Journal of The Korea Society of Computer and Information
Vol. 22 No. 10, pp. 121-128, October 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.10.121

Multi-Label Classification Approach to Location Prediction

Min Sung Lee*

Abstract

In this paper, we propose a multi-label classification method in which multi-label classification estimation techniques are applied to resolving location prediction problem. Most of previous studies related to location prediction have focused on the use of single-label classification by using contextual information such as user's movement paths, demographic information, etc. However, in this paper, we focused on the case where users are free to visit multiple locations, forcing decision-makers to use multi-labeled dataset. By using 2373 contextual dataset which was compiled from college students, we have obtained the best results with classifiers such as bagging, random subspace, and decision tree with the multi-label classification estimation methods like binary relevance(BR), binary pairwise classification (PW).

▸Keywords: Location Prediction, Multi-Label Classification, Data mining, Classifiers

I. Introduction

스마트폰의 등장은 최종 사용자로 하여금 유비쿼터스 환경을 실제로 체험하게 하는데 결정적인 기여를 하고 있다. 사용자가 언제 어디서에서나 스마트폰을 통하여 상당한 수준의 의사결정을 할 수 있게 되면서 장소예측(location prediction)의 중요성이 더욱 커지고 있다.

장소예측이란 사용자가 미래에 방문하고자 하는 장소, 즉 로케이션을 예측하는 것이다. 장소예측을 정확히 하려면 사용자가 과거에 방문하였던 장소에 대한 정보, 사용자의 신상정보, 그리고 다양한 상황정보 (contextual information)가 필요하다. 기존연구를 보면 장소예측은 도시계획[3], 마케팅 [28], 교통예측[16], 광고[6] 분야 등에서 활발히 활용되었다. 장소예측 연구는 대부분 사용자의 이동정보를 이용한다. 예를 들어 사용자의 이동정보를 이용하여 사용자가 특정장소를 언제 방문할 것인가를 예측한다 [21]. 또한 GPS 데이터와 전화기지국 자료와 같은 이동정보를 토대로 장소를 예측하기도 한다 [5, 18].

그러나, 장소예측에 관한 기존연구는 대부분 출력변수가 한 개이면서 클라스가 여러개인 이른바 단일 레이블 분류, 즉 SLC (single-label classification) 기법을 주로 이용하고 있다. 반면에, 사용자들은 활동시 복수개의 장소를 방문하며 장소예측 역시 사용자들에 대해서 복수개의 장소를 방문할 수 있다고 가정하고 장소예측을 하는 것이 자연스럽다. 그러나 이러한 경우 장소예측 문제는 출력변수가 한 개인 SLC 문제가 아닌, 출력변수가 복수개인 멀티 레이블 분류, 즉 MLC (multi-label classification) 문제가 된다. 따라서, 장소예측이 실제로 사용자의 생활패턴과 일치하고 사용자의 행동양식과 부합하는 결과를 내려면 SLC 문제가 아닌 MLC 문제로 변환되어 추정될 필요가 있다. 그러나 기존 장소예측 연구에서는 MLC에 의한 연구결과가 국내외적으로 거의 없다. 따라서, 이같은 연구의 필요를 충족하기 위하여 본 연구에서는 다음과 같은 연구질문 (Research Question : RQ)에 대한 답을 제시하고자 한다.

RQ: 장소예측 문제를 MLC 기법을 적용하여 해결할 때 어떠한 분류기와 예측방법이 가장 좋은 결과를 내는가 ?

본논문의 구성은 다음과 같다. 2장에서는 MLC 방법의 기초개념에 대해 상술한다. 3장에서는 장소예측에 관한 실증자료를 토대로 실험결과를 보여주고 그 의미를 서술한다. 4장에서는 결론 및 향후연구방향을 설명한다.

• First Author: Min Sung Lee, Corresponding Author: Min Sung Lee
*Min Sung Lee (em.analytics.co@gmail.com), EM Analytics Co
• Received: 2017. 09. 27, Revised: 2017. 10. 11, Accepted: 2017. 10. 16.

II. Multi-Label Classification Methods

본 연구의 핵심은 MLC 기법에 의한 장소예측 결과를 실증적으로 분석하는 것이다. MLC는 이미 기존에 많은 연구자들에 의하여 사용되어온 SLC 방법과는 다른 예측방법과 성과측정치를 사용한다. 이에 대해서 상세히 알아보자.

2.1 Fundamentals of MLC

SLC는 일반적으로 단일 분류기와 앙상블 분류기로 구분된다. 단일 분류기에는 로지스틱회귀, 의사결정트리, 인공신경망, 서포트벡터머신, 나이브 베이지안 네트워크 등이 있다. 앙상블 분류기에는 배깅, 랜덤 포레스트, 랜덤 서브스페이스, 부스팅, 스태킹 등이 있다. 이러한 SLC는 스키날씨 예측, 도산예측 등 수많은 의사결정문제에 적용되어 왔고 [7, 19], 최근에는 온라인 리뷰를 토대로 한 다양한 감성분석 연구에 적용되었다 [4, 27].

