

空间数据挖掘关键问题研究^{*}

肖予钦 景 宁 吴秋云 钟志农

(国防科技大学电子科学与工程学院 长沙410073)

The Research on Key Problems of Spatial Data Mining

XIAO Yu-Qin JING Ning WU Qiu-Yun ZHONG Zhi-Nong

(College of Electric Science and Engineering, National University of Defence Technology, Changsha 410073)

Abstract With the rapid development of remote sensing and mapping technology and the widespread application of spatial database system, the spatial data collected and stored by human expand increasingly. These very large data sets far exceed human's capabilities of comprehending and handling, so the requirement of spatial data mining to provide human with valuable information becomes instant need. In this paper, the methods of spatial data handling in spatial data mining are discussed from a database perspective, the key problems of spatial data mining and their solutions in current study are presented, the relations between spatial data mining and geographical information system are analyzed.

Keywords Spatial data mining, Database, GIS

1. 引言

二十世纪八十年代地理信息系统的突破发展,使得专门管理空间数据的数据库系统——空间数据库系统得到前所未有的广泛使用。近年来随着空间信息技术领域内遥感和测绘技术的飞速发展,包括资源、环境和灾害等的各种空间数据呈指数级数增长。但当面对越来越多的迅速膨胀的超级数据库时,却出现了“数据丰富,知识贫乏”的现象。人们无从着手去理解数据中所包含的信息,更难以获得有价值的信息!空间数据挖掘技术的出现,为人类充分利用海量的空间数据提供了工具。空间数据挖掘能够从空间数据库中提取隐含的预测性信息,找出最有价值的知识来指导科学决策,已经成为国际上研究和应用的热点。

空间数据挖掘,是指对空间数据库中非显式存在的知识、空间关系或其它有意义的模式等的提取^[1],可用于对空间数据的理解、空间关系和空间与非空间数据间关系的发现、空间知识库的构造、空间数据库的重组和空间查询的优化。在地理信息系统、遥感、图像数据库探测、医学图像处理、导航、交通控制、环境研究以及许多使用空间数据的领域中有广泛的应用。

空间数据挖掘是数据挖掘技术在空间数据库领域中的应用。由于存储了大量与空间有关的数据,空间数据库有许多与关系数据库所不同的特性,比如它包含有拓扑和距离信息,通常以复杂的多维空间索引结构组织数据,经常需要进行几何计算和空间推理等复杂操作,这些特性使得从空间数据中挖掘信息面临很多挑战性的问题。

自从空间数据挖掘概念的提出至今,经过几年的发展,其研究已取得了一些成果,提出了挖掘空间特征规则^[2]、空间关联规则^[3]、空间分类规则^[4]和空间聚类规则^[5]的算法,有关学者还对此及时进行了总结^[6-9]。但是,目前空间数据挖掘的研

究工作对空间数据的处理方法重视不够,而且大多数研究者是在总结空间数据挖掘的研究成果时,都是按照所发现的知识类型(如空间特征化与对比,空间关联等)来展开的,缺乏对空间数据处理方法的归纳。由于空间数据不同于传统数据挖掘技术所处理的数字、分类和符号等数据,有其自身的特性,使得空间数据的处理方法成为空间数据挖掘研究中最重要的一部分,并因而形成了空间数据挖掘研究不断发展的内在动力。因此,有必要对空间数据挖掘中空间数据的处理方法进行归纳。裴韬等总结了近年来空间数据挖掘领域内针对空间数据复杂性特征的解决方法,但他们主要是从人工智能、机器学习和地理学的角度来论述的^[10]。由于目前数据库系统和数据仓库系统已经成为主流信息系统,空间数据挖掘系统应当和数据库、数据仓库系统实现集成,因而从数据库技术角度探讨空间数据挖掘对空间数据的处理方法既有理论意义,又有现实意义。

