http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.21, No.1, February 2016, pp.147-163
http://dx.doi.org/10.7838/jsebs.2016.21.1.147

스마트 기기의 멀티 모달 로그 데이터를 이용한 사용자 성별 예측 기법 연구

A Study on Method for User Gender Prediction Using Multi-Modal Smart Device Log Data

김윤정(Yoonjung Kim)*, 최예림(Yerim Choi)**, 김소이(Solee Kim)***,
박규연(Kyuyon Park)****, 박종헌(Jonghun Park)*****

초 록

스마트 기기 사용자의 성별 정보는 성공적인 개인화 서비스를 위해 중요하며, 스마트 기기로부터 수집된 멀티 모달 로그 데이터는 사용자의 성별 예측에 중요한 근거가 된다. 하지만 각 멀티 모달 데이터의 특성에 따라 다른 방식으로 성별 예측을 수행해야 한다. 따라서 본 연구에서는 스마트 기기로부터 발생한 로그 데이터 중 텍스트, 어플리케이션, 가속도 데이터에 기반한 각기 다른 분류기의 예측 결과를 다수결 방식으로 앙상블하여 최종 성별을 예측하는 기법을 제안한다. 텍스트 데이터를 이용한 분류기는 데이터 유출에 의한 사생활 침해 문제를 최소화하기 위해 웹 문서로부터 각 성별의 특징적 단어 집합을 도출하고 이를 기기로 전송하여 사용자의 기기 내에서 성별 분류를 수행한다. 어플리케이션 데이터에 기반한 분류기는 사용자가 실행한 어플리케이션들에 성별을 부여하고 높은 비율을 차지하는 성별로 사용자의 성별을 예측한다. 가속도 기반 분류기는 성별에 따른 사용자의 가속도 데이터 인스턴스를 학습한 SVM 모델을 사용하여 주어진 성별을 분류한다. 자체 제작한 안드로이드 어플리케이션을 통해 수집된 실제 스마트 기기 로그 데이터를 사용하여 제안하는 기법을 평가하였으며 그 결과 높은 예측 성능을 보였다.

ABSTRACT

Gender information of a smart device user is essential to provide personalized services, and multi-modal data obtained from the device is useful for predicting the gender of the user. However, the method for utilizing each of the multi-modal data for gender prediction differs according to the characteristics of the data. Therefore, in this study, an ensemble method for predicting the gender of a smart device user by using three classifiers that have text, application, and acceleration data as inputs, respectively, is proposed. To alleviate

본 연구는 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2013R1A2A2A03013947).

* First Author, Department of Industrial Engineering, Seoul National University(yoonj625@gmail.com)
** Corresponding Author, Department of Industrial Engineering, Seoul National University(iangoozh@gmail.com)
*** Co-Author, Department of Industrial Engineering, Seoul National University(kpsinw@gmail.com)
**** Co-Author, Department of Industrial Engineering, Seoul National University(mysnuky91@snu.ac.kr)
***** Co-Author, Department of Industrial Engineering, Seoul National University(jonghun@snu.ac.kr)
Received: 2016-01-12, Review completed: 2016-02-12, Accepted: 2016-02-19

privacy issues that occur when text data generated in a smart device are sent outside, a classification method which scans smart device text data only on the device and classifies the gender of the user by matching text data with predefined sets of word. An application based classifier assigns gender labels to executed applications and predicts gender of the user by comparing the label ratio. Acceleration data is used with Support Vector Machine to classify user gender. The proposed method was evaluated by using the actual smart device log data collected from an Android application. The experimental results showed that the proposed method outperformed the compared methods.

키워드 : 성별 예측, 스마트 기기 로그 데이터, 앙상블 기법, 통계학습, 멀티 모달 데이터
Gender Prediction, Smart Device Log Data, Ensemble Method, Statistical Learning, Multi-Modal Data

1. 서 론

스마트 기기 사용자의 수가 늘어남에 따라 사용자 특성에 기반한 개인화 서비스의 수요가 증가하고 있다[8]. 사용자의 성별을 포함한 인구통계학적 정보는 개인화된 서비스를 제공하기 위해서 고려되어야 할 가장 기초적이고 중요한 정보이다. 일례로 모바일 전자상거래 분야에서는 사용자의 인구통계학적 정보와 관심사 등을 반영한 개인화된 광고를 제공해야 광고 반응률과 효율성을 높일 수 있다[5].

스마트 기기에서 수집할 수 있는 로그 데이터는 텍스트와 어플리케이션, 가속도를 비롯하여 15가지 이상으로 매우 다양하다[14]. 이러한 다양성을 고려하여 스마트 기기로부터 수집 가능한 여러 모달리티의 데이터를 사용하면 성별 예측의 성능을 극대화할 수 있을 것이다. 그러나 각각의 데이터 특성에 따라 처리방식이 상이하다는 문제점을 가진다. 사생활 보호 문제로 인하여 텍스트 데이터는 기기 밖으로의 유출이 지양되고 기기 내에서만 처리되는 것이 바람직하다. 반면 어플리케이션과 가속도 데이터는 사생활 침해 문제로부터 자유롭지만 해당 데이터를 처리하기 위한 계산의 복잡도가 매우 높으므로 기기 내에서의 분석이 제한적이다. 또한 어플리케이션 데이터는 문자 형태지만 가속도 데이터는 숫자 형태이기 때문에 두 종류의 데이터를 개별적으로 다루어야 정보의 소실을 최소화할 수 있다.

이러한 문제점은 앙상블 기법으로 해결이 가능하다. 앙상블 기법은 서로 다른 분류기로 분류된 결과를 종합하여 최종적인 결과를 도출하는 방식이다. 앙상블을 통해 얻어지는 분류 결과는 개별 분류기의 결과 더 높은 성능을 보임이 수리적으로 증명되었다[13]. 또한 Woźniak et al.[23]과 Zenobi and Cunningham[25]에 의하면 분류기들의 모델과 학습 방법, 입력 데이터가 다르면 분류기의 다양성이 보장되어 상호종속도가 낮은 결과를 도출하게 되고 이렇게 생성된 앙상블 분류기는 낮은 일반화 오류를 갖는다. 특히 서로 다른 입력데이터를 가지는 개별적인 분류기를 구축할 수 있기 때문에 데이터 처리 공간과 방식의 차이에도 불구하고 각각의 특성을 고려한 예측이 가능하다[21].

