모바일 라이프 특이성 추론을 위한 베이지안 확률 모델의 자동 학습

황금성, 조성배
연세대학교 컴퓨터과학과
yellowg@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Automatic Learning of Bayesian Probabilistic Model for Mobile Life Landmark Reasoning

Keum-Sung Hwang, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

다양한 기능과 센서를 탑재한 최신 모바일 디바이스는 사용자의 위치, 전화기록, SMS, 사진, 동영상 등 사용자에 관한 다양한 정보를 지속적으로 수집할 수 있기 때문에 개인의 생활을 이해하고 다양한 서비스를 제공할 수 있는 가능성을 가지고 있다. 하지만, 모바일 장치의 성능 제약 및 환경 불확실성으로 인해 아직까지 많은 연구 과제들이 남아 있다. 본 논문에서는 이러한 모바일 환경의 문제를 극복하기 위해 베이지안 네트워크를 이용한 라이프 로그 분석 모델 및 자동 학습 방법을 제안한다. 제안하는 베이지안 네트워크 모델은 모듈화 되어서 계산량은 감소되었으며, 자동 학습 방법을 통해 지속적인 업데이트가 가능하다. 이는 제안하는 방법이 복잡한 확률 모델을 자동으로 분할하는 방법과 분할된 상태에서의 유기적인 추론 방법을 포함하고 있기에 가능하다. 실험에서는 실제 모바일 장치에서 수집된 로그 데이터를 이용하여 제안하는 방법에 의한 실험 결과를 분석하고 분할을 통한 효율성 향상을 논의 한다.

1. 서론

최근 디지털 개인화 기기의 발전과 더불어 모바일 디바이스에서 많은 정보를 다루거나 수집할 수 있게 되었다. 이러한 모바일 로그는 통화, SMS 전송, 사진 촬영, MP3 청취, GPS 이동 기록과 같은 다양한 정보를 포함한다. 그리고 모바일 디바이스는 사용자가 계속해서 몸에 지니고 다니기 때문에 지속적인 관찰 및 로그 수집이 가능한 개인성이 강한 장비이므로 사용자의 일상정보를 효과적으로 수집하고 분석하여 사용자에게 도움을 줄 수 있다[1]. 이러한 모바일 디바이스의 특성은 사용자 편의를 위한 다양한 서비스 제공의 가능성을 열어 주었고, 최근에는 많은 연구자들의 관심을 받고 있다. 특히, 최근 인간 중심의 기술로서 활발하게 연구되고 있는 컨텍스트 어웨어 기술은 모바일 환경에서 더욱 많은 활용 가능성을 보이고 있다[2][3].

하지만, 모바일 디바이스는 PC에 비해서 적은 메모리 용량, 적은 CPU 처리용량, 작은 화면 크기, 불편한 입력 인터페이스, 제한된 배터리 용량 등의 한계를 가지고 있다. 또한, 변화가 많고 불확실한 실세계 환경에서 동작한다. 따라서 효율적인 계산 방법과 효과적인 관리 기법이 요구된다.

본 논문에서는 모바일 환경에서 수집된 로그 정보를 효과적으로 분석하고 효율적으로 고수준의 의미정보, 특히 특이성(Landmark, 특별히 기억에 남을 정보)을 추출하기 위한 방법을 제안한다. 제안하는 방법은 모바일 환경에서 발생하는 다양한 불확실성을 효율적으로 다루기 위해 협력적 모듈형 베이지안 네트워크(Cooperative Modular Bayesian Network)[4][5] 모델을 채택하였으며, 모바일 환경에서 지속적이고 효과적으로 학습하고 동작할 수 있도록 하기 위한 학습 방법을 제안한다.

1.1 이전 연구

앞선 연구[4]에서 저자들은 이미 모바일 라이프 로그를 분석해서 특이성으로 추로하기 위해 협력적 모듈형 베이지안 네트워크를 사용하는 방법을 제안하였다. 본 논문은 그 연구[4]를 바탕으로 확장 진행되었으며, 수집된 로그 데이터를 바탕으로 모듈형 베이지안 확률모델을 자동으로 학습하는 방법을 추가로 다루고 있다.

