

Learning Context Awareness Model based on User Feedback for Smart Home Service

Seongcheol Kwon *, Seyoung Kim**, Kwang Ryel Ryu***

Abstract

Recently, researches on the recognition of indoor user situations through various sensors in a smart home environment are under way. In this paper, the case study was conducted to determine the operation of the robot vacuum cleaner by inferring the user's indoor situation through the operation of home appliances, because the indoor situation greatly affects the operation of home appliances. In order to collect learning data for indoor situation awareness model learning, we received feedbacks from user when there was a mistake about the cleaning situation. In this paper, we propose a semi-supervised learning method using user feedback data. When we receive a user feedback, we search for the labels of unlabeled data that most fit the feedbacks collected through genetic algorithm, and use this data to learn the model. In order to verify the performance of the proposed algorithm, we performed a comparison experiments with other learning algorithms in the same environment and confirmed that the performance of the proposed algorithm is better than the other algorithms.

▶Keyword: Smart home, Situation awareness, Semi-supervised Learning, Genetic Algorithm

1. Introduction

최근 사물 인터넷(Internet of Things, IoT)에 대한 관심이 높아짐에 따라 사물 인터넷의 한 분야인 스마트 홈에 대한 연구가 활발하게 진행되고 있다. 스마트 홈이란 사람의 주거환경에 IT 기술을 융합하여 사용자에게 안전하고 편리한 생활을 제공해주며, 나아가 사용자의 건강한 삶과 경제적 편익을 추구하는 지능화된 주거시설을 의미한다. 스마트 홈에서는 가정 내에 설치된 다양한 센서들로부터 수집된 자료로 사용자에게 서비스를 제공한다. 사용자에게 적절한 서비스를 제공해주기 위해서는 먼저 수집된 센서 데이터로부터 사용자의 상황에 대해 인지하는 기술이 필요하다.

스마트 홈 내 사용자의 상황을 인지하기 위해 다양한 센서 데이터를 활용한 기존 연구들이 있다. [1-2]에서는 카메라, [3]에서는 적외선 센서와 초음파 센서, 그리고 [4]에서는 door contact 센서와 동작 감지 센서 등 다양한 센서를 이용하여 사용자의 상황을 추론하는 모델을 생성했다. 하지만 관련 연구들에 사용된 센서들은 일반적인 가정에 설치하여 사용하기 어렵기 때문에, 본 논문에서는 전자식 전력 측정기인 스마트 미터를 이용하여 스마트 홈 내의 가전기기 동작 데이터를 수집하였다. 스마트 미터는 가정에서 사용하는 전력을 측정하는 장치지만, 스마트 미터에 측정되는 전력의 차이를 계산해 가전기기

• First Author: Seongcheol Kwon, Corresponding Author: Kwang Ryel Ryu

*Seongcheol Kwon (ksc4323@pusan.ac.kr), Dept. of Computer Science and Engineering, Pusan National University

**Seyoung Kim (birdzero@pusan.ac.kr), Dept. of Computer Science and Engineering, Pusan National University

***Kwang Ryel Ryu(krryu@pusan.ac.kr), Dept. of Computer Science and Engineering, Pusan National University

• Received: 2017. 05. 12, Revised: 2017. 06. 23, Accepted: 2017. 07. 18.

• This research was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (2012-0-00265, R0101-17-0129, Development of high performance IoT device and Open Platform with Intelligent Software)

의 동작여부를 실시간으로 알 수 있다.

가전기기 동작 데이터와 당시 사용자 상황 데이터를 이용하면 가전기기 동작에 따른 사용자 상황을 추론하는 모델을 학습시킬 수 있다. 하지만 가전기기 동작 데이터는 스마트 미터를 통해 쉽게 수집이 가능한 반면, 당시 사용자의 상황 데이터를 얻기 위해 대상 사용자를 관찰하며 직접 상황을 기록하는 방법을 사용하였다. 이런 방법은 연구에 적합할 수는 있으나 실생활에 적용하여 사용자에 적합한 상황 인지 모델을 만들기 힘들다는 단점이 있다. 이러한 문제를 해결하기 위해 [6-7]에서는 사용자의 피드백 데이터만으로 자동화 로봇 청소기의 청소 상황 인지 모델을 학습하는 방안을 제안하였다. 피드백 데이터란 로봇 청소기가 동작하고 있을 때 사용자가 동작을 중단시키는 경우, 혹은 로봇 청소기가 동작하지 않을 때 사용자가 직접 동작시키는 경우 수집되는 데이터이다. 피드백 데이터는 그 시점의 가전기기 동작 및 청소 가능 여부를 포함하는 데이터로 수집되는 빈도가 매우 낮아 소량이지만, 사용자가 청소 상황 판단 모델의 판단이 틀렸음을 직접적으로 알려주기 때문에 신뢰성이 매우 높은 데이터이다. [6-8]에서는 사용자의 상황과 가전기기 간 인과관계를 베이지안 네트워크로 표현하였다. [6]에서는 피드백이 발생한 시점의 데이터와 관련된 conditional probability table (CPT) entry들을 조절하는 알고리즘을 이용하여 베이지안 네트워크를 학습시켰다. [7]에서는 CPT entry 학습을 위해 유전 알고리즘을 이용하였다. 먼저 CPT entry들을 해로 표현하고, 수집된 피드백들에 대해 평가하여 피드백들을 최대한 만족시키는 방향으로 해를 진화시켰다. 하지만 두 방안 모두 피드백 데이터만을 이용하여 학습하였기 때문에, 학습된 모델이 소량의 피드백 데이터에 overfitting 하는 문제가 발생하였다. 또한 CPT를 직접적으로 수정하는 방안이기 때문에 실제 데이터와 무관한 모델을 학습할 가능성이 존재하였다.

본 논문에서는 [6-8]과 동일하게 케이스 스터디로 소량의 사용자 피드백 데이터를 이용하여 자동화 로봇 청소기의 청소 상황 판단 모델을 학습시키는 방안을 제안한다. 청소 상황 판단 모델로는 베이지안 네트워크에 비해 복잡도가 낮은 Naive Bayes 모델을 사용하였고, 피드백 데이터 외에도 사용자 상황 데이터가 없는 가전기기 동작 데이터를 학습에 이용하는 방안을 제시한다. 제안 방안의 성능 평가를 위해 [7]의 방안과 준지도 학습 (semi-supervised learning) [9] 방안을 비교 대상으로 선정하였다.

제안 방안의 성능 측정에 사용되는 실험 데이터를 만들기 위해 가정주부와 대학생의 생활 패턴을 가정하여 시뮬레이션을 진행하였고, 시뮬레이션을 통해 얻어진 데이터로 각 방안들의 성능을 측정하였다.

본 논문의 구성은 다음과 같다. 2장에서는 상황 인지에 대한 관련 연구에 대해 소개하였고, 3장에서는 대상 문제에 대해 정의하였다. 4장에서는 제안하는 유전 알고리즘을 이용한 Naive

Bayes 모델 학습 방안에 대해 소개하였고, 5장에서는 실험 결과를 분석하였고 제안 방안의 성능을 확인하였으며, 6장에서 결론을 서술하였다.

