모바일 컨텍스트 로그를 사용한 베이지안 네트워크 기반의 랜드마크 예측 모델 학습

이병길[0] 조성배
연세대학교 컴퓨터과학과
{byulyi[0], sbcho}@sclab.yonsei.ac.kr

Learning Predictive Model of Memory Landmarks based on Bayesian Network Using Mobile Context Log

Byung-Gil Lee[0] Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

유비쿼터스 환경의 발달과 함께 모바일 장비에서 수집되어지는 컨텍스트 로그를 활용한 연구가 활발히 진행되고 있다. 하지만 기존의 컨텍스트 정보를 사용한 연구는 사용자 모델링에 그 초점을 맞추거나 단순하게 수집된 정보를 정리하여 한눈에 알아보기 쉽게 보여주는 정도에 그치고 있다. 본 논문에서는 사용자에게 새로운 서비스를 제공하기 위한 방법으로서 모바일 컨텍스트 로그와 외부 센서를 통해 정보를 수집하여 학습한 베이지안 네트워크를 이용하여 랜드마크를 찾아내는 예측 모델을 제안한다. 베이지안 네트워크 설계는 사전에 수집된 컨텍스트 정보를 요일과 주별로 분류하여 각각에 대한 베이지안 네트워크를 cross validation하여 랜드마크 예측에 대한 정확도를 평가하였다. 그리고 분류에서 가장 많이 사용하고 있는 SVM 방법을 사용하여 제안한 방법과의 성능을 비교·평가하였다. 랜드마크 예측에 대한 정확도는 주간별로 설계한 베이지안 네트워크보다 요일별로 설계한 베이지안 네트워크가 랜드마크를 예측하는데 정확도가 높음을 확인하였고, 베이지안 네트워크를 사용한 방법이 SVM을 사용한 방법보다 예측에 대한 정확성이 우수하였다.

1. 서론

유비쿼터스 컴퓨팅 기술이 발전함에 따라 컨텍스트 기반 응용 서비스에 대한 관심이 높아지고 있다. 이러한 서비스는 사용자 혹은 주변 환경의 변화 등과 같은 컨텍스트 정보를 파악하여 사용자의 직접적 명령 없이도 서비스를 실행, 변경, 정지하는 등의 기능을 제공한다. 컨텍스트(context)란 사전적 의미로는 문맥이나 환경, 정황 등을 의미하는데 유비쿼터스 네트워킹에서 가지는 의미는 사용자가 처한 환경을 컴퓨터가 인식하는 것부터 시작하여 그 환경 아래에서 사용자의 현재 위치, 행동 및 작업 등에 대한 사용자 정보 값과 그 정보들의 변화를 표현한 모든 정보를 통칭하며 이 정보를 얻어내는 과정을 컨텍스트인식(context-awareness)이라 한다[1].

본 논문에서는 모바일 장비로부터 수집 가능한 컨텍스트 정보를 활용하여 개인의 하루 생활에 대해 자동으로 랜드마크를 추론한다[2]. 여기서 말하는 랜드마크의 사전적 의미는 표지물이란 의미로, 주위의 경치 중에서 두드러지게 눈에 띄기 쉬운 특이성을 가지는 사물을 의미한다. 본 논문에서는 사용자가 경험한 많은 이벤트들 중 특별한 에피소드에 대한 기억에 도움을 주는 키워드를 의미한다. 또한 랜드마크는 이벤트가 발생한 장소 및 시간, 같이 경험한 사람, 그리고 이벤트가 일어난 시점 전후 또는 일어난 동안에 저장된 이벤트들에 대한 정보에 의해 구성되고 사용자가 에피소드를 기억하는데 도움을 준다.

2. 관련연구

기존에 진행된 연구를 보면 마이크로소프트 연구소의 경우 온라인 캘린더 정보와 사용자 정보만을 사용하여 베이지안 네트워크 기반에서 랜드마크 예측에 관한 연구를 진행하였고[3], MIT 미디어 연구실에서는 스마트 폰에서 수집할 수 있는 컨텍스트 로그를 사용하여 HMM (Hidden Markov Model) 기반으로 사용자 모델링 연구를 진행하였다[4]. 이 두 연구에 대한 비교는 표 1과 같다.

