时间序列预测与关联规则的维护*)

The Time Series Prediction and Updating of the Associated Rules

王 刚 邱玉辉 程小平
(西南师大计算机科学系 重庆400715)

Abstract This paper discusses how to gain the future predictive events by means of the time serial analysis and discusses the problem in the updating of the associated rules and develops an algorithm to solve these problems. Using an example to verify it can gain more useful, effective associated rules.

Keywords Data mining, Time series prediction, Events, Data sets, Frequency

1 引言

数据开采技术已经引起了国际上人工智能和数据库专家学者的强烈关注，其核心就是要从庞大的数据集里发现知识，为人们管理、决策提供科学依据，而对关联规则的发现一直是数据开采的热门话题，从 Agawal 首先提出 Agrion 算法以来，产生和改进了许多有效的算法和模型[1]。然而，不得不面临的一个现实问题是，面对海量数据以及数据自身、之间复杂的关系、数据集随时间不停地变化、不同时期的数据集之间的联系、怎样才能正确地维护、继承已经开采出来的规则，以期发现更多、更准确的知识。本文在关联规则的维护中，运用时间序列对事件进行预测[2]，并把预测结果用于关联规则的改进，起到了更好的维护关联规则的作用。

2 数据挖掘的相关定义及时间序列事件

设 $I(i_1, i_2, \cdots, i_m)$是一组物品集，D 是一组事务，每个事务 T 是一组物品。关联规则的形式描述，$x \Rightarrow y$，$x \subset I, y \subset I, x \cap y = \Phi$

支持度的定义：$s = s(x) = \text{support}(x) = D(x)/D$，[3]即 D 中有 s%的事务支持物品集 D(x)。本文中用 old_db 表示老的数据集，new_db 表示最近一段时间得到的数据集。设 s1(x)表示 D(x)在 old_db 中的支持度，s2(x)表示 D(x)在 new_db 中的支持度。文中将支持度简化为事件出现的次数与所有事件总数的比。

置信度的定义：$d = d(x) = D(x \cap y)/D(x)$即若 x 的物品集中支持 y 事务的百分比 d%即为 $x \Rightarrow y$ 的置信度。

在现实生活中，时间是影响事件的重要因素，事件定义为 $X = \{x(1), x(2), \cdots, x(t)\}$，t 表示时间，另外，事件在程序中与数据对应。事件的频率定义为单位时间内发生的事件，事件的分类主要有：

周期事件，指相同的事件每隔一定时期又会发生、如图1①所示，其中●表示出现的事件，(下同)。

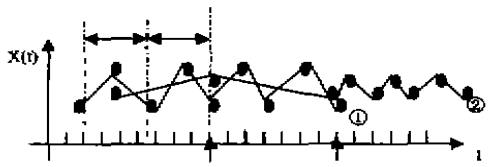

图1

随机事件，事件的出现是随机的，是没有规律的，如图2①所示。

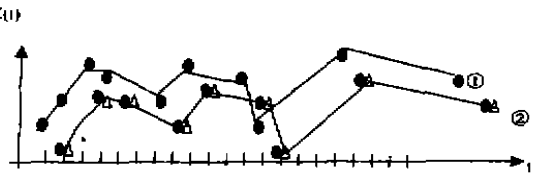

图2

相关联的事件，即如果事件 a 发生，则 b 一定发生，与此相对应的是不相关的事件如图2②所示，其中●，△表示 a，b 事件。

3 关联规则的维护分析[4]

假设对数据集 database1进行了关联规则的挖掘，得到了大项集 cl，在最近一段时期得到数据集 database2。现在总的数据集为 Database1 ∪ Database2，

*)受教育部"现代远程教育关键技术研究重点项目"资助。

要在此基础上得到大项集c2，有两种方法，第一种方法是重新对Database1∪Database2用相关算法重新进行计算，这种方法在理论上是可行的，但实际中，考虑到时间、空间上的复杂性，这种方法是不可行的。第二种方法在理论上是考虑到利用c1的结果，可以避免庞大的计算量，相关算法如IUA[5](increment updating algorithm)、PIUA(parallel increment updating algorithm)、FUP(frequent updating proceed)等，但这些算法基本上考虑的是当前静态的数据集，而忽略了时间因素对将来可能出现的数据集的影响，这样，在算法中对规则集的取舍也就有局限性，可能忽略重要的规则。一个简单的例子是：

从图(如图1②)可见，在点x以前，由于事件d(x)出现的次数较少，大项集中不能包含d(x)，在点x以后出现的频率较高，如在大项集中去掉d(x)，容易造成规则的丢失，而在点y和x之间d(x)已经有频繁出现的可能，因此在对大项集中的取舍时，要考虑到这些情况。

