한국시뮬레이션학회 논문지
Vol. 19, No. 2, pp. 81-89 (2010. 6)

# 동적 로드 밸런싱을 이용한 그리드 기반의 생물학 데이터 마이닝

마용범[1] · 김태영[1] · 이종식[1†]

# Grid-based Biological Data Mining using Dynamic Load Balancing

Yong Beom Ma · Tae Young Kim · Jong Sik Lee

**ABSTRACT**

Biological data mining has been noticed as an issue as the volume of biological data is increasing extremely. Grid technology can share and utilize computing data and resources. In this paper, we propose a hybrid system that combines biological data mining with grid technology. Especially, we propose a decision range adjustment algorithm for processing efficiency of biological data mining. We obtain a reliable data mining recognition rate automatically and rapidly through this algorithm. And communication loads and resource allocation are key issues in grid environment because the resources are geographically distributed and interacted with themselves. Therefore, we propose a dynamic load balancing algorithm and apply it to the grid-based biological data mining method. For performance evaluation, we measure average processing time, average communication time, and average resource utilization. Experimental results show that this method provides many advantages in aspects of processing time and cost.

**Key words** : Dynamic load balancing, Grid computing, Biological data mining

**요 약**

생물학 데이터 마이닝은 생물학 데이터의 볼륨이 급격하게 증가함에 따라 최근 주목받고 있다. 그리드 기술은 계산 자원과 데이터 공유와 활용을 가능하게 한다. 이 논문에서는 생물학 데이터 마이닝과 그리드 기술을 결합한 혼합형 시스템을 제안한다. 특히, 생물학 데이터 마이닝의 처리 효율성을 위해 결정 범위 조정 알고리즘을 사용한다. 우리는 이 알고리즘을 통해 빠르고 자동으로 신뢰할 만한 데이터 마이닝 인식률을 얻는다. 게다가 그리드 환경에서는 지리적으로 분산된 자원들을 연동하기 때문에 통신량과 자원 할당이 이슈가 된다. 우리는 동적 로드 밸런싱을 제안하고 그리드 기반 생물학 데이터 마이닝 기법에 적용한다. 성능 평가를 위해 우리는 평균 처리 시간, 평균 통신 시간, 평균 자원 활용도를 측정한다. 측정 실험의 결과는 제안된 두 알고리즘을 적용한 우리의 기법이 처리 시간과 비용 측면에서 이점을 제공한다는 것을 보여준다.

**주요어** : 동적 로드 밸런싱, 그리드 컴퓨팅, 생물학 데이터 마이닝

## 1. 서 론

최근 생물학 연구가 많은 관심을 받고 급속한 성장을 함에 따라 데이터로부터 유용한 지식을 발견할 수 있는 높은 지능을 갖춘 데이터 분석 방법론의 필요가 증가되었다. 생물학 데이터 마이닝(Wang 등, 2005; Tzanis 등, 2005)은 인간의 감각기관을 통해 인지된 정보를 처리, 분석하여 이해한 정보를 전달하는 형태로 컴퓨터를 통해 인공지능 기술로 사용된다. 또한 생물학 데이터 마이닝은 게놈, 단백질, DNA, RNA와 같은 대용량의 생물학 데이터를 분석하고 처리하기 위한 통계학, 분자생물학, 계산 방법들의 결합을 포함한다. 단백질 데이터와 같은 순차적인 데이터에 관한 연구는 주로 데이터 마이닝을 위한 분류기에 관한 연구가 많이 이루어지고 있다. 그러나 이러한 생물학 데이터는 매우 방대한 데이터이기 때문에 고성능의 컴퓨팅 기술이 필요하다.

그리드 기술은 컴퓨팅과 데이터 접근의 방법을 완전하

*이 논문은 인하대학교의 지원에 의하여 연구되었음.
2010년 1월 26일 접수, 2010년 3월 31일 채택
[1] 인하대학교 정보공학과
주 저 자 : 마용범
교신저자 : 이종식
E-mail; myb112@hanmail.net

게 바꿀 수 있는 잠재력을 지닌 새로운 기술이다. 그리드 컴퓨팅(Foster 등, 1998; Berman 등, 2003)은 서로 다른 개인이나 조직에 의해 소유된 지리적으로 분산된 이기종의 자원들을 활용하여 대용량의 컴퓨팅 문제를 해결할 수 있다. 또한 데이터와 컴퓨팅 파워뿐만 아니라 사람의 전문적 기술까지도 결합할 수 있다. 그리드 컴퓨팅은 데이터를 재사용 할 수 있다는 장점을 갖고 있기 때문에 복잡한 데이터를 처리하는 시스템에 적용함으로써 시스템 성능을 향상시킬 수 있고 데이터를 처리하거나 분석하는 데 필요한 시간과 비용을 줄일 수 있다. 이러한 그리드 컴퓨팅 시스템은 많은 요소로 이루어진 지리적으로 원거리의 시스템들 사이에서 실시간 연결을 요구한다. 그러므로 복잡한 대용량 실행과 지리적으로 분산된 데이터 세트와 컴퓨팅 자원을 협동적으로 공유해야 한다. 그러나 계산 자원 간 통신 속도는 매우 다양하기 때문에 통신량의 축소뿐만 아니라 적절한 자원 할당이 부가적으로 고려되어야 한다. 따라서 그리드 컴퓨팅 환경에서의 효율적인 작업 실행을 위해서는 작업량의 균형 및 적절한 작업 배치가 동시에 수행되어야 한다.

