例外关联规则挖掘*)

印 鉴[1,2] 周祥福[3]

(中山大学计算机科学系 广州 510275)[1]

(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]

(中山大学附属第三医院 广州 510630)[3]

Exception Rule Mining

YIN Jian[1,2] ZHOU Xiang-Fu[3]

(Department of Computer Science, Zhongshan University, Guangzhou 510275)[1]

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)[2]

(The Third Affiliated Hospital, Zhongshan University, Guangzhou 510630)[3]

Abstract Data mining is the process of discovering hidden structure or patterns in large quantities of data by using kinds of analytic tools. The structure or patterns can help decision makers for advantageous actions. This paper introduces the concept of interestingness and reference rules, and uses interestingness to estimate the information included in rule, and then presents a method for mining exception rules while computing the interestingness according to the reference rules. Experiments compared with other methods show that the proposed method has the better effects.

Keywords Data mining, Exception rules, Interestingness, Common sense rules, Reference rules

1 引言

数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

通常，经过某些数据挖掘工具的挖掘后，例如，文[1]所给出的快速算法，我们会得到大量的关联规则。对用户来说，从这些大量的规则中找出自己感兴趣的规则十分困难，而且，也很难知道哪些规则是我们真正感兴趣的。兴趣度依赖于人们的知识，且会因人们知识的不完整和不精确而产生偏差，因此必须首先确定一些我们认为是正确的知识，这些知识可以从数据中挖掘出来，称之为常识规则，依据常识规则，就可以无偏见地来确定所挖掘出来的规则是否是感兴趣的。

有时，与常识相矛盾的规则恰恰是我们感兴趣的，称之为例外规则。例外规则在某些决策中有着重要的作用[2,3]，而且，对数据挖掘工作来说，只产生常识关联规则并没有做得很完善。首先，常识关联规则在其表达形式上并没有考虑各种可能的反面示例的影响，导致知识表达功能不够完善；其次，可能有些常识关联规则是一种常识，对用户来说并没有多大意义(即使它的可信度和支持度都很高)；再次，还有一些知识隐藏在数据库里面，并不能被常识关联规则所表达，这些知识可能对用户做出决策有很大的帮助。

鉴于以上不足，本文引入了例外关联规则的概念，使用兴趣度来衡量例外关联规则所包含的信息量，并给出了一个挖掘例外规则的算法。

2 问题描述与相关工作

2.1 关联规则及其定义

下面首先给出关联规则挖掘问题的形式化描述，设 $I=\{i_1,i_2,\cdots,i_m\}$ 是 m 个不同项目的集合，给定一个事务数据库 D，其中每一个交易 T 是 I 中一些项目的集合，即 $T\subset I$。每一个交易 T 都与一个唯一的标识符 TID 相联。如果对于 I 中的一个子集 X，有 $X\subset T$，我们就说一个交易 T 包含 X。一条关联规则就是一个形如 $X\Rightarrow Y$ 的蕴涵式，其中 $X,Y\subset I$，而且 $X\cap Y=\phi$。X 称作规则的前提，Y 是结果。

为了说明关联规则的正确程度和支持率，人们引入了可信度和支持度两个概念。

关联规则 $X\Rightarrow Y$ 的可信度 C 表示在交易集中，在包含了 X 的所有交易的集合中有 c%包含了 Y。关联规则 $X\Rightarrow Y$ 的支持度 S 表示在交易集 D 中，有 s%的交易包含了 $X\cup Y$。关联规则的可信度和支持度可以计算如下：

令 Pr(X)表示项目集 X 的支持度，即在交易集 D 中有 Pr(X)的交易包含了 X，于是 $X\Rightarrow Y$ 的支持度 $Pr(XY)=Pr(X\cup Y)$，$X\Rightarrow Y$ 的可信度 $Pr(Y|X)=Pr(X\cup Y)/Pr(X)$。

由于关联规则并没有考虑反面示例的影响，例如：买了方糖就不买砂糖的可能性是 80%，这种购买趋势在现实中是很可能存在的，如果不对关联规则的定义加以改进，在挖掘结果中就会遗失这条很有用的信息。