4 数据集及对事件的预测

用时间序列对事件进行预测[6]就是根据目前的事件(数据集)，通过分析，得出未来的可能事件(数据集)。

(1)数据集的分类主要有：动态的，数据时时处于变化之中，与时间密切相关；静态的，数据在某一段时间内表现为不变化；规则的，指数据的变动有规可循，如周期变化(见图1①)，线性变化(见图3①)或遵循其它函数(如图3②)；不规则的，指数据的变化是随机性比较强的，事件的出现是杂乱的(如图3③)。

(2)预测：对规则的数据集可以直接运用相关的函数进行预测[7]。如x(t)=1/2t，就知道在t=4时，事件x(4)=2，可以知道在未来一段时间内发生的事件及事件出现的频率1/T(x)；对不规则的数据集，准确预测是非常困难的，由于在本文中只涉及事件出现的次数，因此可以用前一段时期内某事件出现的次数a与总的次数β的比a/β为该事件出现的频率1/T(x)，利用此频率，来估计在未来的一段时间内，事件将要出现的次数和支持度。

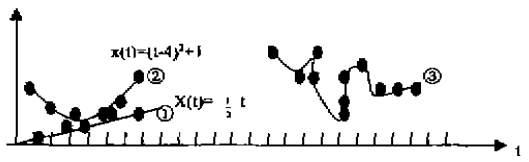

图3

新的支持度S(x_new)=($\frac{T_1(x)}{T(x)}$/r)+s(x_old)，其中$T_1(x)$表示将来的一段时间，r为总的事件数。

5 算法及分析

(1)IUA算法要求上一次对数据集进行采掘得到的大项集在这次采掘中是可以得到的，通过比较最小支持度，确定对大项集的修改，若S(new)>s(old)，则在c(old)中删除小于s(new)的x，若s(new)<s(old)，则在继续一部分c(old)的同时，构建c(new)。

(2)FUP算法则说明若x在new_db和old_db中均为大项集，则x仍为大项集。若x在new_db和old_db中均为小项集，则x仍为小项集。若x在new_db和old_db中任一为小项集，则必须重新计算x新的s(x)，以确定是为大项集或小项集。

(3)以上算法忽略了时间因素，未能够考虑到将来支持度变化的情况，使得在将来一定时间内，支持度能够达到最小支持度的大项集被忽略了。例如c(x)在db_old中是小项集，在db_new中也是小项集，由于数据处于急速变化中，如果在以后一段时间内，c(x)由于x的增加，可能成为大项集而被保留，因此对c(x)应该采用新的支持度。

$$S(x\_new)=(\frac{T_1(x)}{T(x)}/r)+s(x\_old)$$

6 算法描述

在对大项集支持度的判断时引入新的支持度s(x_new)。

```
if(s2(x)>s(x_new))//用新的支持度比较
{
  for(k=1;k<=m;k++)
Lk={c∈Lk|s'(x)>=s(s_new);
    //符合条件者，加入大项集
  Return ∪k L'k
;
```

7 实例

如旧的数据集为{(e1,e4,e5),(e3,e4,e5),(e2,e3,e4)(e1,e2,e5)}，新的数据集为{(e1,e2,e3,e4,e5),(e1,e2,e4,e5),(e1,e4,e5),(e3,e4,e5)}按照没有改进的算法得出的大项集包括：

{(e1),(e2),(e3),(e4),(e5),(e1,e5),(e4,e5),(e1,e4),(e1,e2),(e2,e5),(e3,e4),(e3,e5),(e1,e4,e5)}。

根据原先算法，S(x_new)=3/8而s({e1,e2,e4})=2/8，由于不满足最小支持度3/8，所以该集合被丢弃，根据改进的算法，它们出现的频率为2/4，在将来的一段时间T=2内，其出现的次数2*0.5=1加上原先的次

(下转第73页)

