ISSN: 2508-7894 © 2017 KAIA. http://www.kjai.or.kr
Doi: http://dx.doi.org/10.24225/kjai.2017.5.1.1

Applying CEE (CrossEntropyError) to improve performance of Q-Learning algorithm

Q-learning 알고리즘이 성능 향상을 위한 CEE(CrossEntropyError)적용

[1] Hyun-Gu Kang(강현구), [2] Dong-Sung Seo(서동성),
[3] Byeong-seok Lee(이병석), [4] Min-Soo Kang(강민수)
[1] First Author Department of Medical IT Marketing, Eulji University, Korea,
godec1234@naver.com
[2,3] Department of Medical IT Marketing, Eulji University, Korea,
sds1zzang@naver.com, qudtjr1993@naver.com
[4] Corresponding Author Department of Medical IT Marketing, Eulji University, Korea,
Tel:+82-31-740-7190, E-mail: mskang@eulji.ac.kr

Received: June 16, 2017. Revised: June 19, 2017. Accepted: June 20, 2017.

Abstract

Recently, the Q-Learning algorithm, which is one kind of reinforcement learning, is mainly used to implement artificial intelligence system in combination with deep learning. Many research is going on to improve the performance of Q-Learning. Therefore, purpose of theory try to improve the performance of Q-Learning algorithm. This Theory apply Cross Entropy Error to the loss function of Q-Learning algorithm. Since the mean squared error used in Q-Learning is difficult to measure the exact error rate, the Cross Entropy Error, known to be highly accurate, is applied to the loss function. Experimental results show that the success rate of the Mean Squared Error used in the existing reinforcement learning was about 12% and the Cross Entropy Error used in the deep learning was about 36%. The success rate was shown.

Keywords: Deep-Learning, Q-Learning, Mean squared error, Cross-Entropy error, AI, RL

1. Introduction

제 4 차 산업혁명의 키워드는 '초연결, 초지능, 초실감'이라고 할 수 있다. 그 중 2016 년 전 세계에 이목이 집중된 이세돌 9 단과 구글의 딥마인드 사의 알파고와의 경기는 경기인공지능(AI)에 대한 관심을 높아지게 하였고 경기결과를 통해 나타난 알파고의 능력은 전 세계를 놀라게 하였다. 이를 통해 이른바 초지능(superintelligence)'에 대한 현실가능성이 다가왔음을 알 수 있다.

