

基于空间统计分析的关联规则应用研究^{*}

An Investigation of Association Rules Based on Spatial Statistical Analysis

吕安民 李成名 林宗坚

(武汉大学遥感与信息工程学院 武汉430079)

Abstract This paper leads spatial statistical analysis to data mining. Spatial autocorrelation can be used to analyze spatial distribution feature of spatial data. G statistics can be used to analyze spatial association of spatial data. This paper also analyzes population growth rate in 1982—1990 and 1990—1998 in China. The spatial association of population growth rate in two periods is studied with spatial statistics methods as well. In 1982—1990, China is divided into four regions: lower population growth rate in north sub-region, higher population growth rate in center and west sub-region, lower population growth rate in center and east sub-region and higher population growth rate in south sub-region. In 1990—1998, China is divided into two regions: lower population growth rate in north sub-region and higher population growth rate in south sub-region. All these results are according to the factor of China. From these results we can see that spatial statistical analysis is an effective method of data mining.

Keywords Data mining, Spatial statistics, Spatial association

1 引言

在数据挖掘的研究领域,对于关联规则挖掘的研究开展得比较积极和深入。关联规则挖掘就是要找出隐藏在数据间的相互关系,它展示了数据间未知的依赖关系,根据这种关联性就可从某一数据对象的信息来推断另一数据对象的信息。

文[1~6]对关联规则的挖掘作了有意义的研究。R. Agrawal 等提出了 Apriori 算法^[2]和挖掘多层次关联规则的 Culminate, Stratify 等算法^[3], J. S. Park 等提出了 DHP 算法^[5], J. Han 等提出了面向属性归纳的关联规则挖掘算法 ML-T2L1 等^[6]。

空间统计分析主要用于空间数据的分类和综合评价,它涉及到空间和非空间数据的处理和统计计算。空间统计分析的方法很多,包括主成分分析、层次分析、系统聚类分类、判别分析、空间自相关分析和 G 统计分析。本文试图用空间自相关分析和 G 统计分析方法挖掘数据之间的关联规则。

2 空间统计分析

2.1 空间权重矩阵

n 个对象的邻近关系可以用二进制的空间权重矩阵来表示,采用邻接标准或距离标准。

根据邻接标准,空间权重矩阵的元素 $\{w_{ij}\}$ 当区域 i 和区域 j 相邻时为 1,其它情况下为 0。

$$w_{ij} = \begin{cases} 1 & \text{当区域 } j \text{ 和区域 } i \text{ 相邻时} \\ 0 & \text{其它} \end{cases}$$

根据距离标准,当区域 i 和区域 j 的距离在给定的距离 d 之内时,空间权重矩阵的元素 $\{w_{ij}\}$ 为 1,否则为 0。

$$w_{ij}(d) = \begin{cases} 1 & \text{当区域 } i \text{ 和区域 } j \text{ 的距离小于 } d \text{ 时} \\ 0 & \text{其它} \end{cases}$$

一般情况下的空间权重矩阵可以定义为:

$$w_{ij}^* = w_{ij} X_j / \sum_j w_{ij} X_j$$

同样地,

$$w_{ij}^*(d) = w_{ij}(d) X_j / \sum_j w_{ij}(d) X_j$$

其中 X_i 是面积、人口等属性值。

2.2 空间自相关指标—Moran I

表示全局空间自相关最常用的指标是 Moran I,其公式为:

$$I(d) = \frac{\sum_i \sum_{j \neq i} w_{ij} (x_i - \bar{x})(x_j - \bar{x}) / S^2}{\sum_i \sum_{j \neq i} w_{ij}} \quad (1)$$

其中 $S^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ $\bar{x} = \frac{1}{n} \sum_i x_i$
 x_i 是区域 i 的观测值, w_{ij} 是空间权重矩阵。

根据空间数据分布可以计算 Moran I 的期望值^[7,8]

对于正态分布,

$$E_n(I) = -1/(n-1) \quad (2)$$

$$VAR_n(I) = \frac{n^2 w_1 - n w_2 + 3 w_0}{w_0 (n^2 - 1)} - E_n^2(I) \quad (3)$$