표 1. 관련 연구 비교

	MS Research	MIT Media lab.
목적	랜드마크 정확도 성능 평가	특정위치에서의 상황 예측
특징	사용자별로 모델 평가	해당 요소들 간의 관계분석 오류난 데이터 제거하고 사용
약점	직접 입력하는 정보로 인해 오류 발생 가능성 내포	생활패턴 분석에 한정됨
데이터	온라인 캘린더, 사용자 정보	스마트 폰의 컨텍스트 로그
모델	Bayesian Network	Hidden Markov Model Gaussian mixture model
학습	전체 데이터 중 8/10사용	전체 데이터 중 1/9 사용
테스트	전체 데이터 중 2/10사용	전체 데이터 중 8/9 사용

하지만 각 연구에는 표 1과 같은 문제점이 존재한다. 마이크로소프트 연구소의 경우는 사용자의 변화에 재빨리 대응할 수 없고 또한 사용자 정보를 사용자가 직접 입력해야만 한다. 입력이 잘못될 경우 예상하지 못한 결과가 발생되는 문제를 가지고 있고, MIT 미디어 랩의 경우 특정위치(집, 직장, 그 외)에서 사용자 상황예측의 대상이 연구소 직원이나 그 근처 다른 연구소 직원들로 한정되어 있으며, 특별한 방법을 사용하지 않고 단순히 학습된 상황을 사용해 분석 평가하므로 일반적인 평가가 불가능하다는 문제가 있다.

따라서 본 논문에서는 보다 정확한 랜드마크를 예측하는 모델의 학습을 보다 효율적으로 수행하기 위한 방법을 제안한다. 또, 들어난 문제점을 보완하여 수집된 컨텍스트 로그 데이터로

부터 네트워크를 구성한 후 성능을 비교·평가한다.

3. 베이지안 네트워크를 이용한 랜드마크 예측 모델

3.1 베이지안 네트워크

베이지안 네트워크는 변수들 간의 원인과 결과 관계를 확률적으로 모델링하기 위한 도구로서 불확실한 환경에서 좀 더 신뢰성 있는 결과를 추론하기 위해 쓰이는 대표적인 방법이다. 이 방법은 그림 1과 같이 변수를 나타내는 노드와 변수 사이의 상관관계 또는 인과관계를 나타내는 연결선(arc, edge)으로 이루어져 있는 방향성 비순환 그래프(directed acyclic graph, DAG)이다.

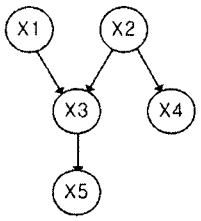

그림 1. 간단한 베이지안 네트워크 구조의 예

각 노드는 조건부확률을 나타내는 테이블을 가지고 베이지안 규칙(Bayes' Rule)을 이용해 계산한다. 만약 두 노드 사이의 연결선이 없다면 두 노드는 서로 독립적이라는 의미로 해석한다. 노드의 부모와 확률값이 정해지면 조건부확률 테이블을 가지고 계산을 하게 되는데, 기존 확률 이론에 비해 훨씬 적은 양의 연산으로 확률 추론이 가능하다. 그림 1에서 결합 확률값인 $P(x_1, x_2, x_3, x_4, x_5)$는 다음과 같이 계산한다.

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_2)P(x_5|x_3)$$

3.2 베이지안 네트워크 예측모델 학습

일반적으로 베이지안 네트워크 학습은 NP-hard 문제로서 해석적 방법으로 해결할 수 없기 때문에 발견적 방법들(heuristic methods)를 사용해야 한다.