Steinhaus(1967)에 의하면 인공지능이란 1956 년 당시의 나온 개념으로 인간의 감각, 사고력을 지닌 채 인간처럼 생각하는 'AI(Artificial Intelligence)'을 말한다. 한편, 머신 러닝은 기본적으로 알고리즘을 이용해 데이터를 분석하고, 분석을 통해 학습하며, 학습한 내용을 기반으로 판단이나 예측을 한다. 딥러닝은 대량의 데이터와 알고리즘을 통해 컴퓨터 그 자체를 '학습'시켜 작업 병행 수행 방법을 하는 것을 말한다. 또한 딥러닝은 인공신경망에서 발전한 형태의 인공지능으로 뇌의 뉴런과 유사한 정보 입출력 계층을 황용해 데이터를 학습한다. 앞서 설명한 머신러닝은 지도학습, 비지도학습, 그리고 강화학습으로 나눌수 있다. 지도학습은 학습 데이터에 정답 레이블이 있는 것을 말한다. 정답 레이블이란 예를 들어, 우리 주변에 있는 사물들을 찍은 사진 속에서, 어떤 사물들이 있는지를 구별하는 태스크가 있다고 하면 가지고 있는 사진들을 학습 데이터라고 하고 사진 속에 있는 사물을 '컵', '책상', '자전거', '고양이'라고 미리 정의해 놓는 것을 말한다. 레이블은 사람은 사진을 보고 정의한 것이기 때문에 학습을 하는 컴퓨터 입장에서는 사람으로부터 지도(Supervised)를 받은 것이 되고 이렇게 학습된 머신러닝을 지도학습이라고 한다. 지도학습에는 분류, 예측 모델이 있다. 반면 비지도학습은 입력 데이터에 레이블이 없다. 따라서 컴퓨터가 사람으로부터 지도를 받은 것이 없기 때문에 이를 비지도 학습이라고 한다. 비지도학습으로는 군집 모델이 있다. 강화학습은 머신러닝의 분류 기준으로 볼 때 지도학습 중 하나로 분류하기도 하고, 또는 독립적으로 세 번째 머신러닝 모델로 분류하기도 한다. MacQueen(1967)에 의하면 강화학습을 지도학습으로 분류하는 이유는 에이전트가 취한 모든 행동에 대해 환경으로부터 보상과 벌칙을 지도 받아 학습하기 때문이다. 하지만 강화학습은 다른 전형적인 지도학습처럼 사전에 사람으로부터 가이드를 받고 학습하지 않을뿐더러 사람이 아닌 환경으로부터 보상과 벌칙을 피드백 받기 때문에 세 번째 머신러닝으로 분류하는 것이 일반적이다. 본 논문에서는 강화학습을 사용하였다. 강화학습 관련 측면을 좀 더 자세히 설명하면 다음과 같다. Lloyd(1982)에 의하면 시스템과 제어입력 간의 상호 작용에 보상값(Reward)으로 표현되는 평가적 신호(Evaluative Feedback Signal)를 활용하여 전체 보상값의 합에 대한 기대값을 최대화하는 최적 제어 전략(Optimal Control Policy)을 데이터를 기반으로 찾아내는 방법으로써, 인공지능 분야의 주요한 도구 중 하나로 자리 잡아 가고 있으며 최근에 발표된 강화학습 기법들은 제어 및 로봇 학습을 비롯한 각종 공학 문제에 성곡적으로 적용되는 모습을 보여준다. 앞에서 예를 들었던 2016 년에 이세돌 9 단과 구글딥마인드의 알파고 역시 강화학습을 통해 이루어졌다. 아래 Figure 1 은 알파고의 시초가 된 구글딥마인드 사의 아타리 게임에 관한 사진으로, 강화학습을 '벽돌깨기'게임에 적용시켰다.

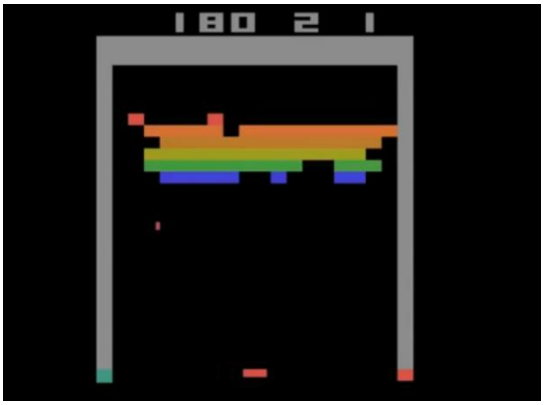


Figure 1. Google Deep Mind's Atari Game

본 논문은 머신러닝의 강화학습 알고리즘 중 기본 알고리즘인 Q-learning 알고리즘을 분석하였다.

Q-learning 알고리즘에서는 평균제곱오차함수(MSE)을 사용한다. 본 논문은 손실함수의 값을 측정하는 함수 중 평균제곱오차함수(MSE)보다 효율이 좋은 교차엔트로피(CEE)를 Q-learning 에 적용시켜 Q-learning 을 비교 분석 검토하였다. 교차엔트로피를 사용하였을 때 정확도가 확인하는 것이 본 논문의 목적이다.

2. Related Study

2.1. Double Q-Learning

2010 년 NIPS 에 발표된 Van Hasselt, Guez, & Silver (2016)의 Double Q-Learning 에 따르면 Q-Learning 은 잘 작동하지 않는데 이러한 이유는 행동 값의 큰 과대 평가에 의해 발생한다고 한다. 이러한 과대 평가는 Q-Learning 이 가장 기대 동작 값의 근사치로 최대 동작 값을 사용하기 위해 도입 된 양수의 바이어스에 유래 한다.

Double Q-Learning 은 Q-learning 의 이러한 발산을 막기 위해 다음과 같은 식 1 을 사용하여 Q-Learning 을 수행한다.