2. 空间数据挖掘研究的关键问题及其解决方法

空间数据自身的特性以及空间数据挖掘系统的设计原则,使得从数据库技术角度探讨空间数据的处理方法成为目前空间数据挖掘研究中面临的最现实的一个问题,也是今后一段时期内研究的重点。有关空间数据挖掘对空间数据处理方法的研究可以归纳为以下几个关键性问题,当前空间数据挖掘的研究工作也应从这几个方面着手展开。

2.1 空间数据的可控制性

Tobler 地理第一定理指出:任何事物之间都存在着相关性,但是距离近的事物之间的相关性比距离远的事物之间的相关性更强。这个定理描述了空间数据的主要特点,那就是距离邻近的空间个体之间存在一定的相互作用。另外,空间数据还具有海量的数据、空间依赖性和空间异构性以及空间维数的增高等复杂性特征,从而使得空间数据挖掘与传统数据挖掘方法之间存在明显的差异^[10,11]。表1给出了传统数据挖掘

^{*} 国防科技大学科研基金资助项目。

和空间数据挖掘的区别,主要包括三个方面:

表1 传统数据挖掘和空间数据挖掘的区别

	传统数据挖掘	空间数据挖掘
数据类型	数字、分类和符号数据	点、线、面等空间对象
处理对象	明确的输入数据	隐含的空间谓词
数据关系	相互独立	有高度的自相关性

空间数据挖掘系统应该增强对空间数据的可控制性,即需要提高对空间数据这类复杂数据类型的处理能力,解决的办法就是综合数据挖掘与空间数据库技术。

(1)将空间数据现有的数据分析技术与数据挖掘的方法集成起来进行研究。空间数据挖掘算法研究中涉及到空间连接、空间合并以及拓扑关系的提取等空间运算^[2,3,12]。比如空间概念分层给出了数据空间上的抽象层次,利用它可以把相邻的空间区域连续地合并成更大的区域,形成更高一层的抽象,从而提取知识。因此能否高效地执行空间合并操作决定了挖掘算法执行的效率。另外在空间数据挖掘领域中,处理空间数据的标准方法是将空间关系实例化成属性,然后用这些新的属性重建模型。这里需要提取空间对象的拓扑关系,高效地计算空间对象的拓扑关系能提高挖掘算法执行的效率。

(2)利用空间数据的访问方法。高效的查询处理是挖掘算法的一个基本需求,空间数据挖掘应用通常处理大量的高维数据,为获得让人满意的性能,多维索引结构可以用来管理数据。如在地理信息系统研究中广泛使用的R树及其变种R*树,可以实现对空间对象的访问^[13,14]。

(3)使用基于最小边界矩形(MBR)的近似计算和精确计算的两步计算优化技术。由于空间数据挖掘算法中常常涉及到计算费用很高的空间运算(如空间连接、提取拓扑关系等),使用两步计算优化技术显得尤为重要。Koperski在研究空间关联规则挖掘算法中就采用了这种优化技术,首先利用R树索引仅对空间对象的MBRs进行概化邻近关系计算,然后对那些满足最小支持度阈值的关系才进行空间对象精确的邻近关系计算^[1]。这样避免了对所有可能的空间关联关系都要进行代价很高的精确邻近关系计算,大大节省了计算费用。

2.2 挖掘算法的有效性和可伸缩性

有效性和可伸缩性是指数据挖掘算法有效处理大量数据的能力。由于空间数据量的不断激增,海量的数据常使一些算法因计算量过大而无法得以实施,一个典型的问题就是所处理的数据不能一次性加载到内存,只能在内存和磁盘高速缓存之间换进换出,这使得挖掘算法的效率变得很低,因而研究新的计算策略和可伸缩算法显得十分重要。

可伸缩性分为行(数据库大小)可伸缩性和列(维)可伸缩性。如果一个数据挖掘系统行数扩大10倍,而执行同样的数据挖掘查询的时间最多也不超过其原来时间10倍的话,则称这个系统是行可伸缩的;如果数据挖掘查询执行时间和列(属性或维)数呈线性增长关系,则称这个系统是列可伸缩的^[1]。

