JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 5: 755~762, May. 2022

클러스터링 알고리즘기반의 COVID-19 상황인식 분석

이강환*

Analysis of COVID-19 Context-awareness based on Clustering Algorithm

Kangwhan Lee*

*Professor, Department of Computer Science Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

요 약

본 논문에서는 학습 예측이 가능한 군집적 알고리즘으로 COVID-19에서 상황인식정보인 질병의 속성정보와 클러스터링를 이용한 군집적 알고리즘을 제안한다. 클러스터링 내에서 처리되는 군집 데이터는 신규 또는 새롭게 입력되는 정보가 상호관계를 예측하기 위해 분류 제공되는데, 이때 새롭게 입력되는 정보가 비교정보에서 오염된 정보로 처리되면 기존 분류된 군집으로부터 벗어나게 되어 군집성을 저하시키는 요인으로 작용하게 된다. 본 논문에서는 COVID-19에서의 질병속성 정보내 K-means알고리즘을 이용함에 있어 이러한 문제를 해결하기 위해 질병 상호관계 정보 추출이 가능한 사용자 군집 분석 방식을 제안하고자 한다. 제안하는 알고리즘은 자율적인 사용자 군집 특징의 상호관계를 분석학습하고 이를 통하여 사용자 질병속성간에 따른 클러스터를 구성해 사용자의 누적 정보로부터 클러스터의 중심점을 제공하게 된다. 논문에서 제안된 COVID-19의 다중질병 속성정보군집단위로 분류하고 학습하는 알고리즘은 적용한 모의실험 결과를 통해 사용자 관리 시스템의 예측정확도가 학습과정에서 향상됨을 보여주었다.

ABSTRACT

This paper propose a clustered algorithm that possible more efficient COVID-19 disease learning prediction within clustering using context-aware attribute information. In typically, clustering of COVID-19 diseases provides to classify interrelationships within disease cluster information in the clustering process. The clustering data will be as a degrade factor if new or newly processing information during treated as contaminated factors in comparative interrelationships information. In this paper, we have shown the solving the problems and developed a clustering algorithm that can extracting disease correlation information in using K-means algorithm. According to their attributes from disease clusters using accumulated information and interrelationships clustering, the proposed algorithm analyzes the disease correlation clustering possible and centering points. The proposed algorithm showed improved adaptability to prediction accuracy of the classification management system in terms of learning as a group of multiple disease attribute information of COVID-19 through the applied simulation results.

키워드 : 상황인식, 클러스터링 알고리즘, K-평균 클러스터링, 비지도학습

Key word : Context-awareness, Clustering Algorithm, K-means Clustering, Non-supervisor learning

Received 8 February 2022, **Revised** 18 March 2022, **Accepted** 23 March 2022

*** Corresponding Author** Kangwhan Lee(E-mail:kwlee@koreatech.ac.kr, Tel:+82-41-560-1356)
Professer, Department of Computer Science Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.5.755 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 대규모 데이터를 처리 및 응용의 사례에 있어 클러스터링 군집 데이터를 처리하기 위한 관리 시스템의 예측정확도 향상을 위해 다양한 분석 기법이 개발되어 제공되고 있는 추세이다[1]. 이러한 대용량의 데이터가 제공되고 분석해야하는 의료정보(Healthcare) 분석과 자율자동차(Autonomous Vehicle), 지능형로봇(Intelligent Robot) 및 비즈니스 애플리케이션 영역과 자연어처리(Natural Language Processing), 기후모델링(Climate Modeling), 영상 및 음성처리에서는 군집화된 데이터의 처리와 이의 상관관계를 해석하고 예측하기 위한 기계학습에 대한 연구가 동반되어야 한다[2-4].

특히 의료정보(Healthcare) 분석에서는 사용자의 질병에 대한 예측과 정확한 정보제공을 위해서는 사용자의 기저질환정보를 기반으로 제공된 질환에 대한 속성정보를 군집단위로 해석 및 분석의 과정이 필요하다. 이러한 군집단위의 제공된 사용자의 정보로부터 각 기저질환이 상호 미치는 영향은 사용자 관리시스템에서 이해하고 분석하는 방법으로 사용자의 특징이 되는 속성정보를 분석하여 군집내에서 기준이 되는 중점데이터군의 이동추이로부터 서비스 제공이 가능 하다.

