基于动态相关性挖掘的信息融合方法

徐凌宇, 张德干, 赵 海
(东北大学信息科学与工程学院, 辽宁沈阳 110006)

摘 要: 本文方法通过学习诸多信源在一定时间段内的变化过程, 挖掘出那些与结论相关的信源及与结论相关时间片段. 形成最终的决策树模式. 该方法适用于大规模多因素动态隐含相关性信息融合.

关键词: 信息融合; 示例学习; 决策树; 挖掘; 动态相关
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2002) 02-0292-03

A Method of Data Fusion Based on Dynamic Association Mining

XU Ling-Yu, ZHANG De-Gan, ZHAO Hai
(School of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110006, China)

Abstract: The relative sources and the time period can be defined by learning all features changes in a period. The set of examples is groups of time series of source data. A decision tree is constructed by the information gain. After the coherence estimation, the same as time series, the irrelative sources are washed out and relative sources are selected to be applied for fusion. The method fit the large scale fusion of dynamic hidden.

Key words: data fusion; learning from examples; decision tree; mining; dynamic association

1 引言

多元动态相关问题以其非线性、非确定性和模糊性而很难用传统数学模型描述. 许多领域的决策依据诸多因素在一定时间段内的变化趋势来确定. 具有学习能力的相关性挖掘方法成为热点, 动态问题表现为诸多信源的连续变化过程, 关键在于如何从诸多因素中筛选出与结论相关的信源子集和相关的时间段进行融合, 淘汰无关因素, 以降低问题的规模与复杂度及减少噪音干扰, 决策树方法是数据挖掘和融合的工具之一, 许多学者对原有 ID3 算法进行了改进[1~5], 以增强其对相关性的关注. 其中文献[3~5]对相关信源子集的选取方法进行了深入地研究, 以处理空间相关性问题. 此外, 很多领域的挖掘考虑到时间相关问题[6~8]及其它相关考虑[9,10]. 本文特点是对现有 ID3 方法加以改进, 使之能够描述动态相关, 给出 FDAM(Fusion Based On Dynamic Association Mining)方法, 将动态相关引入决策树——即认为多元动态问题是时间与空间的融合, 采集多信源的时间序列进行融合, 根据一致性评价结果, 筛选出相关规则, 淘汰虚相关. 使其具有处理多元动态问题行为的能力.

2 定义

2.1 定义示例信息

定义 1 设 m 个过程构成的初始集 $S=\{\boldsymbol{P}_i \mid 1\leqslant i\leqslant m\}$, 其中 $m\geqslant 1$, 当 $m=0$ 时 $S=\Phi$. 矢量 $\boldsymbol{P}_i$ 为过程 i 的时间序列, 有 $n(i)$ 个时刻, $\boldsymbol{P}_i=(X_{i1}, X_{i2}, \cdots, X_{in(i)})$, X_{ij} 为第 i 个过程的第 j 个初始样本, 设由 o 个信源构成, 分为 g 个类, $c_{ij}\in(U_1, U_2, \cdots, U_g)$, 则 $X_{ij}=(x_{ij1}, x_{ij1}, \cdots, x_{ijo}, c_{ij})$, 则 X_{ij} 为第 i 过程的第 j 个数据时间切片(初始样本). 初始集样本总数

$$|S|=\sum_{i=1}^{m} n(i).$$

定义 2 时间序列的长度为由初始时刻到终结时刻包含的所有数据时间切片数. 记为 $len(\ )$.

定义 3 设 $X'_{ij}=(x_{ij1}, x_{ij1}, \cdots, x_{ijo})$, 一个长度为 $\mathrm{len}(E)$ 序列 $E_{ij}=(X'_{ij}, X'_{ij+1}, \cdots X'_{ij+\mathrm{len}(E)-1}, c_{ij+\mathrm{len}(E)-1})$ 为第 i 个过程的第 j 个重构示例.

定义 4 重构示例集 $S'=\{E_{ij} \mid 1\leqslant i\leqslant m, 1\leqslant j\leqslant n(i)-\mathrm{len}(E)\}$, 示例总数 $|S'|=\sum_{i=1}^{m}(n(i)-\mathrm{len}(E)+1)$; 若任取 $i(1\leqslant i\leqslant m)$, 有 $n(i)\leqslant \mathrm{len}(E)$, 则 $S'=\Phi$, $|S'|=0$.