在本文中对关联规则定义改进如下：把反面示例考虑进去，扩展项目集 I，使 I 为项目 i 和项目的反面示例 $\neg i$ 的集合。对交易集中的每一个交易 t，即 i 和 $\neg i$ 这两个项目肯定有且只有一个存在于交易 t 中。于是，对于关联规则 $X\Rightarrow Y$，X 集合和 Y 集合都可以包含项目的反例 $\neg i$。

这样改进之后，虽然挖掘的关联规则会更多，但通过兴趣度进行筛选后，所挖掘到的感兴趣规则会大大减少。

2.2 兴趣度及其定义

为了评价一关联规则所包含的信息量，人们定义了兴趣度这个度量值，兴趣度是一个事件包含的信息量的大小。一关联规则包含的信息量越大，兴趣度就越高。通常定义一个阈值，如果兴趣度大于这个阈值，就说明兴趣度较高，于是该关

*)本文研究得到国家自然科学基金部分资助(69733030)、广东省自然科学基金资助(001264)、广东省教育厅《软件技术》重点实验室研究基金资助。印 鉴 博士，副教授，主要研究方向为数据采掘、人工智能等。

联规则可以称为感兴趣的关联规则或兴趣度高的关联规则。

兴趣度有很多种计算方法[4]，一般，估计兴趣度的方法分为主观方法和客观方法两类：

主观方法就是用户直接运用自身的知识来估计兴趣度，有时用户甚至不了解该领域知识就直接凭借自身的固有观念来估计兴趣度，因此，用主观方法估计出来的兴趣度可能随用户的不同而不同。由于对该领域不完备或不正确的了解，这种估计可能与结果偏离。一个潜在的问题就是，当把主观方法估计兴趣度应用于竞争性强的商业环境中，用户的主观判断很可能是不现实的，因为它只反映了个别人的固有观念。

客观方法就是在估计兴趣度的时候使用从原始数据中提取出来的知识，由于知识是从原始数据中提取的，所以可能发现一种不依赖于用户固有观念的方法来测量兴趣度。

2.3 例外关联规则及其定义

例外关联规则的定义如下[5]：

如果 A⇒X 是常识关联规则(高支持度，高可信度)

B⇒¬X 是参照关联规则(低支持度，可能是低可信度)

则 A，B⇒¬X 是例外关联规则(低支持度，高可信度)

这里，参照关联规则的 B，X 均取自于常识关联规则中的项。

对一个特定的常识关联规则，用户通常找到越来越多的例外关联规则，最后可能误导用户。因此，必须对每一个挖掘出来的例外关联规则评估一下它所包含的信息量的大小，即评估它的兴趣度。

我们知道，一些与用户固有观念冲突的信息一定是兴趣度高的。我们把例外关联规则定义为与用户固有观念冲突的关联规则，如果例外关联规则包含了正确度高(即可信度和支持度都达到了目标值)的信息，那么该例外关联规则也一定是兴趣度高的。例外关联规则可能在关键决策中扮演一个重要的角色，兴趣度高的例外关联规则提供例外的知识，可以使决策者做出有利的决定。但是，例外关联规则又是次要的，它们或者不为人所知，或者被忽略掉。例外关联规则与常识关联规则是对立的，它们有较低的支持度，但它们却有着较高的可信度，这一点和常识关联规则是一样的。

用户可以根据实际情况给例外规则定义一个最小支持度和最小可信度，以保证挖掘出有用的例外关联规则。

2.4 几种例外规则兴趣度的计算方法

通过兴趣度来挖掘例外规则，许多研究者做了这方面的工作，这里，介绍两种取得了较好结果的方法。

在文[6]中，J-measure 方法使用例外规则的支持度和可信度来计算兴趣度，J-measure 方法定义如下：

$$J(X,AB)=\Pr(XAB)\log_2\frac{\Pr(X|AB)}{\Pr(X)}+\Pr(\neg XAB)\log_2\frac{\Pr(\neg X|AB)}{\Pr(\neg X)}$$

