感兴趣度的研究综述*)

An Overview of Measures of Interestingness

杨炳儒　綦艳霞

(北京科技大学信息工程学院计算机系　北京100083)

Abstract In KDD, there are a number of patterns and rules discovered from large database by data mining, but most of them are of no interestingness to the user. How to get those that are useful and interesting is crucial. Therefore it is of great value to discuss the measures of interestingness of discovered patterns. Such measures of interestingness are divided into objective measures and subjective measures. The two kinds of measures of interestingness are respectively discussed in the paper.

Keywords Knowledge discovery, Measures of interestingness, Objective measure, Subjective measure

1 引言

在知识发现过程中,通过数据挖掘算法产生大量的模式和规则,但是大多数用户并不感兴趣。面对如此众多的模式和规则,用户不能很好地去理解它们从而无法把精力集中在其中真正感兴趣的子集上。因此,如何定义和确定好的度量标准,识别出用户感兴趣的和有用的知识成为至关重要的一环,所以对感兴趣的度量标准(感兴趣度)的研究是知识发现的重要内容。

感兴趣度分为客观感兴趣度和主观感兴趣度。客观感兴趣度主要根据模式或规则的形式和数据库中的数据进行定义,属于数据驱动;而主观感兴趣度还要考虑用户的参与等人为因素的影响,属于用户驱动。单纯的只使用客观感兴趣度是不够的,它不能考虑模式和规则的所有方面,同时感兴趣的问题从本质上将是一个主观的问题,不同的用户感兴趣的规则和模式是不一样,因此需要用户的参与,所以要综合使用客观感兴趣度和主观感兴趣度这两种度量标准。比较合理的方法是首先用客观感兴趣度作为第一级过滤器,选出潜在感兴趣的模式,然后再用主观感兴趣度来对它们进行第二级筛选,得到用户真正感兴趣的知识[1]。下面分别针对客观感兴趣度和主观感兴趣度分别进行介绍。

2 客观感兴趣度

2.1 基本的评价准则

目前感兴趣度的研究主要是针对规则的客观感兴趣度,一般的影响规则感兴趣度的数据方面的因素共有三个[1]:(规则 A⇒B)

1)覆盖度:指前件 A 出现的概率 P(A);

2)完全性:指两者 (A 与 B)同 时出现的概率与 B 出现的概率之比 P(A∧B)/P(B);

3)可信度指两者(A 与 B)同时出现的概率与 A 出现的概率之比 P(A∧B)/P(A)。

Piaretsky-Shapiro 提出的规则感兴趣性 RI(Rule Interestingness)度量的三个准则是:

1)如果 P(A∧B)=P(A)P(B),那么 RI=0;

2)当其它参数固定时,RI 随着 P(A∧B)的增加单调递增;

3)当其它参数固定时,RI 随着 P(A)或 P(B)的增加单调递减。

Magor 和 Mangano 提出了第四个准则:

4)当给定的可信度大于允许的可信度时,RI 随着 P(A)的增加单调递增。

另一个通用的评价规则质量的是规则简洁性,主要是要求规则的前件和后件(主要是前件)包含的属性的项数不要太多。即 A 的属性数目越少,规则越简洁,客观感兴趣度越高。一般地,A 包含的属性越少,P(A)越大。

2.2 粗的度量方法

Symth 利用函数[2]:

$$J(A;B)=P(A)(P(A|B)\log(\frac{P(B|A)}{P(B)})+(1-P(B|A))\log(\frac{1-P(B|A)}{1-P(B)}))$$

作为对规则 A⇒B 的简洁性和包含的信息量进行综合度量,考虑了规则的前件 A 和后件 B 的概率分布的相似程度,以及用 A 的出现概率作为前件的简洁性的度量,但是,忽略了 P(B)的作用。函数:

$$Interest(A\Rightarrow B)=\frac{1-P(B)}{(1-P(A))\times(1-P(B\wedge A))}$$ [3]

*)国家自然科学基金重点项目(69835001)资助。

可以度量规则 A⇒B 的新颖性，考虑到了规则 A⇒B 的前件 A 和后件 B 的耦合，以及对大概率事件产生原因的知识已经较多，而对大概率事件导致的结果更加关注等特点，它不对 P(B)大的规则赋予大的感兴趣程度值，并可以判断出在 A⇒B 和 B⇒A 中，对哪一个的感兴趣度更高，因为，一般地，$Interest(A\Rightarrow B)\neq Interest(B\Rightarrow A)$。

在分类规则中，误分类的风险也应该考虑，因为在应用所发现知识的过程中，对于不同的误分类的结果所承担的风险是不一样的。这样就需要在原有的规则感兴趣度的基础上乘以误分类的风险值，记为 MisclassCost，定义为可能误分类费用的总数的倒数：

$$MisclasCost=1/\sum_{j=1}^{k}Prob(j)Cost(i,j)$$

注：其中 Prob(j)是真正属于类 j 的概率，类 i 是被预测的类，Cost(i,j)是把类 j 预测为类 i 的费用，k 是类的数目。

利用挖掘出的关联规则所表示出的信息 $C_R=P(A\wedge B)/P(A)$ 与原数据库支持该信息(用 P(B)表示)的差异定义的感兴趣度为 $[C_R-P(B)]/\max\{C_R,P(B)\}$[4]。它的值介于－1和1之间，如果一条规则的感兴趣度越大于0，那么其实际利用价值越大；如果越小于0，则其反面规则的实际利用价值越大。

