空间例外挖掘算法研究*)

Research on Mining Spatial Outliers

邹力鹏　王丽珍

(云南大学信息学院计算机科学与工程系　昆明650091)

Abstract Spatial outliers are observations which appear to be inconsistent with the most datum in the spatial database. Identification of spatial outliers can lead to the discovery of unexpected, and interesting knowledge. Detecting spatial outliers is useful in many applications of GIS and spatial databases. In this paper, we propose a new formal definition of spatial outliers which overcome the shortcomings of existing definition. We also provide corresponding algorithm to detect spatial outliers and experiment it based on synthetical spatial database to prove that the algorithm is correct and effective.

Keywords Spatial outliers, Spatial data, Local, Spatial predicate

1. 引言

例外挖掘是数据挖掘中的一个重要研究方向，例外数据的挖掘往往可使人们发现一些真实、但又出乎意料的知识。随着数据处理工具、先进数据库技术的不断成熟和数据应用的普及，传统数据挖掘技术已经开始向解决复杂类型数据的方向发展。其中，空间数据挖掘就是一个重要的分支。由于空间例外挖掘对人类社会、科技和经济的发展具有重大的意义，因此针对空间数据的例外挖掘研究是一项重要的、有意义的工作。

本文中，提出了一种新的空间例外模式定义，并以非空间属性集合或空间谓词来评价解释空间例外的形成。同时，提出相应的空间例外模式挖掘算法并以合成数据验证了算法的正确性和有效性。本文第2节介绍相关工作；第3节讨论了空间例外的分类，并给出了相应定义；第4节介绍例外算法；第5节是实验部分，最后总结全文，并给出了本文的后继工作。

2. 相关工作

例外挖掘工作通常由两部分构成：例外数据的发掘和例外内涵知识的获取。

例外数据挖掘主要有如下方法[1~6]：基于分布的方法、基于深度的方法、基于距离的方法和基于密度的方法，但这些方法都侧重例外的发现方面，实际上，例外内涵知识的获取与例外数据的发掘同等重要。例外内涵知识不仅有助于用户评价例外数据的有效性和可靠性，而且有助于加深用户对数据的理解。“为什么这个数据是例外?”才是用户关心的问题，也是寻找例外的最终目的。文[7]进行了这方面的研究，认为一个点的某些属性与其他点有很大差异，就足以使它成为例外，因此关键在于发现是哪些属性导致了它的偏离成为例外。

遗憾的是，上述方法均没有考虑到地理、位置等属性的特殊性，没有区分对象的空间属性和非空间属性。文[8]对空间例外进行了初步研究，给出了一个空间例外 *S_outlier* 例外的一般性定义，并提供了有效的挖掘算法。但该研究存在一定的局限性，例如，所考虑的数据仓库中事实表的维数有限，对空间谓词的使用也仅仅限于传统意义上的“邻接”关系，没有涉及评价机制的研究。

3. 空间例外的分类及其定义

3.1 引例

以澜沧江(湄公河)空间数据仓库中的关于气温的一个小部分数据为例。该空间数据由两部分组成：观测站气温值，即非空间数据；观测站的地理位置，即空间属性(通常为指向图的指针)。对象的空间属性和非空间属性通过表中的空间属性域相联系。在该空间数据中，我们并不涉及到时间属性，所有数据都是在同一时间采集。以1月20日的气温为例：

从不同的角度出发，我们可以得到不同的结果。从满足位于云南省这个空间谓词来说，记录10是一个空间例外，即所有云南省的站点中，中甸的气温异常。从气温值来看，记录9是例外，即所有气温在1度左右的站点中，成都是唯一不在云南省境内的。换言

*)基金项目：云南省自然科学基金资助项目(项目编号：2002F0013)。邹力鹏　硕士生，主要研究领域是数据挖掘方向。王丽珍　教授，研究生导师，主要研究领域是数据库与数据挖掘技术。

之，根据不同的局部定义，进行比较的数据集合不同，记录10是基于空间属性对非空间属性的比较所得出的例外结果；而记录9则是基于某些非空间属性，对空间对象的空间属性进行比较所得的结果。两类结果均是我们所需要的知识类型，均涉及以前讨论较少的空间属性，但两种空间例外的定义显然有所不同。

记录号	观测站	气温	空间属性
1	昆明	16	…
2	玉溪	15	…
3	曲靖	13	…
4	个旧	18	…
5	安宁	16	…
6	呈贡	17	…
7	大理	16	…
8	晋宁	15	…
9	成都	14	…
10	中甸	8	…

