, Spatial data mining and knowledge discovery, Concept hierarchy