식(1)

$$L=(Q(s,a)-(r+\Gamma Q(s_{t+1},\arg\max_{a_{t+1}} Q(s,a_{t+1}))))^2$$

이렇게 Q 값을 곱해주게 되면 Q 의 값이 낮아도 기존의 Q-Learning 보다 더 좋은 결과를 나타낸다.

2.2. Duealing Q-Learning

2015 년 Wang(2015)에 의해 발표된 Dueling network architectures for deep reinforcement learning 의 논문에 따르면 Q-Learning 을 강화학습에 적합하게 바꾸기 위해 연구하였다.

Q-Learning 을 학습 시킬 때 중요한 요인 중 하나가 미래의 Q 값을 예측 하는 것인데 미래의 Q 값을 예측하기는 쉽지 않다 그렇기 때문에 Q 값을 예측하기보다 기준점 V(s)를 설정해 상대적인 Q 값의 차이 A(s,a)를 설정해 Q 값을 아래 식(2)와 같이 구한다.

식(2)

$$Q(s_t,a_t) = V(s) + A(s,a)$$

2.3. Modified Q-Learning

Kim(2008)교수의 논문에 따르면 Q-Learning 은 불연속 상태 공간과 행위 공간을 정의하고 현재 상태로부터 목표 상태에 도달하기 위한 최적의 행위 집합을 구하기 위한 통계적인 해결책으로 제안되었다. 그러나, 연속적인 상태공간과 행위공간을 내포하는 실질적인 환경에 Q-Learning 을 적용하기 위해서는 너무 많은 양의 기억 공간과 학습시간이 필요하게 되기 때문에

Modified Q-Learning 알고리즘을 제안 하였다.

Modified Q-Learning 알고리즘

[초기화]
1. 초기화:여러 파라메타(γ , α , ρ)의 초기화
[반복]
2. 현재상태를받아들임(s 현재상태)
3. 상태 s 와주변상태와의관계를식 (8)를이용하여구한다.
4. 상태 s 에서실행해야할행위는식 (9)에의해구한다. (때로는 Random action 수행)
5. 결정된행위를수행하고, Reward 를받는다.
6. 주변상태들에서의 Q 값은식 (8)를사용하여구한다.
7. 주변상태들에서의최적의행위및 Q 값을식 (8),(9)을사용하여갱신한다.

2.4. Loss Function

손실 함수는 딥러닝에서 사용하는 알고리즘 성능의 “나쁨”을 나타내는 지표로 현재의 신경망이 훈련 데이터를 얼마나 잘 처리하지 ‘못’ 하느냐를 나타내는 함수이다. 그러므로 손실함수 값은 작을수록 정확하게 이루어 진다고 볼 수 있다.

손실함수는 예측의 정도를 평가하는 척도로 평균제곱오차 (MSE), 제곱근 평균제곱오차 (RMSE), 평균절대퍼센트오차(MAPE) 그리고 평균절대오차 (MAE)등이 있으며 강화학습 알고리즘에서는 대부분 평균제곱오차(MSE)를 흔히 사용하며 실험에 사용한 Q-learning 알고리즘 또한 손실함수로 평균제곱오차(MSE)를 사용하고 있다.

2.4.1. Mean Squared Error

평균제곱오차(Mean squared error)는 다양한 딥러닝 알고리즘에 자주 사용 되는 손실함수이다.

평균제곱오차를 간단히 설명하자면 오차(데이터의 정답과 머신러닝 수행 한 뒤에 나오는 출력값과의 차이)의 제곱에 대해 평균을 취하는 것이다.

MSE 식은 아래와 같다.

식(5)

$$E=\frac{1}{2}\sum_{k}(y_k-t_k)^2$$

(y:신경망 출력, t: 정답 레이블,k:차원 수)

MSE 의 값이 작을수록 본래 정답과의 오차가 적게 되는 것으로 추측한 값의 정확성이 높다고 할 수 있다. MSE 는 직관적이며 계산하기 쉬워 많이 사용한 알고리즘이나 훈련 속도가 느리고 다른 loss function 들보다 성능이 조금 떨어져 현재 딥러닝에서는 많이 사용하지 않은 loss function 이다.