3.2 空间例外的分类

通过引例，可以看出空间例外将对象的空间属性进行了充分的考虑，在此基础上，我们将空间例外分成两类：基于空间谓词的空间例外和基于非空间属性的空间例外。

第一类空间例外指的是，非空间属性值和其空间邻域中的其他空间对象明显不同的空间对象。例如，在一个新城市的老城区中的一幢新房就是基于非空间属性房屋年龄的一个空间例外。

另一类空间例外则指，基于非空间属性的空间数据集合的子集中，空间属性和(该子集中)其他对象明显不同的空间对象。例如，在所有的老房子中，房 A 是空间例外，因为其他房子都靠近城市中心，而房 A 则远离城市中心。

3.3 空间例外的定义

讨论前提：

· S 是$\{S_1,S_2,\cdots,S_n\}$是空间对象的集合；

· 非空间属性集合 $P\{P_1,P_2,\cdots,P_t\}$；

· 谓词集合 $Pre\{Pre_1,Pre_2,\cdots,Pre_m\}$；

· 任务相关对象集合 $T\{T_1,T_2,\cdots,T_l\}$。

3.3.1 *基于空间属性的空间例外的定义* 这一类空间例外是一个局部不稳定性，是一个非空间属性值和其邻域极度不同的空间对象，即使对象的非空间属性值在全局中不是那么地不同。我们可以用空间关系距离或拓扑关系来定义空间邻域，对在空间对象的比较则是基于非空间属性的。讨论之前需要如下概念：

空间邻域：S 是$\{S_1,S_2,\cdots,S_n\}$是空间对象的集合，我们把与对象 s 具有相同空间属性的空间对象集合称为 s 的空间邻域，记为 N(s)。对于同一参考对象，谓词不同，任务相关对象不同，所得的空间邻域也是不同的。

类：类 C_t 是数据集 N(s)中相似度不小于 ε 的数据对象 s 的集合。

类支持度：设 C_t 是数据集 N(s)中一个类，C_t 在 N(s)中的类支持度 δ_c 是指 C_t 的数据对象在 N(s)中所占的比重。

类的置信度：设 C_t 是数据集 N(s)中一个类，C_t 在 N(s)中的类置信度 α_c 是类 C_t 的对其最邻近类 $C_{nearest}$的偏离程度，用相异度的比值表示。类的置信度越高，相关类间的耦合性就越低，它偏离最邻近类的程度就越高。

定义3.1(基于空间谓词的空间例外) S 是$\{S_1,S_2,\cdots,S_n\}$是空间对象的集合，对象 s∈S，s 是一个基于空间谓词的空间例外(Spatial Predicate_Based Spatial Outlier)，记为 SPB_SpatialOutlier，当且仅当在 s 的空间邻域 N(s)中，s 所属的类 C 是具有低支持度 δ_c 和高置信度 α_c 的类。即：

$$SPB\_SpatialOutliers=\{s\mid s\in S,C\subseteq N(s),s\in C\wedge\alpha_c\geqslant\Lambda_\alpha\wedge\delta_c\leqslant\Lambda_\delta\}$$

其中，$\Lambda_\delta,\Lambda_\alpha$ 分别为支持度和置信度阈值。

显然，s 所属的类 C 中的所有空间对象都和 s 一样是基于空间谓词的空间例外。

3.3.2 *基于非空间属性的空间例外的定义* 仍然沿用3.2节中的例子，要在所有的老房子中找出与众不同的房子，即基于房屋年龄这个非空间属性上的聚类进行进一步的挖掘工作。而此时，空间对象的比较是基于空间属性的，即是在空间谓词上进行比较。在讨论之前我们需要如下概念：

基于非空间属性集合 P 的类：类 C_t 是数据集 S 中相似度不小于 ε 的数据对象 s 的集合。其中相似度的度量是基于属性集合 P 的。

基于非空间属性集合 P 的类支持度：设 C_t 是数据集 S 中基于非空间属性集合 P 的一个类，C_t 在 S 中的类支持度 δ_c 是指 C_t 的数据对象在 S 中所占的比重。

