数据网格分布式动态分类系统及其在乳腺癌网格中的应用

王 坤, 李三立, 柴云鹏, 王晓英
(清华大学计算机科学与技术系, 北京 100084)

摘 要: 数据网格作为面向服务的架构, 为远程用户提供分布式数据查询、存储和管理等服务, 而数据网格中的数据分类日益成为研究者们所关注的问题. 本文描述了用于数据网格的一种高效的分类系统. 该系统动态综合作为网格服务的多种分类方法(Dynamical Synthesis of Multiple Methods, DSMM), 能够动态地改善传统分类方法的低准确率点, 以负载平衡为前提将分类工作分布于网格中的各个结点上. 另外, DSMM 提供的生命周期管理保障了其作为一个网格应用的鲁棒性和灵活性, 适合于网格的松耦合体系结构. 实验采用了 2927 个乳腺癌患者病例, 结果显示 DSMM 系统的确能够在数据网格环境中发挥其灵活性、高效性并提高分类的准确率.

关键词: 数据网格; 分布式; 分类
中图分类号: TP399 文献标识码: A 文章编号: 0372-2112 (2008) 04-0620-07

A Distributed Dynamical Classification System for Data Grid and Its Application in Breast Cancer Grid

WANG Kun, LI San-li, CHAI Yun-peng, WANG Xiao-ying
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: As a service oriented architecture, data grid provides distributed data access, storage and management for remote users. It has become growing concern among developers about the data classification issue in data grid environment. This paper proposes a high-efficient system used in data grid which makes dynamical synthesis of multiple methods (DSMM), distributes workloads to each node of the grid with the load-balance premise, and improves the low classification accuracy points of traditional methods. In spite of that, as an application in loose coupling grid environment, DSMM, which provides life time management of multiple classification methods, guarantees its robustness and flexibility. Experiment uses 2927 breast cancer cases and proves that the DSMM system has flexibility, high efficiency and increases the classification accuracy in data grid.

Key words: data grid; distributed; classification

1 引言

分类方法[10,11]对于数据网格[17]具有非常重要的意义. 数据网格中的数据集(data set)分布式地存在于网格的各个结点之上, 对数据集进行分类是数据存储和管理的前提. 另外, 很多应用模式都要求将这些异构的、分布式的数据集划分成若干数据类. 例如: 用于疾病普查的医学数据网格[2,5]需要将大量的病例数据划分为阴性、阳性、假阴性或假阳性等普查结果类. 可以说使用网格的分布式计算和存储能力对异构的、地域上分散的医学数据进行分类和分类基础上的存储是网格技术应用于医学领域最为关键的问题之一. 因此, 如何在数据网格上提供高效的分类服务直接关系到网格体系结构在这一类的专业技术环境中的应用广泛性和效率.

开放网格服务体系结构(OGSA)和基于 Web 服务的网格资源架构(WSRF)[4]都定义了一种面向服务基础设施所需要的基本框架结构, OGSA 数据管理与集成(OGSA-DAI)[9]使得现有的数据资源, 例如关系数据库和 XML 数据库能够集成到网格环境中. 这种分布式的异构数据存储和访问方式要求一种分布式的数据分类方法, 这种数据分类方法应该具有以下特点: ①容错性(鲁棒性): 当网格环境中的某个结点出现故障时, 并不会影响分类工作的正常进行; ②高效性: 能够以比较高的准确率进行数据分类. 另外, 在很多研究范围内, 都要

收稿日期: 2007-11-23; 修回日期: 2007-12-17
基金项目: 上海高校网格 e-研究院项目

求分类方法不能出现准确率非常低的分类操作. 例如在医学普查中, 为数不多的分类准确率低点有可能造成假阴性与假阳性的混淆; ③灵活性: 网格中的低准确率分类方法可以被即时删除, 同时也可以添加一些新的分类方法.

本文研究的主要内容是在 Web 服务的网格资源架构(WSRF)下, 基于 OGSA 数据管理与集成(OGSA-DAI)[9], 参照动态更新的分类规则对数据库中的数据集进行分类. 论文提出一种分布式动态分类系统(dynamical synthesis of multiple methods, 简称 DSMM), DSMM 包含四个模块: ①数据分类模块; ②生命周期管理模块; ③负载平衡调度模块; ④数据管理模块.

通过以上四个模块的协同工作, 对分布于网格各个计算结点之上的多个分类服务进行有效的生命周期管理, 及时去掉分类效率低下的分类服务并适时添加分类效率高的分类服务, 同时对各分类服务的分类结果进行基于历史效果的动态加权积分, 使历史效果好蹬分类服务的结果占有更高的权重. DSMM 同时也提供负载平衡调度模块对网格中各个结点的负载情况进行监测, 实时调整负载以发挥系统最优性能. 总结其优点有: ①将分类计算量分配到网格的各个结点之上, 并发地进行计算, 有效地利用了网格的分布式计算资源; ②各种分类服务作为独立的模块可以在网格的结点上进行动态添加或删除. 如果发现分类效果更好的分类服务, 可以把它作为一个新的模块添加进来, 如果发现某个分类服务准确率低下或出现故障, 则可以将它删除掉, 而不影响整个系统的分类工作. 这保证了 DSMM 应用于网格环境不可或缺的高效性和容错性; ③就分类性能而言, 改善了对分类数据集进行分类过程中分类准确率低点的情况, 即提高了分类准确率的下界; ④适合于对数据流分类等有新数据不断加入的情况.

