基于描述逻辑的扩展预测模型标记语言 EPMML

朱小栋[1),2)] 肖芳雄[3)] 黄志球[2)] 沈国华[2)] 靳 玲[2)]

[1)](上海理工大学管理学院信息管理系 上海 200093)

[2)](南京航空航天大学信息科学与技术学院 南京 210016)

[3)](广西财经学院信息与统计学院 南宁 530003)

摘 要 预测模型标记语言PMML正被许多数据挖掘组织作为标准化的数据挖掘模型描述语言.然而,由于数据挖掘技术的不断发展,参与建立PMML的数据挖掘厂商的经验有差异,PMML本身含有的大量语言元素不可避免地带来基于PMML的数据挖掘元数据的语义不一致问题.为解决这个问题,提出了一种基于描述逻辑的扩展预测模型标记语言EPMML,详细分析了EPMML的描述逻辑基础SOIN,设计EPMML的语言元素.基于EPMML描述的数据挖掘元数据可以转化为基于SOIN的知识库,进而进行知识推理以自动发现数据挖掘元数据的内在语义不一致问题.Racer推理实例验证了EPMML语言的良构性,良好表达能力和推理有效性.

关键词 预测模型标记语言;数据挖掘;元数据;描述逻辑;知识推理

中图法分类号 TP311 **DOI号**: 10.3724/SP.J.1016.2012.01644

Description Logic Based Extended Predictive Model Markup Language EPMML

ZHU Xiao-Dong[1),2)] XIAO Fang-Xiong[3)] HUANG Zhi-Qiu[2)] SHEN Guo-Hua[2)] JIN Ling[2)]

[1)](*Department of Information Management, Business School, University of Shanghai for Science & Technology, Shanghai* 200093)

[2)](*College of Information Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016)

[3)](*School of Information and Statistics, Guangxi University of Finance and Economics, Nanning* 530003)

Abstract Predictive Model Markup Language PMML is currently used as standardized description language for data mining model by more and more DMG members. However, different experience of data mining products providers, constant development of data mining techniques, and PMML containing lots of language elements inevitably lead to inconsistency problems in PMML based data mining metadata. Considering this problem, a description logic SOIN is designed in this research. Its syntax and semantics are analyzed. An extended predictive model markup language EPMML is then proposed based on the SOIN infrastructure, the language elements are designed in detail. EPMML based data mining metadata can be transformed into knowledge base of SOIN, and then potential semantic inconsistency problems in the metadata can be automatically discovered by knowledge reasoning upon the knowledge base. Illustrations in the reasoning engineer Racer validate the well-formedness, well expressibility and reasoning efficiency of EPMML.

Keywords predictive model markup language; data mining; metadata; description logic; knowledge reasoning

---

收稿日期:2010-08-17;最终修改稿收到日期:2012-04-26.本课题得到上海市教委科研创新项目基金(12YZ103)、上海市优秀青年教师培养基金(slg10010)、教育部人文社科青年基金(12YJC870037)及上海市重点学科建设项目基金(S30504)资助.**朱小栋**,男,1981年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为数据仓库与数据挖掘、智能数据管理、电子商务.E-mail: zhuxd@nuaa.edu.cn. **肖芳雄**,男,1971年生,博士,高级工程师,主要研究方向为软件工程、云计算和电子商务. **黄志球**,男,1965年生,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据库工程、软件工程. **沈国华**,男,1976年生,博士,副教授,主要研究方向为Web服务、语义Web. **靳 玲**,女,1980年生,硕士研究生,主要研究方向为可信软件、软件度量与测试.

1 引　言

数据挖掘的标准化是目前数据挖掘技术发展亟待解决的重要问题. 提供标准化的数据挖掘元数据和 API 在数据挖掘产品的集成、交换和共享上有着核心的作用. 预测模型标记语言 PMML(Predicative Model Markup Language)是由 DMG 组织开发的用于描述数据挖掘模型的基于 XML 的标记语言. 目前,DMG 组织的许多厂商正致力于用 PMML 作为统一的标准化的数据挖掘模型描述语言. PMML 标准化了常见的数据挖掘算法的模型内容,例如,描述关联规则模型的 PMML 指定了一些标记来描述事务、项与项集以及关联规则的支持度与置信度等. PMML 使得模型的部署、发布、维护、软件包间的模型信息共享交换变得容易. 例如,用一个工具开发的模型可以通过 PMML 转换到另一个工具中用于评测.

在不同的产品和环境中交换预测模型需要对 PMML 规范有共同理解. 然而,连同增加的产品特殊扩展,PMML 包括了大量的语言元素,所以这样的理解并不尽如人意. 结果是,即使有一个详细的 PMML 规范,通过 PMML 定义的模型也可以变化不一致. 缺乏一致性降低了 PMML 的有用性,妨碍了其在数据挖掘团体中的使用. 因此,目前迫切地需要一致性的标准来提高 PMML 模型协同工作的能力,提高 PMML 作为多产品间的模型交换中介的可靠性. Pechter 给出了一种结合 XSD 验证和 XSLT验证来确保 PMML 的正确性的方法[1]. 该方法可以解决 PMML 描述的数据挖掘模型语法层面的错误. 然而,PMML 本身缺乏形式化的语义使得基于 PMML 的数据挖掘模型难以进行自动推理,并难以发现模型内在的语义冲突问题. 而这种冲突伴随着 PMML 描述的数据挖掘模型不断更新和 PMML 自身版本的不断演化显得更加的突出. 这里,为便于读者理解 PMML 语言的局限性,首先给出基于 PMML 的数据挖掘元数据的语义一致性问题和冲突问题的两个实例.

例 1.　在 PMML 语言中,用语言元素 AssociationRule 来声明一个关联规则,而不允许 AssociationRules 作为声明关联规则的语言元素. 然而,我们期望的是 AssociationRule 和 AssociationRules 都可以声明关联规则,表示关联规则类的语义,这更加符合我们的使用习惯.

