第 9 卷 第 3 期 市场与人口分析 Vol. 9 No. 3
2003 年 5 月 Market and Demographic Analysis May. 2003

文章编号:1006 - 4346(2003)03 - 0008 - 05

数据挖掘在人口学中的应用

刘建华
(厦门大学 人口研究所,福建 厦门 361005)

摘要:随着收集数据能力的加强,人口数据和资料日益丰富起来,但同时又普遍感到存在着"数据丰富而信息匮乏"的问题。试图将数据挖掘的技术应用到第五次人口普查当中,通过具体的例子分析来说明在人口学领域如何运用数据挖掘技术,以此探讨如何对五普的数据进行深入分析以及如何提高其应用价值。

关键词:数据挖掘;第五次人口普查;人口学
中图分类号:C92 - 03 **文献标识码**:A

Application of Data Mining in Demography

LIU Jian-hua
(*Institute of Population Research, Xiamen University, Xiamen* 361005, *China*)

Abstract: With our ability of collecting data increasing, we can take advantage of rich demographic data. However, the problem of "rich data but poor information" is in evidence. This paper will explain how to use data mining technique to solve some problems in the fifth census of China by concrete examples. Taking data mining technique, we can analyze demographic data more deeply and improve the value of them.

Key words: data mining; census; demography

在人口学领域中,经常需要各种各样的数据(例如人口普查数据、抽样调查数据、问卷调查数据和经常性的统计登记数据等)来研究和证实人口的特征和规律。随着研究的深入和时间的推移,我们所得到的人口数据资料越来越多,与此同时,我们也发现传统的数据处理和分析手段使我们所获得的信息十分有限,这也就是人们经常提到的"数据丰富而信息匮乏"的问题;在这种情况下,我们便需要依靠新的技术和方法来分析和提取隐含在大量数据中的有用信息,这样,数据挖掘技术便开始受到人们的关注。把数据挖掘技术应用到人口领域,可以极为方便地解决在人口研究领域中遇到的数据处理和分析的问题,并进而可以为决策提供有意义的信息。

收稿日期:2002 - 06 - 03;**修订日期**:2002 - 09 - 05
作者简介:刘建华(1977 -),男,厦门大学人口研究所 2001 级硕士研究生,研究方向为人口、资源与环境系统工程。

1 数据挖掘概念及其技术

1.1 数据挖掘概念及其过程

数据挖掘其实是"知识发现"(knowledge discovery)的通俗说法,指的是通过仔细分析大量数据来揭示有意义的关系、模式和趋势的过程。数据挖掘是数据库发展过程中的一个新阶段,它采用模式识别、统计分析、信息提取、计算机和数学等技术来分析和提取隐藏在数据库或数据仓库中的某一领域或相关领域的大量数据之间的各种有用信息。

数据挖掘所揭示的有用信息主要分为两类:描述性信息和预测信息。其基本处理过程是:

1.2 数据挖掘技术

(1)特征描述(characterization)。系通过对研究对象的各种属性的分析,从数据库中挖掘出(以不同的角度或在不同的层次上的)平均/最小/最大值、总和、百分比等等,并且可以找到与研究对象关系最为密切的属性项。

(2)关联分析(association)。系指指定的数据库挖掘出满足一定条件的依赖性关系。关联规则形如"A1→A2,支持度 = s%,置信度 = c%",其中 s 和 c 是用户指定的支持度和置信度的门限值。这种关联规则采掘可以在不同的抽象概念层次上进行。例如 R1"尿布→啤酒,支持度 = 5%,置信度 50%"与 R2"婴儿用品类→饮料类,支持度 = 25%,置信度 80%"相比,R2 在更高的抽象层次上,更为宏观,因而有较大的支持度和置信度,更适合高层决策需要。

(3)特征分类(classification)。指通过对样本数据的特征和分类结果,为每一个类找到一个合理的描述或模型,然后再用这些分类的描述或模型来对未知的新的数据进行分类。

(4)回归分析(regress)。回归分析与分类相似,其差别在于分类的预测值是离散的,而回归的预测值是连续的。它的模型一般有一定的函数关系:y = f(x)。

(5)聚类分析(cluster)。聚类分析其实是无监督分类(unsupervised classification),其目的在于实事求是地(而不是按照人的主观认识)按被处理对象的特征分类,有相同特征的对象被归为一类。