对于随机分布,

$$E_R(I) = -1/(n-1) \quad (4)$$

$$VAR_R(I) = \frac{n((n^2 - 3n + 3)w_1 - n w_2 + 3w_0^2) - k_2((n^2 - n)w_1 - 2n w_2 + 6w_0^2)}{w_0^2(n-1)(n-2)(n-3)} - E_R^2(I) \quad (5)$$

其中

$$k_2 = (n \sum_i (x_i - \bar{x})^4) / (\sum_i (x_i - \bar{x})^2)^2$$

$$w_0 = \sum_i \sum_j w_{ij} \quad w_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2$$

$$w_2 = \sum_i (w_{i.} + w_{.i})^2$$

$w_{i.}$ 和 $w_{.i}$ 分别表示权重矩阵中 i 行和 i 列之和。

用下式可以检验 n 个区域是否存在空间自相关关系:

$$Z(d) = (I(d) - E(I)) / \sqrt{VAR(I)} \quad (6)$$

当 Moran I 为正时,在距离 d 的范围内的观测值有趋同的趋势;当 Moran I 为负时,在距离 d 的范围内的观测值有不同的趋势;当 Moran I 为零时,观测值属于独立随机分布。

2.3 空间关联—G 统计

Getis 和 Ord^[9]研究了用于空间关联的 G 统计,其计算公

^{*} 本文得到国家计委2000年高技术应用项目“人口地理信息系统建设技术支撑体系”的支持。吕安民 副教授,博士生,研究方向是地理信息系统和数据挖掘的理论及应用。

作进一步的阐述。

·用例模型 在系统分析阶段,由系统开发人员与用户不断地交互讨论,导出一个双方均认同的需求的文档,但通常情况下需求文档并不十分清晰,综合性也不太好。为此,建立复杂一些的用例模型,可以更好地理解用户需求,并驱动分析与设计,甚至测试。

用例模型可以用用例图来进行刻画。从需求规范到用例模型有四个步骤:1)定义出系统的确切范围;2)发现执行者(或者用户);3)寻找用例;4)定义用例间的关系。

·分析对象模型 在确定了用例模型之后,对用例进行详细的分析,开始标识对象。对象是描述问题域以及构造应用的基础。分析对象模型主要包括:构造系统的对象类、对象之间的关联、对象维护的信息(对象的属性)、对象提供的服务(对象的行为)等。可以用UML中的对象图或类图进行描述。

·动态分析模型 描述了应用的动态行为。主要包括:1)应用中的事件流;2)对象的状态改变。动态分析模型可以用UML的行为图与交互图进行描述。

我们通常用分析对象模型来刻画对象的静态结构,而用

动态分析模型来刻画应用的动态行为。这两种模型是紧密相关的,它们反映了同一应用的不同方面。所以在建立这两种模型时不一定是顺序的,这两种模型可以交替进行,同时开发,互相补充。

注意,在分析阶段,分析对象模型和动态分析模型还处于草图状态,具体的细化要到设计部分进行。

4.2 基于UML的MIS设计

在以数据库为核心的MIS设计中,数据库的设计是系统设计的一个中心环节,数据库的结构方式以及效率问题会影响整个系统的应用与扩充,因此设计阶段的其中一个重要任务是利用前面基于UML的MIS分析结果,继续细化,形成设计模型,完成数据库的设计及软件的设计工作。基于UML的MIS设计分以下步骤进行:

1)架构设计。需要定义包(子系统),包间的相关性和基本的通信机制,系统平台选择等;

2)在保证可追踪问题域的前提下,把分析阶段的分析对象模型进一步细化成与实现相关的设计对象模型;

(下转第137页)

(上接第54页)