우리는 그리드 기술과 생물학 데이터 마이닝을 결합한 그리드 기반의 생물학 데이터 마이닝 시스템을 제안한다. 또한, 처리 속도과 비용에 초점을 두고 성능 향상을 위해 결정 범위 조정 알고리즘과 동적 로드 밸런싱 알고리즘을 제안한다. 결정 범위 조정 알고리즘은 자동적인 결정 범위 조정을 통해 처리 효율성을 높이고 동적 로드 밸런싱 알고리즘은 복잡한 데이터를 처리하기 위한 부하를 분산시키는 데 중요한 역할을 한다. 제안하는 두 알고리즘을 적용한 그리드 기반의 생물학 데이터 마이닝 시스템의 성능 평가를 위해 평균 처리 시간, 평균 통신 시간, 평균 자원 활용도 등을 측정한다.

이 논문은 다음과 같이 구성된다. 2장은 로드 밸런싱에 대해 논의한다. 3장은 동적 로드 밸런싱을 가진 그리드 기반 생물학 데이터 마이닝에 대해 자세히 설명한다. 4장은 성능 평가 실험을 통해 제안하는 기법의 유용성을 증명한다. 마지막으로 결론은 5장에서 논한다.

## 2. 관련 연구

### 2.1 로드 밸런싱

이 장에서는 그리드 컴퓨팅 환경과 같은 적절한 자원 할당이 필요한 분산된 컴퓨팅 환경에서 발생 가능한 오버헤드를 해결하기 위해 사용되는 로드 밸런싱(Kumar 등, 1994a, 1994b) 기법들에 대해 소개한다. 로드 밸런싱은 프로세서와 같은 작업을 처리하는 노드들이 작업을 균형있게 분배할 수 있도록 도와줌으로써 시스템 성능을 향상시키는 역할을 한다. 이러한 로드 밸런싱 작업은 크게 정적 로드 밸런싱(Kim 등, 1990)과 동적 로드 밸런싱(Kumar 등, 1994; Wikinson 등, 1999)으로 구분할 수 있다. 정적 로드 밸런싱은 로드 밸런싱 기법 중에서 가장 간단한 기법으로 어떠한 작업을 실행하기 전에 작업 할당이 미리 정의된다. 그러나 정적 로드 밸런싱은 미리 정의해야 하는 기법이기 때문에 실행될 작업에 적합하도록 미리 예측한 작업량의 할당이 실제 수행되는 작업과 다를 수 있다. 따라서 정확하게 작업량을 할당하지 못 한다면 매우 불균형한 작업 할당이 될 수 있으며, 이것은 오히려 시스템의 성능을 저하시키는 결과를 가져온다. 일반적으로 정적 로드 밸런싱은 각각의 작업 처리자가 동일한 작업량을 가지고 있는 경우와 같은 제한적인 환경에서 유용하다. 동적 로드 밸런싱은 작업이 실행되는 동안 해당 작업을 각각의 작업 처리자의 능력에 맞게 작업을 할당하는 기법으로 정적 로드 밸런싱보다 다소 복잡하고 밸런싱 알고리즘을 위한 추가적인 오버헤드가 발생할 수 있지만 일반적으로 더 효과적이다.

생물학 데이터와 같은 대용량 데이터의 처리에 대한 연구가 활발히 이루어지면서 데이터의 처리 시간을 감소시키고 비용을 절감하기 위한 방법으로 동적 로드 밸런싱 기술에 대한 연구도 많이 진행되고 있다. LOADIST (Sureswaran 등, 1995)는 부하 공유의 개념으로 평균 부하에 기반하여 유휴 상태의 작업자에게 작업을 할당하는 동적 로드 밸런싱 방법이다. 에이전트 기반의 동적 로드 밸런싱 기법인 DASH(Rajagopalan 등, 2000)는 임계값을 가지고 노드의 부하가 임계값보다 크면 로드 밸런싱을 실시한다. 게다가 로드 밸런싱은 전송되거나 옮겨질 작업이 먼저 선택되는데 동적으로 작업을 선택하기 위해 작업의 실행 시간을 통계학적으로 예측하는 휴리스틱 알고리즘을 사용한다. Zaki 등(Zaki, 1996)은 실행 시간을 최소화하고 각각의 프로세서의 성능에 적합한 작업을 할당하는 맞춤형 동적 로드 밸런싱 기법을 제안했다.

이 논문에서는 생물학 데이터와 같은 복잡한 데이터를 효과적으로 처리하기 위해 그리드 컴퓨팅 기술과 결합한 그리드 기반의 생물학 데이터 마이닝을 제안한다. 또한, 동적 로드 밸런싱을 그리드 기반의 생물학 데이터 마이닝에 적용한다. 그리드 컴퓨팅 환경 하에서 자원을 제공하는 자원 제공자의 연산 능력을 기반으로 작업 부하를 동적으로 할당하기 때문에 작업 지연 시간을 줄일 수 있고 처리 효율 및 비용 측면에서도 이점을 얻을 수 있다.