由于 J-measure 方法没有使用从原始数据中提取的常识关联规则，所以它并没有很好地估计一个例外规则的兴趣度，

在文[7]中，GACE-方法(Geometric mean of the Average Compressed Entropies)对 J-measure 方法做了改进，使用常识关联规则 A⇒X 和它的例外关联规则 AB⇒¬X 来计算兴趣度。GACE-方法的定义如下：

$$GACE(A\Rightarrow X,AB\Rightarrow\neg X)=\sqrt{J(X,A)J(\neg X,AB)}$$

GACE-方法并没有使用参照关联规则，因此可能会丢掉一些重要的参考信息，必须加以改进。

3 算法

3.1 兴趣度计算方法的改进

本文的研究目的主要是希望挖掘出感兴趣的例外规则，对这些例外规则的兴趣度，我们采用了客观方法来衡量，基本思想是通过已挖掘出来的常识关联规则和参照关联规则来计算其兴趣度，因此，我们对 GACE 的计算方法进行了改进。

关联规则的兴趣度 I 由两部分相加组成，一部分是从可信度计算得到的兴趣度，另一部分是从支持度计算得到的兴趣度。

首先，从可信度来考虑兴趣度。

一般，当没有其它信息时，一个低概率的事件发生会比一个高概率的事件发生提供给我们的信息更多。根据信息理论，可用下式来描述一个事件所包含的信息量：

$I=-\log_2 P$

这里，P 为事件发生的概率。

同理，对一个给定的具有可信度 Pr(X|AB)的规则 AB⇒X，我们可用 $-\log_2\Pr(X|AB)$ 和 $-\log_2\Pr(\neg X|AB)$ 来描述其信息量，于是关联规则 AB⇒X 的期望信息量为：

$$I_c^{AB_0}=-\Pr(X|AB)\log_2\Pr(X|AB)-\Pr(\neg X|AB)\log_2\Pr(\neg X|AB)$$

现在，我们再考虑与 AB⇒X 相关的规则 A⇒X 和 B⇒X，如果这两条规则所描述的信息与 AB⇒X 相差很大，则 AB⇒X 的兴趣度可能会大，可用下式来描述 A⇒X 和 B⇒X 所包含的期望信息量：

$$I_c^{AB_1}=-\Pr(X|AB)(\log_2\Pr(X|A)+\Pr(X|B))-\Pr(\neg X|AB)(\log_2\Pr(\neg X|A)+\Pr(\neg X|B))$$

我们把规则 AB⇒X 从可信度计算得到的兴趣度定义为：

$$I_c^{AB}=|I_c^{AB_1}-I_c^{AB_0}|=\Pr(X|AB)\left|\log_2\frac{\Pr(X|A)\Pr(X|B)}{\Pr(X|AB)}\right|+\Pr(\neg X|AB)\left|\log_2\frac{\Pr(\neg X|A)\Pr(\neg X|B)}{\Pr(\neg X|AB)}\right|$$

基于同样的分析，我们把规则 AB⇒X 从支持度计算得到的兴趣度定义为：

$I_s^{AB}=|I_s^{AB_1}-I_s^{AB_0}|$，其中

$$I_s^{AB_0}=-\Pr(ABX)\log_2\Pr(X|AB)-\Pr(AB\rightarrow X)\log_2\Pr(\neg X|AB)$$

$$I_s^{AB_1}=-\frac{\Pr(ABX)}{\Pr(AX)}\log_2\Pr(X|A)-\frac{\Pr(ABX)}{\Pr(BX)}\log_2\Pr(X|B)-\frac{\Pr(AB\rightarrow X)}{\Pr(A\rightarrow X)}\log_2\Pr(\neg X|A)-\frac{\Pr(AB\rightarrow X)}{\Pr(B\rightarrow X)}\log_2\Pr(\neg X|B)$$

