

# 상황인식기반 선형회귀의 적응적 가중치를 적용한 클러스터링

이강환\*

한국기술교육대학교 ·

## Clustering with Adaptive weighting of Context-aware Linear regression

Kang-whan Lee

Korea University of Technology and Education

E-mail : kwlee@koreatech.ac.kr

### 요 약

본 논문은 이동노드의 클러스터링내에서 보다 효율적인클러스터링을 제공하고 유지하기위한 딥러닝의 선형회귀적 적응적 보정가중치에 따른 군집적 알고리즘을 제안한다. 대부분의 클러스터링 군집데이터를 처리함에 있어 상호관계에 따른 분류체계가 제공된다. 이러한 경우 이웃한 이동노드중 목적노드와는 연결가능성이 가장높은 이동노드를 클러스터내에서 중계노드로 선택해야 한다. 본 연구에서는 이러한 상황정보를 이해하고 동적이동노드간 속도와 방향속성정보간의 상관관계의 친밀도를 고려한 자율학습기반의 회귀적 모델에서 적응적 가중치에 따른 분류를 제시한다. 본 논문에서는 이러한 상황정보를 이해하고 클러스터링을 유지할 수 있는 자율학습기반의 적응적 가중치에 따른 딥러닝 모델을 제시 한다.

### 키워드

Clustering Algorithm, K-means Clustering, non-supervisor learning, Multiple regression, Context-awareness

## I. 서 론

일반적으로 클러스터링 기반의 데이터를 처리함에 있어 데이터의 속성간의 분해와 상관관계에 따른 분류는 클러스터링을 유지하고 분석함에 매우 중요한 요소이다. 특히 상황인식기반에서의 클러스터링 단위 데이터를 분석하고 이를 제공함에 있어 비지도학습에 따른 K-means알고리즘은 가장 많이 적용되는 방법이기도 하다. 본 논문에서는 입력 제공되는 노드의 입력데이터를 분석하고 구분하여 속성단위의 노드 클러스터링을 형성한다. 이때 각 속성정보의 가중치를 적응적으로 적용한 클러스터링과 클러스터링된 네트워크의 노드간 망처리를 함에 있어 노드간의 상관 친밀도를 분석하여 동적이동노드간의 속성정보의 상관관계를 최적상태로 유지하는 것은 자율학습기반 딥러닝의 회귀적 모델을 제시하고자 한다.

## II. 본 론

크기가 비교적 균일한 클러스터를 생성하는 클러스터링 데이터에서 유효한 데이터의 한계성을 극복하는 것은 매우 중요하다. 이 과정에서 일반적으로 K-means클러스터링 알고리즘은 다음과 같은 3가지의 문제점에 대해 고려를 해야 한다[1][2].

- 클러스터내 데이터 속성의 분포가 K-means 군집화의 성능에 어떠한 영향을 미칠 수 있는지를 고려
- K-means 클러스터링에서 데이터의 유사성과 비유사성의 독립적 척도가 제공되는지의 여부
- K-means에 의한 군집화 결과의 분포는 데이터의 중심속성에 의한 유지성의 확보

따라서 위와 같은 문제점을 해결하기 위해 제시된 알고리즘에 따라 분류된 사용자속성 분류의 그룹은 데이터의 특징점 속성에 따라 상당히 다른 모집단의 변동을 비교할 수 있는 속성 가중치 인자가 요구된다. 즉, 변화되는 입력데이터에 대한