II. Preliminaries

1. Related Works

스마트 홈 내 사용자 상황 인지에 관한 연구는 다양한 형태로 수행되어 왔다. [1]은 카메라와 마이크를 이용한 상황 인지 연구로, 카메라는 사용자의 위치와 움직이는 속도, 자세를 검출하기 위해 이용되었고, 마이크는 사용자의 음성을 검출하기 위해 사용되었다. 이렇게 카메라와 마이크에서 얻은 데이터를 통해 사용자의 개인 작업, 낮잠, 게임, 발표 등의 상황을 SVM (Support Vector Machine) 을 이용하여 인지하였다. [4]의 연구는 door contact 센서와 동작 감지 센서를 통해 데이터를 수집하고, [1]과 마찬가지로 SVM 을 상황 인지에 사용하였다. [2]에서는 카메라를 이용하여 머리, 팔, 다리 등 몸의 각 부분과 관절의 움직임을 추적하고, hidden Markov model 을 이용하여 먼저 사람의 동작을 인지하였다. 그리고 시간의 흐름에 따라 인지된 동작들의 조합을 통해 Maximum likelihood 를 계산하여 사람의 상황에 대해 인지하는 방법을 사용하였다. [3], [5]에서는 적외선 센서와 초음파 센서, 조도 센서 등 다양한 센서를 이용하여 사용자의 위치, 자세, 시간 등의 데이터를 수집하여 사용자의 상황을 인지하는 연구를 수행하였다. [3], [5]에서는 사용자의 상황 분류를 위해 오프라인에서 학습한 규칙을 사용하였다. [1-5] 모두 다양한 환경에서 사용자의 상황을 인지하는 연구지만 공통적으로 실생활에 사용되기 어려운 센서를 이용한다는 문제점이 있다. 또한 [1], [4], [5]에서는 분류 대상인 사용자의 상황 데이터를 얻기 위해 직접 사용자를 관찰하고 상황을 기록하는 방법을 사용하였다. 그리고 [3], [5]에서 사용한 미리 학습된 규칙을 이용하여 상황을 분류하는 방안의 경우, 환경이 복잡해질수록 규칙이 증가하며, 오프라인 학습이기 때문에 규칙을 생성하거나 바꾸어 적용하기 힘들다는 단점이 존재한다.

이러한 문제들을 해결하기 위해 [6-7]에서는 스마트 미터를 센서로 이용하여 가전기기들의 동작 상태를 감지하고, 사용자의 피드백을 통해 상황 데이터를 수집하여 온라인 학습을 하는 방안을 제시하였다. 두 연구 모두 베이지안 네트워크를 [10] 이용하여 가전기기와 사용자 상황 간 인과관계를 모델링하였으며 [6]에서는 휴리스틱한 알고리즘을 통해 CPT를 조절하는 방법, [7]에서는 유전 알고리즘을 이용하여 CPT를 탐색하는 방법을 제안하였다.

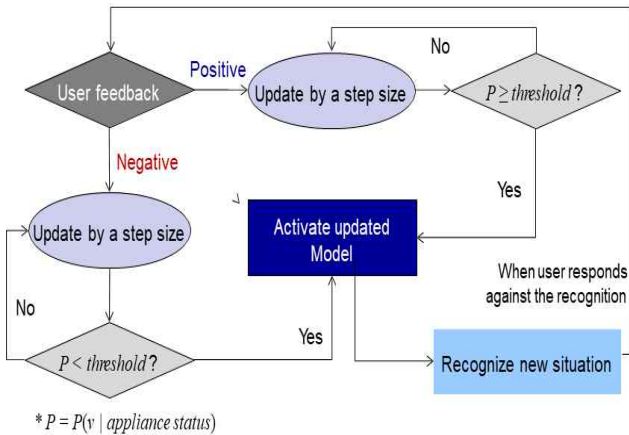


Fig. 1. Online Learning Method Based on User Feedback

[6]의 연구에서는 사용자의 피드백을 받으면 피드백과 관련된 CPT를 조절하는 알고리즘을 통해 학습을 진행하였다. Fig. 1에서 보는 것과 같이 청소 상황 판단 모델이 시간의 흐름에 따라 수집되는 데이터에 대해, 청소 가능 여부에 대해 판단하는 과정에서 모델의 판단이 사용자의 청소 상황과 일치하지 않을 경우 피드백을 주게 된다. 정지 중인 로봇 청소기를 동작시키는 행위를 positive 피드백, 청소 중인 로봇 청소기를 끄는 행위를 negative 피드백으로 정의하였고, positive 피드백을 받으면 threshold 보다 청소 확률이 높아지도록, negative 피드백을 받으면 threshold 보다 청소 확률이 낮아지도록 CPT entry 를 learning rate 만큼 조절하였다. CPT entry 조절 시 피드백과 관련된 노드들을 찾아 수정하였는데,

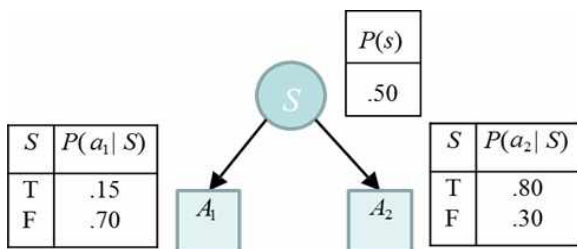


Fig. 2. Bayesian Network Example

예를 들어 Fig. 2와 같은 베이지안 네트워크에서 상단 S 노드의 값과 하단 A₁, A₂ 노드의 값이 모두 true일 때, 사용자의 negative 피드백이 발생하면 P(s|a₁, a₂)의 값은 감소해야 하며 P(¬s|a₁, a₂)의 값은 증가해야 피드백을 만족시키는 모델을 학습할 수 있다. Bayes' rule과 조건부 독립을 이용하면 식 (1), (2)와 같이 변형할 수 있다.

$$P(s|a_1, a_2) = aP(a_1|s)P(a_2|s)P(s).....(1)$$

$$P(\neg s|a_1, a_2) = aP(a_1|\neg s)P(a_2|\neg s)P(\neg s).....(2)$$

변형된 수식에서 P(a₁|s), P(a₂|s), P(s)를 베이지안

네트워크에서 찾아 감소시키고 P(a₁|¬s), P(a₂|¬s), P(¬s)를 증가시키면 negative 피드백을 만족하는 모델을 학습할 수 있다. [6]의 방안에서의 문제점은 초기 생성된 모델에 따라 학습 성능에 큰 차이를 보인다는 것과, 최근의 피드백을 만족시키기 위해 이전 피드백들을 무시하는 문제가 있었다.

feedback	Sensor Variables	V	fitness _i
feed ₁	s ₁ ¬s ₂ ¬s ₃ s ₄ ¬s ₅ s ₆ s ₇ ¬s ₈ ¬s ₉ ¬s ₁₀	Y	P(v feed ₁)
feed ₂	s ₁ s ₂ ¬s ₃ s ₄ s ₅ s ₆ ¬s ₇ ¬s ₈ ¬s ₉ s ₁₀	N	P(¬v feed ₂)
⋮	⋮		
feed _n	¬s ₁ s ₂ s ₃ ¬s ₄ s ₅ ¬s ₆ s ₇ s ₈ ¬s ₉ s ₁₀	Y	P(V feed _n)

$$solution = \underset{solution \in SOLUTION}{\arg \max} \prod_{i=1}^n fitness_i$$

Fig. 3. Evaluation Method for Searching the Cpt of Bayesian Network

[7]에서는 유전 알고리즘을 이용하여 베이지안 네트워크의 CPT를 탐색하는 방안을 제안하였다. 해의 길이는 총 CPT entry의 개수와 동일하며 해의 평가는 Fig. 3과 같이 탐색된 CPT로 모델을 만들어 수집된 모든 피드백 데이터에 대한 확률 곱을 통해 평가한다.

이 평가 방안을 도입하였을 경우 확률 곱을 최대화 하는 방향으로 진화하려는 경향이 있지만 수집된 피드백의 양이 늘어날수록 학습에 걸리는 시간이 증가하고, 최적 해를 찾기 힘들어 진다는 문제가 있었다.

또한 [6-7]의 공통된 문제점으로 베이지안 네트워크의 인과관계가 잘못 설정되면 네트워크를 신뢰할 수 없기 때문에 초기 네트워크 구축 시 관련 분야 전문가의 도움을 받아야 한다는 단점이 있다. 그리고 모델의 복잡도가 높은 반면, 수집할 수 있는 피드백 데이터의 양은 적기 때문에 피드백 데이터에 overfitting할 수 있는 위험성을 가지고 있다.

2. Problem Definition

본 논문에서는 Fig. 4에 보이는 것과 같이 1인 가정을 대상으로 가정에 설치된 센서를 통해 데이터를 수집 및 가공하고, 가공된 데이터를 이용하여 자동화 로봇 청소기의 청소 판단 모델을 실시간으로 학습시키는 문제를 다룬다. 이 주제는 최상위 어플리케이션에 대한 문제이며, 따라서 수집되는 데이터와 사용자의 상황에 대해 여러 가지 가정을 한다. 이 장에서는 대상 문제에 사용된 여러 가정들과 환경에 대해 설명한다.