定义3.2(类中空间谓词关于任务相关对象的置信度) 设 C_t 是数据集 S 中基于非空间属性集合 P 的一个类，C_t 中满足谓词 $Pre_i(x,T_j)$空间对象的百分比，是类 C_t 的空间谓词 Pre_i 关于任务相关对象 T_j 的置信度，记为 $\sigma(Pre_i,T_j)_c$。即：

$$\sigma(Pre_i,T_j)_c=\frac{|\{s\mid s\in C_t,Pre_i(s,T_j)=True\}|}{|C_t|}$$

由于空间谓词和任务相关对象都是多个，因此一个类的空间谓词关于任务相关对象的置信度实际上一个二维数组，其行是对于同一个空间谓词关于不同任务相关对象的置信度，而二维数组的列则是不同空间谓词关于同一个任务相关对象的置信度。

定义3.3(基于非空间属性的空间例外) S 是$\{S_1,S_2,\cdots,S_n\}$是空间对象的集合，对象 s∈S，当

$Pre_i(s,T_j)=True$，且 s 所属的类 C 具有高的类支持度 δ_c 和低的空间谓词 Pre_i 关于任务相关对象 T_j 的置信度 $\sigma(Pre_i,T_j)_c$，或者当 $Pre_i(s,T_j)=False$，且 s 所属的类 C 具有高的类支持度 δ_c 和高的空间谓词 Pre_i 关于任务相关对象 T_j 的置信度 $\sigma(Pre_i,T_j)_c$ 时，s 是一个基非空间属性的空间例外(Nonspatial Predicate_Based Spatial Outlier)，记为 NSPB_SpatialOutlier。即：

$$NSPB\_SpatialOutliers=\{s\mid s\in S,s\in C\wedge\sigma_c\geqslant\Lambda_\sigma\wedge\sigma(Pre_i,T_j)_c\geqslant\Lambda_{\sigma 1}\wedge Pre_i(s,T_j)=False\vee s\in S,s\in C\wedge\sigma(Pre_i,T_j)_c\leqslant\Lambda_{\sigma 2}\wedge Pre_i(s,T_j)=True\}$$

从定义可以知道，s 的空间属性和其他对象不同时，就成为了空间例外。如果 s 所属的类 C 的类支持度小到一定限度时，s 就是一个传统意义上的例外，即现在多数例外研究工作者所能发现的例外，我们不再着眼于这种例外，而是把工作中心放在他们所不能发现的具有相同非空间属性对象中的例外。显然，对基于非空间属性的空间例外的发现是以前例外研究者从未涉及过的。

4. 算法

例外的识别过程与内涵知识的挖掘过程的关系是密不可分的，因此，对空间例外内涵知识的发现和空间例外发现这两个过程是彼此交融，不可分离的。受文[7]影响，我们使用属性集合来解释为什么一个空间对象会成为一个空间例外，针对不同类型的空间例外，其解释也是不同的。对于基于空间谓词的空间例外，用非空间属性集合来评价解释空间例外的形成，而对于基于非空间属性的空间例外则用空间谓词来评价解释空间例外的形成。

4.1 基于空间谓词空间例外提取算法

输入：(1)置信度阈值 Λ_σ 和支持度阈值 Λ_δ；
(2)建立空间邻域的空间谓词 pre；
输出：例外集合 SPB_Outlier；
算法描述：

```
1)Neighhor=Find_neighbor(Si,pre,S);
2)设置 SC 为由 p 个属性组成所有可能属性集合的集合；
3)For all x in SC
  {
4)   Clustering on Neighhor ;//在属性集合 x 上进行聚类，得到 Si 的空间邻域的聚类表示
5)   for each i<=|C
6)        计算 Ci 的支持度
7)   for each i<=|C| and support [i]<Λδ
     {
8)        寻找 Ci 的最邻近类；
9)        计算 Ci 的置信度
10)       if (σc>Λσ)
11)         SPB_Outlier=∪Ci，并标记 x 是导致 Ci 中对象成为例外的属性集合；
     }
}
```

对于一个发现的空间例外 s，导致其成为空间例外的属性集合可能有多个 $X_1,X_2,\cdots,X_d$，我们只关心能解释 s 是 SPB_SpatialOutlier 例外的最小属性集，因此，若有 X_1 属于 X_2，则不再认为 X_2 是导致 s 成为例外的属性集合。因此，在该算法完成之后，还需要对导致对象成为空间例外的属性集合进行裁减。例如：以四个属性 A，B，C，D 为例，SC={A，B，C，D，AB，AC，AD，BC，BD，CD，ABC，ABD，ACD，BCD，ABCD}，该算法执行后得到{A，AB，BCD，ABCD}是导致 s 是例外的属性集合，由于 A 属于 AB，A 属于 ABCD，因此 AB 和 ABCD 均应被删去，最后得到导致 s 是例外的属性集合为{A，BCD}。

