2023 전기 졸업과제 중간 보고서

분과 : 데이터 / SW플랫폼(B)

과제명 : HMM기반Trajectory 예측 시스템

팀명 : 산순이와 정둘이

팀원 : 이성무(201812145), 하연지(201927547), 정재원(201724574)

지도교수 : 권준호 교수님

목차

1. 요구조건 및 제약 사항 분석에 대한 수정사항

1.1 요구조건

- 은닉 마코프 모델 구축에 이용할 "[교통, 안전] 월별 렌터카 위치정보" 의 이상치와 결측치를 처리한 후 분석에 필요한 파생변수(경로, 시간대)를 생성한다.
- "[교통, 안전] 월별 렌터카 체류 빈도" 데이터로 K-Means 클러스터링 방법을 이용해 20개, 35개, 50개와 같이 군집의 개수를 조정하며 관심지점(POI)을 선정한다.

1.2 제약사항

- HMM 학습에 이용하는 데이터는 렌터카 아이디, 위도, 경도, 해당 위치 포인트가 기록된 시간 정보만 있어 개인적 특성을 반영하기 어렵기 때문에 시간 정보를 오전, 오후, 저녁으로 구분하기로 하였으나, 예측 성능을 높이기 위하여 새벽, 오전, 오후, 심야 네 가지로 구분하여 적용한다.
- POI 개수에 따라 계산 시간 복잡도와 예측 정확도에 차이가 있기에 개수를 달리하여 성능을 평가하여 설정한다.
- POI 선정 후 사용자의 GPS 데이터에서 경로 데이터를 도출할 때 기존에는 POI을 기준으로 특정 반경 내에 위치할 경우 해당 POI에 방문한 것으로 보았으나 가장 가까운 POI을 계산하여 해당 지점을 방문한 것으로 처리한다.

2. 설계 상세화 및 변경 내역

본 과제의 Road Map은 Python과 C++을 이용하여 Colab, VSCode 및 MySQL을 통해 수집한 데이터를 처리하여 은닉 마코프 모델을 학습한 후 사용자의 다음 이동 경로를 예측하는 <그림 1>과 같다.

<그림 > 모델 설계도

2.1 데이터 전처리

제주 데이터 허브에서 수집한 "[교통, 안전] 월별 렌터카 위치정보" 를 분석에 이용하였다. 본 과제에서는 2020년, 2021년 데이터를 분석에 이용하고자 하므로 2년치의 데이터를 월별로 통합한다. 해당 데이터는 렌터카 ID, 데이터 수집 일자 및 시간, 위도, 경도 정보를 담고 있다. 이들 중 하나의 정보라도 결측이 발생했을 경우 해당 데이터는 삭제하였다. 제주 데이터 허브에서 제공하는 월별 렌터카 위치 정보는 같은 달 정보끼리 묶여 있다. 그러나 제목의 월 정보와 데이터 수집 일자가 일치하지 않는 경우가 있어 해당 데이터를 삭제하였다. 또한 제주도의 렌터카 위치 정보가 아닌 곳의 데이터가 포함되어 있어 제주도의 왼쪽 끝, 오른쪽 끝 경도에서 위쪽 끝, 아래쪽 끝 위도를 벗어나는 경우 삭제한다. 이상치와 결측치를 삭제한 GPS 데이터 포인트마다 가장 가까운 관심지점을 계산하여 매치한다. 그 결과 각 일자마다 한 사용자의 이동 경로 데이터가 도출된다. 각 경로 포인트마다 수집 시간 정보를 토대로 4개의 시간대에 매칭한다. 데이터 수집 일자, 렌터카 ID 를 tuple로 만들고 PRIMARY KEY 로 하여 경로의 시작점, 도착점, 경로와 시간대 시퀀스를 MySQL 데이터베이스에 저장한다.