\* corresponding author

주어진 속성가중치 인자의 적응적 변화값은 클러스터링 단위에서 속성정보에 대한 데이터의 유사성과 비유사성의 척도를 판별은 클러스터링의 중요한 요소하기도 하다. 이는 평균값에 의해 모집단의 데이터 변화에 따른 오염도를 특정할 필요가 있음을 의미하기도 한다. 본 논문에서는 데이터의 유사성과 비유사성의 척도를 판별할 수 있는 독립적 척도를 제공하여 클러스터링내에서 속성데이터의 유지 및 이탈을 결정하게 된다. 본 연구에서는 데이터의 유효성을 위한 편향값(Biased effect) 내지는 클러스터의 분기된 클러스터에서 적응가중치가 적용된 속성데이터의 독립 속성척도값(Gini Index Property(GIP))을 제공하여 클러스터링내의 데이터를 분석하고자 한다. 이를 위해 본 연구에서는 각 클러스터내에서의 데이터의 유효성에 대한 상대빈도값으로 클러스터(C<sub>i</sub>)에서의 데이터 속성(i)의 상대빈도값 P(C<sub>i</sub>)을 Gini Index(GI)로 표현하여 다음과 같이 나타낼 수 있다.

$$GI(C_i) = 1 - \sum_i P(C_i)^2 \quad (1)$$

이는 임의의 클러스터(C<sub>i</sub>)에서 데이터 속성(i)에 대한 상대빈도율을 의미하는 것으로 GI는 적응수록 데이터의 유효성은 높게 평가된다.

이를 속성데이터의 척도값으로 적용하자면, 만일 어떤 k개의 임의 속성 클러스터에서 해당 클러스터내에서 특정 속성의 데이터가 변화되어 분기되었을 때 전체의 클러스터의 분기된 클러스터에서 각 속성데이터에 대해 가중치가 적용된 속성데이터의 독립적 속성척도값(Gini Index Property(GIP))은 다음과 같이 나타낼 수 있겠다.

$$GIP(C_i) = \sum_{i=1}^k \frac{n_i}{n} GI(C_i) \quad (2)$$

n=클러스터내 전체데이터의 갯수,

n<sub>i</sub>=클러스터(C<sub>i</sub>)에서의 변화된 데이터갯수

따라서, 임의의 클러스터(C<sub>i</sub>)에서 상관관계를 가지는 데이터의 변화에 대한 속성데이터의 독립적 속성척도값(Gini Index Property(GIP))은 클러스터내에서 중심값을 유지해야 하는 K-means에 의한 군집 유지성을 의미하게 된다. 적용하고자 하는 K-means 클러스터링 알고리즘은 비지도 학습방법 알고리즘으로 제공되는 이러한 속성정보로부터 일반적으로 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 비지도 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다. 본 연구에서는 누적된 사용자의 데이터의 속성을 사용하여 노드의 속성정보에 따른 구분을

하려고 한다. 주어진 데이터 집합으로부터 이러한 중심점  $\mu_k$ 의 값을 결정하고 이때 독립적 속성척도값(Gini Index Property(GIP))을 제공하여 클러스터내에서 중심값을 유지하는 것이다. 이때 각 속성정보에 따른 클러스터링 단위에서의 변화되는 입력데이터로는 속성클러스터링의 상관적 관계로 분류되며 입력 노드의 속성정보는 속도와 방향정보를 이용한 보정가중치를 이용하여 상호간의 중심점 노드로부터 클러스터링의 단위변화가 이루어지게 된다.

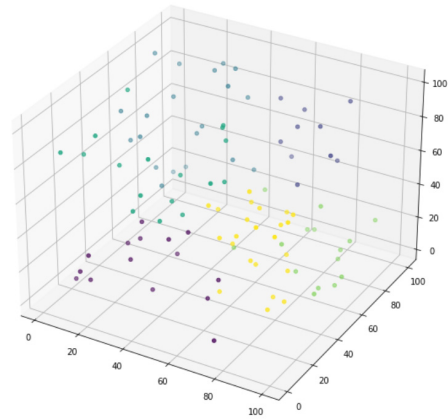


Fig. 1 Convergence of the purpose function according to multiple regression adaptive weighting learning

본 연구에서는 이러한 속성정보에 따른 클러스터링의 예측되는 오차값을 최소화하기 위하여 이동노드의 속도와 방향에 대해 회귀적 오차학습을 적용하여 이의 오차값을 최소화하는 과정의 딥러닝 학습을 적용하여 예측오차를 최소화 할 수 있도록 다음의 그림에서처럼 상황정보의 입력벡터로부터 오차값을 보정하는 절차를 수행하게 된다.

