一种基于模糊集的时间序列挖掘算法的设计与实现

吴煲宁[1] 汪晓刚[2] 林 天[1] 孙志挥[1]
[1](东南大学计算机系,南京 210096)
[2](南京擎天科技公司,南京 210008)
E-mail:wu_bn2000@yeah.net

摘 要 针对商业销售的智能分析需求,文章提出了一种基于模糊集合的数据挖掘时间序列模式算法。该算法已得到有效的应用,对企业的经营决策有一定的参考价值。

关键词 模糊集 数据挖掘 时间序列

文章编号 1002-8331-(2002)20-0196-03 **文献标识码** A **中图分类号** TP311.13

Design and Application of a Time Series Data Mining Algorithm Based on Fuzzy Sets

Wu Baoning[1] Wang Xiaogang[2] Lin Tian[1] Sun Zhihui[1]
[1](Department of Computer Science, Southeast Univ., Nanjing 210096)
[2](Nanjing Sky Science and Technology Ltd., Nanjing 210008)

Abstract: Aiming at intelligent analysis request of business enterprise, this paper provides a time series data mining algorithm based on fuzzy sets. This algorithm has been used efficiently and has some value for management decision

Keywords: Fuzzy sets, Data mining, Time series

1 问题提出

数据挖掘一般指在数据库的基础上,从数据集中识别出有效的、新颖的和潜在的最终可理解的规则和知识。数据挖掘的任务是从大量的数据中发现模式,常见的模式有关联模式、泛化模式、聚类模式、分类模式时间序列模式和回归模式等等。目前数据挖掘的方法较多,它们分别从不同的角度进行规则挖掘。

时间序列是指带有时间标记的数据根据时间顺序排列而得到的数据列值集合。典型的时间序列便是股票的市值:每天的市值均随着市场行情而变化,以时间为坐标,可以得到某种股票的具体变化曲线,这就形成了一个时间序列。所谓时间序列模式是指从经济数据中统计出的某种经常发生的时间序列。例如某超市发现人们总是先购买 A 商品,再购买 B 商品,随后再购买 C 商品,且中间时间间隔大致相当,即可以认为此种购买记录形成了一个时间序列模式。这种模式的发掘,对于商业企业的销售经营决策有着一定的指导和参考价值。

一般对于时间序列模式的发掘可采用类似于关联规则发掘的算法。其主要步骤为:

(1)从原始数据集中找出按顾客的购买记录,形成每个顾客的购买时间序列;

(2)从(1)中的时间序列中找出频繁数据集;

(3)设置最小支持度和关联度,对(2)中的频繁数据集求其支持度与置信度;

(4)输出(3)中的支持度与置信度均大于最小值的频繁数据集,即为所求的时间序列模式。

例如有表 1 所示的销售记录,表中的各个记录是按时间先后排序的。

表 1

顾客 \ 商品	商品 1	商品 2	商品 3
01	A	B	C
02	A	C	D
03	B	C	D
04	A	C	E
05	B	C	F

从表 1 中可以看出,序列 A→C 以及 B→C 均有较高的支持度与置信度,可以认为顾客先买 A 再买 C 以及先买 B 再买 C 均有较大的可能。所以商家可以对买 A 或 B 的顾客推荐 C 产品,从而达到预期促销的效果。

实际运用上述方法进行挖掘时,往往会遇到如何确定频繁集的问题。因为不同的商品有着种类的差异以及购买时间上的不同,尤其是两种商品购买时间间隔上的差异,使得确定频繁集问题较多。该文以模糊集合的思想为指导,提出了一种解决这个问题的方法。

2 原理简介

经典的集合论的基本概念是:对于某一元素 u 以及集合 A,u 要么属于此集合 $u \in A$,要么不属于 A。此集合由真或假两

基金项目:国家中小型企业创新基金项目(编号:00C26213211014)
作者简介:吴煲宁,硕士研究生,主要研究方向:数据库系统,数据仓库与数据挖掘。

种值来决定元素与集合的关系。但现实世界中，个体与个体之间的差别往往不能用这样绝对的0或1来度量。模糊集合便是一种基于隶属度概念的新的集合形式。

用 U 表示被讨论的对象的全体，u,v 等表示论域中的元素，A 表示 U 上的某一集合，$\tilde{A}$ 表示 U 上的模糊集合，则模糊集合可视为全集中的一个具有单位宽度边界的一个圆圈，如图1所示。