따라서 본 연구에서는 스마트 기기 로그 데이터 중 사용자 성별 예측에 개별적으로 사용

되어왔던 텍스트, 어플리케이션, 가속도 데이터로 사용자의 성별을 예측하는 분류기를 각각 구축하고 그 결과를 종합하여 최종적으로 성별을 결정하는 앙상블 기법을 제안한다. 텍스트 데이터에 기반한 분류기는 Kim et al.[12]에서 사용된 방법론을 사용하되, 유사도 계산을 위해서 카이 제곱 통계량 점수와 일치 빈도를 적용한다. 어플리케이션 데이터에 기반한 분류기는 Seneviratne et al.[19]과 같이 웹 상에서 얻어지는 정보를 추가적으로 이용하는 텍스트 마이닝 기법을 적용하여 구축한다. 가속도 데이터 기반 분류기는 Weiss and Lockhart[22]의 방법론을 바탕으로, 실제 로그 데이터의 특성을 반영한 요인을 추가하여 모델을 구축한다. 세 분류기의 분류 결과를 다수결하여 최종적으로 성별을 예측한다.

특히 어플리케이션 데이터 모델은 일반성(generality)를 보장할 수 있도록 독립적인 데이터로 학습된 모델을 사용한다. 기존의 연구[19]를 바탕으로 하되, 남녀를 구분하는 모델을 학습하기 위해 작성자의 성별이 명시되어 있는 웹 문서를 이용하고, 어플리케이션 설명글은 학습된 모델에 기반하여 어플리케이션 사용자의 성별을 예측하는 데에만 사용된다. 이러한 방법은 분류기 학습에 사용 가능한 사용자 데이터가 충분하지 않을 때 발생할 수 있는 과적합(overfitting)이나 편차(bias)를 방지하고 분류기의 일반성을 보장한다.

본 논문은 다음과 같은 순서로 구성된다. 제 2장에서는 스마트 기기 로그 데이터로 사용자의 성별을 예측한 기존 연구를 검토한다. 제 3장에서는 본 논문에서 제안하는 앙상블에 기반한 스마트 기기 사용자 성별 예측 기법을 자세히 서술하고 제 4장에서 실제 데이터에 제안 기법을 적용하여 사용자의 성별을 예측하고 결과를 분석한다. 마지막으로 제 5장에서는 결론 및 향후 연구 방향을 논의한다.

2. 관련 연구

스마트 기기에서 발생하는 로그 데이터를 이용하여 사용자의 성별을 예측하는 연구가 활발히 이루어져 왔다. 연구에 사용된 스마트 기기 로그 데이터로는 통화 기록, 문자 메시지와 문자 메시지의 수발신 기록, 위치 정보, 기기에 저장된 미디어 파일 정보, 가속도, 설치된 어플리케이션 정보 등이 있다. Nokia에서 주관한 Mobile Data Challenge[14]에서는 텍스트를 제외한 스마트 기기 로그 데이터를 공개하고 사용자 성별을 비롯한 인구통계학적 정보를 예측하는 문제가 제시되었으며 문제 해결을 위해 다양한 통계학습 방법론이 제안되었다[3, 17, 24, 26]. 또한 Weiss and Lockhart[22]는 주어진 실험환경에서 스마트 기기 사용자들의 가속도 데이터를 수집하고 평균, 표준편차, 최대, 최소 등의 통계량으로 요인화하여 사용자의 성별을 예측하였다. 기기에 설치된 어플리케이션 목록을 이용해서 사용자 성별을 분류하는 실험도 수행되었는데, 이 경우 웹에서 얻을 수 있는 관련 정보를 추가적으로 활용하여 부족한 정보를 보충하였다[19].

기기 내 텍스트 데이터는 성별 구분에는 유효하지만 개인정보보호 문제로 인해서 활용이 적었으나 최근 기기 내 분석(on-device analytics)에 기반한 연구가 증가하면서 성별 예측에 사용되기 시작했다. 문자 메시지와 인터넷 브라우저 검색 기록은 사용자의 성별에 따라

유의미한 차이를 보이는 것으로 알려져 있다[9, 10]. 그러나 이러한 데이터는 사용자의 개인정보를 포함하기 때문에 사용자들은 사생활 침해를 이유로 텍스트 데이터의 외부 유출을 꺼리는 경향을 보이며[2, 7, 16], 그로 인해 텍스트 데이터는 성별 예측에 사용하는 데에는 어려움이 존재한다. 이를 극복하기 위해 Kim et al.[12]에서는 텍스트 데이터가 기기 내부에서만 처리되는 성별 예측 프레임워크를 제시하여 사용자의 사생활 침해 문제를 방지하였다. 제안된 프레임워크에서는 웹에서 공개적으로 얻을 수 있는 텍스트 데이터를 사용하여 글쓴이의 성별에 따른 특징적 단어 집합을 생성하고 이를 개별 기기로 전송하여 기기 내 텍스트 데이터와의 일치 정도를 확인하여 사용자의 성별을 예측한다.

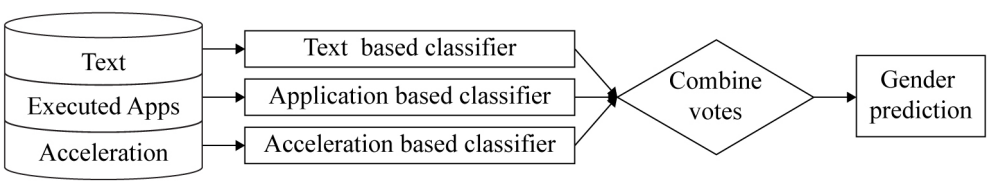
그러나 기존 연구들은 스마트 기기 로그 데이터의 멀티 모달리티 특성을 올바르게 활용하지 못했다는 한계점을 가진다. 많은 연구가 단일 데이터만을 사용하였으며 다양한 종류의 데이터를 사용한 경우에는 데이터를 처리하는 과정에서 각 데이터의 특성을 충분히 고려하지 않아 데이터의 손실이 발생하였다. 손실된 로그 데이터를 보충하기 위해 사용자에 대한 추가적인 인구통계학적 정보를 이용했는데, 이는 실제 상황에서 적용되기 어렵다는 문제점이 있다. 반면에 본 연구에서는 세 가지 개별 데이터의 특성이 반영된 방법론을 제시하며 로그 데이터 이외의 사용자 관련 데이터를 사용하지 않는다.

3. 제안 기법

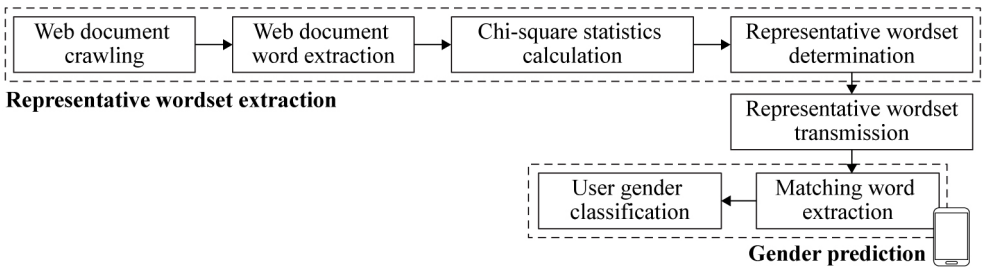
본 연구에서는 스마트 기기 사용자의 성별을 예측하기 위하여 서로 다른 데이터를 입력으로 가지는 분류기를 구출하고 각 분류기의 결과를 토대로 다수결(majority voting)에 의해 최종적으로 성별을 예측한다. <Figure 1>에 나타난 바와 같이 텍스트 데이터, 어플리케이션 데이터, 가속도 데이터를 각각의 입력으로 갖는 분류기들을 이용한다.