2.2 관심 지점 선정

"[교통, 안전] 월별 렌터카 체류 빈도" 는 수집시간 전로그와 후로그의 시각차이가 20분 이상인 경우 전로그의 위치를 나타낸 데이터이다. 해당 데이터에서 위치는 제주도를 50m 격자 상에 올렸을 때의 좌표값으로 나타나며, "[교통, 안전] 월별 렌터카 위치 정보" 데이터와 위치계를 통일하기 위하여 EPSG:5179 좌표계에서 EPSG:4326으로 변환한다. 사용자들의

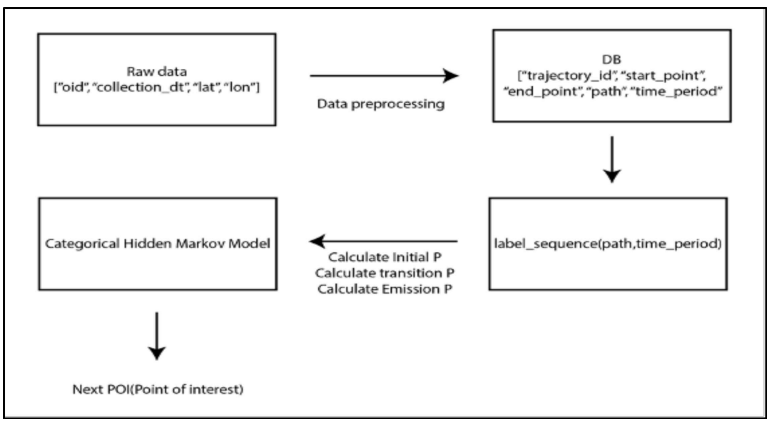
이동 궤적을 K-Means 클러스터링을 통해 주요 장소를 추출한 연구[1]를참고하여 해당 방법론을 통해 관심지점을 선정한다.

2.3 은닉 마코프 모델 구성

사용자의 과거 위치로부터 다음 이동을 예측하기 위해 경로 포인트가 수집된 시간대를 각 장소마다 각각 인덱싱하여 관측값으로 두고, POI ID를 은닉 상태로 하는 은닉 마코프 모델을 분석에 이용한 연구[2]가 있다. 예를 들어, 각 POI의 시간대 새벽, 오전, 오후, 심야를 각각 0, 1, 2, 3에 대응한다. 본 과제에서는 4개의 시간대를 사용하므로, 각 POI ID * 4 + 시간대 인덱스가 관측열을 구성하는 관측값이 된다. MySQL에서 저장한 데이터를 불러와 각 데이터마다 시간대 시퀀스를 인덱싱한다.

은닉 마코프 모델은 한 은닉 상태에서 다른 은닉 상태로의 전이한 누적 횟수를 모든 은닉 상태에서 다른 은닉 상태로 전이한 총 횟수를 나눈 전이 확률과, 특정 상태에서 어떤 관측값이 발생하는 누적 횟수를 모든 상태에서 어떤 관측값이 발생하는 총 횟수로 나눈 방출 확률을 분석에 이용한다. 이때, 은닉 상태를 POI ID로 설정하였으므로 이를 반영하기 위하여 MySQL에서 데이터를 불러와 각각의 확률을 계산한 뒤에 은닉 마코프 생성 시 파라미터로 설정한다.

이전 단계에서 도출한 확률 행렬을 파라미터로 은닉 마코프 모델을 생성한 후 사용자의 현재까지의 위도, 경도, 데이터 수집 일자 및 시간 정보의 전처리 과정을 수행하고 사용자 데이터로 관측열을 생성한다. 수집한 사용자의 위치 정보를 토대로 다음 경로를 예측하기 위해 사용자의 데이터로 관측열을 생성한 뒤 모든 POI ID를 하나씩 뒤에 붙인 시퀀스들의



<그림 > 모델2의 구성도

---

[1] 김용중. "은닉 마르코프 모델을 이용한 스마트폰 사용자의 경로 모델링 및 온라인 학습 기반 목적지 예측." 국내석사학위논문 연세대학교 대학원, 2014. 서울

[2] Pant, N., & Elmasri, R. (2017, April). “Detecting meaningful places and predicting locations using varied k-means and hidden Markov model.” *In Proceedings of the 17th SIAM International Conference on Data Mining (SDM 2017),* Houston, TX, USA (pp. 27-29).

