|  |
| --- |
| **K-Means Clustering을 이용한**  **기상환경 데이터 분석과 음악 추천 시스템 설계** |
| **Music Recommendation by applying the K-means clustering and deep learning to analyze the weather information** |
| **요 약**  많은 제어용 소프트웨어가 지능을 갖도록 설계되고 있다. 또한 방대한 양의 데이터를 고속으로 처리할  수 있는 구축 환경에서 인공지능 기법이 활용되고 있다. 기계에 인공지능을 주기 위해서는 기계를 인간처럼 판단 및 행동할 수 있도록 학습을 해야 하는데, 학습에 필요한 데이터의 분류 단계에서 k-means clustering을 적용하여 환경 데이터의 특성을 찾고 Deep learning을 사용하여 음악을 추천하는 방법을 제안한다. 본 논문에서는 실시간 기상 환경 데이터로부터 최적의 음악을 추천하는 지능 알고리즘을 구현하였다. 구현된 소프트웨어의 검증을 위해 온도, 습도 센서가 장착된 Arduino System을 구성하고 실험을 통해 추천 품질을 분석하였다. | | |

1. **서 론**

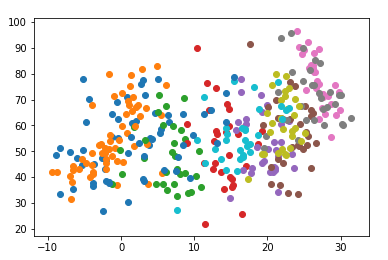
최근 출시된 전자 제품에 인공지능이 안 들어간 제품이 없을 정도로 인공지능은 현재 전자기기 분야에서 많은 각광을 받는 기술 중 하나이다. 인공 지능은 2016년도 ‘알파고’ 를 통해 전세계에 알려졌으며 그로부터 대중들에게 주목받기 시작했다. 인공지능은 인간만이 가진 고유한 특성이라고 생각했던 지능을 기계가 가지게 됐다는 점과 대용량의 데이터를 관리할 수 있게 됐다는 점에서 많은 기기들의 구현에 사용하게 되었다.

공원과 같은 공공장소에 설치된 스피커는 특정 사용자가 정의되어 있지 않고 방치된 상태이므로, 매번 입출력을 선택하여 노래를 재생시키기가 어렵다. 따라서 사전에 기기를 학습시킴으로써 입력에 따른 출력 값을 정해 놓은 인공지능 스피커를 공원의 가로등에 부착시킴으로써 상황에 맞는 노래를 자동으로 추천하는 시스템을 구성하였다.

이 논문에서는 입력으로 넣을 온도와 습도의 자료를 선정 및 기준치를 정하고 이에 따른 출력 값까지 지정하는 과정을 거쳐 인공지능 스피커를 구현하는 방법을 설명할 것이다. 인공지능을 위한 기기를 학습시키기 전에 데이터를 분류하기 위해 k-means clustering을 사용하고 deep learning 기법으로 학습한 후 음악 추천이 되도록 한다.

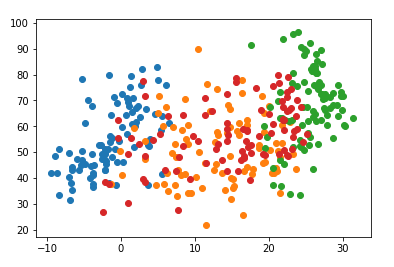
1. **본 론**
   1. **자료 선정**

우리나라는 온도, 습도 변화의 폭이 넓고 계절별 날씨 특성이 뚜렷하다. 하지만 환절기 날씨는 특정 계절로 분류하기 모호하여 계절에 따른 Clustering하기에 어려움이 있다.



<그림 1>

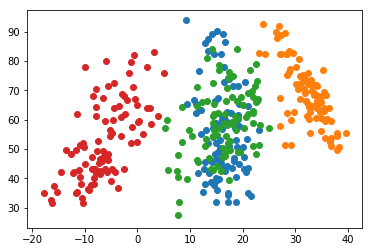
월 별 그래프화



<그림 2>

계절 별 그래프화

데이터 선정을 정확히 하기 위해서 2017년 1월 1일 ~ 2017년 12월 31일까지의 기상청 관측 자료를 이용했다. 사용한 데이터는 일 평균기온, 일 평균 습도이다. <그림 1>은 월을 기준으로 해당 날짜의 온도, 습도를 그래프화 한 것이다. 이 경우 서로 다른 달임에도 불구하고 온도, 습도가 중첩되는 지점이 상당수 존재한다. 이는 앞서 말한 Clustering의 어려움에 해당한다. 그 후 3개월 주기로 계절을 지정한 후 그래프화 한 것이 <그림 2>이다. 이는 <그림 1>의 경우보단 양호하지만 여전히 중첩되는 부분이 발생한다.