일례로, COVID-19의 경우 각 사용자의 기저 질환정보로써 호흡기질환(Respiratory diseases)과 순환질환(Circulatory diseases) 및 사용자의 나이(Age)에 따른 상관관계가 COVID-19에 영향을 미치는 상호 상관관계와 예측 추이에 따른 생존가능성에 대한 이해가 필요하게 된다.

이러한 사용자의 질병 속성정보를 이용한 상황인식 클러스터링 내에서 학습되고 예측이 가능한 군집적 알고리즘은 데이터를 관리함에 있어 보다 효율적인 군집의 데이터 관리를 위한 시스템이 요구되고 있으며 이에 관한 데이터 학습으로부터 시스템의 성능을 향상하고자 하는 연구가 최근 활발하게 진행되고 있다[5-8]. 또한, 사용자의 COVID-19와 같은 질환 질병에 대한 군집적 데이터를 분석하기 위해 기존질병에 대해 자동으로 사용자의 질병속성과 상관관계를 분석 제공 할 수 있는 시스템에서는 군집정보의 특성에 의해 군집내 분류하는 알고리즘에서 오염된 정보로 분류하게 될 경우, 같은 질병에 대한 데이터 값을 가진 사용자의 경우라도 다른 군집으로 분류된 사용자로 인식하는 문제가 발생할 수 있다.

본 논문에서는 이러한 COVID-19 환자의 클러스터링내 속성간 문제를 해결하기 위해 K-means알고리즘을 이용함에 있어 질병 속성 데이터의 유사성과 비유사성의 독립척도에 기준하여 사용자의 기저 질병 속성간의 정보로부터 사용자 클러스터링 정보 추출이 가능한 사용자 군집 분석 방식을 제안하고자 한다. 제안하는 알고리즘은 COVID-19 시스템 내 누적된 기저 질병 속성 정보를 이용하여 자율적인 사용자 군집 특징을 분석하고, 이를 통하여 사용자 속성에 따른 클러스터링을 구성해 사용자를 구분 유지한다. 따라서 기저질환이 있거나 또는 없는 정상적인 다수의 사용자정보를 유사성과 비유사성의 독립척도에 기준하여 사용자 군집 특징이 되는 속성 정보로부터 분석된 클러스터링 방법은 군집 내에서 상호 클러스터링의 기준이 되는 중심속성의 확보를 하고 누적된 클러스터링은 사용자 정보의 학습에 따라 예측율의 정확도가 향상된 서비스의 제공이 가능하게 된다.

Ⅱ. 관련 연구

비지도 학습방법 알고리즘으로 잘 알려진 K-Means 알고리즘은 분할 클러스터링 알고리즘 중 하나이다. 이 방식은 클러스터 수로 분류하는 데 있어 어떤 주어진 사용자에 대한 N개의 데이터를 속성단위로 K개의 클러스터 군집 단위로 분류하는 알고리즘이다. 만일 N개의 처리해야 할 사용자의 데이터가 있다고 가정하자. 이 경우 입력된 데이터는 N보다 작거나 같은 K개의 군집으로 나누게 된다. 이때 사용자의 상호 속성 정보와 정규화된 상호거리에 기반 하여 군집 간 비유사도와 같은 비용 함수를 최소화하는 방식으로 군집화를 형성하게 된다. 따라서 같은 군집 내의 속성 유사도는 증가하고 다른 군집간의 유사도는 상대적으로 감소한다[3-8]. 여기서 데이터베이스에 저장된 사용자에 대한 각 속성정보단위로 분류된 군집단위수인 K는 속성에 따라 분류된 클러스터 수로 주어진 분류속성의 특징에 따라 분류된 속성 군집내의 속성 개체를 의미한다. 또한 데이터베이스에 저장된 사용자에 대한 N개의 데이터는 개개의 속성정보들로 구성된 K-means클러스터링 알고리즘을 구성함에 있어 다음과 같은 문제점에 대한 분석이 동반되어야 한

다[9-10].

