

离群数据挖掘综述^{*}

黄洪宇, 林甲祥, 陈崇成, 樊明辉

(福州大学 福建省空间信息工程研究中心 数据挖掘与信息共享教育部重点实验室, 福建 福州 350002)

摘 要: 通过对当前有代表性的离群数据挖掘算法的分析和比较, 总结了各算法的特性及优缺点, 为使用者选择、学习、改进算法提供了依据。此外, 针对高维数据和空间数据中离群检测的特殊性, 在现有算法的基础上, 分析了高维数据和空间数据离群检测需要注意的一些问题, 以便于研究者提出新的有效的算法。

关键词: 数据挖掘; 离群检测; 异常; 高维离群

中图法分类号: TP391 文献标识码: A 文章编号: 1001-3695(2006)08-0008-06

Review of Outlier Detection

HUANG Hong-yu, LIN Jia-xiang, CHEN Chong-cheng, FAN Ming-hui

(Key Laboratory of Data Mining & Information Sharing of Ministry of Education, Spatial Information Research Center of Fujian, Fuzhou University, Fuzhou Fujian 350002, China)

Abstract: This paper compared and analyzed major outlier detection algorithms. Their features were summarized to help users choose, study and improve algorithm for outlier detection. Attention was paid to high-dimensional data and spatial data because of their unique data structures as better and efficient algorithm is needed to deal with these types of data.

Key words: Data Mining; Outliers Detection; Exception; High-Dimension Outliers

1 引言

自从 20 世纪 90 年代中期数据挖掘引起人们的广泛兴趣以来, 它便得到了迅猛的发展。通常, 数据挖掘被划分为四种类型^[2], 即相关依赖关系的发现、类别的判定、类别的描述、离群或异常(Outlier)数据挖掘。前三类是针对数据集中的大部分数据记录均服从的数据模式, 而离群检测的目的在于找出隐含在海量数据中相对稀疏而孤立的异常数据模式, 这是离群检测与关联规则等传统的面向数据主体的数据挖掘的主要区别。早期, 对数据集进行预处理时, 通常把离群点当作噪声, 或修正离群点的值以减少其对正常数据的影响。尽管离群检测是为了发现数据集中极少数的一些数据, 然而离群数据挖掘常常比其他类型的挖掘来得更有价值, 因为一万个正常的记录很可能只覆盖一条规则, 而十个离群很可能就意味着十条不同的规则。实际生活中, 离群检测有着很广泛的应用, 如网络入侵检测、信用卡恶意透支、贷款证明的审核等。

离群挖掘通常可以看作三个子问题: 什么样的数据是异常, 即离群点的定义; 有效挖掘离群的方法; 离群点的意义, 即离群挖掘结果的合理解释。到目前为止, 离群点还没有一个被普遍采纳的定义, Hawkins^[3]对离群定义在一定意义上揭示了离群点的本质: “离群点与其他点如此不同, 以至于让人怀疑它们是由另外一个不同的机制产生的。”现有的离群点的定义大多是在 Hawkins 定义的基础上给出的一个定量化描

述。

统计学上, 离群数据挖掘与聚类分析一定程度上是相似的, 因为聚类的目的在于寻找性质相同或相近的记录, 并归为一个类, 根据离群的意义, 那些与所有类别性质都不一样的记录则为离群点。因此, 早期的离群检测多见于统计领域, 一些典型的具有离群检测功能的聚类算法有 CLARANS, DBSCAN, OPTICS 等^[4~10]。然而, 离群检测与聚类分析有着本质的区别, 因为聚类的目的主要在于寻找类别, 离群点只是它们的一个附属物, 因此, 由聚类算法挖掘得到的离群点通常是不准确的。

近年来, 研究人员提出了大量的离群检测算法^[1, 2, 13, 19~22], 大致可以把它们归纳为以下几类: 基于统计的方法、基于密度的方法、基于深度的方法、基于距离的方法和基于偏离的方法^[52]。下面我们对各检测算法进行综述。

2 离群检测

2.1 基于统计的离群检测算法

基于统计的方法^[11]是出现得最早的离群点检测方法, 它基于对小概率事件的判别来实现对数据样本异常的鉴别。其主要思想是假定数据集服从某种分布或概率模型, 通过不一致检验把那些严重偏离分布曲线的记录视为离群点。这方面比较有代表性的有 1967 年 Mikey, Dunn & Clark 提出的基于“均数漂移”^[38]模型的单点诊断量^[39], 1970 年 Gentleman & Wilk 提出的群组诊断量^[40], 1972 年 Tietjen & Moore 提出的单样本 k 个离群点的统计量 E_k ^[41], 1985 年 Marasinghe 提出的改进的 E_k 统计量 F_k ^[41], 1989 年 Rosner 提出的单样本多个离群检测算法 ESD(Generalized Extreme Studentized Deviate) 方法^[42],

1991 年 Paul & Fung 改进了 ESD 方法参数 k 选择的主观性, 提出了回归分析的 GESR(Generalized Extreme Studentized Deviate Resi-dual) 方法^[42]。近年来, 多样本的离群检测方法也得到了一定的发展, 总的思路是先尽量得到一个不含离群点的“干净集”^[37], 然后在此基础上对剩余的其他数据点进行逐步离群检测。