3.1 텍스트 기반 분류기

텍스트 기반 분류기는 <Figure 2>와 같이 특징적 단어 집합 추출 단계와 성별 예측 단계로 나뉜다. 단어 집합 추출의 모든 과정은 서버에서 수행된다. 웹 문서의 수집과 단어 추출, 카이제곱 통계량 계산은 큰 저장소 용량을 요구하고 높은 계산 복잡도를 보이기 때문에 서버에서 수행하는 것이 적합하다. 서버에서 결정된 단어 집합은 스마트 기기로 전송되어 기기 내에서 사용자의 성별을 예측한다. 사생활 침해 문제를 최소화하기 위해 텍스트 데이터는 사용자의 기기 내에서만 처리되며 스마트 기기의 한정된 저장 공간과 계산 능력을 고려하여 계산 복잡도가 낮은 방법을 적용한다.



〈Figure 1〉 Overview of the Ensemble Based Gender Prediction Method



〈Figure 2〉 Gender Classification Process of Text Based Classifier

3.1.1 단어 집합 추출

단어 집합 추출은 웹 문서 수집에서 시작된다. 수집되는 웹 문서는 작성자의 성별이 명시적으로 드러나 있는 것으로 한정한다. 사용될 수 있는 웹 문서의 예로는 블로그나 트위터, 페이스북 등의 소셜 네트워크에 게시된 글이 있다. 수집한 웹 문서로부터 어간 추출과 스테밍(steming), 불용어 처리(stop word removal) 과정 등을 거쳐 단어를 추출한다[6]. 남성과 여성을 대표하는 특징적 단어 집합을 결정하기 위해 각 단어의 카이 제곱 통계량을 Kim et al.[12]과 같이 계산한다. 성별마다 카이 제곱 통계량의 크기가 큰 단어들을 해당 성별의 특징적 단어 집합으로 선택한다. 단, 단어 집합의 크기는 주어지며 남녀에 동일하게 적용된다.

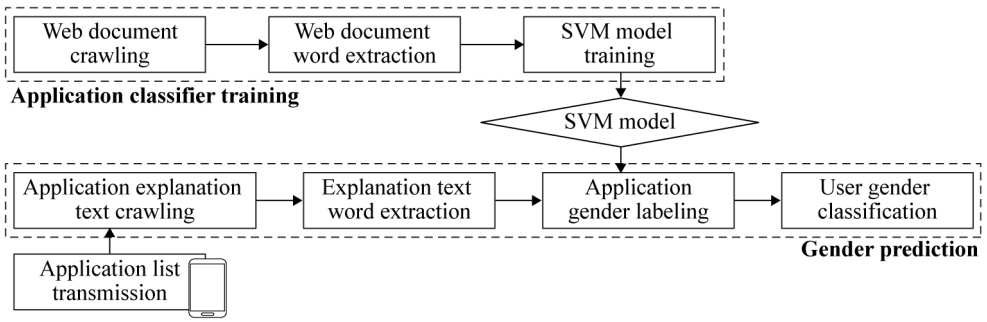
3.1.2 텍스트 기반 성별 예측

서버에서 스마트 기기로 특징적 단어 집합과 단어 별 카이 제곱 통계량이 전송되면 스마트 기기 텍스트 데이터와 단어 집합의 각 단어 사이의 일치 횟수를 추출한다. 먼저, 남녀 단어 집합과의 비교를 위해 스마트 기기 내에서 텍스트 데이터를 수집하고 하나의 문자열로 연결한다. 계산 복잡도를 줄이기 위해서 텍스트 데이터에 대해서는 웹 문서에 사용한 단어 추출 방법을 사용하지 않고, 확인하고자 하는 단어의 길이를 갖는 연속적인 문자열(character sequence)을 조사한다.

그 결과 얻어지는 성별에 따른 일치 빈도 벡터와 단어 별 카이 제곱 통계량을 성분으로 갖는 벡터의 유사도를 측정하기 위해 벡터 유사도 계산에 널리 사용되는 코사인 유사도(cosine similarity)[15]를 도입한다. 마지막으로 높은 유사도를 보이는 성별을 사용자의 성별로 예측한다. 단, 두 코사인 유사도의 값이 동일하면 해당 사용자에 대해 성별을 분류하지 않는다. 따라서 텍스트 기반 분류기를 거치면 사용자의 성별이 남성 혹은 여성으로 분류되거나 미분류 상태로 남을 수 있다.

3.2 어플리케이션 기반 분류기

어플리케이션 기반 분류기는 어플리케이션 설명글을 토대로 어플리케이션에 남성 혹은 여성의 레이블을 부여하고, 사용자가 실행한 어플리케이션들의 성별 레이블을 비교하여 사용자의 성별을 분류한다. <Figure 3>은 어플리케이션 데이터를 이용한 사용자의 성별 분류 과정을 도식화한 것이다. 스마트 기기로부터 사용 어플리케이션 목록을 서버로 전송하는 과정을 제외한 모든 과정이 서버에서 수행된다.



〈Figure 3〉 Gender Classification Process of Application Based Classifier

3.2.1 어플리케이션 분류기 학습

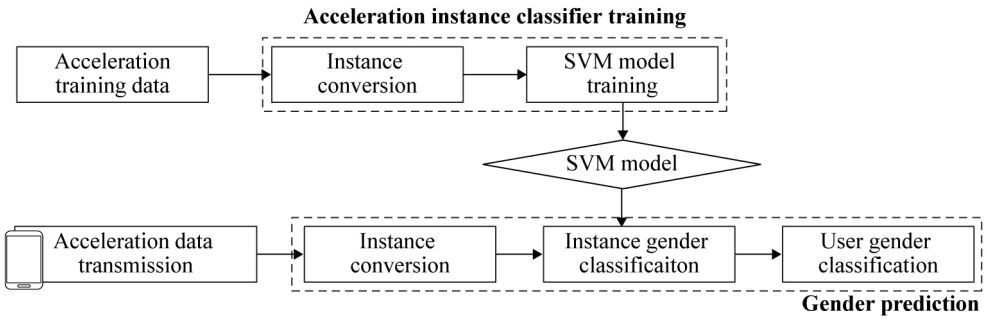
어플리케이션에 남녀 레이블을 부여하기 위한 분류기를 학습하기 위해서 텍스트 기반 분류기와 같이 작성자의 성별이 명시된 웹 문서를 수집한다. 수집한 모든 문서들로부터 단어를 추출하고 모든 문서에서 발생한 단어들로 단어 주머니(bag of words)를 생성한다. 각 문서에 대해 단어 주머니의 단어가 발생하는 빈도수를 성분으로 하는 단어 벡터와 문서 작성자의 성별을 학습데이터로 갖는 Support Vector Machine(SVM)모델을 학습한다.