우도 확률을 계산 후 가장 높은 우도 확률을 가지는 관측값을 도출한다. 그 결과 얻은 관측열을 은닉 마코프 모델에 적용하여 관측열로부터 은닉 상태를 추론하는 디코딩 과정을 거쳐 은닉 상태 시퀀스를 얻어 다음 후보지를 예측한다. 이때 가장 마지막에 오는 은닉 상태는 사용자가 다음 번에 방문할 것으로 추론되는 POI의 ID 이다.

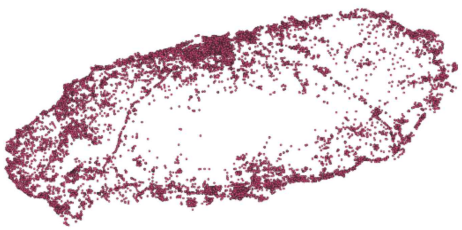
3. 갱신된 과제 추진 계획

3.1 모델의 예측 정확도 개선과 사용자별 특성 반영을 위한 추가 데이터 수집

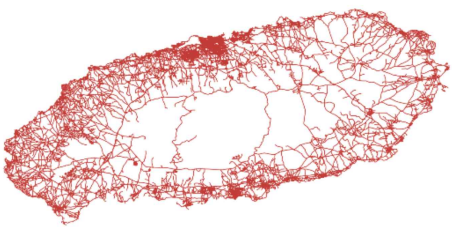
Kim(2013)에 따르면 사용자의 다음 경로 예측 시 이용 가능한 문맥 정보들을 모두 사용했을 때 예측 성능이 가장 뛰어난 것으로 나타났다. 현재까지는 사용자의 위치 데이터가 수집된 시간대 정보만을 활용하였기 때문에 추가적인 데이터 수집이 필요하다. 본 과제에 적용하였을 때 모델 예측 성능을 개선할 수 있는 지 여부를 확인할 필요가 있다. 뿐만 아니라 추가적인 데이터 수집을 통해 예측 모델에 적용함으로써 사용자의 여러 특성 정보까지 반영할 수 있어 사용자의 높은 만족도를 유도할 수 있을 것으로 기대된다.

3.2 패스트 맵 매칭을 활용한 풀스택 환경 개발

패스트 맵 매칭(Fast Map Matching)은 Python과 C++ 로 구성된 오픈 소스 맵 매칭 프레임워크이다. 패스트 맵 매칭의 경우 제주도의 Geojson 파일을 가지고 제주도를 노드와 엣지로 연결하여 위치 정보가 입력되면 그 지점에서의 노드와 엣지와의 거리를 분석하여 맵 매칭을 실행한다. 이를 활용하여 프론트 엔드에서 실시간으로 이동하는 이용자의 경로를 백엔드로 보내어 구성한 은닉 마코프 모델로 다음 경로를 예측하고 예측된 지점의 위치정보를 맵 매칭하여 이동하는 경로에 따라 움직이는 이용자의 모습을 가시화한다.



<그림 > 제주도 nodes



<그림 > 제주도 edges

<table>
<tr><th colspan="3">5월</th><th colspan="5">6월</th><th colspan="5">7월</th><th colspan="5">8월</th><th colspan="4">9월</th></tr>
<tr><td>3</td><td>4</td><td>5</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td colspan="5">데이터 전처리</td><td colspan="18"></td></tr>
<tr><td colspan="6"></td><td colspan="5">모델 구성</td><td colspan="12"></td></tr>
<tr><td colspan="10"></td><td colspan="5">모델 최적화</td><td colspan="7"></td></tr>
<tr><td colspan="11"></td><td colspan="3">중간
보고서
작성</td><td colspan="7"></td></tr>
<tr><td colspan="13"></td><td colspan="5">가시화</td><td colspan="2"></td></tr>
<tr><td colspan="13"></td><td colspan="5">모델 정확도 향상</td><td colspan="2"></td></tr>
<tr><td colspan="13"></td><td colspan="5">데이터 보충 가능 여부 확인</td><td colspan="2"></td></tr>
<tr><td colspan="14"></td><td colspan="4">오류 확인 및 최종 테스트</td><td></td></tr>
<tr><td colspan="17"></td><td colspan="4">최종 보고서 작성
및 발표준비</td></tr>
</table>