基于统计的方法检测出来的离群点很可能被不同的分布模型检测出来, 可以说产生这些离群点的机制可能不唯一, 解释离群点的意义时经常发生多义性, 这是基于统计方法的一个缺陷; 其次, 基于统计的方法在很大程度上依赖于待挖掘的数据集是否满足某种概率分布模型^[29], 模型的参数、离群点的数目等对基于统计的方法都有非常重要的意义, 而确定这些参数通常都比较困难。为克服这一问题, 一些人提出对数据集进行分布拟合, 但分布拟合存在两个问题^[46]: 给出的分布可能不适合任一标准分布; 即使存在一个标准分布, 分布拟合的过程耗时太长。此外, 基于统计的离群检测算法大多只适合于挖掘单变量的数值型数据, 目前几乎没有多元的不一致检验, 对于大多数的应用来说, 例如图像和地理数据, 数据集的维数却可能是高维的。实际生活中, 以上缺陷都大大限制了基于统计的方法的应用, 使得它主要局限于科研计算, 算法的可移植性较差。

2.2 基于距离的离群检测算法

基于距离的离群点最早是由 Knorr 和 Ng^[2,23] 提出的, 他们把记录看作高维空间中的点, 离群点被定义为数据集中与大多数点之间的距离都大于某个阈值的点, 通常被描述为 $DB(pct, d_{min})$, 数据集 T 中一个记录 O 称为离群点, 当且仅当数据集 T 中至少有 pct 部分的数据与 O 的距离大于 d_{min} 。换一种角度考虑, 记 $M = N \times (1 - pct)$, 离群检测即判断与点 O 距离小于 d_{min} 的点是否多于 M 。若是, 则 O 不是离群点, 否则 O 是离群点。基于距离的离群点定义包含并拓展了基于统计的思想, 即使数据集不满足任何特定分布模型, 它仍能有效地发现离群点, 特别是当空间维数比较高时, 算法的效率比基于密度的方法要高得多^[7]。算法具体实现时, 首先给出记录间距离的度量^[20], 常用的是绝对距离(曼哈顿距离)、欧氏距离和马氏距离。在给出了距离的度量并对数据进行一定的预处理以后, 任意给定参数 pct 和 d_{min} 就可以根据离群的定义来检测离群。Rastogi & Ramaswamy 在上面基于距离的离群点定义的基础上, 提出改进的基于距离的 k -最近邻 (k -NN) 离群检测算法^[24,30]。用 $D^k(P)$ 表示点 P 的第 k 个最近邻点的距离, 首先计算出数据集 T 中所有点的 k -最近邻距离, 然后按值大小降序排列, 算法把排在最前面的 n_0 个点标记为 n_0 个离群点。此算法的一个主要缺陷是要计算所有点的 $D^k(P)$, 每计算一个点的 $D^k(P)$ 就要扫描一次数据集, 对于大数据集, 其 I/O 次数常常使得算法的计算效率非常低。针对 k -最近邻算法的这个主要缺陷, 不少人对算法进行了不同程度的改进, 如基于索引的算法引进索引的思想来提高算法的效率^[1,47]; 循环嵌套算法主要从减少操作的 I/O 次数方面来改善算法的效率^[2,47]; 基于单元的算法通过结合点的局部密度的方法来提高异常检测的效率^[1,48]。

基于距离的离群检测方法中, 算法需要事先确定参数 pct 和 d_{min} , 对于不同的数据集这往往是一件比较困难的事情, 特

别是 d_{min} , 不同聚类密度的数据集 d_{min} 会有很大的差异, 而这一般没有规律可循, 因此, 对于给定的不同 d_{min} , 异常检测结果通常具有很大的不稳定性^[65]。另一方面, 基于距离的方法理论上能处理任意维任意类型的数据, 当属性数据为区间标度等非数值属性时, 记录之间的距离不能直接确定, 通常需要把属性转换为数值型^[25,26], 再按定义计算记录之间的距离。当空间的维数大于三维时, 由于空间的稀疏性, 距离不再具有常规意义, 因此很难为异常给出合理的解释。针对这个问题, 一些人通过将高维空间映射转换到子空间的办法来解决数据稀疏的问题, 此方法在聚类算法中用得比较多^[10,18], Agarwal R. 等人^[10] 曾试着用这种投影变换的方法来挖掘离群。总的来说, 基于距离的离群检测方法具有比较直观的意义, 算法比较容易理解, 因此在实际中应用得比较多。

基于距离的方法与基于统计的方法相比, 不需要用户拥有任何领域知识, 与序列异常相比, 在概念上更加直观。更重要的是, 距离异常接近 Hawkins 的异常本质定义。然而, 三种类型的基于距离的离群检测算法中, 基于索引的算法和循环——嵌套算法需要 $O(k \times n^2)$ 的时间开销, 因此在大数据集中还有待于改进; 而基于单元的算法, 虽然与 n 具有线性的时间关系, 但是它与 k 成指数关系, 这限制了它在高维空间中的应用, 此外, 基于单元的算法还需要事先确定参数 pct , d_{min} 以及单元的大小, 这使得算法的可行性比较差; 高维空间中, 基于索引的方法由于需要事先建立数据集的索引, 建立与维护索引也要花大量的时间。因此三种方法对于高维空间中的大数据集, 算法的效率都不高^[47]。