因此，关联规则 AB⇒X 的总兴趣度为 $I=I_c^{AB}+I_s^{AB}$

3.2 算法流程

例外关联规则的定义和兴趣度的计算前面已经给出，下面给出例外关联规则挖掘的算法 CRI。

算法 CRI

```
Begin
  LI=∅ //包含所有的频繁项目集
  LC=∅ //包含所有常识关联规则
  LR=∅ //包含常识关联规则所推出的参照关联规则
  LE=∅ //包含候选的例外规则
  LI←GenerateLargeItemSet()//运行 Apriori 算法，得到频繁项目集
  LC←GenerateAllCommonSense(LI) //运行 Apriori 算法，得到常识关联规则
  For each CSi from LC do //遍历常识关联规则
    A←GetAntecedent(CSi) //得到常识关联规则的前件
    LR←GetReferences(CSi, LC) //从常识关联规则 CSi 推出参照关联规则
    For each RRj from LR do
    B←GetAntecedent(RRj) //得到参照关联规则的前件
    If (A∪B) is not in LI //如果 A∪B 不在频繁项目集 LI 中，
                          //则 AB⇒¬X 一定不会是高支持度
      Insert(A∪B∪¬Consequent(CSi),LE)
    End for
  End for
LE←GenerateExceptions(LE) //筛选出高于一定支持度和可信度
```

```
                                    //阈值的例外关联规则
  EstimateInteresting ness(CS,LE) //计算例外关联的兴趣度
End
```

4 实验

我们首先通过一组简单的交易数据来说明算法的应用及其有效性。

经过对很多数据的测试，决定筛选常识关联规则时使用30%的支持度和70%的可信度。筛选异常关联规则时使用5%的支持度和60%的可信度，如果兴趣度大于2.5，则认为兴趣度较高。

该组数据资料的内容如下，项目集是{1,2,3,4,5}，共有19个交易。

{1 4},{1 4},{1 4},{1 4},{2 4},{2 4},{2 4},{2 4},{1 4},{1 4},{1 4},{1 4},{2 4},{2 4},{2 4},{2 4},{1 2 3 5},{1 2 3 5},{1 2 3}

运行程序得到结果：

频繁项目集：

{1}支持度:57%　　{2}支持度 57%

{4}支持度 84%　　{1,4}支持度 42%

{2,4}支持度 42%

常识关联规则：

{1}⇒{4}支持度:42% 可信度 73%

{2}⇒{4}支持度:42% 可信度 73%

异常关联规则：

{1}∪{2}⇒¬{4} 支持度:15%　可信度:100%　兴趣度:3.852

{2}∪{1}⇒¬{4} 支持度:15%　可信度:100%　兴趣度:3.852

通过观察原始数据，不难看出{1}∪{2}⇒¬{4}确是一条感兴趣的异常关联规则。

表 1

Index	Common Sense Reference (substituted by common sense) Exception	Conf %	Supp %	CRI	J-measure.	GACE
1	{4}→{1} {8,10}→{1} {4,8,10}→{1}	58 55 42	24 15 5	2.50034	-0.00713701	0.0134813
2	{7}→{1} {10}→{1} {7,10}→¬{1}	57 53 42	20 16 5	2.57149	-0.00713701	0.014194
3	{7}→{1} {4,8}→{1} {7,4,8}→¬{1}	57 56 52	20 21 10	2.52952	-0.0182566	0.0227016
4	{10}→{1} {7}→{1} {10,7}→¬{1}	53 57 51	16 20 6	2.57149	-0.00713701	0.00894075
5	{10}→{1} {4,8}→{1} {10,4,8}→¬{1}	53 56 42	16 21 5	2.521	-0.00713701	0.00894075
6	{4,8}→{1} {7}→{1}{4,8,7}→¬{1}	56 57 36	21 20 5	2.52952	-0.0182566	0.0223415
7	{4,8}→{1} {10}→{1} {4,8,10}→¬{1}	56 53 42	21 16 5	2.521	-0.00713701	0.0139688
8	{4,8}→{1} {7,8}→{1} {4,7,8}→¬{1}	56 58 36	21 18 5	2.6588	-0.0182566	0.0223415
9	{4,8}→{1} {8,10}→{1} {4,8,10}→¬{1}	56 55 42	21 15 5	2.64752	-0.00713701	0.0139688
10	{7,8}→{1} {4,8}→{1} {7,4,8}→¬{1}	58 56 36	18 21 5	2.6588	-0.0182566	0.0233533
11	{8,10}→{1} {4}→{1} {8,10,4}→¬{1}	55 58 42	15 24 5	2.50034	-0.00713701	0.0117267
12	{8,10}→{1} {4,8}→{1} {8,10,4}→¬{1}	55 56 42	15 21 5	2.64752	-0.00713701	0.0117267

