基于领域知识的预警规则发现研究

潘洁珠[1,2],吴共庆[1],胡学钢[1],张润梅[1,3]
(1. 合肥工业大学 计算机与信息学院,安徽 合肥 230009;
2. 合肥师范学院 计算机科学与技术系,安徽 合肥 230061;
3. 安徽建筑工业学院 计算机与信息工程系,安徽 合肥 230022)

摘 要:预警有助于及时预防和解决隐患,具有重要的应用价值,受到了广泛的关注。提出了一种新的预警机制,以领域相关的预警知识为基础构建预警系统,重点研究了以实际数据为资源,基于背景领域知识挖掘出预警规则,使得预警系统中的预警知识更丰富、更完备、更具有实际应用价值。将上述研究应用于高校教学教务系统的学生成绩预警,实验结果表明所提出的预警机制和方法是有效的。

关键词:预警;数据挖掘;领域知识;成绩分析
中图分类号:TP18;TP311 **文献标识码**:A **文章编号**:1673-629X(2008)07-0066-03

Research of Mining Early Warning Rules Based on Domain Knowledge

PAN Jie-zhu[1,2], WU Gong-qing[1], HU Xue-gang[1], ZHANG Run-mei[1,3]
(1. School of Computer and Info., Hefei University of Technology, Hefei 230009, China;
2. Dept. of Computer Sci. and Technology, Hefei Teachers College, Hefei 230061, China;
3. Dept. of Computer Sci. and Info. Eng., Anhui Institute of Architecture and Industry, Hefei 230022, China)

Abstract: The early warning is beneficial to prevent and solve hidden trouble in good time, accordingly warning study has important value. The mechanism and method of the early warning are concerned widely. A new warning mechanism was proposed to construct early warning system based on domain early warning knowledge in the paper. Concentrated mainly in mining early warning rules based on background domain knowledge from the real data, to make early warning knowledge richer, more complete and with more practical value. Applying the mechanism and method to resolve early warning problems of student achievement data in a university educational administration management information system. Experimental results suggest that they can work reasonably well.

Key words: early warning; data mining; domain knowledge; achievement analysis

0 引 言

预警(Early-Warning)指的是在警情发生之前对其进行预测报警,即在运用现有知识和技术的基础上,通过对事物发展规律的总结和认识,分析事物的现有状态及特定信息,判断、描述和预测事物的变化趋势,并与预期的目标量进行比较,利用设定的方式和信号,实行预告和示警,以便使预警主体有足够的时间采取相应的对策和反应措施[1]。

广义地说,预警是组织的一种信息反馈机制。它最初起源于军事,随着社会进步的需要,预警所具有的信息反馈机制逐步超越了军事,进入现代经济、技术、政治、教育、医疗、灾变、治安等自然和社会领域。国内外对预警理论的研究涉及经济预警、地震预警、粮食生产预警、银行预警、财务预警等许多方面。常见的预警方法有景气指数法、警兆信号法、警素预警法、ARCH预警方法、基于概率模式分类法、判别分析法、人工神经网络方法、系统动力学方法、专家经验法等[2]。

为了及时地适应实际问题的变化、实际领域的变化而导致预警机制和方法的变化,需探索一种更加灵活的预警机制。随着信息化技术的迅速发展和广泛应用,从已有积累的大量数据中自动提取有价值的预警知识并应用到预警系统成为一个重要的研究问题。从原始资料中自动提取预警知识能减少工作量,并且提取的预警知识更加真实、丰富,更具有应用价值。相关

收稿日期:2007-10-04
基金项目:安徽省重点教学研究项目(2007jyxm011);安徽省教育厅自然基金项目(2006KJ036B);合肥工业大学教学研究项目(XJ200527)
作者简介:潘洁珠(1979-),女,山东藤州人,讲师,硕士研究生,研究方向为数据挖掘;胡学钢,博士,教授,主要从事数据挖掘、机器学习、知识工程研究。