例 2.　基于 PMML 的数据挖掘元数据的一致性可以区分为语法一致性和语义一致性. 经过 XSD 和 XSLT 验证通过的元数据并不能保证其没有语义冲突问题,如 PMML 中存在冗余、引用冲突等等. 图 1 给出了描述关联规则的 PMML 元数据片段,它能通过语法一致性检测,但它不满足关联规则所规定的语义. 如图中描述了一条规则“Beer→Beer”,它违背了在关联规则定义中所要求的前件和后件交集为空的语义要求.

```
〈?xml version="1.0"?〉
  〈PMML version="3.1"〉
    〈Header copyright="www.dmg.org"
      description="example model for association rules"/〉
    …
    〈AssociationModel
      functionName="associationRules"
      numberOfTransactions="4" numberOfItems="3"
      minimumSupport="0.6" minimumConfidence="0.5"
      numberOfItemsets="3" numberOfRules="2"〉
    …
〈!--Two rules satisfy the requirements--〉
    〈AssociationRule support="1.0" confidence="1.0"
      antecedent="Beer" consequent="Diaper"/〉
    〈AssociationRule support="1.0" confidence="1.0"
      antecedent="Beer" consequent="Beer"/〉
    〈/AssociationModel〉
  〈/PMML〉
```

图 1　PMML 元数据中的语义不一致问题

描述逻辑是一阶谓词逻辑的可判定子集,它以结构化和易理解的方式来表示领域知识,目前已成为本体语言如 OWL[2] 的逻辑基础. 以描述逻辑作为基础,本文提出一种扩展的预测模型标记语言 EPMML(Extended Predictive Model Markup Language). 该思路源于用 EPMML 描述的数据挖掘模型可以被转化为基于描述逻辑的知识库,进而基于描述逻辑的知识推理可以自动发现数据挖掘元数据的冲突问题. 理论和实验表明,EPMML 能够作为数据挖掘模型的描述工具,并且 EPMML 具有良好的形式化语义和支持自动推理的能力.

本文第 2 节介绍相关工作;第 3 节设计一种描述逻辑家族的形式逻辑 SOIN,给出它的语法和语义;第 4 节以 SOIN 作为基础,给出支持自动推理的预测模型标记语言 EPMML 的详细设计,分析 EPMML 的语言要素;第 5 节分析 SOIN 的推理复杂性,提出基于 EPMML 的数据挖掘元数据一致性检测框架,并给出 EPMML 支持自动推理发现冲突的示例;最后是本文的总结.

2 相关工作

随着数据挖掘技术的发展,数据挖掘的标准化成为日益关切的问题,数据挖掘元数据在数据挖掘的标准化过程中发挥着越来越重要的作用.许多面向数据挖掘元数据的工业标准被提出,除了预测模型标记语言 PMML,还包括跨行业数据挖掘标准流程 CRISP-DM 和公共仓库元模型 CWM①.CRISP-DM 提供了一个描述整个数据挖掘生命周期的过程标准,目前已成为开发数据挖掘项目的过程的标准方法,但它没有为数据挖掘的元数据制定精确的规范.CWM 是由 OMG 组织的 CWM 工作组负责开发、并由 OMG 采纳的一种使用共享元数据的集成数据仓库和业务分析工具的开放式行业标准.CWM 主要关注商务智能领域,如 OLAP、数据挖掘中元数据的定义.提供 CWM 的目的是为了解决元数据的管理和数据仓库的集成问题,这样不同的应用程序能够在不同的环境中集成.CWM 规范中详细地定义了数据挖掘元模型.然而,参与建立元数据的数据挖掘厂商的不同经验和描述数据的不同角度以及数据挖掘技术的不断更新,不可避免地带来基于 CWM 元数据的冲突问题.Zhu 等人[3]提出了一种基于描述逻辑的策略进行基于 CWM 的数据挖掘元模型和元数据的冲突检测机制,解决了基于 CWM 的自然语言和图形化特点缺乏精确的语义的问题,取得了较好的效果.用现有的数据挖掘元模型来构建面向数据挖掘过程的应用模型的工作包括:Zubcoff 等人[4]利用 CWM 提供的丰富语义信息构建用于数据挖掘分类分析的挖掘元模型,Castellano 等人[5]利用 CWM 元模型构建数据挖掘过程的体系结构,Chaves 等人[6]设计了一种基于 PMML 的评测引擎 Augustus,可以用于进行数据准备和模型分割.

Berners-Lee 等人[7]在 2000 年提出了语义 Web 的概念,其目的是让 Web 上的信息能够被机器理解,从而实现 Web 信息的自动处理.语义 Web 的支撑技术建立在一系列技术标准和规范之上,其中 RDF 和 OWL 是最基本的技术标准.RDF 是一种元数据的数据模型,在该模型下,对资源的描述采用主体、谓词和客体的三元组形式$\{sub, pred, obj\}$陈述②.RDF 在 XML 基础上提供了一定的语义描述能力,但它作为本体语言,其语义描述能力还很有限.DAML+OIL、OWL 是由 RDF(S)扩展的网络本体语言,目前 OWL 已成为 W3C 推荐的网络本体语言.OWL 具有明确的逻辑基础即描述逻辑,它是用 XML 语法、RDF 模型定义的描述逻辑语言.

借鉴 RDF(S)和其它语义 Web 本体语言 OWL、OIL、DAML+OIL 等的设计思路,我们提出基于描述逻辑设计数据挖掘建模语言 EPMML 的理念.这个思路如下:以预测模型标记语言 PMML 为基础,在 PMML 上层扩充 RDF 和 RDFS 以提供数据挖掘领域的资源描述框架,再在 RDF(S)的上层将 PMML 扩充为真正具有语义描述能力的数据挖掘领域的语义本体语言.这种数据挖掘领域的语义本体语言需要明确以一种描述逻辑作为其逻辑基础.在下面第 3 节中,我们首先给出一种合适的描述逻辑 SOIN 作为数据挖掘领域语义本体语言的逻辑基础.据我们所知,目前国内外还没有针对基于 PMML 的数据挖掘元数据应用描述逻辑进行语义扩展和冲突检测的研究.

3 描述逻辑 SOIN

描述逻辑具有正式的基于逻辑的语义和很强的表达能力.Baader 等人[8]指出描述逻辑为语义网提供了必要的逻辑基础.基本的描述逻辑 ALC 的元素是由概念(一元谓词)、关系(二元谓词)、个体(常元)以及在它们上的交$\sqcap$、并$\sqcup$、补$\neg$、存在约束$\exists$、全称约束$\forall$等算子构成.增加 ALC 的构造算子,或者采用不同的构造子组合得到的描述逻辑拥有不同的表达能力和推理复杂性.

然而,知识表示语言的表达能力越强,相应推理问题的复杂性越高.例如,OWL DL 的语义表达能力很强,然而,其语义逻辑基础 SHOIN(D)的推理是 NEXPTIME-complete 问题,这使得其不适合作为数据挖掘元数据的语义描述语言[9-10].针对 PMML 本身的特点,本节设计一种描述逻辑家族的形式逻辑 SOIN 作为 EPMML 语言的逻辑推理基础.这里,S 表示在 ALC 基础上增加关系的传递性,即 S 表示 ALC 的演化 $ALCR^+$.O 表示允许枚举,I 表示允许关系逆,N 表示允许数量约束.

