빅데이터마이닝:머신러닝 과제 2
분류와 군집화의 예시

데이터테크전공

20173204 곽명빈

지도학습 / 비지도 학습 종류

지도학습(Supervised Learning),	Classification	kNN
		Naive Bayes
		Support Vector
		Machine Decision
	Regression	Linear Regression
		Locally Weighted Linear
		Ridge
		Lasso
비지도학습(Unsupervised Learning),		Clustering
		K Means
		Density Estimation
		Exception Maximization
		Pazen Window
		DBSCAN

출처:https://ebbnflow.tistory.com/165

분류와 군집화에 대해 2가지 논문을 사례로 들어 요약 ,정리하려 한다.

분류에서는 KNN(최근접 이웃) 알고리즘을 이용하여 고속도로 통행시간 예측을 알아본다.

군집화는 K Means를 이용하여 개인별 음원 추천 모형에 관한 연구에 대해 알아본다.

분류

KNN 알고리즘을 활용한 고속도로 통행시간 예측

기존의 문제점

보지원시스템을 운영하고 있다. 그러나 교통예보지원시스템에서는 시계열 모형인 지수평활화와 ARIMA (Autoregressive integrated moving average)를 이용하기 때문에 실시간 자료를 반영하지 못하고 있으며, 그 결과 통행패턴이 과거와 상이할 경우 정확도가 떨어지는 문제점이 발생하고 있다.

관련 연구

Smith and Demetsky (1997), Lam et al. (2006), Lim and Lee (2011)는 교통량 예측을 위해 KNN 방법을 이용하였으며, You and Kim (2000)은 KNN 방법에 GIS (Geographic Information Systems)와 Machine Learning 기법을 적용하여 시가지도로의 통행시간을 예측하였다. Park et al. (2006)와 Lim et al. (2013)은 KNN 방법을 이용하여 버스의 통행시간과 일반국도의 통행시간을 예측하였다.

실시간 교통정보는 시공간적 혼잡상황의 변화에 따라 이용 시점에 제공한 통행시간과 실제 경험한 통행시간의 차이가 발생 → 예측정보 제공에 대한 필요성이 증대

분류

KNN 알고리즘을 활용한 고속도로 통행시간 예측

시사점

통행시간 예측과 관련하여 다양한 연구가 진행되었으나 신경망의 경우 학습 과정이 매우 복잡하다는 단점이 있으며, 회귀모형은 다수의 모형(예: 기/종점, 경로 등)을 개발하여야 한다는 문제점이 있다. 반면 KNN의 경우 이러한 문제점을 해결할 수 있으며, 참조할 수 있는 데이터가 충분하다면 타 모형에 비해 정확도가 우수한 장점이 있다(Smith et al., 2002; Nigovski et al., 2005). 또한 한국도로공사는 6년 정도의 고속도로 이력자료를 OASIS에서 보유하고 있어 KNN 방법을 적용하는데 적합한 환경을 보유하고 있다.

KNN

KNN 모형은 실시간으로 수집되는 입력변수의 상태와 가장 유사한 과거 이력자료 내 동일 변수의 상태를 실시간으로 비교하여 유사한 이웃(근접이웃)의 경로통행시간을 가중평균하여 경로통행시간을 추정하는 구조를 가지고 있다. 이러한 모형의 구조를 고려하

K개의 근접이웃을 통해 2개(TCS 교통량, DSRC 구간통행시간)의 T주기 경로통행시간이 산출되므로 본 연구에서는 이들을 가중평균하여 최종 경로통행시간의 예측치를 추정하였다. 이 때 TCS

6년간의 고속도로 이력자료보유 – KNN에 적합
실시간 교통상황을 반영할 수 있는 TCS 교통량과 DSRC 링크통행 시간을 활용

분류

KNN 알고리즘을 활용한 고속도로 통행시간 예측

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y_i})^2}{n}}$$

추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용하는 측도

KNN기반 예측

$$\widehat{P_i} = w \times \frac{\sum_{k=1}^{K} \left(\frac{1}{qS_i^k}\right) p_k^q}{\sum_{k=1}^{K} \frac{1}{qS_i^k}} + (1-w) \times \frac{\sum_{k=1}^{K} \left(\frac{1}{cS_i^k}\right) p_k^c}{\sum_{k=1}^{K} \frac{1}{cS_i^k}} \qquad (3)$$

여기서, $\widehat{P_i}$: i번째 주기의 경로통행시간 예측치

w : 경로통행시간 예측치에 대한 TCS 교통량 가중치

qS_i^k : TCS 교통량 k 근접이웃의 유사성

p_k^q : TCS 교통량 기반 k 근접이웃의 경로통행시간

cS_i^k : DSRC 구간통행시간 k 근접이웃의 유사성

p_k^c : DSRC 구간통행시간 k 근접이웃의 경로통행시간

k : 근접이웃의 수($k = 1,2,...,K$)