3.2.2 실행 어플리케이션 기반 성별 예측

사용자의 스마트 기기에서 실행된 어플리케이션의 목록을 서버로 전송하여 성별 예측 단계를 시작한다. 웹에서 어플리케이션 게시자가 작성한 어플리케이션 설명글을 수집하고 그로부터 단어와 단어의 발생 빈도수를 추출하여 어플리케이션 분류기 학습에 사용된 것과 동일한 형태의 단어 벡터를 생성한다. 학습된 어플리케이션 분류기를 통해 해당 어플리케이션에 성별 레이블을 부여한다. 한 명의 사용자의 성별은 부여된 어플리케이션 성별을 다수결하여 예측한다. 이때, 학습된 SVM 모델의 편향을 보정하기 위해 정규화된 비중으로 다수결한다. 정규화를 위해서 웹 상에 설명글이 존재하는 수집된 모든 어플리케이션의 분류 결과를 사용하였다. 어플리케이션에 부여된 성비가 동일하면 미분류 처리한다. 예를 들어 사용자가 사용한 5개의 어플리케이션 중 1개의 어플리케이션이 남성으로, 4개의 어플리케이션이 여성으로 분류되었다면 이 사용자는 여성으로 예측한다.

3.3 가속도 기반 분류기

가속도 기반 분류기는 [22]를 기반으로 설계되었다. [22]는 실험실 환경에서 얻어진 스마트폰 가속도 데이터를 이용하여 사용자의 성별을 예측하였다. 정해진 크기의 짧은 타임 윈도우(example duration)에서 연속적으로 발생한 가속도 데이터를 사전에 정의된 요인을 갖는 하나의 데이터 인스턴스로 변환하여 분류기의 입력으로 사용하였다. 한 명의 사용자에 대해 여러 개의 인스턴스가 발생하고, 이 인스턴스들을 각각 남녀로 분류한다. 최종적으로는 인스턴스들의 성비를 계산하여 한 명의 사용자의 성별을 예측한다. 이러한 예측 기법을 본 연구에 도입하면 여러 시간대에서 발생하는 사



〈Figure 4〉 Gender Classification Process of Acceleration Based Classifier

용자의 움직임을 반영하여 성별을 예측할 수 있다. <Figure 4>에 가속도 데이터를 이용하여 사용자의 성별을 분류하는 과정을 나타내었다.

3.3.1 가속도 인스턴스 분류기 학습

기존의 연구에서 가속도 데이터 측정 시 반복적인 동작을 취한 것과는 달리 본 연구에서는 실제 생활 속에서 스마트 기기를 사용할 때 발생하는 로그 데이터를 사용한다. 그러므로 기존에 사용된 요인들 중 반복 동작과 관련된 요인을 제외하고, 일주일 중 주중인지 여부(weekday)와 시간대(hour), 정해진 타임 윈도우 안에서 수집된 데이터의 갯수(cnt)를 요인으로 추가한다. 데이터 수집 어플리케이션이 구동되는 안드로이드 운영체제에서는 센서 매니저 API를 사용하여 가속도 데이터를 수집하게 되는데, 일정 시간 동안의 수집 횟수가 가속도 변화와 기기의 특성에 따라 변화한다. 따라서 cnt 요인을 추가하면 사용자의 가속도 변화 특성을 반영할 수 있을 것으로 예상된다. 이외에도 가속도 크기의 평균(norm)과 각 방향 가속도의 평균(avg), 분산(var), 최소값(min)와 최대값(max)을 요인으로 사용한다. 분류기 학습에 사용되는 16개의 모든 요인을 정리하면 <Table 1>과 같다. 16개의 요인으로 구성된 가속도 인스턴스와 그에 해당하는 성별 레이블로 구성된 학습데이터로 SVM 모델을 학습하여 가속도 인스턴스 분류기를 생성한다.

〈Table 1〉 Features Utilized for Acceleration Instance Classifier

Feature	Description	Example
weekday	Weekday/weekend identification	weekday = 0, weekend = 1
hour	Time slot	24
cnt	The number of data collected in a unit time	41 per one minute
{X, Y, Z}min	The minimum value of data collected in a unit time(X, Y, and Z axes)	0.0360107421875
{X, Y, Z}max	The maximum value of data collected in a unit time(X, Y, and Z axes)	6.20950317382813
{X, Y, Z}avg	The average value of data collected in a unit time(X, Y, and Z axes)	7.06655883789063
{X, Y, Z}var	The variance of data collected in a unit time(X, Y, and Z axes)	0.04179000118217
norm	The average value of L2-norm of X, Y and Z axes data in a unit time	4.12770853245654

3.3.2 가속도 기반 성별 예측

새로운 사용자의 가속도 데이터가 서버로 전송되면 성별 예측을 하기 위해 학습데이터와 동일한 요인을 가지는 인스턴스로 변환된다. 각각의 인스턴스는 학습된 가속도 인스턴스 분류기에 의해 남성 혹은 여성으로 분류된다. 사용자의 모든 인스턴스 분류 결과를 종합하여 다수를 차지하는 성별이 가속도 기반 분류기의 분류 결과로 결정된다. 인스턴스의 성비가 동일하면 미분류 처리한다. 또한 예측 대상이 되는 사용자의 가속도 데이터 인스턴스가 매우 적으면 다수결 시 편향이 발생할 수 있으므로 일정 개수 이하의 인스턴스를 가진 사용자도 미분류 처리된다.

3.4 다수결에 근거한 앙상블

본 연구에서 제안하는 기법은 텍스트 기반 분류기와 어플리케이션 기반 분류기, 가속도 기반 분류기로부터 얻어진 결과를 다수결하여 사용자의 성별을 예측한다. 분류기들로부터 얻어진 결과로 다수결에 의해 예측하되 미분류된 결과는 다수결에 포함시키지 않는다. 따라서 동률이 발생할 수 있고, 이 경우에는 결과의 신뢰도가 가장 높은 분류기의 예측 성별을 최종 예측 성별로 선택한다.

각 분류기의 신뢰도는 0과 1사이의 값으로 정의한다. 텍스트 기반 분류기는 가장 높은 코사인 유사도 값을 0°~90° 각도로 변환하고 이를 다시 0과 1사이 값으로 정규화하여 신뢰도로 사용한다. 다수결을 적용한 어플리케이션 기반 분류기와 가속도 기반 분류기의 경우 다수를 차지한 성의 비율이 소수인 성의 비율보다 클수록 신뢰할 수 있는 결과라 할 수 있다. 따라서 두 분류기에 대해서는 두 성별의 비율 차이를 신뢰도로 정의한다.