## 3. 그리드 기반의 생물학 데이터 마이닝

앞서 말했듯이, 생물학 데이터는 데이터의 양이 매우 방대하기 때문에 생물학 데이터 마이닝을 위해서는 이러한 대용량의 데이터 처리에 따른 작업 부하가 발생한다. 이러한 작업 부하 문제를 해결하기 위해 그리드 컴퓨팅과 같은 기술을 사용하여 지리적으로 분산된 자원들을 활용한다. 그리드 컴퓨팅과 같은 분산된 자원을 활용하는 환경에서는 고성능의 연산 능력을 발휘하는 것도 중요하지만 각각의 제공받는 컴퓨팅 파워, 즉 연산 능력에 따라 작업 부하를 균등하게 분배하는 것 또한 매우 중요하다. 각각의 연산 능력에 맞는 작업을 분배하는 것은 해당 작업을 제 시간에 처리할 수 있다는 것을 의미하고 이것은 데이터 처리 시간의 감소와 비용 절감 등의 영향을 미칠 수 있다. 이러한 작업 부하의 연산 능력에 따른 분배를 위해 우리는 그리드 기반의 생물학 데이터 마이닝 시스템에 동적 로드 밸런싱을 적용한다. 동적 로드 밸런싱은 시스템의 상태가 변화할 때마다 프로세서를 재배치하거나 성능에 비례하여 각 프로세스에게 작업들을 할당하고 이를 통해 각 프로세스의 작업 부하를 줄이고 데이터 처리 시간 및 비용을 줄일 수 있는 장점이 있다.

이 장에서 우리는 그리드 기반의 생물학 데이터 마이닝 시스템을 묘사하고 그리드 기반의 생물학 데이터 마이닝 시스템에서 사용된 결정 범위 조정 알고리즘과 동적 로드 밸런싱 알고리즘을 제안한다.

### 3.1 그리드 기반 생물학 데이터 마이닝 설계

우리는 그림 1과 같이 그리드 기반의 생물학 데이터 마이닝 시스템을 제안한다. 그림 1에서 보여지는 것처럼, 그리드 기반의 생물학 데이터 마이닝 시스템은 하는 역할에 따라 3계층으로 분류될 수 있다.

가장 낮은 레벨인 구조 계층에서, 우리는 PC나 슈퍼컴퓨터, 저장 시스템, 여러 가지 형태의 센서들과 같은 다양한 자원들을 가진다. 그 자원들은 소유주가 서로 다르고 서로 다른 연산 능력을 갖고 있다. 구조 계층은 오직 연산 자원의 물리적 상호작용을 인식하고 자원들의 공유는 미들웨어 계층에서 이루어진다. 이 계층에서는 연산 자원 제공을 위해 미들웨어 계층과 통신을 한다.

미들웨어 계층은 연산 자원의 이기종이라는 특징과 분배를 포함하여 그리드 응용 계층을 위한 공동의 인터페이스를 제공한다. 이러한 공동의 인터페이스 제공을 통해 사용자는 연산 자원의 위치나 능력에 있어서 차이를 느끼지 못 한다. 사용자와 그리드 환경 간의 상호 작용을 위해

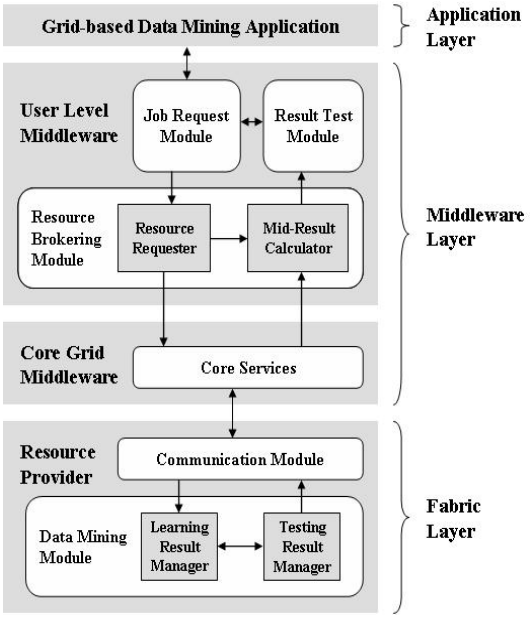


**그림 1.** 그리드 기반 생물학 데이터 마이닝 시스템

미들웨어 계층은 사용자 레벨 미들웨어와 코어 그리드 미들웨어를 포함한다. 사용자 레벨 미들웨어는 응용 프로그램 개발을 위한 환경 및 툴과 응용 프로그램 작업들을 스케줄링하고 자원을 관리하는 자원 중개를 담당한다. 사용자 레벨 미들웨어 내에는 다수의 자원 중개자가 존재하고 이들은 자원 중개 서비스를 담당한다. 코어 그리드 미들웨어는 원격 프로세스 관리, 정보 등록과 발견, 자원 예약, 보안과 같은 코어 서비스들을 제공한다.

미들웨어 계층의 상위 계층은 응용 계층이다. 이 계층에서는 그리드 기반의 데이터 마이닝 응용 프로그램처럼 고성능의 계산 능력을 요구하는 응용 프로그램들을 포함한다.