1.2 관련 연구

최근 로그 정보를 확률적으로 분석하여 향상된 서비스를 제공하고자 하는 연구로, 2006년 A. Krause 등이 수행한 연구가 있다[6]. 이들은 모바일 디바이스에서 수집된 센서 및 로그 정보를 클러스터링하고 사용자의 기호를 반영하는 컨텍스트에 부합하도록 학습시켜 사용자의 상황을 추론 하고, 컨텍스트에 대한 서비스 선택 방법으로 베이지안 네트워크(BN)를 사용하였다.

2005년에는 E. Horvitz 등이 수행한 연구로서[7], 베이지안 네트워크 기술을 기반으로 PC의 로그 데이터를 분석하여 인간 인식 활동을 모델링하고 특이성을 추론하는 방법이 있다.

하지만 이러한 연구들은 전통적인 베이지안 네트워크 모델을 사용하기 때문에 도메인이 커지면 높은 복잡도의 연산을 요구한다. 따라서 모바일 환경에 좀 더 최적화된 방법이 요구된다.

2. 모바일 라이프 로그에서의 특이성 추론

본 논문에서 모바일 라이프 로그를 수집해서 분석하고 시각화하는 과정은 그림 1과 같다. 모바일 장치에서 수집되는 정보는 GPS 위치 정보, Call/SMS 사용 정보, 사진 촬영 정보, MP3 플레이 기록, 장치 이용 기록(충전 정보)이며, 웹에서는 날씨 정보가 수집된다. 사용자에게서 수집되는 정보는 프로필과 PIMS 형태이며, 주로 사용자의 집이나 직장, 학교의 위치와 주변인의 연락처 정보를 얻기 위해 사용된다.

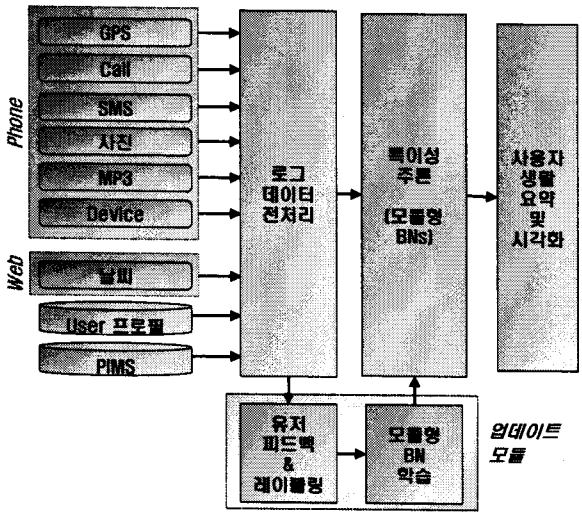

그림 1. 모바일 로그에서의 특이성 추출

2.1 협력적 모듈형 베이지안 네트워크

베이지안 네트워크는 노드의 연결 관계를 표현하는 방향성 비순환 그래프(DAG: directed acyclic graph) 형태이며, 이 구조에 따라 정의된 조건부 확률 테이블(CPT: conditional probability table)에 의해 적은 비용으로 많은 확률 관계를 효율적으로 표현 및 계산할 수 있는 모델이다[5]. 본 논문에서는 이전 논문에서 소개한 바 있는 모바일 환경에서 효율적 동작이 가능한 협력적 모듈형 베이지안 네트워크를 채택하여 사용하였으며, 다음과 같은 두 가지의 특징을 가진다.

첫째, 그림 2와 같이 확률 추론 모델을 분할된 도메인에 따라 모듈화하여 사용한다. 베이지안 네트워크의 특성상 노드와 연결의 수가 많아질수록 더 복잡한 계산을 요구하게 된다. 특히, 하나의 노드에 여러 원인 노드가 연결될 경우 복잡도가 $O(k^N)$ (k는 상태의 수, N은 부모의 수)에 비례하기 때문에 BN이 작을수록 모바일 환경에 유리하다. 따라서 모듈화될수록 복잡도가 감소하고 효율적인 동작이 가능해진다. 제안하는 협력적 모듈형 BN은 가상 연결 방법을 통해 협력적인 추론을 가능하게 한다.