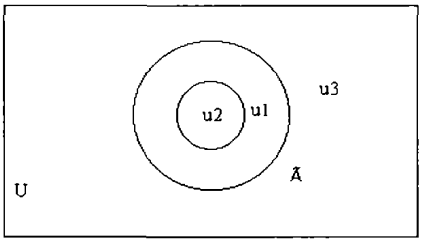

图1

如图1所示，与 $u2$ 类似位于内圈的点以隶属度1属于模糊集合 $\tilde{A}$，与 $u1$ 类似的位于内外层之间的点以隶属度 a 属于集合 $\tilde{A}$（其中 a 为 $u1$ 与外圈的距离），与 $u3$ 类似的位于外圈之外的点则完全不属于集合 $\tilde{A}$，其隶属度为0。

模糊集的严格定义如下：

定义1：所谓给定了 U 上的一个模糊集 $\tilde{A}$ 是指：对任何 $u\in U$，都指定了一个数 $\mu_A(u)\in[0,1]$ 与之对应，它叫做 u 对 $\tilde{A}$ 的隶属度。这意味着作出了一个映射：

$\mu_A(u):U\to[0,1]$

$u\to\mu_A(u)$

这个映射称为 $\tilde{A}$ 的隶属函数。

有了模糊集合的定义，可以用模糊关系来表示元素之间的关系。

定义2：设 U、V 是两个论域，由 U、V 作出的一个新的论域 $U'V$。$U'V$ 上的任一个模糊集合 $R\in F(U'V)$ 都叫做 U 与 V 之间的模糊关系，即

$\mu_R:U\times V\to[0,1]$

$(u,v)\to\mu_R(u,v)$

其中 $\mu_R(u,v)$ 称为 u 与 v 关于 R 的关系强度。当 $U=V$ 时，称 R 为 U 上的模糊关系。

在现实世界中模糊关系的例子很多。例如人的收入水平和购买价格有一定的关系，一般而言，收入越高，其购买商品的价格越高。若用普通的关系表示，则只能得出一个稀疏矩阵如下：（假定收入水平分为A、B、C、D、E五类，而商品价格水平也分为1、2、3、4、5五类）。

表2

收入＼价格	1	2	3	4	5
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1

表2中横行表示商品价格水平，而竖行表示收入水平，从这种表中很难得出收入水平B与价格水平1的关系，如果采用模糊关系来表示，对商品价格水平与收入水平差一个档次的系数用0.8代替0，相差二个档次的用0.5来代替…则模糊关系矩阵如表3所示。

表3

收入＼价格	1	2	3	4	5
A	1	0.8	0.5	0.2	0
B	0.8	1	0.8	0.5	0.2
C	0.5	0.8	1	0.8	0.5
D	0.2	0.5	0.8	1	0.8
E	0	0.2	0.5	0.8	1

在表3中，两个集合之间的元素的模糊关系可用 $R_{n\times m}$ 的模糊矩阵表示，其中 r_{ij} 表示了 u_i 和 v_j 之间的模糊系数。

对于模糊矩阵，可以定义它的λ截关系。

定义3：设 $R\in F(U\times V)$，$\forall\lambda\in[0,1]$，称 R 的λ截关系为：

$\mu_{R_\lambda}v\Leftrightarrow C_{R_\lambda}(u,v)=1\Leftrightarrow\mu_R(u,v)\geq\lambda$

对于表3取λ=0.8的截关系矩阵为：

表4

收入＼价格	1	2	3	4	5
A	1	1	0	0	0
B	1	1	1	0	0
C	0	1	1	1	0
D	0	0	1	1	1
E	0	0	0	1	1

对应于不同的λ，可以得到不同的模糊关系矩阵，从而得到不同的模糊集合。

3 详细算法

对于给定的需要进行时间序列挖掘的数据集，首先进行预处理，即按照时间序列的主体形成各自的子序列。例如对于超市等销售型企业的数据而言，以顾客为主体，求出每个顾客的购买时间序列，并从中选出一些典型顾客的时间序列，作为下一步分析的候选数据；对于股票数据而言，每种股票都可以形成一个时间序列。