2.4.2. Cross Entropy Error

교차 엔트로피(CEE)는 요즘 대부분의 딥러닝 알고리즘의 손실 함수로 사용 되고 있으며 음수 로그 유사도 라고도 한다. CEE 는 쉽게 말해서 2 개의 확률 분포 사이에 정의되는 척도이다.

부호화 방식이 "진정한" 확률분포 p 가 아니라 어떤 소정의 확률 분포 q 에 근거할 경우, 여러 사건에서 한가지 현상을 특정하기 위해 필요한 데이터 수의 평균치이다.

CEE 의 식은 아래와 같다.

식(6)

$$E = -\sum_{k} t_k \log y_k$$ [7]

(y:신경망 출력 데이터, t: 정답 레이블)

딥러닝에서는 보통 한 번에 mini-batch 로 여러 데이터를 묶어 한 번에 처리하기 때문에, 마지막에도 batch size 에 해당하는 개수의 교차 엔트로피 값들이 나오게 된다. 훈련에 사용하기 위해서는 이 값들의 평균을 구해 해당 batch 의 손실 값으로 설정한다.

3. Experiments

3.1. Experimental Environment

Table 1. Experimental Environment

운영체제	Microsoft Windows 10
개발환경	Jupyter Notebook*
개발언어	Python 3.5**
실험데이터	Gym CartPole-V1***
개발라이브러리	TensorFlow****

다음 Table 1 은 실험에서 사용한 실험환경이다.

* Jupyter Notebook 은 오픈 소스 웹 애플리케이션으로 라이브 코드, 등식, 시각화와 설명을 위한 텍스트 등을 포함한 문서를 만들고 공유하도록 한다.

** 파이썬(Python)은 1991 년 프로그래머인 귀도 반 로섬(Guido van Rossum)이 발표한 고급 프로그래밍 언어로, 플랫폼 독립적이며 인터프리터식, 객체지향적, 동적 타이핑(dynamically typed) 대화형 언어이다.

*** OpenAI 사에서 강화학습을 위해 제공하는 Data 의 Carpole-V1 을 사용

**** 텐서플로(TensorFlow)는 구글 제품에 사용되는 머신러닝(기계학습)을 위한 오픈소스 소프트웨어 라이브러리이다.

3.2. Experimental Data

본 논문에서 사용한 실험 데이터는 OpenAI 사에서 강화학습을 할 수 있게 제공하는 gym 라이브러리에 포함된 “CartPole-v0“ 사용 하였다. CartPole-v0 의 목표는 최대한 오랫동안 막대기를 세우고 있는 것이다. CartPole-v0 는 Figure 2 과 같이 구성 되어 있다.

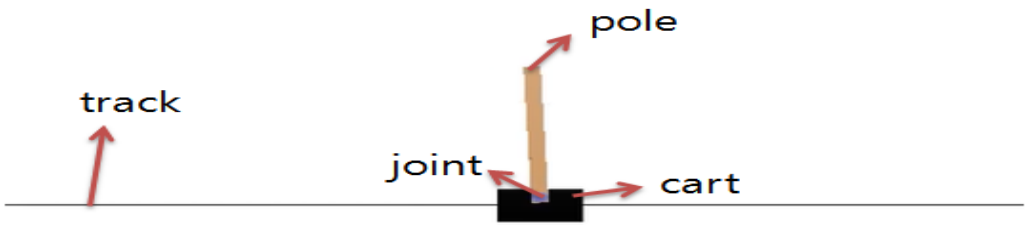


Figure 2. Cart-Pole-v0

cart 는 track 위에서 마찰 없이 움직이고 cart 는 joint 로 pole 과 연결되어 있다. pole 은 joint 를 중심으로 자유롭게 회전이 가능하며 중력이 작용하여 아래로 떨어지게 된다. 따라서 agent 는 pole 이 떨어지지 않게 좌,우로 움직여 pole 의 중심을 잡아주어야 한다. episode 는 pole 이 수직에서 15 도 떨어지거나 가운데로부터 2.4units 만큼 떨어지게 되면 끝나게 되어있다. reward 는 pole 이 세워져 있는 시간 만큼증가하게 된다. 따라서 reward 가 클수록 CartPole-v0 의 목표에 가까워 진다고 할 수 있다.

