

基于时间的模糊关联规则挖掘

张 诚¹, 郑 诚²

(1. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039;

2. 安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘 要:关联规则是数据挖掘研究中的一个重要的主题。一些算法都是假设数据中根本的关联基于时间是稳定的。然而,在现实世界领域,数据具有自己的特征,因此关联随着时间发生巨大的改变。现有的数据挖掘算法没有考虑关联的改变,这导致了严重的性能下降,特别是挖掘出的关联规则被用来分类和预测。尽管关联改变的挖掘是一个重要的问题,因为需要基于过去的历史数据来预测未来,现有的数据挖掘算法不符合这样的工作。文中引入模糊数据挖掘算法来发现基于时间的关联规则的改变。基于挖掘出的模糊规则,能预测关联规则在未来如何改变。实验表明了算法的有效性。

关键词:数据挖掘;时间序列;模糊关联规则

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2007)07-0060-03

Fuzzy Association Rules Mining over Time

ZHANG Cheng¹, ZHENG Cheng²

(1. Ministry of Edu. Key Lab. of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China;

2. School of Computer Science & Technology, Anhui University, Hefei 230039, China)

Abstract: Association rule mining is an important topic in data mining research. Many algorithms have been developed for such task and they typically assume that the underlying associations hidden in the data are stable over time. However, in real world domains, it is possible that the data characteristics and hence the associations change significantly over time. Existing data mining algorithms have not taken the changes in associations into consideration and this can result in severe degradation of performance, especially when the discovered association rules are used for classification (prediction). Although the mining of changes in association is an important problem because it is common that we need to predict the future based on the historical data in the past, existing data mining algorithms are not developed for this task. In this paper, introduce a new fuzzy data mining technique to discover changes in association rules over time.

Key words: data mining; time series; fuzzy association rules

0 引 言

关联规则挖掘是用于发现数据库中不同属性间有趣的关联关系。现有的数据挖掘算法没有考虑关联的改变,这导致了严重的性能下降。然而,在现实世界领域,数据具有自己的特征,因此关联随着时间发生巨大的改变。举个例子,在时间 t_1 ,可能有一个有趣的关联在2个属性间,但在时间 t_2 关联可能就不再有趣了。文中引入模糊数据挖掘算法来发现基于时间的关联规则的改变。基于挖掘出的模糊规则,能预测关联规则在未来如何改变^[1,2]。实验表明了算法的有效性。

1 时间序列模糊离散化处理

效仿文献[3,4]的方法,首先对时间序列进行离散化处理,但这里并非将时间序列形态进行确定性归类,而是能将每一个局部序列软化到代表形态中。

设 $s = (X_1, X_2, \dots, X_n)$ 为一时间序列,将一宽度为 w 的时间窗作用于 s 形成一长度为 w 的子序列 $S_i = (X_i, X_{i+1}, \dots, X_{i+w-1})$,将时间窗在时间序列 s 上从起点至终点进行单步滑移,形成一系列宽度为 w 的子序列 $X_1, X_2, \dots, X_{N-w+1}$, 记

$$W(s, w) = \{s_i \mid i = 1, 2, \dots, N - w + 1\} \quad (1)$$

为由该时间序列 s 用宽度为 w 的滑窗滑移出的子序列集合。

(1) 将 $W(s, w)$ 看作为 w 维欧氏空间中的 $(N - w + 1)$ 个点,并将它们随机地分到 k 类中,计算每类中心。第 j 类中心第 l 坐标值为:

收稿日期:2006-09-18

基金项目:安徽省高校自然科学基金项目(2006KJ055B)

作者简介:张 诚(1980-),男,安徽潜山人,硕士研究生,研究方向为网络下数据库与数据挖掘;郑 诚,博士,副教授,研究方向为网络下数据库与数据挖掘。

$$X_{j,l} = 1/h \sum_{i=1}^k X_{j,l,i}, l = 1, 2, \dots, w; j = 1, 2, \dots, k \quad (2)$$

其中, h 表示第 j 类中的子序列数目, $X_{j,l,i}$ 表示第 j 类中第 i 个子序列第 l 坐标值。

(2) 以这些中心作为每类的代表点, 计算集合 $W(s, w)$ 中每个元素 $s_i, i = 1, 2, \dots, (N - w + 1)$ 属 j 类代表点的隶属度函数 $u_j(s_i)$:

$$U_j(s_i) = (1 / |s_i - x_j|^2)^{1/(b-1)} / \sum_{j=1}^k (1 / |s_i - x_j|^2)^{1/(b-1)}, j = 1, 2, \dots, k; b > 1 \quad (3)$$