4.2 基于非空间属性的空间例外提取算法

根据第3节中基于非空间属性的空间例外的定义，我们将空间例外的提取工作分成如下几部分：基于一个非空间属性或某个非空间属性的集合的聚类；对于每个类计算其针对不同空间谓词和不同任务相关对象的空间谓词置信度；最后才是空间例外的提取。

算法：基于非空间属性的空间例外挖掘算法

输入：(1)用于提取局部数据的非空间属性集合；
(2)类支持度和置信度阈值 Λ_δ 和 Λ_σ；
输出：空间例外，及导致其成为例外的空间谓词集合；
算法描述：

步骤1：根据用户输入的属性集合，得到对象的聚类表示 $\{C_1,C_2,\cdots,C_k\}$；

步骤2：对每个类 $C_i(1\leqslant i\leqslant k)$ 计算支持度；

步骤3：对每个类 $C_i(1\leqslant i\leqslant k)$ 计算其对不同空间谓词 $Pre_a(1\leqslant a\leqslant m)$ 和不同任务相关对象 $T_b(1\leqslant b\leqslant l)$ 的空间谓词置信度 Conference [a][b]；

步骤4：对每个类 $C_i(1\leqslant i\leqslant k)$ 寻找其中的空间例外，执行步骤4.1，4.2，4.3如下：

(1)取谓词集中的第 a 个谓词 Pre_a

(2)取任务相关对象集中的第 b 个对象 T_b

(3)检查：若 Conference [a][b]$<\Lambda_\sigma$，则将 C_i 中所有使 $Pre_a(x,T_b)$ 为真的 x 标记为一个谓词空间例外，同时标记 $Pre_a(x,T_b)$ 为真是导致其成为例外的原因；否则若 Conference [a][b]$>1-\Lambda_\sigma$，则将 C_i 中所有使 $Pre_a(x,T_b)$ 为假的 x 标记为一个谓词空间例外，同时标记 $Prea(x,T_b)$ 为假是导致其成为例外的原因。

5. 实验

我们在 Windows2000平台上，用 Delphi. 0实现了本文中的算法，并用合成数据(以表的形式存储在 SQLSERVER2000中)在内存为192M 的 P-III800个人计算机上进行了测试，测试数据库中共有6个测试用例，其中参考对象的个数从50到800不等，任务相关对象数目、非空间属性数目以及空间谓词都是固定不变的，为4，4，3。

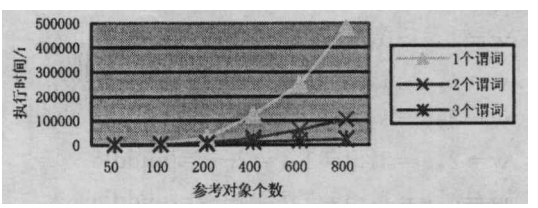

图1 算法4.1执行时间随测试数据量的变化情况

首先，测试算法的正确性和准确性。由于采用的是合成数据，预先知道测试用例中异常数据的分布，

因此，实验结果可以此作为比较标准。通过多次实验，该算法确实能发现空间例外，准确性在89%左右。

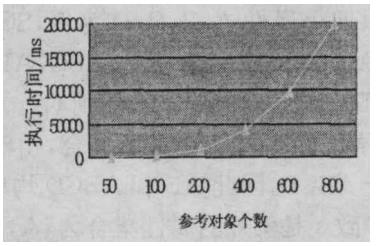

图2 算法4.2执行时间随测试数据量的变化

然后测试算法的效率。算法4.1在对不同数目的参考对象，以及不同数目的空间谓词(用来提取空间邻域)的执行情况如图1所示。算法4.2随数据量的增加其执行时间变化如图2所示，在该算法中，用以聚类的属性集合中的属性的个数对执行时间没有影响。

小结 本文针对现有空间例外研究的不足，提出了全新的空间例外分类以及相应的空间例外的定义，以属性集合或空间谓词的集合概念来解释例外的内涵，并提出了相应的有效算法。实验表明，算法能有效地寻找空间例外并对其进行合理的解释。进一步的工作包括：在真实数据上检验算法的有效性和GIS软件平台的结合。