为了进一步说明算法的有效性，我们把算法CRI与前文所述的算法进行了比较。实验数据采用了文[1]中Apriori算法的数据，数据参数为T5.I2.D100K，且N=10。以支持度为15%，可信度为50%的阈值来挖掘常识关联规则，以支持度为5%，可信度为35%的阈值在已经提取出来的规则中挖掘例外关联规则，使用CRI方法时用2.5的阈值来决定例外关联规则是否是感兴趣的。我们用前面提过的三种方法来估计兴趣度，从而判断CRI方法的有效性：(1)用J-measure方法。(2)用GACE方法。(3)用CRI方法。

实验中通过对比这三种不同方法估计出来的兴趣度，从而确定CRI方法是否达到预期目标：(1)CRI方法能够找到J-measure方法和GACE方法找出的所有兴趣度高的规则。(2)CRI方法能区别J-measure方法和GACE方法所不能区别的不同例外关联规则。(3)两个等价的规则集，CRI方法估

计出来的兴趣度相同。实验结果见表1。

从实验结果可得出如下结论：首先，CRI方法能区别J-measure方法和GACE方法所不能区别的不同例外关联规则。例如，如果使用J-measure方法，对第1,2,5,9个规则估计的兴趣度都是一样的，但是CRI方法能够区分这些规则集相应的兴趣度大小。再看GACE方法，第11,12个规则集估计的兴趣度一样，但是CRI方法能够区分这些规则集相应的兴趣度大小；其次，两个等价的规则集，CRI方法估计出来的兴趣度相同。1和11,2和4,3和6,5和7,8和10,9和12都是等价的规则集，但是J-measure方法和GACE方法却对某些等价的规则集算出不同的兴趣度，而CRI方法对这些等价方法估计出来的兴趣度都是相同的；再次，CRI方法基本上能够找到J-measure方法和GACE方法找出的所有兴趣度高的规则。

结论 本文讨论了例外关联规则的挖掘方法，并引入兴趣度的概念来测量例外关联规则所包含的信息量。本文对估计兴趣度的方法做了改进，根据常识关联规则与参照关联规则来度量一条例外关联规则的相对兴趣度，从而能更好地反映例外关联规则所包含的信息量，挖掘出真正感兴趣的例外关联规则。本文给出了具体的实现算法，实验结果表明该方法达到了较好的效果。

参考文献

1 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), 1994. 487～499

2 Liu H, Lu H, Feng L, Hussain F. Efficient search of reliable exceptions. In Methodologies for Knowledge Discovery and Data Mining, LNAI 1574(PAKDD), Berlin: Springer, 1999. 194～203

3 Padmanabhan B, Tuzhilin A. A belief-driven method for discovering unexpected patterns. In: Proc. Fourth Int'l Conf. Knowledge Discovery and Data mining(KDD), 1998. 94～100

4 周欣，沙朝锋，朱扬勇，施伯乐. 兴趣度——关联规则的又一个阈值. 计算机研究与发展，2000，37(5)：627～633

5 Suzuki E, Kodratoff Y. Discovery of surprising exception rules based on intensity of implication. In Principles of Data Mining and Knowledge Discovery, LNAI 1510 (PKDD), Berlin: Springer, 1998. 10～18

6 Smyth P, Goodman R M. An information theoretic approach to rule induction from databases. IEEE Trans. Knowledge Discovery and Data Eng, 1992, 4(4): 301～316

7 Suzuki E. Discovering unexpected exceptions: a stochastic approach. In: Proc. Fourth Int'l Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery(RSFD), 1996. 225～232

(上接第29页)

1)个性化服务 通过用户聚类，根据同组中其他用户的页面请求的页面为用户动态生成页面；根据页面聚类，决定当前页面包含哪些页面的超链接，自动修改站点的拓扑结构；将当前用户会话与已有的页面聚类进行最佳匹配，为用户推荐其尚未浏览并且与当前页面之间没有直接链接的页面。

