모바일 환경에서 행동추론을 위한 효과적인 데이터 전처리

이영설○ 조성배
연세대학교 컴퓨터과학과
tiras@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Effective Data Preprocessing for Activity Inference in Mobile Environments

Young-Seol Lee○, Sung-Bae Cho
Dept. of Computer Science, Yonsei university

센서와 무선통신 기술의 발달로 우리 주변의 많은 정보들을 정밀하게 수집하는 기술이 발전하고 있다. USN(Ubiquitous Sensor Network)나 RFID(Radio Frequency Identification) 태그, 가속도 센서와 자이로 센서, 그 외에 주변의 온도, 습도, 빛의 세기, 영상정보나 음성 정보를 수집할 수 있는 센서들이 그 대표적인 사례이며 인공지능 분야에서는 이들 센서를 이용하여 인간의 행동이나 상태를 인식하고 분류하기 위한 연구들이 많이 진행되고 있다[1]. 센서를 통해 수집되는 상황정보를 통해 인간의 행동을 기계가 자동으로 인식할 수 있게 되면 인간의 행동이나 상태를 실시간으로 체크할 수 있게 되어 장애 환자나 건강에 문제가 있는 노인들을 돌보는 간호 서비스의 비용과 노력을 절감하고 서비스의 질을 향상시킬 수 있다[2]. 또한 사람의 위치나 상황에 따라서 알맞은 맞춤형 서비스를 제안하거나 추천할 수 있으며 사진이나 동영상 등에 태그 정보로 함께 제공되어 이후에 자신의 기록을 정리, 분석, 검색하는 데 중요한 단서로 이용될 수도 있다. 본 논문에서는 삼성 m-4650 스마트폰을 이용하여 행동을 추론하기 위한 사용자의 생활 기록을 수집하였으며, 휴대용 GPS인 BT-335를 이용하여 GPS 위치 정보를 함께 수집하였다. 수집된 정보는 표 1과 같다.

표 1. 수집하는 정보

구분	수집 모듈	상세 내역
행동	레이블링	시간, 행동, 감정, 상태
통화 내역	휴대 전화	시간, 수신/발신/부재, 전화번호, 전화 상대방 정보
GPS 정보	GPS 장치	위도, 경도, 속도, 고도
MP3 정보	휴대 전화	시간, 곡명, 상태
사진 정보	휴대 전화	촬영 대상, 상태
장소 정보	레이블링	시간, 방문장소
SMS 정보	휴대 전화	시간, 전화번호, 수신/발신, 메시지 내용

수집된 데이터 사이의 상관관계를 분석하고 확률 모델을 구성하기 위해서는 시간을 기준으로 수집된 데이터를 통합하고 연속적인 데이터를 이산화할 필요가 있다. 여기서는 1분 단위로 모든 행동과 정보를 분할하고 각 시간 단위마다 나타나는 정보를 분석할 수 있도록 하나의 테이블로 통합하였다. 하나로 통합된 레코드는 GPS, Place, Activity, Call, Photo, SMS, MP3의 7개 정보를 포함하고 있다. 이 정보들로부터 특이한 상황을 추출하기 위해서 이벤트의 발생 빈도, 지속시간, 그리고 집중도를 계산한다. 이 방법은 이전 연구[3]와 동일한 방법을 이용한다. 빈도와 지속시간, 집중도는 비일상적인 상황에 대한 근거가 될 수 있다.

본 논문에서는 사용자의 행동을 추론하는 확률 모델에 앞에서 제안한 통계분석을 이용하는 것이 얼마나 효과가 있는지 살펴보기 위하여 베이지안 네트워크 확률 모델을 실제로 수집한 데이터로부터 학습하고 통계 분석을 적용한 결과와 적용하지 않은 결과를 비교한다. 이 실험을 위하여 여기서는 Weka[4]를 이용하였다. 실험을 위해 이용된 데이터는 m4650 스마트폰을 이용하여 총 12명의 학부생을 대상으로 약 4주간 수집하였다. 수집된 데이터에는 감정과 행동을 레이블링할 것을 요구하였다. 발생빈도나 지속시간에 대한 정보가 확률 모델의 성능을 향상시키는 지 확인하기 위해서 베이지안 네트워크 확률 모델을 학습시키고10-fold cross validation으로 데이터에 대한 모델의 성능을 비교하였다. 베이지안 네트워크의 학습방법은 모두 동일하게Simluated Annealing을 사용하였으며, 학습 시간을 줄이기 위해서 하나의 속성을 하나의 노드로 변환하고 속성은 기본적으로 Yes 와 No로 구분하도록 하였다. 예를 들어, 전화 상대방을 여기서는 친구, 애인, 가족, 동료, 기타, 모르는 사람 등으로 구분하였는데 이를 '전화 상대방' 이라는 노드의 속성들로 만들지 않고 각각 '친구', '애인' 이라는 노드로 만들어 그 속성 값을 'Yes' 와 'No' 로 구분하였다. 이렇게 데이터를 만들 경우 실제로 확률 모델의 학습을 시작하기 전에 데이터 분포를 보고 손쉽게 현재의 데이터가 추론하려는 행동에 영향을 주는 지 판단할 수 있고 추론에 전혀 영향을 주지 못하는 노드는 학습에서 제외함으로써 학습시간과 최종 학습된 모델의 복잡도를 줄였다. 전체 사용자의 통합 데이터 레코드 개수는 390,345 에 달하지만 이 모든 레코드를 한 번에 학습 시키지 않고 각각의 사용자별로 나누어 학습하고 모델을 생