2　相关工作

目前广泛使用的分类方法主要分为两大类: 传统的经典分类方法和基于经典分类方法之上而作的改进[1, 10, 11, 13~16]. 另外还有一部分专科领域上的分类方法[2, 3, 6~8, 12]. 文献[13]所提方法是在算法的层面上为单个计算结点服务, 作为一个单个的分类方法不可避免地存在分类误差较大的点. 而文献[1]强调使用数据分类技术后对于数据管理和存储性能的改善, 对于数据集分类的准确率则没有作太多限定. 作为医学领域的分类算法研究, 文献[12]采用贝叶斯分类方法对癌症病例进行分类和预测. 网格技术和数据挖掘技术的不断发展需要更能契合网格环境的新分类方法.

目前已经有一些医学网格应用, 如: e-Diamond 项目[15]和 MammoGrid 项目[16], 其中对病例的分类和预测, 也是属于第二类的分类方法, 即对传统方法的改进, 没有专门针对网格体系结构而提出新分类方法. 本文提出的 DSMM 既不属于传统的经典分类方法, 也不属于经典分类方法之上的改进, 它并不仅仅是一个分类的算法, 而是应用于网格分布式计算和存储体系结构之上的分布式分类系统, 改善分类过程中低准确率的分类操作, 拥有生命周期管理机制和负载平衡调度机制, 属于一种新的类别.

3　DSMM 的体系结构

DSMM 由四个模块构成: 数据分类、生命周期管理、负载平衡调度和数据管理, 如图 1 所示.

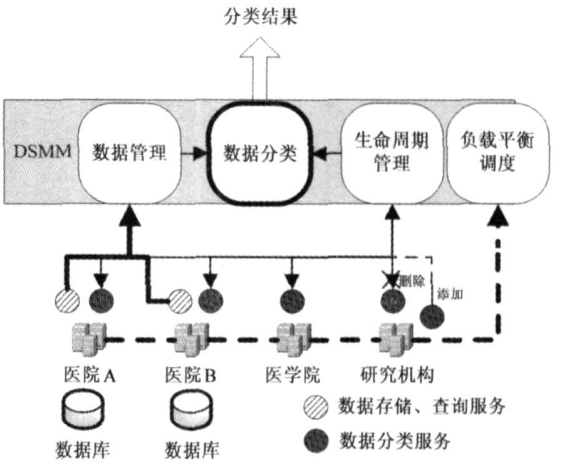

图 1　DSMM 的体系结构

在数据网格的大范围应用中, 数据的存储为分布式的、异构的. 对于这些异构的数据资源, 数据管理模块必须对其进行封装并向数据分类模块提供一个统一的访问接口, 使用 OGSA-DAI[9]. 生命周期管理模块使用分类准确率(下文给出定义)对各结点的数据分类服务工作效率进行衡量, 实时删除效率低的服务, 同时也可以添加新的数据分类服务. 负载平衡调度结点对各结点的负载情况进行监控, 主要包括 CPU 和内存的利用率, 根据监控情况调整各个结点的负载, 避免出现有的结点满负荷而有的结点赋闲的情况. 数据分类模块是 DSMM 的核心, 根据生命周期管理模块调度下的各结点数据分类服务和数据管理模块提供的数据资源, 数据分类模块将数据资源按照本文所设计的算法进行分类, 并提交最终的分类结果.

由于数据管理模块有一些现成的经验可以借鉴[9], 不作为本文论述的重点. 下文将着重介绍 DSMM 中数据分类模块、生命周期管理模块和负载平衡调度模块的工作原理.

4　DSMM 分类模块

DSMM 应用于数据网格环境下, 为大范围内的用户

提供数据集的分类信息. 下文首先定义数据网格中使用分类服务的若干概念, 再对 DSMM 分类模块进行描述并给出算法.

4.1 基本概念

对于在网格中使用分类服务有以下的定义:

定义 1 设数据网格中存储的待分类数据集为 $S=\{pa_1, pa_2, \cdots, pa_m\}$, $|S|=m$, 有属性集合 $A=\{a_1, a_2, \cdots, a_l\}$, $|A|=l$. 用属性集合 A 将数据集 S 分为 $v+1$ 类 $\{R_0, R_1, \cdots, R_v\}$, 有 $\left(\bigcap_{i=0}^{v} R_i=\phi\right) \wedge \left(\bigcup_{i=0}^{v} R_i=S\right)$.

定义 2 $\forall pa_i \in S$, 使用网格结点 $node(\hat{s})$ 上的分类服务 $\hat{s}$ 对数据 pa_i 进行分类, $pa_i \in R_j$ 的概率为 $p(i, j)$. 有: $\sum_{j=0}^{v} p(i,j)=1$.