研究受到了研究者的关注,开展了研究工作并取得一定的成果[3~5]。笔者探索了一种从实际数据中自动提取预警知识的方法,解决了预警系统中预警知识获取困难的问题,在此基础上,以获取的预警知识为基础构建预警系统。

1　基于预警知识挖掘的预警机制

图1描述了基于预警知识挖掘的预警机制。预警系统由输入/输出接口、预处理模块、预警知识库、预警知识挖掘模块和预警模块组成。其中,预警知识挖掘模块根据用户输入的挖掘参数和预警领域知识约束从历史数据中挖掘出预警知识,保存到预警知识库中。由于历史数据可能是动态变化的,挖掘系统需定期从不断更新的历史数据中自动更新预警知识库。预警知识挖掘模块是预警系统的核心部分。对于监测数据,预警模块根据预警知识库和所设计的预警策略决定是否生成预警信息。该模块和预警信息输出接口一起可以提供实时控制接口,集成了多种事件响应接口,可以产生协作进程所需的消息、联动预警等功能。

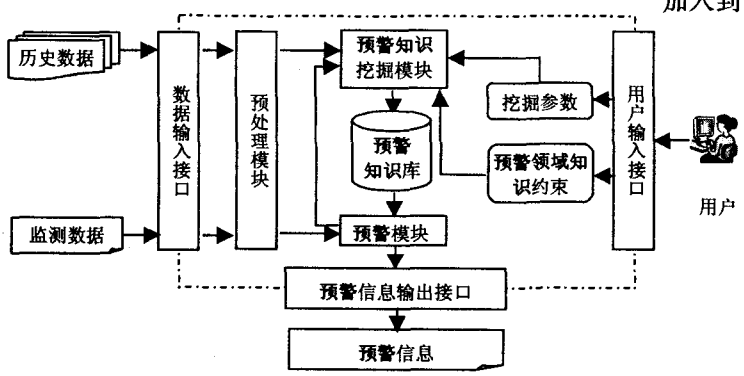

图1　基于预警知识挖掘的预警系统

2　预警规则及其发现算法

2.1　预警规则

预警规则知识反映了预警事件和其他事件之间依赖或关联的关系,预警事件中的项值可以依据与其存在关联的项值进行预测预警。在实际的预警过程中,当系统处理一个监测数据时,总是先通过预处理模块将其转变为系统规定的标准格式。预警模块根据输入数据的特征从预警规则库中提取预警规则集,然后分析预处理后监测数据,对于每一条规则进行匹配分析。匹配预警时,先在监测数据中寻找当前规则的前项,如果找不到,则认为此条规则不适用于这条记录,放弃匹配,转向规则集中的下一条规则;如果找得到,则认为此条规则适用于这条记录,预警模块根据预先设定的预警策略综合匹配结果生成预警信息。

下面给出一种预警知识描述形式:

预警规则:一条预警规则是形如 $X \rightarrow Y$ 的蕴涵式。其中 $X \subset I, Y \subseteq W$。对 $\forall x \in I, \forall y \in W$,有 $(x, y) \in P$。

其中:$I = \{i_1, i_2, \cdots, i_n\}$ 是文字的集合,其中的元素称为项(item),项 i_j 表示某事件的发生,$1 \leqslant j \leqslant n$。$W = \{w_1, w_2, \cdots, w_m\}$ 表示预警项集,其中 $w_i \in I$,$1 \leqslant i \leqslant m$。记项 x 先于项 y 出现为 $x \prec y$,P 为 I 上的项顺序关系集,$P = \{(x, y) \mid x \prec y, x, y \in I\}$。

2.2　基于领域知识约束的预警规则挖掘

上述预警规则可以看作一种受预警领域知识(预警项集和项顺序关系)约束的关联规则。给定一个交易集 D、预警项集 W 和项顺序关系集 P,挖掘预警规则的问题就是产生支持度大于给定的最小支持度(minsupp)、置信度大于给定的最小置信度(minconf)的关联规则 $X \rightarrow Y$,其中 $X \subset I, Y \subseteq W$。对 $\forall x \in I$, $\forall y \in W$,有 $(x, y) \in P$。