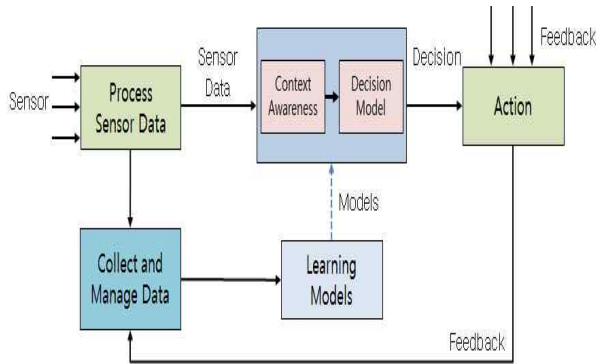


Fig. 4. Feedback Based Automation Robot Cleaner System

2.1 Method of Collecting User’s Data

일반적으로 사용자 상황 인지는 가정에서 얻을 수 있는 데이터와 당시 사용자의 상황 데이터를 이용하여 모델을 학습하고, 이후 들어오는 데이터에 대해 사용자가 어떤 상황인지 분류하는 문제이다. 본 논문에서 가정된 사용자 피드백 기반 학습 데이터 수집과 제안하는 방안의 성능 비교를 위해서는 사용자의 가전 기기 사용 현황과 그 시점의 청소 가능 / 불가능 상황이 명시된 데이터 셋이 필요하며, 이러한 데이터를 수집하는 방법으로는 크게 두 가지가 있다.

첫 번째로는 실측 데이터를 수집하는 방법이 있다. 실제 가정에 센서를 설치하고 사용자를 지속적으로 관찰하며 학습 데이터를 얻는 방법으로, 이 방법을 사용하면 실생활과 가장 밀접한 데이터를 얻을 수 있으며, 기록 시에 실수하지 않는다는 가정 하에 가장 신뢰도 높은 데이터를 수집할 수 있다. 그러나 데이터 수집을 위해 센서 설치비용과 사용자 상황을 모니터링 하는 인력이 요구된다. 따라서 비용이 많이 들고 장기간 데이터를 수집하기 어렵다.

두 번째로는 실제 가정과 유사한 시뮬레이터를 만들고, 가상의 사용자를 시뮬레이션 하여 데이터를 수집하는 방법이 있다. 시뮬레이션을 이용한 데이터 수집은 데이터에 대한 신뢰도가 다소 떨어질 수 있다는 단점이 있으나, 실제 환경 구축에 드는 비용과 데이터 수집에 필요한 인력을 줄일 수 있고, 원하는 단위 시간마다 안정적으로 데이터를 수집할 수 있다는 장점이 있다. 따라서 본 논문에서는 시뮬레이터를 이용하여 사용자 데이터를 수집하는 방법을 선택하였다.

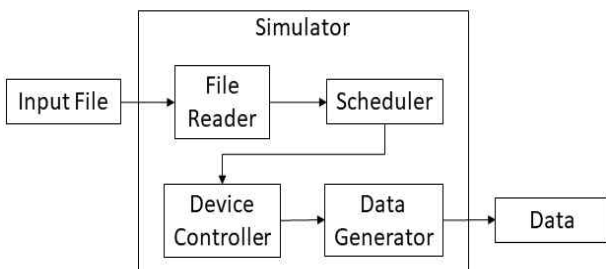


Fig. 5. Simulator Architecture for Collecting User Data

Fig. 5는 사용자 데이터 수집을 위한 시뮬레이터의 구조를 시각화한 것이다. Input File 에는 데이터 생성에 필요한 정보를 입력한다. 먼저 시간대 별 사용자의 상황 패턴을 설정한다. 그리고 상황이 발생할 확률과 상황 별 기기 제어 확률을 설정하여 입력해준다. File Reader 는 Input File 에서 입력 받은 정보를 읽고 Scheduler 로 정보를 전달한다. Scheduler 는 전달 받은 정보 중 시간대 정보와 시간대 별 사용자 상황 패턴 정보를 이용하여 사용자의 하루 일과를 샘플링 한다. 사용자의 하루 일과가 정해지면 Device Controller 로 전송하며 Device Controller 에서는 사용자의 일과에 따라 가전기기의 동작 상태를 결정한다. Data Generator 는 사용자의 일과와 가전기기 데이터를 통합하여 파일로 출력한다.

본 논문에서는 1인 가정을 대상으로 시뮬레이션을 진행하였다. 먼저 시간대를 아침, 낮, 밤으로 나누어 현실과 최대한 유사하도록 사용자의 시간대 별 상황과 상황 별 사용하는 가전기기, 해당 가전기기를 사용할 확률, 사용 후 꺼질 확률에 대한 정보를 생성하였다. 이렇게 생성된 정보가 시뮬레이터에 입력되면 시뮬레이터는 분 단위의 가전기기 동작 데이터와 당시 사용자의 상황과 청소 가능 여부 데이터를 생성한다.

실험에 사용할 데이터를 만들기 위해 생활 패턴과 가전기기 사용 확률, 동일 상황에 대해 청소 가능 여부가 다른 가정주부와 대학생을 시뮬레이션 대상으로 정하였다. 두 대상을 시뮬레이션 하여 잡자는 시간을 제외한 데이터를 수집하였고, 상황 발생의 불확실성을 반영하기 위하여 입력 파일에 설정된 확률에 따라 샘플링 하여 같은 상황에서 다양한 센서 값을 가질 수도 있도록 하였다.

2.2 Network Modeling of Causal Relation between User’s Situation and Operation of Home Appliances

가정 내 설치할 수 있는 센서들은 다양하고 따라서 수집 가능한 데이터 종류 또한 다양하다. 가장 단순한 센서인 온도계, 습도계부터 마이크로폰, 적외선 센서까지 수많은 센서를 설치할 수 있으며, 설치되는 센서에 따라 수집 가능한 데이터의 종류가 달라진다. 하지만 이 센서들 중 사용자의 상황 인지에 사용되는 센서의 종류는 한정적이다. 예를 들어, 온도계나 습도계와 같은 센서는 사용자의 상황이 센싱에 미치는 영향이 매우 작기 때문에 온도와 습도 데이터를 통해 사용자의 상황을 구별해내기는 거의 불가능하다. 가장 일반적으로 사용자 상황 인지에 사용되는 센서로는 카메라가 있으나, 설치에 많은 비용이 들며 설치 위치에 따라 사각지대가 생길 수 있고, 사생활 침해 가능성에 따라 보안에 대한 이슈가 생긴다.

따라서 본 논문에서는 설치하기 쉽고 사용자의 상황에 따라 센싱에 큰 영향을 받는 스마트 미터를 데이터 수집에 사용하였다고 가정한다. 스마트 미터란 전자식 전력 측정기로, 설치된 가정의 시간대 별 전력 사용량을 측정할 수 있다. 가전기기들은 각각 고유한 전력 사용량을 가지고 있기 때문에 스마트 미터에 기록되는 전력 사용량의 차이만을 통해 어떤 기기가 동작하고

있는지 센싱 할 수 있다고 가정하고 시뮬레이션을 통해 단위 시간 당 가전기기 동작 데이터와 사용자 상황 데이터를 수집하였다.