定义 3 定义网格结点 $node(\hat{s})$ 上的分类服务 $\hat{s}$ 对于 pa_i 的 R_j 分类结果为 $\hat{s}(pa_i, R_j)$, $\hat{s}(pa_i, R_j)=p(i,j)$; $\hat{s}$ 对于 pa_i 的最终分类结果为 $\hat{s}(pa_i)$, $\hat{s}(pa_i)=\zeta \Leftrightarrow \exists \zeta \in [0, v] \wedge p(i, \zeta)=\max_{j=0}^{v}(p(i,j))$. $\hat{s}(pa_i)=\zeta \Rightarrow pa_i \in R_\zeta$.

定义 4 设 $\hat{s}(pa_i)=B$, 分类服务 $\hat{s}_n$ 对于数据 pa_i 的分类准确率为 $Accu_{n,i}$.

命题 1 可令 $Accu_{n,i}$ 的取值为:

$$Accu_{n,i}=\sum_{j=0}^{v} \Psi_j \hat{s}(p_i, R_j)=\sum_{j=0}^{v} \Psi_j p(i,j) \tag{1}$$

其中 Ψ_j 为 $p(i,j)$ 在 $Accu_{n,i}$ 中所占的权重. $\forall j \in [0, v] \wedge j \neq B$, $\Psi_j<1$, $\Psi_B=1$.

证明 要证明 $Accu_{n,i}$ 可以作为衡量分类准确率的标准, 需要证明两个条件:

条件 1, $Accu_{n,i} \leqslant 1$. 条件 2, $p(i, B)=1 \wedge p(i,j)=0$. $\forall j \in [0, v] \wedge j \neq B \Leftrightarrow Accu_{n,i}=1$. 由 $\sum_{j=0}^{v} p(i,j)$ 知: $Accu_{n,i}=\sum_{j=0}^{v} \Psi_j p(i,j)=\Psi_1 p(i,1)+\Psi_2 p(i,2)+\cdots+\Psi_v p(i,v) \leqslant \sum_{j=0}^{v} p(i,j)=1 \Rightarrow Accu_{n,i} \leqslant 1$. 条件 1 得证. $p(i, B)=1 \wedge p(i,j)=0, j=0, \cdots, v, j \neq B \Rightarrow Accu_{n,i}=\sum_{j=0}^{v} \Psi_j p(i,j)=\Psi_B p(i, B)=1$. 证得 $p(i, B)=1 \wedge p(i,j)=0, j=0, \cdots, v, j \neq B \Rightarrow Accu_{n,i}=1$.

用反证法, $Accu_{n,i}=1$. 假设 $p(i, B) \neq 1 \vee \exists j \in [0, v]\ j \neq B, p(i,j) \neq 0$, 则 $Accu_{n,i}=\sum_{j=0}^{v} \Psi_j p(i,j)=p(i, B)+\sum_{j=0, j\neq B}^{v} \Psi_j p(i,j)<p(i, B)+\sum_{j=0, j\neq B}^{v} p(i,j)<\sum_{j=0}^{v} p(i,j)=1 \Rightarrow Accu_{n,i}<1$ 与 $Accu_{n,i}=1$ 矛盾. 由此假设不成立, 即有 $p(i, B)=1 \wedge p(i,j)=0, \forall j \in [0, v] \wedge j \neq B$. $Accu_{n,i}=1 \Rightarrow p(i, B)=1 \wedge p(i,j)=0, \forall j \in [0, v] \wedge j \neq B$.

条件 2 得证. 命题得证.

4.2 DSMM 分类模块工作原理

设数据网格中有 n 个结点用于数据分类. 部署于数据网格各个结点的分类服务的集合 $C=\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_n\}$, $|C|=n$, 对 DSMM 分类模块的结果有以下定义:

定义 5 对于 $\forall pa_i \in S$, 定义 DSMM 分类模块判断 $\forall pa_i \in R_j$ 的概率, 即 DSMM 的 R_j 分类结果为 $DSMM_C(pa_i, R_j)$. 有:

$$\sum_{j=0}^{v} DSMM_C(pa_i, R_j)=1 \tag{2}$$

定义 6 对于 $\forall pa_i \in S$, DSMM 分类模块的 R_j 分类结果取值为:

$$DSMM_C(pa_i, R_j)=\frac{\omega_1(i)\hat{s}_1(pa_i, R_j)+\omega_2(i)\hat{s}_2(pa_i, R_j)+\cdots+\omega_n(i)\hat{s}_n(pa_i, R_j)}{\eta(i)} \tag{3}$$

其中 $\omega_1, \cdots, \omega_n$ 为每个分类服务在最后结果中所占的权重. $\eta(i)$ 为调整因子. DSMM 分类模块选取适当的 $\omega_1, \cdots, \omega_n$ 和 $\eta(i)$ 的值, 再根据式 (3) 计算出 $DSMM_C(pa_i, R_j)$. $\omega_1(i), \cdots, \omega_n(i)$ 的取值有以下两种不同的选取方法:

(1) 只参考上一次分类的准确率.

$$\omega_n(i)=Accu_{n,i-1} \tag{4}$$

(2) 考虑历史积累. 当某个分类服务上一次的分类准确率非常不符合整个分类过程中分类准确率的变化趋势时, 只考虑上一次的分类准确率会使本次分类准确率也受到影响, 而考虑历史积累可以更有效地衡量一个分类服务在一段时间内的表现. 有:

$$\omega_n(i)=\int_1^{i-1} \frac{Accu_{n,x}\alpha(x)}{\xi} dx \tag{5}$$

其中 $\alpha(x)$ 用于调整不同时期的分类准确率对于本次分类的影响, ξ 为常数.