우리나라의 일최고기온과 일최저기온의 변동은 겨울철에 가장 크고 여름철에 가장 작다. 여름철의 경우 일최고기온의 감소 경향이 일교차 감소에 영향을 미친다. [1] 이러한 이유로 온, 습도 데이터의 다양성을 넓히기 위해 여름철 데이터를 일 최고기온으로, 반대로 겨울철 데이터를 일 최저기온으로 대체하였다.

<그림 3>

**데이터 기준 변화 후 계절 별 그래프화**

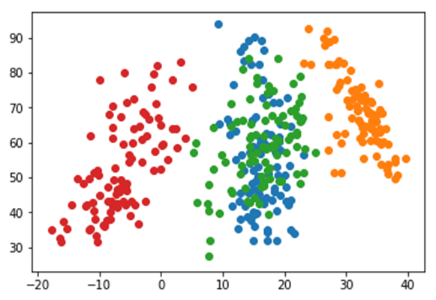
<그림 3>은 3년치 계절별 데이터 약 370개를 그래프화 한 것이다. 일 평균기온으로 그래프화 한 <그림 1>과 <그림 2>에 비해 중첩되는 지점이 감소한 것을 확인할 수 있다.

* 1. **K-Means Clustering**

온,습도 데이터를 군집화하고 그 대푯값을 선정하기 위해 본 논문은 K-means Clustering을 이용하였다. K-means는 간단하면서 구현이 용이하여 새로운 데이터에 대한 계산양이 적다. 데이터와 군집의 중심 좌표(Centroid)와의 거리를 계산하고, 데이터와 중심 좌표 간 거리가 가장 가까운 군집에 각 데이터를 배정하는 반복적 알고리즘이다. [2]

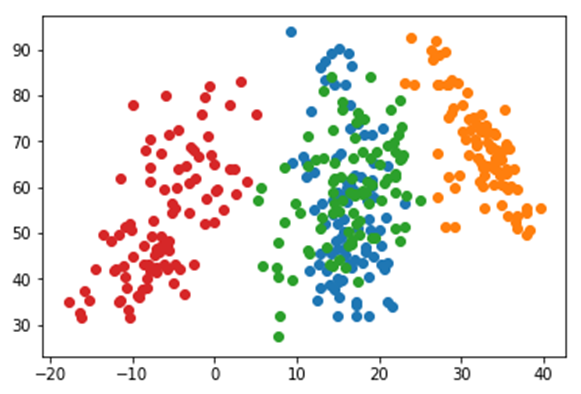
|  |
| --- |
| **Algorithm** : K-Means Algorithm |
| **Input** : K, the number of clusters  **Input** : D, A collection of a set of N vectors (|D| = N)  **Input** : M, the maximum number of iterations  **Output**: A K-partition of D such that  **begin**  Randomly initialize K cluster centers,  **for** j = 1, …, M **do**  **for** each **do**  k’ 🡨  🡨  Recompute cluster centers from the current partition |

각 input point는 반복적으로 가장 가까운 cluster에 할당된다. 총 반복 횟수인 M번 동안, Cluster의 중심 좌표는 현재의 할당된 좌표를 사용하여 재계산된다. [3]



