Journal of The Korea Society of Computer and Information
Vol. 22 No. 11, pp. 97-104, November 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.11.97

# Data mining approach to predicting user's past location

Eun Min Lee*, Kun Chang Lee**

## Abstract

Location prediction has been successfully utilized to provide high quality of location-based services to customers in many applications. In its usual form, the conventional type of location prediction is to predict future locations based on user's past movement history. However, as location prediction needs are expanded into much complicated cases, it becomes necessary quite frequently to make inference on the locations that target user visited in the past. Typical cases include the identification of locations that infectious disease carriers may have visited before, and crime suspects may have dropped by on a certain day at a specific time-band. Therefore, primary goal of this study is to predict locations that users visited in the past. Information used for this purpose include user's demographic information and movement histories. Data mining classifiers such as Bayesian network, neural network, support vector machine, decision tree were adopted to analyze 6868 contextual dataset and compare classifiers' performance. Results show that general Bayesian network is the most robust classifier.

▸Keyword: Location prediction, Data mining, Classifiers, Bayesian network, Contextual dataset

## I. Introduction

최근 스마트폰의 사용이 보편화되면서 장소예측(Location prediction)의 중요성이 증대되고 있다. 장소예측은 사용자의 상황정보(Contextual information)를 분석하여 예측한다. 여기서 상황(Context)이란 사용자가 처한 환경을 나타내는 다양한 정보이다[1]. 일반적으로 장소예측은 일정시점 이후에 사용자가 방문할 장소를 예측한다[2,3]. 그럼에도 불구하고 장소예측 시 사용자가 과거에 방문했던 장소를 예측하는 것은 실제 상황에서 유용하다. 예를 들어 의료분야에서 전염병 원인을 파악하려면 환자가 과거에 방문하였던 장소를 추정할 필요가 있다[4]. 살인 사건의 경우도 마찬가지이다. 범죄학 연구에 의하면 용의자가 과거에 거주하던 장소 근처에서 살인이 발생하는 경우가 많다[5]. 따라서 용의자의 인구통계정보와 이동경로 등을 이용하여 용의자가 과거에 머물렀던 장소를 예측할 수 있다면 용의자의 여죄까지 알아 낼 수 있을 것이다. 또한 대형 공공장소에서 부모가 아이를 잃어버린 경우 해당 아동의 이동경로를 이용하여 과거에 머물렀던 장소를 추정 할 수 있다면 미아 찾기에 도움이 된다[6].

본 연구는 기존의 장소예측 연구와 달리 과거 장소를 예측하고자 한다. 이를 위해 데이터 마이닝에서 널리 사용되는 단일분류기(Single classifiers)를 사용한다. 본 연구에서 분석하고자 하는 연구질문(RQ: Research Question)은 다음과 같다.

RQ: 데이터 마이닝 단일분류기를 이용하여 사용자가 과거에 방문한 장소를 예측한다.

본 연구에서 사용하는 단일분류기는 베이지안 네트워크(Bayesian Network), 의사결정트리(Decision Tree), 인공신경망(Neural Network), 서포트벡터머신(Support Vector Machine)이다.

본 연구의 구성은 다음과 같다. 2장에서는 장소예측에 관한 기존 선행 연구를 소개하고 3장에서는 단일분류기에 대해 서술한다. 4장에서는 실제 장소예측 자료를 토대로 결과를 설명한

• First. Author: Eun Min Lee, Corresponding Author: Kun Chang Lee
*Eun Min Lee (dldmsals159@hanmail.net), Global Business Administration, Sungkyunkwan University
**Kun Chang Lee (kunchanglee@gmail.com), SKK Business School/ SAIHST, Sungkyunkwan University
• Received: 2017. 03. 14, Revised: 2017. 04. 27, Accepted: 2017. 08. 28.

다. 5장에서는 본 연구의 시사점 및 향후 연구방향을 제시한다.

## II. Preliminaries

장소예측(Location prediction)에 관한 기존연구를 살펴보면 다음과 같다.

최근 다양한 분야에서 단일분류기를 통한 예측이 활용되고 있다. 주식 반환예측[7], 실패예측모델구성[8]과 같은 분야에서 단일분류기의 활용이 이루어진다. 장소예측에 관해서도 다양한 연구들이 진행되고 있다. 사용자와 친한 지인의 장소를 바탕으로 사용자의 장소를 예측하는 연구[9]와 모바일 환경에서 사용자의 장소를 예측한 연구[10]가 진행된 바 있다.

장소예측에 관한 연구는 장소 정보에만 제한되어 있으며 다양한 상황정보를 사용하는 것이 필요하다[11]. 장소예측을 위해 행동정보와 관계정보를 이용한 연구도 있다[12]. 그러나 이와 같은 연구는 거의 없다. 또한 기존의 장소예측에 관한 연구는 주로 과거의 정보를 통해 미래의 장소를 예측하는 것에 편중되어 있어[11] 과거 장소 예측에 관한 연구는 현재 까지 진행된 바가 없다. 따라서 본 연구는 이러한 한계를 극복하기 위해, 사용자의 장소정보와 인구통계학적 특성을 이용하여 사용자가 방문한 과거의 장소를 예측하였다.