첫째로, 클러스터에 분포하고 있는 데이터 속성의 분포가 K-means 군집화의 성능에 어떻게 영향을 미칠 수 있는지에 대한 문제 분석이 필요하다. 둘째로는 데이터의 유사성과 비유사성의 독립적 척도를 측정하기 위한 수준의 데이터 클러스터링이 구성되어야 한다. 마지막 셋째로는 군집화의 결과 분포에 따른 데이터의 분포는 데이터의 중심속성에 의해 유지성이 확보되어야 한다. 마지막으로 각 클러스터링의 중심점은 상호 인접한 다른 클러스터의 중심점과 충분한 마진이 확보되어야하며 또한 데이터 학습에 따른 이동 예측이 가능해야 한다. 따라서 위와 같은 조건을 만족하기 위해 제시된 알고리즘에서는 사용자속성 분류 군집단위 데이터의 특징점 속성을 비교 할 수 있는 속성가중치의 인자가 요구되며 이는 상당히 다른 군집단위에서 모집단의 변동을 관찰할 수 있게 된다. 즉, 주어진 속성가중치 인자의 평균값에 의해 군집단위에서 모집단의 데이터 변화에 따른 오염도를 특정할 필요가 있다. 본 논문에서는 모집단 속성인자의 평균가중치를 적용하기 위해 지니(GINI Index) 계수와 모집단 속성인자의 평균가중치를 적용한다[7-9]. 또한 클러스터에 분포하고 있는 사용자 데이터 속성에 따른 데이터의 유사성과 비유사성의 척도를 판별할 수 있는 독립적 척도를 제공하여 클러스터링 내에서 속성 테이터의 유지 및 이탈을 결정하게 된다.

본 논문의 본론에서는 최적의 클러스터링을 구성하기위한 방법을 논하게 될 것이다. 본 내용에서는 군집데이터간의 상대빈도 값으로 지니계수를 도입하여 군집데이터의 독립적 속성 척도값을 제시하고자 한다. 그리고 제시된 알고리즘에 따라 COVID-19 환자의 기저질환 속성에 따른 생존율을 분석한 실험절차 및 결과를 보여주고 마지막으로 사용자의 데이터 학습에 따른 정확도를 분석하고 제시하고자 한다.

Ⅲ. 클러스터링 기반 상황인식 분석

3.1. K-means 클러스터링의 유효성

클러스터에 분포하고 있는 데이터 속성으로부터 유효성 있는 최적의 클러스터 수를 결정은 매우 중요하다. 특히, 유사한 데이터의 상관성으로부터 클러스터링된 결과의 정확성 확인과 관련된 클러스터 유효성의 척도는 K-means 클러스터링에서 매우 중요한 논점중 하나이다. 이러한 클러스터링의 유효성에 관한 연구는 Dunn에 의해 연구된 분리지수(DI)와 Davies-Bouldin이 제안한(DBI)는 클러스터링 알고리즘의 지수척도를 고전적으로 제안하였다[11-12]. 클러스터 간 거리 및 클러스터 내 거리 최소화를 정의하고 이로부터 클러스터링을 유지하는 방법이 초기에 제안된 Dunn의 분리지수(DI) 척도이다. 이로부터 확장된 연구로 제안된 DBI는 각 분리된 클러스터에 있어 분포의 합계에 의한 비율 함수로 클러스터 간 분리 척도를 재 표현하고 있다. 본 논문에서는 이러한 유효성 척도를 비교하며 유효성을 판단하기 위하여 지니(GINI Index)계수를 이용한다. 또한 모집단 속성인자에 대한 평균가중치를 적용하여 K-means 클러스터링에서 데이터의 유사성과 비유사성을 구분하게 된다.

3.2. 클러스터링의 엔트로피 측정

데이터의 속성정보를 포함한 클러스터링에서 크기가 비교적 균일하며 클러스터를 생성하는 유효한 데이터의 한계성을 극복하는 것이 매우 중요하다. 특히, K-means 알고리즘은 크기가 비교적 균등한 클러스터를 생성하고 이때 주어진 데이터를 k개의 클러스터 수로 분할함에 있어 데이터의 속성정보에 따른 유효성과 비유효성간의 독립적 척도가 제공되어야 한다. 본 연구에서는 데이터의 유효성을 위한 편향값(Biased effect) 내지는 클러스터의 분기된 클러스터링 가중치가 적용된 속성데이터의 독립 속성척도값(Gini Index Property (GIP)을 제공하여 클러스터링내의 데이터를 분석하고자 한다. 이를 위해 본 연구에서는 각 클러스터에서의 데이터의 유효성에 대한 상대빈도 값으로 클러스터(Ci)에서의 데이터 속성(i)의 상대빈도값 P(Ci)을 Gini Index(GI)를 이용한 로짓함수로 계산하면 다음과 같이 나타낼 수 있다.