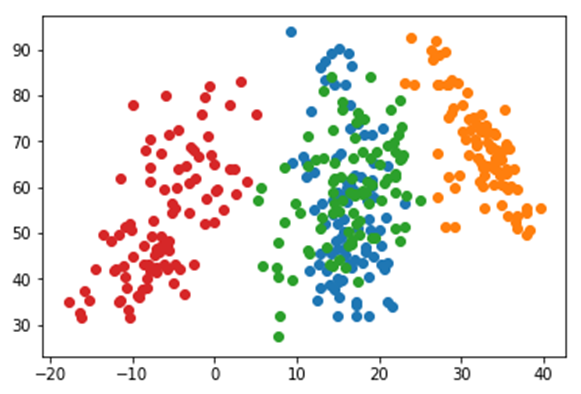
<그림 4>

**특정 날씨 임의 클러스터링**

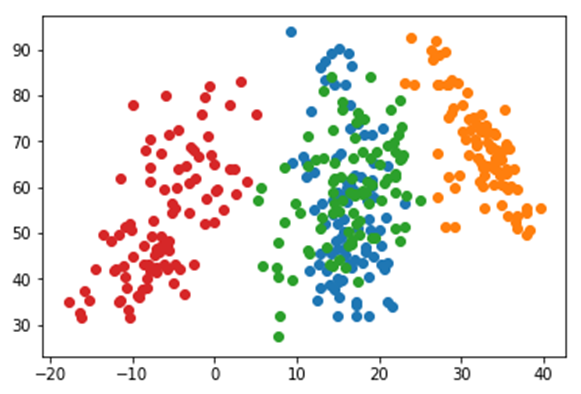


<그림 4>

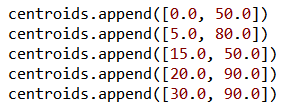
**특정 날씨 임의 클러스터링**



<그림 4>



<그림 4>



<그림 5>

**각 영역 중심 좌표 설정**

<그림 4>는 특정 날씨에 알맞은 노래를 선택하기 위해 전체 데이터를 다섯 영역을 임의로 분류한 것이다. <그림 5>는 그 영역의 중심 좌표로 설정된 영역의 개수만큼 존재하고 유클리드 거리(Euclidean Distance) 계산에 사용된다. [4]

*distance =* **<식 1>**



<그림 6>

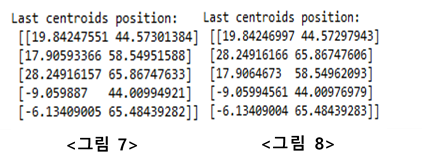
**실제 코드 내 유클리드 거리 계산 수식**



각 영역의 분리는 <식 1>의 유클리드 거리 계산을 이용한다. Point는 Cluster 내의 데이터 벡터를 뜻하고, centroid는 Cluster 내의 기하학적인 중심 좌표를 뜻한다. [5] <그림 6>은 실제 K-Means Clustering에 사용된 일부 코드이다.

* 1. **Clustering 반복 횟수에 따른 결과**

유클리드 거리 계산은 반복 횟수가 늘어날수록 정확도가 높아진다. 어떠한 데이터도 자신이 속한 Cluster가 변하지 않을 때까지 반복한다.

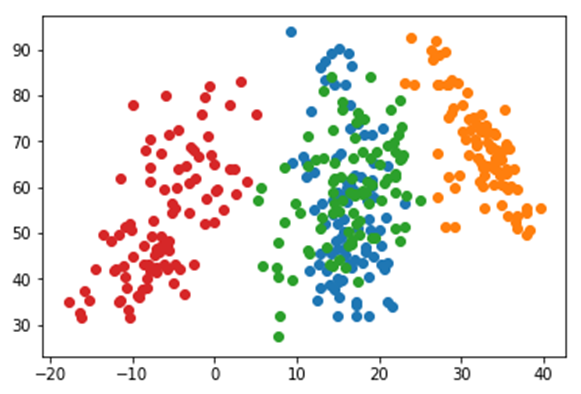


**Iteration : 1**

**Iteration : 6**

<그림 9>

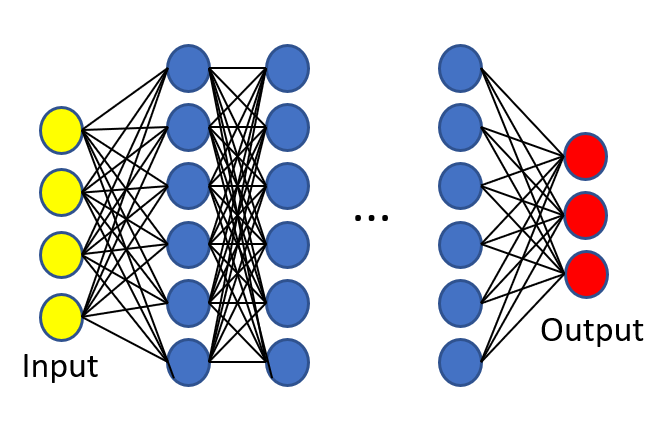
**최종 중심 좌표 표시**



<그림 7>은 iteration이 1일 때, <그림 8>은 iteration이 6일 때의 좌표이고 이때부터 중심 좌표의 위치는 반복 횟수가 늘어나도 변하지 않는다. 이것은 군집화의 최종 중심 좌표다. <그림 9>는 앞서 임의로 지정한 영역에 최종 중심 좌표를 표시한 것으로 알맞게 군집화가 된 것을 확인할 수 있다.

