DOI:10.13546/j.cnki.tjyjc.2009.09.004

数据挖掘技术在经济统计中的应用探索

辛金国[a],柯　芳[b],李绍君[b],夏静波[b]
(杭州电子科技大学 a.管理学院;b.财经学院,杭州 310018)

摘　要:目前,经济统计数据的特点是经济统计数据库的数据庞大而且数据质量问题突出。传统的统计数据处理方式还是停留在统计报表层面,没有智能性地深层次地分析处理,以至于数据间的潜在关系和规则未被发现和利用,并且有部分虚假数据未能识别出来,使得这些经济统计的原始数据对于管理决策支持有限。文章提出将数据挖掘应用于经济统计系统中,并对数据挖掘如何在经济统计中应用作了初步探索,可为经济决策提供支持。

关键词:数据挖掘;经济统计;关联规则;决策树
中图分类号:C32　　文献标识码:A　　文章编号:1002-6487(2009)09-0024-04

经济统计数据库中积累了大量的数据,一般的数据利用主要借助于数理统计知识,资料开发的形式单一,深度不够,仅是对现有资料的整理和保存。在统计过程中,处理数量大、涉及面广的数据,仅仅依靠传统统计方法是无法完成这类数据的分析的。采用数据挖掘对经济统计数据进行深度和广度的挖掘分析,不仅可以提高数据的质量,而且可以从原始数据中获取新的信息内容,对政府和企业都具有重大意义。本文试图就数据挖掘技术如何在经济统计应用进行探索。

1　数据挖掘技术在经济统计中的应用

数据挖掘的整个过程包括数据处理、数据挖掘、结果的输出和评价等步骤。

1.1　数据预处理

经济统计数据预处理是数据挖掘过程中的一个重要步骤,因为现实收集到的数据是不完整的(感兴趣的属性没有值)、含噪声的(数据中存在着错误、或异常即偏离期望值的数据)和不一致的(数据内涵出现不一致情况)。我们都知道没有高质量的数据,就没有高质量的挖掘结果,因此在挖掘之前就一定要对数据进行预处理。

1.1.1 数据清理

数据清理就是通过消除原始数据集中的错误、噪声、缺损、不一致等元组,来提高数据质量。在经济数据的统计过程中,常常会遇到收集的数据缺失或是出现异常值的情况,此时就必须要进行数据预处理。一般我们可以采用如下方法处理空值和噪声。

(1)均值法。如果当前数据点是空值或噪声数据,则可采用均值法进行处理。即用数据库中该属性已知的属性均值填充空缺,具体为当前点前 K(K 可以自定义)个不为空的数据点的平均值来替换。其公式为:

$$C_i=\sum_{i-k}^{i-1} C_j/k$$

其中,C_i 表示当前数据点的值,C_j 表示当前数据点前(后)不为空的数据点,K 表示取多少个数据点。

例如:统计局进行各个企业生产总值统计时,有的企业由于某些原因没有提供当年该指标的数据,此时统计机构可以采用均值法来确定当年的数据,假如该指标前 5 年的数据是 100,110,132,145,163;那么当年的数据就是(100+110+132+145+163)/5=130。

以此类推,统计机构在处理其他缺失值或者含噪声的数据时都可以采用均值法来处理,这种方法处理起来比较简单,其缺点就是没有考虑以前年度的数据对当年实际数据的影响程度。一般来说与当年相隔时间越短的年度的数据对当年的数据影响越大;而较早年度的数据影响会相对小些。

(2)平滑法。如果当前数据点是空值或噪声数据,则取出当前点前 k(k 可以自定义)个不为空的数据点的加权平均值来替换。公式如下:

$$C_i=\sum_{i-k}^{i-1} W_jC_j/\sum_{i-k}^{i-1} W_j$$

其中,C_i、C_j 和 k 的含义与平均值法相同,W_j 表示 C_j 数据点的权值。

仍以上例为例进行说明,统计人员根据自己的职业经验和技术水平估计该企业前五年数据对企业当年数据的权重分别是 0.1、0.1、0.2、0.2、0.4,那么得出当年的数据就是 $\frac{100\times0.1+110\times0.1+132\times0.2+145\times0.2+163\times0.4}{0.1+0.1+0.2+0.2+0.4}=141.6$。平滑法的优点就是考虑不同年度的数据对当年实际数据的不同影响,这是很符合实际情况的。一个企业的生产能力,经营规模,行业水平都是跟最近年度联系很紧密的,所以近年度的数据对当年的影响就相对大些。平滑法克服了均值法的不足之处,如果当年的实际数据受各年的影响不大的时候也可以采用均值法。