최근 진행되었던 장소예측 연구들을 보면 주로 휴대전화를 활용한 위치 정보에 의해 실행된다. 예를 들어 사용자가 SNS에 체크인 활동을 수행한 시점, 장소 데이터를 기반으로 위치예측을 수행하거나 [14], 와이파이 접속 위치에 따른 체크인 데이터, 버스 탑승 정보, 택시 이용 정보를 활용하여 장소 예측을 수행하기도 한다 [28]. 사용자가 '누구와 몇 분 동안 통화하는가'와 같은 통화데이터와 통화시 연결 기지국의 위치를 이용한 장소 예측을 할 수도 있다 [18, 29]. 건물의 용도를 사용해서 장소예측을 수행할 수도 있다 [23]. 최근에는 베이지안 네트워크를 이용하여 장소예측을 하는 연구도 활발하게 진행되고 있다 [9, 10, 11, 12, 13, 18].

이같은 기존연구들은 다양한 방법론을 적용하였으나 여전히 단일 레이블 방식으로 장소예측을 하고 있다. 그러나, 본 연구는 기존연구와 달리 멀티 레이블 방식으로 장소예측을 한다는 점에서 기존연구와 차별화된다. 본 연구에서 사용하는 MLC 방법을 살펴보자.

MLC 방법에는 문제변환방법과 알고리즘 적합화 방법이 있다 [24, 31].

첫째, 문제변환방법이란 MLC 문제를 하나 또는 복수개의 SLC 문제로 변환하여 해결하는 방법이다. 즉, 학습시 멀티 레이블 학습데이터를 단일 레이블 자료로 변환하여 해당 자료를 단일 레이블 분류기로 학습한다. 그리고 테스트 할때 학습된 단일 레이블 분류기를 이용하여 단일 분류 추정치를 계산한 다음 해당 추정치를 멀티 레이블 추정치로 변환한다. 여기에는 다양한 방법론이 존재하는데 대표적인 방법 몇개를 소개한다.

- 바이너리 적합 (Binary Relevance: BR) 방법은 MLC 문제를 복수개의 바이너리 분류문제로 변화하여 문제를 해결한다 [2].

- 레이블 파워셋 (LP: Label Powerset) 방법은 MLC 문제를 멀티 클래스 분류문제로 변환하여 해결하는 방법인데 레이블 결합 (LC: label combination)이라고도 한다 [25].

바이너리 쌍별분류 (PW: binary pairwise classification) 방법은 한쌍의 레이블에 대해서 바이너리 모델을 적용하여 MLC 문제를 해결하고자 하는 방법이다 [20].

- 랜덤 k-레이블셋 (RAkEL: Random k-Labelsets) 방법은 레이블 파워셋 (LP) 분류기를 기반으로 앙상블을 만들어서 최종적으로는 투표방법으로 MLC 추정치를 생성한다 [25].

- 분류기 체인 (CC: Classifier Chains) 방법은 MLC 문제를 일련의 바이너리 분류문제로 변환하여 문제를 푸는 방법이다. 여기에서 체인의 크기는 레이블 갯수와 일치한다. 그러나, 분류기 체인 방법은 해당 체인상에 들어가는 레이블의 순서에 따라 결과에 많은 차이가 발생하는 단점이 있다 [20].

- 분류기 체인 (CC)방법에서 레이블의 순서가 미치는 영향을 최소화 하기 위하여 앙상블 분류기 체인 (ECC: Ensemble CC)방법이 사용된다. 이는 복수개의 분류기 체인을 만들어서 레이블의 순서를 서로 달리한다. 그리고 임계치를 적용하여서 가장 적합한 레이블을 찾아 추정치를 정한다. 또한 몬테카를로 방법을 이용한 MCC방법도 사용된다.

둘째, 알고리즘 적합 방법은 현재 사용되는 분류 알고리즘을 일반화하여 멀티 레이블 데이터를 처리한다. 예를 들어서 멀티 레이블 k-최근접이웃 (MLkNN: Multi-Label k-Nearest Neighbor) [30] 방법은 k-최근접이웃 (k-Nearest Neighbor) 방법을 변환하여 멀티 레이블 자료를 처리할 수 있도록 한 방법이다. 한편, BRkNN은 바이너리 적합 (BR) 개념을 적용하여 MLC 문제해결시 k-최근접이웃 (k-NN)을 적용한다 [22]. PCT는 예측 클러스터링 트리 (Predictive Clustering Trees)의 약자로서 의사결정트리 알고리즘을 이용하여 멀티 레이블 자료를 처리한다 [1]. 반면 RF-PCT는 앙상블 기법인 랜덤 포레스트 (RF: Random Forest)를 이용하여 예측 클러스터링 트리 기법을 확장한 방법이다 [8].

2.2 MLC Performance Metrics

MLC 작업시 사용되는 성과측정치는 대략 16개 정도가 있다 [17]. 이중에서 관련 문헌에서 많이 사용되는 성과측정치는 7개가 있다 [15]. 이같은 성과측정치를 세개의 그룹으로 정리할 수가 있는바 즉, 샘플 기반 성과측정치, 레이블 기반 성과측정치. 그리고 랭킹 기반 성과측정치가 그것이다 [26]. 우선 샘플 기반 성과측정치에는 해밍로스 (Hamming Loss), 서브셋 정확도 (Subset Accuracy), 그리고 샘플 기반 F1 (example based F1)이 있다. 레이블 기반 성과측정치에는 마이크로 F1 (micro F1), 매크로 F1 (macro F1)이 있다. 그리고 랭킹 기반 성과측정치에는 평균 정밀도 (average precision), 그리고 원에러 (one error)가 있다. 성과 측정치에 관한 보다 자세한 설명은 별도의 문헌을 참조하면 된다 [31].