2.2 信源信息分解

定义 5 示例信息二元化, 设数据时间切片有 o 项因素(信源), 第 k 项信源($1\leqslant k\leqslant o$)有 $r(k)$ 个取值, 则将该信源信息分解为 $r(k)$ 个特征, 分解后的每个特征取值为 0、1, 则长度为 $\mathrm{len}(E)$ 重构示例二元化分解后的特征数为 $T=\sum_{k=1}^{o} r(k)*\mathrm{len}(E)$: 重构示例 E 可用二值特征示例 E' 描述: $E'=(u_{111}, u_{112}, \cdots u_{\mathrm{len}(E)or(o)}, c)$, 其中 u_{jkh}. 为第 j 个时刻地 k 个信源的第 $r(k)$ 个特征, 有

$$u_{jkh}.=0 \mid 1 (1\leqslant j\leqslant \mathrm{len}(E), 1\leqslant k\leqslant o, 1\leqslant h\leqslant r(k)).$$

收稿日期: 2001-02-19; 修回日期: 2001-07-10
基金项目: 国家自然科学基金(No. 69873007)

2.3 定义相关

定义 **6**　若问题的行为与 $o(o>1)$ 个信源的 len( ) (len( ) >1) 个数据时间切片相关, 则该问题具有时间空间相关特性.

定义 **7**　集合 S' 中以因素 a 为判据的重构示例子集为 S'_a, 则 $W_a=|S'_a|/|S'|$ 为因素 a 的相关强度, 相应地, 时刻 j 的相关强度记为 W_j^t, 信源 k 的相关强度记为 W_k^s 为,

定义 **8**　设阈值 $\theta(1\leqslant\theta\leqslant1)$, 若相关强度 $W_a\leqslant\theta$, 则 a 为虚相关因素, 相应地, 若 $W_j^t\leqslant\theta^t$, j 为虚相关时刻, $W_k^s\leqslant\theta^s$, k 为虚相关信源.

3　互信息的计算

根据定义 1, 对于 U_b 发生可能性有:

$$P(U_b)=|U_b|/|S| \tag{1}$$

且 $\sum_{b=1}^{g}P(U_b)=1$

U_b 类中在特征, $u_i(1\leqslant i\leqslant T)$ 处取值 V_d 的示例子集合 S_{bd} 的条件概率为:

$$P(V_d|U_b)=|X_{bd}|/|U_b| \tag{2}$$

且 $\sum_{d=1}^{2}P(V_d|U_b)=1$

在特征 $u_i(1\leqslant i\leqslant T)$ 处, 取值 V_d 的示例集合 F_d 的概率为:

$$P(V_d)=|F_d|/|S| \tag{3}$$

且 $\sum_{d=1}^{2}P(V_d)=1$

在特征 $u_i(1\leqslant i\leqslant T)$ 处, 取 V_d 值的示例属于 U_b 类的示例集合 F_{bd} 的概率为:

$$P(U_b|V_d)=|F_{bd}|/|F_d| \tag{4}$$

且 $\sum_{d=1}^{2}P(U_b|V_d)=1$

消息 $U_b(b=1,2,\cdots g)$ 的发生概率 $P(U_i)$ 组成的信源数学模型为:

$$[U,P]=\begin{bmatrix}U_1, U_2, \cdots, U_g\\ P(U_1), P(U_2), \cdots, P(U_g)\end{bmatrix} \tag{5}$$

对应地, 消息 U_i 的自信息为:

$$I(U_b)=\log 1/P(U_b) \tag{6}$$

根据信息熵定义及公式(6)推出:

$$H(U)=\sum_{b=1}^{g}P(U_b)I(U_b)=\sum_{b=1}^{g}P(U_b)\log 1/P(U_b) \tag{7}$$

根据后验熵定义及公式(7), 特征 $u_i(1\leqslant i\leqslant T)$ 取值 V_d 的后验熵为:

$$H_i(U|V_d)=\sum_{b=1}^{g}P(U_b|V_d)\log 1/P(U_b|V_d) \tag{8}$$

根据公式(8)得对应条件熵为:

$$H_i(U|V)=\sum_{d=1}^{2}P(V_d)\sum_{b=1}^{g}P(U_b|V_d)\log\frac{1}{P(U_b|V_d)} \tag{9}$$

根据互信息定义及公式(7), (9)得特征 $u_i(1\leqslant i\leqslant T)$ 平均互信息为:

$$I_i(U|V)=H(U)-H_i(U|V)=\sum_{b=1}^{g}P(U_b)\log\frac{1}{P(U_b)}$$

$$-\sum_{b=1}^{g}\sum_{d=1}^{2}P(V_d)P(U_b|V_d)\log\frac{1}{P(U_b|V_d)} \tag{10}$$

4　挖掘算法.