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

이는 데이터 속성(i)에 대한 임의의 클러스터(C_i)에서 상대분포율을 의미하는 것으로 GI는 적을수록 데이터의 유효성은 높게 평가된다. 즉, 어떠한 속성분포를 지닌 클러스터에서 데이터의 유효성 판단의 척도 기준은 로렌츠곡선으로 나타낼 수 있는데, 만일 로렌츠곡선

에서 아래로 볼록한 형상을 나타내게 되면 속성정보의 비유효한 데이터가 분포도가 증가하게 됨을 의미하고 이로부터 속성정보의 비유효한 데이터가 분포척도 기준을 알 수 있게 된다.

본 논문에서는 이를 속성데이터의 척도 값으로 적용하고 속성데이터의 가중치로 적용하게 되는데, 만일 전체 클러스터의 속성정보의 가중치에 의해 분기된 클러스터에서 어떤 k개의 임의클러스터에서 해당 클러스터에서 특정 속성의 데이터가 변화되어 분기되었을 때 각 속성데이터에 대해 가중치가 적용된 속성데이터의 독립적 속성척도값(Gini Index Property(GIP)을 표현하면 다음과 같이 나타낼 수 있다.

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

n=클러스터에서 전체데이터의 갯수,
n_i=클러스터(C_i)에서의 변화된 데이터갯수

본 연구에서는 K-means에 의한 군집 유지성과 변화의 추적이 가능한 클러스터에서 중심값은 임의의 클러스터(C_i)에서 상관관계를 가지는 데이터의 변화에 따른 속성데이터의 독립적 속성척도값(Gini Index Property (GIP)를 의미하게 된다.

연구로부터 제공되는 독립적 속성척도값(Gini Index Property(GIP)은 K-means에 의한 클러스터에서 중심값을 유지해야 하는 상관관계를 가지는 데이터의 변화에 대한 속성이다. 이러한 사용자 클러스터(C_i)의 분포는 데이터의 중심속성에 의해 데이터의 평균값을 제공하게 된다. 즉, 이는 사용자의 변화 또는 클러스터링을 유지해야 하는 데이터망 또는 분석의 반복적 학습에서 사용되어 데이터 속성의 적응적 가중치를 제공하게 되고 또한 분석된 클러스터에서 중심값은 사용자의 속성에 따른 클러스터링유지에 활용이 된다.

금번 연구의 목적된 실험중 하나로는 COVID-19에서 사용자의 질병속성간 신체정보를 획득하고 이로부터 클러스터에서 시간에 따른 사용자의 변화 추이를 관찰 및 추적하게 한다.

Ⅳ. 실험 및 성능평가

4.1. COVID-19의 군집성 실험절차

본 연구에서는 COVID-19관리 시스템에서 기저질환 사용자 관리서비스를 제공하기 위해서 기존에 누적된 사용자의 데이터를 통한 K-means 클러스터링 알고리즘[5][6]을 사용하여 기저질환별로 사용자 구분을 한다. 제공된 시스템에서의 특징은 독립적 속성척도값(Gini Index Property(GIP)으로 부터 클러스터에서 중심값을 유지하고 이의 시간적 변동사항을 K-means에 의한 군집 유지성을 제공하는 것이다. 적용하고자 하는 K-means 클러스터링 알고리즘의 데이터는 미국질병통제 예방센터(Centers for Disease Control and Prevenction: CDC)의 COVID Data Tracker에서 제공된 데이터베이스 서비스를 활용하였다[13][14].

연구에서 적용된 정보는 각 사용자의 기저 질환정보로써 호흡기질환(Respiratory diseases)과 순환질환(Circulatory diseases) 및 사용자의 나이(Age)에 따른 상관관계가 COVID-19에 영향을 미치는 상호 상관관계를 각각 100개, 500개 및 1000개의 데이터셋을 대상으로 하였다. 이때 각 누적된 사용자의 속성정보로부터 속성정보를 확보하고 이때 입력 변화하는 사용자의 기저질환 정보로부터 속성척도값(GIP)을 제공하였다.