放的环境，很少用于多 Agent 系统中。完全分布的协调方法主要有两种类型：(1)不须交互，但需要其它 Agent 的许多信息和知识。(2)需要各 Agent 直接交互。显然，第一类协调法不适合动态、开放环境中的多 Agent 系统，我们重点分析第二类协调法。应用第二类协调法的多 Agent 系统结构实际上就是 Agent 网络结构。在这样的系统中，当 Agent 数目较大时，有明显的几个缺陷：一是通信代价太大，导致系统的低效率；二是这样的 Agent 网络的实现太复杂；三是每个 Agent 都具备协调管理能力，在系统开发上将产生大量重复劳动，不符合软件重用的思想；四是这样的体系结构不能满足系统动态、开放性的要求，因为 MAS 中的每个 Agent 都要拥有大量其它 Agent 的信息和知识，这在动态，开放的系统中是做不到的；五是这样的协调方法不符合人类社会组织的协调原理，在一个大型社会组织里一定有一个或多个协调管理中心，这种组织的工作才是高效的。因此，一般的多 Agent 系统都使用协调管理 Agent，这将大大提高系统性能和效率，减少系统实现的复杂度，有利于软件重用。综上所述，完全分布的协调方法不适合于动态、分布环境中的多 Agent 系统，特别是大型多 Agent 系统。最后，我们考察一下集中与分布相结合的协调方法。这类协调方法是指系统中的 Agent 组成层次结构，上层的协调管理 Agent 对下层的 Agent 有部分的控制能力，和完全集中的协调相比，既提高了系统整体性能和效率，更接近人类社会的特征，也体现了系统中 Agent 的自主性，适合应用于动态、开放的多 Agent 系统中[6~8]。正是基于上述考虑，我们的协调策略都是集中与分布相结合的。另外，我们根据 Internet 环境的特点，建立了用户模型、业务模型和协调模型，并在此基础上建立了系统的仿真模型，通过仿真分析解答本文开头提出的问题。用仿真的方法进行多 Agent 系统的性能分析尚不多见。

本文的工作为在 Internet 环境下设计和开发面向自动文摘的多 Agent 系统提供了依据。而且，只要对诸如 MAS/ABS 适用的领域数量、系统的访问量或负载等参数作适当的修改，我们可以得到更多具有指导意义的结果。正是在上述工作的基础上，利用北京邮电大学智能研究中心研制的几个基于单机的自动文摘系统，我们在 Internet 环境下建造了一个面向自动文摘的多 Agent 系统。目前，系统采用的是 MinQ 协调算法。

参 考 文 献

1 Hu Shungeng, Zhong Yixin, Wei Chaocheng. An automatic abstracting architecture based on multiagent technologies. In: Proc. of Intl Conf on MT & CLIP. Beijing, China, 1999
2 胡舜耕，钟义信，魏超成. 基于多 Agent 技术的自动文摘研究. 计算机工程与应用(已录用)
3 Zipf G K. Human Behaviour and the Principles of Least Effort. Cambridge, Mass.: Addison-Wesley, 1949
4 Colajanni M, et al. Analysis of task assignment policies in scalable distributed web-server systems. IEEE Trans. on Parallel and Distributed Systems, 1998, 9(6): 585~600
5 Cunha C, et al. Characteristics of WWW Client-Based Traces: [Technical Report BU-CS-96-010]. Computer Science Dept., Boston Univ., Apr. 1995
6 李建民，石纯一. DAI 中多 Agent 协调方法及其分类. 计算机科学，1998, 25(2)
7 Miao X, et al. A Normative-Descriptive Approach to Hierarchical Team Resource Allocation. IEEE Trans on Sys. Man. and Cyb., 1992, 32(3)
8 Genesereth M R, et al. Software Agents. Communications of the ACM, 1994, 37(7): 48~53

(上接第68页)

数其最小支持度为3/8，可以满足最小支持度，因此应该在大项集中加入这个集合(如图4)。图中 * 表示预测要发生的事件。

图4

小结 本文把时间序列预测用于关联规则的维护过程，分析了用时间序列来预测事件的发生，把事件与数据集对应起来。分析了关联规则的维护在基于动态变化的数据集时的不足，并加以改进，得到了新的对大项集的判别算法，用实例证明了对关联规则维护的改进过程，得到了更有意义的关联规则，这对关联规则的准确维护起到了重要作用。由于预测的准确性影响着规则的可靠性，这还需要以后不断地改进和探索。

参 考 文 献

1 Ester M. Algorithm For Characterization And Trend Detection. In: Spatial Database Process of 4th Intel Conf. On KDD, 1998. 1~4
2 Povinelli R J. Time Series Identify temporal patterns for characterization and prediction of time series events. 1999, 12: 15~40
3 陆玉昌. 数据挖掘与知识发现. 中国计算机用户，1999, 10: 29~32
4 钱治欣，陈奇. 关联规则采掘综述. 计算机应用研究，2000, 1: 1~5
5 冯玉才，冯剑林. 关联规则的增量式更新算法. 软件学报，1998, 9(4): 301~306
6 董力. 现代经济管理预测与决策. 地震出版社，1999. 66~99
7 时间序列用于经济预测的方法. Available at: http://mba.netbig.com/teach/course/973/2000628/ts/d064.htm