用传统关联规则挖掘算法挖掘出来的关联规则作为预警规则,会出现许多无用的规则。这些规则一旦加入到预警规则库中,在进行预警时就会出现错误,降低预警的准确率。为了使预警规则库中的规则更加有意义,便于预警,在挖掘过程中作了一些限定:

(1)支持度阈值限定。支持度是反映项集在数据库中的普遍规律。为了保证频繁项集更具有普遍的规律,可以根据需求设置支持度阈值,使频繁集更加有意义。

(2)置信度阈值限定。如果置信度阈值设置太小,则信任度非常小的规则也将被挖掘出,如果加入到预警规则库中,则会降低预警准确率,同时增加预警过程的开销。

(3)规则后件预警项集限定。一般情况下,关联规则的左部和右部项的属性是不作限定的,而在预警时,预警规则右部的项必须限定为预警项集中的元素。在挖掘算法中加入预警项集约束关联规则的右部项的属性,可将无效的规则剔除。

(4)规则前件和后件的项顺序关系限定。关联规则左部的项和右部的项之间项顺序关系是不作限定的,而预警规则中,规则前件中项事件需发生在规则后件项事件之前。因此,为了保证挖掘出的预警规则的有效性,须在挖掘算法中加入项顺序系约束条件。

文中的预警规则挖掘以 Agrawal 等人提出的 Apriori 算法为框架[6]。算法伪代码描述见 GenEarly-WarningRules。算法的第1~9行根据最小支持度约束计算出所有频繁项集,算法的第10~17行根据置信度

约束、预警项集约束和项顺序关系约束生成预警规则集 R。经过上述步骤的处理,所挖掘出的关联规则是比较符合实际的,于是将这些规则加入到预警规则库中。

经典的关联规则挖掘出来的是大批量的规则,而基于约束条件的挖掘可以挖掘出用户感兴趣的规则,实现定向挖掘。最小支持度约束从约束性质上属于一种反单调约束,可推进到频繁项集的挖掘过程中[7],而置信度约束、预警项集约束和项顺序关系约束难以推进到频繁项集的挖掘过程中,GenEarlyWarningRules 算法采用"事后约束"的方式来实现。

Procedure GenEarlyWarningRules

/* Input:事务集 D,最小支持度 minsupp,最小置信度 minconf,预警项集 W,项顺序关系集 P

Output:预警规则集 R */

//C_k 表示大小为 k 的候选集,L_k 表示大小为 k 的频繁项集

(1) L_1 = 初始频繁 1 - 项集;

(2) for($k = 2; L_{k-1} \neq \emptyset; k++$)do begin

(3) C_k 为 L_{k-1} 中产生的新候选集;

(4) 根据最小支持度 minsupp 对 C_k 进行剪切;

(5) for 所有事务(记录) do begin

(6) 遍历包含在记录项 T 中的候选集 C_k 中的所有候选项并计算支持计数;

(7) end for;

(8) $L_k = C_k$ 中所有支持度大于最小支持度 minsupp 的候选集;

(9) end for;

(10) for all 频繁模式 $l_k, k \geqslant 2$ do begin /* 生成规则 */

(11) for all subset $s_m: (s_m \subset l_k) \wedge (s_m \subseteq W) \wedge (\forall x \in (l_k - s_m), \forall y \in s_m$,有$(x, y) \in P)$ do begin

(12) conf = support(l_k)/support($l_k - s_m$);

(13) if conf ≥ 最小置信度 minconf then begin

(14) $R = R \cup \{(l_k - s_m) \rightarrow s_m\}$;

(15) end if;

(16) end for;

(17) end for

3 应用实例

随着高校招生规模的扩大和信息化程度的提高,高校教学教务系统数据库中积累了大量的数据,这些数据中蕴涵有大量有价值的规律,挖掘这些规律并应用于高校教学教务的决策工作和辅助应用工作,可促进教育决策和教育辅助应用的科学化、合理化、系统化[8]。为了验证所提出的预警机制和预警规则挖掘方法的有效性,以高校教学教务信息数据挖掘为背景,以学生成绩预警为目标,将文中的预警机制和方法应用于高校教学教务系统的预警功能模块。从学生成绩数据库中挖掘成绩预警规则,以发现的预警规则为依据,分析学生成绩,提出有针对性的课程成绩预警信息,以提高教学管理的预知性,为同学合理地调整课程学习计划、分配学习时间提供参考依据。