定义 7 对于 $\forall pa_i \in S$, 定义 DSMM 分类最终结果为 $DSMM_C(pa_i)$. $DSMM_C(pa_i)=\zeta \Leftrightarrow \exists \zeta \in [0, v] \wedge DSMM_C(pa_i, R_\zeta)=\max_{j=0}^{v}(DSMM_C(pa_i, R_j))$.

设网格结点 $node(\hat{s})$ 上的分类服务 $\hat{s}$ 对 $\forall pa_i \in S$ 进行分类所需时间为 $T_{\hat{s}}(pa_i)$, 由于 DSMM 分类模块之下的各个分类服务在网格环境下分布式并行地进行计算, $T_{DSMM}(pa_i)+\max_{j=1}^{n} T_{\tilde{s}_j}(pa_i)+\Delta t$, Δt 为计算式 (3) 所用的时间. 在负载平衡模块的调度下, 最好情况有 $T_{DSMM}(pa_i)=\sum_{j=1}^{n} T_{\tilde{s}_j}(pa_i)/n+\Delta t$.

5 DSMM 生命周期管理模块

DSMM 生命周期管理原理为: 部署于数据网格各个结点分类服务的集合 $C=\{\hat{s}_1, \hat{s}_2, \cdots \hat{s}_n\}$, $|C|=n$, 初始

阶段将分类服务 $\hat{s}_i(i\in[1, n])$ 的生命时间 $T_{life}(i)$ 设为 $T_{initial}$. 每隔 Δt 时间, 对各结点分类服务的分类准确率 (以 ω_i 作为标准) 进行测量, 重新赋予 $T_{life}(i)$ 值. 若有新分类服务 $\hat{s}_i(i> n)$ 则为 $T_{life}(i)$ 赋值 $T_{initial}$; 若 ω_i 为 0, 则证明分类服务 $\hat{s}_i$ 出现故障, 将此服务从系统中删除; 若 ω_i 小于设定的低准确率阈值 *THRESHOLD_low*, 证明 $\hat{s}_i$ 的分类准确率低于分类准确率最低要求, 缩短其生命时间 $T_{life}(i)$; 反之若 ω_i 大于设定的高准确率阈值 *THRESHOLD_high*, 则延长其生命时间. 以伪代码形式表示如下:

if $i> n$, then $T_{life}(i)= T_{initial}$

if $\omega_i= 0$, then $T_{life}(i)= 0$;

else if $\omega_i< THRESHOLD\_low$, then $T_{life}(i)= T_{life}(i)- \Omega$;

else if $\omega_i> THRESHOLD\_high$, then $T_{life}(i)= T_{life}(i)+ \theta$;

其中 Ω 为生命时间减小的步长, θ 为其延长的步长. 若分类服务 $\hat{s}_i$ 存活时间达到或超过 (对于故障结点) $T_{life}(i)$, DSMM 生命周期管理模块将其删除.

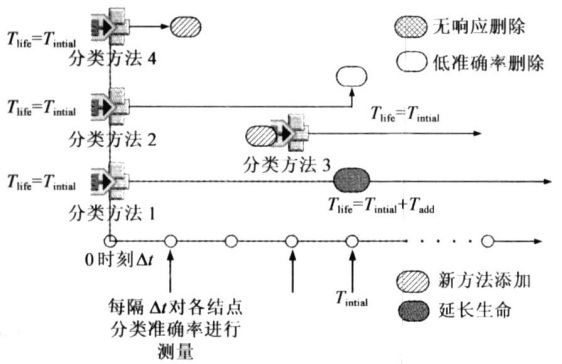

图 2　DSMM 生命周期管理模块

以图 2 所示为例, 有四个分类服务, $C= \{\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4\}$, $|C|= 4$. DSMM 生命周期管理基于对各个结点分类服务分类准确率的监测. 在第 1 个 Δt 时间测量到某结点上分类服务 $\hat{s}_4$ 没有响应, 可以判定为故障, 将 $\hat{s}_4$ 从系统中删去; 在第 3 个 Δt 时间接受新分类服务 $\hat{s}_3$ 的请求, 将其加入到系统中, 生命时间设为 $T_{initial}$; 在第 4 个 Δt 时间测量到 $\hat{s}_1$ 和 $\hat{s}_2$ 的生命时间已达到预设值, 根据测量结果 $\omega_1> THRESHOLD\_high$, 证明 $\hat{s}_1$ 是一个具有高分类准确率的分类服务, 因此延长其生命时间 $T_{life}(1)= T_{life}(1)+ T_{add}$, 而 $\omega_2< THRESHOLD\_high$, 从系统中删除 $\hat{s}_2$.