4. 실험 및 결과

4.1 실험 데이터

본 연구에서 제안한 스마트 기기 사용자의 성별 예측 기법의 성능을 평가하기 위해 안드로이드 어플리케이션을 통해 텍스트 데이터인 주소록에 등록된 이름, 문자메시지, 브라우저 북마크, 검색 기록과 사용자가 실제로 실행한 어플리케이션, 기기에 설치된 어플리케이션, 주기적인 가속도 데이터를 수집하였다. 사용자의 실제 성별은 사용자가 직접 입력하도록 하였다. 사용자가 실제로 실행한 어플리케이션을 조사하기 위해서 Böhmer et al.[1]의 연구 결과를 바탕으로 시스템의 가장 상위에서 구동되고 있는 프로그램을 1분 간격으로 수집하였다. 가속도 데이터는 15분마다 30초씩 수집되어 30초간의 데이터가 하나의 인스턴스를 이루도록 하였다.

20명의 스마트 기기 사용자로부터 데이터를 수집하였으며 실험에 참여한 사용자는 11명의 남성과 9명의 여성으로 구성되었다. 수집한 스마트 기기 로그 데이터의 특성을 <Table 2> 상단에 요약하였다. 평균 텍스트 데이터의 길이는 여성 사용자가 남성 사용자의 약 3배에 달하는 것으로 나타났다. 설치된 어플리케이션의 수는 여성이 다소 많았으며 가속도 데이터 인스턴스도 여성 사용자로부터 더 많이 수집되었다. 가속도 데이터 인스턴스의 경우 일부 여성 사용자의 데이터 수집 기간이 다른 사용자

〈Table 2〉 Summary of Collected Data

Data type	Category	Male	Female	Overall
Smart device log data	The number of subjects	11	9	20
	The average character length of text data	45,633	140,668	88,399
	The average number of installed applications	325	359	340
	The average number of acceleration instances	251	472	347
Blog data	The number of blogs	97	65	162
	The number of documents	135,745	53,382	88,399
	The number of extracted words	137,743	117,941	141,509
	The average character length of a document	1,327	1,520	1,381

와 비교하여 상대적으로 길었던 것으로 나타났다.

텍스트 기반 분류기와 어플리케이션 기반 분류기는 작성자의 성별이 알려져 있는 웹 문서를 필요로 한다. 본 실험에서는 이를 위해 성별에 따라 서로 다른 관심사와 사용 단어를 포착할 수 있는 블로그 문서를 수집하였다. 다양한 주제를 갖는 블로그들 중 문서 작성자인 블로거의 성별이 명시되어 있는 블로그에 한하여 문서를 수집하였다. 수집한 블로그 데이터를 <Table 2>의 하단에 요약하였다. 텍스트 기반 분류기에서는 141,509개의 단어 중 전체 문서에서의 출현 빈도수가 5회 미만인 단어는 카이 제곱 통계량 계산에서 제외되었다.

4.2 실험 환경 및 평가 지표

텍스트 기반 분류기의 단어 집합 크기는 단일 분류기 실험 결과 가장 좋은 성능을 보인 10으로 고정하였다. 모든 SVM 모델은 선형(linear) SVM을 사용하였으며 모델의 변수는 패키지가 제공하는 기본값을 그대로 사용하였다. 모든 SVM 모델 학습을 위해 데이터를 정규화하였고, 비복원 랜덤 샘플링하여 동일한 수의 남녀 학습데이터를 사용하였다. 또한 한 사용자에 대해 수집된 가속도 인스턴스의 수가 정상 수집 시 하루 동안 수집 가능한 96개 미만인 경우 가속도 데이터로 성별을 분류하지 않았다.

제안된 기법의 성능은 텍스트, 어플리케이션, 가속도 데이터에 기반한 세 분류기와 비교하여 정확도와 미분류율의 지표로 평가되었다. 정확도는 분류 성능을 평가하기 위해 보편적으로 사용되는 지표로서[18], 본 연구에서는 전체 사용자에 대한 정확도와 남녀 사용자에 대한 예측 정확도로 세분화하였다. 미분류율은 사용자의 성별이 분류되지 못한 경우를 성능으로 평가하기 위하여 도입하였으며, 전체 피실험자 중에서 제안 기법으로 성별을 예측하지 못한 사용자의 비율로 계산된다. 따라서 미분류율이 낮을수록 좋은 성능을 보인다고 할 수 있다.

가속도 기반 분류기의 성능을 평가하기 위해 leave-one-out 교차타당화(cross-validation) 방법을 적용하였다. 텍스트와 어플리케이션에 기반한 분류기는 학습과정이 수집된 스마트 기기 로그 데이터와 무관하지만 가속도 기반 분류기는 수집된 데이터로 학습되기 때문에 정당한 성능 비교를 위해 해당 피실험자의 데이터를 제외한 데이터로 학습된 분류기를 사용하고 그 결과를 평가에 반영하였다.

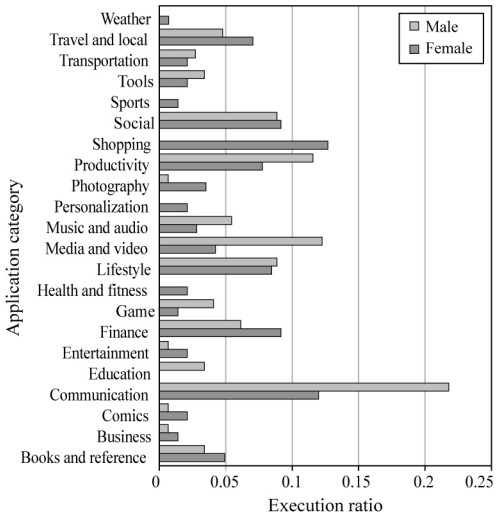
본 실험은 Java 프로그래밍 언어를 사용하여 수행되었다. 수집한 문서로부터 단어를 추출하기 위해 루신 아리랑 분석기와 꼬꼬마 형태소 분석기[20]를 이용했다. 어플리케이션 기반 분류기의 SVM은 학습데이터의 희소 행렬 특성을 고려하여 SVM Light 라이브러리[11]를 사용하였고 가속도 기반 분류기의 SVM은 LibSVM 라이브러리[4]로 구현되었다.

4.3 실험 결과

4.3.1 데이터 탐색

<Table 3>는 수집된 텍스트 데이터에서 추출한 단어들 중 출현 빈도가 가장 높은 10개의 단어를 남녀에 대해 나타낸 것이다. 남성의 텍스트 데이터에서 발생한 단어들로는 사용자의 성별을 직관적으로 파악할 수 없었지만 여성의 경우 사용자의 성별을 예상할 수 있는 '오빠', '언니' 등의 단어가 빈번히 발생하였다.