本节主要讨论有关行可伸缩性的问题,即有关海量数据的处理问题。有关列可伸缩性问题,即有关高维数据挖掘算法的问题,我们将在2.4节进行详细的讨论。

针对海量数据的处理,目前提出的解决办法主要有以下5种:

(1)改变算法运行的策略,采用并行计算环境,实施并行算法。

如已经实现了关联规则、序贯模式和SPRINT分类算法

的并行化^[15~17]。

(2)使用划分技术,将所分析的数据集划分成若干个非重叠的部分,使得每一部分都能够放入内存。这样在用算法分别处理每一部分时,只需要一次性读取数据,降低了I/O开销。最后将各部分的处理结果进行综合就形成了最终的结果。

Guha等在处理非球形聚类和存在孤立点的问题时,提出了利用代表点聚类的算法CURE(Clustering Using Representatives),该算法采用划分的方法。首先将一个随机样本划分,对每个划分进行部分聚类,然后这些结果簇被聚类产生最后的结果^[18]。

(3)使用采样技术,即在给定数据集的一个子集上进行分析。选取给定数据集的随机样本,然后用算法对该样本进行分析。这里样本选取的好坏非常重要,好的样本可以提高算法的效率,但差的样本会导致结果质量的下降。样本选取的原则应该是能尽量忠实地反映原始数据的信息。另外,随机样本大小的选取,应该使得可以在内存中进行处理。用这种方法,牺牲了一些精度换取了有效性。

Easter等利用空间数据结构提出了一种使用高质量采样技术的聚类算法CLARANS聚焦法(CLARANS with focusing on representatives),该算法仅对R*树叶节点最中央的对象进行聚类^[19]。由于很好地反映了原始数据的空间分布特征,所有这些R*树叶节点最中央的对象构成了一个高质量的样本。

使用划分和采样的策略,其结论的准确性不如一次使用所有数据的方法高,但当对精度的要求不是很严格时,这两种方法能够有效地处理大数据集。尤其当效率最为重要时,采样方法特别合适。

(4)采用数据预处理技术,在实施算法前使用新的数据结构来重新描述数据。新的数据结构应该能够加载到内存中,并且要很好地保留原始数据内在的信息。

适用于对非常大的训练集进行判定树归纳的算法SLIQ(Supervised Learning In Quest)^[20]和SPRINT(Scalable Parallelizable INduction of decision Trees)^[17],都定义使用新的数据结构来构造判定树。SLIQ使用多个驻留磁盘的属性表和一个驻留内存的类表来描述数据,SPRINT使用不同的属性表数据结构来存放记录标识符和类标号。

利用层次方法的平衡迭代消减聚类算法BIRCH(Balanced Iterative Reducing and Clustering)引入了两个概括描述聚类的概念:聚类特征(clustering feature, CF)和聚类特征树(CF-tree)^[21]。CF给出了对象子聚类信息的汇总描述,从而使子聚类的表示可以总结为对应的一个聚类特征,而不必再用所有的对象来表示。CF-树是一个高度平衡的树,它存储了层次聚类的聚类特征,可以被看作是数据的多层压缩,试图保留数据内在的聚类结构。CF-树可以动态构造,当CF-树建好后,就可以利用任何聚类算法对叶节点进行聚类。

(5)基于约束的挖掘,使挖掘在用户提供的各种约束的指导下进行^[22~24]。它提供了额外的控制方法,能在增加用户交互的同时改进挖掘处理的总体效率。基于约束的挖掘允许用户说明和使用约束,提供元规则(模式模板)和集合/子集联系、变量的常量初始化以及聚集函数等附加规则,引导数据挖掘系统对感兴趣模式的搜索,以得到更有效的和更有作用的挖掘。