예측결과

Table 3. Result by Applied KNN Method on Seoul TG ~ Daejeon IC

Day	RMSE (Min)		
	w=1.0	w=0.8	w=0
Mon	5.7	5.7	4.1
The, Wed, Thu	5.6	5.5	4.8
Fri	4.2	3.8	3.1
Sat	8.9	11.4	6.9
Sun	6.2	6.0	4.2
Average	6.1	6.5	4.6

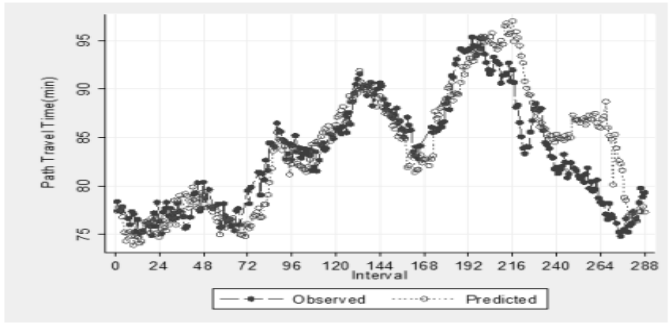


Fig. 3. Predicted Result on Seoul TG ~ Daejeon IC (Jan 14th 2011)

분류

KNN 알고리즘을 활용한 고속도로 통행시간 예측

결론

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용하는 측도

KNN기반 예측

$$\hat{P}_i = w \times \frac{\sum_{k=1}^{K} (\frac{1}{qS_i^k}) p_k^q}{\sum_{k=1}^{K} \frac{1}{qS_i^k}} + (1-w) \times \frac{\sum_{k=1}^{K} (\frac{1}{cS_i^k}) p_k^c}{\sum_{k=1}^{K} \frac{1}{cS_i^k}} \quad (3)$$

여기서, $\hat{P}_i$: i번째 주기의 경로통행시간 예측치

w : 경로통행시간 예측치에 대한 TCS 교통량 가중치

qS_i^k : TCS 교통량 k 근접이웃의 유사성

p_k^q : TCS 교통량 기반 k 근접이웃의 경로통행시간

cS_i^k : DSRC 구간통행시간 k 근접이웃의 유사성

p_k^c : DSRC 구간통행시간 k 근접이웃의 경로통행시간

k : 근접이웃의 수($k = 1,2,...,K$)

- KNN기반 예측 모형이 기존 시계열모형의 예측 오차에 비해 작음
- 시계열모형은 실시간 상황을 제대로 반영하지 못하기 때문에 향후 데이터가 축적될 경우 예측 오차는 감소
- 현재자료와 유사한 과거 이력자료를 통해 분류하기 때문에 오차가 작다고 할 수 있음

예측결과

Table 3. Result by Applied KNN Method on Seoul TG ~ Daejeon IC

Day	RMSE (Min)		
	w=1.0	w=0.8	w=0
Mon	5.7	5.7	4.1
Tue/Wed/Thu	5.6	5.5	4.8
Fri	4.2	3.8	3.1
Sat	8.9	11.4	6.9
Sun	6.2	6.0	4.2
Average	6.1	6.5	4.6

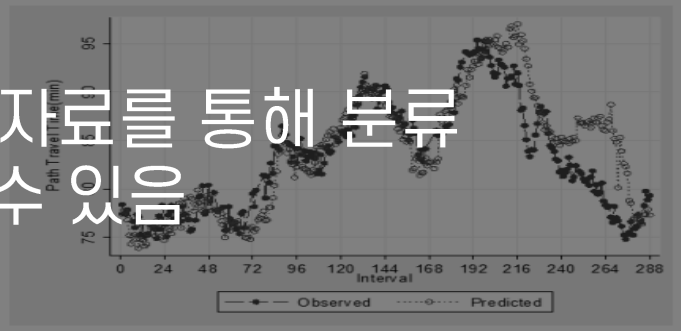


Fig. 3. Predicted Result on Seoul TG ~ Daejeon IC (Jan 14th 2011)

군집화

K-Means을 활용한 개인별 음원 추천 모형에 관한 연구

연구 목적

이러한 방안은 사회적 문제가 되고 있는 음원 사재기, 순위조작 등 병폐를 개선하기 위한 새로운 음원 추천방식을 위한 모형개발과 클러스터링 분석을 활용으로 보다 개인별 취향 및 기호에 적합한 음원 추천 및 관리 서비스의 제시를 그 목적으로 하고 있으며, 아래 그림2와 같이 개인이 소장하고 있는 음원을 클러스터링 분석을 통해 유사 음원을 군집화하여 개인 소유 음원을 자동으로 관리를 하고 가장 많은 음원이 속한 군집의 특징을 기반으로 신규로 출시되는 음원을 추천한다.

