

智能数据挖掘引擎的设计与实现^{*}

Design and Implementation of A Intelligent Data Mining Engine

张英朝 邓 苏 张维明 刘青宝

(国防科技大学管理科学与工程系 长沙 410073)

Abstract The basic characters of a high-performance data mining system are to support high-speed data processing and provide enough flexibility and extensibility. In this article, according to the basic requirements of the high-performance data mining system, the substrate details of the data mining engine are discussed. In the discussion, the intelligent integration and call of the data mining algorithms are emphatically illustrated. Based on this discussion, a system architecture of the intelligent data mining engine is provided. On the basis of its architecture, some key techniques of the intelligent data mining engine are analyzed in detail and the implementation of this engine is briefly illustrated. In the end, the applications of the intelligent data mining engine are introduced.

Keywords Intelligent data mining engine, Data mining algorithms, Intelligent selection, Metaknowledge

1. 引言

数据挖掘就是从大量数据中提取或“挖掘”有用信息,进而提供决策支持。在互联网时代,商业竞争日趋激烈,就企业而言,对于企业所积累的大量数据进行高效挖掘、智能分析以辅助企业决策就变得越来越重要。因此,基于人工智能和统计技术的数据挖掘系统日益成为数据挖掘研究的焦点。高性能数据挖掘系统应支持高速数据处理并且提供足够的灵活性和可扩展性。基于此,在构建数据挖掘系统时,所要考虑的一个关键因素就是如何使数据挖掘系统具有很高的通用性和足够的灵活性以满足用户的不同需求。实现数据挖掘系统通用性的基础就是提供多种挖掘分析方法技术(数据挖掘引擎),以使得用户能够选择适合于所要分析数据特点和抽取信息使用方式的挖掘分析技术。

所谓引擎,是借用了机器工业的同名术语,表明其在整个系统中的核心地位。在本文把它称之为“支持数据挖掘应用的底层算法函数集”^[1],或者说是特定应用的一种抽象。

引擎最大的特点就是具有“驱动性”,即引擎仅仅在功能上支持某个方面的应用,但具体的实现细节则依赖于应用本身。其次,引擎具有“完整性”的特点,即引擎是能完整实现某个方面功能的算法函数集。第三,引擎还具有“独立性”的特点,即引擎可以不依赖于具体的应用而独立存在。

数据挖掘引擎是数据挖掘系统核心组成部分,由一组算法集和功能模块组成,用于特征化、关联、分类、聚类分析以及演变和偏差分析^[2]。本文从数据挖掘算法模块通用性和使用灵活性的角度出发,设计了一种智能数据挖掘引擎,同时初步探讨了该挖掘引擎的实现。

2. 引擎的系统结构

引擎是一个比较泛化的概念,在不同的应用场合其核心技术也有所不同。本文研究的数据挖掘引擎侧重于多种数据挖掘算法的智能化集成及调用,通过构造一个通用的数据挖掘算法集成系统,来实现该挖掘引擎的驱动性、完整性、独立

性、灵活性和通用性。智能数据挖掘引擎的系统结构如图1所示。

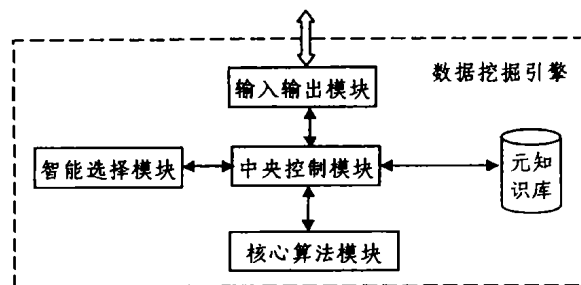


图1 智能数据挖掘引擎系统结构

本文所设计的智能数据挖掘引擎由核心算法模块、智能选择模块、输入输出模块、元知识库、中央控制模块五个功能单元组成。

(1) 核心算法模块^[3]

核心算法模块完成数据挖掘引擎所需要的最基本的处理功能,是引擎中最为重要的部分。核心算法模块主要由关联规则算法、基于记忆推理算法(MBR)或基于实例推理算法(CBR)、聚类分析算法、分类算法和判定树、人工神经网络和遗传算法等几种主要的数据挖掘算法组成。这些算法单元各自有各自的优缺点和适用范围,数据挖掘算法的多样性直接构成了数据挖掘引擎的多功能性。