기존 연구에서 단일분류기는 장소예측에 많이 활용되었다. 빌딩내의 장소 예측에 베이지안 네트워크와 인공신경망이 이미 활용된 바 있다[13]. 의사결정트리를 사용하여 장소를 예측한 연구도 있다[14]. 또한 단백질 세포위치의 예측에서 서포트벡터머신이 활용된 연구가 있다[15]. 따라서 본 연구는 다양한 단일분류기를 이용하여  과거 상황을 예측하였다.

본 연구에서 사용한 자료는 기존 위치 예측연구에서 다양하게 활용된 바 있다. 앙상블 방법을 통해 사용자의 위치를 예측한 연구[25]가 있으며, 유비쿼터스 환경에서 GBN(General Bayesian Network)을 기반으로 한 위치 예측모형을 구축을 하고[26,27] 그 예측모형을 평가한 연구[28]도 진행된 바 있다. 또한 다양한 베이지안 네트워크를 사용하여 사용자의 위치를 예측하는 연구[29], 의사결정트리를 사용해 위치 예측의 규칙을 찾아내는 연구[30]도 진행되었다. 동일한 자료를 사용한 위의 연구들과 달리  본 연구는 앙상블 방법이 아닌 다양한 분류기를 활용한다. 의사결정트리와 베이지안 네트워크뿐만 아니라 인공신경망, 서포트벡터머신을 사용한다. 또한 본 연구는 일원배치 분산분석을 통해 단일분류기의 예측력 차이를 비교한다.

## III. Proposed Schemes

### 3.1 Data and Variables

본 연구는 대학생들을 대상으로 수집한 6868개의 자료를 바탕으로 하였다. 자료는 나이, 성별, 전공 등 인구통계학적 특성에 따라 대학생들을 대상으로 약 한달간 수집하였다.

본 연구는 캠퍼스 내에서 학생이 머물렀던 과거의 건물을 예측하는 모델을 구축하고자 한다. 종속변수를 제외하고 인구통계학적 특성(성별, 나이, 전공, 학년, 군복무 상황, 흡연 여부, 종교, 연애관계, 전공, 주중, 주말 여가활동 등)과 지나간 경로와 도착한 건물 등 30개의 유의미한 독립변수를 선정하였다. 다음으로 특성추출을 이용하여 변수를 2차적으로 추출하였다. 이를 통해 과거 장소예측에 영향을 미치는 변수를 확인하였다. 그 다음, 분산분석을 통하여 특성추출 전과 후를 비교하였다. 마지막으로는 Tukey검정을 이용하여 단일 분류기들을 비교하였다. 아래 Table 1은 특성추출을 적용하기 전의 변수를 나타낸 표이다.

Table 1. Variables

| Variable | Definition |
|---|---|
| Day of Week | Day went to school |
| Arrival Time | Time when arrived at building |
| Depart Time | Time when departed from building |
| Path1 | Way passed firstly |
| Path2 | Way passed secondly |
| Path3 | Way passed at third time |
| Path4 | Way passed by fourth time |
| Path5 | Way passed by fifth time |
| Building Name | Building visited |
| Activity Code | An activity played in building |
| Gender | Gender |
| Age | Age |
| Major | Major |
| Grade | Grade |
| Military | Before and after of military service |
| Religion | Religion believing |
| Monthly Allowance | Monthly allowance by parents |
| Smoking | Smoking or not smoking |
| Lover | Has a lover or not |
| Weekday Leisure | Type of leisure in weekday |
| Weekend Leisure | Type of leisure in weekend |
| Vacation Leisure | Type of leisure at vacation |
| Lunch Leisure | Type of leisure at lunch time |
| Leisure Utility | The use of leisure |
| Leisure Satisfaction | Satisfaction of leisure |
| Housing | Housing type |
| Arrival Transportation | Transportation use when go to school |
| Departure Transportation | Transportation use when leave school |
| Average Study Time | usual study time without lecture and period of exam |
| Monthly Mobile Phone Fee | Monthly mobile phone fee |

#### 3.1.1 Feature Selection (FS)

특성추출은 선택한 타겟변수에 대해 적절한 특성을 추출해준다. 특성추출은 분류모델의 성과를 결정하기 때문에 매우 중

요하다. 특성추출에 관한 연구는 두 가지 방향으로 진행된다. 하나는 새로운 방법의 개발이고 다른 하나는 이미 존재하는 방법 간의 비교이다. 그러나 모든 자료에 적합한 방법을 개발하는 것은 어려움이 따른다.

