제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

GHSOM을 이용한 대용량 데이터 처리의 군집화 방법

김만선[1], 이상용[2]
공주대학교
e-mail : {mansun[1], sylee[2]}@kongju.ac.kr

A Clustering Method using GHSOM for Processing Large Data

Man-sun Kim[1], Sang-yong Lee[2]

Dept. of Computer Engineering, Kongju National University[1],
Division of Information and Communication Engineering, Kongju National University[2]

요 약

최근 대용량의 데이터베이스로부터 유용한 정보를 발견하고 데이터간에 존재하는 연관성을 탐색하고 분석하는 데이터 마이닝에 관한 많은 연구들이 진행되고 있다. 실제 응용분야에선 수집된 데이터는 시간이 지날수록 데이터의 양이 늘어나게 되고, 중복되는 속성과 잡음을 갖게 되어 마이닝 기법을 이용하는데 많은 시간과 비용이 소요된다. 또한 어느 속성이 중요한지 알 수 없어 중요한 속성이 중요하지 않은 속성에 의해 왜곡되거나 제대로 분석되지 않을 수 있다.

본 연구는 이러한 문제점들을 해결하기 위해 GHSOM을 이용한 계층적 신경망 군집화 방법을 제안한다. 제안하는 방법은 미리 군집의 개수를 정해줄 필요가 없고, 다양한 레벨의 군집들을 얻을 수 있는 계층적 군집화를 이루어낸다는 장점을 갖는다. 본 논문에서는 신경망 GHSOM의 구조와 특성에 대해 간략히 살펴보고 시스템 처리과정에 대해 설명한다.

1. 서 론

지식탐사 프로세스의 핵심적인 역할을 담당하는 데이터 마이닝 단계에서는 여러 가지 목적에 따라 알고리즘을 선택하여 사용한다. 데이터 마이닝에서 클러스터링 방법은 기존의 통계, 기계학습, 패턴인식에서 쓰이던 방법에 부가적으로 데이터베이스 지향적인 사항들을 첨가시킨 것으로써, 다양한 다차원 데이터를 효율적으로 분류해 나가기 위한 방안으로 연구되고 있다.

클러스터링은 입력 데이터집합을 유사한 관찰값들의 군집들로 구분하여 데이터집합 속에 존재하는 의미있는 정보를 얻는 과정이다[1][2]. 최근에는 대용량의 데이터베이스로부터 유용한 정보를 발견하고 데이터 간에 존재하는 연관성을 탐색하고 분석하는 데이터 마이닝에 관한 많은 연구들이 진행되고 있다. 또한 마이닝 과정의 속도를 향상시키고 효율을 높이기 위해 중요한 속성을 선택(feature selection)하고 가중치(weight)를 조절하는 연구가 진행되고 있다[3].

실제 응용분야에선 수집된 데이터는 시간이 지날수록 데이터의 양이 늘어나게 되고, 중복되는 속성과 잡음을 갖게 되어 마이닝의 기법을 이용하는데 많은 시간과 비용이 소요된다. 또한 어느 속성이 중요한지 알 수 없어 중요한 속성이 그렇지 않은 속성에 의해 왜곡되거나 제대로 분석되지 않을 수 있다.

이 논문은 이러한 문제점들을 해결하기 위해 GHSOM을 사용한 계층적 신경망 군집화 기법을 제안한다. 미리 군집의 개수를 정해줄 필요가 없고, 다양한 레벨의 군집들을 얻을 수 있다.

1) 공주대학교 컴퓨터공학과 박사과정
2) 공주대학교 정보통신공학부 부교수

2. 관련 연구

클러스터링 알고리즘은 일반적으로 통계적 클러스터링 방법과 인공지능적 클러스터링 방법의 두 가지로 분류할 수 있다.

2. 1 통계적 클러스터링 방법

통계적 클러스터링 방법에는 분할적 클러스터링과 계층적 클러스터링이 사용된다.

1) 분할적 클러스터링

분할적 클러스터링(partitioning clustering)은 계층적 클러스터링과 달리 중첩된 분할 계층구조가 아닌 평평한 하나의 분할 구조로 형성된 군집을 생성한다.

여러 가지 분할적 클러스터링 알고리즘 중에서 k-means는 유클리디안 거리(Euclidean distance)[4]를 이용하여 가깝게 위치한 점들을 찾아 군집으로 묶어주는 기법으로 차원의 제약이 전혀 없고 간단하다는 장점 때문에 널리 사용되는 방법이다.

k-means는 임의의 초기 분할로부터 시작하여 군집의 중심값과 데이터 개체들과의 유사도에 근거하여 목적함수가 수렴 조건을 만족할 때까지 데이터의 소속 군집을 재할당한다. 이렇게 발견된 군집은 중심값(centroid)으로 표현되는데, 이는 해당군집에 속하는 데이터들의 평균 혹은 중앙값이다.