#### 3.1.1 작업 요청 모듈

작업 요청 모듈은 응용 계층으로부터 데이터 마이닝 응용 프로그램 작업들을 전달받아 작업을 각 자원 중개자에게 할당하는 역할을 한다. 데이터 마이닝 작업은 고성능 연산을 필요로 하는 대용량의 작업이기 때문에 부하가 크다. 따라서, 작업 요청 모듈은 작업 부하 분산을 위해 각 자원 중개 모듈에게 동일하게 작업을 분산시킨다. 게다가 작업 요청 모듈은 기존의 그리드 미들웨어와 같이 SOAP(Simple Object Access Protocol) 메세징 기법을 사용하며 인터넷을 통한 연결이 용이하다.

#### 3.1.2 결과 테스트 모듈

결과 테스트 모듈은 모든 중간-결과가 도착한 경우, 데이터 마이닝 작업의 통합된 결과를 테스트한다. 만약, 데이터 마이닝 작업이 신뢰할 만한 데이터 마이닝 인식률을 보인다면 결과 테스트 모듈은 현재의 결정 범위와 작업 결과를 작업 요청 모듈에게 전송한다. 그렇지 않다면, 결과 테스트 모듈은 결정 범위를 조정하고 작업 요청 모듈에게 업데이트된 결정 범위를 전송한다. 결과 테스트 모듈에서의 결정 범위 조정 알고리즘은 3.2장에서 나타난다.

#### 3.1.3 자원 중개 모듈

자원 중개 모듈은 사용자 레벨 미들웨어의 핵심 모듈로 응용 프로그램과 자원 사이에서 중개자 역할을 한다. 자원 중개 모듈의 기본적인 기능은 자원 사용에 대한 요청과 중간 결과 통합이다. 이 기능을 수행하기 위해 자원 중개 모듈은 자원 요청기와 중간-결과 산출기 두 가지 하위 모듈을 포함한다. 자원 요청기는 작업 요청 모듈으로부터 데이터 마이닝 작업 일부를 전송받고 동적 로드 밸런싱을 통해 자원 사용을 요청한다. 동적 로드 밸런싱 알고리즘은 3.2장에서 자세히 설명된다. 각 자원 중개 모듈은 동적으로 자원 사용을 요청한 각 자원 제공자로부터 결과를 전송받고 중간-결과 산출기는 이러한 중간-결과를 산출한다. 각 자원 중개 모듈로부터 산출된 중간-결과는 결과 테스트 모듈에게 전송된다. 결과 테스트 모듈이 신뢰할 만한 인식률을 얻을 때까지 자원 중개 모듈은 자원 요청과 중간-결과 산출 작업을 반복한다.

#### 3.1.4 코어 서비스

코어 서비스는 자원들에 안전하고 투명한 접근을 제공하기 위한 모듈이다. 코어 서비스 모듈은 구체화되고, 상호작용, 통신, 관리되기 위한 서비스들의 기본 토대가 되는 방법들을 제공한다.

#### 3.1.5 통신 모듈

통신 모듈은 코어 그리드 미들웨어와 자원 제공자 사이의 통신을 관리한다. 이 모듈은 코어 그리드 미들웨어의 코어 서비스와 연결되고 각 데이터 마이닝 모듈의 데이터 마이닝 결과들을 코어 서비스에게 전달한다.

#### 3.1.6 데이터 마이닝 모듈

데이터 마이닝 모듈은 할당된 작업에서 필요로 하는 실제 자원을 제공하여 데이터 마이닝 작업을 처리한다. 각 데이터 마이닝 모듈은 서로 다른 연산 능력을 가지고 있으며 자신의 연산 자원을 이용하여 학습과 테스팅을 수행한다. 두 가지 주요 기능을 수행하기 위해 데이터 마이닝 모듈은 학습 결과 관리기와 테스팅 결과 관리기를 포함한다. 학습 결과 관리기는 데이터 마이닝을 위해 현재의 결정 범위를 통해 학습을 반복하고 그 결과를 테스팅 결과 관리기에게 전송한다. 테스팅 결과 관리기는 학습 결과 관리기로부터 전송된 결정 범위와 학습 결과를 토대로 테스팅을 실시한다. 각 테스팅 결과 관리기는 통신 모듈을 거쳐 미들웨어 계층의 자원 중개 모듈에게 결과를 전송한다.

### 3.2 결정 범위 조정 알고리즘과 동적 로드 밸런싱 알고리즘

이 장에서 우리는 자동적으로 신뢰할 만한 인식률을 찾기 위한 결정 범위 조정 알고리즘과 대용량의 데이터 처리에 따른 작업 부하를 분산시키기 위한 동적 로드 밸런싱 알고리즘을 자세히 설명한다.

#### 3.2.1 결정 범위 조정 알고리즘

이 논문에서는 생물학 데이터 마이닝에서 쓰이는 방법 중에서 표준 통계적 방법보다 더 효과적이고 반복적인 학습을 통해 에러를 감소시키고 예측에 의해 문제를 해결하는 학습 기반의 신경망 모델을 이용하는 것에 초점을 둔다. 이러한 모델들에서 학습은 기존의 예제나 경험에 의해 얻어진 데이터를 통해 이루어지고 무수히 많은 신경들 간 시냅틱 연결의 수렴을 통해 자동으로 새로운 학습 지식을 얻기 때문에 분류하는 방법에 대한 상세가 필요하지 않다. 또한, 반복적인 학습을 통해 일반화된 결과를 얻을 수 있다. 그러나 이러한 반복적인 학습을 수행하는 신경망 모델들은 학습에 따른 신경들 간 연결 가중치를 갖는 경우가 많다. 연결 가중치는 데이터를 분류하기 위해 시냅틱 연결의 수렴을 돕는 역할을 하며 데이터 분류와 밀접한 관계에 있다.

