网格环境下分布式元学习任务的设计

吕 品[1],陈年生[2],董武世[2]
(1.武汉工程大学 计算机科学与工程学院,湖北 武汉 430073;
2.湖北师范学院 计算机系,湖北 黄石 435002)

摘 要:元学习方法是采用集成学习的方式来生成最终的全局预测模型。该方法的基本思想是从已经获得的知识中再进行学习,从而得到最终的数据模式。网格能有效地为元学习提供高性能和分布式的基础设施。文中根据知识网格的概念,在Globus Toolkit的基础上,分析了知识网格的体系结构和它的主要组件。根据分布式元学习的一般过程,设计了在知识网格体系结构下的元学习任务。

关键词:分布式元学习;知识网格;网格服务
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2006)10-0014-03

Design of Distributed Meta-Learning Task on Grid

LÜ Pin[1], CHEN Nian-sheng[2], DONG Wu-shi[2]
(1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430073, China;
2. Department of Computer Science, Hubei Normal University, Huangshi 435002, China)

Abstract: Meta-learning method makes use of ensemble learning to obtain the global predication module. Its primary idea is to learn again according to discovered knowledge, finally to get data patterns. The grid can be effectively exploited for deploying meta-learning because of the high-performance it can offer and its distributed infrastructure. According to the concept of knowledge grid, analyzes the knowledge grid architecture and its main components on top of the Globus Toolkit, designs a kind of software modules according to the process of data mining, and presents the modules should provide services. Presents the design of meta-learning task of the grid on basis of the process of distributed meta-learning.

Key words: distributed meta-learning; knowledge grid; grid services

0 引 言

各相关学科的飞速发展,尤其是Internet的广泛使用,使得实际应用要求数据挖掘系统具有更好的可扩展性。例如:研究某种疾病在某地的发病情况与气候的关系,需要疾病控制数据库和环境数据库;金融组织间通过合作防止信用卡欺诈,需要数据共享;大型跨国公司营销策略的制定,由于销售点分散,数据仓库构造十分耗时。于是人们将数据挖掘技术与分布式计算的有机结合,应用于分布式环境下的数据模式发现[1]。

元学习的概念正是在这种背景下由Prodromidis等人于2000年首先提出的,该方法采用集成学习(ensemble learning)的方式来生成最终的全局预测模型(即元分类器)。该方法的基本思想是从已经获得的知识中再进行学习,从而得到最终的数据模式。由于元学习是在分布式数据挖掘的基础上实现的,而网格能有效地提供高性能应用和分布式的基础设施,因此,研究网格环境下的元学习将具有重要的意义。

收稿日期:2006-01-03
基金项目:湖北省自然科学基金资助项目(2004ADA023)
作者简介:吕 品(1973-),女,湖北黄石人,硕士,讲师,研究方向为数据挖掘、算法分析与设计、软件工程。

1 元学习的基础

为了有效地利用网格挖掘出大量数据中隐藏的知识,必须在网格中增加数据挖掘的工具和相应的服务。这个目标可以通过网格技术和数据挖掘技术的结合来实现。这样利用网格技术统一实现数据管理和知识发现就构建了一个基于知识的网格[2,3]。

目前,这种知识网格的体系结构定义在网格工具包和服务的基础上,知识提取的实现基于Globus toolkit[4,5]。如图1所示,知识网格由两个层次组成:核心知识网格层和高层知识网格层。核心知识网格层的服务在一般网格服务的基础上实现;高层知识网格层用来描述、开发和执行在知识网格上的分布式知识发现。

●核心知识网格层由两个主要的服务组成:

(1)知识目录服务:这个服务主要负责维护知识网格上所有数据和工具的元数据描述。元数据信息用XML文档来描述,并且存储在知识元数据仓库中。此外,知识网格中还有一个存储基本知识的仓库,其作用是用这些基