2)系统改进 用户对网站服务的满意程度与系统的性能密切相关，包括服务器缓存、网络传输、负载平衡、数据分布等。根据代理服务器的访问日志预测用户的请求在时间和空间的分布，这种预测可以帮助代理服务器选择页面预取和缓存的策略；从服务器日志中挖掘路径配置文件，作为创建动态HTML页面的依据，所以用户请求动态页面之前已经将动态页面生成，从而减少服务器的响应延迟时间。

3)网站拓扑结构改进和构建自适应站点 根据用户访问模式所描述的用户访问站点时的行为信息，可以修改站点页面内容、链接结构和构建自适应站点的信息。

4)网络安全 随着Internet的迅速发展，网站的安全性成为关注的焦点，同时也成为制约基于网络的应用的主要因素。由于Web日志挖掘技术可以用于监视非法登录、黑客入侵等，所以将Web日志挖掘技术和网络技术相结合是提高门户网站安全性的较佳的解决方案。

5 Web用户访问模式挖掘的发展方向

Web用户访问信息挖掘是一个较新的研究领域，具有广阔的发展和应用前景，应该指出的是，面对日益增加的商业需求，Web用户访问信息挖掘技术还有许多问题需要解决，有待这一领域的研究者深入研究。将来很有用的几个研究方向是：

(1)Web日志挖掘中内在机理及新的挖掘体系和结构的研究；

(2)用户访问模式库的动态维护和更新、模式(知识)的评价体系和评价方法；

(3)挖掘算法在海量数据挖掘时的适应性和时效性研究；

(4)智能站点服务个性化和性能最优化的研究；

(5)关联规则和序列模式在构造自组织站点的研究[8]；

(6)分类在电子商务市场智能提取中的研究[8]。

结束语 目前，国内外在Web用户访问信息挖掘领域的研究尚处于起步阶段，还没有形成比较成熟的理论和统一的体系。本文通过阐述Web用户访问信息挖掘中数据预处理的主要流程和技术方法、用户访问信息挖掘算法、模式分析方法以及Web用户访问信息的应用，旨在对Web日志挖掘的框架、流程、技术方法及用户访问信息挖掘的研究和发展方向做全面概括。

随着Internet的进一步发展，Web用户访问信息挖掘在个性化的信息服务、改进门户站点的设计和服务、开展有针对性的电子商务、电子政务、构建智能化Web站点、提高网站的声誉和效益等方面将起到极其重要的作用，Web用户访问信息挖掘技术将成为重要的研究课题和方向。

我们正在承担教育部科技司重点项目："远程教育网关键技术——信息挖掘和智能搜索工具的研究"(教技司[2000]175)。结合我们在数据挖掘内在机理研究方面所取得的成果，试图提出一种新型的Web用户访问信息挖掘的结构框架——基于双库协同机制的Web用户访问信息挖掘，我们将在以后的文章中详细论述。

参考文献

1 Kosala R, Blockeel H. Web Mining Research: A Survey. SIGKDD Exllorations, 2000, 2(1)

2 Pitkow J. In search of relable usage data on the www. In: Sixth Intl. World Wild Web Conf. Santa Clara, CA, 1997. 451～463

3 Pirolli P, Pitkow J, Rao R. Silk from a sow's ear: Extracting usable structures from the web. In: Proc. of 1996 Conf. on Human Factors in Computing Systems (CHI-96), Vancouver, British Columbia, Canada, 1996

4 Cooley R, Mobasher B, Srivastava J. Grouping web page references into transactions for mining world wide web browsing patterns: [Technical Report TR 97-021]. University of Minnesota, Dept. of Computer Science, Minneapolis, 1997

5 Catledge L, Pitkow J. Charactering browsing behaveious on the World Wild Web, Computer Networks and ISDN Systems, 1995, 27(6): 1065～1073

6 Cooley R, et al. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1999, 1

7 Srivastava J, et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Appear in SIGKDD Explorations, 2000, 1(2)

8 王实，高文，李锦涛. Web数据挖掘. 计算机科学，2000，27(4)