본 연구의 실험을 위해 구성된 시스템의 흐름도는 그림 1과 같다. 제시된 흐름도에서 나타난 것과 같이 우선적으로 군집의 수를 지정하기 위해 시스템에서 사용할 사용자의 수를 입력받는다. 여기서는 기저질환을 포함한 사용자로 구성된 기저질환 데이터들을 K-means 클러스터링 알고리즘에 적용하여 속성에 따라 사용자 입력 데이터를 구분한다. 이때 COVID-19에서 사용자의 생존율에 대한 정보 분석을 위해 백신의 1차 및 2차 접종여부에 따른 생존율을 분석하고 백신을 접종한 경우와 접종하지 않은 경우에 대한 요인을 추가하였는데 백신의 경우 현재 질병관리청 통계를 기준으로 효과를 설정하였다. 본 연구에서 적용되는 K-means 클러스터링 알고리즘은 비지도 학습방법이며, 제공되는 이러한 기저질병의 속성정보로부터 사용자의 주어진 속성 데이터를 유효한 k개의 클러스터로 묶는 알고리즘으로 시작된다.

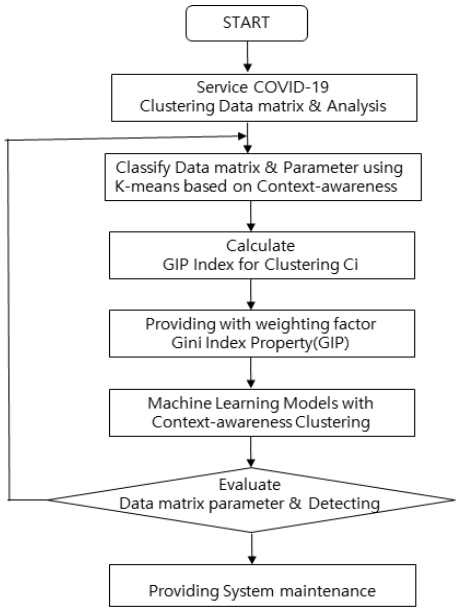

Fig. 1 COVID-19 Flow chart of proposed user's context-awareness system in K-means

여기서 각 클러스터의 동작과 구성은 거리 차이의 분산을 최소화하는 방식으로 이루어지며, 본 연구에서 적용된 방식은 레이블이 달려 있지 않은 입력 데이터에 구분자를 달아주는 비지도 자율 학습의 일종으로 역할을 수행한다[7-8].

본 연구에서는 CDC에서 2021년 1월에서 4월에 제공한 사용자의 기저질환을 포함한 다양한 사용자의 누적된 속성 데이터를 다각적으로 구분 분석 위해서 날짜별 데이터 통계자료를 이용하여 군집분석을 진행하였다. 현재 27만명의 사용자에 대한 샘플 데이터를 월별로 분리 가공 활용하여 사용자를 나타내는 3개의 속성정보를 이용하였고 이때 클러스터 내에서 중심값을 유지해야 하는 K-means에 의한 기저질환 속성간의 군집 유지성을 확보하고 변화를 관찰 하였다.