Table 2. Variables arranges by feature selection

| Rank | Variable | Correlation |
|---|---|---|
| 1 | Grade | 0.148 |
| 2 | Age | 0.128 |
| 3 | Path1 | 0.093 |
| 4 | Average study time | 0.092 |
| 5 | Military | 0.089 |
| 6 | Building Name | 0.085 |
| 7 | Housing | 0.078 |
| 8 | Path3 | 0.069 |
| 9 | Arrival Transportation | 0.065 |
| 10 | Path4 | 0.06 |
| 11 | Monthly Mobile Phone Fee | 0.059 |
| 12 | Path2 | 0.057 |
| 13 | Lover | 0.05 |
| 14 | Major | 0.05 |
| 15 | Departure Transportation | 0.047 |
| 16 | Path5 | 0.045 |
| 17 | Smoking | 0.042 |
| 18 | Religion | 0.031 |
| 19 | Weekday Leisure | 0.027 |
| 20 | Weekend Leisure | 0.026 |
| 21 | Leisure Utility | 0.024 |
| 22 | Leisure Satisfaction | 0.024 |
| 23 | Gender | 0.024 |
| 24 | Activity Code | 0.022 |
| 25 | Day Of Week | 0.021 |

어떠한 방법도 모든 상황에서 적용되지 않는다. 따라서 기존의 방법들을 비교하는 특성추출 연구가 많이 이루어진다. 또한 상황에 적합한 특성추출 방법을 찾는 것에 대한 관심도 늘어나고 있다.

본 연구는 웨카(WEKA)소프트웨어를 이용하여, 상관속성평가법(Correlation attribute evaluation)에 의한 특성추출을 적용하였다. 속성 검색 방법으로는 랭커(Ranker)를 사용했다. 특성추출을 적용한 결과 선택된 변수는 Table 2에 제시된 상위 총 17개 항목이다.

## 3.2 Classifier

본 연구는 베이지안 네트워크, 의사결정트리, 인공신경망, 서포트 벡터 머신을 사용하였다. 위의 4가지 데이터 마이닝 기법은 모두 단일분류기다. 모든 분석 방법은 10겹 교차타당성 검증(10-fold Cross Validation)방식으로 시행한다.

### 3.2.1 Single Classifier

#### 베이지안 네트워크(Bayesian Network)

베이지안 네트워크는 불확실성이 있는 문제를 해결하기 위해 사용된다. 베이지안 네트워크의 종류로는 가장 일반적인 형태인 GBN(General Bayesian Network)과 단순한 형태인 NBN(Naïve Bayesian Network), 그리고 확장된 형태인 TAN(Tree Augmented Naïve Bayesian Network)이 있다.

NBN은 단순하고 효율적이다. NBN은 조건 확률의 정확성에 의해 좌우되고 노드들이 상황 내에서 서로 독립이라는 가정을 한다. 하지만 NBN의 정확성은 독립 가정이 위반 되면 낮아진다. 또한 자료들이 거의 없을 때, NBN은 정확하게 측정되기 어렵다[16].

TAN은 NBN의 특성에 연결선들을 추가함으로써 NBN을 확장하여 상호관계를 베이지안 네트워크로 표현한다[16]. TAN에서는 모든 요소들이 독립적이지 않고 상호의존적이다.

베이지안 네트워크 중 가장 일반적인 GBN은 타겟 노드에 해당하는 변수를 일반 노드와 동일시한다. GBN에서는 타겟 노드도 부모 노드들을 가질 수 있기 때문에 변수 간의 인과관계를 자연스럽게 표시할 수 있다.

모든 변수를 동일하게 취급하는 GBN의 종류로는 크게 두 가지가 있다. 첫 번째는 GBN-K2이고 두 번째는 GBN-HC (Hill-Climber)이다. K2는 변수의 부분집합에서 최적의 모집합을 찾는다[17]. K2는 한 노드에서 다른 노드로 선을 뻗어 가며 단계마다 검증력을 높인다. 또한, 변수의 순서에 의해 제한을 받는다. 한편, HC(Hill-Climber)는 비어있는 그래프에서 시작하여 모든 경로(Arc)의 추가 또는 삭제마다 각각 다른 베이지안 네트워크를 구성한다. 또한 HC는 변수의 순서에 의해 제한을 받지 않아 비교적 자유롭다. 따라서 모든 단계에서 HC는 가장 효율적인 방법을 선택한다.

다양한 분야의 미래를 예측하는데 베이지안 네트워크(Bayesian Network)는 효과적으로 활용되고 있다. 소프트웨어 실패율 예측[18], 내부장소 예측[19]과 같은 분야에서 베이지안 네트워크는 성공적인 방법을 제시한다.

#### 의사결정트리(Decision Tree: DT)

의사결정트리는 데이터마이닝에서 가장 유명한 방법 중의 하나이다. 의사결정트리의 결과는 의사결정트리 구축의 효율성에 따라 달라진다. 의사결정트리는 두 가지 단계로 이루어져 있다. 첫 번째는 성장 단계이고 두 번째는 가지치기 단계이다. 성장 단계는 순차적으로 2개 혹은 3개의 부분집합으로 데이터를 나누는 과정이다. 가지치기 단계에서는 다양한 방법들이 의사결정트리의 복잡성을 줄이기 위해 사용된다. 의사결정트리는 데이터와 패턴을 분석하는 다양한 분야에서 효과적인 도구로 사용되고 있다[20]. 본 연구는 의사결정트리 방법 중 의사결정 분야에서 널리 쓰이고 있는 C4.5를 사용하여 과거 장소를 예측하였다. C4.5는 하나의 뿌리에서 다수의 가짓수를 가진다. 또한 하나의 하위트리에서 또 다른 하위트리로 가지치기가 가능하기 때문에 실패율이 낮다.