(3)预测法。采用回归、拟合、插值、判定树归纳等方法,推断空值或噪声数据属性最可能的取值。它通过考虑其它属性的值,最大限度的保持填入的属性值和其它属性及属性值之间的联系。

基金项目:浙江省统计局 2008 年课题资助项目

(4)频率统计法。此方法既可以用于离散数据,也可用于经过离散化的连续数据的数据缺损处理。具体方法为:设数据库 D 中的属性存在空值或噪声数据,属性 a 的值域为$\{V_{a1}, V_{a2}, \cdots\cdots, V_{am}\}$,$P(V_{ai})$表示值 V_{ai} 在该信息系统中出现的频率。可以用最大出现频率的值 $\max\{P(V_{ai})\}$来填充。

如果某个企业或者某个地区一些指标的数据统计不到的情况下,这时统计人员可以采用频率统计法来处理,将同行业中各方面条件都比较类似的企业的数据进行统计,将出现概率最大的那个数据作为该企业的数据。

1.1.2 数据集成

数据挖掘常常涉及到来自多个数据源的数据,这样就需要我们把这些数据结合在一起形成统一的数据集合,也就是数据集成。数据集成是经济统计中经常用到的,统计部门在统计数据时,先由各个地方的统计局收集统计本地区的经济数据,然后由各个地区的数据进行集成。在数据集成过程中我们主要考虑到以下几个方面问题。

(1)模式集成。即如何使来自多个数据源的现实世界的实体相互匹配,这就涉及到实体识别问题。例如,如何确定一个数据库中的"std_id"与另一个数据库中的"std_no"是否表示同一实体。一般可以使用数据库与数据仓库包含元数据来帮助避免在模式集成时发生错误。各个地区在收集数据时所使用的数据库和模式有可能不一致,以及各个企业登记统计数据时所使用的数据库和模式也可能不一致,这就需要进行模式集成。

(2)冗余问题。若一个属性可以由其它属性推演出来,那么这个属性就是冗余属性。比如人均国内生产总值就是冗余属性,它可以根据国内生产总值和总人口属性计算出来。除此之外属性命名的不一致也会导致数据集成后出现不一致情况。我们可以利用相关分析来发现一些数据冗余情况。例如,给定两个属性,则可以根据这两个属性的数值分析出这两个属性间的相互关系。属性之间的相互关系可以根据以下计算公式来分析:

$$r_{A,B}=\frac{\sum(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}$$

其中,n 表示元组的个数,$\bar{A}$和$\bar{B}$分别代表属性 A 和 B 的平均值;σ_A 和 σ_B 分别表示属性 A 和 B 的标准方差。如果 $r_{A,B}>o$,则属性 A,B 之间是正关联,也就是 A 增加,B 也增加;该值越大,说明两个属性的正关联关系越密。如果 $r_{A,B}=0$,则两个属性相互独立,没有关系。如果 $r_{A,B}<O$,则两个属性之间是负关联,也就是 A 增加,B 也减少;$r_{A,B}$ 的绝对值越大,说明两个属性的负关联关系越密切。

在以下方法的应用介绍中,主要以浙江省为例,类似地,可以推广到其他省份以及国家统计中。

以浙江省生产总值和浙江省国内旅游收入为例,说明上述方法的具体应用。浙江省生产总值和浙江省国内旅游收入 2001 年到 2006 年的数据如表 1 所示。

表 1　2001~2006 年浙江省生产总值和浙江省国内旅游收入数据

	浙江省生产总值(单位:亿元)	浙江省国内旅游收入(单位:亿元)
2001	6898.34	529
2002	8003.67	633.8
2003	9705.02	695.3
2004	11648.8	1012.5
2005	13437.85	1239.7
2006	15742.51	1519.6