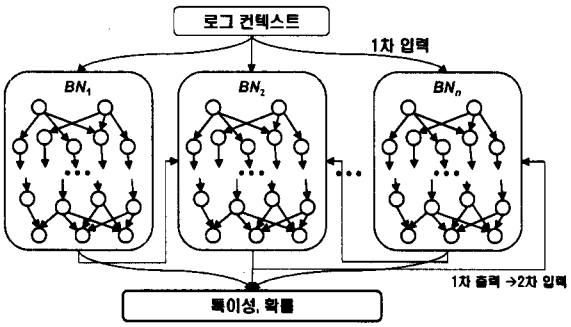

그림 2. 모듈화된 베이지안 네트워크. 점선은 가상 연결에 의한 정보 전달을 표현한다.

두 번째, 모듈화된 BN에서의 상호 인과성을 반영하기 위해 가상 연결에 의한 협력 추론을 지원한다. 이 방법은 다른 노드에서 얻어진 확률 정보를 다른 BN의 확률 증거로 반영하여 사용하는 방법으로 그림 3과 같은 2가지 형태의 버전을 가지고 있다[4]. 왼쪽은 모든 노드에서 사용 가능하며, 가상 노드를 자식으로 두고 뒤 CPT 값을 증거로 세팅하는 방법이고, 오른쪽은 루트 노드인 경우 가상노드 없이 CPT 값을 직접 수정하는 방법이다.

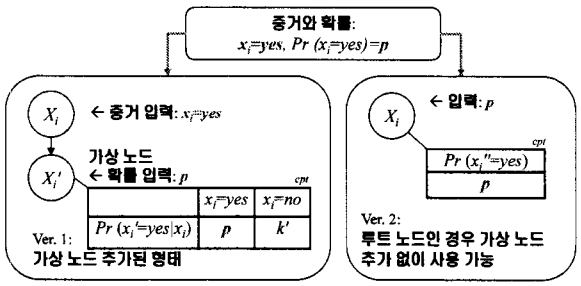

그림 3. 베이지안 네트워크에 확률 증거를 사용하는 방법. 증거로 x_i=yes (확률 p)가 들어온 경우 사용법을 나타낸다.

2.2 베이지안 네트워크의 학습

베이지안 네트워크는 노드의 연결 관계를 표현하는 방향성 비순환 그래프(DAG: directed acyclic graph) 형태를 가지고 있으며, 이 구조에 따라 정의된 조건부 확률 테이블(CPT: conditional probability table)에 의해 적은 비용으로 많은 확률 관계를 효율적으로 표현한다. 베이지안 네트워크 모델은 네트워크 구조를 나타내는 B_s와

파라미터 집합을 나타내는 θ를 이용해 (B_s,θ) 쌍으로 정의할 수 있다. 여기서 $\theta=\{B_\Phi, B_p\}$는 조건부 확률 테이블 BΦ와 초기 확률 분포 B_p로 구성된다. 본 논문에서는 B_s를 제안하는 구조 학습을 통해 구성하고, 파라미터 θ는 학습 데이터 집합 D로부터 수식 (1)과 같은 방법으로 계산한다.

$$\theta^* = \arg\max_{\theta} P(D|\theta)P(\theta) \quad (1)$$

여기서 $P(\theta)$는 초기 확률을 의미한다. 기본적인 학습 과정은 다음과 같다. 만약 $Z_T=\{z_1, z_2, \cdots, z_T\}$가 T개의 상태 변수를 나타내고, Y_T가 실제로 측정된 T개의 변수라면 수식 (2)와 같은 관계를 가지게 된다.

$$P(Z_T,Y_T,\theta) = P(Y_T|Z_T,B_\phi)P(Z_T|B_p) \quad (2)$$

여기서 조건부 확률 테이블 B_ϕ는 측정값과 상태변수의 조합이 얼마나 있었는지의 빈도를 조사하여 정의가 가능하다. 즉, 학습 데이터의 분포에 대한 히스토그램을 분석하여 빈도를 확률로 계산하고 베이지안 네트워크의 파라미터를 학습한다.