2.3 空间依赖性和异构性的描述

空间依赖性和空间异构性是空间数据复杂性特征的表现

之一。有些研究者指出,忽略这些特性将影响空间数据挖掘导出的模式。空间依赖性,在空间统计学中也叫做空间自相关,是指空间位置相近的点之间的属性存在关联关系。空间异构性,是指大部分地理现象的变化具有非平稳性,全局参数不能很好地反映局部现象。空间异构性表明空间数据的变化量是位置的函数,可以通过局部的空间自相关量来度量空间异构性^[25]。

传统上,研究空间依赖性的主要技术是空间统计学。空间统计学扩展了经典的统计学方法,考虑了空间自相关性。空间自回归模型(Spatial Autoregressive Model, SAR)就是一种对线性回归模型进行扩展以处理空间数据的方法,通过引入描述邻域之间相互影响的邻接矩阵的概念, SAR 利用回归分析的方法可以对依赖变量进行较好的拟合。但是,由于空间统计学方法往往要通过复杂的数字解法或者基于马尔可夫链蒙特卡罗采样的方法得到结果,计算费用偏高,这就使得它不能在空间数据挖掘领域内得到广泛的应用^[11]。

从数据库技术的角度来看,因为空间依赖性的处理涉及到对邻近对象的访问,因此高效地处理邻近关系(包括拓扑关系、度量关系和方向关系)就会加快挖掘算法的执行效率。Ester 等扩展了空间数据库系统的数据结构和操作,引入邻近图和邻近路径的概念以及相应的一些操作来高效地处理隐含的邻近关系^[26]。

2.4 高维数据的挖掘算法

随着现代信息获取技术的飞速发展,空间数据的属性增加极为迅速。在遥感领域,由于感知器技术的飞速发展,波段的数目由几个增加到几十甚至上百个,如何从几十甚至几百维空间中提取信息、发现知识是对空间数据挖掘研究的一个巨大挑战。

对高维数据的挖掘处理,关键是要实现高效的查询处理。目前,处理高维数据查询的索引方法可以分为两类:一类是映射技术,即把一个高维问题先映射到等价的一维问题,然后利用现有的一维索引结构(如 B⁺ 树)进行查询。这类映射有 Z 排序, Hilbert 曲线, Gray 编码和 Pyramid 树。另一类就是建立高维索引,即直接用高维索引来解决高维问题。这类索引有 TV 树, SS 树, SR 树和 X 树。

映射技术的主要问题在于,当维数无限增加时,由高维到一维的映射会掩盖数据原有的信息,使得原先在高维空间中存在的特征和差异在一维空间内变得混杂而难以区别。因而映射技术的关键就在于寻找合适的映射方法,将高维空间中数据的信息尽可能忠实地映射到一维空间。

高维索引方法因为考虑了高维数据的特点,所以更适合处理高维查询。但是,由于高维索引结构复杂并且必须实现动态更新,因而在现有数据库系统中实现高维索引是一件非常复杂的工作。Berchtold 等指出,采用目前的对象-关系数据库系统或者纯关系数据库系统存储多维属性,唯一可行的办法就是将多维属性映射为关系模式^[27]。这是因为,虽然目前关系数据库系统得到广泛应用,但它不能有效管理数据挖掘研究的高维数据;对象-关系数据库系统提供了抽象数据类型以实现在数据库中管理高维属性,但是实现高维索引结构需要访问数据库系统的块管理器,而这在很多商业数据库系统中是不允许的。比如 ORACLE 通用服务器不提供任何数据库的面向块的接口,抽象数据类型只能通过 SQL 语句来访问关系。因此,采用目前的对象-关系数据库系统对于高维索引的建立也不是很有帮助。