根据表 1 数据我们可以得出 $r_{A,B}=\frac{\sum(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}=0.98$,由此可知国内旅游收入与省生产总值的正关联关系很密切,国内旅游收入增加时,全省生产总值也在增加。

(3)数据值冲突检测与消除。如对于一个现实世界实体,其来自不同数据源的属性值或许不同。产生这样问题原因可能是表示的差异、编码的差异或比例的不同等。例如,经济统计属性在一个系统中采用人民币,而在另一个系统中却采用美元。同样属性不同系统采用不同计量单位,这些差异为数据集成提出许多问题。

1.1.3 数据变换

数据转换就是将数据转换成一个适合数据挖掘的描述形式,包括数据泛化和数据规范化等。数据泛化就是用更抽象或更高层次的概念来取代低层次或数据层的数据对象。例如经济发展水平,就可以映射到更高层次概念,如低水平、中等水平和发达水平。数据规范化是将一个属性取值范围投射到一个特定区间之内,比如 1 到 10。主要是为了消除数值型属性因大小不一而造成挖掘结果的偏差。规范化的方法有许多,如最大最小规范化、零均值规范化和十基数变换规范化方法等。现介绍最大最小规范化的具体做法。

设 minA 和 maxA 为属性的最小值和最大值。最大最小规格化方法将属性 v 的一个值映射为 v′ 且有任意 v′ [new_minA,new_maxA],其映射计算公式如下:

$$v'=\frac{v-\min_A}{\max_A-\min_A}(new\_max_A-new\_min_A)+new\_min_A$$

实例分析:浙江省的人均生产总值从 2000 年到 2006 年的数据如下表所示,从这些数据中不能一眼看出人均生产总值增长的快慢,以及相邻年份变化的大小,但是把这些数据映射到 1~100 的范围就能明显的看出这些数据的变化情况。下面我们采用最大最小规范化线性变换公式把原始数据规范映射到 1~100 之间:

$$v'=\frac{v-\min_A}{\max_A-\min_A}(100-1)+1$$

最大最小规范化以 2000 年的数据为最小值,2006 年的数据为最大值,规范化后的数值如表 2 所示。这样就可以清楚地看出各年的人均生产总值与 2000 年和 2006年相比的水平。

表 2　浙江省人均生产总值规范化表

年份	2000	2001	2002	2003	2004	2005	2006
浙江省人均生产总值(单位:元)	13416	14713	16978	20444	24352	27703	31874
规范化后的值	1	8	20	39	60	78	100

资料来源:www.zj.stats.gov.cn

1.1.4 数据离散化和概念分层

离散化技术方法可以通过划分区间,来减少一个连续(取值)属性的取值个数。可以用一个标号来代替一个区间内的实际数据值。概念分层通过利用较高层次概念替换低层次概念来减少原来数据集。现实生活中的数据是连续型的,而许多知识发现算法对于连续的取值无法处理。虽然在数据泛化过程中消失了一些细节,但这样所获得的泛化数据会更易于理解、更有意义。在消减后的数据集上进行数据挖掘与在

未概化的数据上挖掘相比，显然所需的 I/0 操作更少，并且效率更高。

数据离散化的方法有固定区间法和等分区间法。

(1)固定区间法。如对于浙江省各地区的经济发展水平，可以将人均生产总值增长率高于 20%的地区设定为“经济发达地区”，[10%，20%]设定为“经济中等发达地区”，低于 10%的地区设定为“经济欠发达地区”，则对于浙江省内的每个地区，我们都可以将他们归入三类发展水平中的一类，得到各个地区对应的经济发展水平。这样政府就可以有针对不同地区施政，以提高政府施政的效率。