#### 인공신경망(Neural Network: NN)

인공신경망 방법은 패턴인식을 할 때 주로 사용되는 비선형적 알고리즘이다. 인공신경망은 생물의 신경망에서 영감을 얻은 것이다. 인공신경망은 각 입력변수에 대응되는 마디를 갖는 입력층(Input layer), 여러 개의 은닉 마디를 갖는 은닉층

(Hidden layer) 그리고 목표변수에 대응하는 마디를 갖는 출력층(Output layer)으로 구성되어있다. 인공신경망은 브랜드 포지셔닝과 같은 마케팅 분야에서도 활용되고 있다[21]. 본 연구는 인공신경망 중 다중퍼셉트론(Multilayer Perceptron)을 이용한다. 다중퍼셉트론은 가장 유명하고 널리 쓰이는 인공신경망이다. 다중퍼셉트론은 각각 하나의 입력층과 출력층으로 구성되어 있으며 두개의 층 사이에 하나 이상의 은닉층이 있다. 입력층과 달리, 은닉층과 출력층은 데이터를 처리하고 출력층이 인공신경망 방법의 결과를 나타낸다.

**서포트 벡터 머신(Support Vector Machine: SVM)**

서포트 벡터 머신은 비교적 최근에 만들어졌으며 특성추출에서 일반적으로 사용되는 분류기다. 예측 모델을 형성할 때, 서포트 벡터 머신과 인공신경망은 비슷한 구조를 가지고 있다. 그러나 서포트 벡터 마신과 인공신경망의 차이점은 다음과 같다. 인공신경망은 비선형 문제를 해결하기 위해서 다층 연결을 사용한다. 하지만 서포트 벡터 머신은 비선형 문제를 해결하기 위해 비선형 함수를 사용한다. N개의 클래스가 있다면, 서포트 벡터 머신은 비선형 함수를 사용해 모든 클래스에 대해 학습한다. 단일 서포트 벡터 머신 모델이 분류에서 더 좋은 성과를 낼 수 있지만, 단일 서포트 벡터 머신 모델은 클래스의 설정에 따라 변하기 쉽다. 따라서 최근에는 다중 클래스 문제들을 위해 서포트 벡터 머신 중 다중 이진 분류기를 널리 이용한다[22]. 본 연구는 다중 클래스 문제들이 존재하므로 WEKA의 다중 이진 분류기인 SMO를 사용하였다.

### 3.3 Evaluation Criteria

ROC 곡선 밑의 영역으로 알려져 있는 AUC는 예측 모델의 정확도를 평가한다. ROC 곡선은 기준치가 모두 0과 1사이에서 구성된다. 만약에 AUC가 0.5보다 높으면 차별성이 있다. 또한 AUC가 0.5이면 불완전하고 차별성이 없다. AUC의 장점은 대상들을 나눌 수 있는 기준치가 있다는 점이다. AUC에서는 0.5가 기준치이다[23]. 본 연구는 단일분류기의 AUC를 비교하였다. 특성추출 후의 단일분류기 AUC를 비교한 그래프는 그림 1과 같다.

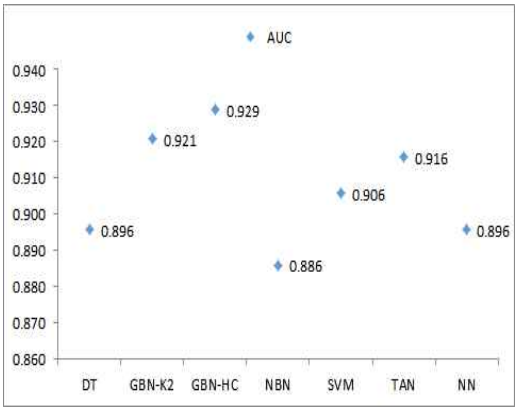



Fig. 1. AUC per single classifiers after FS

Accuracy는 측정할 때마다 기대되는 결과를 얻는 정도이다. 측정치 평균의 분포가 참의 값과 차이가 없으면 Accuracy가 높다. Precision은 반복된 측정치가 분포되는 정도이다. 분포의 정도가 작을수록 Precision은 높다. Kappa통계치는 동일한 정보에 대한 평가 일치 확률을 측정한다. Kappa통계치는 -1부터 1(완전일치) 사이의 값을 가진다. TPR(True Positive Rate)은 실제값이 참일 때 예측값이 참일 확률이다. FPR(False Positive Rate)은 실제값이 거짓일 때 예측값이 참일 확률이다. F-measure는 Precision과 TPR을 합쳐 정확도를 나타낸다. F-measure은 0과 1사이의 값을 가진다. 본 연구에서는 분류기의 성과비교를 위하여 이상과 같은 7가지 성과변수를 사용하였다. 성과변수에 대한 식은 다음과 같다[24].