定义 1. SOIN 语法.用 A 和 P 分别表示原子概念和原子关系,符号::=表示定义.SOIN 上的概

① OMG, Object Management Group. Common Warehouse Metamodel Specification, Version 1.1 [EB/OL]. (2003-03-01) [2011-12-12]. http://www.omg.org/spec/CWM/1.1/

② W3C. Resource Description Framework(RDF): Concepts and Abstract Syntax [R/OL]. (2003-02-10)[2011-12-12]. http://www.w3.org/TR/rdf-concepts/

念 C 和关系 R 递归定义如下：

$$C::=\top_1|\bot_1|A|\neg C|C_1\sqcap C_2|C_1\sqcup C_2|\exists R.C|\forall R.C|(\geq nR)|(\leq nR)|\{a_1,\cdots,a_n\}|(\geq nR.C)|(\leq nR.C),$$

$$R::=\top_2|\bot_2|P|\neg R|R_1\sqcap R_2|R_1\sqcup R_2|R^-.$$

定义 2.　SOIN 语义. SOIN 的解释是一个二元对 $I=(\Delta^I,\cdot^I)$，Δ^I 是论域的非空集合，$\cdot^I$ 是解释函数. 令 **card** 表示一个集合的基数，具体的语义如下所示：

(1) $(\top_1)^I=\Delta^I$；

(2) $(\bot_1)^I=\varnothing$；

(3) $(\neg C)^I=\Delta^I\backslash C^I$；

(4) $(C_1\sqcap C_2)^I=C_1^I\bigcap C_2^I$；

(5) $(C_1\sqcup C_2)^I=C_1^I\bigcup C_2^I$；

(6) $(\exists R.C)^I=\{a\in\Delta^I\mid\exists b((a,b)\in R^I\wedge b\in C^I)\}$；

(7) $(\forall R.C)^I=\{a\in\Delta^I\mid\forall b((a,b)\in R^I\rightarrow b\in C^I)\}$；

(8) $(\geq nR)^I=\{a\in\Delta^I\mid\mathbf{card}\{b\in\Delta^I\mid(a,b)\in R^I\}\geq n\}$；

(9) $(\leq nR)^I=\{a\in\Delta^I\mid\mathbf{card}\{b\in\Delta^I\mid(a,b)\in R^I\}\leq n\}$；

(10) $\{a_1,\cdots,a_n\}^I=\{a_1^I,\cdots,a_n^I\}$；

(11) $(\top_2)^I=\Delta^I\times\Delta^I$；

(12) $(\bot_2)^I=\varnothing\times\varnothing$；

(13) $(\neg R)^I=(\Delta^I\times\Delta^I)\backslash R^I$；

(14) $(R_1\sqcap R_2)^I=R_1^I\bigcap R_2^I$；

(15) $(R_1\sqcup R_2)^I=R_1^I\bigcup R_2^I$；

(16) $(R^-)^I=\{(b,a)\in\Delta^I\times\Delta^I|(a,b)\in R\}$.

定义 3.　SOIN 的知识库 K 是一个三元组：

$$K=\{Tbox,Rbox,Abox\}.$$

$Tbox$ 是描述领域结构的公理的集合，含有引入概念名称的公理和声明包含关系的公理，分别记为 $A\equiv C$ 和 $A\sqsubseteq C$. $Rbox$ 是描述关系间等价和包含的语法结构的公理集合，分别用 $R\equiv S$ 和 $R\sqsubseteq S$ 的形式来描述关系定义和关系包含的集合. $Abox$ 是描述具体情形的公理的集合，包含概念断言和关系断言. 概念断言是表示一个对象是否属于某个概念，用 $a:C$ 的形式描述，关系断言是表示两个对象是否满足一定的关系，用 $\langle a,b\rangle:R$ 的形式表示.

由定义 3 可得如下定理 1.

定理 1.　令 iff 表示当且仅当，以下 3 个命题成立.

(1) 一个解释 I 满足：$a:C$ iff $a^I\in C^I$，$\langle a,b\rangle:R$ iff $\langle a^I,b^I\rangle\in R^I$.

(2) 一个解释 I 满足 $Abox$ $\mathcal{A}$ iff 它满足 $\mathcal{A}$ 中的每个公理，记为 $I\models\mathcal{A}$.

(3) 一个解释 I 满足 SOIN 知识库 $K=\{Tbox\ \mathcal{T},Rbox\ \mathcal{P},Abox\ \mathcal{A}\}$，iff 它同时满足 $\mathcal{T},\mathcal{P}$ 和 $\mathcal{A}$，记为 $I\models K$.

4　扩展预测模型标记语言 EPMML

EPMML 的设计理念如图 2 所示. 该体系结构是一个层次化的结构. Unicode 和 URI 是国际统一化字符集和资源标识手段. EPMML、命名空间和 EPMML schema 定义了 EPMML 的语法层面的互操作的标准. RDF(S)描述和定义 EPMML 的资源. SOIN 是 EPMML 逻辑基础，并提供严格的可判定性的形式化机制，支持自动推理. EPMML 可以为数据挖掘应用提供标准化的具有语义的描述，该描述可称为基于 EPMML 的数据挖掘元数据. EPMML 的元素是描述数据挖掘模型的元类、属性、元类的实例以及这些实例之间的关系. 下面各小节按照这些描述对象详细地分析 EPMML 的语言元素.

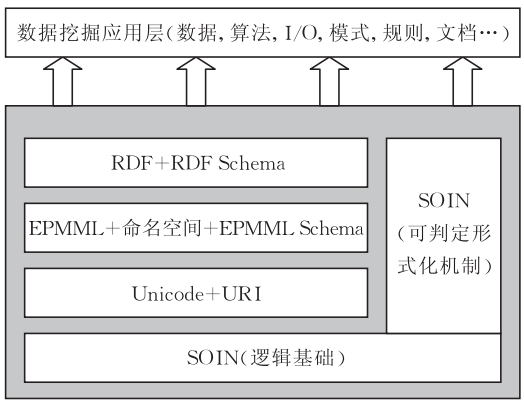

图 2　EPMML 的体系结构

4.1　EPMML 元类

EPMML 元类由元类名称和一个限制列表构成. 例如：

```
〈epmml:Class rdf:ID="AssociationRules"/〉
〈epmml:Class rdf:ID="Itemset"/〉
〈epmml:Class rdf:ID="Item"〉
  〈rdfs:subClassOf rdf:resource="#Itemset"/〉
〈/epmml:Class〉
```

EPMML 元类的逻辑基础是 SOIN 中的概念，包括原子概念和简单复合概念. 在 4.2 节中介绍的 EPMML 复杂元类的逻辑基础是 SOIN 中的复杂复合概念.