2.5 空间数据挖掘系统和数据库、数据仓库系统的集成

数据库和信息产业界研究和开发历经数十年,数据库系统和数据仓库系统已经成为主流信息系统。海量数据和信息已经存储集成在这些系统中。此外,全面的信息处理和分析基础已经或将持续不断地、系统地围绕数据库系统和数据仓库构造。这包括多个异种数据库的访问、继承、统一和转换, ODBC/OLE DB 连接, Web 访问和服务工具, 报告和 OLAP 分析工具。在这种情况下,数据挖掘系统设计的一个重要问题就是应当实现和数据库系统或数据仓库系统的集成,以各种组件的形式无缝地集成到一个信息处理环境中。

基于不同的结构设计,可以采用以下耦合模式将数据挖掘系统与数据库和数据仓库系统进行集成^[1]:

(1)不耦合(no coupling):即数据挖掘系统不利用数据库、数据仓库系统的任何功能。它可能由特定的源(如文件系统)提取数据,使用某些数据挖掘算法处理数据,然后再将挖掘结果存放到另一个文件中。

(2)松散耦合(loose coupling):即数据挖掘系统将使用数据库、数据仓库系统的某些工具,从这些系统管理的数据存储中提取数据,进行数据挖掘,然后将挖掘的结果或者存放到文件中,或者存放到数据库或数据仓库的指定位置。

(3)半紧密耦合(semitight coupling):意味除了将数据挖掘系统连接到一个数据库、数据仓库系统之外,一些基本数据挖掘原语可以在数据库、数据仓库系统中实现,一些频繁使用的中间结果也可以预计算,并存放在数据库和数据仓库系统中。

(4)紧密耦合(tight coupling):意味数据挖掘系统平滑地集成到数据库、数据仓库系统中,数据挖掘子系统被视为信息系统的一个部分。

不耦合的系统在使用大的数据集或者存储在数据库中的数据时会有困难。在松散耦合的系统中,使用 SQL 语句进行查询,数据首先被数据库或数据仓库返回到缓冲区或内存,然后利用数据挖掘功能进行分析。由于这种耦合是基于内存的,挖掘本身并不使用数据库、数据仓库系统提供的数据结构和查询优化方法,对大数据集,松散耦合的系统很难获得可伸缩性和良好的性能。半紧密耦合的系统只对少数几个数据挖掘原语提供了有效实现。最理想的是实现数据挖掘系统和数据库、数据仓库系统的紧密耦合,通过把数据挖掘查询优化成循环的挖掘处理和检索过程,将二者结合起来。

从以上分析可知,为了提高空间数据挖掘系统的性能和处理海量数据的能力,就必须要实现空间数据挖掘应用和数据库系统的紧耦合。针对采用 SQL 语句进行挖掘查询的缺点, Agrawal 等提出的解决方法是把部分有关查询数据运算的程序放在数据库系统中去执行,而不是将数据从数据库运行的地址空间拷贝到应用程序运行的地址空间后再进行运算^[28]。商业数据库产品中的存储过程提供了这一思想实现的一种可能,如在 Oracle 中可以创建和存储用 PL/SQL 写成的过程,并将之命名为对象。另外,应用程序也可以给服务器发送匿名的 PL/SQL 块,并由服务器编辑执行该块。

2.6 空间数据挖掘结果的可视化

数据挖掘结果的可视化是指将数据挖掘后得到的知识和结果用可视化的形式表示出来,这些形式包括表、图、曲线、决策树、关联规则以及其它可视化输出。可视化技术的质量和灵活性严重地影响了数据挖掘系统的使用、解释能力和吸引力。由于空间数据具有复杂的空间特性,因此如何更有效地表示

挖掘结果就成为空间数据挖掘需要解决的问题之一。

地图是表现空间数据特征的最有效手段,地理信息系统近二十年来的发展已在自动生成地图方面有了很大进展。借助地理信息系统强大的图形功能,空间数据挖掘系统可以通过更直观、更易于理解的方式向用户表达数据中所包含的信息。MacEachren 等就提出集成空间数据挖掘和地理信息系统的观点,同时给出了将所发现的知识在地理空间中可视化的数据库方法^[29]。另外,空间数据挖掘还需要考虑如何实现背景知识(概念分层)和可视化技术的结合以及如何合理有效地在地图上表示挖掘结果等问题。