6　DSMM 负载平衡调度模块

DSMM 的负载平衡调度建立在对结点 CPU 和内存使用情况的监测基础之上. 负载平衡调度模块每隔 Δt 时间对网格中各结点的 CPU 利用率和内存利用率进行监测, 设定阈值 *THRESHOLD_CPU* 和 *THRESHOLD_MEM* 用于衡量节点是否超负荷. 若有结点的测量值高于这两个阈值, 则将此结点上的一个分类服务迁移到 CPU 和内存利用率最低的节点之上.

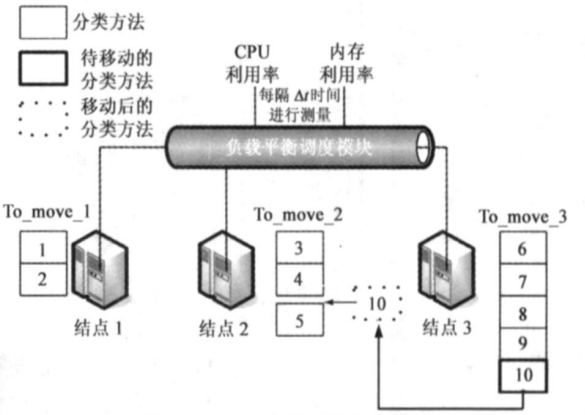

图 3　DSMM 负载平衡调度模块

以如图 3 所示为例, 数据网格中 $C= \{\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4, \hat{s}_5, \hat{s}_6, \hat{s}_7, \hat{s}_8, \hat{s}_9, \hat{s}_{10}\}$, $|C|= 10$.

DSMM 负载平衡调度模块维护所有结点上待迁移服务的队列. 在图 3 中, To_move_1 为 $node_1$ 的待迁移服务队列, 若要迁移 $node_1$ 中的一个分类服务到别的网格结点, 则优先迁移 $To\_move\_1[0]= \hat{s}_2$. 在 Δt 时间的某个整数倍时测量到 $node_3$ 的 CPU 利用率超过 *THRESHOLD_CPU* 或内存利用率超过 *THRESHOLD_MEM*, 则迁移 $node_3$ 上的 $To\_move\_3[0]$ 到 CPU 和内存利用率最低的结点 $node_2$. 在 $node_2$ 上部署 $\hat{s}_{10}$, 将其插入 To_move_2 队列, 最后将 $To\_move\_3[0]= \hat{s}_{10}$ 从 To_move_3 队列中删除. 下面将 DSMM 运用到乳腺癌网格中, 验证上文分析.

7　DSMM 在乳腺癌网格中的应用

下文首先介绍用于实验的乳腺癌网格平台、实验数据来源和相关的行业规定, 再用 DSMM 对实验数据进行分类, 以验证其有效性.

7.1　乳腺癌网格项目

本文用于实验的数据网格是乳腺癌网格, 作为上海高校网格中的一个项目, 有包括上海市瑞金医院、上海复旦大学肿瘤医院和上海交通大学医学部等数家医学研究机构参与其中, 得到了大量实际病例数据的支持, 包括分布于上海市复旦大学附属肿瘤医院的 1004 个病例和上海市瑞金医院的 1923 个病例, 共 2927 个病例, 每个病例占存储空间约 290MB, 其体系结构如图 4 所示.

乳腺癌网格作为一种特定领域上的数据网格, 存储了普查所产生的大量病例数据 (以病例形式、DICOM[14] 输出格式和胶片扫描格式为主), 这些病例数据可以分为几类: 患癌、未患癌或疑似患癌等等. DSMM 作为医学服务之一, 对这些数据集进行分类不管是对于辅助医生的诊断, 还是对于乳腺癌问题的研究都具有非常重要的

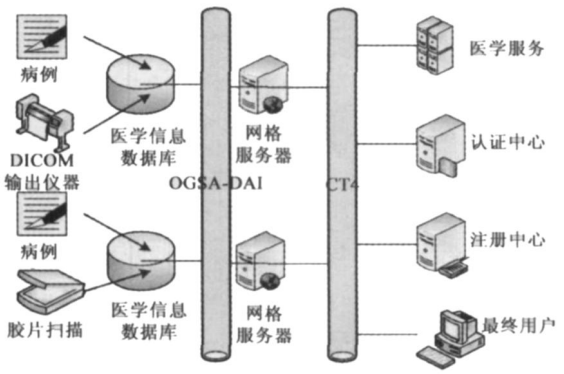

图 4 乳腺癌网格体系结构

意义[3]. 乳腺癌网格严格遵循相关的行业标准: 美国放射学会乳腺影像报告和数据系统(Breast Imaging Reporting And Data System, BI-RADS) 提出乳腺癌的诊断常规共分为七级[5]. 这意味着所有应用于乳腺癌网格中的分类服务都必须能够将某个病例划分为其中的某一级. 将以上行业规定引入到分类服务中, 有 $v = 6$. $pa_i \in R_j$ 代表病例 pa_i 的 BI-RADS 分级为 j, 实验在已有病例基础之上提取一部分属性字段作为训练样本, 包括放射科的诊断结果, 其中 BI-RADS 分级为 0 至 3 的病例与 BI-RADS 分级为 4 以上的病例数量比为 2∶1, 即.