학습으로 구성된 베이지안 네트워크에서 주어진 증거 집합 E의 추론 결과 h에 대한 확률 Bel(h)는 Bayes' rule에 의해 수식 (3)과 같이 계산된다[5].

$$Bel(h) = P(h|E) = \frac{P(E|h)P(h)}{P(E)} = \frac{P(h \wedge E)}{P(E)} \quad (3)$$

2.3 베이지안 네트워크의 모듈화

본 논문에서는 학습된 베이지안 네트워크를 그대로 사용하지 않고 모듈화해서 사용한다. 이를 위한 규칙은 그림 4와 같으며, 어떠한 BN 구조 학습 알고리즘을 쓰더라도 다음과 같은 과정을 통해 모듈 분화가 가능하다. 그림에서 'E'는 증거 노드, 'L'은 특이성 노드, 'V'는 가상 증거 노드를 나타낸다.

2.4 분류 기준치 최적화

분류기의 성능은 분류 기준치에 좌우된다. 일반적으로 베이지안 네트워크를 이용한 분류기의 경우 확률값이 50%를 넘는 상태를 결과 상태로 선택하지만, 모바일 라이프 로그 데이터의 경우 실세계의 확률 분포를 얻을 수 있을 만큼 충분한 학습 데이터를 수집하기에 어려움이 있기 때문에 분류 기준치의 고려가 필요하다. 따라서 본 논문에서는 결과노드의 최적 분류기준을 학습 데이터의 테스트를 통해 결정되도록 하였다. 예를 들어 '이동중' 노드의 경우 기준치를 10%로 했을 경우 학습데이터에 대해 최적의 분류 성능을 보였다.

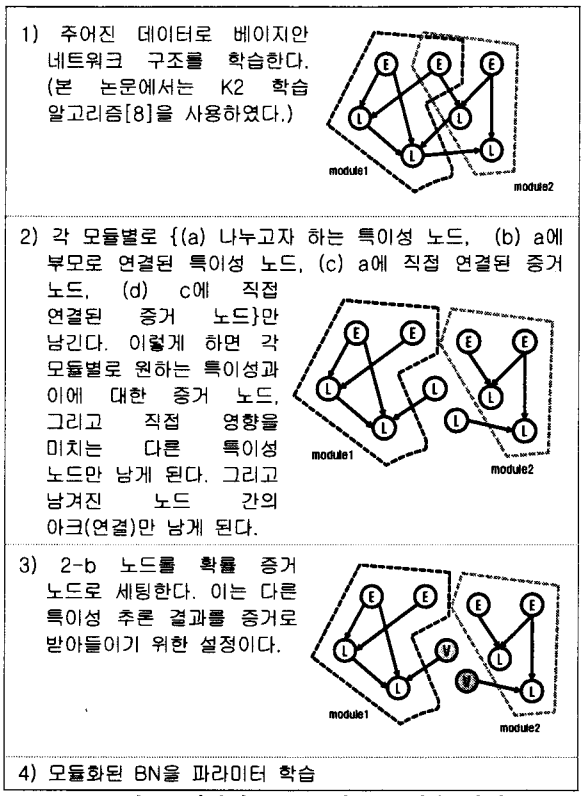

그림 4. 제안하는 모듈형 BN 학습 과정.

3. 실험 및 결과

본 장에서는 제안하는 방법을 이용한 실험 결과를 통해 성능을 평가한다. 이를 위해 실제 모바일 환경에서 수집된 데이터를 사용하였다. 실험에서는 모듈화된 BN과 모듈화되지 않은 BN의 성능을 비교하였다.

3.1 실험 데이터

실험을 위해 수집된 로그 데이터는 {GPS, call, SMS, 사진, MP3 Player, 날씨}이다. 3명의 여대생이 로그 수집 기능이 있는 스마트폰을 가지고 약 3주간 생활하였으며, 그 중에서 GPS 로그 정보가 제대로 수집된 16일을 선택하여 실험하였다. 선택의 기준은 다음과 같다.

1) 하루의 시작 지점과 끝 지점은 1km 이내이다.
2) 하루 동안 수집된 GPS 데이터는 적어도 5분 이상 머문 3종류 이상의 다른 장소를 포함한다.

실험에 참여한 사용자는 매일 자신이 행동한 내용을 시간과 함께 일지로 기록을 남겼으며, 이를 바탕으로 로그에 특이성 정보를 레이블링하여 실험을 수행하였다.