해밍 로스란 레이블중에서 오분류된 비율을 말한다. 따라서 해밍 로스값은 0에서 1 사이에서 움직이며 되도록 작을 수록 좋다. 서브셋 정확도는 주어진 학습자료의 레이블 셋이 해당 샘플의 그라운드 참 레이블 셋과 일치하는 비율을 말한다. 이 역시 0과 1사이의 값을 가지며 클 수록 좋다. 샘플 기반 F1값이란 샘플기반 정밀도 (precision)와 샘플 기반 리콜 (recall)의

조화평균을 의미한다. 마이크로 F1은 마이크로 정밀도와 마이크로 리콜의 조화평균이다. 이때 샘플기반 정밀도는 추정 레이블값과 그라운드 참 레이블값의 합의 규모를 추정 레이블값 규모로 나눈 값이다. 샘플기반 리콜도 유사하게 정의된다. 마이크로 정밀도는 모든 샘플과 레이블에 대해서 평균값으로 만든 정밀도를 말한다. 마이크로 리콜도 같은 방법으로 정의된다. 매크로 F1은 모든 레이블에 대한 정밀도와 리콜의 조화평균의 평균을 말한다. 평균 정밀도는 하나의 샘플에 포함되는 실제 레이블보다 더 높게 랭크된 레이블의 평균비율을 말한다. 커버리지는 하나의 샘플에 대해서 모든 실제 레이블을 커버하기 위하여 가장 높은 단계에서 아래로 평균 몇단계나 내려가는지를 계산한 것이다. 커버리지가 높다는 것은 그만큼 주어진 실제 레이블을 많이 커버하는 것이기 때문에 커버리지가 클 수록 좋다. 원에러는 샘플중에서 가장 높게 랭크된 레이블이 관련 적정 레이블에 속하지 않는 샘플들의 비율을 말한다. 따라서 원에러는 0과 1 사이의 값을 가지며 작을 수록 좋다.

본 연구에서는 이같은 성과측정치중에서 장소예측에 적합하다고 판단되는 성과측정치인 해밍로스, 원에러, 서브셋 정확도, 평균 정밀도, 그리고 세 개의 F1값 (즉, 샘플기반 F1, 마이크로 F1, 매크로 F1)를 적용한다.

III. Experiment and Results

3.1 Data Sample

본 연구에 사용된 자료는 서울의 한 사립 대학교 인문사회 캠퍼스 내의 학생들을 대상으로 학교 내 방문 건물, 이동경로, 건물 내 활동, 인구통계학적 특성들을 설문을 통해 2주간 조사한 상황자료다. 그리고 본 자료는 기존연구에서 다양한 분류기를 이용한 단일 레이블 분류문제로 사용된 바 있으나 [9, 10, 11, 12, 13], 본 연구에서는 해당 자료를 MLC 문제에 맞게 다시 수정하여 모두 2373개의 상황예측 자료로 구성하였다.

MLC 문제에서는 동일인이 여러 건물을 방문한 것이라면 이를 별도의 자료로 간주하지 않는다. 오히려 이를 같은 하나의 자료내에서 멀티 레이블로 표기된 건물에 표기를 한다. 따라서, MLC문제는 단일 레이블 분류 문제와 비교하여 주어진 학습자료의 개수를 줄이는 효과가 있다. 그렇기 때문에 본 자료를 단일 레이블 분류문제로 풀때에는 6868개의 자료였지만 이를 MLC 문제로 변환하면서 동일인이 방문한 여러건물을 멀티 레이블을 이용하여 하나의 자료로 간주할 수가 있어서 2373개의 자료로 재구성된 것이다.

본 연구에서는 인구통계학적 특성들과 건물 내 활동내역을 토대로 사용자가 방문한 건물을 추정하기 위해 표 1에 정리된 22가지 변수를 사용했다. 변수 내 범주의 수는 정의 칸의 괄호 안에 표기했다.

본 상황자료는 개인정보 보호를 위하여 설문자가 대학 캠퍼스에 들어왔을때에 한하여 수집한 자료이다. 참가자들에게는 일정액의 인센티브를 지급하였다. 본 연구에서 대상으로 하는 타겟변수는 방문한 건물을 예측하는 것이다. 즉, 표 1에서 첫번째 변수인 건물명 (Building Name)이 타겟변수로 사용되었다. 해당 타겟변수는 19개의 건물명을 클래스로 갖고 있다. 표 2는 사용자가 방문한 건물명을 정리하여 보여준다.