2.3 基于密度的离群检测算法

基于密度的离群检测算法^[12,66] 一般都建立在距离的基础上, 某种意义上可以说基于密度的方法是基于距离的方法中的一种, 但基于密度的异常观点比基于距离的异常观点更贴近 Hawkins 的异常定义, 因此能够检测出基于距离的异常算法所不能识别的一类异常数据——局部异常^[66]。基于密度的方法主要思想是将记录之间的距离和某一给定范围内记录数这两个参数结合起来, 从而得到“密度”的概念, 然后根据密度判定记录是否为离群点。

Breunig 等人提出的基于局部离群因子的异常检测算法 LOF^[4,13,41] 是基于密度方法的一个典型例子。它首先产生所有点的 MinPts-邻域及 MinPts-距离, 并计算到其中每个点的距离; 对低维数据, 利用网格进行 k -NN 查询, 计算时间为 $O(n)$; 对中维或中高维数据, 采用如 X-树等索引结构, 使得进行 k -NN 查询的时间为 $O(\log n)$, 整个计算时间为 $O(n \log n)$; 对特高维数据, 索引结构不再有效, 时间复杂度提高到 $O(n^2)$ ^[66,67]。然后计算每个点的局部异常因子, 最后根据局部异常因子来挖掘离群。LOF 算法中, 离群点被定义为相对于全局的局部离群点, 这与传统离群的定义不同, 离群不再是一个二值属性(要么是离群点, 要么是正常点), 它摒弃了以前所有的异常定义中非此即彼的绝对异常观念, 更加符合现实生活中的应用。LOF 算法中充分体现了“局部”的概念, 每个点都给出了一个离群程度, 离群程度最强的那几个点被标记为离群点。此外, Aggarwal^[18] 也提出了一个结合子空间投影变换的基于密度的高维离群检测算法。

2.4 基于深度的离群检测算法

基于深度的离群点检测算法^[21, 22]的主要思想是先把每个记录标记为 k 维空间里的一个点, 然后根据深度的定义(常用 Peeling Depth Contours^[22] 定义) 给每个点赋予一个深度值; 再根据深度值按层组织数据集, 深度值较小的记录是离群点的可能性比深度值较大的记录大得多, 因此算法只需要在深度值较小的层上进行离群检测, 不需要在深度值大的记录层进行离群检测。基于深度的方法比较有代表性的有 Struyf 和 Rousseeuw 提出的 DEEPLOC^[19] 算法。虽然, 理论上基于深度的识别算法可以处理高维数据, 然而实际计算时, k 维数据的多层操作中, 若数据集记录数为 N , 则操作的时间复杂度为 $\Omega(N^{\lceil k/2 \rceil})$ 。因此, 当维数 $k \leq 3$ 时处理大数据集时还有可能是高效的, 而当 $k \geq 4$ 时, 算法的效率就非常低。也就是说, 已有的基于深度的离群点检测算法无法挖掘高维数据, 只有当 $k \leq 3$ 时计算效率才是可接受的。

2.5 基于偏移的离群检测算法

基于偏移的离群检测算法(Deviation-based Outlier Detection)^[14 ~16, 28]通过对测试数据集主要特征的检验来发现离群点。目前, 基于偏移的检测算法大多都停留在理论研究上, 实际应用比较少。以下三种是比较有代表性的: Arning^[14] 采用了系列化技术的方法来挖掘离群, 由于算法对异常存在的假设太过理想化, 因此并没有得到普遍的认同, 对于现实复杂数据, 其效果不太好, 经常遗漏了不少的异常数据; Sarawagi^[15] 应用 OLAP 数据立方体引进了发现驱动的基于偏移的异常检测算法; Jagadish^[16] 给出了一个高效的挖掘时间序列中异常的基于偏移的检测算法。虽然, 基于偏移的离群检测算法理论上可以挖掘各种类型的数据, 但是由于要事先知道数据的主要特征, 而现实世界中的数据集一方面由于数据量比较大, 另一方面由于属性比较多, 因此这方面的特征往往不容易发现, 当确定记录之间的相异度函数时, 如果选择不合适, 则得到的离群挖掘结果很可能不尽人意, 所以本方法在实际问题中应用得比较少。表 1 列出了上述各种类型的离群检测方法的一些主要特性。