其中 $b > 1$ 是一个可以控制聚类结果的模糊程度的常数, $|s_i - x_j|^2$ 表示每一点到第 j 类代表点距离的平方。

(3) 用当前的隶属度函数更新计算各类中心:

$$X_{j,l} = \sum [u_j(s_i)]^b x_{j,l,i} / \sum [u_j(s_i)]^b \quad j = 1, 2, \dots, k; l = 1, 2, \dots, w \quad (4)$$

重复以上(2)、(3)步的计算, 直到各个样本的隶属度稳定。并且将代表集合记作 $D = \{x_1, x_2, \dots, x_k\}$, 其中 x_j 表示第 j 个代表点。

2 模糊关联规则的挖掘

经过模糊离散化处理后, 得到 k 个代表点和各个子序列到每个代表点的隶属度。每个代表点也就是每个代表形态, 并且每个子序列到各个代表形态的隶属度之和为 1, $\sum [u_j(s_i)] = 1$ 。定义模糊关联规则的形式为: “如果 A 发生, 那么在时间 T 内 B 发生”, 记作 $A \Rightarrow B, A, B \in \{x_1, x_2, \dots, x_k\}$ 。形态 A 发生的频数 $F(A)$ 定义为: $F(A) = \sum u_A(s_i)$, 其中, $u_A(s_i)$ 为 s_i 点属第 A 个代表形态的隶属度。

模糊规则 $A \Rightarrow B$ 的可信度为:

$$C(A \Rightarrow B) = F(A, B, T) / F(A)$$

其中 $F(A, B, T)$ 为形态 A 发生后, 紧跟着在 T 时间内 B 发生的频数:

$$F(A, B, T) = | \{ i (s_i = A) \wedge (B \in \{s_{i+w+1}, s_{i+w+2}, \dots, s_{i+w+T-1}\}) \} | = \sum u_A(s_i) \cdot [u_B(s_{i+w+1}) \vee u_B(s_{i+w+2}) \vee \dots \vee u_B(s_{i+w+T-1})]$$

3 有效规则的选择

经过以上的处理后, 得到大量具有不同可信度的规则。为了选择最有价值的规则, 用 Smyth^[5,6] 等人提出的 J -measure 方法对所得规则的有效性进行排序。规则 $A \Rightarrow B$ 的 J -measure 定义为: $J(Br:A) = p(A) \cdot \{p(B_T | A) \log(p(B_T | A) / p(B_T)) + [1 - p(B_T | A)] \log[1 - p(B_T | A) / (1 - p(B_T))]\}$, 其中,

$p(A)$ 表示第 A 种形态出现的频率, 也就是形态 A 发生的频数 $F(A)$ 和总子序列数之比, $p(A) = F(A) / (N - w + 1)$ 是任意时间窗之后在时间 T 内 B 发生的频率:

$$P(Br) = \sum [u_B(s_{i+w+1}) \vee u_B(s_{i+w+2}) \vee \dots \vee u_B(s_{i+w+T-1})] / (N - w + 1)$$

即先验概率 $p(B_T | A)$ 是 A 形态出现之后在时间 T 内 B 形态发生的频率, 亦即模糊规则 $A \Rightarrow B$ 的可信度, 即后验概率。直观来看, 公式右边的第一部分 $p(A)$ 是希望这种形态出现的次数更多一些, 公式右边的第二部分是熵, 表示从先验概率 $p(Br)$ 到后验概率 $p(B_T | A)$ 的信息获得。