Table 1. Variables

Variable	Remarks(# of classes)
Building Name	Building names which users visited (19)
Activity Code	Activities done in buildings (17)
Gender	male or female (2)
Age	age (2)
Major	majoring departments (15)
Grade	grades (4)
Military	whether military services were done (2)
Religion	name of religions (5)
Monthly Allowance	monthly allownace amounts (8)
Smoking	smoking or not (2)
Lover	whether users have girl (boy) friends (2)
Weekday Leisure	types of leisures in weekday(5)
Weekend Leisure	types of leisures on weekend(5)
Vacation Leisure	types of leisures in vacation days (5)
Lunch Leisure	types of leisures in the lunch time (5)
Leisure Utility	degree of perceived utility from leisure activities (5)
Leisure Satisfaction	degree of perceived satisfaction from leisure activities (5)
Housing	types of housing (5)
Arrival Transportation	types of transportation when commuting to school (5)
Departure Transportation	types of transportation when going home after school(5)
Average Study Time	amount of time spent for daily study (8)
Monthly Mobile Phone Fee	average mobile phone fee paid on a monthly basis (8)

Table 2. Name of Buildings

ID#	Building Name
1	Hoam Hall
2	Toegye Hall Of Humanities
3	International Hall
4	Suseon Hall
5	Dasan Hall Of Economics
6	Business Building
7	600th Anniversary Building
8	Law Building
9	Central Library
10	Suseon Hall Annex
11	Student Union
12	East Gate
13	Front Gate
14	Outside Campus
15	Basketball Court
16	Large Playground
17	Geumjandi Square
18	Faculty Hall
19	Rear Gate

3.2 Results

본 연구에서는 장소예측을 위한 작업을 위하여 2373개의 자료를 토대로 다음과 같은 분석을 하였다.

첫째, MLC작업을 위하여 사용된 분류기로는 우선 단일 분류기와 앙상블 분류기를 골고루 사용하였다. 단일 분류기로는 로직스틱 회귀(LR), 의사결정트리 (DT), 인공신경망 (NN), 나이브 베이지안망 (NB), 서포트벡터머신(SVM), k최근접이웃(kNN) 분류기를 사용하였다. 앙상블 분류기로는 랜덤포레스트(RF), 부스팅(AB), 배깅(BA), 스태킹(ST). 랜덤서브스페이스(RS)를 사용하였다.

둘째, MLC추론방법중 본연구에서는 기본적인 벤치마킹 추론방법으로 널리 사용되는 BR, CC, RAkEL외에 이를 개선한 PW, MCC 를 사용하였다.

셋째, 성과측정치로는 해밍로스 (HL), 원에러 (OE), 서브셋 정확도(SA),평균 정밀도 (AP), 샘플기반 F1 (EF), 마이크로 F1(MiF), 매크로 F1(MaF)를 적용하였다.

3.3 Implications

표 3에 정리된 연구결과는 각 성과지표별로 상위 11개 값을 정리한 것이다. 체계적인 결과해석을 위해 다음과 같이 세가지 경우로 나누어서 보자.

첫째, 우선은 오차위주의 해석을 하기 위하여 해밍로스(HL)와 원에러(OE)를 비교하고 그에 따른 분류기와 MLC추론방법의 적용정도를 알아본다. 해밍로스와 원에러를 정렬을 시켜서 어떤 분류기와 어떤 MLC추론방법에서 상대적으로 더 유리한 해밍로스와 원에러가 나오는지를 보고자 한다.

둘째, 정확도와 정밀도 위주로 보기 위하여 서브셋 정확도와 평균 정밀도를 비교하고 그에 따른 분류기와 MLC추론방법의 적용정도를 비교한다. 이를 위하여 서브셋 정확도 SA와 평균정밀도 AP를 정렬하여 보다 정확한 의미를 해석한다.

셋째, F1값을 중심으로 해석한다. 이를 위하여 샘플기반 F1(EF), 마이크로 F1(MiF), 매크로 F1(MaF)값을 구한 후에 이들 간 비교를 통하여 어떤 분류기와 어떤 MLC추론방법이 상대적으로 기여를 더 많이 하는지를 분석한다.

우선, 해밍로스와 원에러 관점에서 표 3의 결과를 요약하면 표 4와 같다. 각 분류기 및 MLC 추론방법 옆의 숫자는 해당 분류기와 MLC 추론방법이 해밍로스와 원에러 측면에서 유리한 성과를 보여준 횟수를 의미한다. 이는 가장 우수한 결과로부터 상위 11개 순위까지를 카운트한 숫자이다.

표 4는 분류기 관점으로 성과를 보여주고 있다. 즉, 해밍로스의 경우 배깅과 랜덤 서브스페이스가 좋은 성과를 보이고 있고, 원에러의 경우 배깅이 상대적으로 좋은 성과를 보이고 있다. 결국 분류기 관점으로 보면 배깅과 랜덤 서브스페이스의 상대적으로 우수한 성과를 보이고 있음을 알 수 있다. MLC 추론방법 관점으로 보면 해밍로스에서는 BR이 원에러에서는 PW와 BR의 성과가 좋다. 정리하면 해밍로스와 원에러 관점에서 볼 때 분류기에서는 배깅이 공통적으로 좋은 성과를 보이고 MLC 추론방법으로는 BR이 공통적으로 좋은 성과를 보인다. 해밍로스와 원에러와 같이 MLC를 적용할때 발생되는 오차를 줄이기 위해서는 배깅 분류기를 이용하여 BR 추론방법을 적용하는 것이 상대적으로 유리하다.

서브셋 정확도와 평균 정밀도 관점에서 표 3의 결과를 요약하면 표 5와 같다. 각 분류기 및 MLC 추론방법 옆의 숫자에 대한 해석은 표 5와 같다. 즉, 해당 분류기와 MLC 추론방법이 서브셋 정확도와 평균 정밀도 측면에서 가장 우수한 결과로부터 상위 11개 순위까지를 카운트하였을 때 포함된 숫자이다.