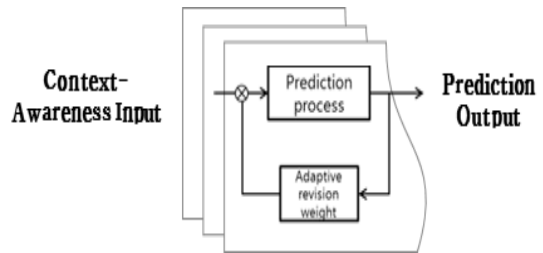


Fig. 2 The application of adaptive regressive weighting vector in Context-awareness

이에 따라 구성된 시스템의 흐름도는 그림 3과 같다. 흐름도에서 나타난 것과 같이 우선적으로 군집의 수를 지정하기 위해 시스템에서 사용할 사용자

의 수를 입력받고 이의 클러스터링과 중심속성을 확보하고 유지한다.

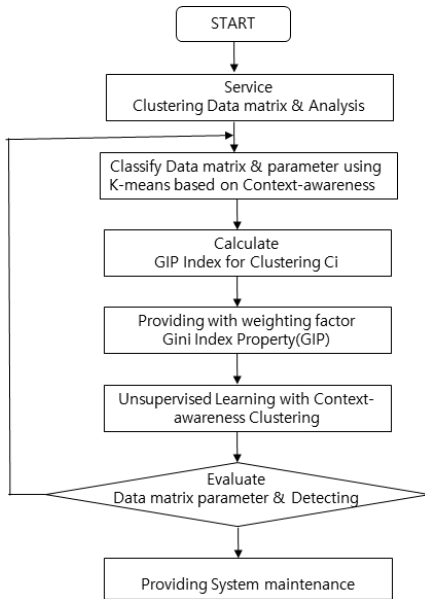


Fig. 3 Flow chart of proposed user's context-awareness unsupervised learning system in K-means clustering

### III. 결 론

논문에서는 상황인식정보를 이용한 클러스터링내에서 효율적인 속성정보에 따른 군집 알고리즘을 제안하였다. 일반적으로 헤드노드를 중심으로 관리되는 클러스터링은 종속된 그룹데이터의 유사도 측정이 필요하며 이로부터 실질적인 클러스터링 유지관계가 성립된다. 즉, 입력된 속성 데이터 분석에 따라 클러스터링된 어느 한 속성인자에 대해 사용자의 정보가 급격하게 변화하더라도 주어진 GIP와 가중치 분석에 의해 같은 사용자로 인식이 가능하여 정보의 수정이 필요가 없어 유연하게 반응 할 수 있었다. 본 연구에서는 이러한 연구의 결과를 이동통신망에서 클러스터링을 수행하고, 노드의 중심 속성정보를 이용한 클러스터링 유지 알고리즘연구에 연구결과를 적용 할 수 있을 것이다[4].

### References

[1] A. Likas, N. Vlassis and J.Verbeek, "The global k-means clustering algorithm," Department of Computer Science, University of Ioannina, 45110

Joannina, Greece, pp 452-456, 2003.  
 [2] L. Xue and W. Luan, "Improved K-means Algorithm in User Behavior Analysis," Dalian Maritime University, pp.339-341, 2015.  
 [3] Lindgren, "Probabilistic Routing in Intermittently Connected Networks", ACM SIGMOBILE Mobile Computing and Communications Review, Vol. 7, no. 3, pp. 19-20. 2003.  
 [4] X. Meng, G. Xu, T. Guo, Y. Yang, W. Shen, and K. Zhao, "A Novel Routing Method for Social Delay-Tolerant Networks", Tsinghua Science and Technology, Volume 24, Number 1, February 2019.