上述的规则感兴趣度是把规则的前提作为一个整体来考虑的，并没有考虑其中的各个属性，从这个意义上讲这些度量方法是粗精度的。但是，对于在粗的度量标准下相同的两条规则，用户感兴趣程度是不同的。所以还要考虑前件中的单个属性的感兴趣度，这属于精度比较细的度量。

2.3 细的度量方法

对于发现的规则，当要使用它时，必须首先获得规则中前件的各个属性的值。但是得到不同属性值的花费是不一样的，往往差别很大。所以在知识发现过程中要考虑确定属性值的费用，称为属性的费用，用规则前件中所有属性的花费的算术平均的倒数，记为：

$$AttUsef=1/(\sum_{i=1}^{k}Cost(A_i)/k)$$

作为规则的属性费用的度量标准。

另一个要考虑的是单个属性的新颖性，利用 $AttSurp=1/(\sum_{i=1}^{k}InfoGain(A_i)/k)$ 可度量规则的属性的新颖性。其中 $InfoGain(A_i)$ 是第 i 个属性的信息增益，它定义为规则后件作为一个类的信息熵减去当给定了这个属性的属性值的情况下这个类的熵，在其它不变的情况下，前件包含低的信息增益的规则比包含高的信息增益的规则更具有新颖性，用户对它更感兴趣。

在考虑上述两个细的度量标准时，只须在求出其它粗的综合度量后，将它们作为附加的乘积项加进去即可。

3 主观感兴趣度

3.1 基本的评价标准

仅仅根据客观感兴趣度选取用户所关注的模式，难于获得用户真正感兴趣的模式，还需要人为的参与。目前，对主观感兴趣度的研究尚不多见。

从主观的角度，能使用户对发现的模式产生兴趣的原因主要有两个：

(1)意外性[5~6]：如果所获得的模式使用户感到出乎意料，则表明它具有意外性。

(2)实用性[7~8]：如果用户可以利用所获得的知识采取一定的行动，从而提高工作效率或带来一定的经济效益，则认为它具有实用性。

这两者是紧密相关的，一般来讲，大多数具有意外性的模式也具有实用性；具有实用性的模式也具有意外性。尽管实用性是用户主要的目标，但是直接衡量模式的实用性是很困难的，这是因为：1)必须把所有的模式划分成一系列的类，给每个类指定能给用户带来效益的相应的行动。但是用户并不知道全部的模式，即使知道，给它们分类并分配行动的过程是很复杂的。2)采取的行动和它与模式类之间的映射关系是随时变化的，这样重新把模式分配给相应的行动便成为一项很艰巨的任务。况且它是与应用领域密切相关的，适用范围不广，因此便通过评价模式的意外性实现对它的实用性的度量。模式意外性的度量标准主要根据模式与用户的信念系统和各种类型知识之间的关系来制定，下面分别进行介绍。

3.2 信念系统意义下的感兴趣度

信念分为两种：硬信念和软信念。硬信念是指那些固定不变的信念，它不会因为新模式的产生而改变。如果发现的模式与硬信念相对立，说明在获取这个模式的过程中出现了错误。同时硬信念是主观的，不同的用户硬信念不同。软信念可根据新模式的产生有相应的变化，用户为每个软信念都赋予一个数值 d，来衡量用户相信它的程度。这个数值可以通过几种方法给出，有贝叶斯方法、Dempster-Shafer 方法、频率的方法、Cys's 方法和统计的方法[6]。