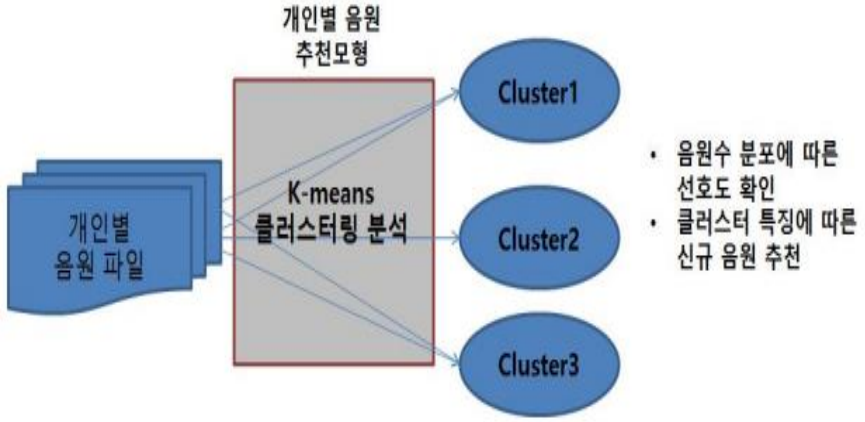


그림 2. 개인별 음원 추천 모형의 활용 구성

→ 개인별 취향에 맞는 음원 추천 및 관리 서비스 방안 모델

군집화

K-Means을 활용한 개인별 음원 추천 모형에 관한 연구

전처리

본 연구에서는 음원을 데이터마이닝 기법인 클러스터링 분석을 수행하기 위하여 우선적으로 음원을 클러스터링 분석이 가능하도록 분석이 가능한 데이터로의 변환이 필요하다. 이를 위하여 본 연구에서는 푸리에 변환(Fourier transform)을 적용하였다.

$$u(t)=\frac{1}{2\pi}\int_{-\omega}^{\omega}U(\omega)e^{j\omega t},\ U(\omega)=\frac{1}{2\pi}\int_{-\omega}^{\omega}u(t)e^{-j\omega t}dt,\ \omega=2\pi f$$

U(t) : 시간함수
U(w) : 주파수 함수

그림 9. 푸리에 변환(Fourier transform)

K-Means

본 연구에서 개인별 음원 추천 모형을 수립하기 위하여 사용한 K-means 분석 알고리즘에서 가장 최적의 K값(클러스터링 수)를 찾기 위한 방법으로 Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation (Ray and Turi, 1999)을 이용하였다.

N은 측정대상 음원 수의 좌표, K는 클러스터의 개수이고, Z_i는 클러스터의 Ci 중심의 좌표이다.

$$\text{intra} = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}\|x - z_i\|^2$$

$$\text{inter} = \min(\|z_i - z_j\|^2),\ i = 1,2,\ldots,K-1\quad j = i+1,\ldots,K$$

그림 12. 클러스터링 내부거리(intra), 클러서터링간 거리(inter) 관계식

군집화

K-Means을 활용한 개인별 음원 추천 모형에 관한 연구

K-Means

표 8. 모형 검증을 위한 개인음원의 임의의 K에 따른 Validity 값

K	Validity	클러스터 구분
2	0.34881277	1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1
3	0.401186657	1,1,1,1,1,0,1,1,1,2,1,1,1,1,1,1,1,1,1,1,1
4	0.354002265	1,1,1,1,1,0,1,1,1,2,1,1,1,1,1,1,1,1,1,1,3
5	0.475570292	1,1,1,1,1,0,1,1,1,2,1,1,1,1,1,1,1,4,1,1,3
6	0.596728179	1,1,1,1,1,0,1,5,5,2,1,1,1,1,1,1,1,4,1,1,3
7	0.507328243	3,1,1,1,1,6,1,0,0,2,1,1,1,1,1,1,1,1,5,1,4
8	0.446800037	5,0,0,0,0,6,0,1,1,4,0,0,0,0,0,0,0,0,0,7,3,0,2
9	0.375625632	5,7,0,0,0,4,0,3,3,1,0,0,0,0,0,0,0,0,0,8,6,0,2