(2)等分区间法。从所选取的数据源中，找出最大值和最小值，按照所需要的区间数目对其进行等分，得到增长率的离散区间。假设浙江省各个地区的批发零售贸易商品销售总额的增长率最低值是 10%，最高值是 40%，若划分为 3 个区间，则[30%，40%]设定为“3”，[20%，30%]设定为“2”，[10%，20%]设定为“1”。这样对于每个地区的批发零售贸易商品销售总额的增长率，我们都能得到它的对应离散值。

概念分层是指从低层概念的集合到它们所对应的更高一层的映射，此映射将概念集合以偏序方式组织。一个概念分层可以在一个属性域上或一个属性域集合上定义。假设一个分层 H 是定义在域的集合 D1，…，Dk 上，其中不同的概念层次组成一个分层，概念分层通常从一般到特殊的顺序以偏序的形式排列。定义格式为：HI：Di×…×Dk⇒HI-1⇒H0，其中 HI 表示为最原始的概念集；HI-1 表示比 HI 更高一层的概念；H0 为最高一层的概念。

本文是针对经济统计数据进行的，所以我们就着重介绍一下数值属性的概念分层。概念分层的方法有很多，实际中用到的方法有自然分段法。自然分段法需要将数值区间划分为归一的、易读懂的间隔，以使这些间隔看起来更加自然直观。3-4-5 规则可以用于将数值数据划分成相对一致和“自然”的区间。该规则根据最重要的数字上的值区域，递归地和逐层地将给定的数据区域划分为 3，4 或 5 个等宽区间。该规则如下：

(1)如果一个区间在最高有效位上包含 3，6，7 或 9 个不同值，则将该区间划分成 3 个区间 (对于 3，6 和 9 划分成 3 个等宽区间；而对于 7，按 2 一 3 一 2 分组，划分成 3 个区间)；

(2)如果它在最高有效位上包含 2，4 或 8 个不同的值，则将区间划分成 4 个等宽区间；

(3)如果它在最高有效位上包含 1，5 或 10 个不同的值，则将区间划分成 5 个等宽区间。

该规则可以递归地用于每个区间，为给定的数值属性创建概念分层。由于在数据集中可能存在特别大的正值和负值，最高层分段简单地按最小和最大值可能导致扭曲的结果。为解决这一问题，顶层分段可以根据代表给定数据大多数的数据区间(例如，第 5 个百分位数到第 95 个百分位数)进行。越过顶层分段的特别高和特别低的值将用类似的方法形成单独的区间。

1.2 关联规则和决策树的应用

关联规则挖掘是数据挖掘领域研究的一个重要课题，也是最活跃的研究方向之一。关联规则反映了大量数据中属性之间有趣的关联或相关联系，查找存在于属性集合或对象集合之间的频繁模式、关联、相关性或因果结构，以发现隐藏在数据之中的、不易被人们发现的、甚至与人们的意志相违背的关联事件。

在经济统计数据中应用关联规则，可以发现各个地区以及全国的经济发展特点和关系密切的行业。我们可以通过各个属性的数据，来寻找有关联关系的属性。有些属性之间的关系是早已经被大家所知晓的，例如工业总产值的大幅增加必然会导致生产总值的增加，我们把这样的规则称为平凡的规则。利用平凡规则可以检验统计数据的真伪和质量，也就是说利用我们熟悉的一些关联关系来检验统计得到的数据是否吻合这些关系，如果不吻合则说明统计到的数据是不真实的，从而需要对这些数据进行专业处理，以消除不真实的水分。假如统计机构统计到浙江省某市的第二产业生产总值比前一年下降了 20%，而第一产业和第三产业没有特别明显的因素促使本年会大幅增加产出的情况下，统计得到的该市国内生产总值却比前一年有较大的增幅。此时，统计人员应该对这样的数据表示怀疑。因为第二产业是浙江省经济发展的支柱产业，占全省生产总值的绝大部分。根据第二产业生产总值与浙江省国内生产总值正相关的平凡规则得出该市的国内生产总值存在虚假成分，这可能是统计人员统计过程中的失误，也可能是政府部门为了提高自己的政绩在造假。平凡的规则是人们通过过去的一些研究成果和经验所熟悉的，而数据挖掘的主要贡献是可以发现人们所不熟悉的而又是对人们有帮助的一些关系。下面以浙江省为例介绍关联规则在经济统计数据中的具体应用。