4.2. 속성척도값을(GIP) 이용한 K-means 클러스터링

COVID-19 사용자 관리 시스템에서 적용한 K-means 클러스터링 알고리즘은 각 사용자의 기저 질환정보로써 호흡기질환(Respiratory diseases)과 순환질환(Circulatory diseases) 및 사용자의 나이(Age)에 따른 3가지의 특성들을 K-means 클러스터링 알고리즘에서 속성정보로 적용된다. 이런 정보를 가지는 n명의 데이터셋을 $X_1, \cdots, X_n$ 으로 하면 총 N개의 데이터로 D개의 차원을 가지며 k개의 클러스터로 나눌 것이다[4]. 이러한 D개의 차원을 가지는 벡터를 μ처럼 표기하고 이 벡터가 k번 째 클러스터에 속하는 경우 μ_k와 같이 표기한다. 본 연구에서는 누적된 사용자의 데이터의 속성을 사용하여 COVID-19에서 사용자의 기저질환에 따른 상관관계를 이용하여 사용자를 구분 하려고 한다. 주어진 데이터 집합으로부터 이러한 중심점 μ_k의 값을 결정하고 이때 독립적 속성척도값(Gini Index Property(GIP)을 제공하여 클러스터 내에서 중심값을 계산하고 변화를 관찰하는 것이다. 위와 같은 문제는 데이터 집합을 사용자의 속성정보단위에서 상호관계를 이용한 특정 클러스터에 할당하는 것으로 우선 변수 $r_{nk} \in 0,1$을 정의한다. 이때의 k 값은 $k = 1, \cdots, k$로 정의 된다. 만일 사용자의 속성 n 번째 샘플 X_n이 k번째 클러스터에 속하는 경우 $r_{nk} = 1$이고 아닌 경우 0이 된다. 이로부터 왜곡 측정(distortion measure) 된 목적 함수를 정의하면 식(3)과 같이 정의된다.

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|X_n - \mu_k\|^2 \tag{3}$$

다음의 표1에서는 COVID-19에서 사용된 기저질환 및 사용자 정보의 상관관계를 분석한 실험에서 사용된 변수를 보여주고 있다.

Table. 1 COIVD-19 Variable of distortion measure

Variable	Attribute
n_i	Number of total disease fatal data set
k	Number of dividing clustering
x_n	User's disease data set of x_n
μ_k	Means value of Corelation clustering μ_k
r_{nk}	Clustering Rank value for μ_k

여기서 COVID-19에서 클러스터링의 특징을 나타낼 수 있는 목적함수가 최소가 될 때까지 수렴 r_{nk}와 μ_k의 값을 구해야할 필요가 있다. 따라서 아래의 식 (4)를 이용해 가장 최적의 상관관계를 가지는 상황인식 기반의 클러스터링 r_{nk}의 값을 구한다.

$$r_{nk} = \begin{cases} 1 & \text{if} \quad k = \arg\|X_n - \mu_k\|^2 \\ 0 & otherwise \end{cases} \tag{4}$$

즉, 각 클러스터 중심과 샘플의 거리를 측정해서 가장 가까운 클러스터를 선택한 다음으로 μ_k의 값을 구축한다. 이때 r_{nk}의 값을 고정하고 목적함수를 미분하여 최소인 점을 알아 낼 수 있다.

$$2\sum_{n=1}^{N} r_{nk}(X_n - \mu_k) = 0 \tag{5}$$

이를 가장 최적의 상관관계를 가지는 상황인식 기반의 클러스터링μ_k에 대해 전개하면 다음의 식(6)이 됨을 알 수 있다.

$$\mu_k = \frac{\sum_n r_{nk} X_n}{\sum_n r_{nk}} \tag{6}$$

이러한 두 단계를 거치는 동안 데이터는 각각의 클러스터에 다시 할당이 되고, 이렇게 재 할당된 데이터를 이용하여 평균값을 다시 계산하는 과정을 반복하게 된다. 이의 의미는 결과적으로 k클러스터에 속한 값들의 평균을 구하게 되고 이의 의미로부터 클러스터링 속성의 중심점을 표시하게 되며 이로부터 누적된 사용자의 상관관계를 가지는 기저질환 정보로부터 각 클러스터링간의 중심점 이동을 추적 관찰 할 수 있다.

4.3. K-means 클러스터링 중심값의 이동 관찰 및 학습 예측 실험결과(Experimental Results)

제안한 알고리즘을 실행하기 위한 COVID-19 시스템에서 기저질환 속성에 따른 사용자구분은 그림 2와 같이 기본적으로 분류되고 구분되어 나타난다. 본 논문에서는 적용된 알고리즘을 적용하여 학습하면서 처음 입력된 데이터량이 100개, 200개, 500개, 1000개 및 5000개로 증가하여 학습실험을 하였다. 이때 임의의 사용자 값들이 클러스터링 된 곳으로 이동하고 사용자를 기저질환의 속성에 따라 변화된 중심점을 특정 지을 수 있게 되며 이는 그림 3과 같이 나타난다. 여기서 만약 사용자가 기저질환에 있어 변화가 발생할 경우 독립적 속성척도값(Gini Index Property(GIP)가 적용되면 클러스터에서 중심값을 이용하여 다른 특징들의 가중치를 이용해서 사용자의 기저질환과 상관관계를 새롭게 특정 지을 수 있다.