Table 3. Results by Error, Accuracy, Precision, F1 Measures

Hamming Loss and One Error				Subset Accuracy & Average Precision				F1 Measures					
Method	HL	Method	OE	Method	SA	Method	AP	Method	EF	Method	MiF	Method	MaF
BA_BR	0.076	DT_PW	0.389	BA_PW	0.426	BA_BR	0.69	BA_PW	0.517	BA_PW	0.537	DT_PW	0.399
BA_CC	0.077	BA_PW	0.39	BA_MCC	0.389	RS_BR	0.676	DT_PW	0.504	DT_PW	0.504	AB_PW	0.381
BA_MCC	0.077	BA_BR	0.398	DT_PW	0.388	RF_BR	0.638	AB_PW	0.467	BA_MCC	0.472	BA_PW	0.381
RS_PW	0.077	RS_BR	0.403	DT_CC	0.382	AB_BR	0.635	BA_MCC	0.447	AB_PW	0.469	DT_CC	0.32
RS_BR	0.078	AB_PW	0.436	DT_MCC	0.382	DT_PW	0.615	DT_CC	0.441	DT_CC	0.463	DT_MCC	0.32
NN_BR	0.08	BA_MCC	0.446	BA_CC	0.372	AB_PW	0.613	DT_MCC	0.441	DT_MCC	0.463	AB_CC	0.32
RS_CC	0.08	RF_BR	0.461	RS_PW	0.37	DT_BR	0.605	NN_PW	0.432	DT_BR	0.462	AB_MCC	0.319
RS_RAkEL	0.08	AB_BR	0.466	AB_MCC	0.368	NN_BR	0.603	AB_MCC	0.431	BA_CC	0.462	AB_BR	0.313
RS_MCC	0.08	BA_CC	0.466	AB_CC	0.366	ST_BR	0.576	AB_CC	0.429	NN_PW	0.461	DT_BR	0.302
DT_BR	0.081	NN_BR	0.467	NN_PW	0.359	LR_BR	0.573	BA_CC	0.428	BA_BR	0.46	AB_RAkEL	0.295
BA_RAkEL	0.081	DT_BR	0.473	RF_MCC	0.352	BA_PW	0.568	RS_PW	0.425	RS_PW	0.455	BA_MCC	0.285

Table 4. Contributions of Classifier and MLC Estimation Methods in Hamming Loss and One Error

	Hamming Loss		One Error		Sum
Classifier	BA	4	BA	4	8
	RS	5	RS	1	6
	AB	0	AB	2	2
	RF	0	RF	1	1
	DT	1	DT	2	3
	NN	1	NN	1	2
	LR	0	LR	0	0
	ST	0	ST	0	0
MLC Estimation Methods	PW	1	PW	3	4
	CC	2	CC	1	3
	MCC	2	MCC	1	3
	BR	4	BR	4	8
	RAkEL	2	RAkEL	0	2

Table 5. Contributions of Classifier and MLC Estimation Methods in Subset Accuracy and Average Precision

	Subset Accuracy		Average Precision		Sum
Classifier	BA	3	BA	2	5
	RS	1	RS	1	2
	AB	2	AB	2	4
	RF	1	RF	1	2
	DT	3	DT	2	5
	NN	1	NN	1	2
	LR	0	LR	1	1
	ST	0	ST	1	1
MLC Estimation Methods	PW	4	PW	3	7
	CC	3	CC	2	5
	MCC	4	MCC	0	4
	BR	0	BR	8	8
	RAkEL	0	RAkEL	0	0

분류기 관점으로 정리하면 배깅, 부스팅, 그리고 의사결정트리가 유리한 성과를 보이고 있음을 알 수 있다. 배깅은 해밍로스와 원에러에서도 우수한 성과를 보이는 분류기이기 때문에 배깅은 에러는 줄이는 목적과 정확도와 정밀도를 높이는 목적 달성에 유리한 분류기임을 알 수 있다. MLC 추론 관점으로 보면 BR, CC와 같은 전통적인 MLC 추론기법이 유리하고 아울러 PW, MCC 추론기법도 유리한 결과를 보이고 있다. 따라서, 서브셋 정확도와 평균 정밀도 관점으로 보면 MLC 추론기법은 RAkEL을 제외하고는 모두 다 좋은 성과를 보여준다. 한가지 특기할 만한 사실은 MLC 추론방법에서 BR은 평균 정밀도 향상에는 유리하지만 서브셋 정확도 향상에는 기여를 못하고 있다. 반면 MCC는 서브셋 정확도 향상에는 유리하지만 평균 정밀도 향상에는 기여를 못하고 있다. 만약 서브셋 정확도와 평균 정밀도를 동시에 향상시키는 것이 목적이라면 MLC 추론방법 중 CC와 PW를 적용하는 것이 유리하다.