## IV. Results

상관관계 기반에 의한 특성추출을 이용하여 30개의 독립 변수중 타겟변수를 의미있게 설명하는 특성을 추출하였다. 그 결과, Correlation의 값이 상대적으로 구분이 되고 포함된 속성의 수가 적당한 구간은 0.042에서 0.031로 낮아지는 구간으로 선택하였다. 이렇게 하여 선택된 변수, 즉 특성은 모두 17개이며 이는 Table 2에 정리되어 있다.

$$\text{Accuracy}: \frac{TP+TN}{N} \quad (1)$$

$$\text{Precision}: \frac{TP}{TP+FP} \quad (2)$$

$$\text{Kappa통계치}: \frac{p_0-p_e}{1-p_e} \quad (3)$$

*$p_o$ : 평가의 일치 확률, $p_e$ : 평가가 확률적으로 일치할 확률

$$\text{TPR}: \frac{TP}{TP+FN} \quad (4)$$

$$\text{FPR}: \frac{FP}{FP+TN} \quad (5)$$

$$\text{F-measure}: \frac{2\times Precision\times Recall}{Precision+Recall} \quad (6)$$

위 Table 1 그리고 Table 2에서 알 수 있듯이, 특성추출을 이용하여 변수를 다시 선정한 결과 그 수가 30개에서 17개로 줄어 감소함을 알 수 있다. 다음으로 특성추출 전과 후 변수의 예측력을 비교하였다.

Table 3에는 특성추출 전과 후의 Accuracy와 AUC가 나타나있다. GBN-HC의 특성추출 전후 Accuracy와 AUC의 수치가 가장 높다. Table 3을 보면, 특성추출 전의 Accuracy보다 특성추출 후의 Accuracy의 수치가 모든 단일분류기에서 훨씬 높음을 알 수 있다. 또한 DT을 제외하고 모든 특성추출 후의 단일분류기의 AUC가 더 높다는 것을 알 수 있다. 특성추출을

Table 3. Comparison between before FS and after FS

(FS: Feature Selection)

| Single classifier(Before FS) | | | Single classifier(After FS) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Name | Accuracy (%) | AUC | Name | Accuracy (%) | Precision (%) | Kappa | AUC | TPR | FPR | F-measure |
| DT | 73.1509 | 0.907 | DT | 74.3114 | 57.9245 | 0.4938 | 0.896 | 0.5747 | 0.0826 | 0.5569 |
| GBN-K2 | 66.5685 | 0.901 | GBN-K2 | 71.7824 | 70.9927 | 0.6793 | 0.921 | 0.7308 | 0.0571 | 0.7135 |
| GBN-HC | 74.1382 | 0.928 | GBN-HC | 74.4847 | 72.7464 | 0.6991 | 0.929 | 0.7463 | 0.0474 | 0.7305 |
| NBN | 59.0161 | 0.878 | NBN | 60.1420 | 66.8849 | 0.6253 | 0.886 | 0.6849 | 0.0630 | 0.6700 |
| SVM | 68.4046 | 0.893 | SVM | 71.2848 | 71.4131 | 0.6834 | 0.906 | 0.7325 | 0.0500 | 0.7116 |
| TAN | 64.3166 | 0.891 | TAN | 70.9149 | 66.6191 | 0.6364 | 0.916 | 0.6954 | 0.0626 | 0.6717 |
| NN | 61.0821 | 0.858 | NN | 71.0203 | 70 .6005 | 0.6594 | 0.896 | 0.7105 | 0.0577 | 0.7032 |

이용하여 변수를 선정한 결과 Accuracy와 AUC의 수치가 상대적으로 높아졌다. 따라서 특성추출 후의 결과가 예측력이 상대적으로 높다는 것을 알 수 있다.

본 연구는 특성추출 전후의 단일분류기 예측력 차이를 검증하기 위해 일원배치 분산분석을 이용하였다. 일원배치 분산분석으로 특성추출 전후 단일분류기의 Accuracy, Precision, Kappa통계치, AUC, TPR, FPR 그리고 F-measure를 비교한다. Table 4는 일원배치 분산분석으로 특성추출 전과 후 전체 단일분류기의 차이를 비교한 결과이다. 95%의 신뢰수준에서 AUC의 유의확률이 0.277이고 Accuracy, Precision, Kappa통계치, TPR, FPR 그리고 F-measure의 유의확률은 모두 0.000으로 0.05보다 낮다. 따라서 AUC를 제외한 6개의 측면에서 특성추출 전의 단일분류기와 특성추출 후의 단일분류기가 차이가 있다. Table 3을 보면 Accuracy의 측면에서 특성추출 전보다 특성추출 후의 값이 전체적으로 높은 것을 알 수 있다. 또한 특성추출 전과 후의 AUC 차이는 무의미하다. 따라서 특성추출을 활용하면 실제 분류에 필요한 정보의 수를 감소시키면서 예측력을 유지 또는 상승시킬 수 있다. 다음으로 가장 예측력이 우수한 알고리즘을 개발하기 위하여 특성추출 후의 단일분류기들을 95%의 신뢰수준에서 일원배치 분산분석과 Tukey검정을 이용하여 비교한다.