사용자가 사용한 어플리케이션도 성별에 따라 차이를 나타냈다. <Figure 5>는 각 성별에 대해 카테고리에 따른 어플리케이션 실행 비율을 보여준다. 그래프에서 x축은 데이터 수집 기간 동안 사용한 모든 어플리케이션 중에서 각 카테고리의 어플리케이션이 차지하는 비율을, y축은 구글에서 지정한 어플리케이션 카테고리의 명칭을 나타낸다. 남성 사용자의 데이터를 살펴보면 생산성(productivity)과 미디어와 비디오(media and video), 게임(game), 커뮤니케이션(communication) 등의 카테고리 사용비율이 여성보다 높은 것으로 나타났다. 반면 여성 사용자는 남성 사용자는 실행하지 않은 쇼핑(shopping) 카테고리에 속하는 어플리케이션을 높은 비율로 이용하였다.

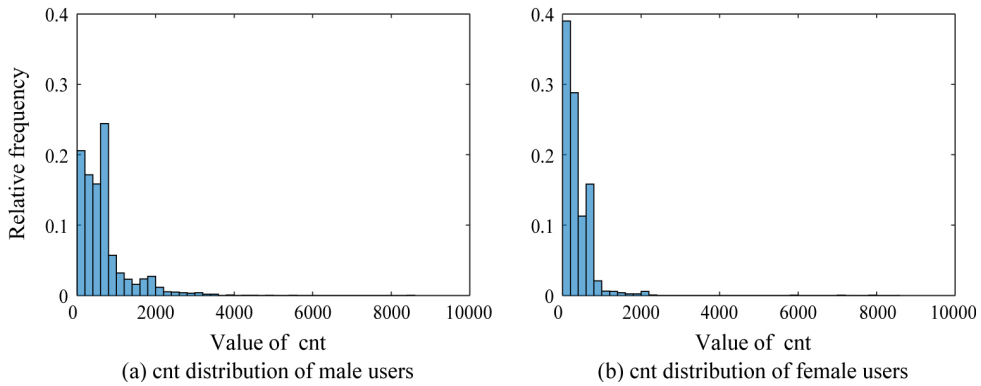


〈Figure 5〉 Execution Ratio of Applications by Category

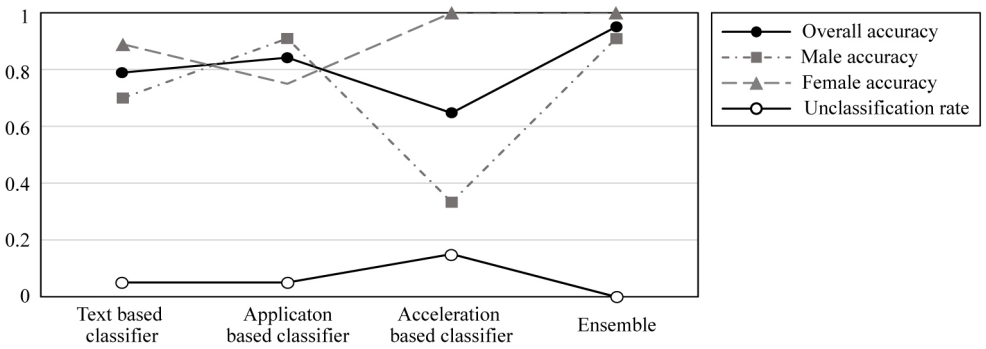
가속도 데이터 수집 결과 cnt 요인의 분포가 성별에 따른 차이를 나타내었다. <Figure 6>은 cnt 요인 분포를 보여준다. 그림에서 알 수 있듯이 여성 사용자의 데이터에서 낮은 cnt값을 가지는 인스턴스가 남성에 비해 높은 빈도로 관측되었다.

〈Table 3〉 Top Ten Most Popular Words in Collected Smart Device Text Data

Gender	Words
Male	Naver(네이버) Web cartoon(웹툰) Confirmation(확인) Living(생활) Data(데이터) Taxi(택시) Use(이용) Phone call(통화) Basics(기본) Completion(완료)
Female	Nate(네이트) News(뉴스) Older brother(오빠) Today(오늘) Phone call(통화) Older sister(언니) Sweet heart(쟈기야) Tomorrow(내일) Now(지금) Thanks/gratitude(감사)



〈Figure 6〉 Histogram of cnt Feature Distribution



〈Figure 7〉 Accuracy and Unclassification Rate of the Classifiers and the Proposed Ensemble Method

4.3.2 성능 비교

텍스트, 어플리케이션, 가속도 기반 분류기와 최종적으로 제안된 앙상블 기법의 실험 결과를 <Figure 7>에 나타내었다. 제안된 앙상블 기법의 결과가 전체 정확도 0.95, 미분류율 0으로 가장 높은 성능을 보였다. 남성 정확도와 여성 정확도 또한 앙상블한 결과가 세 분류기들 보다 높았다.

제안 방법론을 구성하는 개별 분류기들은 기존 방법론에 근간을 둔 것으로 이들과의 비교를 통해 제안 방법론과 기존 연구 간의 간접적인 비교 평가가 이루어진다고 할 수 있다. 다만, 개별 분류기들은 데이터 수집 방식 및 양의 차이로 기존 방법론을 그대로 도입할 수 없기 때문에 본 문제에 적합하게 개량되었다. 텍스트 기반, 어플리케이션 기반, 가속도 기반 분류기들은 0.6에서 0.85사이의 정확도를 보여 0.95의 정확도를 보인 제안 방법론보다 낮은 성능을 나타냈다. 이는 앙상블 기법의 도입을 통해 데이터의 멀티 모달리티를 고려하여 예측에 사용되는 정보의 양을 극대화 하였으며, 동시에 정보의 손실을 최소화하기 위해 데이터를 각각의 특성에 맞게 분류에 이용하였기 때문이다.

어플리케이션 기반 분류기를 구축하기 위해 피실험자의 스마트 기기에 설치된 모든 어플리

케이션을 조사한 결과, 구글에 등록되어 설명글이 존재하는 어플리케이션은 총 663개였다. 이 중 77%의 어플리케이션이 남성으로 분류되었고 23%는 여성으로 분류되었다. 어플리케이션 기반 분류기의 정확도를 살펴보면, 전체 정확도는 세 가지 분류기 중 가장 높은 0.8421을 기록하였다. 또한 남성 사용자에 대한 정확도가 여성 사용자에 대한 정확도보다 높았다. 이는 위에서 언급한 바와 같이 웹 문서로 학습된 SVM 모델이 80%에 가까운 어플리케이션을 남성 어플리케이션으로 분류하였기 때문인 것으로 분석된다. 비록 모델에 의해서 남녀로 레이블이 부여된 설치 어플리케이션들의 비율을 고려하여 다수결 시 정규화하였지만 모든 안드로이드 어플리케이션이 고려된 것은 아니기 때문에 편향이 발생한 것으로 보인다.