$$\frac{|R_0 \cup R_1 \cup R_2 \cup R_3|}{|R_4 \cup R_5 \cup R_6|} = \frac{2}{1}.$$

平台的搭建基于 Globus 项目组的 Globus Toolkit 4.0 (简称 GT4. 0)[4] 工具包, 所采用的结点包括上海大学的自强 3000 集群式高性能计算机(354 个 Xeon 3.06 GHz CPU, 内存 1GB, 互联网为 10Gbps 的 Infiniband, HP DL360 Cluster/ Cluster Platform) 和上海交通大学医学部结点(3 个 P4 2.66GHz, 内存 512MB, Microsoft Windows XP Professional) , HTTP 服务器软件使用 Tomcat 4.1.29, 结点之间使用 100M 以太网互联. 在自强 3000 的 15 个结点上部署了 15 个分类服务. 由上海交通大学医学部的一个结点做分类结果的整合.

实验首先分别获取分类服务的分类准确率, 进而计算出每种分类服务在 DSMM 模块中所占的权重, 最后由 DSMM 给出最终分类结果 $DSMM_C(pa_i)$.

7.2 分类准确率和耗时

表 1 所示为三个分类服务所用的属性集合 A.

表 1 分类的属性字段

基本信息	钼靶			超声		*MRI*	
姓名	肿块	钙化	结构扭曲	乳腺组织常规描写	病变	检查表现	病变
普查号	数量	数量	数量	位置	大小	伪影	平扫抑脂
初潮年龄	部位	部位	部位	象限	位置	植入物	形态
绝经状态	位置	位置	位置	厚度	形态	有无植入物	信号
末次月经年龄	大小	形态	形态	结构层次	肿块长轴方向	位置	信号是否均匀
首次分娩年龄	形态	分布	外伤史	弥漫改变	边缘	内容物和种类	异常强化病灶
累计哺乳时间	是否伴钙化	范围大小	是否伴钙化		边界	位置	位置
哺乳乳房	伴钙化描述	单发多发	伴钙化描述		内部回声	腺体分型	大小
使用口服避孕药物时间	非对称致密	特殊表现	前片比较		后方回声	腺体分布	病变类型
雌激素替代治疗时间	球形非对称	密度	有无前片		周围组织	前片比较	伴随征象
夜班频率	局限性非对称	皮肤收缩	检查项目		类型	有无前片	动态曲线
烟酒习惯		乳头内陷	检查日期		钙化分布	检查项目	起始期
既往胸部放疗史		皮肤增厚	比较结果		彩色多普勒血流	检查日期	延迟期
既往化疗史		小梁增粗				肿块样强化	检查印象
乳腺炎		皮肤病变				边缘	强化

对于不同的 B, 考虑到不同情况下相邻的分级情况对于分类准确率的影响, $\Psi_i \times 49$ 的选取如图 5 所示.

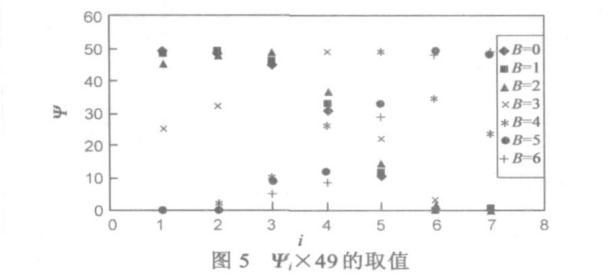

图 5 $\Psi_i \times 49$ 的取值

7.3 生命周期管理

如图 6 所示为生命周期管理模块调度下的四类情况. 其中 $s_1 = \{\hat{s}_5, \hat{s}_{11}\}$; $s_2 = \{\hat{s}_2, \hat{s}_3, \hat{s}_7\}$; $s_3 = \{\hat{s}_1, \hat{s}_4, \hat{s}_6, \hat{s}_{10}, \hat{s}_{12}, \hat{s}_{14}, \hat{s}_{15}\}$; $s_4 = \{\hat{s}_8, \hat{s}_9, \hat{s}_{13}\}$. s_1 类方法出现故障, 在第 1 个 Δt (100s) 时间内被删除; s_2 类方法分类准确率小于阈值, 在 $T_{\text{initial}} = 600$s 后被删除; s_3 类方法分类准确率大于阈值, 生命周期被延长, s_4 类方法在 100s 时加入, 分类准确率大于阈值, 生命周期被延长.

