분산 환경에서 신경망을 응용한 데이터 서버 마이닝

박민기* · 김귀태* · 이재완*

*군산대학교

Data Server Mining applied Neural Networks in Distributed Environment

Min-Gi Park* · Gui-Tae Kim** · Jae-Wan Lee***

*Kunsan National University

E-mail : sopiru@kunsan.ac.kr

요 약

오늘날 인터넷은 하나의 거대한 분산 정보 서비스센터의 역할을 수행하며 여러 가지 많은 정보들과 이를 관리 운영하는 데이터 베이스 서버들은 분산된 네트워크 환경 속에서 광범위하게 존재하고 있다. 그러나 우리는 데이터 특성에 따라 입력 데이터를 처리할 서버를 결정하는데 여러 가지 어려움을 겪고 있다.

본 논문에서는 분산 환경 속에 존재하는 수많은 데이터들 가운데 신경망을 이용해 입력 데이터 패턴을 가장 효율적으로 처리할 수 있는 목적지 서버를 마이닝하는 기법과 이를 기반으로 한 지능적 데이터 마이닝 시스템 구조를 설계하였다.

그 결과로서 새로운 입력 데이터패턴이 신경망으로 구현된 동적 바인딩 방법에 따라 목적지 서버를 결정한 후 처리됨을 보였다. 이 기법은 데이터 웨어하우스, 통신 및 전력부하패턴 분석, 인구센서스 분석, 의료데이터 분석에 활용될 수 있다.

ABSTRACT

Nowaday, Internet is doing the role of a large distributed information service center and various information and database servers managing it are in distributed network environment. However, the we have several difficulties in deciding the server to disposal input data depending on data properties.

In this paper, we designed server mining mechanism and Intellectual data mining system architecture for the best efficiently dealing with input data pattern by using neural network among the various data in distributed environment.

As a result, the new input data pattern could be operated after deciding the destination server according to dynamic binding method implemented by neural network. This mechanism can be applied Datawarehous, telecommunication and load pattern analysis, population census analysis and medical data analysis.

키워드

Datamining, Neural network, Distributed system, Data server

Ⅰ. 서 론

최근 대부분의 컴퓨팅 환경은 중앙집중적인 클라인언트와 서버간 관계 시스템에서, 미들웨어의 역할이 차지하는 부분이 큰 분산처리 환경으로 급속하게 전환되었다. 네트워크 환경이 발달하면서 컴퓨팅 자원, 정보, 각종 서비스들이 모두 인터넷을 중심으로 연결되어 있고 이런 분산된 자원들은 각각의 분리된 영역의 시스템에서 관리되어 운영되고 있다.

이런 가운데 새로운 과제로 떠오른 문제는 분산된 자원들 속에서 유도된 새로운 데이터 모델을 발견하여 미래에 실행 가능한 정보를 추출해 내고 이 정보를 분산환경에서 효율적으로 처리할 수 있는 서버를 결정하는 과정이다. 이 문제를 해결하기 위한 기존의 데이터 마이닝 알고리즘들 가운데에서도 신경망 기술은 입력된 정보를 지능적으로 분석하여 처리할 서버를 마이닝하는 데 유용한 기법이다.

많은 입력 데이터들을 처리할 데이터 서버를

마이닝하는데 중요시 다루어야할 것은 서버를 인식하는데 걸리는 지연시간이다. 지연시간을 개선하기 위해 한번 결정된 데이터 패턴에 대한 서버 결정을 유지하여 중복된 데이터 패턴을 처리하는데 걸리는 지연시간을 감소시켜야 한다. 이를 위해 본 논문에서는 참조리스트를 이용한 정적 바인딩과 신경망에 의한 동적 바인딩을 통해 입력된 데이터 패턴을 처리하는데 지연시간을 최소화하였다.