가속도 기반 분류기로 사용자의 성별을 분류한 결과, 앞에서 살펴본 다른 분류기보다 분류 정확도가 낮았다. 낮은 성능의 원인은 가속도 기반 분류기가 다른 분류기들과는 달리 수집된 스마트 기기 로그 데이터에 영향을 받는 통계적 학습 기법을 사용하는 데에 있다. 어플리케이션 기반 분류기는 웹에서 얻어진, 스마트 기기 로그 데이터와는 별개인 데이터를 사용하여 SVM 모델을 학습한다. 그러나 가속도 기반 분류기의 경우 스마트 기기 사용자로부터 수집된 데이터를 바탕으로 SVM 모델을 학습하고 성별을 예측한다. 본 실험에서 수집된 가속도 데이터의 부족으로 SVM이 충분히 학습되지 못했고 그 결과 가속도 데이터 인스턴스를 올바르게 분류하지 못하였다.

4.3.3 사례연구

피실험자 A는 세 가지 분류기에 의해서 만장일치로 올바르게 여성으로 예측되었다. 우선, 사용자 A의 텍스트 기반 분류기의 분류 신뢰도는 다른 사용자들과 비교하여 가장 높았다. <Table 4>는 카이 제곱 통계량을 기준으로 선택된 10개의 단어와 정규화된 점수, 사용자의 텍스트 데이터에서 실제로 발생한 횟수를 보여준다. 남성 단어 집합과 일치하는 단어는 87회 출현하는데 그쳤지만 여성 단어 집합과 일치하는 단어는 1,673회 출현하였다. 특히 점수의 비중이 높은 '엄마'와 '언니'가 가장 많이 출현하여 여성의 특징적 단어 집합과의 비교로 계산된 코사인 유사도가 높았다.

〈Table 4〉 Case Study: Matching Words

Male wordset		Female wordset	
Word	Frequency	Word	Frequency
Director(감독)	2	Mom(엄마)	332
Movie(영화)	34	Older sister(언니)	691
Answer(정답)	0	Completeness(완전)	175
Problem(문제)	47	Husband(신랑)	0
Movie release(개봉)	1	Child(아이)	84
Emergence(등장)	1	Our/my(저희)	46
Production(제작)	1	I(제가)	81
Appearance(출연)	1	We(우리)	227
Action(액션)	0	These days(요즘)	35
Original work(원작)	0	This time(요번)	2

피실험자 A의 실행 어플리케이션 중 구글 어플리케이션 마켓에 등록되어 있는 어플리케이션은 <Table 5>와 같았다. 이 사용자에게 여성 레이블이 부여된 어플리케이션이 남성 레이블의 어플리케이션보다 많음을 알 수 있다. 실제 여성이 많이 사용하는 소셜 커머스의 모바일 어플리케이션(Coupang)과 사진으로 대화하는 메신저(Snapchat) 등이 여성으로 분류되었다. 마지막으로 이 사용자의 가속도 데이터는 여성으로 분류된 인스턴스가 246개, 남성으로 분류된 인스턴스가 10개로, 올바른 분류 결과를 보였다.

5. 결 론

본 연구에서는 스마트 기기 로그 데이터를 이용하여 기기 사용자의 성별을 예측하는 앙상블 기법을 제안하였다. 제안된 앙상블 기법은 로그 데이터 중 텍스트와 어플리케이션, 가속도 데이터를 입력으로 갖는 세 가지 분류기를 바탕으로 다수결에 의해 최종적으로 사용자의 성별을 예측한다. 또한 스마트 기기 로그 데이터 수집 어플리케이션을 개발하고 이를 이용하여 데이터를 수집하여 실험을 통해 제안 기법의 성능을 평가하였다.

텍스트 기반 분류기는 사용자의 성별을 잘 구분할 수 있을 것으로 예상되지만 타인에게 노출될 경우에 사생활 침해 문제가 발생하는 문자메시지, 검색 기록 등을 사용하기 위하여 기기 내에서 데이터 처리가 가능하도록 하였다. 어플리케이션 기반 분류기는 웹에서 수집한 어플리케이션의 설명글을 사용하여 어플리케이션에 성별 레이블을 부여하고 이를 바탕으로 사용자의 성별을 예측하였다. 사용자가 데이터 수집기간 동안 실행한 어플리케이션의 성별을 분류한 결과를 이용하였으며 사용자의 실제 어플리케이션 사용 성향을 반영하여 제안된 세 가지 분류기 중 가장 높은 분류 정확도를 얻었다. 가속도 기반 분류기는 스마트 기기 로그 데이터와 데이터 수집 플랫폼의 특성을 반영하는 요인을 입력으로 갖는 SVM을 학습하여 가속도 인스턴스에 대해 성별을 분류하여 그 결과의 다수결에 따라 사용자의 성별을 예측하였다.

제안된 앙상블 기법은 위의 세 분류기의 결과를 모두 사용하여 사용자의 성별을 최종적으로 예측하였다. 수집된 데이터를 통해 성능을 확인한 결과, 각 분류기들과 비교하여 더 높은 분류 정확도를 보였으며 남성과 여성 사용자에 대해 모두 높은 분류 성능을 나타내었다.

〈Table 5〉 Case Study: Executed Applications

Application name	Category	Description	Labeled gender
KakaoTalk	Communication	Mobile messaging service	Male
NH Smart Banking	Finance	Mobile banking service	Male
Coupang-discount, mart	Shopping	Social commerce	Female
Naver Music	Music and audio	Music streaming service	Female
PhotoWonder	Photography	Photo editing service	male
Naver Calendar	Productivity	Personal schedule management	Female
Snapchat	Social	Photo conversation service	Female
Kakao Home	Personalization	Android launcher	Female

더불어, 제안된 기법은 텍스트와 어플리케이션, 가속도 데이터 중 일부가 수집될 수 없는 상황에서도 사용자의 성별을 분류할 수 있다. 높은 분류 성능과 이러한 기법의 특성은 실제 스마트 기기 로그 데이터를 이용한 사용자 성별 예측에 제안된 기법이 실제로 적용 가능함을 의미한다.

추후 연구로는 두 가지 방향이 있다. 첫째, 제안 방법론의 성능을 극대화시키기 위해 다양한 종류의 앙상블 기법을 도입할 수 있다. 본 연구는 스마트 기기 로그 데이터의 멀티 모달리티를 앙상블하는 초기 연구로 앙상블 방식 중 가장 단순한 다수결 방식을 사용하였으나, 다양한 방식의 비교를 통해 성능을 향상 시킬 것으로 기대한다. 둘째, 사용자의 다양한 속성에 대한 기법을 개발 할 수 있다. 연구의 배경이 된 스마트 기기를 이용한 개인화 서비스에 필요한 정보는 사용자의 성별 외에도 많은 인구통계학적 정보를 포함한다. 나이와 직업, 종교, 가족구성원의 수 등은 성별과 더불어 매우 중요한 사용자 정보이며, 이 속성들을 위한 기법 또한 연구되어야 할 것이다.

References

[1] Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G., "Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage," Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services, 2011.