Table 4의 일원배치 분산분석 결과, Accuracy, Precision, Kappa통계치, TPR, FPR 그리고 F-measure의 유의확률이 모두 0.000으로 0.05보다 낮음을 알 수 있다. 따라서 모든 측면에서 단일분류기 사이의 차이가 있기 때문에 Tukey검정을 이용하여 단일분류기를 비교한다. Tukey검정의 결과가 Table 5에 나타나 있다. Table 3에서 GBN-HC가 모든 측면에서 가장 높았기 때문에 GBN-HC를 중심으로 비교한다. 우선, Accuracy의 측면에서 GBN-HC가 가장 높다. 그 다음으로는 DT과 GBN-K2가 높았고 NBN이 가장 낮다.

Table 4. ANOVA-test Results

| F-test between before and after FS | | F-test after FS | |
|---|---|---|---|
| Name | P-value | Name | P-value |
| Accuracy | 0.000 | Accuracy | 0.000 |
| Precision | 0.000 | Precision | 0.000 |
| Kappa | 0.000 | Kappa | 0.000 |
| AUC | 0.277 | AUC | 0.000 |
| TPR | 0.000 | TPR | 0.000 |
| FPR | 0.000 | FPR | 0.000 |
| F-measure | 0.000 | F-measure | 0.000 |

또한 Table 5를 보면, GBN-HC는 TAN, DT, NBN, NN와 유의미한 차이를 보인다. Precision의 측면에서도 GBN-HC가 가장 높았으며 다음이 SVM, GBN-K2 순서이다. GBN-HC의 Precision은 TAN, DT 그리고 NBN과 유의미한 차이를 보인다. Kappa통계치와 AUC에서는 GBN-HC가 가장 높고 다음이 GBN-K2이다. GBN-HC의 TPR과 F-measure의 수치가 가장 높고, SVM, GBN-K2의 순서이다.

GBN-HC는 Kappa통계치, TPR 그리고 F-measure의 측면에서 TAN, DT, NBN, NN 그리고 AUC의 측면에서 TAN, DT, SVM, NBN, NN와 유의미한 차이를 보인다. 낮을수록 좋은 FPR에서도 GBN-HC가 가장 낮으며 SVM, GBN-K2가 다음으로 낮다. FPR에서 GBN-HC는 GBN-K2, TAN, DT 그리고 NBN과 유의미한 차이를 보인다.

Table 5. Tukey test results based on GBN-HC

| Name | Classifier 1 | Classifier 2 | Mean Difference (I-J) | P-value |
|---|---|---|---|---|
| Accuracy | GBN-HC | GBN-K2 | 0.0155 | 0.444 |
| | | TAN | 0.0509 | 0.000 |
| | | DT | 0.1716 | 0.000 |
| | | SVM | 0.0138 | 0.585 |
| | | NBN | 0.0614 | 0.000 |
| | | NN | 0.0358 | 0.001 |
| Precision | GBN-HC | GBN-K2 | 0.0175 | 0.458 |
| | | TAN | 0.0612 | 0.000 |
| | | DT | 0.1482 | 0.000 |
| | | SVM | 0.1333 | 0.755 |
| | | NBN | 0.0586 | 0.000 |
| | | NN | 0.0214 | 0.223 |
| Kappa | GBN-HC | GBN-K2 | 0.0205 | 0.325 |
| | | TAN | 0.0634 | 0.000 |
| | | DT | 0.2061 | 0.000 |
| | | SVM | 0.0164 | 0.593 |
| | | NBN | 0.0746 | 0.000 |
| | | NN | 0.0404 | 0.001 |
| AUC | GBN-HC | GBN-K2 | 0.0104 | 0.161 |
| | | TAN | 0.0207 | 0.000 |
| | | DT | 0.0656 | 0.000 |
| | | SVM | 0.0201 | 0.000 |
| | | NBN | 0.0332 | 0.000 |
| | | NN | 0.0372 | 0.000 |
| TPR | GBN-HC | GBN-K2 | 0.0155 | 0.444 |
| | | TAN | 0.0509 | 0.000 |
| | | DT | 0.1716 | 0.000 |
| | | SVM | 0.0138 | 0.585 |
| | | NBN | 0.0614 | 0.000 |
| | | NN | 0.0358 | 0.001 |

| | | | | |
|---|---|---|---|---|
| FPR | GBN-HC | GBN-K2 | -0.0096 | 0.008 |
| | | TAN | -0.0151 | 0.000 |
| | | DT | -0.0351 | 0.000 |
| | | SVM | -0.0125 | 0.956 |
| | | NBN | -0.0155 | 0.000 |
| | | NN | -0.0002 | 1.000 |
| F-measure | GBN-HC | GBN-K2 | 0.0170 | 0.396 |
| | | TAN | 0.0587 | 0.000 |
| | | DT | 0.1735 | 0.000 |
| | | SVM | 0.0188 | 0.276 |
| | | NBN | 0.0604 | 0.000 |
| | | NN | 0.0272 | 0.027 |