Table 1. User's Indoor Situation

Situation	Description	Vacuum?	
		Housewife	College Student
Toilet	Using the toilet	No	No
Bath	Taking a bath	No	No
Make up	Wearing clothes or makeup	No	No
Heating	Using boiler for heating	No	No
Laundry	Doing laundry	No	No
Dish wash	Washing dishes	No	No
Out	Going out	Yes	Yes
Work	Studying or reading a book	Yes	No
Cooking	Preparing for a meal	No	No
Dining	Having a meal	No	No
Rest	Taking a rest	No	Yes
Fold laundry	Folding laundry	No	No

가전기기 동작 데이터와 함께 수집되는 사용자 상황 데이터의 종류는 Table 1과 같다. 인지해야 할 상황은 사람이 가정 내에서 취할 수 있는 상황 중 가장 빈도가 높은 12가지의 상황으로 정의하였다. 이 12가지 상황에 대해 자동화 로봇 청소기가 청소가 가능한 상황과 불가능한 상황으로 다시 구분한다. 청소 가능 / 불가능 상황을 구분할 때 시뮬레이션 대상인 가정주부와 대학생이 선호하는 청소 상황이 다르다고 가정하였다. 공통적으로 지속시간이 짧은 상황인 toilet, bath, makeup, heating, laundry, dish wash, cooking, dining, fold laundry 상황들에 대해서는 가정주부와 대학생 모두 청소 불가능하다고 가정하였다. 두 대상의 청소 가능 여부는 대학생의 경우 work 상황에서는 청소를 할 수 없는 반면 rest 상황에서는 청소가 가능하고, 가정주부는 rest 상황에서 청소가 불가능, work 상황에서 가능한 것에서 차이를 보인다.

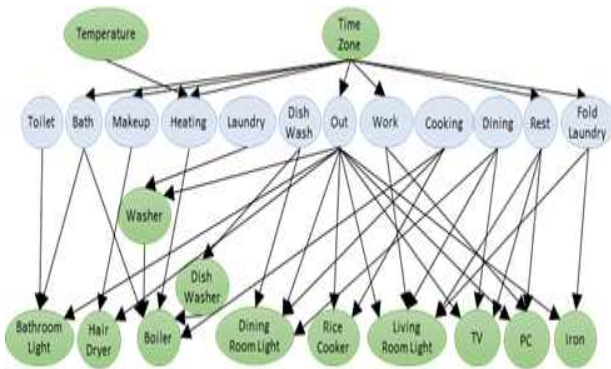


Fig. 6. Causal Network Between User Situation and Home Appliance Operation

사용자 상황과 가전기기 동작 간 인과관계를 반영하여 생성한 네트워크는 Fig. 6과 같다. Fig. 6의 네트워크에 표시된 중단의 하늘색 노드들은 Table 1에 정의한 사용자의 상황을 나타낸다. 하단과 상단의 녹색 노드들은 스마트 미터로 관측 가능한

가전기기들과 쉽게 측정 가능한 시간대나 온도를 나타낸다. 관측 가능한 가전기기는 총 11가지 종류로, 가정에서 주로 사용되는 기기들을 가정주부와 대학생 공통으로 사용한다고 가정하였다. Fig. 6 노드 사이의 간선은 노드 간 인과관계가 존재한다는 것을 나타낸다. 예를 들어, Time Zone 노드의 상태는 Bath 상황 발생에 영향을 끼치며 Bath 노드는 Bathroom Light 와 Boiler 의 동작 확률에 영향을 끼친다.

시뮬레이션을 통한 사용자 데이터 수집 과정에서 가정주부와 대학생 모두 공통적으로 Fig. 6의 네트워크 구조를 이용하나, 상황 발생이나 가전기기 사용 확률에서 차이를 보인다. 다음은 가정주부와 대학생을 시뮬레이션 할 때 이용한 Input File의 정보이다.

	Housewife	College Student
Morning	06:15 (±10min)	07:30 (±10min)
Afternoon	13:00 (±30min)	12:30 (±30min)
Evening	18:30 (±10min)	18:30 (±10min)
Sleeping	23:40 (±15min)	23:40 (±60min)

Fig. 7. Start Time Information of Time Zone

Fig. 7에서 볼 수 있듯이 가정주부와 대학생은 시간대의 구분에서 차이를 보인다. 각 시간대의 시작 시간은 주어진 시간과 편차 내에서 정해지며, 가정주부의 기상시간은 이르고 취침시각이 비교적 규칙적이다. 반면 대학생은 늦은 시간에 기상하고 취침시각이 불규칙하다.

		1	2	3	4	5	6
Housewife	Morning	Toilet	Cooking	Dining	Bath	Makeup	Fold laundry
		Dining	Toilet	Bath	Makeup		
		Toilet	Bath	Makeup	Work		
		Cooking	Dining	Toilet	Bath	Makeup	
	Afternoon	Toilet	Bath	Makeup	Fold laundry	Work	
		Toilet	Dining	Bath			
		Cooking	Dining	Toilet	Fold laundry		
		Cooking	Dining	Work	Toilet		
	Evening	Cooking	Dining	Toilet	Bath		
		Cooking	Dining	Toilet	Bath	Makeup	Work
		Cooking	Dining	Bath	Makeup	Toilet	
		Cooking	Dining	Toilet	Fold laundry	Bath	
College Student	Morning	Toilet	Cooking	Dining	Bath		
		Toilet	Dining	Bath	Makeup		
		Dining	Bath	Makeup			
		Toilet	Bath				
	Afternoon	Dining	Bath	Makeup	Toilet	Fold laundry	Work
		Toilet	Bath	Makeup	Fold laundry	Work	
		Toilet	Dining				
		Cooking	Dining	Toilet	Fold laundry	Work	
	Evening	Bath	Makeup	Cooking	Dining	Toilet	Work
		Toilet					
		Dining	Toilet				
		Toilet	Cooking	Dining	Makeup	Fold laundry	Work

Fig. 8. Behavior Patterns of Housewife and College Student

	Housewife		College Student	
	Probability	Duration	Probability	Duration
Toilet	0.9	15±2 min	0.9	10±2 min
Bath	0.99	25±3 min	0.99	15±2 min
Heating	0.95	300±60 min	0.95	300±60 min
Laundry	0.99	100 min	0.99	40 min
Dish wash	0.99	60 min	0.99	60 min
Cooking	0.99	45±5 min	0.7	45±5 min
Dining	0.95	25±3 min	0.7	25±3 min
Out	0.99	90±15 min	0.99	120±15 min
Rest	1	No restrictions	1	No restrictions
Makeup	0.99	20±5 min	0.99	15±5 min
Fold laundry	0.85	20±5 min	0.85	20±5 min
Work	0.8	120±10 min	0.7	150±30 min

Fig. 9. Situation Probability and Duration of Housewife and College Student

Fig. 8에서는 가정주부와 대학생의 시간대 별 행동 패턴을 보여준다. 각 사용자는 시간대 별로 5개의 행동 패턴을 가지고 있는데, 매일 임의로 행동 패턴 중 하나를 선택하여 시뮬레이션 한다. 각 사용자는 무조건 선택된 행동 패턴에 따르지 않고 각 상황 별로 상황 발생 확률에 따라 행동 패턴이 달라진다. Fig. 9는 사용자 별 상황 발생 확률과 상황이 발생 했을 때의 지속시간을 나타내고 있다. Rest 상황은 각 사용자가 시간대 별 행동을 마치고 남은 시간에 발생한다. 따라서 확률이 1이며 지속시간에 제약이 없다.

	Home Appliance	Housewife		College Student	
		On Prob.	Off Prob.	On Prob.	Off Prob.
Toilet	Bathroom light	0.98	0.95	0.9	0.95
Bath	Bathroom light	0.95	0.95	0.9	0.95
	Boiler	0.9	0.99	0.9	0.99
Heating	Boiler	0.99	0.99	0.99	0.99
Laundry	Boiler	0.99	0.99	0.1	0.99
	Boiler	0.95	0.99	0.95	0.99
Dish wash	Dining room light	0.1	0.99	0.2	0.99
	Boiler	0.8	0.99	0.1	0.99
Cooking	Dining room light	0.95	0.99	0.95	0.99
	Rice cooker	0.9	0.99	0.85	0.99
Dining	Dining room light	0.1	0.99	0.95	0.99
	Living room light	0.95	0.99	0.1	0.99
Rest	TV	0.95	0.95	0.1	0.95
	Living room light	0.95	0.99	0.1	0.99
	TV	0.95	0.95	0.3	0.9
Makeup	PC	0.05	0.9	0.8	0.9
	Hair dryer	0.9	0.99	0.9	0.99
Fold laundry	Living room light	0.95	0.99	0.9	0.99
	Iron	0.9	0.99	0.1	0.99
Work	PC	0.3	0.99	0.9	0.99
	Living room light	0.9	0.99	0.7	0.99

Fig. 10. Home Appliance On / Off Probability of Housewife and College Student

Fig. 10은 가정주부와 대학생의 상황 별 기기 On / Off 확률을 나타낸다. On 확률은 상황 별 기기가 동작할 확률이고 Off 확률은 상황이 끝난 후 사용자가 기기를 끄 확률이다. 사용자의 시간대 별 발생하는 상황이 정해지면 Fig. 10의 확률데이터에 따라 가전기기가 켜지고 꺼지는 것이 결정된다.