(6)可视化(visual)。可视化主要是解决如何将挖掘出的知识或信息表达出来,它拓宽了传统的图表功能,使用户可以对数据的剖析结果了解的更清楚、更形象、更直观。例如,把数据库中的多维数据变成多种图形,这对揭示数据的状况、内在本质及规律性起到了很强的作用。

2 数据挖掘技术在人口学中的应用

人口普查为全面的了解人口的各种信息提供了丰富的数据来源,但是人口的研究并不能仅局限于人口领域本身,而应该为制定政策、拟制计划和日常行政管理服务,为社会经济研究服务,为工商企业服务。以前的人口普查资料仅仅是按照一些规定指标中的统计项目作了归纳与整理,并没有更好的应用到社会的各个领域中去。这一方面是因为对如何与其他领域相结合的研究工作做得不够深入,另外一个重要的原因是没有更好的处理手段和方法。下面结合第五次人口普查的指标谈谈如何应用数据挖掘技术来分析和提取对工商企业有用的人口信息。

2.1 采用关联规则技术分析妇女文化程度与生育子女数的相关性

我们都知道一般来说妇女文化程度与生育子女数之间是负相关关系,我们找到妇女文化程度与生育子女数的相关性,便能够从数理上对此加以证实和说明。

建立关联规则:education(X,E) = > birth(X,B)

其中 X 代表妇女,E 代表某一种文化程度,B 代表生育的子女数。

设定最小的支持率 Sup-Min ,以及最低的可信度 Con - Min(支持率指的是在所有的样本数据中出现的概率 ,可信度是指在某一支持度水平下关联规则成立的可信程度。这两个值一般是根据分析的结果动态取值 ,尽量达到在最小的支持度下有最大的可信程度) 。文化程度可以分得非常细 ,如文盲、半文盲、小学、初中、高中、专科、本科、研究生。但为了使分析的结果更具有说服力 ,一般在更抽象的层次上进行分类 ,如 :小学及以下、中学、大学、研究生。生育子女数分为 :0 - 1 人、2 - 3 人、4 - 6 人等 ,当然这些分类按照需要都可以更改。

我们从人口普查数据中提取所有妇女的文化程度和生育情况 ,同时根据文化程度以及生育子女数的分类 ,将各条记录的文化程度和生育情况相应归类 ,然后在提取出的数据的基础上挖掘出在给定支持度条件下的各类文化程度各生育情况的可信度及 Cij ,将求得的各种可信度填入对比表。

表 1　可信度对比表

	0 - 1 人	2 - 3 人	4 - 6 人
小学及以下	C11	C12	C13
中学	C21	C22	C23
大学	C31	C32	C33
研究生	C41	C42	C43

我们还可以将结果用图表来反映 ,例如坐标系(见图 1) 。

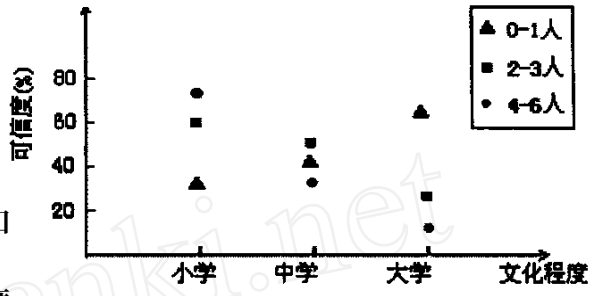


图 1　某一支持度下的可信度对比图

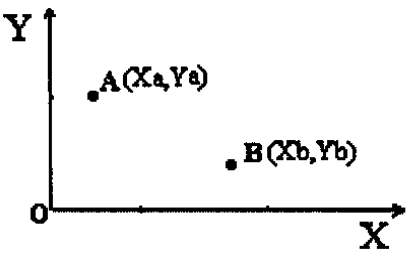
当然也可以将各种支持度下的可信度情况作出对比 ,这样更能说明两个变量之间的相关程度。