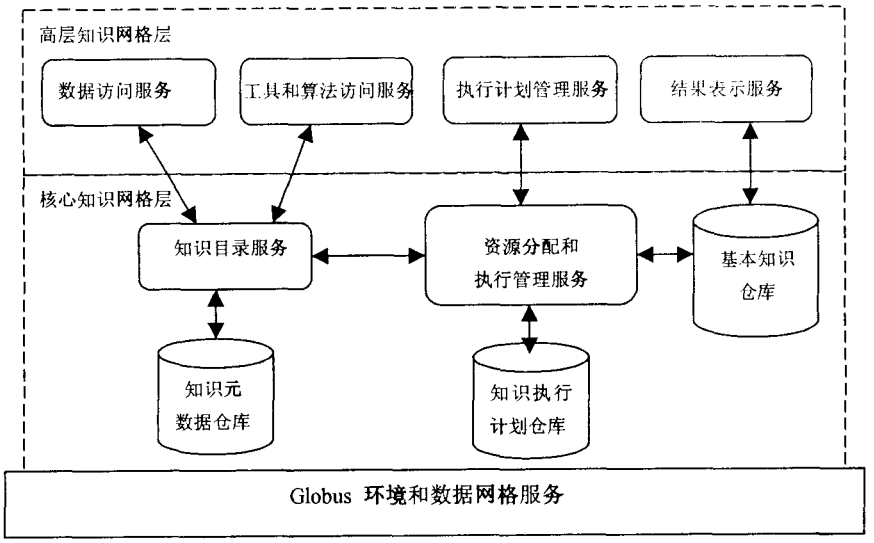

图 1　知识网格的体系结构

本知识与挖掘出来的结果进行比较。

(2)资源分配和执行管理服务。这些服务主要用来寻找可执行计划与可利用资源之间的一种映射关系,目的是满足应用的需求(如:计算能力、存储量、内存容量、数据库、网络带宽以及网络延迟)和网格约束。这种映射可通过有效地定位资源来获得。当执行计划开始后,该层就要直接利用 Globus GRAM(Globus Resource Allocation Manager)服务协调和管理应用执行的情况。一个数据挖掘程序的资源需求用资源说明语言来表达。执行计划的分析和处理过程将产生全局的资源请求,反过来,这些请求又会被翻译成局部的资源说明语言被局部的 GRAM 所用。

●高层知识网格层有 4 个主要的服务:

(1)数据访问服务:它主要负责查找、选择、提取、转换和传送被挖掘的数据。查找和选择是基于核心的知识目录服务,根据用户的需求和约束,数据访问服务能自动地查找并找到有用的数据源给数据挖掘工具。提取、转换和传送不仅基于 Globus GASS 服务,还利用了知识目录服务。

(2)工具和算法访问服务:它主要负责查找、选取和下载数据挖掘工具和算法。这些工具和算法都存储在每一个知识网格结点的局部存储器中。如果一个用户想在本网格结点传送数据挖掘工具给其它网格结点用户,那么他可以利用核心知识网格层中存储在知识目录服务中的知识元数据仓库,找到相应的参数、输入输出数据的格式、数据挖掘算法的类型、资源需求和约束等等。

(3)执行计划管理服务:一个执行计划由一个图形数据结构来表示,该图描述了执行计划内部的相互作用和数据源、提取工具、数据挖掘工具、可视化工具以及在基本知识仓库中的知识结果存储之间的数据流。然而,由于数据访问服务和工具及算法访问服务结果的变化,也会产生不同的执行计划。因而,该服务是一个半自动的工具,它需要用户的参与。执行计划存储在知识执行计划仓库中,便于迭代的知识发现过程的实现。比如,周期性地分析同样的数据集,或同一执行计划分析不同的数据集等。

(4)结果表示服务:在数据挖掘过程中,结果可视化是一个非常重要的步骤,有助于用户对模型的理解。它定义了如何产生、表示和可视化被提取的知识。

2　元学习方案

根据以上知识网格体系结构,下面描述如何利用知识网格实现元学习。元学习的思想就是运用学习算法收集分布式的数据集产生大量独立的局部分类器,然后把这些局部分类器合并成一个全局的分类器[1]。