3.2. Experimental Result

실험결과는 다음과 같다. 먼저 기존에 Q-learning 알고리즘에 MSE(Mean Squared Error)을 적용한 결과는 다음 Figure 3 과 같다.

```
Mean Reward: 11.8 Total Steps: 105632 p: 0.7496605000020657
Mean Reward: 11.31 Total Steps: 106763 p: 0.7445710000021077
Mean Reward: 12.24 Total Steps: 107987 p: 0.7390630000021532
Mean Reward: 12.06 Total Steps: 109193 p: 0.733636000002198
Mean Reward: 11.78 Total Steps: 110371 p: 0.7283350000022417
Mean Reward: 12.26 Total Steps: 111597 p: 0.7228180000022872
Mean Reward: 11.84 Total Steps: 112781 p: 0.7174900000023312
Mean Reward: 12.61 Total Steps: 114042 p: 0.711815500002378
Mean Reward: 11.89 Total Steps: 115231 p: 0.7064650000024222
Mean Reward: 12.13 Total Steps: 116444 p: 0.7010065000024672
Mean Reward: 13.25 Total Steps: 117769 p: 0.6950440000025164
Mean Reward: 14.22 Total Steps: 119191 p: 0.6886450000025692
Mean Reward: 13.5 Total Steps: 120541 p: 0.6825700000026194
Mean Reward: 14.13 Total Steps: 121954 p: 0.6762115000026718
Mean Reward: 14.42 Total Steps: 123396 p: 0.6697225000027254
Mean Reward: 15.4 Total Steps: 124936 p: 0.6627925000027826
Mean Reward: 14.14 Total Steps: 126350 p: 0.6564295000028351
Mean Reward: 14.01 Total Steps: 127751 p: 0.6501250000028871
Percent of succesful episodes: 12.9397%
```

Figure 3. Success Rate of Mean Squared Error

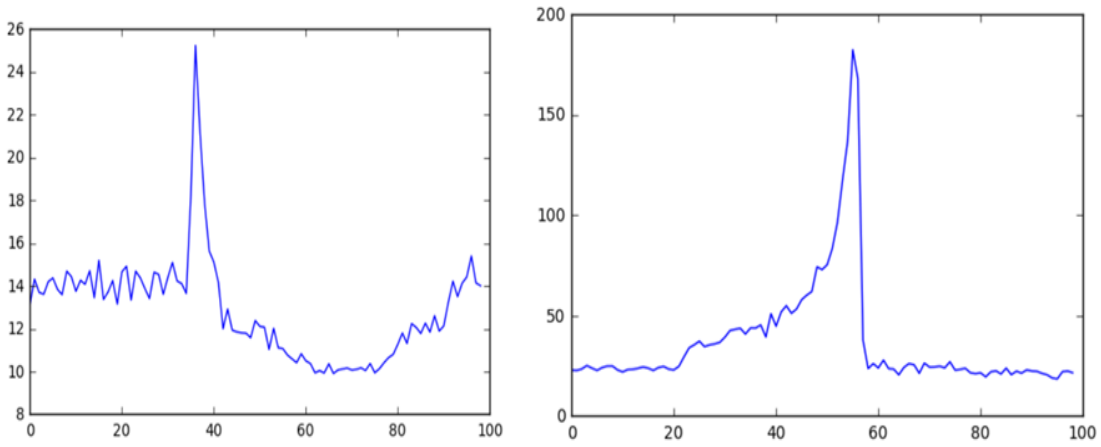
Figure 3 에서 알 수 있듯이 100 번의 학습을 통한 성공률은 약 13%의 수치를 얻었다. 아래 Figure 4 는 Q-learning 에 손실함수 교차엔트로피(Cross Entropy Error)를 적용한 성공률이다.

```
Mean Reward:  21.55 Total Steps:  327002 p:  0.09999550000334036
Mean Reward:  19.56 Total Steps:  328958 p:  0.09999550000334036
Mean Reward:  22.07 Total Steps:  331165 p:  0.09999550000334036
Mean Reward:  22.54 Total Steps:  333419 p:  0.09999550000334036
Mean Reward:  21.0 Total Steps:  335519 p:  0.09999550000334036
Mean Reward:  23.88 Total Steps:  337907 p:  0.09999550000334036
Mean Reward:  20.65 Total Steps:  339972 p:  0.09999550000334036
Mean Reward:  22.37 Total Steps:  342209 p:  0.09999550000334036
Mean Reward:  21.41 Total Steps:  344350 p:  0.09999550000334036
Mean Reward:  22.99 Total Steps:  346649 p:  0.09999550000334036
Mean Reward:  22.46 Total Steps:  348895 p:  0.09999550000334036
Mean Reward:  22.3 Total Steps:  351125 p:  0.09999550000334036
Mean Reward:  21.3 Total Steps:  353255 p:  0.09999550000334036
Mean Reward:  20.57 Total Steps:  355312 p:  0.09999550000334036
Mean Reward:  18.98 Total Steps:  357210 p:  0.09999550000334036
Mean Reward:  18.52 Total Steps:  359062 p:  0.09999550000334036
Mean Reward:  22.13 Total Steps:  361275 p:  0.09999550000334036
Mean Reward:  22.44 Total Steps:  363519 p:  0.09999550000334036
Mean Reward:  21.62 Total Steps:  365681 p:  0.09999550000334036
Percent of succesful episodes:  36.741%
```