2.7 新的有意义模式的发现

根据所挖掘的知识类型,目前传统数据挖掘方法可分为:特征化和比较、关联规则发现、分类和预测以及聚类分析。由于空间数据具有不同于其它数据类型的特征,因此针对空间数据的特点,提出新的数据挖掘方法也是空间数据挖掘需要解决的问题之一。

目前在很多应用中,越来越多的空间数据和时间相关,因此发现有关时空方面的知识逐渐得到人们的关注^[30,31]。另外,检测空间数据中的孤立点(即误差)有可能得到意想不到的知识,这将在诸如天气预报及利用信用卡进行欺诈活动等领域的空间数据分析中发挥作用。随着空间数据越来越深入地渗透到人类生活的各个领域,空间数据挖掘研究需要针对不同的应用背景需求,提出新的数据挖掘方法以发现新的有趣的模式。

3. 空间数据挖掘和地理信息系统的关系

空间数据挖掘和地理信息系统(GIS)之间既有区别,又有联系。

一方面,专职处理空间数据的 GIS 在近十几年来虽得到了广泛的应用,并在空间数据的存储、查询以及显示等方面有了较快的发展,但其空间分析多以图形操作为主(如缓冲区操作、空间叠加、邻近分析和空间连接等),在空间信息的深入提取和知识发现等方面的功能仍相对薄弱,而空间数据挖掘技术的研究正是为解决这一问题而兴起的。空间数据挖掘和 GIS 在对空间数据的处理过程、结果类型以及表示这三个方面都有很大不同^[32]。表2给出了它们的比较。

表2 空间数据挖掘和地理信息系统的比较

	空间数据挖掘	地理信息系统
处理过程	系统产生假设并自动进行分析	用户产生假设并指导查询过程
处理结果	对数据进行抽象后得到的有用的模式和知识	直接的数据查询结果
处理过程及结果的表示	在抽象空间中可视化	在地理空间中可视化

另一方面,空间数据挖掘技术仅侧重于研究在空间数据中发现感兴趣模式的算法,它必须借助 GIS 强大的数据库和图形功能来加快前进的步伐,以便更有效地处理数据和表示挖掘结果。

从以上的分析可以看出,空间数据挖掘和地理信息系统可以认为是描述和分析空间数据的两种互补的技术。二者的发展相互交融,相互促进,在今后的研究中也必然会相互结合,成为智能空间决策支持系统的重要组成部分,为智能化空间信息处理提供新的解决办法。

结论 空间数据挖掘是一项诱人而充满挑战性的前沿技术,空间数据挖掘技术及其相关的空间决策支持系统的发展将会给空间信息处理领域带来一个新的浪潮。研究空间数据的处理方法,特别是从数据库技术角度探讨空间数据挖掘对空间数据的处理方法具有非常重要的理论和现实意义。