图 2 是一个分布式的元学习方案。原始数据集 DS 存储在结点 A 中,在结点 Z 上获得全局分类器 GC。

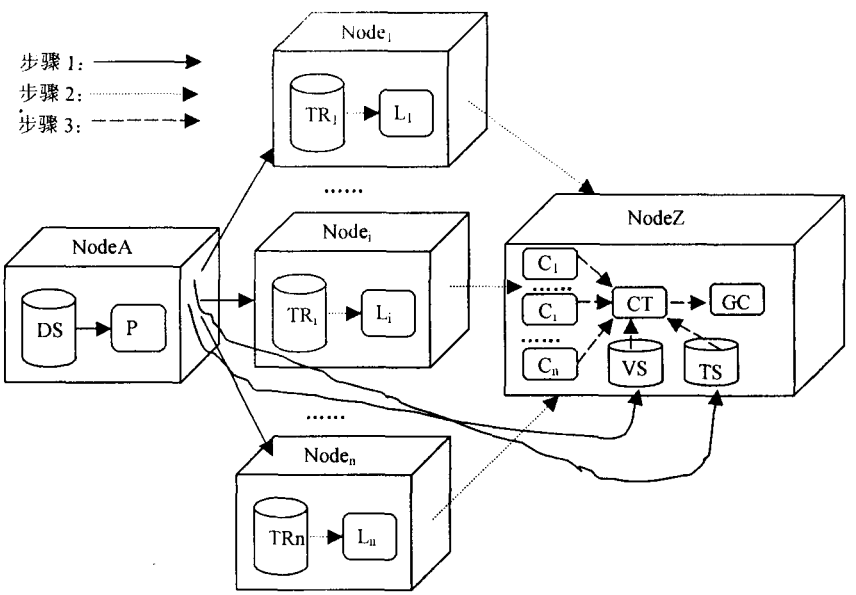

图 2　分布式元学习方案

元学习的过程分为三步:第一步,在结点 A 上,利用分类方法 P 从原始数据集 DS 中抽取训练集 $TR_1,\cdots,TR_n$,测试集 TS 和校验集 VS,然后,训练集 $TR_1,\cdots,TR_n$ 分别从结点 A 移动到 n 个结点 $Node_1,\cdots,Node_n$ 上,测试集 TS 和校验集 VS 从结点 A 移动到结点 Z 上;第二步,在每一个结点 $Node_i(i=1,\cdots,n)$ 上,利用学习算法 L_i,从

训练集 TR_i 中得到分类器 C_i，然后把每一个分类器 C_i 从 $Node_i$ 移动到结点 Z 上；第三步，在结点 Z 上，n 个分类器首先合并，然后经过测试集 TS 测试，校验集 VS 验证，最后得到全局分类器 GC。

如果有趣的数据已经分布在不同的结点上，元学习的第一步可以省略。

3 网格上元学习任务

基于上述思想，元学习的基础是知识网格。假定结点 A 是一个知识网格结点，P 是分类方法，L 是学习算法且 $L_1 = L_2 = \cdots L_n = L$；结点 S 是一个知识网格结点，合并－测试器 CT 可在多种平台上运行。