성하였다. 다음은 한 명의 사용자에 대해서 학습한 결과이다 추론하고자 하는 행동은 '수업' (Full-Time Class)이며, 추론에 사용된 노드는 총 9개로 각각의 노드 가운데 빈도, 지속시간 등을 기준으로 생성된 노드들을 제거하면서 확률모델의 성능을 비교하였다 성능 비교를 위해서 여기서는 10-fold cross validation을 이용하였으며, 총 4번의 실험을 진행하였다

표 2. '수업'을 추론하는 모델의 구분

구분	1	2	3	4
모든 단서 사용	O			
장소 방문 빈도 제거		O	O	O
장소 방문 지속시간 제거			O	O
MP3 음악 청취 시간 제거				O

그림 5는 실험에서 나온 확률 모델의 성능을 비교한 것이다그림 5는 통계 분석 결과를 제거할수록 정확도가 낮아진다는 것을 보여준다. 이는 통계분석을 통해 얻은 부가 정보가 분류성능에 영향을 미친다는 것을 보여 주고 있다

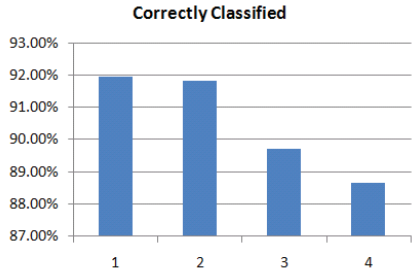

그림 1. '수업' 행동의 분류 정확도

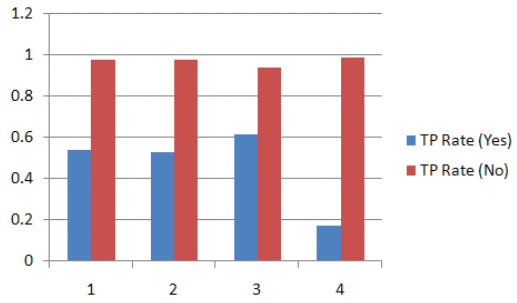

그림 2. True Positive 비율

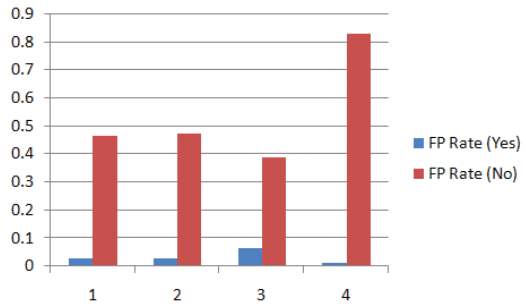

그림 3 . False Positive 비율

본 논문에서는 이전 연구에서 제안한 스마트폰의 개인정보에 대한 통계적 분(석 벤트의 발생 빈도, 지속시간, 집중도 등)이 실제로 사용자의 행동을 추론하기 위한 확률 모델을 학습하는데 의미가 있음을 보였고그 결과를 검증하기 위하여 Weka를 이용하여 생성된 확률모델의 성능을 비교하였다 향후 연구로는 스마트폰에 존재하는 기본적인 정보들뿐만 아니라, RFID 와 가속도 센서, 온도, 습도 등의 부가적 센서를 이용하여 행동 추론의 정확도를 높여볼 것이다. 동시에 추론된 정보를 이용하여 사람의 생활을 효과적으로 시각화하고 공유하기 위한 연구가 진행되어야 할 것이다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-(C1090-0801-0046))

참고문헌

[1] T. Choudhury, M. Philipose, D. Wyatt and J. Lester, "Towards activity databases: using sensors and statistical models to summarize people's lives," *IEEE Data Engineering Bulletin,* Vol. 29, No. 1, pp. 49 - 56, 2006.

[2] J. Lester, T. Choudhury and G. Borriello, "A practical approach to recognizing physical activities," *Lecture Notes in Computer Science : Pervasive Computing*, pp. 1-16, 2006.

[3] Y.-S. Lee, M.-C. Jung and S.-B. Cho, "Collection and construction of user's context in smart phone," *Proc. of KCC*, vol. 33, no. 1(B), pp. 115-117, 2006.

[4] H. I. Witten and E. Frank, "Nuts and bolts: Machine learning algorithms in Java," *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, pp. 265-320, Morgan Kaufmann, 1999.