2.2　采用聚类分析方法探讨家庭结构与购房倾向的关系

我们知道不同的家庭结构对住房的需求也不一样 ,但是这仅仅是一个很模糊的概念 ,究竟这不一样体现在哪里 ? 是否能够更进一步将家庭结构与购房倾向之间的关系具体形象的表现出来 ? 采用聚类分析可以在这一方面做些工作 ,聚类依据的是相同或接近的家庭结构对住房的需求相同或相近。我们可以根据大量的普查数据将不同家庭结构的购房倾向作一些分类 ,从中找到一些规律。

在第五次人口普查数据当中 ,增加了对家庭户的房屋状况的普查项 ,这样便有了考察的数据源。可以从普查资料中提取购买了房屋的家庭户的房屋状况(住房建筑面积、购建住房费用、住宅层数等) 以及该家庭户的家庭结构数据。

下面用分裂的聚类方法对住房面积与家庭结构的聚类为例来分析它们之间的关系。



先引入一个抽象的距离概念 ,因为住房面积与家庭结构都可以量化 ,也就是说可以建立一个坐标系 ,横纵轴分别为家庭结构、住房面积。所有提取出来的数据都可以看成是一个坐标对 ,在坐标系中有相应的位置 ,这样不同的点之间就有距离可言。如右图所示。

图中 ,A 点与 B 点之间的距离可用欧氏距离 : $D = \sqrt{(Xa - Xb)^2 + (Ya - Yb)^2}$,或者曼哈顿距离 : $D = |Xa - Xb| + |Ya - Yb|$ 来表现。

这样就可以利用距离来对所有的数据进行分类 ,分裂法分类的原理阐述如下 :

首先假定将所有的 n 条记录分成 k(k ≪n)类,并从 n 条记录中任意挑出 k 条记录作为这 k 类的聚类中心,然后将剩下的记录归并到各类中去。归类所遵循的原则是“同类间距离最短,异类间距离最长”。计算一条记录到各类的聚类中心之间的距离,比较之后,把该记录分入离它距离最近的聚类中心所在的类,接着更新该类的聚类中心的位置,把该类所有记录的中心位置作为新的聚类中心。如此重复,直到分类结束。

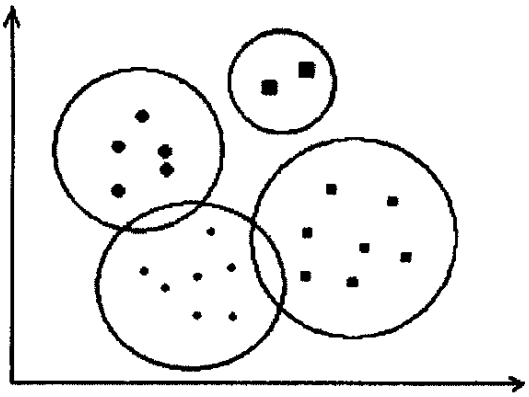


图 2　聚类结果示意图

在具体的分类过程中 k 的值可以根据分类的结果进行调整,同时亦可以采用门限值技术,将离散度的很大类的裂变成两类或者将聚类中心离的很近的两类合并成一类。

聚类过程结束之后,我们可以用图形来描述所有记录的聚类情况,示意(如图 2)。

将最后所得到的各聚类中心描述出来,我们便挖掘出了隐含在大量数据中的住房面积与家庭结构之间的关系。

2.3　采用可视化及数据挖掘语言分析人口因素与商业网点布局的关系

在商业选址的时候,要考虑的一个重要因素便是人口。人们在选址之前往往要分析潜在的客户群分布在哪里,然后尽量选一个离客户群近的地方来布局商业网点。

以书店的选址为例,来分析如何挖掘出潜在的客户群。在布局书店的位置之前,要先分析买书者的特征,我们可以采用抽样调查的方法对买书者的年龄、职业、文化程度等进行比较研究后找到比较一般的特征。一般来说,买书者主要是文化程度相对较高的成人。这样就有必要了解具有这些特征的人的分布情况。我们可以利用数据挖掘技术以及专题图的表现手法对此加以分析。