4.1　学习算法

(1)示例集 $S'=\{S0+S1\}$, 其中, $S0$ 为基本集, $S1$ 为测试集.

(2)选取 $S0$, 调用建树算法

(3)若 $S1\neq\{\Phi\}$, 任取 $E'_i\in S1$, $E'_i=(u_1, u_2, \cdots u_T, c_i)$, 有: $S1=S1-E'_i$, 按(2)中决策树判定为 c'_i 类, 若 $c_i=c'_i$ 则返回(3); 若 $c_i\neq c'_i$ 则 $S0=S0+E'_i$ 返回(2)

4.2　决策树算法

(1)对当前子集中所有特征 $u_i(1\leqslant i\leqslant T)$, 根据公式(7)、(9)、(10)求出互信息集 $I=\{I_i(U|V)|1\leqslant i\leqslant T\}$, 取 $I_{max}=\max(I)$ 对应的特征记为 u_{max}.

(2)生成两个子集:

$W_P=\{E'_i|E'_i\in S1 \text{ AND } u_i=VP\}$, $P=1,2$

若 E'_i、$E'_j\in W_P$, E'_i 为 c_i 类, E'_j 为 c_j 类, 且 $c_i\neq c_j$, 返回(1); 若 $\forall E'_I$、$E'_j\in W_P$, E'_i 为 c_i 类, E'_j 为 c_j 类, 有 $c_i=c_j$, 则对应分支为 c_j 类, 返回调用处.

4.3　基于阈值的一致估计的相关规则筛选算法

(1)相关时序长和相关信源筛选, 建立时刻划分 $S^t=\{S_1^t, S_2^t, \cdots S_{\text{len}(E)}^t\}$, 建立信源划分 $S^s=\{S_1^s, S_2^s, \cdots S_o^s\}$, 初始状态: $S_1^t\cup S_2^t\cup\cdots S_{\text{len}(E)}^t\cup S_1^s\cup S_2^s\cup\cdots S_o^s=\Phi$, 从 S' 中取 E', $S'=S'-\{E'\}$, 将示例 E' 用决策树匹配, 匹配过程中, 若当前匹配结点(特征)为 u_{jkh} 则相应地 j 时刻相关集 $S_j^t=S_j^t\cup\{E'\}$, 相应地 k 信源相关集 $S_k^s=S_k^s\cup\{E'\}$, 若 $j(1\leqslant j\leqslant\text{len}(E))$, 且 $S_j^t=\{\Phi\}$, 则 j 为判别无关时刻, 若 $k(1\leqslant k\leqslant o)$, 且 $S_k^s=\{\Phi\}$, 则 k 为判别无关信源, 筛选后 $S^t=\{S_j^t|S_j^t\neq\{\Phi\}\text{ AND }(1\leqslant j\leqslant\text{len}(E))\}$, $S^t=\{S_k^t|S_k^t\neq\{\Phi\}\text{ AND }(1\leqslant k\leqslant o)\}$.

(2)淘汰虚相关　设虚相关时刻集 $J=\{\Phi\}$, 虚相关信源集 $K=\{\Phi\}$, 以问题的精度要求和数据质量作为阈值的选取原则, 设时间阈值为 θ^t, 则 $\forall S_j^t\in S^t$, 若 $|S_j^t|/|S^t|\leqslant\theta^t$, 则判定 j 时刻虚相关时刻, $S^t=S^t-S_j^t$, $J=J+S_j^t$; 设空间阈值为 θ^s, 则 $S_k^s\in S^s$, 若 $|S_k^s|/|S^t|\leqslant\theta^s$, 则判定 k 信源为虚相关信源, $S^t=S^t-S_j^t$, $K=K+S_k^t$. 重复(2), 直到无虚相关因素

(3)剪枝　先序遍历二叉树, 设当前访问结点特征为 u_{jkh}, 若 $j\in J$ OR $k\in K$ 则该结点为虚判定结点, 删除以该结点为根的子树. 选择该子树上数目最多的叶子类属代替该子树, 对于不确定性问题可按人类专家经验解决.