Figure 4. Success Rate of Cross Entropy Error

Figure 4 에서의 학습 역시 100 번의 학습을 통해 통제변수를 같이 하였고 성공률은 약 36%로 기존의 MSE 보다 약 3 배 정도의 높은 성공률을 얻을 수 있었다.

Figure 5 에서 볼 수 있듯이 같은 학습횟수이지만 보상측면에서 CEE(CrossEntropyError)가 약 3 배 정도 높은 보상을 받는 다는 것을 알 수 있다.



(x=Learning times, y=Reward)

Figure 5. MSE(Left) CEE(Right)

3. Conclusion

본 연구에서는 온라인상에서 증명서를 발급할 수 있는 통합시스템을 구축하여 one-stop 으로 보험료를 받을 수 있는 시스템을 구축하였다. 구축 된 시스템은 온라인상에서 안전하고 편리한 증명서 발급 서비스 제공하고 병원에서 증명서를 발급받아 보험회사로 접수하는 번거롭고 어려운 보험료 청구를 국가공인 메일인 #-메일을 사용하여 온라인상에서 보험료 청구까지 한번에 처리할 수 있었다. 이렇게 개발된 솔루션은 증명서 발급비용 절감효과와 paperless 차원의 종이문서 발생을 억제하고 인증절차 간소화 및 원무행정을 간소화시킬 것이다. 나아가 민영의료보험사로부터 복잡한 절차 없이 보험금을 지급 받음으로써 민영의료보험 혜택의 사각지대를 해소하고 사회자본을 절약할 수 있을 것으로 기대된다.

References

Chang, J. Y. (2013). Automatic retrieval of SNS opinion document using machine learning technique. *The Journal of The Institute of Internet, Broadcasting and Communication, 13*(5), 27-35.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21,* 768-769.

Hartigan, J. A., & Hartigan, J. A. (1975). *Clustering algorithms* (Vol. 209). New York: Wiley.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics*), 28(1), 100-108.

Kim, W., & Kim, S. (2014). Document Clustering Technique by K-means Algorithm and PCA. *Journal of the Korea Institute of Information and Communication Engineering, 18*(3), 625-630.

Kim, Y. J. (2008). Modified Q-Learning for Intelligent System. *The Journal of the KICS*, 33(2), 82-87. Retrieved from http://kics.or.kr

Lee, G. S., & Kim, I. K. (2016). A Study on Simplification of Machine Learning Model. *JIIBC, 16*(4), 147-152.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory, 28(*2), 129-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1, 801- 804

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. In AAAI Conference on Artificial Intelligence(pp.2094-2100). NIPS

Wang, Z. (2016). Dueling network architectures for deep reinforcement learning. Retrieved from https://arxiv.org/pdf/1511.06581.pdf

Wikipedia (2017a). Retrieved from https://en.wikipedia.org/wiki/DBSCAN

Wikipedia (2017b). Retrieved from https://en.wikipedia.org/wiki/K-means_clustering