1. *Initialization:*
(a) **for** $i=1$ **to** n **do:** $Pa(x_i)=\emptyset$
(b) **for** $i=1$ and $j=1$ **to** n **do:**
if $(i \neq j)$ **then** $A[i,j]=f(x_i, x_j)-f(x_i, \emptyset)$

2. *Loop:*
(a) repeat
i. Select two indexes (i,j), such that,
$(i,j) \leftarrow \arg\max_{i',j'} A[i',j']$
ii. **if** $(A[i,j]>0)$ **then** $Pa(x_i)=Pa(x_i)\cup x_j$
iii. $A[i,j]=-\infty$
iv. **for all** $x_a \in Ancestors(x_i)\cup x_j$
and $x_b \in Descendants(x_i)\cup x_i$ **do:**
$A[a,b]=-\infty$
v. **for** $k=1$ **to** n **do:**
if $(A[i,k]>-\infty)$
then $A[i,k]=f(x_i, Pa(x_i)\cup x_k)-f(x_i, Pa(x_i))$
until $\forall i,j\,(A[i,j]\leq 0 \ or \ A[i,j]=-\infty)$

* f: scoring metric
* A[i,j]: adjacency matrix
* Pa(x_i): instantiated to its jth value

그림 2. ACO를 사용한 BN 학습

가장 많이 사용하고 있는 발견적 방법으로 탐욕적 검색 방법이 있다. 본 논문에서는 예측모델 학습방법으로 탐욕적 탐색 방법을 최적화한 알고리즘인 Ant Colony Optimization (ACO) 방법을 사용하였다[5]. 기본적인 구조는 그림 2에서 보듯이 알고리즘 B를 사용하여 베이지안 네트워크의 구조를 생성하고 K2 알고리즘을 사용하여 네트워크의 조건부 확률 테이블을 작성한다.

ACO 방법은 가장 짧은 경로를 탐색할 수 있는 협동적 행동에 근거하여 처음에는 다양한 경로를 조사하고 설정하지만 데이터에 의한 학습이 계속 진행될수록 가장 짧은 경로만을 설정하게 되는 최적화 방법으로 최근 많이 사용되고 있는 방법 중 하나이다. 그림 3은 ACO 방법을 사용하여 본 논문에서 사용한 데이터 집합을 가지고 학습된 예측 모델 중 한 예이다.

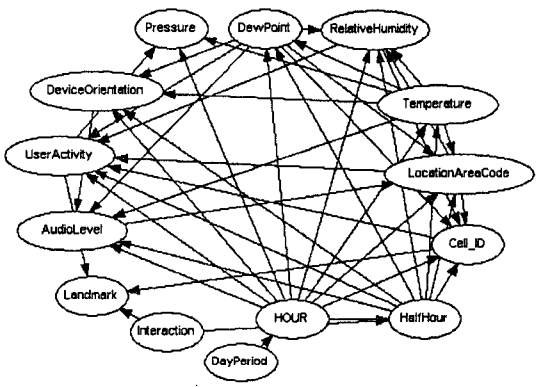

그림 3. 학습된 예측 모델의 예

그림 3의 예측 모델 구조에서 아크가 복잡하게 설정된 것은 데이터 집합을 통하여 베이지안 네트워크를 학습 시킬 경우 상관관계가 다양하게 나올 수 있기 때문에 복잡하게 구성되어진다.

3.3 베이지안 네트워크 상에서의 랜드마크 예측

본 논문에서 사용한 랜드마크 예측 과정은 그림 4와 같다.

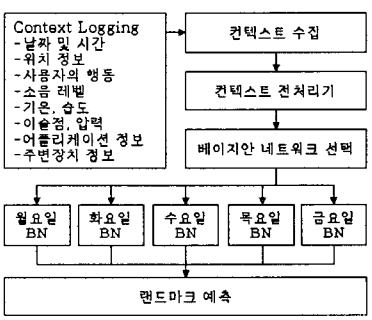

그림 4. 제안한 랜드마크 예측 과정

그림 4에서 보면 입력된 컨텍스트 정보를 베이지안 네트워크에 적용하기 위하여 전처리를 하고 전처리된 컨텍스트 정보에 해당하는 베이지안 네트워크를 선택하여 랜드마크인지 아닌지에 대하여 추론을 한다. 본 논문에서 사용한 추론 방법은 베이지안 네트워크에서 일반적으로 사용되는 정션트리 알고리즘이다. 이 알고리즘은 그래프 이론과 확률 이론 사이의 연결 분석에 기반을 두는 알고리즘으로 실질적인 DAG 대신에 정션트리라 불리는 데이터 구조를 사용하여 추론을 한다.