(2) 智能选择模块

智能选择模块负责根据用户所要挖掘数据的特点和进行抽取所得出信息的使用方式来决定选用合适的数据挖掘算法,进行数据挖掘,以达到最优的挖掘效果。智能选择模块是数据挖掘引擎核心智能模块,直接决定了数据挖掘引擎的工作效率和高效性。

(3) 输入输出模块

输入输出模块负责从数据挖掘集市(数据源)中输入数据和接受来自数据挖掘系统控制器的信息^[3],并将这些数据和

^{*}国家自然科学基金项目(批准号:60003013)。张英朝 博士生,主要研究方向为信息管理与智能决策技术。邓 苏 教授,主要研究方向为数据库技术、智能决策技术。张维明 教授,主要研究方向为信息系统工程、智能决策技术。刘青宝 副教授,主要研究方向为数据挖掘技术。

信息通过引擎中央控制模块提交给引擎智能选择模块,使得智能选择模块可以利用这些数据与信息进行工作。同时,输入输出模块还负责将数据挖掘引擎核心算法模块的处理结果提交给数据挖掘系统控制器^[3]。

(4) 中央控制模块

中央控制模块是数据挖掘引擎核心控制单元,负责整个挖掘引擎各模块之间的协调、控制。它通过控制引擎的输入输出模块,来控制引擎与整个数据挖掘系统的交互工作,进而实现引擎的完整性和独立性。

(5) 元知识库

元知识库主要存储关于数据挖掘引擎构建及工作过程、核心算法模块算法和智能选择模块的知识,中央控制模块负责对其进行更新和控制。元知识库是实现数据挖掘引擎智能

性的基础单元。

3. 引擎中的若干关键技术

智能数据挖掘引擎的设计涉及到众多领域的各种前沿技术,本文将主要从数据挖掘引擎的驱动性、通用性和灵活性出发,着重就引擎的数据挖掘算法和实现智能选择这两个方面做一个简要的介绍。

3.1 数据挖掘算法^[4]

数据挖掘算法是整个数据挖掘引擎的技术核心,不同的挖掘算法其应用领域及特点不尽相同,这就要求数据挖掘引擎必须拥有多种算法,以供用户在数据挖掘时进行选择。本文所设计的数据挖掘引擎采用几种业界公认的挖掘算法,同时对于其中的某些算法(如 MBR 等)进行了改进,如表 1 所示。

表 1 引擎主要数据挖掘算法

数据挖掘算法	
关联规则	关联规则算法识别原因和结果的关系,同时以“支持度”和“置信度”等概率因子来支持挖掘结论 ^[2] 。关联规则的形式为“if(conditions), then(conclusion)”,用来做出预测或者估计未知属性值。该算法能够自动抽取出现概率较高的关系 ^[5] 。
基于记忆推理算法(MBR) ^[6] 和基于实例推理算法(CBR)	MBR 和 CBR 算法针对新记录和新状态,在历史记录中找到与其最为相似的记录,再将这些相似的记录提交给一个基于权重的“大多数原则投票过程”,用以估计、预测新记录的值或者新状态的结果。本文所设计的数据挖掘引擎采用了一种改进型的 MBR 算法:EMBR ^[3] 。
聚类分析算法	聚类分析算法根据数据记录间的距离将不同种类的数据分成相似的或者半相似的数据子群 ^[7] 。这种算法首先创建一个分类树图,通过在所创建的树中定义感兴趣的簇的序号或者不一致性,进而对于数据进行分簇。根据计算数据记录间距离方法的不同,聚类算法也有所不同,本引擎主要支持 Ward 算法 ^[7] ,该算法基于欧氏距离将归并后能够减小群内平方数量级分类错误的记录/簇合并起来。
分类算法和判定树	判定树将数据分成代表领域状态的树结构 ^[2] 。这种算法基于一个目标变量来决定数据的自然分离,能够从已经创建的树中很容易地抽取 if-then 规则以辅助用户理解分析结果。由于判定树是基于领域创建的,因此本算法不适合于领域间非线性数据的分析。常用的判定树算法有 C5.0 ^[8] 、CART ^[9] 、CHAID ^[10] 和 PDT 算法 ^[11] 等,本引擎采用 PDT 判定树算法。
人工神经网络 ^[12]	人工神经网络是由大量类似于人脑神经元的处理单元相互联接而成的非线性复杂网络系统。应用一组输入变量、数学活化函数和输入权重来预测目标变量的值。通过一个反复的学习过程,神经网络不断地调整权重直到预测的输出与实际的数据值相一致。一旦学习完成后,该网络就作为用于预测新数据值的模型。由于人工神经网络能够应用其概括能力在学习数据中插入新的值,因此比较适合于连续值数据的挖掘问题。本引擎采用基于 B-P 模型的神经网络系统 ^[13] 。
遗传算法	遗传算法应用一个高度反复的选择过程,通过交叉和变异等操作来创建连续的后代模型。通过一个适当的函数来保持特定的成员模型,同时丢弃其他的成员模型 ^[4] 。遗传算法易于并行,业已用于分类和其他优化问题(如优化神经网络的布局和权重),同时,还可独立作为数据挖掘算法模型。