4 多维时间序列模糊关联规则的挖掘

在一维时间序列模糊关联规则的基础上, 进一步研究多维时间序列模糊关联规则的挖掘。对于 m 维时间序列, 经过滑窗处理后得到子序列集合:

$$W(S, w) = \{s_i^h | i = 1, 2, \dots, (N - w + 1); h = 1, 2, \dots, m\}$$

对 $W(S, w)$ 中的 m 个子集:

$$\{s_i^h | i = 1, 2, \dots, (N - w + 1)\}, h = 1, 2, \dots, m$$

均用模糊离散化方法处理后, 则每个子集均得到 k 个代表点和该集中各个子序列到该集各个代表点的隶属度。每个子集的代表点集合记作 $Dh = [x_1^h, x_2^h, \dots, x_k^h], h = 1, 2, \dots, m$, 其中 x_j^h 表示第 h 个子集的第 j 个代表点。定义模糊关联规则的形式为: “如果在 V 时间内 A^1 和 $A^2 \dots$ 和 A^p 和 \dots 和 A^h 发生, 那么在时间 T 内 B 发生”, 记作 $A^1 \wedge A^2 \dots \wedge A^p \wedge \dots \wedge A^h \Rightarrow B$ 。其中 $A^p \in D^p, D^p \in \{D^1, D^2, \dots, D^m\}, p = 1, 2, \dots, h, B \in D', D' \in (D^1, D^2, \dots, D^m)$, 并且对于 $i, j = 1, 2, \dots, h$ 和 $i \neq j$, 有 $D^i \cap D^j = \emptyset$ 。 V 时间 A^1 和 A^2 和 \dots 和 A^h 发生的频数定义为:

$$F(A^1 \wedge \dots \wedge A^h, V) = | \{ i A^1 \in \{a_i^1, a_{i+1}^1, \dots, a_{i+w-1}^1\} \wedge \dots \wedge A^h \in \{a_i^h, a_{i+1}^h, \dots, a_{i+w-1}^h\} \} | = \sum \{ [u_{A^1}(s_i^1) \vee \dots \vee u_{A^h}(s_{i+w-1}^h)] \dots [u_{A^h}(s_i^h) \vee \dots \vee u_{A^h}(s_{i+w-1}^h)] \}$$

模糊规则 $A^1 \wedge A^2 \wedge \dots \wedge A^p \wedge \dots \wedge A^h \Rightarrow B$ 的可信度为:

$$C\{A^1 \wedge A^2 \wedge \dots \wedge A^p \wedge \dots \wedge A^h \Rightarrow B\} = F(A^1 \wedge A^2 \wedge \dots \wedge A^p \wedge \dots \wedge A^h, V; B, T) / F(A^1 \wedge A^2 \wedge \dots \wedge A^p \wedge \dots \wedge A^h, V)$$

其中 $F(A^1 \wedge A^2 \wedge \dots \wedge A^p \wedge \dots \wedge A^h, V; B, T)$ 为在 V 时间内 A^1 和 $A^2 \dots$ 和 A^p 和 \dots 和 A^h 发送后, 紧跟

着在 T 时间内 B 发生的频数:

$$F(A^1 \wedge A^2 \wedge \cdots \wedge A^p \wedge \cdots \wedge A^h, V; B, T) = |\{i \mid A^1 \in \{a_i^1, a_{i+1}^1, \dots, a_{i+w-1}^1\} \wedge \cdots \wedge A^h \in \{a_i^h, a_{i+1}^h, \dots, a_{i+w-1}^h\} \wedge B \in \{a_{i+w}^r, \dots, a_{i+w+r-1}^r\}\}| = \sum \{[u_A(s_i^1) \vee \cdots \vee u_A(s_{i+w-1}^1)] \cdots [u_{A_h}(s_i^h) \vee \cdots \vee u_{A_h}(s_{i+w-1}^h)] \cdot [u_B(s_{i+w}) \vee \cdots \vee u_B(s_{i+w+r-1})]\}$$

多维情况出现的问题是潜在规则成级数增长,为此需采用 Agrawal^[7]等人提出的 Apriori 算法首先对非频繁规则进行删剪。

5 参数问题

以上的模糊关联规则挖掘是在滑窗参数 w 和离散类数 k 固定的条件下进行的, w 和 k 的大小直接影响到最终结果,我们的目标是发现有意义的关联规则。一般情况是,研究时间序列的短形态关系时, w 值应取的较短;研究长形态关系时, w 值应取得长一些。分得类数太多,每一类中的子序列数目太少,不利于计算可信度;分得类数太少,每一类中的子序列的形态相差太远,类中心的代表性太差。一种简单的方法是不考虑参数选取合适与否,而是选取不同 w 和 k 值,一个好的参数应是能够取得具有意义的形态之间的关系。