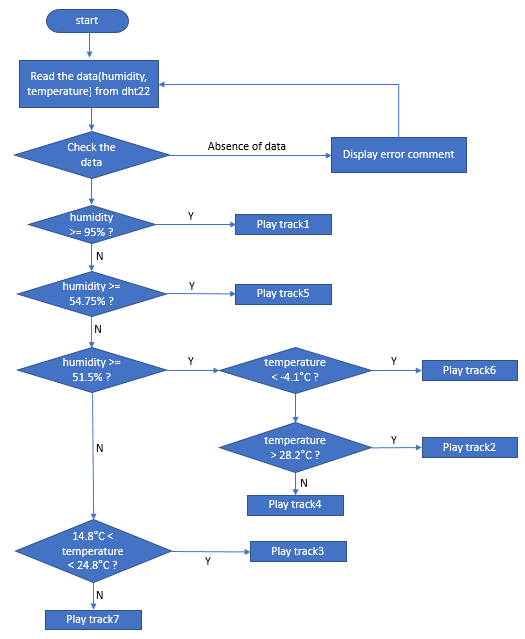
* 1. **Deep learning을 이용한 인공지능 스피커 구현**

Deep learning 클러스터는 k-means clustering 방법으로 구한 결과를 활용한다. 클러스터 구성 요소가 입력이며 속한 클러스터 번호가 출력 값으로 나온다.[6]



**<그림 10> Deep learning 개념도**[7]

본 논문에서 구현하는 인공지능 스피커는 공원에 설치된 것이므로, 실외에서 입력 값으로 가장 적절한 것은 온도, 습도이다. 앞서 정한 바와 같이 온도와 습도의 수치 값을 기반으로 하여 그룹화하며 각 입력이 들어올 때마다 그 입력 값이 속한 그룹을 찾아낸다. 각 그룹별로 테마를 정하고 그에 맞는 노래들을 미리 선정하여 한 트랙으로 묶어 놓는다. 이 트랙들이 각 그룹에 대한 출력 값으로 설정되는 것이다.



**<그림 11> rule 방식 제어 흐름도**

<그림 11>은 온습도에 따른 인공지능 스피커의 알고리즘 순서도를 나타내며, 다음과 같이 처리된다. [8]

단계 1 : dht22 센서로부터 온도와 습도 값을 읽어온다.

단계 2 : 올바른 데이터인지 확인하고 만약 데이터를 읽어오지 못했을 시에는 다시 1단계로 돌아간다.

단계 3 : 여러 if-then-else문을 통과하며 현재 온도, 습도가 포함되는 날씨 그룹을 찾는다.

단계 4 : 현재 날씨가 해당된 그룹의 출력 값으로 지정된 노래 트랙을 스피커에 재생시킨다.

예를 들어, 비가 오는 날이면 습도가 100%이므로, 단계 3의 첫번째 if문에 의해 그룹이 결정되며 비에 관련된 노래로 이루어진 track 1이 재생된다.

위와 같은 알고리즘을 통해 날씨에 맞는 노래를 추천하고 재생하는 인공지능 스피커가 작동된다.

1. **결론**

본 논문에서는 인공지능 입력 값을 위해 기간 및 기준치의 범위를 다양하게 하여 가장 적합한 자료를 선정했다. 조사한 자료를 토대로 k-means clustering 알고리즘을 이용하여 각 날씨 그룹의 대표 값을 선정했다. 입력 값에 따라 그룹을 설정하고 그룹별 출력 값을 지정해 놓음으로써 입력 값에 따른 출력 값을 미리 설정해 deep learning 학습에 적용하였다.

실험적 검증을 위해 구현한 결과, 기상 환경에 따라 적합한 노래가 추천됨을 알 수 있다.

# 참고 문헌

|  |  |
| --- | --- |
| [1] | 이은주, “K-means 군집과 유클리드 거리에 의한 선박 부식상태 진단,” *한국정보과학회*, p.2, 2013. |
| [2] | 이승호, “우리나라의 일교차 변화에 관한 연구,” *한국지역지리학회지*, pp.5-8, 2011. |
| [3] | D. Ganguly, “A Fast Partitional Clustering Algorithm based on Nearest Neighbours,” p.112, 2018. |
| [4] | 이수현, “Fuzzy K-means 군집분석을 위한 다양한 유효성 지수 개발,” *한국경영학회 경영학연구*, p.8, 2017. |
| [5] | K. R. Zalik, “An efficient k'-means clustering algorithm,” *Pattern Recognition Letters*, p.1385, 2007. |

[6] 이기남, “휴머노이드 로봇의 마라톤 경기 및 전략,” *한국지능시스템학회*, p.66, 2016.

[7]He, Ji, “**Deep Reinforcement Learning in Natural Language Scenarios,” p.9, 2017.**

[8] Schulman, John, “Optimizing Expectations: From Deep Reinforcement Learning to Stochastic Computation Graphs. [electronic resource],” pp.38-42, pp.78-84, 2016.