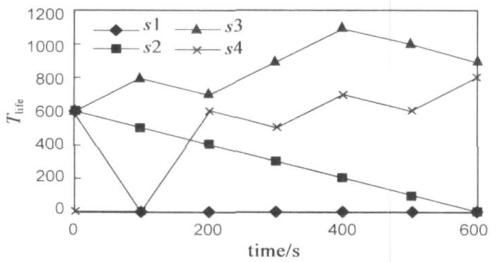

图 6　生命周期管理

7.4　负载平衡管理

负载平衡监控 $\Delta t = 100s$, 实验记录在两个结点之间进行负载调度的情况, 如图 7 所示. 0s 时, 结点 1 过载而结点 2 几乎空载. 经过两次调度后, 结点 1 和结点 2 的负载基本平衡.

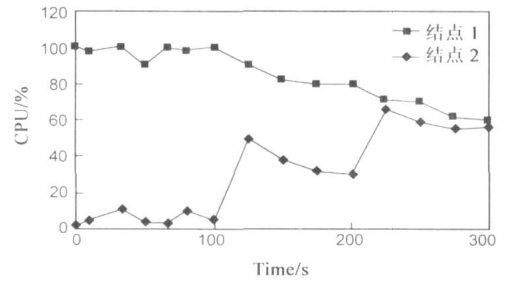

图 7　负载平衡管理

7.5　DSMM 的分类准确率

在实验中, DSMM 采用考虑历史积累的权重选择方法, 即 $\omega_n(i) = \int_1^{i-1} \frac{A\alpha cu_{n,x}\alpha(x)}{\zeta} dx$, 选择 $\alpha(x)$ 为 $2^{x/500}$, $\zeta = \sum_{j=1}^{i-i} 2^{j/500}$. 使时间越近的分类准确率对权重的影响越大, 符合分类服务的分类准确率逐渐增大的过程.

DSMM 的最终结果可由(6)算得:

$$DSMM_C(pa_i, R_j) = \frac{\omega_1(i)\tilde{s}_1(pa_i, R_j) + \omega_2(i)\tilde{s}_2(pa_i, R_j) + \cdots + \omega_{30}(i)\tilde{s}_3(pa_i, R_j)}{\eta(i)} \quad (6)$$

中, 取 $\eta(i) = \sum_{\zeta=1}^{30} \omega_\zeta(i)$. DSMM 对待分类数据集 S 进行分类的分类准确率 $Accu_{DSMM,i}$ 如图 8 所示.

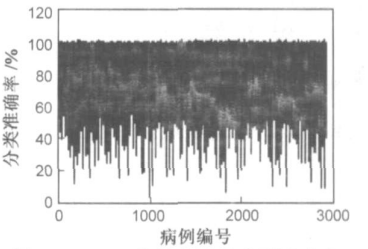

图 8　DSMM 分类方法的分类准确率

采用经典分类方法: 判定树分类方法、贝叶斯分类方法和贝叶斯分类方法的分类准确率如图 9, 图 10 和图 11 所示.

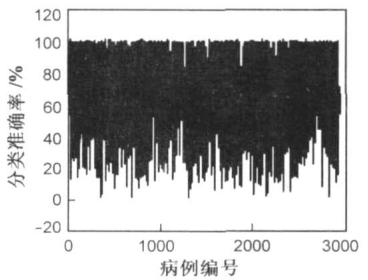

图 9 判定树分类方法

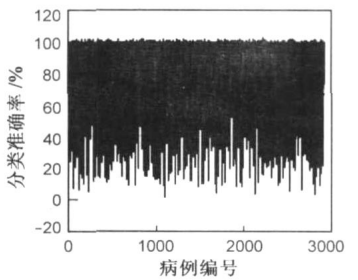

图 10 贝叶斯分类方法

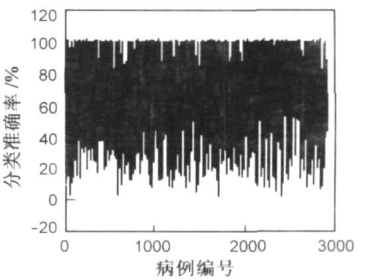

图 11 贝叶斯网络

三种分类服务耗时的平均值分别为: 985ms, 836ms, 1012ms. DSMM 的分类准确率 $Accu_{DSMM,i}$ 与三种分类服务比较如表 2 所示.

表 2　DSMM 与其余三种的比较

分类准确率 / 分类服务	0– 20 (%)	20– 40 (%)	40– 60 (%)	60– 80 (%)	80– 100 (%)
DSMM	22 人	122 人	559 人	650 人	1574 人
判定树分类方法	91 人	268 人	791 人	1038 人	739 人
贝叶斯分类方法	108 人	362 人	551 人	606 人	1300 人
贝叶斯网络	88 人	243 人	740 人	1086 人	770 人

可见, DSMM 大幅度降低了分类准确率在 0% – 20% 和 20% – 40% 的病例数目, 同时增加了分类准确率在 80% – 100% 的病例数目. DSMM 的分类准确率确实高于其它三种分类服务单独使用的分类准确率, DSMM 平均耗时为 1089ms, 与其它三种分类服务的耗时: 985ms, 836ms, 1012ms 相比较, 在几乎不影响时间消耗的前提下, 很大程度上提高了系统分类准确率.