若一个网格用户需要利用结点 S 上的元学习工具实现对数据集 DS 的分类。这个任务在知识网格上的执行过程如图 3 所示。

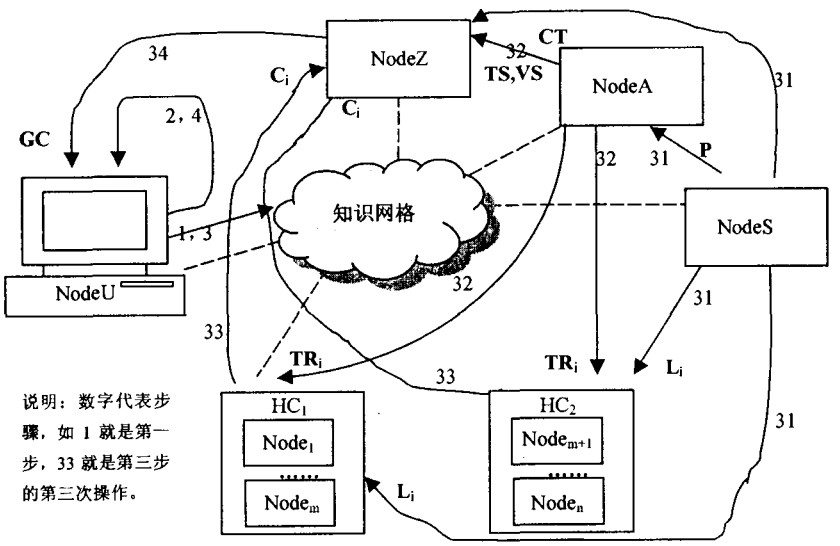

图 3 网络环境下分布式元学习任务

网格上元学习任务分为以下几个步骤：

步骤 1：网格用户在知识网格结点 Node$U$ 上查找元学习任务需要的可计算资源。这个查找过程由核心知识网格层的知识目录服务完成。定位学习过程中需要的资源则分别由 HC_1 和 HC_2 Globus 结点完成，这些 Globus 结点还是一个计算机簇，学习过程可在 Globus 结点上并行进行。Node$Z$ 是 Globus 结点，它负责执行合并过程。查找过程要比较关于 L 和 CT 程序的元信息，以与 HCs 和结点 Node$Z$ 对照。

步骤 2：网格用户构建一个元学习任务执行计划，指定算法实现的工具和数据移动的策略。这个执行计划通常由执行计划管理服务创建，然后存储在本地知识执行计划仓库中。

步骤 3：网格用户将执行计划提交给资源定位管理服务后，元学习过程就按以下顺序执行：

(1)结点 Node$S$ 利用工具和算法访问服务负责将 P 安装在Node$A$ 上，L_i 安装在 HC_i 上，CT 安装在 Node$Z$ 上；

(2) 结点 Node$A$ 利用数据访问服务将 P 从数据子集中提取 $TR_1, \cdots, TR_n$ 以入 TS 和 VS，然后，将 $TR_1, \cdots, TR_n$ 传送给 HC_1 和 HC_2，TS 和 VS 传送给结点 Node$Z$；

(3)HC_1 和 HC_2 上的每一个计算单元会根据 L_i 产生一个局部分类器 C_i，然后 C_i 被传送到结点 Node$Z$ 上；

(4) 结点 Node$Z$ 利用 CT 产生全局分类器 GC，然后 Node$U$ 上的结果表示服务将 GC 存储在基本知识仓库中。

步骤 4：网格用户利用结果表示工具将挖掘的最终结果可视化，并对这个挖掘过程进行评价。

4 结束语

分布式数据挖掘和计算网格是未来高性能计算的两个重要方面。网格上的工具和应用将会越来越完全、复杂和多样化。网格正在将面向数据的、高水平的信息管理服务和面向计算的服务结合在一起。这种趋势扩大了网格的应用领域，为高性能分布式的数据挖掘和知识发现提供了机遇。知识网格就是这种趋势下的一个重要构件。

利用知识网格实现分布式元学习有着广泛的应用，如何利用它实现对网络入侵检测系统中数据包的分类分析，生命信息科学中白质分子的折叠以及将数据挖掘系统整合到知识网格环境中以实现复杂的挖掘任务，是下一步要重点研究的问题。

参考文献：

[1] Prodromidis A L, Chan P K, Stolfo S J. Meta－learning in distributed data mining systems: issues and approaches[A]. In: Kargupta H, Chan P(Eds.). Advances in Distributed and Paralles Knowledge Discovery[C]. Boston: AAAI Press/MIT Press, 2000.81－87.

[2] Cannataro M, Congiusta A, Pugliese A, et al. Distributed Data Mining on Grids: Services, Tools, and Applications[J]. IEEE Transactions on Systems, Man, Cybernetics, Part B, 2004, 34(6): 1－15.

[3] Hoschek W, Martinez J J, Samar A, et al. Data management in an international data grid project[A]. In: Proceedings of the IEEE/ACM International Workshop on Grid Computing Grid'2000, LNCS[C]. Berlin: Springer, 2001.77－90.

[4] The Globus Toolkit[EB/OL]. http://www.globus.org/toolkit, 2001.

[5] Foster I. Building the grid: an intergrated services and toolkit architecture for next generation networked applications[EB/OL]. http://www.gridforum.org/building_the_grid.htm., 2000.