이렇게 수집된 데이터는 10분 단위로 분할되어 하루에 144개씩 총 2,304개가, 그리고 중복된 데이터를 제거하여 총 779개가 남았다. 실험을 위해 레이블링 된 특이성은 총 48종류가 사용되었다. 실험을 위해 사용되는 데이터의 수가 BN의 규모에 비해 부족하였기 때문에, 1개의 테스트 데이터를 번갈아가면서 검증하는 Leave-One-Out 검증 방법을 사용하였다.

3.2 특이성 추론 성능 평가

실험에서 모듈로 분할하기 위해 사용된 도메인은 4종류이며 각각 {감정 및 상태 BN, 일상생활 BN, 이벤트 BN, 학교생활 BN}으로 학습이 되었다. K2 학습 알고리즘에서 학습 파라미터로 정의되는 부모의 수 제한은 8로 정의하였다. 이때, 실험에서 노드의 이름은 표 1과 같은 인덱스를 사용하여 짧게 표현하였다.

표 1. 학습된 BN의 노드 이름 인덱스

ID	이름	ID	이름	ID	이름	ID	이름	ID	이름	ID	이름
N001	공공기관	N021	01시	N041	21시	N061	사진찍음	N081	수업	N101	식당
N002	공원	N022	02시	N042	22시	N062	음악듣기	N082	식사	N102	지하철
N003	관람지역	N023	03시	N043	23시	N063	음악오래들기	N083	줄거움	N103	편의점
N004	대학강의동	N024	04시	N044	날씨=맑음	N064	이동중	N084	친구만남	N104	술품
N005	대학교	N025	05시	N045	낮	N065	찾은전화	N085	길기리는 구장	N105	시험
N006	대학교강의동	N026	06시	N046	목요일	N066	전화받기	N086	라스토랑	N106	졸림
N007	도로	N027	07시	N047	밤	N067	전화아주많음	N087	쇼핑지역	N107	집책만남
N008	버스정류장	N028	08시	N048	봄	N068	전화안받음	N088	외식지역	N108	집책집
N009	영화관	N029	09시	N049	새벽	N069	출근중	N089	유흥지역	N109	학생회관
N010	이동속도=10 150km	N030	10시	N050	아침	N070	평균속도20 km미만	N090	커피점	N110	휴업상담
N011	이동속도=10 20km	N031	11시	N051	오전	N071	교통체증발생	N091	사진많이찍음	N111	교회
N012	이동속도=10 3km	N032	12시	N052	오후	N072	퇴기	N092	나갈준비	N112	친구집
N013	전람역	N033	13시	N053	일정시간	N073	부딪힘	N093	데이트	N113	과외
N014	중학교	N034	14시	N054	저녁	N074	외식	N094	산역	N114	결혼식장
N015	집	N035	15시	N055	주중	N075	집안생활	N095	식사(외식)	N115	지각
N016	학교	N036	16시	N056	현식	N076	걷기	N096	고등학교	N116	집중
N017	학교생활	N037	17시	N057	혼잡상태=친한친구	N077	공부	N097	피곤	N117	식당(중식)
N018	학교일	N038	18시	N058	GPS잡힘	N078	등교	N098	등대문	N118	자택식사
N019	휴식지역	N039	19시	N059	SMS상대=모르는사람	N079	버름	N099	휴일		
N020	00시	N040	20시	N060	call사용	N080	쇼핑	N100	휴일		

그림 5~7은 K2 학습 알고리즘과 제안하는 모듈형 BN 학습 알고리즘을 이용하여 학습한 결과 얻은 베이지안 네트워크를 나타낸다. 그리고 표 2는 학습된 BN의 크기 및 복잡도 비교를 나타낸다. 일체형 BN은 115개의 노드와 298개의 부모의 수, 그리고 4,396개의 확률 파라미터를 가지고 있었으며, 모듈형 BN은 BN당 평균 40개의 노드, 182개의 부모, 그리고 1,336개의 확률 파라미터를 가지고 있었다. 따라서 $O(k^N)$ (k는 상태의 수, N은 부모의 수)에 비례하는 BN의 계산 복잡도를 고려할 때 모듈형 BN의 경우 계산 복잡도가 크게 줄어들 것임을 알 수 있다.