[2] Baek, S. I. and Choi, D. S., "Exploring User Attitude to Information Privacy," The Journal of Society for e-Business Studies, Vol. 20, No. 1, pp. 45-59, 2015.

[3] Brdar, S., Ćulibrk, D., and Crnojević, V., "Demographic Attributes Prediction on the Real-World Mobile Data," Proceedings of Mobile Data Challenge by Nokia Workshop, 2012.

[4] Chang, C.-C. and Lin, C.-J., "LIBSVM: A Library for Support Vector Machines," ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, p. 27, 2011.

[5] Chen, P.-T. and Hsieh, H.-P., "Personalized Mobile Advertising: Its Key Attributes, Trends, and Social Impact," Technological Forecasting and Social Change, Vol. 79, No. 3, pp. 543-557, 2012.

[6] Croft, W. B., Metzler, D., and Strohman, T., Search Engines: Information Retrieval in Practice, Pearson, 2009.

[7] Delany, S. J., Buckley, M., and Greene, D., "SMS Spam Filtering: Methods and Data," Expert Systems with Applications, Vol. 39, No. 10, pp. 9899-9908, 2012.

[8] Ha, S. H., Oh, J., and Lee, B. G., "The Analysis of Advertisement Effect in Smart Phone Environment: The Comparison of Users with Providers of Commercial," The Journal of Society for e-Business Studies, Vol. 16, No. 4, pp. 221-239, 2011.

[9] Hu, J., Zeng, H.-J., Li, H., Niu, C., and

Chen, Z., "Demographic Prediction based on User's Browsing Behavior," Proceedings of the International Conference on World Wide Web, 2007.

[10] Igarashi, T., Takai, J., and Yoshida, T., "Gender Differences in Social Network Development via Mobile Phone Text Messages: A Longitudinal Study," Journal of Social and Personal Relationships, Vol. 22, No. 5, pp. 691-713, 2005.

[11] Joachims, T., "Making Large-Scale SVM Learning Practical," in Advances in Kernel Methods-Support Vector Learning, ed Cambridge, Massachusetts: MIT Press, pp. 169-184, 1999.

[12] Kim, S., Choi, Y., Kim, Y., Park, K., and Park, J., "On-Device Gender Prediction Framework Based on the Development of Discriminative Word and Emoticon Sets," KIISE Transactions on Computing Practices, Vol. 21, No. 11, pp. 733-738, 2015.

[13] Kuncheva, L. I., Combining Pattern Classifiers: Methods and Algorithms, John Wiley and Sons, 2004.

[14] Laurila, J. K., Gatica-Perez, D., Aad, I., Blom, J., Bornet, O., Do, T. M. T., Dousse, O., Eberle, J., and Miettinen, M., "From Big Smartphone Data to Worldwide Research: The Mobile Data Challenge," Pervasive and Mobile Computing, Vol. 9, No. 6, pp. 752-771, 2013.

[15] Lee, D. and Shim, J., "Survey on Vector Similarity Measures: Focusing on Algebraic Characteristics," The Journal of Society for e-Business Studies, Vol. 17, No. 4, pp. 209-219, 2012.

[16] Lee, Z., Choi, H., and Choi, S., "Study on How Service Usefulness and Privacy Concern Influence on Service Acceptance," The Journal of Society for e-Business Studies, Vol. 12, No. 4, pp. 37-51, 2007.

[17] Mohrehkesh, S., Ji, S., Nadeem, T., and Weigle, M. C., "Demographic Prediction of Mobile User from Phone Usage," Proceedings of Mobile Data Challenge by Nokia Workshop, 2012.

[18] Roh, J.-H., Kim, H.-j., and Chang, J.-Y., "Improving Hypertext Classification Systems Through WordNet-based Feature Abstraction," The Journal of Society for e-Business Studies, Vol. 18, No. 2, pp. 95-110, 2013.

[19] Seneviratne, S., Seneviratne, A., Mohapatra, P. and Mahanti, A., "Your Installed Apps Reveal Your Gender and More!," SIGMOBILE Mobile Computing and Communications Review, Vol. 18, pp. 55-61, 2015.

[20] Shim, K.-S., "MADE: Morphological Analyzer Development Environment," Journal of Internet Computing and Services, Vol. 8, No. 4, pp. 159-171, 2007.

[21] Walkowiak, K., Sztajer, S., and Woźniak, M., "Decentralized Distributed Computing System for Privacy-Preserving Combined Classifiers-Modeling and Optimization," Proceedings of the International Conference on Computational Science and Its

Applications, 2011.

[22] Weiss, G. M. and Lockhart, J. W., "Identifying User Traits By Mining Smart Phone Accelerometer Data," Proceedings of the International Workshop on Knowledge Discovery from Sensor Data, 2011.

[23] Woźniak, M., Graña, M., and Corchado, E., "A Survey of Multiple Classifier Systems as Hybrid Systems," Information Fusion, Vol. 16, pp. 3-17, 2014.

[24] Ying, J. J.-C., Chang, Y.-J., Huang, C.-M. and Tseng, V. S., "Demographic Prediction based on Users Mobile Behaviors," Proceedings of Mobile Data Challenge by Nokia Workshop, 2012.

[25] Zenobi, G. and Cunningham, P., "Using Diversity in Preparing Ensembles of Classifiers based on Different Feature Subsets to Minimize Generalization Error," Proceedings of the European Conference on Machine Learning, 2001.

[26] Zhong, E., Tan, B., Mo, K., and Yang, Q., "User Demographics Prediction Based on Mobile Data," Pervasive and Mobile Computing, Vol. 9, No. 6, pp. 823-837, 2013.

저 자 소 개

김윤정 (E-mail: yoonj625@gmail.com)
2013년 서울대학교 산업공학과 (학사)
2015년 서울대학교 산업공학과 (석사)
관심분야 통계학습, 모바일 데이터

최예림 (E-mail: iangoozh@gmail.com)
2010년 서울대학교 산업공학과 (학사)
2010년~현재 서울대학교 산업공학과 (석박사 통합과정)
관심분야 사물인터넷 및 빅데이터 기반의 인간 모델링

김소이 (E-mail: kpsinw@gmail.com)
2014년 포항공과대학교 산업경영공학과 (학사)
2014년~현재 서울대학교 산업공학과 (석사과정)
관심분야 데이터마이닝, 추천 시스템

박규연 (E-mail: mysnuky91@snu.ac.kr)
2015년 서울대학교 산업공학과 (학사)
2015년~현재 서울대학교 산업공학과 (석사과정)
관심분야 모바일 서비스

박종헌 (E-mail: jonghun@snu.ac.kr)
1990년 서울대학교 산업공학과 (학사)
1992년 서울대학교 산업공학과 (석사)
2000년 Georgia Institute of Technology 산업시스템공학과 (박사)
2004년~현재 서울대학교 산업공학과 교수
관심분야 모바일 인텔리전스, 산업 데이터 애너리틱스