데이터 마이닝을 위한 신경망 클러스터링 기법에 관한 연구

김만선†, 이상용††
공주대학교
e-mail : {mansun†, sylee††}@kongju.ac.kr

Hybrid Neural Network Clustering Using SOM and BP for DataMing

Man-sun Kim†, Sang-yong Lee††

Dept. of Computer Engineering, Kongju National University†,
Division of Information and Communication Engineering, Kongju National University††

요약

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.

이 논문은 이러한 문제점들을 해결하기 위해, 대용량의 데이터에 적용할 수 있고 데이터에서 알려지지 않은 패턴을 발결할뿐만 아니라, 사용자가 얻고자 하는 출력을 생성할 수 있는 혼합형 신경망 클러스터링 기법을 제안 한다. 그리고 알고리즘의 타당성을 검증하기 위해 몇 가지 벤치마크데이터를 이용하여 본 논문의 타탕성을 보인다.

1. 서론

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 것으로서, 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 알수 없어 중요한 속성이 그렇지 않은 속성에 의해 왜곡되거나 제대로 분석되지 않을 수 있다.

이 논문은 이러한 문제점들을 해결하기 위해, 대용량의 데이터에 적용할 수 있고 데이터에서 알려지지 않은 패턴을 발견할뿐만 아니라, 사용자가 얻고자 하는 출력을 생성할 수 있는 혼합형 인공신경망 클러스터링 기법을 제안한다.

2. 관련 연구

클러스터링 알고리즘은 크게 계층적 알고리즘, 분할적 알고리즘, 신경망 모델로 나눌 수 있다.

계층적 알고리즘은 군집들이 단지 데이터에 의해서만 결정되고 군집의 수가 계층을 내려가거나 올라가는 것에 의해 증가할 수도 줄어들 수도 있다는 점에서 특징이 있다[4].

분할적 알고리즘은 주어진 목적함수를 최적화 하도록 데이터 집합을 k개의 군집으로 나누는 것이다. 임의의 초기 분할로부터 시작하여 데이터 개체에 대한 소속 군집의 재할당 과정과 목적 함수의 평가를 반복적으로 수행하여 목적함수를 최적화하지만 초기 분할에 민감하다는 특징이 있다[5].

신경망 모델을 이용한 클러스터링 방법에는 kohonen네트워크, Carpente와 Grossberg네트워크가 있다. kohonen네트워크에는 자기조직화지도(SOM)와 LVQ알고리즘이 있으며, 이 방법은 패턴과 클러스터의 중심값과의 거리를 최소화시키는 학습 알고리즘에 따라 클러스터링한다는 점에서 k-means알고리즘과 대응된다고 할 수 있다.

최근 발표된 새로운 클러스터링 알고리즘에는 BIRCH, CURE처럼 전처러과정을 거쳐 세부적인 클러스터링을 수행하는 알고리즘이 많이 있다.

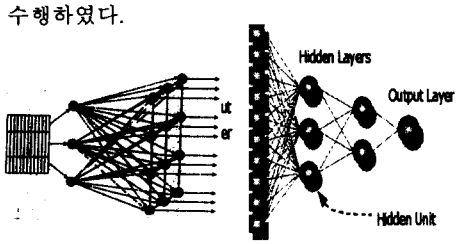
BIRCH[1]는 원시데이터를 직접 다루지 않고 군집에 속한 데이터 개체의 수, 개체들의 선형 합, 개체들의 제곱 합으로 구성된 군집의 요약정보인 군집특징을 이용한다. 먼저 전체 데이터에 대한 전클러스터링(preclustering)를 한 후 중심값에 기반한 계층적 클러스터링을 수행한다. 이 방법은 새로운 데이터가 추가되면 새로운 군집 특징은 이전의 군집 특징으로부터 계산할 수 있는 특징이 있어 유연한 구조를 갖고 있지만, 중심값에 기반한 계층적 클러스터링 기법으로 큰 군집과 작은 군집이 존재할 때 큰 군집은 작게 쪼개고 작은 군집은 합치는 현상을 나타내기 때문에 크기가 다른 군집이 존재할 때는 올바른 클러스터링을 수행하지 못하는 단점이 있다.

CURE[6]는 큰 데이터를 다루기 위해 무작위 표본추출과 계층적 클러스터링 기법을 혼합한 방법이다. 각각 군집으로부터 잘 분포된 몇 개의 데이터 개체를 선정하고 이 점들이 군집의 중심값을 향해 일정 비율 만큼 모이도록 한 후 군집을 나타내는 대표값으로 사용한다. 그러나 대표값들에 대한 최단연결법으로 유사도를 판별하기 때문에 군집간의 상호 연관성을 판별하기에 부족하다.