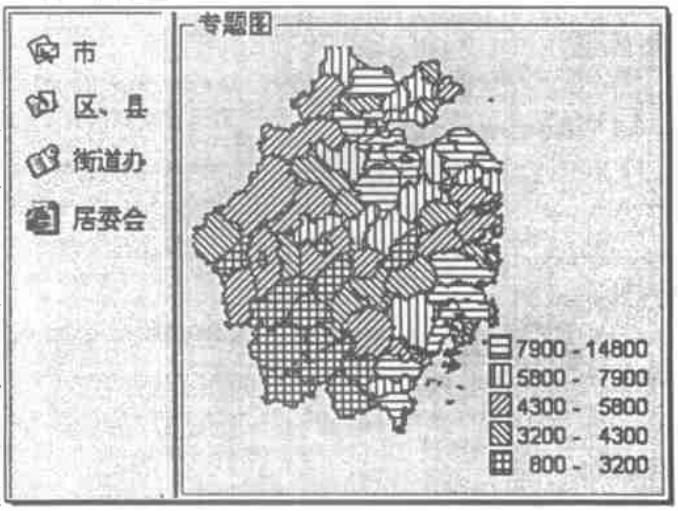


图 3　某地高中文化程度人口分布的密度专题图

在第五次人口普查中地址码共有 14 位,并且可以根据地址码对照表找到它的更高级的所属关系。这样可以按照:调查小区 ≫居(村)委会 ≫街道办 ≫区县 ≫市 ≫省的级别关系来组织。通过这种级别关系,我们可以很清楚地从整体了解到局部的各种分布情况。

这里主要用到的是结构化查询语言(SQL),我们先给定条件,在这里条件为“address = ‘xxxxxxxxx’,and education = ‘1’, and age > = 18”(地址码可以根据不同的级别来相应的给定,文化程度可以是单个值,也可以是全部的文化程度分类),条件给定之后,通过数据库中的结构化查询工具,可以很快得到满足该条件的记录数。当所有的记录数得到之后,我们可以通过专题图

的方式将结果直观、形象、明了地表达出来(如图 3、图 4 所示)。

通过地图上呈现出来的专题图,我们不但可以很快知道潜在的客户分布在哪里,并且可以知道客户的容量大小。

3　结论

数据挖掘在人口领域还有更多的应用,例如对问卷调查的设计、横(纵)向人口研究、微观人口研究等等。以上对数据挖掘技术在人口领域中的应用仅作了一些探讨,随着研究的深入,我相信数据挖掘技术会给人口研究工作带来更多的便利。

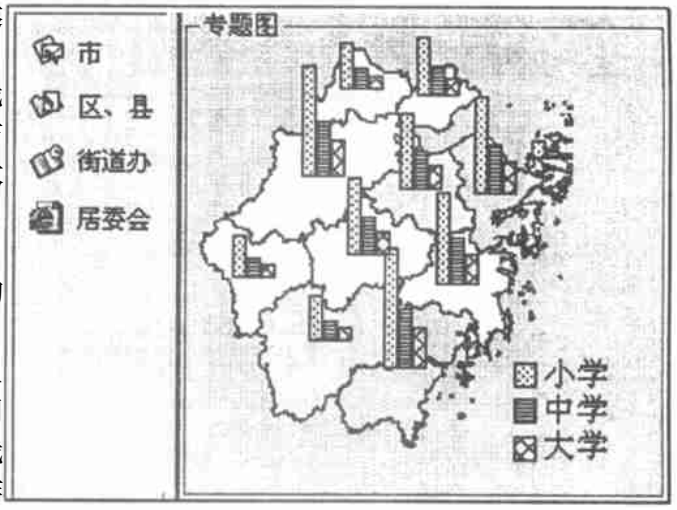


图 4　某地各类文化程度分布的柱状专题图

参考文献

[1]Jiawei Han、Micheline Kambr. *Data Mining: Concepts and Techniques*, Higher Education Press, 2001。

[责任编辑:陈功]

(上接第 7 页)

综上所述不难得知,该系统在人口变动分析与预测工作中,有一定的理论意义和应用价值。

参考文献:

[1]彭祥光、郭文燕. 邓氏灰色动态模型在人口变动和预测中的应用[J]. 统计教育,(增刊)1996:52 - 53.

[2]武汉市统计局. 武汉五十年[Z]. 武汉市. 武汉统计学会. 1999.

[责任编辑:李涌平]