表 1 各种离群检测算法比较

检测方法	典型算法	主要优点	适合数据集类型	主要缺陷	高维空间	时间效率分析
基于统计的方法	$E_k \cdot F_k$ 统计量 ESD 算法	统计意义上的离群, 视严重偏离分布模型的点为离群点, 对挖掘结果的意义比较容易解释	单变量的服从特定概率模型的数据集	只适合于单变量离群的检测, 需要知道较多的关于数据集的先验知识	不适合	根据数据集的大小以及采用的概率分布模型而定, 一般超过 $O(n \times \log n)$
基于深度的方法	DEEPLOC, FindOut 算法	如仅适于处理四维以下的数据集	数据属性适合于分层的低维数据集	当数据维数大于 3 时, 算法的可行性非常差	不适合	$\Omega(N \log N)$, 当属性维数增加时, 算法的效率变得很低, 趋于 ∞
基于距离的方法	一般距离方法 Index-based 算法 NL 算法 Cell-based 算法	从数据索引查询及 I/O 操作次数角度提高算法效率	属性维比较少且参数 p_{cst} 和 d_{min} 比较容易确定的数值属性数据集	数据的稀疏性和离群意义难以解释, I/O 代价通常比较高	算法的时空效率不高	Index-based 和 NL 需要 $(k \times N^2)$, Cell-based 需要 $O(c^{k+n})$
	基于密度的方法 LOF 算法	提出离群强度的概念, 量化了异常程度	数据集聚类特性比较明显, 求局部密度时的 I/O 代价比较低			低维空间: $O(n^3)$; 中高维空间: $O(n \times \log n)$; 高维空间: $O(n^2)$
基于偏移的方法	知识、发现驱动检测方法	对数据集的类型没有特殊要求	知道数据集的一些特征	实用性不强	不适合	与数据集大小呈线性关系

备注: n 是数据集记录数, k 是属性维数, c 是依赖于单元数目的常数。

多基于数据挖掘的思想, 它们不需要对数据集的分布特征附加任何假设, 而只需从数据集本身出发对数据进行检测, 发现其中包含的异常模式, 这使得算法具有较大的可行性和可移植性。然而, 大部分检测方法多存在一个缺陷, 即把算法重点放在离群点的识别上, 而常常忽略了用户最关心的离群点的现实意义。此外, 实际生活中的数据往往具有较大的噪声, 异常模式只存在于低维子空间中, 而高维空间中的异常都比较难确定, 传统的离群检测算法在数据特性完全不同的高维数据中性能急剧下降。例如 Knorr & Ng 等人^[27] 提出的基于距离的 FindAllOutsD; Joshi 等人^[35] 提出的基于机器学习方法的二阶段规则推导算法 NP-rule; Yu 等人^[60] 提出的基于小波变换的 Find-Out 和 Breuning 等人^[13] 提出的基于密度的 LOF 等没有充分考虑离群点检测问题的特点, 在内存有限的条件下, 缺乏数据的预处理, 经常导致频繁的数据交换和难以容忍的空间复杂度, 且算法一般均有 $O(N \times \log N)$ 以上的时间复杂度。在处理海量高维数据中, 算法经常无法获得令人满意的响应速度, 有些算法甚至失效。

3 高维离群检测算法

到目前为止提出的大多数离群检测算法对高维数据的异常检测效果都不是很理想。高维空间离群检测与其他数据集的离群检测差别甚大的原因主要有两个: 高维空间数据的稀疏性经常导致所有记录在传统距离的意义上都很有可能是离群点^[10], 因此, 高维空间中很难找到一个通用的离群产生机制; 高维空间中的离群检测算法的计算复杂度通常都比较高, 算法的执行效率极低^[44]。高维空间中的异常检测通常采用 Aggarwal & Yu 提出的高维数据异常检测方法^[45], 基本思想是把高维空间投影到子空间, 采用演化计算来寻找最优子空间, 降低空间的维数, 然后利用传统的检测算法来发现异常。具体实现时, 高维空间中异常模式的检测通常是非常困难的, 一个最简单的办法是考虑数据维的所有可能组合, 对所有子空间进行异常模式的搜索。这种方法虽然能找出全局最优的离群点, 然而对于具有 n 个属性的数据集, 可能的维数组合数为 2^n , 对于高维数据, 这个数字将是一个天文数字, 因此算法的执行效率非常差。针对这一缺陷, 近年来 Aggarwal & Yu 运用遗传算法^[43] 优化了这个问题的求解, 并取得良好效果。其他的典型算法还有: SLOT(Subspace Local Outlier Test)^[13], 它通过对低维子空间中离散系数上下限的估计, 预先去除大量不可能成为异常的数据, 从而大大提高了算法的效率。Bially^[34, 60, 61] 提出了利用分形中的空间填充曲线 Hilbert 曲线, 将多维空间中的对象映射到低维子空间, 根据变换保持空间聚集特性不变来解决挖掘高维空间中的离群点。

Aggarwal^[18] 提出了高维空间中值得思考的几点建议: 处理好高维空间中数据的稀疏问题; 合理地解释离群产生的原因; 必须确定适当的度量, 以解释 K 维子空间中离群的物理意义; 超高维数据离群检测算法的时空效率问题; 当决定某个点是否是离群点时, 必须能为局部数据行为提供价值。因此, 要提高高维空间中离群检测的效率, 我们必须进一步改善获得的投影子空间。在传统的直接映射到某个低维子空间, 或按属性加权平均求得一个综合倾斜度的投影之前, 先对数据进