* $validity = \frac{intra}{inter}$

K-Means 결과

표 9. 개인음원의 클러스터 구성과 음원의 개인선호도 비교

Cluster	음원 제목	개인 선호도
0	어차피 잘 될놈	13
1	바람이나좀쐐2002	4
	바람이나 좀 쐐(Feat.MIWOO)	11
2	**너에게 배운다**	**1**
	주마등	**2**
	엉덩이	**5**
	또 다시 사랑	**6**
	눈물	**7**
	목소리	**8**
	Upton Funk	**9**
	Lion Heart	**14**
	Shake it	**15**
	우리 사랑하지 말아요	**17**
	I'm Not The Only One	**10**
	사랑했다치자	**21**
3	잘 나가서 그래	18
4	뱅뱅뱅	12
5	댄싱게놈	22
6	이유 갖지 않은 이유	19
7	쩔어	16
8	I Feel You	20
9	나란 놈은 답은 너다	3

표9와 개인이 선호하는 음원의 순위와 유사성이 높음

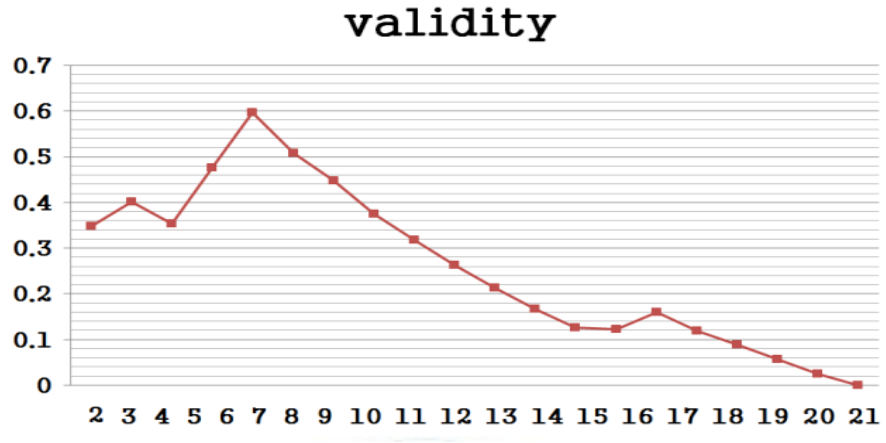
군집화

K-Means을 활용한 개인별 음원 추천 모형에 관한 연구

결론

본 연구에서는 음원의 클러스터링 분석을 활용하여 연구를 수행한 결과로 K-means 클러스터링 음원의 추천 및 관리하는 방식이 개인의 음원 선호도를 파악하는데 활용이 가능하다는 결론을 얻었다.

이를 통하여 개인이 보유한 음원을 활용하여 개인별 선호하는 음원를 파악할 수 있다고 판단되며, 또한 개인이 보유하고 있는 음원을 분석하여 개인 취향에 부합하는 음원을 추천하는데 있어서 시간적 경제적 비용을 줄이는데 기여할 수 있다.



K-Means를 통해 이상적인 클러스터를 찾아 개인이 보유한 음원을 군집화 하여 선호도를 파악하여 유의성이 있음을 검증

결론

느낀점

분류와 군집화를 2가지 논문을 바탕으로 살펴 보았다.

- KNN 알고리즘을 활용한 고속도로 통행시간 예측에서 분류(KNN)을 사용한 이유는 과거 데이터가 있기 때문에 과거 데이터와 유사한 데이터를 분류하기 때문에 지도학습 중 하나인 KNN을 사용했다고 생각한다.

- K-means 클러스터링 마이닝기법을 활용한 개인별 음원 추천 모형에 관한 연구에서 군집화(K-Means)를 사용한 이유는 음원간 Label이 존재하지 않아 분류를 할 수 없기 때문에 비지도학습 중 하나인 K-Means를 사용했다고 생각한다.

- 머신러닝을 할 때, 지도학습인지, 비지도학습인지 판단하는 능력이 중요할 것 이라고 느꼈으며, 다양한 분류, 회귀, 군집화 과정을 통해 최적의 예측을 하는 모델을 찾아 적합 시키는 것이 중요할 것 같다는 생각을 하였다.

출처

이창해. "K-means 클러스터링 마이닝기법을 활용한 개인별 음원 추천 모형에 관한 연구." 국내석사학위논문 연세대학교 공학대학원, 2016. 서울

Shin, Kangwon, Shim, Sangwoo, Choi, Keechoo, & 김수희. (2014). KNN 알고리즘을 활용한 고속도로 통행시간 예측. 대한토목학회논문집, 34(6), 1873–1879. https://doi.org/10.12652/KSCE.2014.34.6.1873