5　FDAM 与 C4.5 实验对照

5.1　实验对照

本文融合方法用于对小丰满水电站 4# 主变压器运行状态的综合评价. 选初始示例集 $S=1500$ 例, 分别训练成 FDAM 与 ID3 两棵决策树. 对丰满电站运行仿真机生成 300 例分别进行故障预测情况对照见表 1.

5.2　性能评价

(1) 两种方法对于处于正常状态的比较规律的数据均能给出正确判断.

表 1 实验效果对照表

	工况	类	示例	对判	错判	正确率%
4# 主变	空载 100	C_1	36	32	4	88.94
		C_2	64	57	7	89.0
	带载 200	C_1	62	53	9	85.5
		C_2	138	107	31	78.1
10# 主变	空载 100	C_1	36	29	7	80.5
		C_2	64	51	13	79.6
	带载 200	C_1	62	41	21	66.1
		C_2	138	78	60	56.5

(2) 表 2 中两种方法均能发现参数超标的危险和故障, 对于隐含于数据内动态相关 FDAM 比 C4.5 更敏感.

(3) 与 C4.5 相比, FDAM 具有动态过程的融合能力.

(4) 与 C4.5 相比, FDAM 经过一致性评价时空筛选, 增加了对有用信源和序列长的提取能力, 减少不相关信源和不相关时刻信息的干扰, 取得有意义结果.

6 结论

多元动态问题的特点是问题的行为不能仅由信源信息值确定, 常常与变化趋势相关. 本文基于时间空间数据思想能够确切描述此类问题的行为; 通过示例重构的方法生成时间序列, 初始样本不必考虑时间相关问题; 有时动态系统融合考虑的因素(信源)的多少以及时间序列长度不宜预先选定, 需进行筛选淘汰虚相关, 剪枝后的决策树, 能够在诸多信息中自动选取相关信源和时间序列长, 提高了融合对有用信息的提取能力.

参考文献:

[1] Roiger R J, Azarbod C, Sant R R. A majority rules approach to data mining [A]. In: Proc of Intelligent Information Systems IIS′97 [C], 1997: 100- 107.

[2] Marsala C, Bouchorr Meunier B. An adaptable system to construct fuzzy decision trees [A]. In: Proc of 18th International Conference on North American Fuzzy Information, 1999 [C], 1999: 223- 227.

[3] Piramuthu S. Evaluating feature selection methods for learning in data mining applications [A]. In: Proc of the Thirty-First Hawaii International Conference on System Sciences [C], 1998, 5: 294- 301.

[4] Wang X Z, Tsang E C C, Yeung D S. A problem of selecting optimal subset of fuzzy-valued features. Systems [A]. In: Proc of Man, and Cybernetics, IEEE SMC′99 [C], 1999, 3: 361- 366.

[5] Skrypnik I, Terziyan V, Puuronen S, Tsymbal A. Learning feature selection for medical databases [A]. In Proc of 12th IEEE-Conference on Computer-Based Medical Systems [C], 1999: 53- 58.

[6] 周斌, 吴泉源. 序列模式挖掘的一种渐进算法 [J]. 计算机学报, 1999, 22(8): 883- 887.

[7] 唐常杰, 于中华, 游志胜等. 基于时态数据库的 Web 数据周期规律的采掘 [J]. 计算机学报, 2000, 23(1): 52- 59.

[8] 欧阳为民, 菜庆生. 在数据库中发现具有时态约束的关联规则 [J]. 软件学报, 1999, 10(5): 527- 532.

[9] 何友, 陆大 , 彭应宁. 多传感器数据融合系统中两种新的航迹相关算法 [J]. 电子学报, 1997, 25(9): 9- 13.

[10] 程洪伟, 周一宇, 孙仲康. 多目标关联中的多特征数据融合方法 [J]. 电子学报, 1999, 27(3): 136- 139.

作者简介:

徐凌宇 男. 1965 年生于辽宁省沈阳市. 东北大学副教授, 在职博士生, 主要研究方向为数据挖掘, 信息融合.

赵 海 男. 1959 年生. 东北大学教授, 博士生导师, 博士, 主要研究方向为信息融合, 网络通信.