下一步是在预处理过的时间序列值中求出候选频繁集。不同的应用对象往往可以选出不同的候选频繁集。例如对于超市而言，候选频繁数据集可以是商品形成的序列，如顾客总是先购买A商品，再购买B商品，且有一定比例的顾客遵循着这样的9序列；对于股票等实时数据，候选频繁集可以是股票的价格的类似的波动状况。

直接从候选的频繁数据集中较难得出所需的频繁数据集，这是因为时间序列中的元素的相互关系较为复杂，例如购买A商品与购买B商品的时间间隔对于不同顾客而言往往是不同的，加上商品类别等其它因素的影响，所以不能像求关联规则那样得出精确的频繁数据集。这里采用模糊关系矩阵来表示候选的频繁数据集元素之间的关系，并由λ截关系确定出候选的频繁数据集。

对于某一候选的频繁数据集中的时间序列中的元素而言，只要求出它们之间的模糊系数，就可以表示出该时间序列元素之间的模糊关系矩阵。例如候选的频繁数据集中甲顾客依次购买了A、B、C、D四种商品，则对此序列可形成4*4的模糊关系矩阵，由于以时间为序列方向，所以此矩阵只要是三角阵就可以了，即只要求出A→B，A→C，A→D，B→C，B→D，C→D的模糊系数（此处先求二维的关联关系，多维的如AB→C，AB→CD等模式是在二维的基础上再求得的）。对于求A→B的模糊系数，有较多的方法，常用的有数量积法、相关系数法、算术平均最小法以及主观评定法等。上例中提到的顾客购买时间序列，

由于 A 和 B 两种商品之间的影响因素较多且不易量化，所以可以采用主观评定法与相关系数法相结合的办法。例如可以根据实际情况或商家的经验或评判感觉，将商品种类和时间间隔等因素编成数值表，数值越接近表明类型或时间等越接近。例如家电类数值为 1，文具类数值为 2，衣物类数值为 3；一月份购买的为 1，二月份购买的为 2…等等。对于 AB 商品，分别求出 A 与 B 的各个因素的对应数值，形成两个向量 X_i 与 X_j，再根据相关系数的公式：

$$r_{ij}=\frac{\sum_{k=2}^{m}(x_{ik}-\overline{x_i})(x_{jk}-\overline{x_j})}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\overline{x_i})^2}\sqrt{\sum_{k=1}^{m}(x_{jk}-\overline{x_j})^2}}$$

其中 r_{ij} 为相关系数，而 x_{ik} 与 x_{jk} 分别表示 Xi 与 Xj，两个向量的相应的第 k 个值。用此相关系数可以作为 A 和 B 的模糊系数，类似可以求得 A 和 C、B 和 C 等的模糊系数。由此形成模糊矩阵，例如上例中甲顾客的 ABCD 四种商品的可能的模糊矩阵为：

表 5

	A	B	C	D
A	1	0.7	0.5	0.5
B		1	0.4	0.6
C			1	0.2
D				1

若取 λ=0.5，则可以认为是序列的有 A→B，A→C，A→D 和 B→D。则统计频繁数据集时只有上述四种序列参加计数，而其余的如 B→C、C→D 等由于其模糊系数小于 λ，所以不能进入实际的频繁数据集统计。假设乙顾客也有 A→B 的购买序列，其模糊关系矩阵如表 6 所示：

虽然乙顾客也有先买 A 再买 B 的情况，但由于其模糊系数小于 0.5，可以认为通过模糊理论判断 A→B 的序列对乙不成立。则计算 A→B 的支持度时，只有甲顾客的序列参加计数，而乙的不参加。由于 λ 可以取不同的值，所以根据最初设定的不同的值，可以得到不同的模糊矩阵，最终得到的时间序列模式的支持度与置信度也是不同的。