신경망 모델에서 모든 데이터 값들은 0~1 사이의 값을 가지고 두 개 혹은 세 개의 클래스로 구분될 수 있다. 결정 범위는 클래스를 구분하는 데 결정적인 역할을 하는 매개변수이기 때문에 데이터 마이닝에서 인식률 즉, 임의의 데이터가 속하는 클래스로 분류될 확률과 밀접한 관계에 있다. 결정 범위 값이 클 경우, 넓은 범위의 데이터들이 하나의 클래스에 포함될 수 있고 모든 클래스들의 인식 범위가 넓기 때문에 데이터 분류는 모호해질 수 있다. 반대로, 결정 범위가 작을 경우, 좁은 범위의 데이터들이 포함되고 정확한 데이터들만 인식될 수 있다. 그러나 결

```
if ( CurrentDataMiningRate < ReliableDataMiningRate – α )
   DecisionRange = DecisionRange – δ
elseif ( CurrentDataMiningRate > ReliableDataMiningRate + α )
   DecisionRange = DecisionRange + δ
elseif (CurrentDataMiningRate ≤ |ReliableDataMiningRate| )
   return DecisionRange
else
   DecisionRange is initialized
end if
```

**그림 2.** 결정 범위 조정 알고리즘

정 범위가 너무 작으면 해당 클래스에 포함되더라도 인식이 되지 못 하는 경우가 발생한다. 따라서 우리는 신뢰할 만한 데이터 마이닝 인식률을 보일 때까지 결정 범위를 자동으로 반복하여 조정하는 결정 범위 조정 알고리즘을 제안한다. 결정 범위 조정을 위한 의사코드는 그림 2에서 주어진다.

그림 2에서 α는 데이터 마이닝 인식률 한계 상수를 나타내고 δ는 결정 범위 변동 계수를 나타낸다. 둘 모두 0부터 1 사이의 값을 가지고 데이터 마이닝 인식률에 따라 바뀐다. 데이터 마이닝 인식률은 알파를 통해 상한값과 하한값을 가지고 이것은 신뢰할 만한 데이터 마이닝 인식률의 범위를 나타낼 수 있게 한다. 현재의 데이터 마이닝 인식률이 신뢰할 만한 데이터 마이닝 인식률의 하한값 보다 낮은 경우는 아직 신뢰할 만한 인식률을 나타내지 못한다는 것을 의미한다. 그러므로 현재의 결정 범위는 결정 범위 변동 계수를 빼서 구할 수 있다. 앞에서 언급했듯이, 결정 범위가 작아지면 인식률은 높아진다. 만약 신뢰할 만한 데이터 마이닝 인식률의 상한값 보다 높은 경우 결정 범위는 결정 범위 변동 계수를 더한다. 결정 범위 변동 계수는 데이터 마이닝 인식률과 신뢰할 만한 데이터 마이닝 인식률의 차에 비례한다. 따라서 이 알고리즘은 자동적이고 빠르게 신뢰할 만한 데이터 마이닝 인식률을 얻을 수 있게 해준다.

### 3.2.2 동적 로드 밸런싱 알고리즘

데이터 마이닝 작업은 자원 제공자의 결과로부터 계산된 통합된 결과가 얻어졌을 때 테스트될 수 있다. 따라서 자원 제공자들의 상태에 따라 작업을 할당하는 것은 매우 중요하다. 이 장에서 우리는 대용량의 데이터 마이닝 작업의 부하를 분산시키기 위해 맞춤형 동적 로드 밸런싱 알고리즘을 제안한다. 이 논문은 대용량 데이터 마이닝 작업의 효율적인 처리에 초점을 두고 있기 때문에 우리는

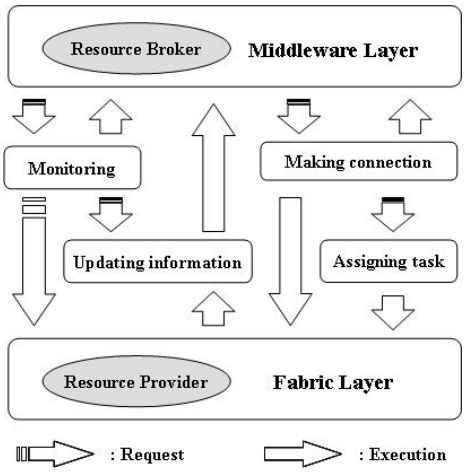


**그림 3.** 동적 로드 밸런싱 알고리즘

그리드 자원들의 다양한 특성은 고려하지 않고 단지 연산 능력만 고려한다. 다른 특성들은 차후 연구에서 고려될 것이다. 그리드 기반의 생물학 데이터 마이닝 작업은 미들웨어 계층에서 몇 명의 자원 중개자에게 동일하게 할당된다. 그 결과, 모든 자원 중개자는 동일한 양의 작업을 할당받게 될 수 있고 할당된 작업은 하위의 자원 제공자에게 분산되어 재할당 된다. 자원 중개자는 각 자원 제공자들의 연산 능력을 모니터링하고 이 정보를 업데이트한다. 그리고 각 자원 제공자와 새로운 연결을 만들고 연결된 자원 제공자에게 작업을 할당한다. 이 때, 각 자원 중개자는 서로 다른 연산 능력을 가진 자원 제공자들의 연산 자원을 기반으로 하여 작업을 할당한다. 위의 과정을 그림으로 나타내면 그림 3과 같다.