实现文中预警机制和方法的软件环境如下:操作系统为 Windows 2000,采用 SQL Server 2000 数据库管理系统,以 Java 为开发语言。记 $D_{Traning}$ 为训练数据集合,D_{Test} 为测试数据集合,$R = \{r_1, r_2, \cdots, r_n\}$ 为算法 GenEarlyWarningRules 从 $D_{Traning}$ 中挖掘出的预警规则集。定义规则 $r: X \rightarrow Y$ 的预警准确率为 $P(r)$,规则集 R 的预警准确率为 $PR(R)$。

$$P(r) = \frac{|\{T: X \cup Y \subseteq T, T \in D_{Test}\}|}{|T: X \subseteq T, T \in D_{Test}|} * 100\%$$

$$PR(R) = \frac{\sum_{r \in R} P(r)}{|R|}$$

以某高校某专业某年级 47 位同学以超过 90% 的比例选修的 25 门专业课成绩为训练数据源,以下一级 49 位同学的成绩为测试数据源,挖掘出的规则数作为评估挖掘结果的数量指标,规则集的预警准确率作为评估挖掘结果的质量指标,验证该预警机制和方法的有效性。

设置最小支持度分别为 0.30、0.25 和 0.20,最小置信度分别为 0.90、0.85、0.80,表 1 列出了不同参数条件下挖掘出的规则数和规则集的预警准确率。从表 1 中可以发现:minsupp 和 minconf 的值越小,挖掘出的规则数目越多,挖掘出知识的量越大。然而,预警准确率和 minsupp、minconf 的值之间不存在这种反比例关系。文中所提的模型和方法在上述现实数据源上可达到 65% 以上的预警准确率,通过合理设置参数,最高可达到 96% 以上的预警准确率。实验结果表明,该预警机制及其方法在实践上是有效的。

另外,在实验过程中发现,挖掘出是一种统计意义上的规律,有时在学习内容上不明显相关的课程却具有较强的关联,这种规律应用于实际的数据也能取得较好的预警效果。分析其原因,有可能是这些课程在学习的思维方式、学习方式、学习方法等方面具有相似性。

表 1 挖掘出的规则数和平均预测准确率

Minsupp	minconf=0.90		minconf=0.85		minconf=0.80	
	规则数	准确率(%)	规则数	准确率(%)	规则数	准确率(%)
0.30	1	90.91	3	96.97	25	67.69
0.25	2	67.68	7	81.24	29	68.04
0.20	4	78.45	11	76.24	47	65.89

(下转第 73 页)

下面要根据具体的事例对三种情况进行说明:

1) 若给出 $E = (\{2\},\{5\},\{1\})$,按包含度公式分别计算 E 和 F_1、F_2、F_3 的近似度。

$D(F_1/E) = 2/3, D(F_2/E) = 4/5, D(F_3/E) = 0$

F_2 对应的包含度最大,于是得到规则:

If$(a_1,2) \wedge (a_2,5) \wedge (a_3,1)$ then $d = 2$ (0.8)

显然这与已有的知识矛盾,但是可信度大于 0.5 属于小部分矛盾,可以近似推导。

2) 当 $E = (\{3\},\{2\},\{3\})$ 时,根据公式 $D(F_2/E) = 1$,而尽管$(\{3\},\{2\},\{3\})$ 不属于 M 中分量,但是由于等价类中冗余的存在,按照粗糙集原理同样可推导 $E = (\{3\},\{2\},\{3\})$ 为确定性规则,即

If$(a_1,3) \wedge (a_2,2) \wedge (a_3,3)$ then $d = 2$ (1.0)