4.2　EPMML 复杂元类

在 EPMML 中，复合概念通过设计元类的交、

并、补等来构造，它们的 SOIN 基础是概念的交、并、补. 用 epmml:intersectionOf、epmml:unionOf 和 epmml:complementOf 来声明. 例如：

```
〈epmml:Class rdf:ID="Abnormal"〉
  〈epmml:complementOf rdf:resource="#Normal"/〉
〈/epmml:Class〉
```

在数据挖掘模型中，有一些概念可以通过枚举实例来描述. 这种描述枚举概念的枚举元类是一种特殊的构造子，用 epmml:oneOf 来声明. 例如：

```
〈epmml:Class rdf:ID="Wether"〉
  〈epmml:oneOf rdf:parseType="Collection"〉
    〈epmml:Thing rdf:about="#Fine"/〉
    〈epmml:Thing rdf:about="#Cloudy"/〉
    …
  〈/epmml:oneOf〉
〈/epmml:Class〉
```

4.3 EPMML 属性

一个 EPMML 属性是一个二元关系，在 EPMML 中，属性区分为对象属性和数据属性. 对象属性描述了元类的实例之间的关系，用〈epmml:ObjectProperty〉宣称对象属性，用〈rdfs:domain〉和〈rdfs:range〉指出该对象属性的定义域和作用域. 例如：

```
〈epmml:ObjectProperty rdf:ID="HasAntecedent"〉
  〈rdfs:domain rdf:resource="#AssociationRules"/〉
  〈rdfs:range rdf:resource="#Antecedent"/〉
〈/epmml:ObjectProperty〉
```

区别于对象类型属性，数据类型属性的值域是数据类型. 在 EPMML 中，使用 PMML3.2.0 版本的 xsd 文件中定义的数据类型，如 xs:string，xs:integer. 用〈epmml:DataTypeProperty〉宣称数据属性，用〈rdfs:domain〉和〈rdfs:range〉指出该数据属性的定义域和值域. 例如：

```
〈epmml:DatatypeProperty rdf:ID="HasSupport"〉
  〈rdfs:domain rdf:resource="#AssociationRules"/〉
  〈rdfs:range rdf:resource="http://www.dmg.org/
v3-2/pmml-3-2.xsd#PROB-NUMBER"/〉
〈/epmml:DatatypeProperty〉
```

这里 PROB-NUMBER 是 pmml-3-2.xsd 中定义的 0～1 之间的小数类型.

EPMML 属性的逻辑基础是 SOIN 中的关系. 为了提高 EPMML 的语义表达能力，SOIN 包含了关系传递和关系逆的构造算子，这两个算子是 EPMML 的属性约束的逻辑基础，参见 4.5 节.

4.4 EPMML 个体

EPMML 除了描述数据挖掘模型的元类和属性之外，需要描述数据挖掘模型中具体的个体以及个体之间的关系. 用〈epmml:Thing〉来宣称一个个体，用 rdf:type 来指明该个体所属的元类. 例如，在描述一个关联规则模型中，指定 Cracker 是一个项 item 元类的实例：

```
〈epmml:Thing rdf:ID="Cracker"〉
  〈rdf:type rdf:resource="#Item"〉
〈/epmml:Thing〉
```

与 PMML 不同，EPMML 中的个体不是语言元素，这大量地约简了 PMML 的语言元素复杂性. 为了减少推理的复杂性，在 EPMML 中，不允许出现一个资源是元类并且是个体.

4.5 EPMML 属性约束

属性是特殊的二元关系，根据二元关系的理论，属性可以具有自反性、对称性、传递性以及函数性等特性，并且属性可以有逆属性. 然而在描述逻辑中，增加属性的特性，必然会增加逻辑推理的复杂性，甚至导致推理不可判定. 根据描述数据挖掘模型的 PMML 特点，在 EPMML 中，不增加属性的自反性和函数性，但允许属性具有传递性，并允许与其它属性互逆. 例如传递属性 Is_Part_Of. 在 EPMML 中，用 epmml:TransitiveProperty 来声明属性具有传递性. 例如：

```
〈epmml:ObjectProperty rdf:ID="Is_Part_Of"〉
  〈rdf:type rdf:resource="&epmml:TransitiveProperty"/〉
〈/epmml:ObjectProperty〉
```

用 epmml:InverseOf 声明属性的逆属性. 例如：

```
〈epmml:ObjectProperty rdf:ID="BeAntecedentOf"〉
  〈epmml:InverseOf rdf:resource="HasAntecedent"/〉
〈/epmml:ObjectProperty〉
```

在 PMML 中，对数据挖掘模型的描述中，一些属性不仅指明了定义域和作用域，而且有明确的数量限制. 例如，为了描述关联规则的支持度和置信度都是 0～1 之间的小数，在关联规则的 PMML 模型中，需要添加若干语言元素，而在 EPMML 中，不需要添加语言元素，并且可以对其赋予语义，告诉机器支持度和置信度的数量是 0～1 之间的小数. 在 EPMML 中，用 epmml:someValuesFrom，epmml:allValuesFrom 来声明属性值域约束，用 epmml:minCardinality，epmml:maxCardinality 来声明属性的基数约束. 它们的 SOIN 逻辑基础分别是 $\exists R.C$，$\forall R.C$，$(\geq n\,R)$ 和 $(\leq n\,R)$ 概念描述. 例如：

```
〈epmml:Restriction〉
  〈epmml:onProperty rdf:resource="#hasWether"/〉
  〈epmml:allValueFrom rdf:resource="#Wether"/〉
〈/epmml:Restrition〉
```

```
〈epmml:Restriction〉
  〈epmml:onProperty rdf: resource="#hasSupport"/〉
  〈epmml:minCardinality rdf:datatype="http://www.
dmg.org/v3-2/pmml-3-2.xsd#PROB-NUMBER"〉0.0
〈/epmml:minCardinality〉
  〈epmml:maxCardinality rdf:datatype="http://www.
dmg.org/v3-2/pmml-3-2.xsd#PROB-NUMBER"〉1.0
〈/epmml:maxCardinality〉
〈/epmml:Restriction〉
```

4.6 EPMML 辅助语言元素

在 EPMML 中，兼容了 XML 的注释等语言元素. 为了增加语义可理解性，减少推理的复杂性，版权、版本、命名空间等 EPMML 的辅助语言元素用 EPMML 数据属性来描述. 例如：

```
〈epmml:DatatypeProperty rdf:ID="HasCopyRight"〉
  〈rdfs:domain rdf:resource="#EPMML"/〉
  〈rdfs:range rdf:resource="http://www.nuaa.edu.cn"/〉
〈/epmml:DatatypeProperty〉
〈epmml:DatatypeProperty rdf:ID="HasVersion"〉
  〈rdfs:domain rdf:resource="#EPMML"/〉
  〈rdfs:range rdf:resource="1.0.0"/〉
〈/epmml:DatatypeProperty〉
```

在设计 EPMML 中，显然一个资源只允许以一种语言元素的形式出现. 例如，设定一个资源是个体，则不允许其是元类；反之也是如此. 同时，同一个命名空间下的资源不允许重名，但是不在一个命名空间下的资源可以允许重名，但必须加以引用. 基于 SOIN 的 EPMML 具有严格的形式化语义，这为 EPMML 支持自动推理提供了完备的形式逻辑基础.