## 4. 실험 및 결과

그리드 기반의 생물학 데이터 마이닝 시스템의 성능을 평가하기 위해 우리는 그림 4와 같이 RTI(DoD, 1998) 실행을 기반으로 하는 시뮬레이션 환경을 구축하고 로드 밸런싱을 전혀 하지 않는 그리드 기반의 생물학 데이터 마이닝 기법(Grid-based Biological Data Mining; GBDM)과 정적 로드 밸런싱을 이용하는 그리드 기반의 생물학 데이터 마이닝 기법(Grid-based Biological Data Mining with Static Load Balancing; GBDM with SLB), 동적 로드 밸런싱을 이용하는 그리드 기반의 생물학 데이터 마이닝 기법(Grid-based Biological Data Mining with Dynamic Load Balancing; GBDM with DLB)을 비교한

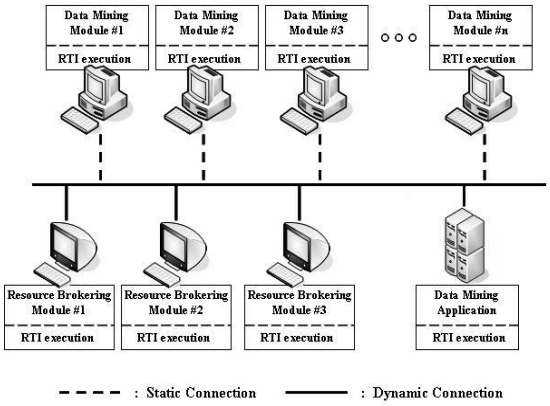


**그림 4.** RTI 기반의 시뮬레이션 환경

다. 각각의 데이터 마이닝 모듈과 자원 중개 모듈은 RTI 실행을 통해 동작하며 데이터 마이닝 응용 프로그램 역시 하나의 모듈과 같이 RTI 실행을 통해 동작한다. RTI는 분산된 운영 시스템의 실행시간 동안 각 모듈 간 상호 작용을 위해 응용 프로그램에 공통된 인터페이스 서비스들을 제공한다. 또한, 신뢰성 있는 실험을 위해 GBDM, GBDM with SLB, GBDM with DLB 모두 GPCR(Watson 등, 1994) 데이터를 사용하여 두 개의 클래스로 분류를 하였고 실험 상에서 신뢰할 만한 인식률은 90%로 가정하였다.

### 4.1 평균 통신 시간

통신 시간은 크게 두 가지로 구분된다. 하나는 데이터 마이닝 응용 프로그램과 자원 중개 머신 간 통신 시간이고, 다른 하나는 자원 중개 머신과 자원을 제공하는 학습 머신 간의 통신 시간이다. 통신 시간은 네트워크의 상태와 같은 외부 영향을 받기도 하지만 실험에서 외부로부터 발생되는 변수는 고려하지 않는다. 평균 통신 시간은 데이터 마이닝 응용 프로그램과 자원 중개 머신 간 통신 시간과 각 학습 머신의 자원 중개 머신과 학습 머신 간의 통신 시간을 평균하여 계산할 수 있다. 따라서 $CTam_i$가 응용 프로그램과 자원 중개 머신 간 통신 시간, $CTmr_i$이 자원 중개 머신과 $i$ 번째 학습 머신 간의 통신 시간, $n$이 학습 머신의 수를 나타내는 식 (1)에 의해 얻어질 수 있다.

$$Average\ Communication\ Time = \left\{\sum_{i=1}^{n}(CTam_i + CTmr_i)\right\}/n \quad (1)$$

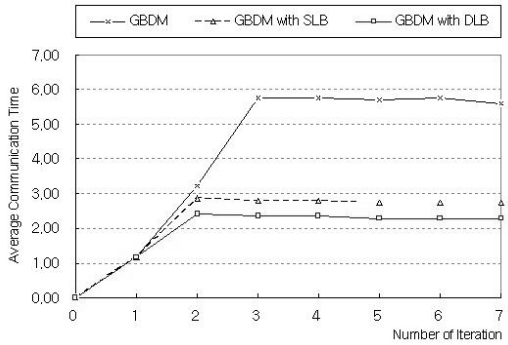


**그림 5.** 평균 통신 시간

그림 5에서 나타나듯이, 로드 밸런싱을 하지 않는 GBDM의 경우 통신 시간이 대략적으로 평균 5.7 시뮬레이션 시간으로 매우 높고 정적 로드 밸런싱을 하거나 동적 로드 밸런싱을 하는 GBDM with SLB나 GBCDM with DLB의 경우는 각각 평균 2.8, 2.5 시뮬레이션 시간을 기록하였다. GBDM with SLB와 GBDM with DLB는 연산 능력에 따라 정적, 혹은 동적으로 연결하는 점이 다르고 이것은 통신 시간에는 큰 영향을 미치지 않는 것을 알 수 있다.