表3 GI统计及其Z值

编号	名称	Gi(d)	E(Gi)	VAR(Gi)	Zvalue
1	黑龙江	0.053	0.069	0.000	-0.720
2	内蒙古	0.266	0.275	0.000	-0.428
3	新疆	0.116	0.103	0.000	0.734
4	吉林	0.760	0.103	0.000	-1.304
5	辽宁	0.092	0.103	0.000	-0.621
6	甘肃	0.212	0.207	0.000	0.298
7	河北	0.242	0.241	0.000	0.042
8	北京	0.073	0.069	0.000	0.281
9	山西	0.144	0.138	0.000	0.401
10	青海	0.140	0.138	0.000	0.152
11	天津	0.082	0.069	0.000	1.229
12	陕西	0.244	0.241	0.000	0.154
13	宁夏	0.103	0.103	0.000	0.002
14	西藏	0.133	0.138	0.000	-0.231
15	山东	0.141	0.138	0.000	0.236
16	河南	0.213	0.207	0.000	0.307
17	江苏	0.116	0.138	0.000	-1.217
18	安徽	0.188	0.207	0.000	-1.130
19	四川	0.282	0.276	0.000	0.166
20	湖北	0.197	0.207	0.000	-0.564
21	上海	0.046	0.069	0.000	-2.012
22	浙江	0.169	0.172	0.000	-0.098
23	江西	0.204	0.207	0.000	-0.143
24	湖南	0.209	0.207	0.000	0.148
25	云南	0.138	0.138	0.000	0.035
26	贵州	0.129	0.138	0.000	-0.630
27	福建	0.099	0.103	0.000	-0.276
28	广西	0.148	0.138	0.000	0.545
29	广东	0.192	0.172	0.000	0.710
30	海南	0.045	0.034	0.000	1.243

区域,中西部高人口增长率区域,中东部低人口增长率区域和南部高人口增长率区域。

结束语 本文将空间统计方法引入到数据挖掘领域,空间自相关分析可以用来研究数据的空间分布特征,空间关联分析可以分析空间数据的空间关联性。同时本文用这两种空间统计分析方法研究了两个时期人口增长率的空间关联关系。1982—1990年中国可分为北部人口低增长、中西部高人口增长率、中东部低人口增长率和南部高人口增长率等四个区

域;1990—1998年中国可分为北部低人口增长率和南部高人口增长率两个区域。两个时期的空间聚类虽然不完全相同,但它们有共同的特点,南部和西部的人口增长率都比较高。这些结论比较符合中国的实际情况,这就说明空间自相关和G统计分析是有效的数据挖掘方法。

参考文献

- 1 Agrawal R, Imielinski T, Iyer B, et al. Mining association rules between sets of items in large databases. In: Proc. of ACM SIGMOD Conf. on Management of Data, 1993. 207~216
- 2 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: the Proc. of VLDB—1994, Chile, Sep. 1994. 487~499
- 3 Agrawal R, Srikant R. Mining generalized association rules. In: the Proc. of VLDB—1995, Switzerland, 1995. 407~419
- 4 Fayyad U, Piatetsky-shapiro G, Smyth P. From data mining to knowledge discovery: An overview. Advances in knowledge discovery and data mining, MIT Press, 1996. 1~34
- 5 Park M, Chen M, Yu P. An effective hash based algorithm for mining association rules. In: the Proc. of ACM SIGMOD 1995. 175~186
- 6 Han J, Fu Y. Discovery of multiple-level association rules from large databases. In: Proc. of VLDB—1995, Switzerland, 1995. 420~431
- 7 Cliff & Ord. Spatial processes: Models and applications. London: Pion, 1981
- 8 Goodchild. Spatial autocorrelation (CATMOG47). Norwich, UK: Geobooks, 1986
- 9 Getis and Ord. The analysis of spatial association by the use of distance statistics. Geographical analysis, 1992, 24: 189~206
- 10 Ord & Getis. Distributional issues concerning distance statistics. Working paper, The Pennsylvania State University and San Diego State University, 1994
- 11 Ding & Fotheringham. The integration of spatial analysis and GIS. Computers, Environment an urban systems, 1992, 16(1): 3~19
- 12 Kehris. Spatial autocorrelation statistics in ARC/INFO. [Research Report NO. 16]. North west regional research laboratory, Lancaster University, 1991