总的来说, 以上各离群检测算法, 除基于统计的方法外, 大

行旋转变换, 求出数据的一个基准生成器, 再把数据映射到一个改进的基准坐标面上, 按各属性加权进行离群检测, 这样得到的离群点通常都能比较真实地反映离群, 也能比较好地解释离群点的意义。然而, 基于投影的高维空间离群检测仅能对数值属性数据进行分析处理, 对名称标度等非数值属性, 如何选择一个合适的投影子空间还有待于进一步研究。总之, 要构造好的高维空间离群检测算法, 我们必须充分利用高维数据自身的特性, 处理好数据集的主体聚类性与稀疏性之间的矛盾, 以提高异常检测算法的效率。

4 空间数据的离群检测

空间数据自身的独特性使得其与关系数据存在着巨大的差别, 除了包含属性数据外, 还包含了空间关系数据。因此, 空间数据集中的异常检测比关系数据中的离群异常检测来得复杂, 因为空间对象经常受到邻近对象的影响, 因此空间异常检测只有充分考虑了对象的近邻点的影响才能获得有用的知识^[63, 64]。空间例外是指那些非空间属性值和邻域中其他空间对象的非空间属性明显不同的空间对象^[49, 63], 两个空间对象的差异程度通常用相异度来衡量。由于空间数据自身的特殊性, 空间离群点一般是局部不稳定的, 这种局部意义上的离群点在全局中不一定仍为离群点。空间例外挖掘是空间数据挖掘领域的一个重要研究内容, 例外知识的发现能为我们提供丰富的信息, 在地理信息系统、遥感图像数据勘测、公众安全与卫生、交通控制、基于地理位置的服务^[55]等各种领域有着广泛的应用。

空间数据集的离群检测方法大致可以分成三类^[57, 63]: 基于集合的例外、基于多维属性空间的例外和基于图的例外。基于集合的例外是在不考虑空间关系的数据集合中, 属性值和其他数据对象属性值不一致的数据对象。严格上讲, 基于多维属性空间的例外和基于图的例外才是真正意义上的空间例外, 即属性值和其空间邻域中的其他对象明显不同的空间对象, 两者的不同在于空间邻域的定义不同, 前者的空间邻域定义是基于距离, 而后的定义是基于图的连接性。基于多维属性空间的例外检测算法则对空间例外与其他数据之间的区别进行检验, 代表性的算法有 Scatterplot^[50, 52] 和 Moran Scatterplot^[53]; 基于图的例外检测算法在空间数据的可视化基础上, 将空间离群点突出地显示出来以达到检测的效果, 典型的算法有 Variogram Clouds 和 Pocket Plots^[51, 54] 以及 Shekhar 等人在文献[56, 57] 中引进的在图形数据库中基于点与邻域点的关系的空间离群点检测方法。与非空间数据的离群检测类似, 这些算法有个缺点, 即可能忽略某些空间离群点, 同时, 算法可能错判一些空间离群点。针对这个缺陷, Lu 等人在文献[49] 中通过多次循环的方式对算法进行了一定的改进, 主要思想是每次循环都只识别一个空间离群点, 并修正离群点的值以使它不影响后续的离群点检测。

5 结论与展望

异常检测在很多领域中有着非常重要的应用, 不久的将来它将在更多邻域中发挥更重要的作用。现有的大多离群检测算法在高维空间、时间序列以及地理数据中的效率还比较低,

算法还或多或少地存在各种不同的缺陷, 如何发挥数据集自身的主体聚类性和完备性, 最大化相异数据之间的差异, 最小化相似数据之间的差异, 提高算法的时空效率和可移植性是离群检测面临的具有挑战性的工作。最新文献表明, 地学数据的离群检测算法、动态环境下异常的增量式挖掘算法、长时间序列离群检测算法以及基于人工智能的离群检测算法将是未来一段时间内离群数据挖掘领域的一个主要研究方向。

参考文献:

[1] an Jiawei, Kamber M. Data Mining: Concepts and Techniques[M] . Academic Press, 2001.

[2] Knorr E M, Ng R T. Algorithms for Mining Distance-based Outliers in Large Datasets[C] . New York: Proc. of Int. Conf. Very Large Databases(VLDB '98) , 1998. 392-403.

[3] Hawkins D. Identification of Outliers[M] . London: Chapman and Hall, 1980.

[4] Ng R T, Han J. Efficient Clustering Methods for Spatial Data Mining [A] . Bocca J B, Jarke M, Zaniolo C. Proc. of the 20th International Conference on Very Large Data Bases[C] . Santiago: Morgan Kaufmann, 1994. 144-155.

[5] Ester M, Kriefel H P, Sander J, *et al.* A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[A] . Simoudis E, Han J, Fayyad U M. Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining[C] . Portland: AAAI Press, 1996. 226-231.

[6] Zhang T, Ramakrishnan R, Linvy M. Birch: An Efficient Data Clustering Method for Very Large Databases[A] . Jagadish H V, Mumick I S. Proc. of the ACM SIGMOD International Conference on Management of Data[C] . Montreal: ACM Press, 1996. 103-114.