표 6은 표 3의 내용을 세가지 종류의 F1값 기준으로 정리한 것이다. 이를 분류기 관점으로 정리하면 배깅, 부스팅, 의사결정트리가 상대적으로 좋은 성과를 보여주고 있다. 보다 세부적으로 조사해보면 배깅이 모든 F1값에서 좋은 성과를 보여주고 있고, 의사결정트리 역시 모든 F1값에서 유리한 성과를 보여주고 있다. 반면 부스팅은 샘플기반 F1과 매크로 F1에서 상대적으로 더 좋은 성과를 보여주고 있다. MLC 추론방법 관점으로 보면 PW, CC, MCC가 모든 F1값에 대해서 좋은 성과를 보여주고 있다. 그중에서도 PW는 다른 추론방법에 비해 더 좋은 성과를 보이고 있다. BR도 어느정도 성과를 보이고는 있으나 CC에 비해서는 성과가 떨어지는 것으로 나타났다.

Table 6. Contributions of Classifier and MLC Estimation Methods in F1 Measures

	Example F1		Micro F1		Macro F1		Sum
Classifier	BA	3	BA	4	BA	2	9
	RS	1	RS	1	RS	0	2
	AB	3	AB	1	AB	5	9
	RF	0	RF	0	RF	0	0
	DT	3	DT	4	DT	4	11
	NN	1	NN	1	NN	0	2
	LR	0	LR	0	LR	0	0
	ST	0	ST	0	ST	0	0
MLC Estimation Methods	PW	5	PW	5	PW	3	13
	CC	3	CC	2	CC	2	7
	MCC	3	MCC	2	MCC	3	8
	BR	0	BR	2	BR	2	4
	RAkEL	0	RAkEL	0	RAkEL	1	1

성과측정치별로 베스트 분류기와 MLC추론방법을 정리하면 표 7과 같다. 해밍로스, 원에러 기준으로 볼 때 가장 좋은 분류기로는 배깅, 랜덤 서브스페이스이고, 가장 좋은 성과를 보이는 MLC 추론방법은 BR, PW이다. 해밍로스, 원에러와 같이 오차 관점으로 볼 때 가장 좋은 성과를 보이는 분류기는 모두 앙상블 임이 특기할 만하다. 반면, 서브셋 정확도, 평균 정밀도와 같이 성과측면으로 보면 가장 좋은 분류기는 배깅과 의사결정트리이다. 그리고 BR, PW가 가장 좋은 서브셋 정확도와 평균정밀도를 산출하는 MLC 추론방법이다. 이 경우 단일 분류기에서는 유일하게 의사결정트리가 포함되어 있음이 주목된다. 샘플기반 F1, 마이크로 F1, 매크로 F1 관점으로 보면 가장 좋은 성과를 보이는 베스트 분류기는 의사결정트리이다. 그리고 그 뒤를 이어 배깅과 부스팅이다. 그리고 베스트 MLC 추론방법은 PW, MCC이다.

결국, 성과측정치 측면으로 볼 때 제일 유리한 분류기는 배깅이 유일하며, 아울러 제일 유리한 MLC 추론방법은 PW이다. 반면, 의사결정트리는 서브셋 정확도, 평균 정밀도, 샘플기반 F1, 마이크로 F1, 매크로 F1 측면에서 좋은 성과를 보여주고 있다.

표 8은 성과측정치별로 성과가 불리한 분류기와 MLC 추론방법을 정리한 것이다. 해밍로스와 원에러 관점에서 볼 때 가장 불리한 분류기는 로지스틱 회귀와 스태킹이며 이는 서브셋 정확도와 평균정밀도의 경우도 마찬가지이다. 그리고 해밍로스, 원에러 관점에서 가장 불리한 MLC 추론방법은 RAkEL이며 이 역시 서브셋 정확도와 평균 정밀도 측면에서도 가장 불리한 MLC 추론방법이다. F1 값 기준으로 볼 때 가장 불리한 분류기

는 로지스틱 회귀, 스태킹, 그리고 랜덤 포레스트 순으로 나타났다. 그리고 가장 불리한 MLC 추론방법은 RAkEL이다. 결론적으로 성과측면에서 볼 때 가장 불리한 결과를 보여주는 분류기는 로지스틱 회귀와 스태킹이며, 가장 불리한 MLC 추론방법은 RAkEL이다.

Table 7. Best Classifiers and MLC Estimation for Performance Metrics

Performance Metrics	Classifier	MLC Estimation
Hamming Loss, One Error	Bagging, Random Subspace	BR, PW
Subset Accuracy, Average precision	Bagging, Decision Tree	BR, PW
Example F1, Micro F1, Macro F1	Decision Tree, Bagging, Boosting	PW, MCC

Table 8. Worst Classifiers and MLC Estimation for Performance Metrics

Performance Metrics	Classifier	MLC Estimation
Hamming Loss, One Error	Logistic Regression, Stacking	RAkEL
Subset Accuracy, Average precision	Logistic Regression, Stacking	RAkEL
Example F1, Micro F1, Macro F1	Logistic Regression, Random Forest, Stacking	RAkEL