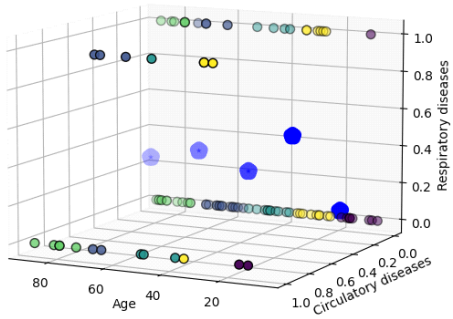

Fig. 2 User classification apply proposed algorithm(n=100) characteristic central point

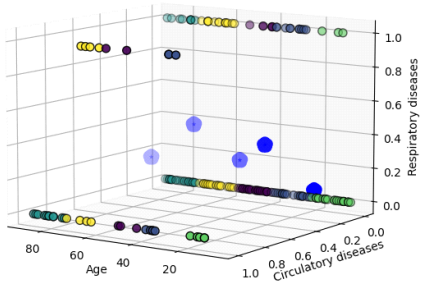

Fig. 3 User classification apply proposed algorithm(n=500) characteristic central point

다음의 그림4와 그림5는 백신을 접종한 사용자 경우에 있어 기저질환 사용자를 각각 100명, 200명, 500명 및 1000명의 누적된 데이터를 학습하고 이때 학습된 데이터의 정확도를 분석하였다. 이 경우 사용자의 기저질환에 따른 데이터를 각각 학습하는 과정에서 예측의 정확도가 97.33%에서 99.2%로 향상됨을 보여주었다. 그 요인 중에서는 나이의 가중치가 1.29에서 1.71, 순환계 질환의 가중치가 0.32에서 1.56 및 호흡계 질환의 가중치가 1.03에서 2.83로 증가하는 것을 확인 할 수 있었다. 특히 백신접종에 대한 가중치가 처음에는 1.18정도로 높지 않았지만, 1000개의 데이터를 학습하면서 가장 큰 가중치인 2.68로 증가함을 볼 수 있었다. 즉 제공된 예측 모델은 기저질환 또는 기저질환이 없는 모든 경우의 사용자에 대해 백신을 접종한 경우와 접종하지 않은 경우에 대해서도 생존율에 대한 예측 분석의 제공도 가능 할 것이다.

```
100 Data Trained Model Accuracy: 97.33%

Age = -1.291
Vascular and unspecified dementia = 0.019
Alzheimer disease = 0.129
Circulatory diseases = -0.322
Diabetes = 0.0
Malignant neoplasms = 0.072
Obesity = 0.159
Renal failure = 0.019
Respiratory diseases = -1.038
Sepsis = 0.121
Vaccinated = 1.182
```

Fig. 4 User classification of Intentionally changing (n=100) with Accuracy

```
1000 Data Trained Model Accuracy: 99.2%

Age = -1.716
Vascular and unspecified dementia = 0.148
Alzheimer disease = -0.018
Circulatory diseases = -1.568
Diabetes = 0.0
Malignant neoplasms = -0.08
Obesity = 0.174
Renal failure = -0.299
Respiratory diseases = -2.83
Sepsis = -0.056
Vaccinated = 2.688
```