[7] Wang W, Yang J, Muntz T R. Sting: A Statistical Information Grid Approach to Spatial Data Mining[A] . Jarke M, Carey M J, Dittrich K R, *et al.* Proc. of Bases[C] . Athens: Morgan Kaufmann, 1997. 186-195.

[8] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multi-resolution Clustering Approach for Very Large Spatial Databases[A] . Gupta A, Shmueli O, Widom J. Proc. of the 24th International Conference on Very Large Databases[C] . New York: Morgan Kaufmann, 1998. 428-439.

[9] Hinneburg A, Keim D A. An Efficient Approach to Clustering in Large Multimedia Databases with Noise[A] . Agrawal R, Stolorz P E, Piatetsky-Shapiro G. Proc. of the 4th International Conference on Knowledge Discovery and Data Mining[C] . New York: AAAI Press, 1998. 58-65.

[10] Agrawal R, Gehrke J, Gunopulos D, *et al.* Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications [A] . Haas L M, Tiwary A. Proc. of the ACM SIGMOD International Conference on Management of Data[C] . Seattle: ACM Press, 1998. 94-105.

[11] Barnett V, Lewis T. Outliers in Statistical Data[M] . New York: John Wiley & Sons, 1994.

[12] Breunig M M , Kriegel H P, Ng R T, *et al.* Optics of: Identifying Density-based Local Outliers[A] . Zytkow J M, Rauch. Proc. of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, Lecture Notes in Computer Science 1704 [C] . Prague: Springer, 1999. 262-270.

[13] Breuning M M, Kriegel H P, Ng R T, *et al.* LOF: Identifying Density-based Local Outliers[A] . Chen W, Naughton J F, Bernstein. Proc. of the ACM SIGMOD International Conference on Management

of Data[C] . Dallas: ACM Press, 2000.93-104.

[14] Arning A, Agrawal R, Raghavan P. A Linear Method for Deviation Detection in Large Databases[C] . Proc. of 1996 Int. Conf. Data Mining and Knowledge(Special Issue on High Performance Data Mining) , 2000.

[15] Sarawagi S, Agrawal R, Megiddo N. Discovery-driven Exploration of OLAP Data Cubes[C] . Valencia: Proc. of Int. Conf. Extending Database Technonlogy(EDBT '98) , 1998.168-182.

[16] Jagadish H V, Koudas N, Muthukrishnan S. Mining Deviants in a Time Series Databases[C] . Edinburgh: Proc. of Int. Conf. Very Large Databases(VLDB '99) , 1999.102-113.

[17] Aggarwal C C, Yu P. Outlier Detection for High Dimensional Data[A] . Aref W G. Proc. of the ACM SIGMOD International Conference on Management of Data[C] . Santa Barbara: ACM Press, 2001.37- 47.

[18] Aggarwal C C, Yu P. Finding Generalized Projected Clusters in High Dimensional Spaces[A] . Chen W, Naughton J F, Bernstein P A. Proc. of the ACM SIGMOD International Conference on Management of Data[C] . Dallas: ACM Press, 2000.70-81.

[19] Struyf A, Rousseeuw P J. High-dimensional Computation of the Deepest Location[J] . Computational Statistics and Data Analysis, 2000, 34: 415-426.

[20] 袁志发,周静芋. 多元统计分析[M] . 北京: 科学出版社, 2002. 242-243.

[21] Tukey J W. Exploratory Data Analysis[M] . MA: Addison-Wesley and Sons, Inc. , 1994.

[22] Preparata F, Shamos M. Computational Geometry: An Introduction [M] . Springer-Verlag, 1988.

[23] Knorr E, Ng R. Finding Intensional Knowledge of Distance-based Outliers[C] . Scotland: Proc. of the 25th VLDB Conference Edinburgh, 1999.211-222.

[24] Knorr E, Ng R. A Unified Approach for Mining Outliers: Properties and Computation[C] . Newport Beach: Proc. of Int. Conf. Knowledge Discovery and Data Mining(KDD '97) , 1997.219-222.

[25] S D Bay, M Schwabacher. Mining Distance-based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule[C] . Washington DC: SIGKDD '03, 2003.

[26] S Ramaswamy, R Rastogi, K Shim. Efficient Algorithms for Mining Outliers from Large Data Sets[C] . Proceedings of the ACM SIGMOD Conference, 2000.427-438.

[27] E M Knorr, R T Ng, V Tucakov. Distance-based Outliers: Algorithms and Applications[J] . VLDB Journal: Very Large Databases, 2000,237-253.

[28] 郑建国,焦李成. 偏差检测挖掘方法研究[J] . 计算机工程, 2001, 27(8) : 33-35.

[29] 史东辉,张春阳,蔡庆生. 离群数据的挖掘方法研究[J] . 小型微型计算机系统, 2001,22(10) : 234-236.

[30] F Angiulli, C Pizzuti. Fast Outlier Detection in High Dimensional Spaces[C] . Proceedings of the 6th European Conference on the Principles of Data Mining and Knowledge Discovery, 2002.15-16.