3.2 智能选择控制技术^[13]

智能选择控制技术是实现数据挖掘引擎通用性的关键技术,引擎中智能选择以元数据库为支持,通过推理机机制实现。本文所设计引擎的智能选择采用较为常用的产生式规则推理机制^[14],主要通过收集来自用户的数据挖掘需求和数据源数据的特征,进而依托元知识进行 if-then 式推理,来判定选取适当的数据挖掘算法进行数据挖掘。

3.3 元知识库

元知识库用来存储和管理引擎中的各数据挖掘算法、智能选择模块和中央控制模块以及引擎工作控制过程等的知识信息。元知识库中知识的存储和管理方式对于引擎的整体性能至关重要。一般来说,组织这些知识可以通过单个文件,也可以通过文件簇的形式实现。本文设计的数据挖掘引擎中,元知识以面向对象的方式进行多维存储(即元知识库从某种程度上讲是一个多维数据库);元知识以基于功能和面向应用的方式进行管理。这种面向对象组织知识存储方式极大地方便

了元知识库的维护和升级,而基于功能的管理方式对于提高整个挖掘引擎智能选择效率至关重要。

4. 引擎的设计与实现

4.1 对象的构造

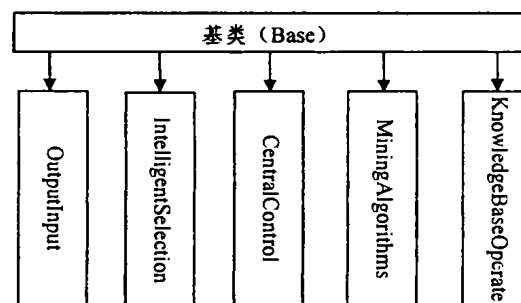


图 2 数据挖掘引擎类结构

引擎中对象的构造是基于引擎模块的。在引擎的各模块

中最为重要的无疑是核心算法模块。引擎的基本类结构如图2所示。

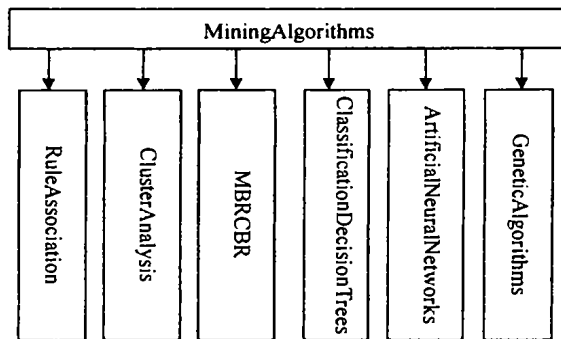


图3 MiningAlgorithms 子类结构

数据挖掘引擎核心处理算法模块中算法类型比较多,因而 MiningAlgorithms 类比较复杂,它又分为如图3所示的几种子类。

4.2 初始化流程

这个过程包括设定状态参数、数据及用户要求预处理、加载元知识库以及某些模块的初始化等等。简要描述如下:

(1)导入数据挖掘引擎的状态参数配置文件,设定引擎的状态;状态参数配置文件是一个文本文件,用来描述引擎缺省的参数。

(2)采集需要挖掘数据的特征信息和用户要求信息。

(3)加载元知识库知识。

(4)将相关的参数和元知识传送到各个相应的功能模块。

(5)初始化中央控制模块。

(6)初始化智能选择模块。

最后,智能选择模块在中央控制模块的控制下,根据元数据库提供的元知识、需要挖掘数据的特征信息和用户要求,选择出适当的数据挖掘算法模块。数据挖掘算法模块在中央控制模块的控制下开始数据挖掘工作。

5. 应用实例

5.1 战场决策支持

随着现代科学技术的飞速发展以及大量高新技术应用于军事领域,使得现代战争具有许多新的内容,战场信息多种多样、数量巨大;另一方面,多兵种的联合协同作战日益成为现代战争的主要作战模式,这就使得战场决策变得异常复杂。如何从种类繁多的海量战场信息中为战场指挥员获取有用信息以辅助战场决策已成为决定战争胜负的关键。

数据挖掘引擎能够很好地应用于战场信息的数据挖掘,为指挥员提供进行决策的有用数据。本文所设计的引擎在进行改进的基础上,可以成功地应用于战场信息数据挖掘,提供战场决策支持。

5.2 金融证券 CRM

CRM 是一个前景非常好的领域,也是数据挖掘引擎的主要应用方向之一。通过对于客户的数据信息进行挖掘,得出企业所需要的有用信息,以使企业可以根据挖掘得到的有益客户信息来调整企业客户策略,确保企业在日趋激烈的市场竞争中始终立于不败之地。本文所设计的数据挖掘引擎已成功应用于金融证券行业 CRM 系统中,取得了很好的效果。

结论 本文在重点考虑引擎通用性和完整性的前提下,从底层设计并实现了一个面向决策支持应用的智能数据挖掘引擎。该引擎的设计与实现侧重于多种数据挖掘算法的智能化集成及调用,采用了目前数据挖掘领域较为成熟和广泛应用的算法,对于各种算法的具体实现细节及改进工作,本文只做了初步的探讨。数据挖掘算法是数据挖掘引擎的基石,挖掘算法的先进性、高效性直接决定了数据挖掘引擎的性能。因此,数据挖掘算法的具体实现及改进将是我们进一步工作的重点。

参考文献

- 1 Kim S-M, Kim J-D, Hong J-H, Nam D-E. A System for Association Rule Finding from an Internet Portal Site. INFORMS & KO-RMS - 999 - Seoul 2000 (Korea)
- 2 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc., 2001
- 3 Yaginuma Y. High-performance Data Mining System. FUJITSU Sci. Tech. J. 36. 2. Dec. 2000. 201~210
- 4 Collier K, Carey B. A Perspective on Data Mining. <http://www.nau.edu/cdi>. July 1998
- 5 Srikant A R. Fast Algorithms for Mining Association Rules. In: Proc. of the 20th Int'l. Conf. on Very Large Databases. Santiago, Chile. Sept. 1994
- 6 Stanfill C, Waltz D. Toward Memorybased Reasoning. Communications of the ACM, 1986, 12
- 7 Ward J H. Jr. Hierarchical Grouping to Optimize an Objective Function. J. American Statistical Association, 1963, 58 (301): 235~244
- 8 Quinlan R. RULEQUEST RESEARCH Home Page. <http://www.rulequest.com>.
- 9 Breiman L, et al. Classification and Regression Trees. C. J. Champion & Hall, International Thomson, Publishing. 1984
- 10 Kass G V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 1980, 29: 119~127
- 11 Yugami N. Decision Tree Induction based on an Expected Accuracy. In: Proc. of JSAT'96. 1996. 87~90
- 12 Yasui. A new method to remove re-dundant connections in back-propagation neural networks: Introduction of parametric lateral inhibition fields. In: Proc. of IJCNN'92. II360/ 367, Beijing
- 13 陈文伟. 智能决策技术. 电子工业出版社, 1998
- 14 林尧平, 马少平. 人工智能导论. 清华大学出版社, 1989