분할적 클러스터링은 군집이 벡터 평면상에서 구(sphere)의 형태를 가지고 있어야 효율이 좋다고 알려져 있다. 그러나 군집의 결과가 항상 구의 형태를 나타낸다고 볼 수 없기 때문에 본 논문에서는 계층적 클러스터링을 사용하였다.

2) 계층적 클러스터링

계층적 클러스터링(hierarchical clustering)은 가장 유사한 두 개체들을 선택하여 병합해 가는 방법과 가장 먼 개체들을 선택하여 나누어 나가는 방법이 있다. 두 군집의 유사도를 측정하는 기준에 따라 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법 등으로 나뉜다.

계층적 클러스터링은 상향식 방법(bottom-up)인 병합적 계층군집방법과 하향식 방법(top-down)방법인 분할적 계층적 군집방법이 있다. 각각의 모든 데이터 개체가 하나의 분할을 이루는 최하위 계층에서부터 모두 하나의 군집으로 합쳐진 최상위 계층까지 중첩된 분할의 계층순서를 생성한다.

2. 2 인공지능적 방법

신경망 SOM(Self-Orgnizing Map)은 분할 군집화 방법의 하나로서, 각 문서 그룹에 대응되는 K개의 뉴런들로 구성된 1차원 혹은 2차원 단일 계층의 신경망이다. SOM에서는 하나의 문서모델이 입력벡터로 주어지면 각 뉴런의 가중치벡터들과 비교하여 입력과 가장 유사한 가중치를 갖는 뉴런에 입력문서가 배정되고, 이 뉴런과 이웃한 뉴런들의 가중치를 입력에 가깝게 조정하는 학습과정이 되풀이된다. 이와 같은 학습과정을 거쳐 SOM은 입력되는 문서들에 대해 뉴런별로 하나의 군집을 형성하게 되고, 또한 유사한 문서 군집들의 뉴런은 거리적으로도 가까이 배치되어 군집화의 결과를 시각적으로 이해하기 좋은 장점을 갖는다. 그러나 SOM을 이용하기 위해서는 미리 적당한 문서 군집수를 예측하고, 그 수만큼의 뉴런들로 신경망을 구성해야하는데, 문서들의 내용을 모두 확인하지 않은 상태에서는 이러한 문서 군집수를 정하는 일은 적절하지 못하다.

성장하는 SOM(Growing SOM)은 이러한 SOM의 문제점을 극복하기 위해, 입력되는 문서들의 양과 이질성에 따라 뉴런의 수를 늘여감으로써 스스로 군집의 개수를 정해가는 특징을 가지고 있다. 한편 계층적 SOM은 단순 SOM이나 계층적 SOM의 경우와는 달리 주어진 문서들을 단순히 K개의 군집으로 분할하지 않고, 유사성과 크기가 다른 연속된 여러 계층의 군집들을 생성해내는 특징이 있다.

3. GHSOM을 이용한 군집화 방법

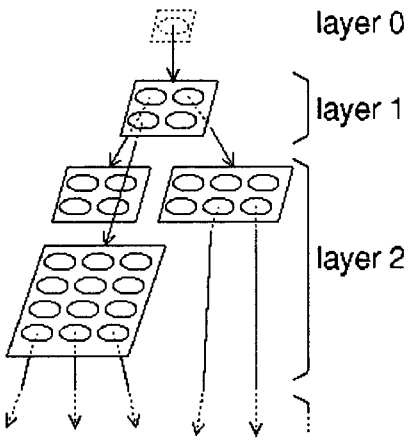
3. 1 GHSOM

GHSOM(Growing Hierarchical SOM)[5][6]은 성장 SOM과 계층적 SOM의 장점을 결합하여 만들어진 군집화 알고리즘이다.

GHSOM의 기본구조는 [그림 1]과 같이 여러 계층의 서로 독립적인 SOM들로 구성된 계층적 구조로 되어있고, 기존 SOM과는 달리 입력문서들에 따라 맵(map)의 크기와 계층 수가 스스로 늘어나는 성질을 가지고 있다.

GHSOM에서 계층은 map 계층 0, 계층 1 그리고 나머지 부분으로 구분할 수 있다. 우선 계층 0은 가상의 계층으로 1개의 유닛(unit)을 포함하고 있으며, 이 유닛의 가중치벡터는 $m_0=[\mu_{01}, \mu_{02}, .., \mu_{0n}]^T$와 같이 표현되며, 모든 입력 데이터의 평균으로 초기화된다. 입력 데이터 x와 이 유닛과 편차는 식 1과 같이 나타낼

수 있다.