표 3은 특이성 추론 및 평가 결과를 보여준다. 일체형 BN(monolithic BN)과 모듈형 BN(modular BN)의 성능을 비교하였으며, True Positive, True Negative, False Positive, False Negative 오류와 함께 정확률(precision rate, TP/(TP+FP))과 재현율(recall rate, TP/(TP+FN), 그리고 정답 일치율(hit rate, (TP+TN)/(TP+TN+FP+FN))을 평가하였다.

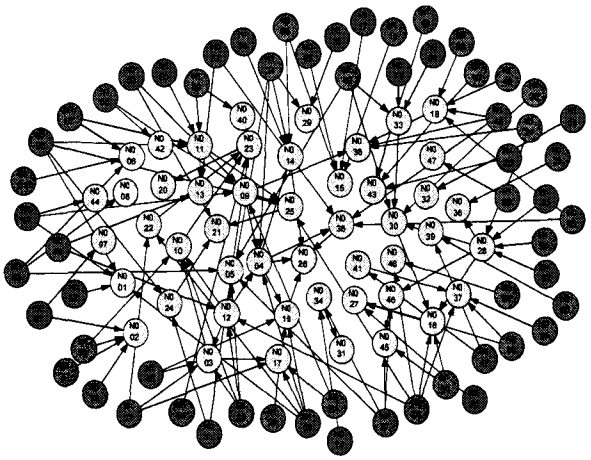

그림 5. 학습된 일체형 BN.

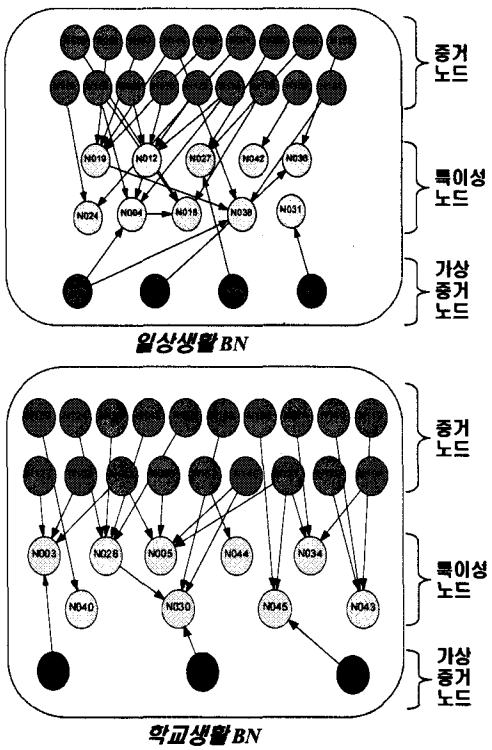

그림 6. 제안하는 방법으로 학습된 모듈형 BN 중 일상생활 BN과 학교생활 BN.

실험 결과를 보면 제안하는 모듈형 BN이 모듈화하기 전의 BN과 비슷한 성능을 보이고 있음을 알 수 있다.

일치율은 같았으며, 재현율은 0.001의 감소를 보였으나 정확률은 오히려 0.026이 상승한 것을 알 수 있다. 모듈화할 경우 BN 대상인 확률 파라미터의 수가 줄어들기 때문에 충분하게 주어지지 학습 데이터에 대해 더 좋은 학습 효과를 얻을 수 있기 때문인 것으로 보인다.

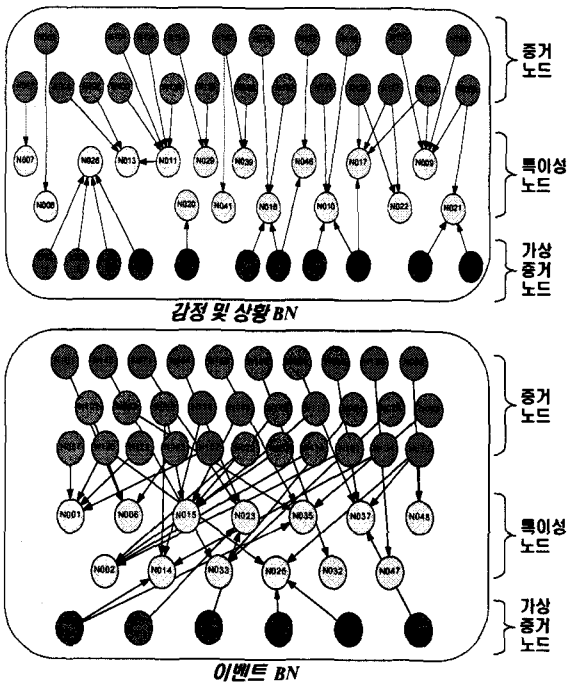

그림 7. 제안하는 방법으로 학습된 모듈형 BN 중 감정 및 상황 BN과 이벤트 BN.