II. 관련 연구

1. 분산환경에서 데이터 마이닝

분산된 네트워크 환경에서 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 데이터 마이닝이라하며 이 알고리즘을 적용하기 앞서 데이터 전 처리 과정을 통해 해결해야 할 일들로는 접근 데이터에 대한 모델 생성, 관련 없는 데이터나 잡음을 삭제하는 데이터 정제와 여과, 개별적인 데이터를 트랜잭션 단위로 묶는 그룹화, 이용자 등록정보와 같은 다른 데이터와의 통합 등이 있다.[1][2]

데이터의 전처리 과정이 끝나고 나면 분석자의 필요에 따라 경로 분석(path analysis), 연관규칙(association rule)과 순차 패턴(sequential pattern)발견, 군집화(clustering)와 분류(classification)와 같은 다양한 접근 패턴 마이닝 작업이 수행되고 패턴 발견 도구를 이용하여 일단 분산된 환경 속에서 입력 데이터패턴들을 발견한다.[3][4][5]

그림 1의 마지막 단계에서처럼 이 패턴들을 이해하고 해석한 새로운 입력 패턴을 분산된 서버들 가운데 처리할 서버를 마이닝하여 효율적으로 처리해야 한다.

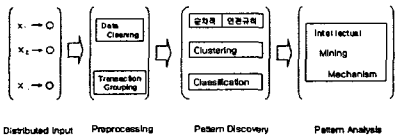

그림 1. 데이터 마이닝 프로세스

2. 데이터 마이닝을 위한 신경망

신경망 모형은 신경생리학 분야에서 사람의 두뇌 활동을 이해하고자 하는 목적 하에 신경의 작업을 설명하려는 시도에서 출발하여 생물학적인 프로세스를 컴퓨터를 이용하여 모형화 하려는 노력에서 비롯된 것이다. 신경망 모델들은 시각적인 패턴인식, 광학, 문자인식, 음성분석, 로봇학 그리고 데이터 마이닝에 이용되고 있고 신경망은 위상, 노드의 특성, 학습 규칙 등으로 특징 지워지며 훈련 방법에 따라 각각 지도학습(supervised learning)모델과 자율학습(unsupervised learning)모델로 나누어진다.[6]

지도학습 모델중의 하나인 홉필드 네트워크나 퍼셉트론과 같은 신경망 모델들은 연상 기억장치(Associative memory)나 분류기(classifier)로 사용된다.[7]

감독 없이 훈련되는 코호넨(T.Kohonen)의 형상지도(feature-map)는 훈련기간 중 정확한 클래스(class)에 대한 정보 없이 자기조직화로 클러스터링을 한다.[8]

분산 환경에서 발견한 입력 데이터패턴들을 처리할 서버를 마이닝 하기 위해서는 자율 신경망 모델 중의 경쟁학습 알고리즘을 이용한다. 이 신경망 모델은 자기조직화(self-organize)하는 것을 비교적 단순하면서도 효율적으로 보여준다.

본 논문에서는 경쟁학습 알고리즘을 통해 훈련모델을 간결하고 이해하기 쉽게 하였으며 자기 조직화를 통해 입력 데이터패턴을 처리할 서버를 결정하는 지능적 데이터 마이닝 시스템을 설계하였다.

III. 지능적 데이터 마이닝 시스템

1. 지능적 데이터 마이닝(IDM) 시스템 구조

지능적 데이터 마이닝(Intellectual Data Mining)을 위한 시스템 구조는 그림 2와 같으며 클라이언트로부터 전달된 입력 데이터패턴은 각각의 입력 노드인 패턴핸들러에 전달되고 모니터는 입력 데이터 패턴과 참조 리스트를 비교하여 두 가지 바인딩 방법중 한가지를 선택하고 그 방법에 따라 목적지 서버를 결정하여 목적지 서버의 참여자에게 입력된 데이터 패턴을 전달하여 처리한다.