表 6

	A	B	E
A	1	0.4	0.5
B		1	0.4
E			1

与用 RS 算法计算关联规则类似，当两两关系系数确定后，可以对多元关系确定模糊系数，从而得到多元的模糊矩阵，再根据具体的 λ 值得到相应的 λ 截矩阵最终得到多元的时间序列模式。唯一的区别是求多因素的模糊系数往往用到正则化向量的欧氏距离来计算。

4 工程实践

在国家科技部资助的中小型企业创新基金项目智能商务软件的工程开发中，该算法作为该项目中数据挖掘模块的一部分，用模糊数学的思想加上 RS 算法对建材类销售企业的数据仓库进行了时间序列模式的分析，取得了较理想的效果。通过对建材类企业的销售数据进行时间序列模式挖掘所得出的规则，对于该类企业的决策有着一定的参考价值。

（收稿日期：2001 年 9 月）

参考文献

1.孙志挥，肖利.知识发现与数据挖掘.东南大学研究生教材，2000
2.李洪兴等.工程模糊数学方法及应用[M].天津科学技术出版社，1991
3.张文修.模糊数学基础[M].西安交通大学出版社，1984

（上接 94 页）

准，凡符合标准的特征点删除，其余的给予保留。保留下来的特征点以链码方式记录它们之间的相对位置关系，用以与指纹库中的数据比对匹配。

3.3.3 指纹的比对

在进行指纹比对之前，一定要先建立指纹数据库。建立指纹数据库，一般要采集同一枚指纹的 3~5 个样本，分别对这些样本进行预处理和特征抽取，由特征点间的相互位置关系确定样本图象是否两两匹配，根据特征点被匹配上的次数，确定该特征点的匹配权值，从所有样本图象中找出权值大于一给定阈值的特征点，以这些特征为模板建立指纹数据库样本。对于待匹配的指纹图象，经预处理和特征提取后，形成一个坐标链码记录，根据这些特征的相互位置关系与指纹数据库中的的样本做图形匹配[3]，得到最终的识别结果。

3.4 指纹识别实验

利用硅晶体电容指纹传感器采集指纹 50 枚，笔者利用自己开发的自动指纹识别系统进行处理，形成指纹样本库，在对同样的指纹提供人采样 200 枚，作为待识别的指纹输入识别系统，得出匹配结果。错误识别的指纹数为 0，正确识别的指纹为 189。尽管拒识率达到了 5%，但因为误识率极低，充分证明了该系统的可靠性与安全性。

4 结束语

文章主要介绍了自动指纹识别系统中指纹提取和指纹识别的方法和手段。当前用于自动指纹提取的设备都还存在着多多少少的问题，指纹识别要走向市场，还需降低指纹提取设备的成本，提高指纹图象的质量。另外，为加快指纹识别的速度，还应该考虑简化图象的预处理，实现在灰度图上直接抽取指纹细节特征。随着指纹提取设备的小型化，指纹识别算法也要能够根据最少量的特征对指纹进行匹配，并能克服因指纹旋转等带来的偏差，增强算法的鲁棒性。（收稿日期：2001 年 9 月）

参考文献

1.Mehtre B M，Chatterjee B.Segmentation of fingerprint image-a composite method[J].Pattern Recognition，1989；22(4)：381~385
2.Hardic R C，Boncelet C A.Class of rank-order-based filters for smoothing and sharping[J].IEEE transaction on signal processing，1993；41(3)：1061~1075
3.Isenor D K，Znky S G.Fingerprint Identification Using Graph Matching[J].Pattern Recognition，1986；19(2)：113~122
4.刘少聪.新指纹学[M].合肥：安徽人民出版社，1984
5.徐建华.图象处理与分析[M].北京：科学出版社，1992
6.冯星奎等.方向加权中值滤波算法[J].中国图形图象学报，2000；5(7)：609~611
7.李晓昆.基于结构特征的指纹识别[J].计算机工程与科学，1999；21(2)：25~29
8.Veridicom 公司.FPS110 指纹传感器说明书.America，1999