4. 구성원별 진척도

이름	진척도
이성무	데이터 전처리 완료 은닉 마코프 모델 구성 완료 모델 최적화와 성능 개선 방안 수집 중
정재원	데이터 전처리 완료 클러스터링을 활용한 관심 지점 군집 선정 완료 데이터 시각화 진행 중
하연지	데이터 수집 및 전처리 완료 학습 데이터 구축 완료 모델 최적화와 성능 개선 방안 수집 중

5. 보고 시점까지의 과제 수행 내용 및 중간 결과

5.1 데이터 전처리

2020년 및 2021년의 데이터를 분석에 이용하기 위하여, 각 월별로 2년의 데이터를 통합하였다. 이 과정에서 "20200307110532500" 형식으로 저장되어 있던 데이터 수집 시점을 "2020-03-07 11:05:33 AM" 형태로 바꾸어 주었다.

<표 > 데이터 가공 전

oid (렌터카 ID)	collection_dt (위치 정보 수집 시점)	longitude (경도)	latitude (위도)
46100018	20200307110532599	126.505399	33.5012831
46100018	20200307110802621	126.4975278	33.4977589

<표 > 데이터 가공 후

oid (렌터카 ID)	collection_dt (위치 정보 수집 시점)	longitude (경도)	latitude (위도)
46100018	2020-03-07 11:05:32 AM	126.505399	33.5012831
46100018	2020-03-07 11:08:02 AM	126.4975278	33.4977589

제주도의 끝을 대략적으로 나타내었을 때, 경도는 [126.143480, 126.973814] 위도는 [33.567186, 33.112476]이다. 각 렌터카 ID와 데이터 수집 시점 별로 이를 벗어난 데이터를 갖는 경우 삭제하였다.

<표 > 제주도를 벗어난 데이터 예시

oid (렌터카 ID)	collection_dt (위치 정보 수집 시점)	longitude (경도)	latitude (위도)
461000f7	20200329000006927	180	90
46100780	20200329000058356	129.164406	35.236036

이러한 데이터 처리 과정을 통해, 3,4,5월(봄)의 약 1300만개 데이터에서 약 60만개의 이상치를 제거했다.

5.2 관심 지점 선정

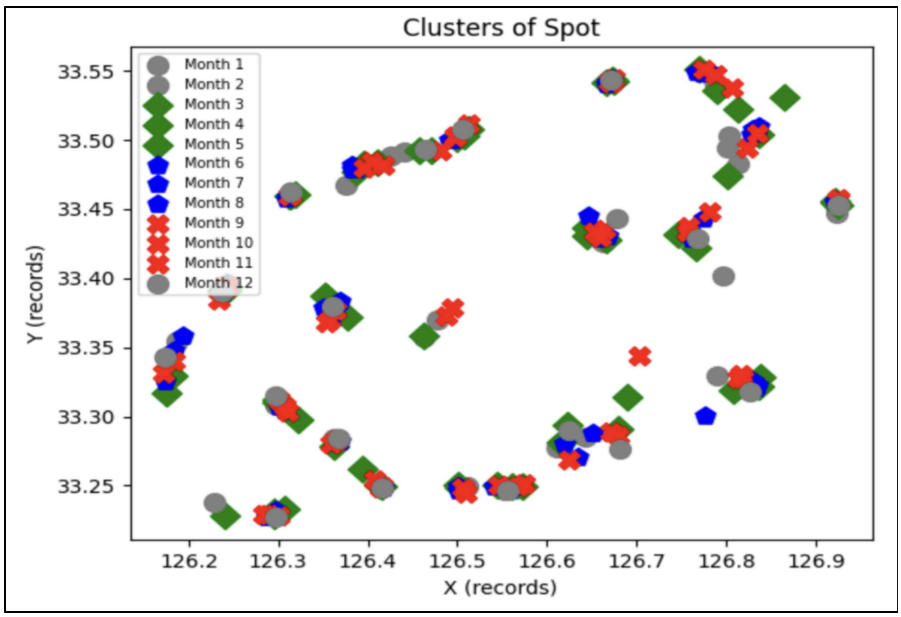
제주 데이터 허브에서 제공하는 "[교통, 안전] 월별 렌터카 체류 빈도" 데이터는 50m 격자로 위치를 나타낸 것으로, 격자 ID, 격자의 왼쪽, 위쪽, 오른쪽, 아래쪽 위치 정보와