6 实验

选用香港资产评估公司的资产数据库。资产数据库是从数据仓库中提取出来的,它包含了 1991~2000 年间的买卖交易细节。数据库包含 11176 条记录。每条记录代表一套公寓的买卖交易,每条记录有 11 个属性(如表 1 所示)。

表 1 数据库资产属性

Attribute	Description
TRAND	交易日期
PHASE	公寓所在方位
BLOCK_NO	街区数
FLOOR_NO	层数
DIRET	方向
DERV_SIZE	公寓面积
BUILD_AGE	建筑年代
BW	凸窗大小
BEDRM	卧室数量
LIVRM	起居室数量
PRICE	公寓价格

把资产数据库分为 10 个部分, D_1, \dots, D_{10} (D_1 包含 1991 年的交易)。然后把量化属性域分为 5 个区间,使用 Apriori 算法从最初的 9 个数据库部分 D_1, \dots, D_9 挖掘出 9 个关联规则集 R_1, \dots, R_9 。应用模糊数据挖掘算法发现模糊规则集来表示每个在 $R(R_1 \cup \dots \cup R_9)$ 中的关联规则的支持度和置信度改变的规律性。

利用挖掘出的模糊规则,预测在 R 中每个关联规则的支持度和置信度如何改变。这导致了一个关联规则集 R'_{10} , 每个在 R'_{10} 中的规则的支持度和置信度被预测基于 1991~1999 年被挖掘出的关联规则的改变。使用 CBA^[8] 来预测 D_{10} 中每条记录的 PRICE 的值。给定 D_{10} 中的一条记录, CBA 可以把它归入一个区间,这个区间的中点被认为是 PRICE 的值。

在实验中,使用出错率来作为性能的度量。设 N 为 D_{10} 中的记录数,任意 $i \in D_{10}$, t_i 是 PRICE 的目标值, o_i 是 CBA 的预测值。出错率被定义为 $\text{error} = \sum |t_i - o_i| / o_i \mid / N$ 。为了进一步评价方法的性能,使用 D_9 中被挖掘出的关联规则来预测 D_{10} 中记录 PRICE 的值。把 D_{10} 分成 2 个数据集(训练集和测试集),训练集包含 80% 的记录。从训练集中挖掘出一个关联规则集 R_{10} , 使用这些规则来预测测试集记录中 PRICE 的值。这一步骤重复 10 次。由于训练集和测试集是从 D_{10} 中随意取出,所以预测是理想的。实验中最小支持度阈值 1%, 最小置信度 50%, 滑窗参数 w 为 5。

实验结果如表 2 所示。基于资产数据库的实验结果。

表 2 资产数据库实验结果

Rule Set	Percentage Error
R_0	16.7%
R'_{10}	15.1%
R_{10}	14.2%

7 结 语

现有的数据挖掘算法没有考虑关联的改变,这导致了严重的性能下降,文中引入模糊关联规则挖掘方法来发现基于时间的关联规则的改变。基于挖掘出的模糊规则,能预测关联规则在未来如何改变。实验证明了方法的可行性和优越性。

参考文献:

- [1] Weigend A S, Gershenfeld N A. Time Series Prediction: Forecasting the Future and Understanding the Past[M]. St Louis, Missouri: Addison Wesley Longman, 1994.
- [2] Povinelli R, Feng X. Temporal Patterns Identification of Time Series Data Using Pattern Wavelets and Genetic Algorithms [C]//In: Proceedings of Artificial Neural Networks in Engineering. St Louis, Missouri: [s. n.], 1998: 691-696.
- [3] Guralink V, Srivastava J. Event Detection from Time Series Data [C]//In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. San Diego, CA, USA: [s. n.], 1999: 33-42.

果表明,用文中算法去噪后的语音,其语谱图能清楚地看到浊音共振峰时变过程,特别是时间分辨率高。

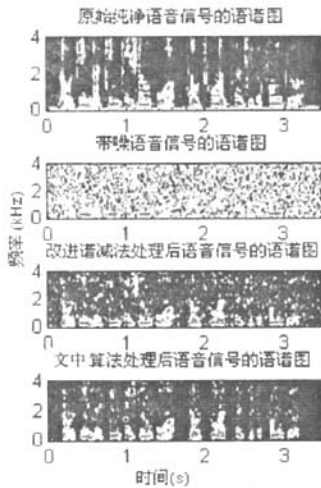


图 5 输入 SNR = 5dB 的实验结果

语音去噪效果的客观评价是以语音系统的输入信号和输出信号之间的误差大小来判别语音质量的好坏。信噪比(SNR)^[9]是衡量针对宽带噪声失真的语音去噪效果的常规方法。假设 $y(n)$ 为带噪语音信号, $s(n)$ 为其中的纯净语音信号, $\hat{s}(n)$ 为去噪后的语音信号, L 为语音信号的长度,则输入信噪比为:

$$SNR_{in} = 10 \lg \frac{\sum_{n=0}^L s^2(n)}{\sum_{n=0}^L [y(n) - s(n)]^2} \quad (13)$$