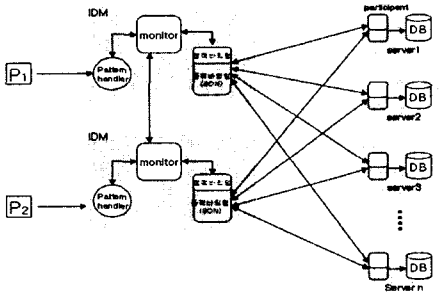

그림 2. IDM 시스템 구조

2. IDM의 패턴핸들러(Pattern Handler)

패턴핸들러는 그림 3에서 처럼 크게 두 부분으로 나뉘는데 수신부와 전송부로 구성된다. 수신부에서는 발견한 패턴을 수신큐를 통해 순서대로 수신하며 수신확인 신호를 모니터에 전달한다. 전송부는 송신기(Sender)와 제어기

(Controller)로 나뉘며 제어기는 모니터의 전송 허락 신호를 기다리다가 전송허락 신호가 들어오면 송신기에게 수신부에 입력된 패턴을 모니터에 전송하도록 한다.

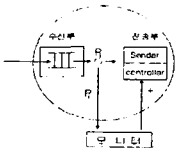

그림 3. 패턴 핸들러

3. IDM의 Monitor

모니터는 입력패턴을 분산환경의 서버들에게 전달하기 위한 중재자역할을 담당하며 그림 4와 같은 구조로 이루어져 있다. 패턴 핸들러에 입력 데이터 패턴이 전달되면 모니터객체는 래퍼에 보관되어 있는 참조리스트와 비교를 하여 정적바인딩을 할 것인지 아니면 동적 바인딩을 할 것인지 결정한다.

3.1 목적지 서버 마이닝 방법

입력 데이터패턴을 처리할 서버를 마이닝하는 방법에는 두 가지가 있으며 첫 번째 방법은 정적 바인딩으로써 정적 바인딩을 할 경우는 새로 입력된 데이터 패턴에 대한 참조리스트가 존재하면 이를 참조해 기록된 서버를 참조하여 처리할 서버를 선택한다. 그리고 두 번째 방법은 동적바인딩으로써 입력 데이터 패턴이 없을 경우 신경망을 이용해 처리할 서버를 선택하게 된다. 동적바인딩 하는 순서는 먼저 입력 데이터패턴을 신경망인 SON에 보내고 SON에 의해 서버를 선택한 후 선택된 서버의 참여자에게 입력데이터 패턴을 전달한다. 모니터 안에 존재하는 래퍼는 참여자에 대한 정보를 참조 리스트에 기록하고 참여자는 모니터의 참조리스트를 기록한다.

3.2 모니터간 참조 방법

입력 데이터 패턴을 처리할 서버를 찾지 못하는 경우가 발생할 수 있다. 이를 해결하기 위해 모니터간의 참조를 통하여 근접한 모니터에서 처리하도록 하였다. 모니터간의 참조는 모니터 내에 존재하는 대리자를 통해 이루어지며 대리자가 처리할 요청을 받아 처리할 서버를 선택하여 입력 데이터 패턴을 처리시킨다. 처리가 완료되면 모니터와 참여자간의 참조 리스트를 작성하고 모니터 각자의 대리자에 대한 참조리스트를 모니터의 대리자가 기록하여 복구할 수 있는 모든 입력패턴의 경로를 보관한다. 래퍼런스를 기록하여 보관하고 있다는 것은 중복된 입력데이터 패턴의 경우 기억된 경로를 따라 처리할 서버를 찾는데 시간을 절약할 수 있다.