특징적인 차이점으로 대학생은 아침과 낮에 자주 Out 상황

이 발생하는 반면, 주부는 Out 상황의 발생 빈도가 낮다. 가전기기 사용 확률 면에서 두 사용자의 차이는 Rest 상황에서 가정주부는 TV 시청을 선호하는 반면, 대학생은 PC 이용을 선호한다. 또 Work 상황에서 대학생의 PC 사용 빈도가 가정주부에 비해 높다고 가정하였다.

2.3 Data Processing and Scenario Definition

가정주부와 대학생을 대상으로 시뮬레이션 하는 과정에서 Fig. 6에서 정의된 인과관계 네트워크에 따라 확률적으로 샘플링 된 가전기기 동작 데이터와 청소 가능 / 불가능 상황 데이터가 수집된다.

Temperature	Time Zone	Bathroom Light	Hair Dryer	Boiler	Dining room Light	Rice Cooker
Boolean	String	Boolean	Boolean	Boolean	Boolean	Boolean
Living room Light	TV	PC	Iron	Washer	Dish Washer	Vacuum
Boolean	Boolean	Boolean	Boolean	Boolean	Boolean	Boolean

Fig. 11. Structure of Collected Data

수집되는 데이터의 실제 구조는 Fig. 11과 같다. 총 13개의 Boolean 값과 하나의 string 값을 가지는 벡터로, Vacuum 은 청소 가능 / 불가능 클래스를 나타내며 나머지는 속성을 나타낸다. 속성들 중 Time Zone 은 "Morning", "Afternoon", "Evening" 세 가지 string 값을 가지며 시간대를 나타내고, 다른 속성들은 가전기기의 on / off 여부를 나타내는 Boolean 값을 가진다.

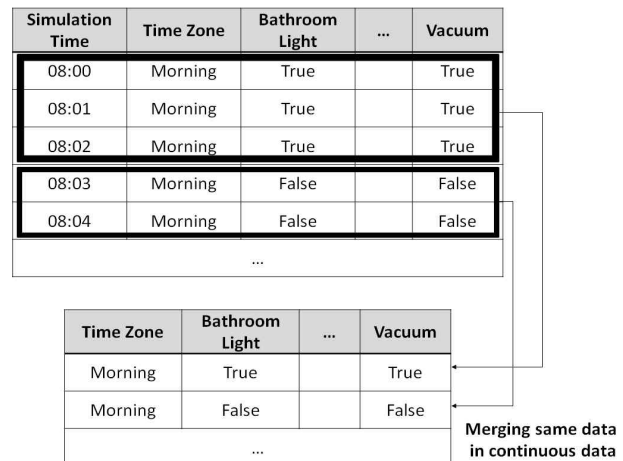


Fig. 12. Data Processing Method

Fig. 12와 같이 시뮬레이션은 1분 단위로 수행되나, 관측되는 센서 값에 변화가 생기는 시점의 데이터만 추출하는 데이터 가공 과정을 거친다. 예를 들어, 사용자가 다른 기기가 모두 꺼진 상태에서 Bathroom Light 만 10분 사용하였다면 관측 값에 변화가 없기 때문에 10분간의 데이터를 병합하여 하나의 데이

터로 본다. 이러한 데이터 가공은 지속시간이 긴 상황에 대해서 많은 데이터가 수집되어 알고리즘의 성능을 평가할 때 비정상적으로 많은 한 기기 조합의 데이터로 인해 성능 측정에 큰 영향을 받는 것을 방지한다.

시뮬레이션 결과로 수집되는 데이터는 총 90일 간의 데이터로 사용자의 실시간 피드백을 받아 청소 상황 판단 모델을 학습하는 사용자 데이터로 이용되며, 같은 확률의 인과관계 네트워크를 시뮬레이션 하여 얻은 90일 간의 테스트 데이터를 통해 학습된 모델의 성능을 측정한다.

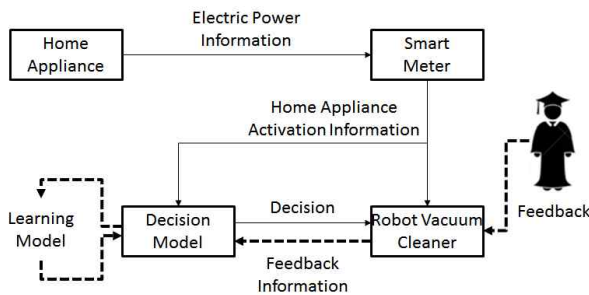


Fig. 13. Online Learning Scenario of Automation Robot Cleaner

자동화 로봇 청소기 온라인 학습 과정은 Fig. 13과 같다. 가정 내 로봇 청소기는 실시간으로 스마트 미터를 통해 가전기기 동작 상태에 대한 센서 데이터를 받으며, 센서 조합이 변할 경우 이 센서 조합을 입력으로 받아 청소 상황 판단 모델을 통해 청소 가능 / 불가능 상황을 분류한다. 초기 청소 상황 판단 모델은 임의로 주어지며, 모델의 분류 결과가 시뮬레이션을 통해 얻은 사용자 데이터의 청소 가능 / 불가능 상황과 일치하지 않을 경우 일정 확률로 피드백이 발생한다. 즉 청소 상황 판단 모델의 오판에 대해 사용자가 실시간으로 직접 관여하는 형태이며, 동작 중인 로봇 청소기를 끄는 행위 혹은 동작 중이지 않은 로봇 청소기를 켜는 행위를 피드백으로 정의한다. 피드백은 사용자가 직접 불편을 느끼고 청소 가능 / 불가능 상황을 바꾼 것이기 때문에 신뢰성이 높은 데이터이다. 따라서 모델 학습을 위한 학습 데이터로 사용되며, 소량의 피드백을 통해 좋은 성능을 내는 청소 상황 판단 모델을 만드는 것이 목적이다. 피드백 데이터의 구조는 Fig. 11의 구조와 같으며, 피드백이 발생하면 피드백 데이터를 통해 실시간으로 학습을 진행하고, 학습이 완료되면 학습된 모델을 적용하여 시나리오를 계속 진행하는 과정을 반복한다.

분류해야 하는 클래스는 결과적으로 청소 가능과 불가능 두 가지이기 때문에 피드백 발생 확률에 대한 설정이 없을 경우, 피드백을 받지 않는 상황에서는 모델의 판단이 정답이라 가정하면 매우 쉬운 문제로 바뀐다. 본 논문에서 가정한 피드백 확률은 50%로, 청소 상황 판단 모델이 오판을 할 경우 50%의 확률로 피드백이 발생한다.

피드백이 발생하지 않는 시점에 획득한 데이터는 청소 가능 여부를 알 수 없는 unlabeled 데이터이다. Unlabeled 데이터의

구조는 Fig. 11에서 Vacuum을 제외한 구조이며, 본 논문에서는 unlabeled 데이터를 중요한 정보로 보고 모델 학습에 활용한다.

III. The Proposed Scheme

본 논문에서는 unlabeled 데이터의 label을 유전 알고리즘을 통해 탐색하고, 탐색 결과로 얻은 label을 데이터에 적용하여 Naive Bayes 모델을 학습하는 방안을 제안한다.