5　基于 SOIN 的自动推理演示示例

在这一节里，我们首先分析描述逻辑 SOIN 的推理复杂性，然后设计了基于 EPMML 的数据挖掘元数据一致性检测框架，并通过示例验证 EPMML 支持自动推理的正确性和有效性.

5.1 SOIN 的推理复杂性

定理 2.　SOIN 上的推理可以规约到 SOIN 的可满足性问题.

证明.　SOIN 的推理问题包括 5 类.

(1) 知识库的可满足性：给定一个 SOIN 知识库 K，如果存在一个解释 I，使得 $I \models K$.

(2) 概念的可满足性：关于 $TBox\ T$，如果概念 C 非空，即存在一个解释 I，其中 $I \models T$，满足 $C^I \neq \varnothing$.

(3) 概念的包含关系：关于 $TBox\ T$，如果概念 C_1 包含概念 C_2，即对任意解释 I，其中 $I \models T$，满足 $C_2^I \subseteq C_1^I$，记作 $T \models C_2 \sqsubseteq C_1$.

(4) 实例检测：关于 SOIN 的知识库 K，如果个体名 a 属于概念 C，即对任意解释 I，其中 $I \models K$，满足 $a^I \subseteq C^I$，记作 $K \models C(a)$.

(5) 查询检索：关于 SOIN 知识库 K，找到概念 C 的所有个体名 a，使得 $K \models C(a)$.

对于两个概念 C 和 D，有 C 不可满足$\equiv C$ 包含于 $\perp$；C 和 D 相等$\equiv C$ 包含于 D，且 D 包含于 C；C 和 D 相离$\equiv C \sqcap D$ 包含于 $\perp$. 根据实例检测的含义，实例检测可以规约到 $a^I \subseteq C^I$ 是不可满足的. 查询检索可以通过实例检测实现. 所以 SOIN 上的推理问题都可以规约到包含关系的判断，如果存在判断包含关系的算法，必然存在解决其它推理问题的算法，且判断包含关系的复杂度是其它推理问题复杂度的上界.

进一步地，两个概念 C 和 D，有 C 包含于 $D \equiv C \sqcap \neg D$ 是不可满足的；C 和 D 相等$\equiv C \sqcap \neg D$，$\neg C \sqcap D$ 都不可满足；C 和 D 相离$\equiv C \sqcap D$ 不可满足，所以 SOIN 上的推理问题都可以规约到可满足性的问题，如果存在判断可满足性的算法，必然存在解决其它推理问题的算法.

根据二元关系理论，关系是概念的笛卡尔积的子集. 因此，自然地可将上面概念的可满足性问题演化到 $Rbox$ 上的关系可满足性问题.

综上所述，SOIN 上的推理可以规约到 SOIN 的可满足性问题.　　证毕.

定理 3.　SOIN 的推理问题可以规约到 $Abox$ 上的一致性检验问题.

证明.　概念 C 是可满足的当且仅当$\{C(a)\}$是一致性的，这表明知识库的可满足性、概念的可满足性以及概念的包含关系可以规约到 $Abox$ 的一致性检验. $A \models C(a) \equiv A \sqcup \neg C(a)$ 是不一致的，这表明实例检测可以规约到 $Abox$ 一致性检验，查询检索可以通过实例检测实现，所以查询也可以通过一致性检验实现.　　证毕.

定理 4.　SOIN 是可判定的，并且是 EXPTIME-Complete 问题.

证明.　SOIN 在 ALC 上增加关系传递、关系逆、绝对数量约束算子，枚举的描述逻辑推理是可判定的，且是 EXPTIME-Complete 问题[11-12]. SOIN 可以映射为 SHOIQ 的一个子集，并且 SOIN 在基本的 ALC 描述逻辑基础上仅增加了具体域、关系传递、关系逆和绝对数量约束. 而 SHOIQ 的 Tableaux 可判定推理算法是 EXPTIME-Complete 问题[9,13]，

所以 SOIN 是可判定的，且是 EXPTIME-Complete 问题. 证毕.

表 1 给出了 EPMML 语言与其它标记语言的比较. 作为数据挖掘领域的本体语言，我们并不需要 OWL 这样表达能力过强、而推理复杂性大的语言.

表 1 EPMML 与其它语言的比较

	语义表达能力	逻辑基础	语法基础	推理可判定性	推理复杂性
XML	无	无	XML(S)	无	无
PMML	无	无	XML(S)	不可判定	无
RDF	弱	三元组(谓词逻辑)	XML(S)	可判定	EXPTIME- Complete
EPMML	中等	描述逻辑 SOIN	XML(S)+RDF(S)	可判定	EXPTIME- Complete
OWL Lite	强	描述逻辑 SHIF(D)	XML(S)+RDF(S)	可判定	NEXPTIME- Complete
OWL DL	最强	描述逻辑 SHOIN(D)	XML(S)+RDF(S)	可判定	NEXPTIME- Complete

表 1 中 EXPTIME-Complete 表示确定型图灵机上指数时间完全问题，NEXPTIME-Complete 表示非确定型图灵机上指数时间完全问题. 这两类问题的复杂性层次关系是 EXPTIME-Complete $\subseteq$ NEXPTIME-Complete.

数据挖掘是从大量数据中发现潜在的能为决策者服务的规则和知识的过程. 描述数据挖掘应用层的数据、算法、规则、模式的 EPMML 元数据因此庞大且多变. 这些特征不容许使用表达能力强而推理能力复杂的描述逻辑如 SHOIQ 和 SHOIN(D)作为语义表达基础. 相比 OWL，EPMML 的推理复杂性大大减小.

总的来说，作为面向数据挖掘领域的预测模型标记语言 EPMML，它不能使用领域无关的语义本体语言 OWL 来取代，而需要针对数据挖掘领域的特点，有针对地建立描述逻辑 SOIN，在此基础上设计合适的具有语义描述功能的预测模型标记语言. 正是因为描述逻辑是 EPMML 语言的逻辑基础，EPMML 是一种具有逻辑推理能力的数据挖掘领域相关的语义本体语言. 进一步地，我们明确描述逻辑、本体和 EPMML 语言三者的关系是：描述逻辑是本体的逻辑基础、EPMML 语言是本体的表现形式，这里本体是指数据挖掘领域相关的本体.