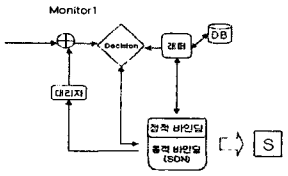

그림 4. 모니터의 구성도

4. 자기조직화 신경망(SON)

자기조직화 신경망 구조는 그림 5와 같으며 초기자는 연결강도를 임의의 수로 초기화한다. 계산처리기에서 입력 데이터 패턴과 연결강도를 이용해 입력 데이터 패턴을 처리할 서버 패턴의 거리를 계산하고 출력선택자는 계산된 서버 패턴의 거리들 가운데 최소 거리에 있는 것을 최종 서버로 선택한다. 서버 패턴과 그 이웃들의 연결강도를 재조정하기 위해 연결강도 조정자에 선택된 서버패턴을 전달한다. 마지막으로 조정된 연결강도값을 초기자에게 전달하여 연결강도를 조정하며 새로운 입력 데이터패턴이 들어왔을 경우 위 과정을 반복한다. 각각의 구체적인 구성요소는 다음과 같다.

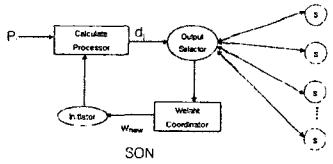

그림 5. SON의 구성도

Initiator : 연결강도 초기자는 연결강도를 작은 값의 임의의 수로 초기화한다. 그리고 계산처리기에 초기화한 연결강도를 계산 처리기에 전달한다.

Calculator Processor : 계산 처리기는 입력된 데이터 패턴과 연결강도를 이용하여 모든 패턴간의 거리를 구하는 식에 의해 계산하고 그 결과를 출력 선택자에 전송한다.

Output Selector : 출력 선택자는 최소 거리에 있는 서버 패턴을 선택하여 입력 데이터 패턴을 선택한 서버로 전송하며 이 때 선택된 서버와 이웃들간의 연결강도를 재조정하기 위해 연결강도 중재자에 전송한다.

Weight Coordinator : 연결강도 중재자는 선택된 서버 패턴과 이웃들의 연결강도를 구하는 다음 식에 의해 재조정하고 새로운 연결강도 (W_{new})를 연결강도 초기자에 전송하여 연결 강도를 재 설정한다.

5. 지능적 데이터 마이닝 알고리즘

지능적 데이터 마이닝 알고리즘은 아래 표 1과 같으며 모니터는 패턴핸들러로부터 입력 데

이터 패턴(P_i)을 전달받아 래퍼의 참조리스트와 비교하여 $P_i \neq P_{i+1}$이라면 모니터는 동적 바인딩 방법을 선택하여 입력 데이터 패턴을 처리할 서버를 선택하고 입력 데이터 패턴을 보낸다. 만약 비교한 결과가 $P_i = P_{i+1}$이라면 모니터는 래퍼에 기록되어 있는 참조 리스트를 이용해 정적 바인딩 방법으로 입력 데이터 패턴을 처리할 서버를 선택하고 입력 데이터 패턴을 보내어 처리한다.

표 1. 지능적 데이터 마이닝 알고리즘

```
Procedure Intellectual Data Mining (result)          /* data pattern : Pi */
    for all i∈n /* n : integer */
        H ← Pi                                       /* receive Pi into Pattern Handler(H) */
    for i ←1 to n
        if Pi ≠ null then
            M ← Pi
        while(M← Pi)   /* Send Pi into Monitor(M) */
        if Pi ≠ Pi+1 then
        select(SOM)                                  /* choose method */
                                    initialize Wij
                                    if CP ← Pi then /* receive Pi into Calculate Processor */
                                        Xi ← Pi
                                        while Xi ≠ 0 do
                                                    N-1
                                            dj = Σ (Xi(t) - Wij(t))^2
                                                    i=0
                                    if OS ← dj then /* select server in Output Selector */
                                        if dj < dj-1 then select(server)
                                                    send(Pi)
                                                    record(reference list)
                                    if WC ← result then /* send result in Weight Coordinator */
                                        Wij(t+1) = Wij(t) + α(Xi(t) - Wij(t))
                                        Wnew ← Wij(t+1)
                                    if Wij ← Wnew then /* coordinate new weight */
                                        coordinate(Wnew)
                                    else
                                        /* Pi = Pi+1 */
                        select(static)      /* choose method */
                        for i←1 to n
                          while Pi ≠ result ← reference list do
                                    select(server)
                                    send(Pi)
                                update(reference list)
                        end if
End Intellectual Data Mining
```

IV. 구현 및 결과

본 논문에서 제시한 지능적 데이터 마이닝을 위한 구현환경은 다음과 같다.