기존 동일 환경에서 연구하였던 베이지안 네트워크 CPT 탐색 방안 [7]의 가장 근본적인 문제점이었던 복잡한 모델을 비교적 단순한 모델로 교체하기 위한 Naive Bayes, Decision Tree, Logistic Regression 모델의 성능 비교 실험이 기존 연구 [8]에서 수행되었고, 결과적으로 Naive Bayes 모델이 성능과 학습 시간 면에서 다른 모델에 비해 우수하다는 결론을 내렸다. 따라서 본 논문에서는 Naive Bayes 모델을 사용하여 모델의 복잡도를 줄이고 학습에 소요되는 시간을 단축시키고자 한다.

또한 [6-7]에서는 피드백 데이터만을 이용하여 학습을 진행하였기 때문에, 피드백 데이터뿐만 아니라 수집되는 가전기기 동작 데이터를 활용하면 모델의 성능향상을 기대해 볼 수 있다고 생각하였다. 본 논문에서는 유전 알고리즘으로 가전기기 동작 데이터의 label을 탐색하는 방안을 통해 unlabeled 데이터를 활용하는 방법을 제안한다.

1. Naive Bayes

Naive Bayes는 Bayes theorem을 기반으로 분류에 사용되는 지도 학습 분류기 중 하나이다 [11]. 1950년대 이후 광범위하게 연구되고 있으며, 주로 문서에 나타나는 단어를 통해 문서의 종류를 분류하는 문제에 사용되는 방법으로 잘 알려져 있다. Naive Bayes의 기본적인 원리는 조건부 확률에 Bayes' theorem을 적용하고, 각 속성이 등장할 확률에 대해 독립성을 가정하여 입력 데이터를 확률적으로 분류하는 것이다.



Fig. 14. The Proposed Naive Bayes Model

본 논문에서 제안하는 Naive Bayes 모델의 구조는 Fig. 14와 같다. 하단의 노드들은 스마트 미터를 통해 감지할 수 있는 가전기기들을 나타낸 것이고, 이 가전기기의 동작 상태를 입력으로 받아 상단부의 청소 상황을 추론한다. 기존 연구였던 [7]에서는 3.2의 Fig. 6의 인과관계 네트워크를 그대로 베이지안 네트워크로 모델링하여 청소 상황을 판단하기 위해서는 사용자의 상황을 먼저 추론하는 과정이 필요하였다. 때문에 모델이 매우 복잡했고 센서 정보에는

사용자의 상황 데이터가 없이 청소 가능 / 불가능 데이터만 존재하기 때문에 학습을 하여도 사용자의 상황에 대한 CPT 는 제대로 학습될 수 없는 구조였다. 본 논문에서 제안하는 Naive Bayes 모델은 사용자의 상황 추론 과정을 없애고 청소 상황을 직접 추론하기 때문에 모델의 복잡도가 매우 낮다.

$$P(vacuum|Sensors) = \frac{P(Sensors|vacuum) \cdot P(vacuum)}{P(Sensors)} \dots (3)$$

제안하는 Naive Bayes 모델에서의 청소 가능 여부는 Vacuum 노드가 True일 확률과 False일 확률을 계산하여 높은 확률을 가지는 클래스로 추론한다. 식 (3)의 좌변은 조건부 확률로 Sensor 데이터가 주어졌을 경우 Vacuum 노드가 True일 확률을 구하는 수식이며, 우변은 좌변의 식을 Bayes' theorem 을 이용하여 변형한 식이다.

Naive Bayes 모델은 결과적으로 얻어진 청소가능 확률과 청소 불가능 확률 중 더 큰 쪽으로 청소 상황에 대한 판단을 내리게 된다. 청소 상황 판단 과정에서 $P(vacuum|Sensors)$ 를 계산할 때, Bayes' theorem에 의해 식을 변형하고 각 속성은 독립적이라는 가정을 추가하면 비교적 간단하게 확률 계산을 할 수 있다.

2. Method of Searching Labels Based on Naive Bayes Model

본 논문에서는 제안한 Naive Bayes 모델에 기반을 두어 소량의 labeled 데이터로 unlabeled 데이터의 label을 탐색하는 방안을 제안한다. 데이터의 label은 청소 가능 / 불가능 2가지로 나뉘며, 사용자의 피드백을 받을 경우 유전 알고리즘을 통해 label 탐색을 실시하고, 탐색된 레이블을 적용하여 Naive Bayes 모델을 학습하는 방안으로, 해의 표현은 Fig. 15와 같다. 해의 길이는 수집되는 unlabeled 데이터의 종류에 따라 점차적으로 증가하며, 청소 가능 / 불가능의 이분화 된 label을 탐색하기 때문에 해는 1 또는 0으로 인코딩 된 binary gene을 사용한다. Unlabeled 데이터를 처리할 때 센서 조합이 같은 데이터를 함께 처리하기 때문에 unlabeled 데이터의 절대적인 수가 늘어난다 하더라도 그 조합은 한정적이다.

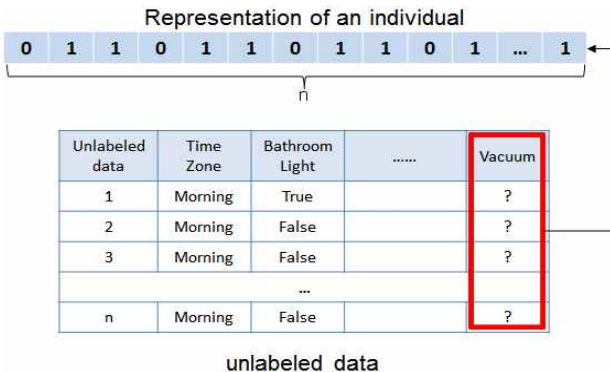


Fig. 15. Representation of Individuals

[7]에서는 베이지안 네트워크의 CPT를 직접 탐색하는 방안을 제안하였으나, 이 방안은 탐색해야 하는 해의 길이가 길고 실수해에 대해 탐색해야 하기 때문에 무한한 탐색 공간을 가지게 된다. 따라서 탐색 결과로 도출된 해는 실제 최적 해가 아닐 수도 있다. 또한 CPT를 직접 탐색하기 때문에 수집되는 데이터와 무관한 결과를 도출할 위험성이 존재한다. 제안 방안에서는 binary gene을 사용함에 따라 탐색 공간을 대폭 줄였으며, 데이터의 label을 탐색하기 때문에 수집된 데이터의 범위 내에서 만들어질 수 있는 모델에 대해서만 탐색하게 된다.

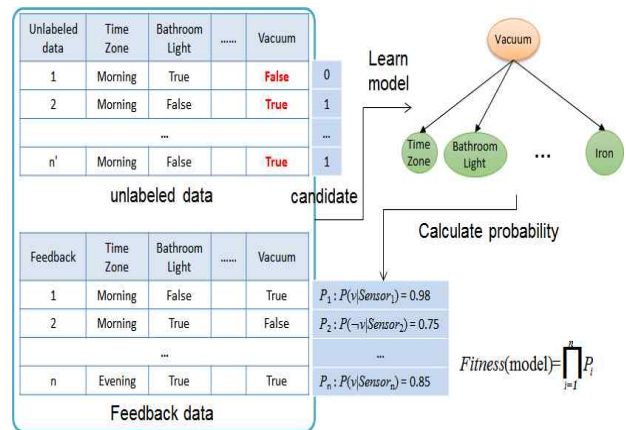


Fig. 16. Evaluation method of Individuals

해의 평가는 피드백 시점까지 수집된 모든 피드백 데이터를 기반으로 실시한다. Fig. 16에서 볼 수 있듯이 후보 해를 통해 unlabeled 데이터에 label을 부여하여 전체 데이터를 통해 Naive Bayes를 학습 시키고 나면, 학습된 모델을 통해 피드백 데이터에 적용시켜 likelihood 를 계산한다. 청소 가능 상황에 대한 피드백에는 $P(vacuum|Sensors)$ 를 계산하고, 청소 불가능한 상황에 대해서는 $P(-vacuum|Sensors)$ 를 계산하여 계산된 모든 확률을 곱한 후 이를 통해 해의 적합 도를 평가한다. 피드백 시점까지 모든 피드백 데이터의 확률 곱을 통해 평가하기 때문에 [6]에서 최근 피드백만 만족시키도록 학습하던 경향과는 달리 이전에 받은 모든 피드백을 최대한 만족하려는 방향으로 해가 진화하게 된다.