Fig. 5 User classification of Intentionally changing (n=1000) with Accuracy

Ⅴ. 결 론

논문에서는 상황인식정보를 이용한 클러스터에서 COVID-19의 질병 속성정보를 이용한 효율적인 군집 학습예측 알고리즘을 제안하였다. 일반적으로 K-Means 알고리즘에서 데이터의 중심점으로부터 관리되는 클러스터링은 종속된 군집데이터의 유사도 측이 필요하다. 또한 종속데이터간의 상관관계 분석이 필요하며, 이로부터 실질적인 클러스터링 유지관계가 성립된다. 이를 위해 본 연구에서는 제안된 알고리즘을 적용한 일례로 COVID-19의 질병 속성정보로 부터 K-means 클러스터링 알고리즘을 사용하여 학습 예측이 가능한 군집적 알고리즘을 제안하였다. 사용된 CDC데이터로부터 K-means 클러스터링 알고리즘을 구축하고 사용자의 기저질병에 대한 속성정보를 활용해 질병 상호관계 정보 추출이 가능한 사용자 군집 분석 방식을 제안하고 사용자의 시스템 내 누적된 정보를 이용하여 자율적인 사용자 군집 특징을 분석학습하고, 이를 통하여 사용자 질병속성간에 따른 클러스터를 구성해 시스템에서 질병 속성의 변화 임계치에 따라 사용자의 구분을 명확히 할 수 있었다. 본 논문에서는 적용된 3종류 속성을 가진 사용자 기저질병 정보를 이용하여 사용자를 구분하기 때문에 어느 한 속성인자에 대해 사용자의 정보가 변화할시 주어진 GIP와 가중치 분석에 의해 같은 군집내 사용자로 인식이 가능하여 최종 학습된 데이터의 정확도를 향상 시켜주었다. 본 연구에서는 사용자의 누적된 데이터를 3종의 기저질환 데이터로 한정하여 사용하였지만 질병상호관계 추출이 가능한 사용자 질병속성간의 클러스터링에서 명확한 유사도 측정 정보가 제공된다면 비지도 학습 알고리즘의 클러스터링과 속성유사도의 제공된 측정정보로부터 각 기저질환의 클러스터링을 통해 사용자에 대한 치명율 및 백신을 접종한 경우와 접종하지 않은 경우에 대한 치명율도 분석 할 수 있을 것이다.

ACKNOWLEDGEMENT

This research was supported by the education research promotion program of Koreatech in 2020

REFERENCES

[1] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. -W. On, W. Aslam, and G. S. Choi, "COVID-19 Future Forecasting Using Supervised Machine Learning Models," *IEEE Access*, vol. 8, pp. 101489-101499, May. 2020.

[2] G. Adami, P. Avesani, and D. Sona, “Clustering documents in a web directory,” in *Proceedings of the 5th ACM international workshop on Web information and data management*, New Orleans, USA, vol. 54, no. 3, pp.66-73, Sep. 2015.

[3] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” *PLoS ONE*, vol. 13, no. 3, pp. 1-26, Mar. 2018.

[4] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long. “A survey of clustering with deep learning: From the perspective of network architecture,” *IEEE Access*, vol. 6, pp. 39501 – 39514, Jul. 2018.

[5] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A

K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, Jan. 2012.

[6] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451-461, Feb. 2003.

[7] L. Xue and W. Luan, "Improved K-means Algorithm in User Behavior Analysis," in *Ninth International Conference on Frontier of Computer Science and Technology*, Dalian, China, pp. 339-342, 2015.

[8] H. Xiong, J. Wu, and J. Chen, "K-means Clustering versus Validation Measures: A Data Distribution Perspective," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39, no. 2, pp. 318-331, Dec. 2008.

[9] Kangwhan-Lee, "Context-awareness User Analysis based on Clustering Algorithm," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 24, no. 7, pp. 943-948, Jul. 2020

[10] T. Obichi, Y. Okaie, T. Nakano, T. Hara, and S. Nishio, "Inbody Mobile Bionanosensor Networks Through Non-diffusion-based Molecular Communication," in *2015 IEEE International Conference on Communications*, London, U.K, pp. 1078-1084, 2015.

[11] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224-227, Apr. 1979.

[12] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *J. Cybernetics*, vol. 3, no. 3, pp. 32- 57, Apr. 2008.

[13] CDC [Internet]. Available: https://covid.cdc.gov/covid-data-tracker/support.html

[14] NYC Health [Internet]. Available: https://www1.nyc.gov/site/doh/covid/covid-19-data.page#daily

이강환(Kangwhan Lee)

1987년 한양대학교 전자공학과 학사
1989년 중앙대학교 전자공학 석사
2002년 중앙대학교 전자공학 박사
1989년 한국전자통신연구원 통신시스템연구단 선임연구원
2005년~현재 한국기술교육대학교 컴퓨터공학부 교수
2018년 San Diego State University 교환교수

※관심분야 : 지능형차세대이동통신기술, CDNN, WSN, Ad-hoc network, Wireless SoC & UoC