입력 데이터패턴을 보내는 두 개의 지능적 데이터마이닝 시스템과 입력 데이터패턴을 처리할 100개의 서버로 정하였다. 학습율은 0.5에서 시작하여 0.01에서 끝나고 훈련은 훈련집합의 5000회의 반복으로 수행되며 서버선택은 매 500회 실행마다 이루어지며 그림 6에서와 같이 근접한 서버 개수를 경쟁학습에 의해 줄여가며 최종 서버개수가 1이 되었을 때 프로그램은 종료된다. 그리고 그 결과로 그림 7과 같이 출력 서버 파일(outserver.dta)을 작성하며 동시에 참조 리스트(list.dta)를 작성하여 저장한다.

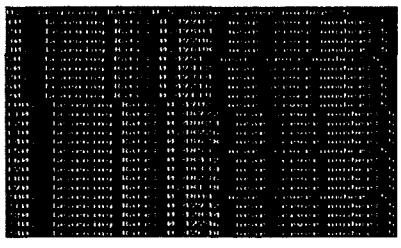

그림 6. 서버결정을 위한 훈련과정

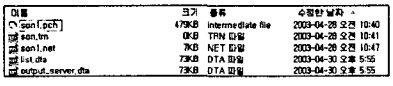

그림 7. 생성된 출력 서버파일

V. 결 론

지능적 데이터 마이닝 시스템은 데이터 패턴의 중복성을 고려하여 새로운 입력 데이터 패턴일 경우 동적 바인딩 방법으로 처리되었고 기존의 입력 데이터 패턴일 경우 참조 리스트와 비교하여 정적바인딩 방법을 이용해 데이터 서버를 마이닝하였다. 동적바인딩 방법은 신경망의 자기형상화에 기반을 두어 데이터 패턴을 처리할 서버를 결정짓기 위해 사용되었다.

결과로서, 새로이 데이터 패턴이 입력될 경우 신경망에 의해 구현된 동적바인딩 방법으로 서버를 선택하여 처리되는 결과를 보였다. 본 논문에서 제시한 지능적 데이터 마이닝 시스템은 데이터 웨어 하우스 구축, 전력부하균등, 인구센서스 분석, 의료데이터 분석 등에 활용될 수 있다.

향후 연구과제로는 지능적 데이터마이닝 시스템의 구체적인 구축방안을 제시할 수 있는 모습으로 발전시켜 나갈 것이며 더욱 방대한 정보에 대한 입력 데이터의 패턴을 처리할 능력을 제공하도록 확대해 나갈 것이다.

참고문헌

[1] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 21 - 30, 2001

[2] P. Pirolli, J. Pitkow, and R. Rao, Silk from a Sow's Ear: Extracting Usable Structures from the Web, Proceedings of 1996 Conference on Human Factors in Computing Systems(CHI-96), 1996.

[3] Mannila, H., Toivonen, H. On an algorithm for finding all interesting sentences, Proceedings EMCSR '96, 1996

[4] Apte, C., and Hong, S. J, In Advances in Knowledte Discovery and Data Mining, eds., 514-560, AAAI Press, 1996

[5] Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J, Classification and Regression Trees, AAAI Press, 1984

[6] 김대수, 신경망 이론과 응용, 하이테크정보, 169-189, 1999

[7] J. J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, Proc. of the National Academy Science 81, 3088-3092, 1984

[8] T.K Kohonen, Self-Organization and Associative Memory, Springer-Verlag, 1984.