IV. Conclusions

본 연구에서는 사용자의 인구통계학적 정보와 활동정보를 토대로 미래에 방문할 장소를 멀티 레이블 분류방법, 즉 MLC 추론방법을 이용하여 예측했다. 본 연구에서 사용한 MLC 추정방법은 BR, CC, MCC, PW, RAkEL 등 모두 5개 방법을 적용하였다. 이들 5개 방법은 알고리즘 적합기법과 문제 변환방법을 동시에 커버하므로써 MLC 방법의 핵심사항을 모두 적용하였다. 그리고 본 연구에서 적용한 성과측정치는 모두 7개이고 이는 기존 MLC방법에서 널리 사용되는 성과측정치들이다. 즉, 샘플 기반 성과측정치로서 해밍로스 (Hamming Loss), 서브셋 정확도 (Subset Accuracy), 그리고 샘플 기반 F1 (example based F1)을 적용하였다. 그리고 레이블 기반 성과측정치로서 마이크로 F1 (micro F1), 매크로 F1 (macro F1)을 적용하였고 랭킹 기반 성과측정치로서는 평균 정밀도 (average precision)와 원에러 (one error)를 적용하였다. 아울러 MLC 분석을 위하여 기존 연구에서 사용한 모든 주요 분류기를 골고루 적용하였다. 즉, 단일 분류기에서 로지스틱회귀, 의사결정트리, 인공신경망을 사용하였고 앙상블 분류기로는 배깅, 스태킹, 부스팅, 랜덤 포레스트, 랜덤 서브스페이스를 적용하였다.

본 연구의 결과와 의의를 정리하면 다음과 같다.

첫째, 기존 연구와 달리 본 연구에서는 사용자의 인구통계정보와 활동정보를 토대로 MLC추론방법을 적용하여 결과를 도출하였다. 이같은 분석시도는 기존 장소예측 분야에서는 처음으로 시도되었다는 점에서 학술적, 실무적 의의가 있다.

둘째, 해밍로스, 원에러 (이상 오차 측면 성과치), 서브셋 정확도, 평균 정밀도 (이상 정확도와 정밀도 측면의 성과치), F1 값 등 성과측정치 관점에서 유리한 분류기와 MLC 추론방법, 불리한 분류기와 MLC 추론방법을 제시하였다. 그 결과 모든 성과측면에서 유리한 분류기는 배깅이고 모든 성과측면에서 유리한 MLC 추론방법은 PW인 것으로 확인되었다. 반면, 의사결정트리는 서브셋 정확도, 평균 정밀도와 F1값 측면에서 유리한 분류기였다.

셋째, 성과측면에서 불리한 분류기와 MLC 추론방법도 분석하였다. 그 결과 해밍로스, 원에러, 서브셋 정확도, 평균 정밀도, F1 값등 모든 성과측면에서 볼 때 가장 불리한 결과를 보여주는 분류기는 로지스틱 회귀와 스태킹이며, 가장 불리한 MLC 추론방법은 RAkEL인 것으로 확인되었다.

넷째, 기존 장소예측 연구는 성과측정치가 정확도나 AUC 정도에 국한되었지만 본 연구는 MLC에서 도출되는 다양한 성과측정치를 파악할 수가 있었다. 이같은 MLC기법의 특성은 향후 장소예측에 관한 연구에서 많은 기여를 할 것으로 기대된다.

본 연구는 장소예측에 관한 연구중에서 MLC 기법을 적용한 최초의 연구로 파악된다. 본 연구와 관련되어 향후 연구주제는 다음과 같다.

첫째, SLC와 MLC간 성과비교를 보다 면밀하게 할 필요가 있다. 주어진 장소예측 자료의 성격에 따라 SLC와 MLC가 서로 상충된 결과를 보일 수 있기 때문이다.

둘째, MLC에 적합한 속성추출 (Feature Selection) 방법을 연구할 필요가 있다. SLC에는 이미 다양한 속성추출 방법이 기존연구에서 제안된 바 있으나 MLC에서는 아직 속성추출에 관한 연구가 많지 않기 때문이다.

향후 본 연구가 장소예측에서 MLC 기법이 보다 활발하게 적용되는 기본연구로 이용될 수 있다.

REFERENCES

[1] H. Blockeel, L. De Raedt, and J. Ramon, “Top-down induction of clustering trees”, Proceedings of the 15th International Conference on Machine Learning, pp. 55-63, July 1998.

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, “Learning multi-label scene classification.” Pattern Recognition, Volume 37, Issue 9, pp.1757-1771, September 2004

[3] M. Dash, H. L. Nguyen, C. Hong, G. E Yap, M. N. Nguyen, X. Li, S. P. Krishnaswamy, J. Decraene, S. Antonatos,

Y. Wang, D. T. Anh, and A. Shi-Nash, "Home and work place prediction for urban planning using mobile network data", In IEEE 15th Mobile Data Management, Vol.2, pp.37-42, July 2014

[4] N. F. F. da Silva, E. R. Hruschka, E. R. Hruschka Jr., "Tweet sentiment analysis with classifier ensembles", Decision Support Systems, Vol.66, 170-179. October 2014

[5] J. Hu, Y. Wang, and Y. Zhang, "IOHMM for location prediction with missing data," In IEEE Data Science and Advanced Analytics. pp.1-10. October 2015

[6] G. A. Johnson, R. A. Lewis, and D. Reiley, "Location, Location, Location: Repetition and Proximity Increase Advertising Effectiveness," Available at SSRN: https://ssrn.com/abstract=2268215, October 2017.