IV. Experiment

1. Experiment Setting

본 논문의 실험을 수행하기에 앞서 데이터를 얻기 위해 시물레이션을 진행하였다. 대상은 가정주부와 대학생으로, 생활 패턴이나 가전기기 사용 확률이 다른 사용자에게 대해 평균적인 성능 검증을 위해 두 대상을 가정하고 시물레이터의 입력 파일을 작성하였다. 작성된 입력 파일로 각 대상 별 90일간 시물레이

션을 11번 진행하였고, 10개의 데이터 셋은 피드백을 통한 모델 학습을 위해, 남은 1개의 데이터 셋은 학습된 모델의 성능을 측정하기 위해 사용하였다.

수집한 데이터 셋을 통해 본 논문의 제안 방안을 평가하기 위해서 비교 방안 3가지를 설정하고 비교 실험을 수행하였다. Laplace correction 을 거친 피드백 데이터의 relative frequency 계산을 통해 Naive Bayes 모델을 학습하는 방법 (NB) [11] 과 기존 연구 [7] 의 CPT 탐색 방안 (BN+GA), 그리고 label이 없는 가전 기기 동작 데이터를 활용하는 간단한 방법인 semi-supervised learning 의 self-learning 방안과 (NB+Semi) [9] 제안 방안 (NB+GA) 을 비교한다. NB+Semi 방안은 labeled 데이터로 Naive Bayes 모델을 학습한 뒤 학습한 모델을 이용하여 unlabeled 데이터에 label 을 부여하고, 모든 데이터로 다시 모델을 만들어 unlabeled 데이터에 다시 label 을 부여하는 과정을 모델이 수렴할 때까지 반복하는 방법이다. 기존 연구 [6] 은 [7] 의 연구에서 초기 모델에 따라 학습 성능이 크게 좌우되며, BN+GA 방안에 비해 좋지 않은 성능을 가지는 것을 실험적으로 확인하였기에 비교 방안에서 제외하였다.

모든 방안은 스마트 홈 환경의 시뮬레이션을 통해 모델을 학습한다. 시뮬레이션 중에 사용자의 피드백이 발생할 때마다 각 방안으로 모델을 학습 및 갱신한다. 이렇게 학습한 모델의 성능을 평가하기 위해 학습 모델을 테스트 데이터에 적용하여 청소 가능 여부를 추론한다. 알고리즘의 성능을 측정하기 위해 상황 판단 결과를 이용하여 성능 평가 지표를 계산한다. Table 2는 성능 평가 지표와 그에 대한 설명을 나타낸 표이다.

Table 2. Performance Measures

Measure	Description & Formula
Accuracy	The proportion of correctly classified examples $(TP+TN)/(TP+TN+FP+FN)$
Precision	The proportion of examples labeled True among examples classified as True $TP/(TP+FP)$
Recall	The proportion of examples classified True among examples labeled as True $TP/(TP+FN)$
False Positive Rate	The proportion of examples labeled False among examples classified as True $FP/(TP+FP)$
False Negative Rate	The proportion of examples labeled True among examples classified as False $FN/(TN+FN)$
F-measure (F1 score)	The harmonic mean of Precision and Recall $2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$

최종적인 성능 비교는 10회 실험에서 피드백을 15개 획득하는 시점까지의 성능 평가 지표들의 평균을 통해 수행하였다. 피드백이 증가함에 따라 각 방안들의 학습 양상을 비교하였고 그에 따른 실험 결과를 분석하였다. BN+GA 와 NB+GA 의 경우 모델을 학습시키기 위한 과정에서 유전 알고리즘을 이용하였

다. 두 방안의 공정한 비교를 위해 유전 알고리즘의 파라미터는 [7] 에서 사용하였던 설정을 동일하게 이용하였다.

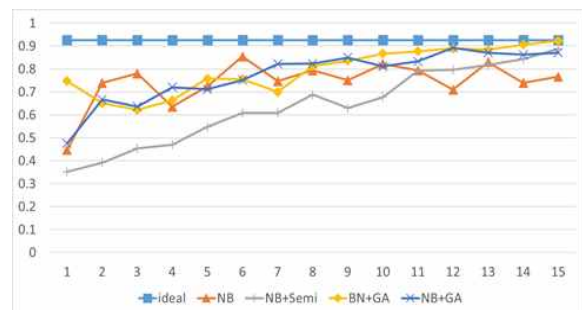
	BN+GA	NB+GA
Population size	100	
Decision variable	164 (real value)	15 (binary value)
Objective function	Multiplication of likelihoods about feedbacks	
Number of generation	100	
Selection method	Tournament selection (tournament size = 2)	
Crossover method	SBX (simulated binary crossover)	Discrete crossover
Mutation method	Polynomial mutation	Flip mutation

Fig. 17. GA Parameters

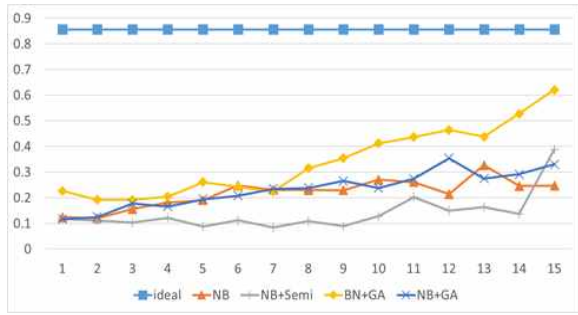
BN+GA 방안의 crossover method 와 mutation method 는 [7] 의 설정을 따라 SBX [12], polynomial mutation [13] 을 사용하였으나, NB+GA 는 해가 binary encoding 되어있기 때문에 crossover method 로는 discrete crossover [12], mutation method 로는 flip mutation [13] 을 이용하였다. 마지막으로 selection method 로는 BN+GA 와 NB+GA 모두 tournament selection [14] 을 이용하였다. (Fig. 17) 또한 추가적으로 NB+GA 와 BN+GA 의 CPU process time 을 측정하여 학습에 걸리는 시간을 비교하였다.

2. Experiment Result

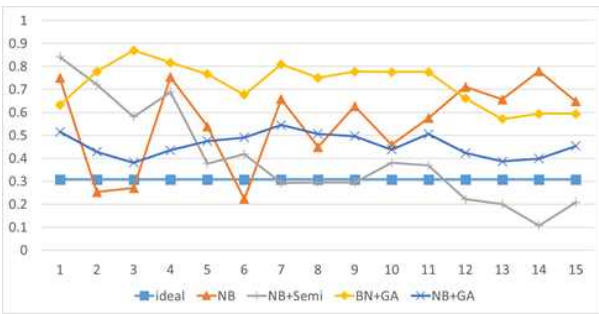
먼저 실험 데이터 중 가정주부 데이터에 대해 각 학습 방안을 적용하여 실험을 진행하였다.