浙江省经济在上世纪 90 年代崛起以后，经济发展非常迅速，成为现在全国的经济大省。各界人士普遍将私营企业和民间资本视为浙江省经济崛起的重要原因。但是，在浙江省内我们也应该看到各市地的经济发展水平仍存在着很大的差异。

表 3　　2006 年浙江省各地区的原始数据

地区名称	人均 GDP	居民储蓄存款	社会消费品零售总额	实际利用外资
杭州市	51878	2473.69	1112.37	225536
宁波市	51460	1752.04	882.54	243018
嘉兴市	40206	850.43	429.72	122178
绍兴市	38540	1045.04	440.64	97188
舟山市	34682	204.9	113.97	5003
湖州市	29527	386.94	273.8	75702
金华市	27108	940.16	482.08	51013
台州市	26026	858.2	511.57	41354
温州市	24390	1476.34	779.15	46273
衢州市	15740	227.85	153.86	3985
丽水市	14104	257.88	145.86	1926

资料来源：www.zj.stats.gov.cn

将浙江省各地区按人均 GDP 水平划分为三个等级，总共有 11 个地区的数据(见表 1)，按人均 GDP 从高到底排序，前三个为“经济发达地区”取值为“A”；中间四个为“经济中等发达地区”取值为“B”；最后四个为“经济不发达地区”取值为“C”。由于各地区所处地理位置、经济发展基础不同，以及所处环境、发展阶段等多方面因素的影响，其他三个指标的个体差异还是比较大的。因而采取简化措施，以此 11 个地区为准，计算各指标的平均值，各地区数值与平均值比较大小后的逻辑值作为待挖掘分析的数据。

实际计算平均值为居民储蓄存款 952.13、社会消费品销

售总额 484.14、实际利用外资 83016。其中关系比较运算符三个都是大于号。比较大小后的逻辑值真用“1”表示，假用“0”表示，实际计算结果如表 4 所示。

表 4　　三级分类三种指标计算结果

地区名称	人均 GDP	居民储蓄存款	社会消费品零售总额	实际利用外资
杭州市	A	1	1	1
宁波市	A	1	1	1
嘉兴市	A	0	0	1
湖州市	B	0	0	0
绍兴市	B	1	0	1
舟山市	B	0	0	0
金华市	B	0	0	0
温州市	C	1	1	0
衢州市	C	0	0	0
台州市	C	0	1	0
丽水市	C	0	0	0

关联分析应用的目标是希望通过对各地区的经济指标量化处理后，挖掘出各地区经济指标与分类结果的关联关系，以及各经济指标的关联程度。如果设最小支持度阈值为 20%(≤20%)，即在 11 个地区中出现的次数≤2。按照 Apriori 算法的计算比较过程如表 5。其中 I 表示项集，c 表示计数。

扫描所有数据项，得到 1 一项集的计数，与最小支持度比较均满足。连接 1 一项集得到 2 一项集，再扫描所有数据项，得到 2 一项集的计数，与最小支持度比较均满足。连接 2 一项集得到 3 一项集不满足最小支持度，被删除。

表 5　　Apriori 算法的计算比较过程

I	C	I	C	I	C
1	4	1,2	3	1,2,3	2
2	4	1,3	3		
3	4				

对于上述结果我们可以得到在已知分类结果的情况下，了解到 I1 到 I3 三项指标各自对每一分类结果的影响程度，即支持度，以及三项指标综合起来对每一分类结果的支持度。至于置信度，由于局限在各类子样本集合中，包含 Ij(j=1，2，3)和 I 的事务个数必与 lj 的个数相等，故全为 100%。计算支持度的结果如表 6。