본 연구에서 분석한 결과 Accuracy는 GBN-HC, DT, GBN-K2, 그리고 SVM의 순서이다. 또한 AUC의 측면에서는 GBN-HC, GBN-K2, TAN, 그리고 SVM의 순서이다. Precision, TPR, FPR 그리고 F-measure의 수치는 GBN-HC, SVM, GBN-K2의 순서이다. Kappa통계치는 GBN-HC가 가장 높고 두 번째로는 GBN-K2인 것을 알 수 있다. 위의 결과들을 바탕으로 GBN-HC와 GBN-K2가 예측력이 가장 높은 세 개의 기법 중에 항상 포함되어 있다는 것을 알 수 있다. 따라서, 과거 상황예측 시 GBN이 매우 효과적이다.

Table 6. Best 3 classifiers based on analysis

| Name | Comparison after FS | Value |
|---|---|---|
| Accuracy(%) | GBN-HC | 74.4847 |
| | DT | 74.3114 |
| | GBN-K2 | 71.7824 |
| Precision(%) | GBN-HC | 72.7464 |
| | SVM | 71.4131 |
| | GBN-K2 | 70.9927 |
| Kappa | GBN-HC | 0.6991 |
| | SVM | 0.6834 |
| | GBN-K2 | 0.6793 |
| AUC | GBN-HC | 0.929 |
| | GBN-K2 | 0.921 |
| | TAN | 0.916 |
| TPR | GBN-HC | 0.7463 |
| | SVM | 0.7325 |
| | GBN-K2 | 0.7308 |
| FPR(low) | GBN-HC | 0.0474 |
| | SVM | 0.0500 |
| | GBN-K2 | 0.0571 |
| F-measure | GBN-HC | 0.7305 |
| | GBN-K2 | 0.7135 |
| | SVM | 0.7032 |

## V. Concluding Remarks

본 연구는 기존의 연구와 달리 사용자가 지나온 경로, 방문한 건물 그리고 인구통계정보를 활용하여 사용자의 과거 장소를 예측하였다. Accuracy, Precision, Kappa통계치, AUC, TPR, FPR 그리고 F-measure의 결과값 비교, ANOVA test와 Tukey test의 결과를 통해 본 연구는 기존의 연구결과에서 많이 언급된 GBN-HC와 GBN-K2의 예측력이 여타 단일 분류기보다 상대적으로 우수하다는 것을 통계적으로 검증하였다.

본 연구관련 향후 연구주제는 다음과 같다. 첫째, SNS정보를 과거 장소예측에 활용할 수 있다. SNS정보에는 사용자의 감정 그리고 개인적인 판단이 포함되어 있어 과거 장소예측 시 효과적이기 때문이다. 둘째, 텍스트 마이닝(Text Mininxg)을 활용하여 사용자의 감성분석(Sentiment analysis)을 과거 장소예측 시 적용할 수 있다. 사용자의 감성에 따라 방문하는 장소가 크게 좌우될 수 있기 때문이다.

본 연구에서 채택한 방법론과 결과가 빅데이터의 활용도가 증가되고 있는 시점에서 다양한 분야들, 예를 들어 기업의 고객관리와 헬스케어(Health Care) 분야에 좋은 선행연구가 되기를 기대한다.

## REFERENCES

[1] Da Rosa, J. H., Barbosa, J. L., and Ribeiro, G. D., "ORACON: An adaptive model for context prediction," Expert Systems with Applications, Vol.45, pp.56-70, 2016.

[2] Scott, J., Bernheim Brush, A. J., Krumm, J., Meyers, B., Hazas, M., Hodges, S., and Villar, N., "PreHeat: Controlling home heating using occupancy prediction," ACM, pp. 281-290, Beijing, China, Sep. 2011.

[3] Sigg, S., Gordon, D., von Zengen, G., Beigl, M., Haseloff, S., and David, K., "Investigation of context prediction accuracy for different context abstraction levels," IEEE Transactions on Mobile Computing, Vol.11, No.6, pp.1047-1059, 2012.

[4] Cohen, M. L., "Changing patterns of infectious disease," Nature, Vol.406, pp. 762-767, 2000.

[5] Bernasco, W., "A sentimental journey to crime: Effects of residential history on crime location choice," Criminology, Vol.48, No.2, pp.389-416, 2010.

[6] Cha, M. Q., Jung, D. K., Kim, Y. K., and Chong, H. J., "A USN Based Mobile Object Tracking System for the Prevention of Missing Child," Journal of KIISE: Information Networking, Vol.35, No.5, pp.453-463, 2008.

[7] Rather, A. M., Agarwal, A., and Sastry, V. N., "Recurrent neural network and a hybrid model for prediction of stock returns," Expert Systems with Applications, Vol.42, No.6, pp.3234-3241, 2015.

[8] Bala, A., and Chana, I., "Intelligent failure prediction models for scientific workflows," Expert Systems with Applications, Vol.42, No.3, pp.980-989, 2015.