[31] 史东辉,蔡庆生,倪志伟,等. 基于规则的分类数据离群挖掘方法 [J] . 计算机研究与发展,2000,37(9) : 109-110.

[32] 姜灵敏. 基于相似系数和检测孤立点的聚类算法[J] . 计算机工程,2003,29(11) : 183-185.

[33] 陆声链,林士敏. 基于距离的孤立点检测及其应用[J] . 计算机与数字工程,2003,12(22) : 31-33.

[34] Bially T. Space-filling Curves: Their Generation and Their Application to Bandwidth Reduction[J] . IEEE Trans. on Info. Theory, 1969,15(6) : 658-664.

[35] 魏黎,宫学庆,钱卫宁,等. 高维空间中的离群点发现[J] . 软件学报, 2002,13(2) : 280-290.

[36] Yang Feng-Zhao, Zhu Yang-Yong. IncLOF: An Incremental Algorithm for Mining Local Outliers in Dynamic Environment[J] . Computer Research and Development, 2004,41(3) : 477-484.

[37] Karjalainen L P, Somani M, C Porter. Regression and Solute Drag Models for the Activation Energy of Static Recrystallisation in Hot-worked Steels[J] . Materials Science Forum, 2003,426-432(2) : 1181-1188.

[38] Maneesh K Singh, Narendra Ahuja. Mean-shift Segmentation with Wavelet-based Bandwidth Selection[C] . Proceedings of the 6th IEEE Workshop on Applications of Computer Vision(WACV '02) , 2002.

[39] Beckman R J, Cook RD. Outliers[J] . Technometrics, 1983,25: 119-163.

[40] Gentleman J F, Wilk MB. Detecting Outliers in a Two-way Table: Statistical Behavior of Residuals[J] . Technometrics, 1975,17: 1-14.

[41] Marasinghe MG. A Multi-stage Procedure for Detecting Several Outliers in Linear Regression[J] . Technometrics, 1985,27: 395-399.

[42] Paul S T, Fung K Y. A Generalized Extreme Studentized Residual Multiple-outlier-detection Procedure in Linear Regression[J] . Technometrics, 1991,33: 339-348.

[43] Li Ai Bing. Evaluation on Slope Stability Using Visual Quick Nerval Network[J] . Kuangye Yanjiu Yu Kaifa, 2001,21(5) : 8-11.

[44] Berchtold C, Bohm, H P Kriegel. Improving the Query Performance of High-dimensional Index Structures by Bulk Load Operations[C] . Proc. of EDBT, 1998.

[45] N Katayama, S Satoh. The SR-tree: An Index Structure for High-dimensional Nearest Neighbor Queries[C] . Proc. of ACM SIGMOD Int. Conf. on Management of Data, 1997.369-380.

[46] 范大昭,雷蓉,张永生. 从地理数据库中探测奇异值[J] . 测绘科学, 2004,29(5) : 12.

[47] S D Bay, M Schwabacher. Mining Distance-based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule[C] . Washington, DC: SIGKDD '03, 2003.

[48] F Anguylli, C Pizzuti. Fast Outlier Detection in High Dimensional Spaces[C] . Proceedings of the 6th European Conference on the Principles of Data Mining and Knowledge Discovery, 2002.15-16.

[49] Lu C T, Chen D, Kou Y. Algorithms for Spatial Outlier Detection [C] . The 3rd IEEE International Conference on Data Mining(IC-DM '03) , 2003.597-600.

[50] R Haining. Spatial Data Analysis in the Social and Environmental Sciences[M] . Cambridge University Press, 1993.

[51] J Haslett, R Brandley, P Craig, *et al.* Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies[J] . The American Statistician, 1991,45: 234-242.

[52] A Luc. Exploratory Spatial Data Analysis and Geographic Information Systems[M] . M Painho. New Tools for Spatial Analysis, 1994.45-54.

[53] A Luc. Local Indicators of Spatial Association: LISA[J] . Geographical Analysis, 1995,27(2) : 93-115.

[54] Y Panatier, Variowin. Software for Spatial Data Analysis in 2D[M] . New York: Springer-Verlag, 1996.

[55] S Shekhar, S Chawla. A Tour of Spatial Databases[M] . Prentice Hall, 2002.

[56] S Shekhar, C T Lu, P Zhang. Detecting Graph-based Spatial Outlier: Algorithms and Applications(A Summary of Results) [C] . Proc. of the 7th ACM-SIGKDD Int '1 Conference on KDD Mining, 2001.

[57] S Shekhar, C T Lu, P Zhang. Detecting Graph-based Spatial Outlier [J] . Intelligent Data Analysis: An International Journal, 2002,6(5) : 451-468.

[58] 王丽珍, 何婧. 基于图的空间例外检测算法研究[J] . 云南大学学报, 2003, 25(5) : 398-400.

[59] Martin Ester, Hans-Peter Kriegel, Jörg Sander. Algorithms and Applications for Spatial Data Mining[C] . Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis, 2001.

[60] Zheng Binxiang, Du Xiuhua, Xi Yugeng. Outliers Mining in Time Series Data Sets[J] . Journal of Systems Engineering and Electronics, 2002, 13(1) : 93-97.