显然这属于已有知识等价类的子集,可以准确推导,属于确定性规则。

3) 但是当任意给定的条件属性值 E 与 F 的比较中使得规则的可信度为零时,则完全属于新知识,要增加新的等价类。这也表明不能从现有规则推导实际的决策规则,必须重新构建分量,也就是往前面的每个向量中添加可区分的等价类,从而才可以推导新的决策规则。对于这种新知识不但条件属性值要划分新的等价类,决策值也要添加新的值,这难免就存在估计的问题,因为给定的只有条件分量的值,没有决策属性值。因此要考虑分类质量对等价类的划分。

5　结束语

利用粗糙集理论在数据库中解决普遍性的问题很难,因为寻找约简集的算法是一个 NP 难题,因此只能在具体问题中选择建立适当的模型来解决问题。文中创新点是:在充分考虑区分矩阵法和基于属性重要性的启发式算法上,用一个构造函数计算属性的加权频率,按属性频率的大小在区分矩阵重新排序后获得约简集。通过对静态规则和决策算法的分析,结合动态样本新增知识的各种情况,用分类质量标识其变化的结果程度;用集合向量包含度依次得到决策规则的极大可信度,并形成近似决策规则。

参考文献:

[1] Pawlak Z. Rough Sets[J]. International J. of Computer and Sciences, 1982,11(5):341－356.

[2] 张文修,吴伟志.粗糙集理论与方法[M]. 北京:科学出版社,2001.

[3] Zhang W X, Leung Y. Theory of including degrees and Its applications to uncertainty inferences[M]// Soft Computing in Intelligent Systems and Information Processing. New York: IEEE,1996:496－501.

[4] Hu X. Knowledge discovery in databases: an attribute－oriented rough set approach[D]. Canada: University of Regina, 1995.

[5] 胡可云.基于概念格和粗糙集的数据挖掘方法研究[D].北京:清华大学,2001.

[6] 张文修,仇国芳.基于粗糙集的不确定决策[M]. 北京:清华大学出版社,2005.

[7] Kryszkiewicz M. Comparative study of alternative types of knowledge reduction in inconsistent systems[J]. International journal of Intelligence Systems,2001,16:105－120.

(上接第 68 页)

4　结束语

提出了一种基于领域知识和数据挖掘技术的预警机制,给出了一种预警知识描述形式,设计并实现了预警规则挖掘算法。以现实数据为数据源,对某高校某专业的学生成绩进行基于领域知识约束的预警规则挖掘,获取预警规则集。以预警规则集为基础,根据预警算法生成预警信息,能取得较好的效果。在以后的工作中,将研究如何根据数据的动态变化自动设置参数提高预警准确率。另外,研究将预警项集以及项顺序关系等约束推进到频繁项集的生成过程中以提高挖掘的效率,也是一个感兴趣的问题。

参考文献:

[1] 李本建. 关于建立航空工业经济预警系统的初步设想[J]. 航空系统工程, 1993(5):24－28.

[2] 王耀中, 侯俊军, 刘志忠. 经济预警模型述评[J]. 湖南大学学报,2004,18(2):27－31.

[3] 柳炳祥. 基于数据挖掘的危机管理及其预警方法研究[D]. 南京:东南大学, 2003.

[4] 姚靠华, 蒋艳辉. 基于决策树的财务预警[J]. 系统工程, 2005,23(10):102－106.

[5] 胡华平, 张　怡, 陈海涛, 等. 面向大规模网络的入侵检测与预警系统研究[J]. 国防科技大学学报,2003,25(1):21－25.

[6] Agrawal R, Imielinskia T, Swami A. Mining Association Rules between Sets of Items in Large Databases[C]// Proc. of the ACM－SIGMOD 1993 Int'l Conference on Management of Data. Washington D.C.:[s.n.],1993:207－216.

[7] 卢炎生, 杨　芬, 赵　栋. 带单调约束的关联规则挖掘[J]. 计算机工程,2004,30(15):78－80.

[8] 张玉林. 数据挖掘技术在教学过程中的指导作用[J]. 西安通信学院学报,2006,5(2):38－40.