8　结论

本文设计实现了一种基于数据网格的 DSMM 系统, 并将其运用到分布在数据网格各结点的 2927 个乳腺癌病例数据之上. 实验证明, 由于单个分类服务在网格环境中只能使用单个结点, 不可避免地产生了较多分类准确低的分类结果, 而采用 DSMM 对数据网格中存储的数据集进行分类时, 在耗时没有明显增加的前提下, 将分类准确率低的分类结果数量大大降低, 从整体上提高了分类准确率. DSMM 充分利用了网格中各结点的计算能力, 进行了生命周期管理和负载平衡调度, 契合于网格环境, 更能发挥网格作为一个分布式系统

的分布式计算能力. DSMM 适应网格松耦合的特性并最大程度地合理利用了结点的资源.

参考文献:

[1] K Chelluri, V Kumar. Data classification and management in very large data warehouses[A]. Advanced issues of E-Commerce and web-based information systems, WECWIS, Third International Workshop[C]. Washington, DC, USA: IEEE Computer Society, 2001. 52-57.

[2] E Lamma, P Mello, A Nanetti, et al. Artificial intelligence techniques for monitoring dangerous infections[J]. IEEE Transactions on Information Technology in Biomedicine, 2006, 10(1): 143-155.

[3] B Kovalerchuk, E Vityaev, J F Ruiz. Consistent knowledge discovery in medical diagnosis[J]. IEEE Transactions on Engineering in Medicine and Biology Magazine, 2000, 19(4): 26-37.

[4] Globus[OL]. http://www.globus.org, 2006.

[5] American college of radiology. BI-RADS: mammography[S]. In: Breast Imaging Reporting and Data System: BI-RADS Atlas. 4th ed. Reston, Va: American College of Radiology. 2003.

[6] M Mertik, P Kokol, B Zalar. Gaining features in medicine using various data-mining techniques[A]. IEEE 3rd International Conference on Computational Cybernetics: ICCC[C]. IEEE Press, 2005. 21-24.

[7] N Bogunovic, V Marohnic, Z Debeljak. Efficient gene expression analysis by linking multiple data mining algorithms[A]. Engineering in Medicine and Biology Society: IEEE-EMBS[C]. IEEE Press, 2005. 4830-4833.

[8] F A Thabtah, P Cowling, YH Peng. MMAC: a new multi-class, multi-label associative classification approach[A]. Data Mining, Proceedings. Fourth IEEE International Conference[C]. IEEE Press, 2004. 217-224.

[9] OGSA-DAI[OL]. http://www.ogsadai.org.uk, 2006.

[10] F Pan, B Wang, X Hu, et al. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis[J]. Biomedical Informatics, 2004, 37(4): 240-248.

[11] P Radivojac, N V Chawla, et al. Classification and knowledge discovery in protein databases[J]. Biomedical Informatics. 2004, 37(4): 224-239.

[12] X B Zhou, K Y Liu, S T C Wong. Cancer classification and prediction using logistic regression with Bayesian gene selection[J]. Biomedical Informatics, 2004, 37(4): 249-259.

[13] A Li, Y M Yang. Using recursive classification to discover predictive features[A]. In Proc ACM symposium on Applied computing[C]. ACM Press, 2005. 1054-1058.

[14] DICOM homepage[OL]. http://medical.nema.org, 2003.

[15] EDiaMoNd[OL]. http://www.ediamond.ox.ac.uk, 2006.

[16] Mammogrid[OL]. http://mammogrid.vitamib.com/, 2006.

[17] A Chervenak, I Foster, C Kesselman, et al. The data grid: towards an architecture for the distributed management and analysis of large scientific data sets[J]. Network and Computer Applications, 2001, 23(3): 187-200.

作者简介:

王 坤 女, 1982 年 11 月生于贵州兴义, 2004 年毕业于清华大学计算机系, 其后在清华大学计算机系高性能所攻读博士学位. 现为硕博连读生, 主要研究方向为网格计算技术和高性能计算等.
E-mail: wangkun00@mails.tsinghua.edu.cn

李三立 男, 教授, 中国工程院院士, 博士生导师, 1935 年生于上海, 1955 年毕业于清华大学无线电系, 1960 年获苏联科学院博士学位, 以后在清华大学任教至今. 现任清华大学计算机科学与工程研究所所长, 兼任上海大学计算机学院院长. 研究方向为网格计算技术和高性能计算技术等. E-mail: lsl-dcs@mail.tsinghua.edu.cn

柴云鹏 男, 1983 年生于内蒙古, 2004 年毕业于清华大学计算机系, 其后在清华大学计算机系高性能所攻读博士学位. 现为硕博连读生, 主要研究方向为网格计算技术和流媒体处理等. E-mail: chaiyunpeng@china.com

王晓英 女, 1982 年生于内蒙古, 2003 年毕业于清华大学计算机系, 其后在清华大学计算机系高性能所攻读博士学位. 现为硕博连读生, 主要研究方向为网格计算技术和高性能计算技术等. E-mail: wangxy@tirc.cs.tsinghua.edu.cn