[61] Zhang Tong, Zhang Xibin, Zhang Shiyong. Testing for Outliers in Time Series Using Wavelets[J] . Journal of Systems Science and Complexity, 2003, 16(4) .

[62] Seng Chuan TAY, Wynne HSU, Kim Hwa LIM. Spatial Data Mining: Clustering of Hot Spots and Pattern Recognition[M] . Toulouse France: International Geoscience & Remote Sensing Symposium, 2003.

[63] Vandana Janeja, Vijayalakshmi Atluri, Nabil R Adam. Outlaw: Using GeoSpatial Associations for Outlier Detection and Visual Analysis of Cargo Routes[C] . Proceedings of the 2nd National Conference on Digital Government, 2002.

[64] Stephen D Bay, Mark Schwabacher. Near Linear Time Detection of Distance-based Outliers and Applications to Security[C] . SIAM Data

Mining Conference, Workshop on Data Mining for Counter Terrorism and Security, 2003.

[65] Shengyi Jiang, Qinghua Li, Kenli Li, *et al.* GLOF: A New Approach for Mining Local Outlier[C] . International Conference on Machine Learning and Cybernetics, 2003. 157-162.

[66] Spiros Papadimitriou, Hiroyuki Kitagawa, *et al.* LOCI: Fast Outlier Detection Using the Local Correlation Integral[C] . The 19th International Conference on Data Engineering, 2003. 315.

[67] Yang Fengzhao, Zhu Yangyong, Shi Baile. IncLOF: An Incremental Algorithm for Mining Local Outliers in Dynamic Environment[J] . Journal of Computer Research and Development, 2004, 41(3) : 477-484.

作者简介:

黄洪宇(1971-), 男, 福建泰宁人, 硕士, 主要研究方向为资源与环境信息工程、空间数据仓库与空间数据挖掘等; 林甲祥(1982-), 男, 福建安溪人, 硕士研究生, 主要研究方向为空间数据挖掘; 陈崇成(1968-), 男, 福建闽清人, 副教授, 博士, 主要研究方向为环境与资源信息工程、空间决策支持系统、自然资源与环境遥感、地理信息网络共享技术、虚拟地理环境、空间数据仓库与可视化挖掘等; 樊明辉(1974-), 男, 湖北黄冈人, 博士研究生, 主要研究方向为环境与资源信息工程、空间决策支持系统、网络地理信息系统、空间数据仓库及数据挖掘、网格计算及应用等。

(上接第 3 页) 日语、法语、荷兰语、西班牙语), 上海交通大学的多语言天气预报发布系统。 面向实际应用的开发, 如国内有北京交通大学和北京颐和园的导游系统、中国科技大学的机器人足球现场解说系统^[7] 以及人机接口^[8] 等; 国外的有英国 inburgh大学 Michaelo’ Donnell 所设计的在线文件剪接系统。 在生成过程中对所要表达的信息进行语义和句法方面的聚合亦是目前研究重点之一。当前语言生成的研究方向主要是在语言表示形式、信息内容规划以及语言生成模型等^[9] 方面。自然语言生成的研究将继续在诸多语言学科、计算机领域和其他学科的通力协作下获得新的成果。

3 结束语

本文对自然语言生成的体系结构进行了介绍, 重点对内容规划、微观规划和表层生成三个基本模块的设计以及内部核心功能部件进行了详细描述, 受篇幅限制, 不当之处请予以指正。

参考文献:

[1] 郭忠伟, 徐延勇, 周献中. 基于 Schema 和 RST 的自然语言生成混合规划方法[J] . 计算机工程, 2003, 29(6) : 113-115.

[2] 黄友能. 可移植的自然语言生成系统中知识库的设计[J] . 北京交通大学学报, 2004, 28(5) : 20-23. .

[3] 许龙飞, 杨晓昀, 唐世渭. 基于受限汉语的数据库自然语言接口技术研究[J] . 软件学报, 2002, 13(4) : 537.

[4] Ehud Reiter, Roberty Dale. Building Natural Language Generation Systems[M] . Cambridge University Press, 2000.

[5] Daniel Jurafsky, James H Martin. Speech and Language Processing [M] . Prentice Hall, 2005.

[6] 张冬荣, 李锦乾, 等. 汉语自然语言生成的句子结构优化[J] . 计算机工程, 1998, 24(7) : 16.

[7] 许华东, 刘贵金, 陈小平, 等. 一个机器人足球现场解说系统[J] . 计算机工程, 2002, 28(3) : 179-181.

[8] 沈军, 顾冠群. CIMS 领域中面向问题的自然语言人机接口研究[J] . 微机发展, 2002, 12(6) : 43-47.

[9] 陈华, 韩近强, 邓海清, 等. 面向特定领域人机对话模型研究与实现[J] . 计算机工程与应用, 2004, 40(26) : 82-85, 100.

作者简介:

张建华(1973-), 男, 硕士研究生, 主要研究方向为自然语言生成、自动文摘; 陈家骏(1963-), 男, 教授, 博导, 研究方向为自然语言处理和软件工程。