表 6　　三项指标单独对每一分类结果关联关系

I	S(I1,I)	S(I2,I)	S(I3,I)	S(I1,,I2,I3,I)
A	67%	67%	100%	67%
B	25%	0	25%	0
C	25%	50%	0	0

对于表中的支持度的解读：

(1)四个为零的孤立值。经济中等发达地区的社会消费品零售总额都没有达到平均水平；经济不发达地区的实际利用外资情况都没有达到平均水平；三项指标都达到平均水平的情况下，不可能是经济中等发达地区和经济不发达地区。

(2)正常情况，三项指标随着分类结果的级别的降低，其支持度也应该逐级降低。实际情况是在经济发达地区→经济中等发达地区→经济不发达地区，居民储蓄存款对分类的支持度分别为 67%→25%→25%，此时后面两个分级支持度相同可能是因为取样的局限性所致；社会消费品零售总额对分级的支持度分别为 67%→0%→50%，貌似与我们的结论矛盾，其实这是符合浙江省经济特点的；实际利用外资对分级的支持度分别为 100%→25%→0%，这是符合结论的；三项指标综合对分级的支持度分别为 67%→0%→0%，这说明综合指标对分级是非常支持的。

决策树算法是另一种常用的、直观的快速分类方法。其关键是决策树的构建，决策树的构建包括建树和剪枝两个阶段。用决策树进行分类分两步走。第一步是利用训练集建立并精化一棵决策树，建立决策树模型。第二步是利用生成完毕的决策树对输入数据进行分类。对输入的记录，从根节点依次测试记录的属性值，直到到达某个叶子节点，从而找到该记录所在的类。决策树建树算法是一个递归的过程，直到满足某种终止条件。停止分割的条件有两个：一个是当一个节点上的数据都是属于同一个类别；另一个是没有属性可以再用于对数据进行分割。剪枝的目的是降低由于训练集存在噪声而产生的起伏。决策树以及决策规则属于以逻辑模型方式输出的分类方法。主要用来解决数据挖掘中的分类和预测问题。

在决策树的生成阶段要对决策树进行必要的修剪。我们常用的修剪技术有两种：一种是预修剪；另一种是后修剪。而决策树的质量主要依赖于停止规则，获取大小合适的树常用的方法是后剪枝。后剪枝的方法主要有三种：训练和验证集法、使用统计的方法和最小描述长度准则的方法，而其他的剪枝方法主要有限制最小结点规模、两阶段研究、不纯度的阈值、将树转变为规则和 Tree reduction 等方法。

在经济统计的大量数据中，我们可以应用决策树来进行各种分类以及各种预测。可以选择多个指标进行构建决策树也可以选择少量指标构建决策树。

2　结论

数据挖掘技术在经济统计中的应用是未来统计工作的必然趋势，是信息技术高速发展的结果。它可以进行深层次的数据处理分析，提高数据质量，由此帮助政府更有效地制定政策、拟制计划和日常行政管理。并且可以为工业企业以及社会经济研究服务，创造巨大的社会和经济效益。

在经济统计工作中引进数据挖掘技术，将该技术在经济统计中的应用普遍化，能有效地提高统计工作的效率，降低统计成本。数据挖掘的成果将会在统计系统的工作中发挥重要作用，具有良好的应用前景。同时，利用数据挖掘技术从大量的数据中挖掘出有价值的信息，有利于相关部门在大量经济数据的基础上制定各种经济政策，促进我国经济良性发展。

参考文献：

[1]Jiawei Han,Micheline Kamber.数据挖掘概念与技术[M].北京：机械工业出版社，2001.

[2]Pang-Ning Tan,Michael Steinbach,Vipin Kumer.数据挖掘导论[M].北京：人民邮电出版社，2006.

[3]Kantardzic M. 数据挖掘——概念、模型和算法[M].北京:清华大学出版社，2003.

[4]朱玉全等.数据挖掘技术[M].南京：东南大学出版社，2006.

[5]王斌会.数据挖掘技术及其应用现状[J].统计与决策,2006，(10).

[6]刑莉.统计分析的新模式——数据挖掘技术[J].统计与咨询，2006，(4).

[7]行智国.数据挖掘及其在官方统计中的应用前景[J].江苏统计，2003，(2).

(责任编辑/亦　民)