4. 실험 및 결과

4.1 실험 환경

실험에서 사용된 구조학습 방법은 앞서 소개한 ACO 방법이며, 예측 방법은 정션트리 알고리즘을 사용하였다. 실험에 사용한 노키아 데이터는 6주에 걸쳐 토요일과 일요일을 제외한 27일(11/3~12/12)동안 기록된 데이터이다. 총 15개의 속성, 약 21만개로 구성된 데이터를 전처리를 통하여 13개의 속성, 약 3만여개의 데이터로 재구성하였다. 단 데이터는 결측치(missing value)가 존재하지 않는다.

데이터는 요일에 맞게 분류한 경우와 로그가 발생한 시간 순서(주간별)로 분류한 경우로 나누어 비교 실험을 진행하였다. 다른 기계학습 도구와 성능을 비교하기 위해서 분류 문제를 해결할 때 가장 많이 사용되어지고 있는 SVM(Support Vector Machine) 방법을 사용하였다. 각 요일에 맞게 분류된 데이터 집합은 학습을 통하여 설계된 각각의 베이지안 네트워크에 파일 수 만큼 cross validation을 통하여 성능을 측정하고, 주간별로 학습한 경우는 트레이닝 데이터를 1주, 2주, …, 5주까지 점차적으로 늘려가면서 베이지안 네트워크를 설계하여 마지막 6주에 해당하는 데이터를 테스트 데이터로 사용하여 랜드마크 예측에 대한 성능을 측정한다.

4.2 예측 성능 비교

그림 5의 베이지안 네트워크와 SVM의 요일별 랜드마크 모델에서 예측 정확도와 표준편차를 분석해 보면 모든 요일에 대해 본 논문이 제안한 베이지안 네트워크 예측 방법이 SVM 방법보다 예측 정확도가 높게 나타났고 상대적으로 표준편차는 작게 나타나 성능이 더 우수하다는 것을 보여준다.

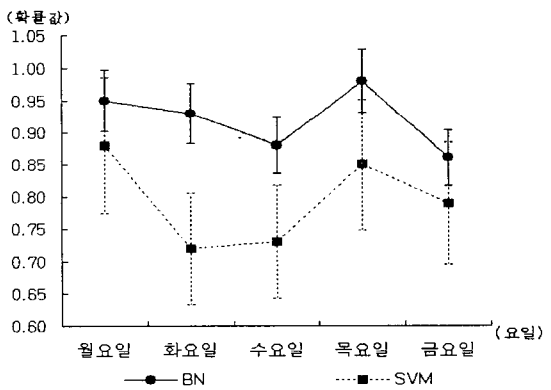

그림 5. BN과 SVM의 요일별 랜드마크 예측 정확도

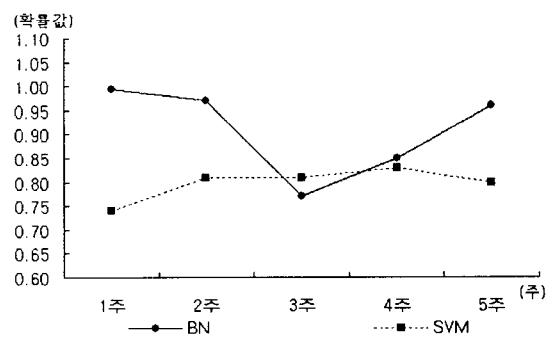

그림 6. BN과 SVM의 주간별 랜드마크 예측 정확도

그림 6에서 베이지안 네트워크와 SVM에 대한 주간별 비교를 보면 1~3주의 경우 베이지안 네트워크의 경우 SVM과 다르게 예측 정확도 성능이 점점 떨어지는 것을 알 수 있다. 이는 학습된 데이터의 양이 적을 경우 정확한 예측이 불가능하다는 것을 보여준다. 하지만 학습 데이터가 3주 이상 될 경우 주간별에 대해서도 베이지안 네트워크 모델이 SVM에 비해 높은 성능을 나타낸다.