### 4.2 평균 자원 활용도

그리드 환경에서 자원은 지리적으로 분산되어 있으며 거의 무한대에 이르지만 그 활용도는 천차만별이다. 자원 활용도는 해당 학습 머신이 일련의 할당된 작업을 처리하는 데 있어서 얼마만큼의 자원을 사용했는가를 나타내는 척도이다. 이 논문에서는 실제 그리드 환경 하에서 자원이 사용되는 것을 각각의 자원 제공자의 연산을 위해 자원을 사용한 시간을 이용하여 측정하였다. 또한, 일련의 작업이 학습 머신에서 처리되어 자원 중개 머신에 그 결과가 전달되는 것을 하나의 작업이 정상적으로 처리되는 것으로 간주하고 이 때 각 학습 머신이 자원을 활용한 시간으로 나타내어 식 (2)와 같이 평균 자원 활용도를 계산한다. 식에서 $CT_i$는 학습 머신과 자원 중개 머신과의 통신 시간을 나타내며, $ET_i$는 학습 머신이 자원을 사용한 시간을 나타낸다. 또한, n은 학습 머신의 수를 나타내며 평균 자원 활용도는 각각의 학습 머신의 자원 활용도를 평균하여 계산할 수 있다.

$$Average\ Resource\ Utilization = \left\{\sum_{i=1}^{n}\frac{ET_i}{CT_i + ET_i}\times 100(\%)\right\}/n \quad (2)$$

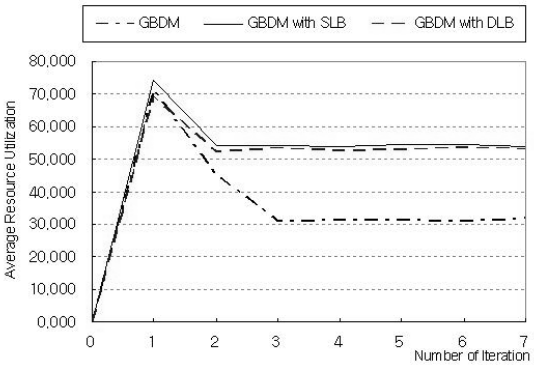


**그림 6.** 평균 자원 활용도

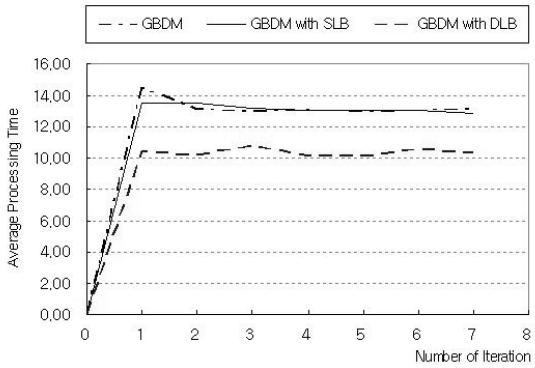


**그림 7.** 평균 처리 시간

그림 6에서 나타나듯이, 로드 밸런싱을 하지 않는 경우(GBDM) 통신 시간이 많이 필요하기 때문에 자원을 활용하는 시간이 평균적으로 다른 두 경우보다 낮다. 앞서 말했듯이 정적 로드 밸런싱과 동적 로드 밸런싱은 연산 능력에 따라 동적으로 연결하는 점에서 차이가 나고 통신 시간에는 큰 영향을 미치지 않기 때문에 GBDM with SLB와 GBDM with DLB는 평균 자원 활용도가 거의 비슷한 것을 알 수 있다.

### 4.3 평균 처리 시간

처리 시간은 오직 데이터 마이닝 작업을 위해 사용된 학습과 테스팅 시간을 의미한다. 이 실험에서 우리는 동적 로드 밸런싱의 유용성을 평가하기 위해 평균 처리 시간을 측정했다. 평균 처리 시간은 $Tl_i$가 학습 시간이고 $Tt_i$가 테스팅 시간, $n$은 학습 머신의 수를 나타내는 식 (3)에 의해 얻어질 수 있다. 그림 7에서 알 수 있듯이, GBDM, GBDM with SLB, GBDM with DLB 모두 동일한 결정 범위 조정 횟수에 신뢰할 만한 데이터 마이닝 인식률을 찾지만 동적 로드 밸런싱을 가진 경우(GBDM with DLB) 가장 낮은 평균 처리 시간을 보인다. GBDM은 데이터 마이닝 작업을 처리하기 위해 대략적으로 평균 10 시뮬레이션 시간을 기록하고 GBDM with SLB와 GBDM with DLB는 동일한 작업을 대략적으로 평균 13 시뮬레이션 시간을 기록하였다. GBDM은 데이터 마이닝 작업을 그리드 컴퓨팅을 통해 지리적으로 분산된 자원을 활용하지만 각각의 자원 제공자들은 서로 다른 연산 능력을 가지고 있기 때문에 할당된 작업에 따라 학습과 테스팅에 필요한 시간이 다르다. 따라서 다른 두 기법보다 더 많은 처리 시간이 필요하게 된다. 이러한 결과는 동적 로드 밸런싱을 통해 동일한 비용으로 더 많은 데이터의 처리가 가능하다는 것을 증명한다.