5.2 EPMML 元数据一致性检测框架

图 3 给出了基于 EPMML 的数据挖掘元数据的一致性检测框架. 图中箭头表示组件间的流程方向. 框架的初始状态是用户输入的基于 EPMML 的数据挖掘元数据. XML 验证过程是检测 EPMML 元数据语法层的一致性，如果结果不合法，则元数据一致性检测任务终止，并向用户返回不合法的错误信息，以供修正 EPMML 元数据语法错误. 经过 XML 验证通过的合法 EPMML 元数据，首先经过映射过程转化为基于 SOIN 的数据挖掘元数据知识库. 然后进入知识推理过程，这一过程的推理原理使用描述逻辑的 Tableaux 算法. 经过知识推理的结果返回用户交互界面. 如果存在冲突信息，则对合法的 EPMML 元数据进行语义修正；否则任务结束.

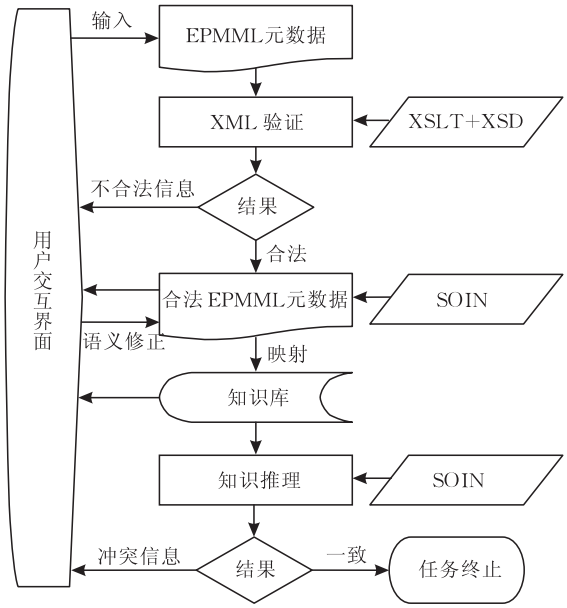

图 3 EPMML 元数据一致性检测框架

5.3 一致性检测示例

我们在推理示例中选择 RacerPro1.90 作为推理工具(http://www.racer-systems.com)，这是考虑到推理引擎 Racer 的 Tableaux 算法是可靠完备的，并且目前 RacerPro1.90 具备了描述逻辑知识库 *Tbox*，*Abox* 和 *Rbox* 的建立界面，且查询推理语言 RQL 具有良好的表达能力. 我们同时选择 Protégé 作为数据挖掘元数据的构建工具，然后调用 Racer 推理引擎进行知识推理，以检测数据挖掘元数据的冲突问题. 表 2 给出了 SOIN 语法、EPMML 语法和推理工具 RacerPro 语法之间的映射关系.

(1)语义一致性实例. 在 PMML 语言中，用语言元素 AssociationRule 来声明一个关联规则，则不允许使用 AssociationRules 的语言元素. 然而，在描

述关联规则的 EPMML 元数据中，我们允许 AssociationRule 和 AssociationRules 都表示关联规则元类，这更加符合我们的使用习惯. 这个语义匹配可以通过增加元类的匹配来实现. 下面的定义可以实现该元类语义的匹配，保证该语义使用的一致性.

```
〈epmml:Class rdf:ID="AssociationRules"〉
  〈epmml:equivalentClass rdf:resource="#AssociationRule"/〉
〈/epmml:Class〉
```

表 2　SOIN 语法与 RacerPro 语法的映射

SOIN 语法	EPMML 语法	RacerPro 语法
$\top_1$	epmml:Thing	Top-Concept
$\bot_1$	epmml:Nothing	Nothing
$\neg C$	epmml:comlementOf	not C
$C_1 \sqcap C_2$	epmml:intersectionOf	and $C_1 .. C_2$
$C_1 \sqcap C_2$	epmml:unionOf	or $C_1 .. C_2$
$\exists R.C$	restriction(R someValueFrom (C))	some R C
$\forall R.C$	restriction(R allValueFrom (C))	all R C
$\{a_1, \cdots, a_n\}$	epmml:oneOf	one-of $[a_1, \cdots, a_n]$
$\geq n\, R$	restriction(R minCardinality (n))	at-least n R
$\leq n\, R$	restriction(R maxCardinality (n))	at-most n R
R^-	epmml:InverseProperty	inverse[R]

(2) 冲突检测实例. 在描述关联规则的 EPMML 元数据中，经过 XSD 和 XSLT 验证通过的元数据并不能保证其没有冲突问题，如 PMML 中存在冗余、引用冲突等等. 下面给出描述关联规则的 EPMML 元数据中描述一条关联规则的 EPMML 片断，但其违背了关联规则前件和后件交集为空集的语义要求.

```
〈AssociationRules rdf:ID="AssociationRules_1"〉
  〈HasAntecedent〉
    〈Item rdf:ID="Beer"〉
      〈BeAntecedentOf rdf:resource="#AssociationRules_1"/〉
    〈/Item〉
  〈/HasAntecedent〉
  〈HasConfidence rdf:datatype="http://www.dmg.org/
      v3-2/pmml-3-2.xsd#PROB-NUMBER"〉0.8
  〈/HasConfidence〉
  〈HasSupport rdf:datatype="http://www.dmg.org/
      v3-2/pmml-3-2.xsd#PROB-NUMBER"〉0.2
  〈/HasSupport〉
  〈HasSubsequent〉
    〈Item rdf:ID="Beer"〉
      〈BeSubsequentOf rdf:resource="#AssociationRules_1"/〉
    〈/Item〉
  〈/HasSubsequent〉
  〈HasSubsequent〉
    〈Item rdf:ID="Diaper"〉
      〈BeSubsequentOf rdf:resource="#AssociationRules_1"/〉
    〈/Item〉
  〈/HasSubsequent〉
〈/AssociationRules〉
```

关联规则中要求规则的前件项集和后件项集的交集是空集，然而，在传统的 PMML 元数据中，不能标识这样的语义. 在 EPMML 中，我们可以声明在一个规则中，前件项集和后件项集是分离的，即没有交集. 如果元数据中出现了交集，则发生元数据冲突. 首先将 EPMML 元数据映射为描述逻辑知识库. 图 4 是关联规则模型的描述逻辑知识库的 TBox 层，图中$\sqsubseteq$表示概念之间的包含关系，$\equiv$表示概念之间的等价关系.

$$
\begin{aligned}
&\text{Item} \sqsubseteq \text{Itemset}\\
&\text{Subsequent} \sqsubseteq \text{Itemset}\\
&\text{Antecedent} \sqsubseteq \text{Itemset}\\
&\text{AssociationRule} \equiv \text{AssociationRules}\\
&\text{AssociationRule} \sqsubseteq \text{EPMML}\\
&\text{AssociationRules} \sqsubseteq \text{EPMML}\\
&\text{Itemset} \sqsubseteq \text{EPMML}
\end{aligned}
$$

图 4　描述逻辑表示的关联规则模型知识库

然后用 RacerPro 的查询推理语句 nRQL 可以发现上述冲突.

```
;;========A-box Reasoning========
;;Check the consistency of and Abox w.r.t. a Tbox
(abox-consistenct? EPMML-data-mining-metadata)
;; Retrieve individuals that satisfy certain conditions
(Retrieve (?x)(?Antecedent)
```

我们得到如下的冲突信息：

```
Error:Abox EPMML-data-mining-metadata is incoherent.
```

这是因为 Antecedent 和 Subsequent 两个类是分离的. 个体 Beer 不能同时属于 Antecedent 和 Subsequent.