空间数据挖掘的研究发展至今,已经取得了一定的成果。希望本文提出的有关空间数据挖掘研究的关键问题,能为今后研究工作的开展指明方向。

参考文献

- Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000
- Lu W, Han J, Ooi B C. Discovery of general knowledge in large spatial databases. In: Proc. Far East Workshop on Geographic Information Systems, Singapore, June 1993. 275~289
- Koperski K, Han J. Discovery of spatial association rules in geographic information databases. In: Proc. 4th Int. Symp. Large Spatial Databases, SSD'95, Portland, ME, Aug. 1995. 47- 66
- Koperski K, Han J, Stefanovic N. An Efficient Two-Step Method for Classification of Spatial Data. 1998
- Ng R, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB'94, Santiago, Chile, 1994. 144~155
- Koperski K, Adhikary J, Han J. Spatial Data Mining: Progress and Challenges. In: Proc. 1996 SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery, DMKD'96, Montreal, Canada, June 1996. 1~10
- Koperski K, Han J, Adhikary J. Mining knowledge in geographical data. Communication of ACM, 1999
- 石云, 孙玉方, 左春. 空间数据采掘的研究与发展. 计算机研究与发展, 1999, 36(11): 1301~1309
- 李德仁, 王树良, 史文中, 王新洲. 论空间数据挖掘和知识发现. 武汉大学学报, 2001, 26(6): 491~499
- 裴韬, 周成虎, 等. 空间数据知识发现研究进展评述. 中国图象图形学报, 2001, 6A(9): 854~860
- Chawla S, Shekhar S, et al. Modeling spatial dependencies for mining geospatial data: An introduction. In Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 2001
- Zhou X, Truffet D, Han J. Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining. In: Proc. 6th Int. Symp. Large Spatial Databases, SSD'99, Hong Kong
- Guttman A. R-tree: A dynamic index structure for spatial searching. In: Proc. 1984 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'84, Boston, MA, June 1984. 47~57
- Beckmann N, Kriegel H-P, et al. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. In: Proc. 1990 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'90, Atlantic City, NJ, 1990
- Agrawal R, Shafer J. Parallel mining of association rules: Design, implementation and experience. Research Report RJ 10004, IBM Almaden Research Center, San Jose, California, 1996
- Cheng D W, Xiao Y. Effect of data distribution in parallel mining of association. Data Mining and Knowledge Discovery, 1999, 3(3): 291~314
- Shafer J, Agrawal R, Mehta M. Fast serial and parallel classification of very large data bases. In: Proc. 22th Int. Conf. Very Large

- Data Bases, VLDB'96, Bombay, India, Sept. 1996
- 18 Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases. In: Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'98, Seattle, WA, June 1998. 73~84
 - 19 Ester M, Kriegel H-P, Xu X. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In: Proc. 4th Int. Symp. Large Spatial Databases, SSD'95, Portland, ME, Aug. 1995. 67~82
 - 20 Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining. In: Proc. 1996 Int. Conf. Extending Database Technology, EDBT'96, Avignon, France, Mar. 1996
 - 21 Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'96, Montreal, Canada, June 1996. 103~114
 - 22 Wang K, Zhou S, Han J. Pushing support constraints into association mining. In: Proc. 26th Int. Conf. Very Large Data Bases, VLDB'00, Cairo, Egypt, Sept. 2000
 - 23 Pei J, Han J. Can We Push More Constraints into Frequent Pattern Mining? In: Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining, KDD'00, Boston, MA, Aug. 2000
 - 24 Lakshmanan L V S, Ng R, et al. Optimization of constrained frequent set queries with 2-variable constraints. In: Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'99, Philadelphia, PA, June 1999
 - 25 Bittenfield B, Gahegan M, et al. Geospatial Data Mining and Knowledge Discovery. UCGIS White Paper on Emergent Research Themes, 2001
 - 26 Ester M, Kriegel H-P, Sander J. Spatial data mining: A database approach. In: Proc. 5th Int. Symp. Large Spatial Databases, SSD'97, Berlin, Germany, July 1997. 47~66
 - 27 Berchtold S, Bohm C, et al. Implementation of Multidimensional Index Structures for Knowledge Discovery in Relational Databases. In: Proc. 1999 Int. Conf. on Data Warehousing and Knowledge Discovery, DaWaK'99, 1999
 - 28 Agrawal R, Shim K. Developing tightly-coupled data mining applications on a relational database system. In: Proc. 2th Int. Conf. Knowledge Discovery and Data Mining, KDD'96, Portland, OR, Aug. 1996
 - 29 MacEachren A M, Wachowicz M, et al. Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. International Journal of Geographical Information Science, 1999, 13 (5): 311~334
 - 30 Miller H J, Han J. Discovering geographic knowledge in data-rich environments: [Report of NCGIA Varenus Workshop on Discovering geographic knowledge in data-rich environments]. Kirkland, Washington, March 1999
 - 31 Abraham T, Roddick J F. Research issues in spatio-temporal knowledge discovery. In: Proc. SIGMOD'97 Workshop on Data Mining, Arizona, USA, 1997
 - 32 May M. Spatial Knowledge Discovery: The SPIN! System: [GMD Technical Report]. 2000