参考文献

1 Knorr E M, Ng R T. Algorithms for Mining Distance-Based Outliers in Large Databasets. In: Proc. of the 24th VLDB Conf. New York: VLDB Endowment, 1998. 392～403
2 Arning A, Agrawal R, Raghavan P. A linear method for deviation dection in large databases. In: Proc. of the 2nd Int Conf. Knowledge Discovery &Data Mining. Portland: VLDB Endowment, 1996. 164～169
3 Ramaswamy S, Rastongi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data, vol29, ACM, 2000. 427～438
4 Markus M. breuing, Hans-Peter Kriege. LOF: Identifying Density-Based Locak Outliers. In: Proc ACM SIGMOD 2000 Int. Conf. on Management of Data, Dalles, TX, 2000
5 Anthony W J, Tung K H. Mining Top-n Locak Outliers in Large Database, KDD2001 San Francisco, California USA
6 宋擒豹，沈钧毅，等. 基于距离的例外模式挖掘算法，2002，38(7)：25～29
7 Knorr E M, Ng R T. Finding Intensional Knowledge of Distance-Based Outlier. In: Proc. of the 25th VLDB Conf. Edinburgs, 1999. 211～222
8 Shekhar S, Lu C T, Zhang P. A Unified Approach to Spatial Outliers. http://www.cs.umn.edu/Research/shashi-group

(上接第257页)

经过数据的预处理和条件属性的约简后，就可以进行规则的生成。根据数据预处理生成的数据字典，education_level 的值对应关系为：Partial High School(0)，High School Degree(1)，Partial College (2)，Bachelors Degree (3)，Graduate Degree (4)。management_role 的值对应关系为：Store Temp Staff(0)，Store Full Time Staff(1)，Middle Management (2)，Store Management (3)，Senior Management (4)。store_type 的值对应关系为：Small Grocery (0)，Mid-Size Grocery (1)，Supermarket (2)，Gourmet Supermarket (3)，Deluxe Supermarket(4)，HeadQuarters(5)。salary 的值对应关系为：1000 美元以下(low)，1000 美元至 10000 美元(middle)，10000美元以上(high)。worktime 用 w 表示，education_level 用 e 表示，management_role 用 m 表示，store_type 用 t 表示，salary 用 s 表示，所提取出来的规则有：

w=4,e=0,m=0,t=0→s=low
w=4,e=1,m=1,t=0→s=low
w=2,e=1,m=0,t=2→s=low
w=2,e=2,m=1,t=2→s=middle
w=4,e=1,m=1,t=4→s=middle
w=2,e=3,m=3,t=3→s=middle
w=4,e=2,m=1,t=2→s=middle
w=2,e=2,m=1,t=5→s=middle
w=3,e=3,m=3,t=4→s=high
w=5,e=4,m=4,t=4→s=high
w=3,e=3,m=4,t=5→s=high

上述所有规则对用户而言，不一定全部都有实际意义。我们还需要根据实际情况对所得规则进行评估，找出最有价值的规则集，也就是知识。

结论 本文的元信息采用渐增方法生成，通过对数据逐个处理，不断进行调整，最终生成反映当前数据信息的元信息，因而避免了传统粗集方法对大量数据作一次性处理而造成的内存不足，同时又具备处理动态增加数据的能力。并且，重用元信息将大大减少数据挖掘的时间，提高了数据挖掘系统的性能。从时间维上看，元信息的渐增生成将数据挖掘的任务分散到各个时间段上，从而避免传统粗集方法一次性处理所有数据对计算机资源的大量要求。通过设置检查点来定期存储元信息，可以有效地避免长时间的数据挖掘中系统崩溃而造成任务失败。当系统崩溃时，本方法可以在最近可用的元信息的基础上进行后续的挖掘任务，从而提高了系统的鲁棒性。

参考文献

1 Pawlak Z, et al. Rough Sets. Communications of the ACM, 1995, 38(11)
2 曾黄麟. 粗集理论及其应用. 重庆大学出版社
3 Bazan J G. A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In Polkowski and Skowron [169], chapter 17: 321～365
4 常犁云，吴渝，等. 一种基于 Rough Set 理论的属性约简及规则提取方法. 软件学报，1999，10(11)
5 苏健，高济. 基于元信息的粗糙集规则增式生成方法. 模式识别与人工智能，2001，14(4)