根据发现的冲突，我们可以进一步找出冲突的原因，并修复关联规则模型的元数据. 上面的元数据中，去掉后件中的项 Beer 描述，重新运行查询推理语句，冲突解决.

此外，基于 EPMML 的数据挖掘元数据，还可以进一步得到直观的数据挖掘模型的语义图. 图 5 是在关联规则知识库上作出的基于 EPMML 的关联规则模型的元类之间的语义图. 图中箭头标识元类与元类之间的包含关系 is_a.

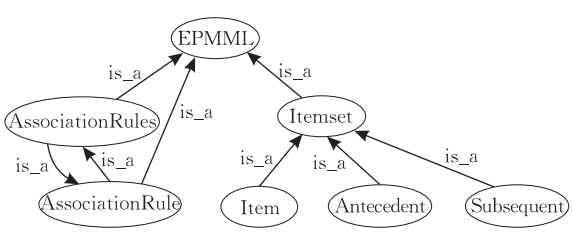

图 5　关联规则模型的语义图

6 EPMML 的应用

目前,笔者探索将 EPMML 语言应用于数据挖掘领域的系统建模,初步的效果是令人满意的.

如何构建快速、高效和智能的数据流挖掘系统,实现数据流挖掘算法的动态灵活扩展、数据的透明集成、挖掘结果模式的迭代精化,是当前数据流挖掘研究的一个焦点问题.软件工程已推进企业进入软件"工业化"生产时代,不断采用构件技术是未来软件生产力提高的主要来源[14].数据和算法是数据挖掘不可或缺的两个组成部分.在面向构件的软件体系结构中,我们将数据组件和算法组件作为数据流挖掘系统的两个相辅相成的构件.本节综合分析基于 EPMML 语言的数据流挖掘系统建模的作用.图 6 描述了 EPMML 元数据在数据流挖掘系统的作用,图中箭头表示了系统中的 EPMML 元数据流向.关于 EPMML 怎样进行知识表示和推理以及怎样应用到系统的数据管理组件和算法管理组件,读者可以参考文献[15].

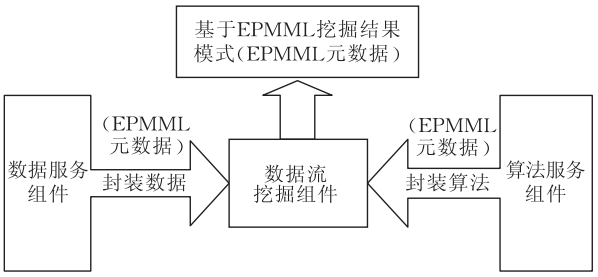

图 6 数据流挖掘系统内的 EPMML 元数据

(1)挖掘结果模式的表示与推理.由数据挖掘组件产生的结果模式通过 EPMML 语言进行描述和及时部署,以方便用户获取以及与其它应用程序共享、交换和集成.EPMML 描述的结果模式不仅支持良结构化的知识表示,而且支持模式的推理,这样便于发现模式的内部语义不一致性问题以及进行模式的迭代更新和维护.

(2)封装数据服务组件中的数据流资源,实现数据流资源的透明集成.数据流挖掘系统中数据服务组件的元数据收集模块收集当前待处理的数据流资源上下文参数,将这些参数用 EPMML 语言进行描述形成 EPMML 元数据,以供数据服务组件中的数据语义服务层来完成数据注册服务,以及数据管理子组件对数据流资源的持续地快速访问.

(3)封装算法服务组件中的算法资源,实现算法资源的动态扩展.数据流挖掘系统中算法服务组件将算法提供者提供的各种数据流挖掘算法封装为 EPMML 语言描述的算法服务,对外提供算法访问接口.当用户有挖掘任务时,将访问服务的命令和数据流资源上下文参数一起发送到算法管理子组件中,然后由领域适配模块解析数据流的上下文参数,再由服务发现模块自动地寻找最适合的数据流挖掘算法或算法组合,由服务调用组合模块组合相关算法,执行数据挖掘任务.

此外,EPMML 语言在系统之间起着元数据交换的作用.OMG 组织通过制定统一建模语言 UML、元对象设施 MOF、XML 元数据交换 XMI 和公共仓库元模型 CWM 等标准,来实现模型驱动架构 MDA 的蓝图,OMG 使用可扩展标记语言 XML 作为 CWM 元数据生成交换格式的规范.

扩展预测模型标记语言 EPMML 是基于 XML 的数据挖掘内容描述语言,所以本质上说,EPMML 是一种 XML 语言.图 7 描述了 EPMML 作为数据流挖掘系统间的元数据交换语言连接的两个数据流挖掘系统.图中数据流挖掘系统由数据流服务层、算法服务层、挖掘层和用户界面层 4 个层次组成,EPMML API 函数提供了立刻访问 EPMML 元数据的方法,EPMML 解释器提供对 EPMML 元数据的解释,EPMML 推理器提供对 EPMML 元数据的知识推理.数据流挖掘系统之间的元数据交换采用 EPMML 语言作为交换规范.

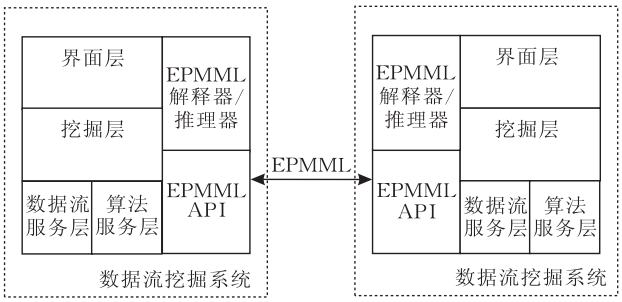

图 7 数据流挖掘系统间的 EPMML 元数据交换

7 结 论

我们不仅需要数据挖掘能够为用户挖掘出数据中潜在的规则和模式,并利用这些模式进行预测,而且,我们希望这些规则和模式能够方便地与其它应用程序共享、交换和集成.要设计数据挖掘领域的语义描述功能的标记语言,不能使用复杂的语义网络本体语言 OWL 来取代.EPMML 语言是在 PMML 语言的基础上增加了语义描述,兼容了 PMML 语言的良结构化特点和具有的数据挖掘模型描述能

力；同时，约简了 PMML 语言元素的复杂性并扩展了 PMML 语言众多语言元素之间的语义. SOIN 是 EPMML 能够支持自动推理的逻辑基础.