(上接第42页)

种类之间的误识情况,例如将鳞癌细胞错识为腺癌细胞或小细胞癌细胞的情况,所以,在表3中,总错误率要高于假阴性率与假阳性率之和。

从表3可以看出,使用本文所提出的基于特征提取神经网络集成方法后,对肺癌识别的准确率比单纯使用单级神经网络时有显著提高,其平均准确率在87.5%左右,假阳性率也得到了较为有效的抑制。而且,由于各级神经网络的输入针对的是某一类图像特征,其收敛速度也得到了提高,减少了训练代价。实验结果分析表明,对于肺癌图像分类识别,基于图像特征提取的神经网络集成方法是行之有效的,提高了识别准确率,达到了预期的效果。

结论 LCDS系统在Windows 2000环境下,用C++编程实现。该系统可以辅助肺癌细胞病理专家提高对肺癌细胞的识别效率,特别是在缺乏细胞病理专家的医院,实用性很好。LCDS系统已在南京肿瘤医院、中国人民解放军八一医院等4家医院应用,其效果很好。目前,我们正从事利用神经网络集成方法抽取用于肺癌分类识别的有效规则,既可以大大增加神经网络集成分类方法的可理解性,又有助于进一步提高识别准确率,提高系统的实际应用价值。本文的后续工作将致力于从神经网络集成中抽取肺癌分类识别规则的研究。

参考文献

- 1 Cox G S, Hoare F J, de Jager G. Experiments in lung cancer nodule detection using texture analysis and neural network classifiers. In: Third South African Workshop on Pattern Recognition, Nov. 1992. 136~142
- 2 Li X, Bhide S, Kabuka M R. Labeling of MRI Brain Images Using

- Boolean Neural Network. IEEE Transactions on Medical Imaging, 1996, 15: 628~638
- 3 Osareh A, Mirmehdi M, Thomas B, Markham R. Automatic Recognition of Exudative Maculopathy using Fuzzy C-Means Clustering and Neural Networks. In: Proc. Medical Image Understanding and Analysis Conf. BMVA Press, July 2001. 49~52
- 4 Shyu C, Brodley C E, Kak A, et al. Assert: a physician-in-the-loop content-based image retrieval system for hrct image databases. Computer Vision and Image Understanding, 1999, 74: 111~132
- 5 Zurada J M. Introduction to Artificial Neural Systems. West Publishing Company, 1992
- 6 Hansen L K, Salamon P. Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993~1001
- 7 Freeman. Computer Processing of Line-drawing Image. Computing Surveys, 1974, 6(1): 57~97
- 8 Miyahara M, Yasuhida Y. Mathematical Transform of (R,G,B) Color Data to Munsell (H,V,C) Color Data. Visual Communications and Image Processing'88, SPIE, 1988, 1001: 650~657
- 9 Hornik K M, Stinchcombe M, White H. Multilayer Feedforward Networks Are Universal Approximators. Neural Networks, 1989, 2(2): 359~366
- 10 Gutta S, Wechsler H. Face Recognition Using Hybrid Classifier Systems. In: IEEE Intl. Conf. on Neural Networks, NY: IEEE, 1996. 1017~1022
- 11 Shimshoni Y, Intrator N. Classification of Seismic Signals by Integrating Ensembles of Neural Networks. IEEE Transactions on Signal Processing, 1998, 46(5): 1194~1201
- 12 Schapire R. The Strength of Weak Learnability. Machine Learning, 1990, 5: 197~227
- 13 Breiman L. Bagging Predictors. Machine Learning, 1996, 24(2): 123~140