针对硬信念和软信念，分别给出相应的模式感兴趣度的定义：

(1)硬信念：如果一个模式与硬信念相悖，那么用户一直对它感兴趣。

(2)软信念：因为信念系统中各个信念的重要程度不同，用户对于改变某些信念感兴趣的程度会高于其

它的信念，所以形式上为信念系统 B 中的每个软信念 a_i 规定一个权值 w_i，并且它们的累加和为1。

新模式 p 相对于一个软信念系统 B 和以前的证据 ζ 的感兴趣的形式化定义为：

$$I(p,B,\zeta)=\sum_{a\in B_s} w_i|d(a_i|p,\zeta)-d(a_i|\zeta)|$$

通过计算新模式 p 改变信念的程度来定义它的感兴趣度。

3.3 通过用户输入各种类型的知识(主要是各种规则)进行感兴趣度的分析

针对关联规则感兴趣度的分析，用户输入的知识根据其准确程度划分为三种类型[3]。它的划分标准是对知识表达的准确程度。它们分别是：广义的描述、合理的概念和准确的知识。广义的描述表示了用户的一些模糊的感觉，指出某些属性之间存在着关联关系，但是关联的方向没有给出，即不能确定哪些为前件哪些是后件；合理的概念不但给出了属性之间的关联关系，并且同时给出了它们的关联方向。这两类知识都属于用户不很明确的概念，对于它们的可信度等客观性的度量最小阈值是可以选择的。用这两类知识对规则的分析主要是进行语法上的处理。而准确的知识是指出用户确信特定的属性之间，以明确的关联方向按确定阈值的客观感兴趣度构成关联规则。用这类知识对规则进行感兴趣度的分析，不仅考虑语法上的匹配，还要进行语义上的处理，即要进行客观感兴趣度的比较。以上的对领域知识的处理符合人类的认知结构，反映了用户模糊的概念和准确的知识一种综合。

通过用上述三类知识对关联规则进行感兴趣度分析，分别在三个层面上得到四种类型的规则：完全符合的规则、结论不符合的规则、前提不符合的规则及前提和结论均不符合的规则。然后把分析的结果通过可视化界面来与用户进行交互，用户可以根据需要和喜好对不同层次上的规则进行筛选。同时可以提醒用户以前忘掉的知识，对输入的知识进行修改。尽管需要进行可视化分析的规则也是很多的，但是通过对它们进行分类，用户每次只需分析一类规则，这是用户完全可以胜任的。

3.4 与基于查询的方法的比较

传统地，基于查询的方法用于帮助用户识别或产生感兴趣的规则[10~13]。尽管查询语言可能很不同，但是一个查询基本上定义了一种特定类型规则的集合(或对找到的规则进行限制)；去"执行"一个询问意味着去找到满足查询的所有的规则。基于查询的方法因下面两个原因而表明它是不够的。

1)很难找到真正意外的规则，只能找到那意料之中的规则，因为用户查询的仍然在他/她的已有的知识空间中。

2)用户常常不知道或不能完全指明对什么感兴趣，需要激发他/她的兴趣。基于查询的方法不能主动地完成这任务，因为它只能得到满足查询的那些规则。

通过用户输入各种类型的知识进行感兴趣度的分析技术不仅能找到象用基于查询方法得到的符合用户期望出现的规则，也能找到三种意想不到的规则。它还能帮助用户通过提醒可能忘掉了什么来为系统提供更多的知识，同时利用可视化部分可以让用户简单、快速地找到那些感兴趣的规则。

结束语 尽管长期以来人们认识感兴趣度问题是知识发现中的一个重要的内容，但是大多数的知识发现系统很少专门用来解决这一问题。实际上，目前知识发现系统重点在于从给定数据库中发现所有的模式，很少使用感兴趣的度量构件对大量的模式和规则进行筛选，导致用户无法把精力集中在真正感兴趣的知识上。目前感兴趣度的研究主要集中在对客观感兴趣度的研究上，但是因为感兴趣度的问题从本质上讲是一个主观的问题，这就需要加入背景知识和个人的喜好。它们的表示以及如何将这些主观的因素更有效地融入到所关注的模式的发现过程中，都是进一步要研究的内容。

参 考 文 献

1 Freitas A A. On rule interestingness measures. Knowledge-Based Systems, 1999, 12: 309~315
2 Symth P, Goodman R M. An information theoretic approach to rule induction from databases. IEEE Trans. Knowledge and Data Eng., 1992, 4(4): 301~316
3 程继华，郭建生，等. 挖掘所关注规则的多策略方法研究. 计算机学报，2000，23(1)：47~51
4 周欣，沙朝峰，等 兴趣度—关联规则的又一个阈值. 计算机研究与发展，2000，37(5)，629~633
5 Liu B, Hsu W. Post-analysis of learned rules. In: Proc. AAAI-96, 13th Nat'l Conf. Artificial Intelligence, Portland, Ore., 1996. 828~834
6 Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems. IEEE Trans. Knowledge and Data Eng., 1996, 8(6), 970~974
7 Piatetsky-Shapiro G, Matheus C J. The interestingness of deviations. In: Proc. AAAI'94, Workshop Knowledge Discovery in Databases 1994. 25~36
8 Liu B, Hsu W, Chen S. Using general impressions to analyse discovered classification rules. In: Proc. of the 3rd int'l Conf on Knowledge Discovery and Data Mining. 1997. 31~36
9 Liu B, Hsu W, et al. Visually aided exploration of interesting association rules. In: Proc of the 3rd Pacific-Asia Conf on Knowledge Discovery and Data Mining. 1999. 380~389
10 Klemettinen M, Mannila H, et al. Finding interesting rules from large sets of discovered association rules. CIKM-94, 1994. 401~407
11 Imielinski T, Virmani A, Abdulghani A. DataMine: application programming interface and query language for database mining. KDD-96, 1996
12 Han J, Fu Y, et al. DMQL: a data mining query language for relational databases. SIGMOD Workshop on KDD, 1996
13 Adomavicius G, Tuzhilin A. Discovery of actionable patterns in databases: the action hierarchy approach KDD-97, 1997. 111~114