이러한 단점을 보완하기 위하여 대용량의 데이터에 적용할수 있고 사용자가 얻고자 하는 출력을 생성할 수 있는 있는 감독학습을 적용한 혼합형 신경망 클러스터링 기법을 제안한다.

3. 혼합형 신경망 알고리즘

이 기법은 자기조직화지도 방법과 백프로파게이션 방법을 접목한 방법으로 두 단계를 거쳐 클러스터링을 수행하였다.



dataset →1차군집→ 소군집 →2차군집→ 최종군집
SOM BP

그림 1 개 요 도

첫 번째 단계는 초기의 소군집을 발견하기 위한 단계로 자기조직화 지도를 이용하고, 두 번째 단계는 사용자의 의도에 맞는 출력을 생성하기 위하여 가중치를 조절하는 백프로파게이션 방법으로 최종 군집을 생성하였다. 이와 같은 소군집들의 반복적인 병합과정을 통해 원하는 군집을 발견해 낸다.

제안하는 신경망 알고리즘은 자기조직화지도(SOM)과 백 프로파게이션(BP)를 혼합하여 두 방법의 장점을 이용한다.

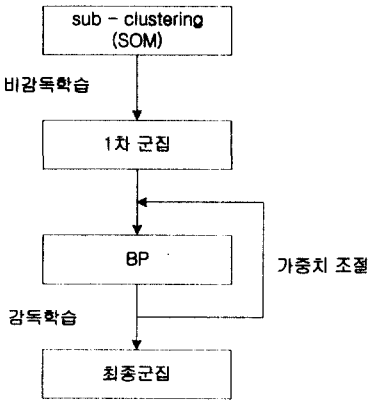


그림 2 알고리즘

자기조직화지도는 입력벡터와 가장 유사한 연결강도벡터를 갖는 뉴런인 승자뉴런 뿐만 아니라 위상적

으로 주변의 이웃관계에 있는 뉴런의 연결강도벡터까지도 조정할 수 있다. 소클러스터링을 수행하여 대용량의 데이터를 대표값으로 표현되는 소군집으로 요약하고 이 정보를 이용해 사용자의 의도에 따라 가중치를 조절하여 다양한 특징을 갖는 군집을 발견할 수 있는 클러스터링을 수행한다.

SOM을 수행하는 단계는 다음과 같다.

step 1. 연결강도를 초기화한다.
step 2. 새로운 입력 벡터를 제시한다.
step 3. 입력 벡터와 모든 뉴런들간의 거리를 계산한다.
step 4. 최소거리에 있는 출력 뉴런을 선택한다. 그 출력 뉴런의 이웃뉴런들도 선택한다.
step 5. 승자 뉴런과 이웃 뉴런들의 연결강도를 조정한다.

$$W(t+1) = W(t) + N * (X - W(t))$$

t : 시간
W(t) : t시점의 입력벡터
N : 학습률
step 6. 단계 2로 가서 반복한다. 모든 뉴런들이 변화가 없을 때 종료한다.

4. 실험 및 분석

실세계 데이터에 대한 실험을 위해 [표1]와 같이 UCI Machine Learming Repository의 데이터 집합인 Australian, Credit, Pima indians diabetes, Heart disease, Iris, Soybean, Zoo 데이터를 사용하였다.

표 1 실험 데이터

실험데이터	특징수	데이터개체수	군집수
Australian	14	690	2
Iris	4	150	3
diabetes	8	768	2
Heart	13	270	2
Soybean	35	47	4
Zoo	17	101	7

5.결론 및 향후 연구방향

제안한 알고리즘은 대용량의 데이터에 적용할 수 있고 데이터에서 알려지지 않은 패턴을 발견할 뿐만 아니라 사용자가 얻고자 하는 출력을 생성할 수 있는 있는 감독학습을 적용한 혼합형 신경망 클러스터링 기법이다. 또한 출력에 영향을 미치는 속성에 가중치를 조절하여 양질의 군집을 발견할 수 있다.

향후 연구로는 CRM과 접목하여 범주속성을 갖는 데이터에 데이터마이닝을 이용하여 고객 중심 서비스를 위한 연구가 필요하다.

참고문헌

[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]CRM을위한 데이터마이닝 Alex Berson 대청

[5]Kaufman, Leonard and Rousseuw, Peter J., Finfing Groups in data - An Introduction to Cluster Analysis, Wiley Series in Probability and Mathematical Statistics, 1990.

[6]S. Guha, K. Shim, and R. Rastogi. ROCK: A robust clustering algorithm for categorical attributes. Data Engineering, 1999