표 2. 학습된 BN의 구조 및 복잡도 비교.

BN	노드 #	노드 $\#_{avg}$	부모 #	부모 $\#_{avg}$	cpv $\#_{avg}$
Mono BN	115	115	298	2.59	4,396
Mod. BNs	160	40	182	1.14	1,336

※ # - 개수, #avg - 평균개수, mod. - 모듈형

표 3. 추론된 특이성의 정확도 비교. 밑줄로 표시된 부분은 동일한 성능을 보일 경우를 나타낸다.

BN	TP	TN	FP	FN	Prc.	Rec.	Hit rt.
Mono BN	135	35,845	64	1,348	0.678	0.091	0.962
Mod. BN	133	35,853	56	1,350	0.704	0.090	0.962

※ TP - true positive, TN - true negative, FP - false positive, FN - false negative, Prc. - 정확률, Rec. - 재현율, Hit rt. - 정답 일치율

5. 결론 및 향후 연구

본 논문에서는 모바일 장치로부터 수집한 사용자의 로그 정보를 분석하여 특이성을 추론하는 모듈형 베이지안 네트워크를 자동으로 학습하는 방법을 소개하였다. 제안하는 방법을 이용해 실제로 수집된 모바일 라이프 로그를 분석하여 특이성을 추론하고 그 성능을 평가하였다. 제안하는 모듈형 베이지안 네트워크 모델링 방법을 사용할 경우, 여러 개의 베이지안 네트워크로 분할되어 동작하기 때문에 추론 복잡도를 감소시키는 효과를 얻을 수 있다. 실제 실험에서는 비슷한 특이성 추론 성능을 보였으며, 일부는 더 좋은 성능을 보이기도 하였다.

본 논문에서 사용한 학습데이터는 사용자가 아닌 연구자가 모든 특이성을 레이블링 한 다음 실험 데이터로 사용하였다. 하지만, 실제 환경에서는 사용자가 모든 데이터에 레이블링을 수행하기는 어렵다. 따라서 향후에는 사용자가 부분적으로 제공한 정보나 피드백을 통해 베이지안 네트워크를 부분적으로 학습하는 방법에 대한 연구가 필요하다.

감사의 글

본 연구는 LG전자의 산학협동과제를 통해 지원되었습니다."

참고문헌

[1] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen, "ContextPhone: A prototyping platform for context-aware mobile applications," *IEEE Pervasive Computing*, vol. 4, no. 2, pp. 51-59, 2005.

[2] A. Oulasvirta, "Finding meaningful uses for context-aware technologies: The humanistic research strategy," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, pp. 247-254, 2004.

[3] G.D. Abowd and E.D. Mynatt, "Charting past, present, and future research in ubiquitous computing," *ACM Trans. Computer-Human Interaction*, vol. 7, no. 1, pp. 29-58, 2000.

[4] K.-S. Hwang, S.-B. Cho, "A Bayesian inference model for landmarks detection on mobile devices," *Journal of Korea Information Science Society: Computing Practices*, vol. 13, no. 1, pp. 35-45, 2007. 02.

[5] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, 2003.

[6] A. Krause, A. Smailagic, and D. P. Siewiorek, "Context-aware mobile computing: Learning context-dependent personal preferences from a wearable sensor array," *IEEE Trans. on Mobile Computing*, vol. 5, no. 2, pp. 113-127, 2006.

[7] E. Horvitz, P. Koch, R. Sarin, J. Apacible, and M. Subramani, "Bayesphone: Context-sensitive policies for inquiry and action in mobile devices," *Proc. of the Conf. on User Modeling*, pp. 251-260, 2005.

[8] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309-347, 1992.