요일별 데이터와 주간별 데이터를 통한 표 2의 분석을 보면 주간별로 학습시켜 베이지안 네트워크에서 랜드마크를 예측하는 모델은 정확도에서는 요일별 평균과 큰 차이를 보이지 않았지만 표준 편차는 약 2배 정도의 차이가 생긴다. 이것은 시간 순으로 전체 요일에 대하여 학습시킨 주간별 데이터에 대한 성능이 입력되는 요일에 맞는 예측 모델을 사용한 모델보다 예측에 대한 정확도가 떨어진다는 것을 의미한다. 이런 결과가 발생하는 원인은 MIT 미디어 연구실에서 수행한 연구 결과에 비추어 봤을 때 사람의 생활 패턴은 요일별로 유사하게 나타나기 때문이다.

표 2. 요일별 vs 주간별 성능비교

	요일별		주간별	
	BN	SVM	BN	SVM
랜드마크 예측율	0.92	0.79	0.91	0.80
표준편차	0.05	0.12	-	-
학습 및 추론 평균 소요 시간	0:06:13	0:00:34	01:43:59	0:00:35

5. 결론 및 토의

본 논문에서 베이지안 네트워크 기반의 랜드마크 예측 모델 학습 실험을 진행한 결과 대다수의 사람들은 규칙적인 라이프스타일을 가지면서도 그것을 구성하는 요소들은 상당히 불규칙한 요소들로 구성된다는 것을 알 수 있었다. 그러나 이러한 불규칙적인 요소들을 잘 분석하여 사용자가 요구하는 것에 정확한 결과를 보여주기 위해서는 네트워크를 구성할 때 많은 데이터가 필요한데 각 사용자에 맞는 데이터를 구성하기란 쉽지 않다. 이런 점에 비춰볼 때 충분한 데이터가 수집되기 이전에는 사용자가 하나하나 확인하여 그 정보를 다시 네트워크 구조에 반영할 수 있도록 동적인 구조로 설계하는 것이 더 효율적일 것이다.

향후 과제로는 본 논문에서 분석된 랜드마크 항목을 일반 사용자에게 쉽게 보여주고 이해할 수 있도록 하는 어플리케이션 개발이 필요하고 실시간으로 사용자에게 랜드마크 정보를 보여 주기 위하여 베이지안 네트워크 학습 및 추론에 대해 경량화하는 연구가 필요할 것으로 보인다.

6. 참고문헌

[1] A. K. Dey, "Context-aware computing: The cyber desk project," *Proc. of the AAAI 1998 Spring Symposium on Intelligent Environments (AAAI Technical Report SS-98-02)*, pp. 51-54, Mar 1998.

[2] R. M. Shiffrin and M. Steyvers, "A model for recognition memory: REM retrieving effectively from memory," *Psychonomic Bulletin & Review*, vol. 4, no. 2, pp. 145-166, 1997.

[3] E. Horvitz, S. Dumais, and P. Koch, "Learning predictive models of memory landmarks," *CogSci 2004: 26th Annual Meeting of the Cognitive Science Society*, Chicago, August 2004.

[4] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Journal of Personal and Ubiquitous Computing*, June 2005.

[5] L. M. de Campos, J. M. Fernrandez-Luna, J. A. Gamez, and J. M. Puerta, "Ant colony optimization for learning Bayesian networks," *Int. Journal of Approximate Reasoning*, pp. 291-311, 2002.