EPMML 在对 PMML 的扩展时注意两个要点：(1) 明确了向 PMML 增加语义描述是为了使得 PMML 语言描述的数据挖掘元数据能够支持自动推理，以自动发现这种数据挖掘模型内在的语义冲突问题. (2) 针对 PMML 的特点，向 PMML 增加语义需要确保增加的语义表达能力与逻辑基础 SOIN 的逻辑推理能力之间的权衡. 目前，我们基于 EPMML 的元数据描述和自动推理取得的结果是令人满意的，并初步应用 EPMML 语言到数据挖掘领域的系统建模，本文成果有助于增强数据挖掘元数据标准的稳定性，保障数据挖掘元数据集成的可靠性. 进一步的工作将包括完善 EPMML 对数据挖掘模型的描述功能，以及根据自动推理发现的不一致性问题进行数据挖掘元数据自动修正的研究.

参 考 文 献

[1] Pechter R. Conformance standard for the predictive model markup language//Proceedings of the 4th Workshop on Data Mining Standards, Services and Platforms (DM-SSP'06). associated with 12th ACM SIGMOD International Conference on Knowledge Discovery & Data Mining (KDD'06). Philadelphia, Pennsylvania, USA, 2006: 6-13

[2] Kalyanpur A, Golbeck J, Banerjee J. OWL: Capturing semantic information using a standardized web ontology language. Multilingual Computing and Technology Magazine, 2004, 15(7): 1-8

[3] Zhu X, Huang Z, Shen G. Description logic based consistency checking upon data mining metadata//Proceedings of the 3rd International Conference on Rough Sets and Knowledge Technology (RSKT 2008). Chengdu, China, 2008: 475-482

[4] Zubcoff J, Trujillo J. Conceptual modeling for classification mining in data warehouses//Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2006). Cracow, Poland, 2006: 566-575

[5] Castellano M, Pastore N, Arcieri F, Summo V, Grecis G. A model-view-controller architecture for knowledge discovery//Proceedings of the 5th International Conference on Data Mining. Malaga, Spain, 2004: 383-392

[6] Chaves J, Curry C, Grossman R L, Locke D, Vejcik S. Augustus: The design and architecture of a PMML-based scoring engine//Proceedings of the 4th Workshop on Data Mining Standards, Services and Platforms (DM-SSP'06). Associated with 12th ACM SIGMOD International Conference on Knowledge Discovery & Data Mining (KDD'06). Philadelphia, Pennsylvania, USA, 2006: 38-46

[7] Berners-Lee T, Hendler J. Publishing on the semantic Web-The coming Internet revolution will profoundly affect scientific information. Nature, 2001, 410(6832): 1023-1024

[8] Baader F, Horrocks I, Sattler U. Description logics as ontology languages for the semantic Web//Hutter D, Stephan W. Lecture Notes in Artificial Intelligence 2605, Berlin: Springer-Verlag, 2005: 228-248

[9] Horrocks I, Patel-Schneider P F. Reducing OWL entailment to description logic satisfiability. Journal of Web Semantics, 2004, 1(4): 345-357

[10] Lutz C. An improved NExpTime-hardness result for description logic ALC extended with inverse roles, nominals, and counting. Dresden University of Technology, Germany: LTCS-Report LTCS-04-07, 2004

[11] Wessel M. Decidable undecidable extensions of ALC with composition-based role inclusion axioms. University of Hamburg, Germany: Technical Report FBI-HH-M-301/01, 2000

[12] Lutz C. The complexity of reasoning with concrete domains. RWTH Aachen, Germany: LTCS-Report LTCS-99-01, 2002

[13] Horrocks I, Sattler U. A Tableaux decision procedure for SHOIQ//Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005). Edinburgh, UK, 2005: 448-453

[14] Yang F. Development of software engineering: Co-operative efforts from academia, government and industry//Proceedings of the 28th International Conference on Software Engineering. Shanghai, China, 2006: 2-11

[15] Zhu Xiao-Dong. Research on modeling of data streams mining systems based on extended predictive model markup language[Ph. D. dissertation]. Nanjing University of Aeronautics and Astronautics, Nanjing, 2009(in Chinese)
(朱小栋. 基于扩展预测模型标记语言的数据流挖掘系统建模研究[博士学位论文]. 南京航空航天大学，南京，2009)

ZHU Xiao-Dong, born in 1981, Ph. D., lecturer. His current research interests include data warehouse and data mining, intelligent data management and electronic business.

XIAO Fang-Xiong, born in 1971, Ph. D., senior engineer. His current research interests include software engineering, cloud computing and electronic business.

HUANG Zhi-Qiu, born in 1965, professor, Ph. D. supervisor. His main research interests are database engineering and software engineering.

SHEN Guo-Hua, born in 1976, Ph. D., associate professor. His current research interests include Web services and semantic Web.

JIN Ling, born in 1980, M. S. candidate. Her current research interests include trusted software, software metrics and testing.

Background

Data mining standardization become a new focus problem in database field in recent years. DM-SSP Workshop associated with the KDD Conference has been hold for six years and it focuses on the data mining standards, services and platforms. Data mining metadata play a kernel role in the standardization of data mining. Data mining products providers look forward common data mining metadata for the exchanging, sharing, integration and standardization of data mining products. PMML can facilitate the exchange of data mining models from one environment to another. It is more attractive and acceptable than CWM and CRISP-DM. However, as pointed out in the KDD Workshop DM-SSP'06, the lacks of conformity of PMML reduce its usefulness of PMML and hamper the growth of its usefulness in the community of data mining. In this research, we proposed an extended PMML language EPMML. A description logic SOIN is designed as the infrastructure of EPMML, which makes EPMML has not only structured XML characteristics, but also semantic expressive ability. EPMML has strict formal mechanism so it supports automatically reasoning to discover the latent inherent inconsistency, which cannot be found with XML validation upon traditional PMML. This research is supported by the Excellent Youth Scholars of Ministry of Education of Shanghai under Grant No. slg10010, the Innovation Program of Shanghai Municipal Education Commission under Grant No. 12YZ103, the Humanity and Social Science Youth Foundation of Ministry of Education of China under Grant No. 12YJC870037, and Leading Academic Discipline Project of Shanghai Municipal Government "Management Science and Engineering" under Grant No. S30504.