输出信噪比为:

$$SNR_{out} = 10 \lg \frac{\sum_{n=0}^L s^2(n)}{\sum_{n=0}^L [\hat{s}(n) - s(n)]^2} \quad (14)$$

表 1 语音去噪效果实验结果比较

输入信噪比 SNR _{in} (dB)	-5	0	5
改进谱减法 SNR _{out}	1.5480	3.8227	8.4300
文中方法 SNR _{out}	5.2653	7.6118	10.8180

表 1 给出了在语音信号输入信噪比分别为 -5dB,

0dB 和 5dB 的情况下,由文中方法得出的实验结果输出信噪比的比较。

4 结 论

与直接采用改进的谱减法进行语音去噪相比,文中提出的语音去噪方法使得噪声得到很明显的抑制,提高了语音质量。实验表明,特别是在输入信噪比较低的情况下,采用文中算法,更好地提高了带噪语音的输出信噪比,明显削弱了谱减法所带来的“音乐”噪声,抑制加性噪声效果最好、运算量小、系统简单。但语音的可懂度却受到一定的损失,略微存在失真,针对这个不足还在不断地研究中。

参考文献:

- [1] 杨行峻,迟惠生. 语音信号与数字处理[M]. 北京:电子工业出版社,1995:391-400.
- [2] 曹晓琳,张素莉,吴平,等. 基于 MATLAB 的谱相减语音增强算法的研究[J]. 计算机仿真,2006,23(3):278-283.
- [3] 易克出,田斌,付强. 语音信号处理[M]. 北京:国防工业出版社,2000.
- [4] 金学骥. 语音增强算法的研究与实现[D]. 杭州:浙江大学,2005:12-13.
- [5] Preuss R D. A Frequency Domain Noise Canceling Preprocessor for Narrowband Speech Communications Systems [C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. [s. l.]: [s. n.], 1979:212-215.
- [6] Berouti M. Enhancement of Speech Corrupted by Acoustic Noise [C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. [s. l.]: [s. n.], 1979:208-211.
- [7] Soo Kim Nam, Chang Joon-Hyuk. Voice activity detection based on complex laplacian model[J]. Electronics Letter, 2003 (7):632-633.
- [8] 李富强,万红,黄俊杰. 基于 MATLAB 的语谱图显示与分析[J]. 微计算机信息,2005,21(30):172-174.
- [9] 王振力,张雄伟,郑翔,等. 一种新的子波域语音增强方法[J]. 信号处理,2006,22(3):327-328.

(上接第 62 页)

- [4] Das G, Lin K, Mannila H, et al. Rule Discovery from Time Series [C]//In: Proceeding of the 3rd International Conference of Knowledge Discovery and Data Mining. Gregory P, California: [s. n.], 1998:16-22.
- [5] Mark L, Klein Y, Kandel A. Knowledge Discovery in Time Series Databases [C]//In: IEEE Transaction on Systems, Man, and Cybernetics - Part B: Cybernetics. Gregory P, California: [s. n.], 2001:20-30.
- [6] Smyth P, Goodman R M. Rule Induction Using Information

Theory [C]//In: Gregory P, William J. Knowledge Discovery in Databases. Cambridge: MIT Press, 1991:159-176.

- [7] Agrawal R, Mannila H, Srikant R, et al. Fast Discover of Association Rules [C]//In: Fayyad M, Piatetsky-Shapiro G, Smyth P. Advance in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI/MIT Press, 1996:307-328.
- [8] Liu B, Hsu W, Ma Y. Integrating Classification and Association Rule Mining [C]//in Proc. of the 4th Int'l Conf. on Knowledge Discovery and Data Mining. New York, NY: [s. n.], 1998.