격자의 중간 지점의 x, y좌표 정보로 구성되어 있다. 체류 지점의 데이터를 QGIS(Quantum GIS)를 사용해 지도에 나타내었을 때, 제주도 대부분 지역이 표시되었다. 따라서 유의미한 POI을 얻기 위해 해당 지점이 나타난 빈도 수(Oid_count)가 10번 이상인 지점만을 가지고 K-Means 클러스터링을 적용하였다. 좌표계 또한 "[교통, 안전] 월별 렌터카 체류 빈도" 에서 사용한 기존의 EPSG:5179에서 ESPG:4326으로 변경하였다. <표4>는 K-Means 클러스터링 결과 정보의 형식이다.

X (격자 중간 지점 X 좌표)	Y (격자 중간 지점 Y 좌표)	lon(GPS 상 경도)	lat(GPS 상 위도)
906782.2916	1501389.5833	126.4964	33.5003
898900.0	1473647.2222	126.4147	33.2494

<표 > K-Means 클러스터링 한 관심 지점 정보

사용자가 이동한 날이 주중인지 주말인지에 따라 모델을 달리 구성하였을 때 예측 성능이 향상된 연구[3]결과를 참고하여, 제주도를 여행할 때 계절에 따라 사용자들이 방문하는 POI에는 차이가 있을 것이란 가설을 세우고 계절별로 모델을 구축하기 위하여 1월부터 12월까지의 데이터를 3개월씩 나누어 POI을 선정하였다. 결과는 <그림 5>와 같다.



<그림 > 계절별 체류 지점 클러스터링 결과 (K=20)

[3] 김종환, 이석준, & 김인철. (2014). "이동 사용자의 다음 장소 예측을 위한 맵리듀스 기반의 분산 데이터 마이닝." *한국정보처리학회 학술대회논문집*, 21(1), 777-780.

5.3 은닉 마코프 모델 구성

효율적인 데이터 관리 및 사용을 위하여 MySQL을 이용했다. 계절마다 별도의 스키마(Schema)를 생성하고, 사용자의 경로 아이디를 PRIMARY KEY로 설정하여, 경로 시작점, 경로 도착점, 경로 시퀀스, 시간대 시퀀스를 저장했다. 경로 아이디는 (데이터 수집 일자, 렌터카 ID) 형식으로 변환했다. 2.3절에서 설명한 과정을 거친 후 데이터 베이스에 저장되는 정보의 형식은 표 5와 같다.

<표 > 데이터 베이스 저장 시 정보 형식

trajectory_id (사용자 경로 ID)	start_point (경로 시작점)	end_point (경로 도착점)	path (경로 시퀀스)	time_period (시간대 시퀀스)
(2020, 3, 14, '46100055')	POI7	POI4	['POI7', 'POI13', … , 'POI16', 'POI4']	['오전', '오전', …, '오후', '오후']