[9] Gong, Y., Li, Y., Jin, D., Su, L., and Zeng, L., "A location prediction scheme based on social correlation," IEEE, pp.1-5, Budapest, Hungary, May. 2011.

[10] Yavaş, G., Katsaros, D., Ulusoy, Ö., and Manolopoulos, Y., "A data mining approach for location prediction in mobile environments," Data and Knowledge Engineering, Vol.54, No.2, pp.121-146, 2005.

[11] David, K., Kusber, R., Lau, S. L., Sigg, S., and Ziebart, B., "3rd workshop on recent advances in behavior prediction and pro-active pervasive computing," ACM, pp. 415-420, Seattle, United States, Sep. 2014.

[12] Gao, H., Tang, J., and Liu, H., "Exploring social-historical ties on location-based social networks," In International Conference on Web and Social Media, June. 2012.

[13] Petzold, J., Bagci, F., Trumler, W., and Ungerer, T., "Next location prediction within a smart office building," Cognitive Science Research Paper-University of Sussex, Vol.577, 2005.

[14] Ying, J. J. C., Lee, W. C., Weng, T. C., and Tseng, V. S., "Semantic trajectory mining for location prediction," ACM, pp.34-43, Chicago, United States, Nov. 2011.

[15] Hua, S., and Sun, Z., "Support vector machine approach for protein subcellular localization prediction," Bioinformatics, Vol.17, No.8, pp.721-728, 2001.

[16] Diab, D. M., and El Hindi, K. M., "Using Differential Evolution for Fine Tuning Naïve Bayesian Classifiers and its Application for Text Classification," Applied Soft Computing, Vol.28 pp.1-60, December 2016.

[17] Liu, H., Zhou, S., Lam, W., and Guan, J., "A New Hybrid Method for Learning Bayesian Networks: Separation and Reunion," Knowledge-Based Systems, 2017.

[18] Bai, C. G., Hu, Q. P., Xie, M., and Ng, S. H., "Software failure prediction based on a Markov Bayesian network model," Journal of Systems and Software, Vol.74, No.3, pp.275-282, 2005.

[19] Petzold, J., Pietzowski, A., Bagci, F., Trumler, W., and Ungerer, T., "Prediction of indoor movements using Bayesian networks," ACM, pp.211-222, Oberpfaffenhofen, Germany May. 2005.

[20] Lee, S., "Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls," Decision Support Systems, Vol.49, No.4, pp.486-497, 2010

[21] Ferilli, G., Sacco, P. L., Teti, E., and Buscema, M., "Top corporate brands and the global structure of country brand positioning: An AutoCM ANN approach," Expert Systems with Applications, Vol.66, pp.62-75, 2016.

[22] Zhou, L., Lai, K. K., and Yu, L., "Least squares support vector machines ensemble models for credit scoring," Expert Systems with Applications, Vol.37, No.1, pp.127-133, 2010.

[23] Ballings, M., and Van den Poel, D., "Kernel Factory: An ensemble of kernel machines," Expert Systems with Applications, Vol.40, No.8, pp.2904-2913, 2013.

[24] Liu, X. Y., Wu, J., and Zhou, Z. H., "Exploratory undersampling for class-imbalance learning," IEEE Transactions on Systems, Man, and Cybernetics, Vol.39, pp.539-550, 2009.

[25] Lee, K. C., and Cho, H., "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, No.3, 2010.

[26] Lee, K. C., and Cho, H., "Integration of general Bayesian network and ubiquitous decision support to provide context prediction capability," Expert Systems with Applications, Vol.39, No.5, pp.6116-6121, 2012.

[27] Lee, S., and Lee, K. C., "Context-prediction performance by a dynamic bayesian network: Emphasis on location prediction in ubiquitous decision support environment," Expert Systems with Applications, Vol.39, No.5, pp.4908-4914, 2012.

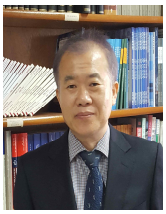
[28] Lee, K. C., and Cho, H., "Designing a Ubiquitous Decision Support Engine for Context Prediction: General Bayesian Network Approach," International Journal of u-and e-Service, Science and Technology, Vol.3, No.3, pp.25-36, 2010.

[29] Lee, S., Lee, K. C., and Cho, H., "A dynamic Bayesian network approach to location prediction in ubiquitous computing environments," In International Conference on Advances in Information Technology, pp.73-82, Springer Berlin Heidelberg, Germany, Nov. 2010.

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

## Authors

Eun Min Lee is a student in Global Business Administration at Sungkyunkwan University, Korea. He is interested in big data-mining based location prediction, cloud computing, marketing strategy and artificial intelligence.



Kun Chang Lee is a distinguished professor in SKK Business School at Sungkyunkwan University. He is now in charge of Creativity Science Research Institute (CSRI) and Health Mining Research Center (HMRC) as well. He holds a joint appointment professor in SAIHST (Samsung Advanced Institute for Health Sciences & Technology), Sungkyunkwan University. His recent research interests lie in data mining, health informatics, creativity science, Human-Robot Interaction (HRI), and artificial intelligence techniques in decision making analysis.