[그림 1] GHSOM의 구조

$$mqe_0 = \frac{1}{d} \cdot \| m_0 - x \|, \ d: x\text{의 수} \quad (식\ 1)$$

msq_0를 계산한 후에 GHSOM의 첫 번째 SOM으로부터 훈련이 시작된다. 첫 번째 계층의 맵은 유닛의 수보다 적은 수의 유닛으로 초기화된다. 각 유닛 i는 식과 같이 n-차원의 벡터 m_i로 정의된다.

$$m_i = [\mu_{01}, \mu_{02}, .., \mu_{0n}]^T, \ m_i = R^n \quad (식\ 2)$$

각 유닛들은 랜덤한 값으로 초기화되며, SOM의 학습규칙은 식 3과 같다. 여기서 α는 학습률이고, h_{ci}는 이웃함수이고, x는 현재의 입력 패턴이다. 그리고 c는 t 만큼 반복한 후의 승자유닛이 된다.

$$mi(t+1) = mi(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - mi(t)] \quad (식\ 3)$$

일정한 회수만큼의 훈련이 이루어진 후에 식 4에 의해 맵의 MQE(Mean Quantization Error)[7]가 계산된다. 여기서 u는 SOM m에 포함된 유닛 i의 개수이고, mqe_i에 의해서 계산된다.

$$MQE_m = \frac{1}{u} \cdot \sum_i mqe_i \quad (식\ 4)$$

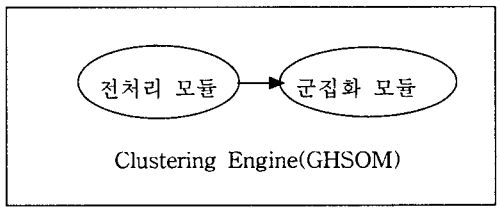
그리고 식 5의 조건이 만족하는 동안 mqe_e가 가장 큰 유닛 e에 새로운 열이나 행을 삽입함으로써 맵은 계속 성장한다.

$$MQE_m \geq \tau_m \cdot mqe_0 \quad (식\ 5)$$

$MQE_m \geq \tau_m \cdot mqe_0$이 되어 한 계층의 성장이 종료되면, 이 맵은 다음 계층으로 확장을 시도한다. 이때 매우 높은 mqe를 가진 이 유닛들은 다음 계층의 새로운 맵에 추가된다. 그리고 각 유닛 i는 식 6과 같은 조건을 만족하면 확장하게 된다.

$$MQE_m > \tau_m \cdot mqe_0 \quad (식\ 6)$$

3. 2 시스템 처리과정



[그림 2] 시스템의 구조

내부적으로 군집화에 필요한 전처리 모듈과 실제 군집화를 하는 군집화 모듈로 구성되며 [그림 2]와 같은 시스템 구조로 보여질 수 있다.

전처리 모듈에서는 초기 군집단계를 형성하는 단계로 각 data에 사용된 단어를 기초로 TFIDF 알고리즘을 이용한 벡터를 파일로 생성한다.

군집화 모듈에서는 GHSOM의 핵심모듈로 각 군집간의 유사도를 측정하여 계층적 군집화를 수행하여 사용자에게 보여준다.

4. 결론 및 향후 과제

본 논문에서는 신경망 GHSOM의 구조와 특성에 대해 간략히 살펴보았다. 제안한 시스템은 미리 군집의 개수를 정해줄 필요한 없는 시스템으로 기존의 단순한 계층적 군집화 방법의 문제점을 극복할 수 있을 것으로 기대된다.

향후 연구 과제로는 GHSOM의 시스템을 좀더 확장하여 보다 높은 정확도를 얻을 수 있는 시스템 구현과 end-user에게 편리한 인터페이스를 제공할 수 있는 연구가 필요하다.

참 고 문 헌

[1]Tian Zhang, Raghu Ramakrishnan, and Miron, "Birch : an efficient data clustering method for very large database", the ACM SIGMOD Conference on Management of Data, Montreal,

Canada,June 1996
[2]Tian Zhang, Raghu Ramakrishnan, and Riron, "BIRCH: A New Data Clustering Algorithm and Its Application". Data Mining and Knowledge Discovery, 1,141-182, 1997
[3]Fayyad, Piatetsky-Shapiro, Smyth, "Advances in knowledge discovery and data mining", 1996.
[4]http://www.palantir.swarthmore.edu/~maxwell/loicz/workshop.3-00/distances.htm
[5]Micheal Dittenbash, Dieter Merkl and Andeas Rauber, "The Growing Hierarchical Self Organizing Map", Proc of IJCNN2000, pp.15-19, 2000
[6] Micheal Dittenbash, Dieter Merkl and Andeas Rauber, "Regent Advance with the Growing Hierarchical Self Organizing Map", Springer Verlag, 2001
[7]http://www.ifs.tuwien.ac.at/~mbach/ghsom/introduction.html