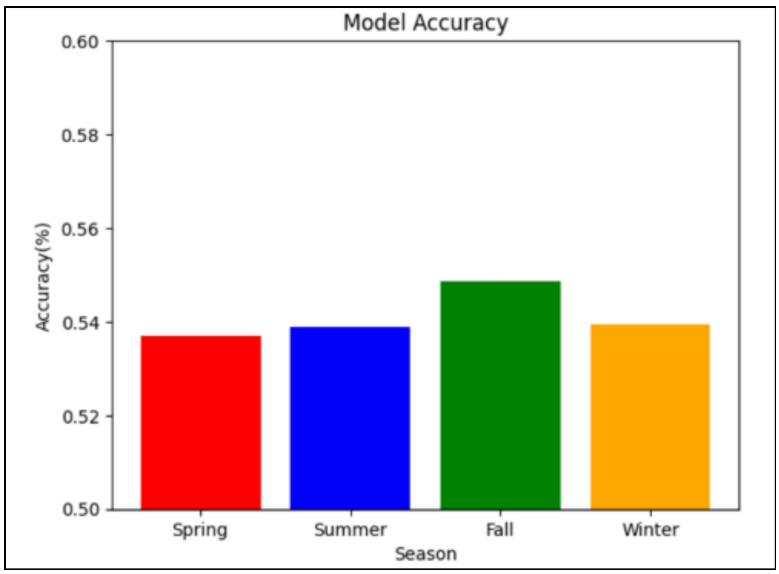
모델 학습을 위해 초기값, 전이 확률 행렬과 방출 확률 행렬을 계산하기 전, 모델 검증을 위해 MySQL에 저장된 계절별 데이터를 학습 데이터(Train Data)와 검증 데이터(Test Data)를 7:3의 비율로 나누었다.

은닉 상태는 알려지지 않았다는 가정을 하고 관측값을 각 경로 포인트마다 관측값 = (POI ID – 1) * |시간대 수| + (시간대 ID - 1)로 설정하여 분석에 이용하였으나, 관측값 종류의 개수에 비해 관측열의 길이가 짧아 기대하는 예측 성능을 얻지 못했다.
따라서 2.3절에서 설명한 바와 같이 POI ID * 4 + 시간대로 인덱싱하여 관측열 시퀀스를 생성하여 초기값, 전이 확률 행렬과 방출 확률 행렬을 구했다.

은닉 마코프 모델을 사용하기 위해 Python 의 hmmlearn 패키지를 이용하였으며 본 과제에서 사용하는 데이터의 특성에 따라 CategoricalHMM 모델을 분석에 이용했다. 앞서 구한 초기값, 전이 확률 행렬과 방출 확률 행렬을 파라미터로 사용하여 은닉 마코프 모델을 생성했다.

5.4 중간 결과

학습 데이터(Train Data)로 생성한 모델을 검증하기 위해 검증 데이터(Test Data)의 관측열을 디코딩하여 얻은 경로 시퀀스와 검증 데이터의 실제 경로 시퀀스를 비교하여 약 99%의 정확도를 확인하였다. 그러나 본 과제의 목표는 사용자의 이동 정보가 주어졌을 때, 다음 이동 경로지를 예측하는 것이기에 검증 데이터(Test Data)의 관측열에서 마지막 관측값을 제외하고, 다음 후보 경로를 붙인 새로운 관측열을 생성하여 가장 큰 우도 확률을 가지는 관측열을 디코딩하였다. 그 결과 계절별 모델의 정확도는 <그림 6>과 같이 약



<그림 > 계절별 모델의 정확도

53%에서 54%의 정확도를 보인다.

추후 정확도 향상을 위한 요소들을 분석하여 모델의 예측 정확도를 높여 나갈 것이며 패스트 맵 매칭을 활용한 풀스택 환경 또한 구축하여 본 과제의 목표를 달성하고자 한다.

6. 참고문헌

1. 김용중. (2013). "은닉 마르코프 모델을 이용한 스마트폰 사용자의 경로 모델링 및 온라인 학습 기반 목적지 예측." [석사, 연세대학교 대학원].

2. 김종환, 이석준, & 김인철. (2014). "이동 사용자의 다음 장소 예측을 위한 맵리듀스 기반의 분산 데이터 마이닝." *한국정보처리학회 학술대회논문집*, 21(1), 777-780.

3. Pant, N., & Elmasri, R. (2017, April). "Detecting meaningful places and predicting locations using varied k-means and hidden Markov model." *In Proceedings of the 17th SIAM International Conference on Data Mining (SDM 2017),* Houston, TX, USA (pp. 27-29).