[7] M. A. King, A. S. Abrahams, and C. T. Ragsdale. "Ensemble methods for advanced skier days prediction", Expert Systems with Applications, Vol.41, Issue 4, pp.1176-1188, March 2014

[8] D. Kocev, C. Vens, J. Struyf, and S. Dz˘eroski, "Ensembles of multi-objective decision trees." Proceedings of the 18th European conference on machine learning, pp. 624-631, January 2007

[9] J. S. Lee and E. S. Lee, "Exploring the usefulness of a decision tree in predicting people's locations," Procedia-Social and Behavioral Sciences, Vol.140, pp.447-451. August 2014

[10] K. C. Lee, and H. Cho, "Performance of ensemble classifier for location prediction task: emphasis on Markov Blanket perspective." International Journal of u-and e-Service, Science and Technology, Vol.3, Issue.3, October 2010

[11] K. C. Lee, and H. Cho, "Integration of general Bayesian network and ubiquitous decision support to provide context prediction capability," Expert Systems with Applications, Vol.39, Issue.5, pp.6116-6121, April 2012

[12] K. C. Lee, and H. Cho, "Designing a Ubiquitous Decision Support Engine for Context Prediction: General Bayesian Network Approach." International Journal of u-and e-Service, Science and Technology, Vol. 3, Issue. 3, pp.25-36, September 2010

[13] S. Lee, K. C. Lee, and H. Cho, "A dynamic Bayesian network approach to location prediction in ubiquitous computing environments." Proceeding of the 4th International Conference on Advances in Information Technology, pp73-82, November 2010

[14] D. Lian, X. Xie, V. W. Zheng, N. J. Yuan, F. Zhang, and E. Chen, "CEPR: A collaborative exploration and periodically returning model for location prediction," ACM Transactions on Intelligent Systems and Technology, Vol. 6, Issue.1, April 2015

[15] S. M. Liu, M. J. H. Chen. "A multi-label classification based approach for sentiment classification", Expert Systems with Applications, Vol. 42, Issue 3 pp. 1083-1093, February 2015

[16] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with big data: a deep learning approach," IEEE Transactions on Intelligent Transportation Systems Vol.16, Issue 2, pp.865-873, April 2015

[17] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dz˘eroski. "An extensive experimental comparison of methods for multi-label learning." Pattern Recognition, Vol.45, Issue 9, pp.3084-3104 September 2012

[18] D. Matekenya, M. Ito, R. Shibasaki, and K. Sezaki, "Enhancing location prediction with big data: evidence from dhaka," Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp.753-762, September 2016

[19] D. L. Olson, D. Delen, and Y. Meng. "Comparative analysis of data mining methods for bankruptcy prediction", Decision Support Systems, Vol. 52, Issue 2, pp.464-473, January 2012

[20] J. Read, B. Pfahringer, G. Holmes, and E. Frank. "Classifier chains for multilabel classification. Machine Learning.", Vol. 85, Issue 3, pp. 333-359, December 2011

[21] J. Scott, A. J. B. Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar, "PreHeat: controlling home heating using occupancy prediction," Proceedings of the 13th ACM international conference on Ubiquitous computing, pp.281-290, September 2011

[22] E. Spyromitros, G. Tsoumakas, and I. Vlahavas. "An empirical study of lazy multilabel classification algorithms." Proceedings of the 5th Hellenic conference on artificial intelligence: Theories models and applications. pp.401-406, October 2008

[23] A. Thomason, M. Leeke, and N. Griffiths, "Understanding the impact of data sparsity and duration for location prediction applications," International Internet of Things Summit. Springer International Publishing, pp.192-197, October 2014

[24] G. Tsoumakas, L. Katakis, and I. Vlahavas. "Mining multi-label data." In Data mining and knowledge discovery handbook, pp. 667-685. September 2010

[25] G. Tsoumakas, L. Katakis, and I. Vlahavas. "Random k-labelsets for multilabel classification." IEEE Transactions on Knowledge and Data Engineering, Vol. 23, Issue 7, pp. 1079-1089, July 2011

[26] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. "Mulan: A java library for multi-label

learning." Journal of Machine Learning Research, Vol. 12, pp. 2411-2414, February 2012

[27] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu. "Sentiment classification: The contribution of ensemble learning", Decision Support Systems, Vol.57, pp. 77-93, January 2014

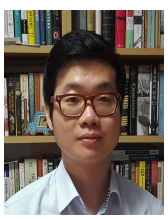
[28] Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui, "Regularity and conformity: Location prediction using heterogeneous mobility data," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1275-1284. August 2015

[29] D. Zhang, D. Zhang, H. Xiong, L. T. Yang, and V. Gauthier, "NextCell: Predicting location using social interplay from cell phone traces." IEEE Transactions on Computers, Vol. 64, Issue. 2 pp.452-463. February 2015

[30] M. L. Zhang, and Z. H. Zhou. "Ml-knn: A lazy learning approach to multi-label learning." Pattern Recognition, Vol. 40, Issue. 7, pp. 2038-2048, July 2007

[31] M. L. Zhang, and Z. H. Zhou. "A review on multi-label learning algorithms." IEEE Transactions on Knowledge and Data Engineering. Vol. 26, Issue. 8, pp.1819-1837, August 2014

Authors



Min Sung Lee is in charge of EM Analytics Co. The company's business model lies in big data analytics, customer sentiment analysis, providing IoT-driven consulting service to clients. His research interests include python-based deep learning mechanisms, ensemble CNN (convolutional neural network), and development of novel feature selection methods for multi-label classification. Currently, he is also writing scholarly articles about start-up entrepreneurship by using PLS (Partial Least Square).