$$\textit{Average Processing Time} = \left\{\sum_{i=1}^{n}(Tl_i + Tt_i)\right\}/n \quad (3)$$

## 5. 결 론

이 논문에서는 동적 로드 밸런싱을 가진 그리드 기반 생물학 데이터 마이닝이 제안되었다. 생물학 데이터 마이닝은 대용량의 데이터 처리를 위해 고성능 컴퓨팅을 필요로 한다. 그리드 기술은 분산된 컴퓨팅 자원을 활용하여 대용량의 데이터 빠르고 효율적인 처리를 가능하게 한다. 우리는 생물학 데이터 마이닝과 그리드 기술을 결합하여 생물학 데이터 마이닝을 문제점을 해결한다. 게다가 생물학 데이터 마이닝에 결정 범위 조정 알고리즘을 적용하여 처리 시간과 비용 측면에서 장점을 얻는다. 또한, 우리는 대용량의 데이터 마이닝 작업의 부하를 분산시키는 동적 로드 밸런싱 알고리즘을 제안한다. 성능 측정을 위해 그리드 테스트베드 위에서 신경망 모델을 이용한 데이터 마이닝 실험이 이루어졌다. 동적 로드 밸런싱을 가진 그리드 기반 생물학 데이터 마이닝의 성능은 로드 밸런싱을 하지 않는 그리드 기반 생물학 데이터 마이닝, 정적 로드 밸런싱을 가진 그리드 기반 생물학 데이터 마이닝과 비교되었다. 실험에서 동적 로드 밸런싱을 가진 그리드 기반 생물학 데이터 마이닝 기법이 다른 두 기법보다 우수한 성능을 낸다는 것을 보여주었다. 향후 연구는 연산 능력 외에 더 많은 요소를 고려하여 우선 순위를 매기고 우선 순위에 기반한 더 효과적인 기법들을 연구할 계획이다.

## 참 고 문 헌


1. Berman, F., G. Fox and T. Hey, Grid Computing: Making the Global Infrastructure a Reality, J. Wiley, New Jersey, 2003.
2. Department of Defense, High Level Architecture Run-Time Infrastructure Programmer's Guide 1.3 Version 5, 1998.
3. Foster, I. and C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, San Francisco, 1998.
4. Kim, C. and H. Kameda, “An Algorithm for Optimal Load Balancing in Distributed Computer Systems”, IEEE Transactions on Computers, Vol. 41, No. 3, pp. 381-384, 1990.
5. Kumar V., A. Grama, A. Gupta and G. Karypis, Introduction to Parallel Computing: Design and Analysis of Algorithms, The Benjamin/Cummings Publishing Company, San Francisco, 1994.
6. Kumar, V., A. Grama and V. N. Rao, “Scalable Load Balancing Techniques for Parallel Computers”, Journal of Distributed Computing, Vol. 7, 1994.
7. Rajagopalan, A., S. Hariri, “An agent based dynamic load balancing system”, Proceedings of the International Workshop on Autonomous Decentralized Systems, pp. 164-171, 2000.
8. Sureswaran, R., M. Samaka and J. Knaggs, “LOADIST: a distributed processing environment based on load sharing”, IEEE SICON/ICIE '95 International Conference on Networks and Information Engineering, pp. 518-522, 1995.
9. Tzanis, G., C. Berberidis and I. Vlahavas, “Data Mining in Biological Data”, Encyclopedia of Database Technologies and Applications, IDEA Group Publishing, 2005.
10. Wang, J. T. L., M. J. Zaki, H. T. T. Toivonen and D. Shasha, Data Mining in Bioinformatics, Springer, Berlin, 2005.
11. Watson, S. and S. Arkinstall, The G-protein Linked Receptor Facts Book. Academic Press, Burlington, 1994.
12. Wikinson and Allen, Parallel Programming Techniques & Applications using Networked Workstations and Parallel Computers, Prentice Hall, New Jersey, 1999.
13. Zaki, M. J., W. Li and S. Parthasarathy, “Customized Dynamic Load Balancing for a Network of Workstations”, 5th IEEE International Symposium on High Performance Distributed Computing (HPDC '96), 1996.



**마 용 범** (myb112@hanmail.net)

2005 인하대학교 컴퓨터공학부 학사
2007 인하대학교 컴퓨터정보공학과 석사
2007～현재 인하대학교 정보공학과 박사과정

관심분야 : 그리드 컴퓨팅, 모델링 및 시뮬레이션, 분산 컴퓨팅



**김 태 영** (silverwild@gmail.com)

2007 인하대학교 컴퓨터공학부 학사
2009 인하대학교 정보공학과 석사
2009～현재 인하대학교 정보공학과 박사과정

관심분야 : 시스템 모델링 & 시뮬레이션, 분산처리

**이 종 식** (jslee@inha.ac.kr)

1993 인하대학교 전자공학과 학사
1995 인하대학교 전자공학과 석사
2001 애리조나대 전기・컴퓨터공학과 박사
2001～2002 캘리포니아 주립대학교 전기・컴퓨터공학과 전임강사
2002～2003 클리블랜드 주립대학교 전기・컴퓨터공학과 조교수
2003～2006 인하대학교 컴퓨터공합부 조교수
2006～현재 인하대학교 컴퓨터공학부 부교수

관심분야 : 시스템 모델링&시뮬레이션, 분산 컴퓨팅