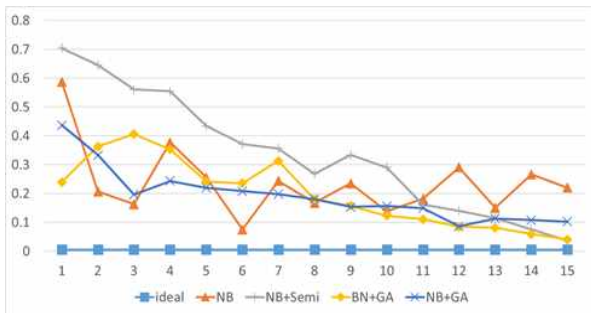
(가) Accuracy



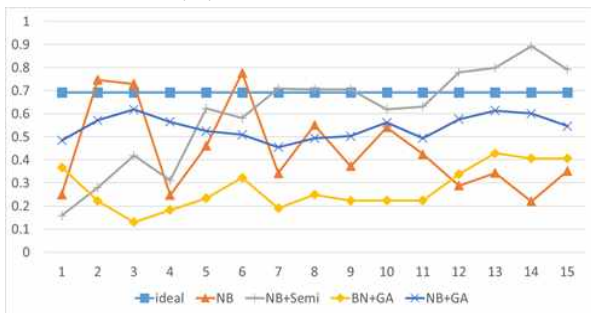
(나) Precision



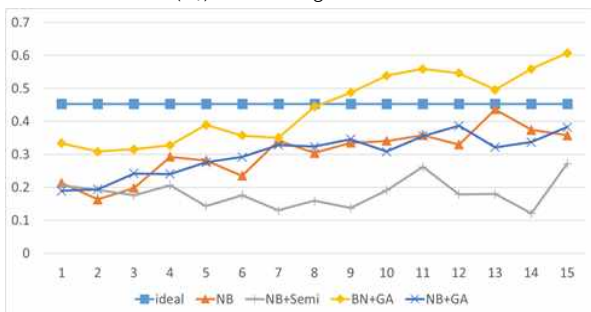
(다) Recall



(라) False Positive Rate



(마) False Negative Rate



(바) F-measure

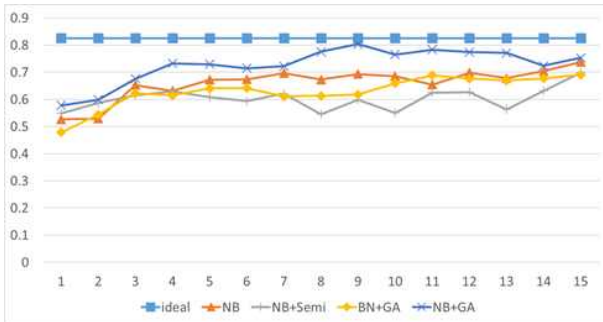
Fig. 18. Experimental Results Applied to Housewife Data

Fig. 18은 15개의 피드백을 받는 동안 6가지 성능 평가 지표를 그래프로 나타낸 결과이다. 각 그래프의 가로축은 피드백의 개수를 나타내고 세로축은 성능을 나타낸다. 가장 먼저 관찰할 수 있는 결과로 NB의 경우 (다) recall, (라) false positive rate, (마) false negative rate 그래프에서 초기 모델의 성능 지표가 진동하는 모습을 보인다. NB는 피드백 데이터의 relative frequency로 학습하기 때문에 학습 데이터가 적은 초반부의 학습에 피드백의 영향이 매우 크다는 것을 실험적으로 확인하였다.

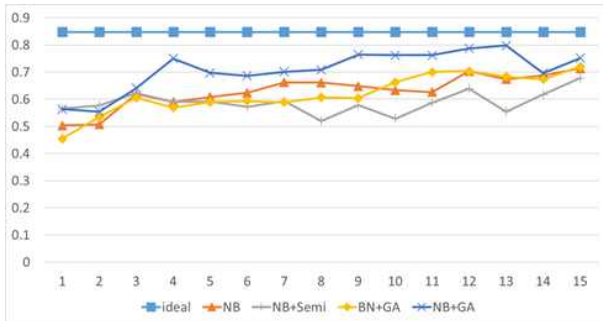
가정주부 데이터를 가공하여 통계를 내면 청소 불가능한 상황이 가능한 상황보다 약 9배정도 많다. 모든 데이터의 label을 알고 학습한 결과인 ideal 그래프를 보면 (가) accuracy는 높지만 (나) precision에 비해 (다) recall이 낮은 점으로 보아 대부분의 청소 가능한 상황에 청소하지 않는다는 점을 확인할 수 있다. 이는 모든 데이터로 학습할 경우 각 클래스에 속하는 데이터의 개수가 모델의 prior probability에 영향을 미쳐, 가정주부 데이터와 같이 한 클래스의 데이터가 월등히 많을 경우 해당 클래스로 대부분 판단을 내리게 된다. 하지만 사용자에게 받는 피드백 데이터는 청소 가능 / 불가능 상황의 비율이 약 5:5이다. 따라서 (다) recall과 (나) precision, (마) false negative 성능 지표에서 ideal 그래프와 다른 방안의 그래프가 상당한 차이를 보임을 확인할 수 있다. 가정주부 데이터와 같이 한 클래스로 편향된 데이터 비율에서는 accuracy가 성능 지표로서 큰 역할을 하지 못한다. 청소를 전혀 하지 않는 모델을 만들었다고 가정하여 가정주부 데이터에 적용하면 accuracy는 약 0.9로 측정될 것이다. 따라서 이러한 데이터에서는 (가) accuracy보다 (바) f-measure가 의미 있는 성능 지표이다. F-measure는 precision과 recall의 조화평균으로, 청소 상황에서 올바른 청소 판단을 내리고 또 청소 판단을 내릴 경우 그 판단이 틀리지 않아야 좋은 f-measure 값을 가지게 된다. (바) f-measure를 보면 다른 알고리즘들의 f-measure 수치는 실험환경 하에서 0.45를 넘지 못하였지만, BN+GA는 9회의 피드백 획득만으로 ideal 그래프의 성능을 넘어 가정주부 데이터에 대해 가장 우수한 성능을 보인다. NB+Semi는 (가) Accuracy 그래프만 본다면 성능이 우수한 것으로 보이지만, (다) recall과 (마) false negative rate 성능 지표를 보면 과도하게 청소를 하지 않는 모델로 학습하려는 경향을 보인다. (나) Precision과 (다) recall 성능 지표를 보면 NB+Semi 방안이 초반부에 과도하게 청소하려는 경향을 보이며, 때문에 이후 사용자가 대부분 청소 불가에 대한 피드백을 주어 청소를 점차하지 않는 모델로 바뀌는 것을 확인하였다.

가정주부 데이터에 대해 NB+GA는 BN+GA보다는 성능이 좋지 않지만 안정적인 성능 그래프를 보여준다. 전체적인 성능 지표를 보면 다른 방안에 비해 피드백 개수에 따른 성능 변화 폭이 작음을 알 수 있고, 적은 피드백에 의해 모델이 과도한 영향을 받지 않음을 확인하였다. Unlabeled 데이터를 이용한다는 관점에서는 NB+Semi와 비슷한 학습 방안이지만,

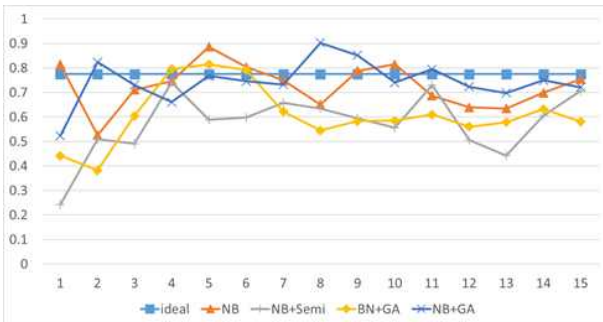
NB+Semi 는 초기 모델에 따라 일정한 방향으로 탐색하고 수렴하기 때문에, 수집되는 피드백 데이터가 원본 데이터를 제대로 반영하지 못하면 학습이 잘 되지 않는 반면, 제안 방안은 유전 알고리즘을 통해 다양한 모델에 대해 탐색하기 때문에 더 좋은 모델을 찾을 가능성이 높아진다.



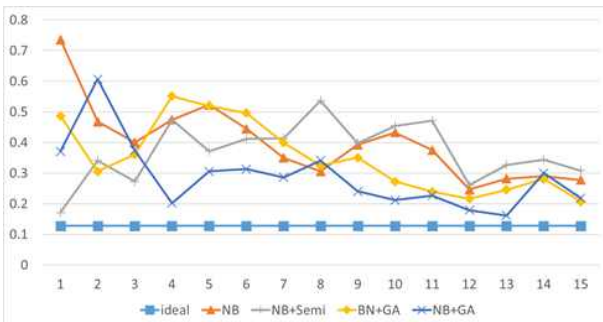
(가) Accuracy



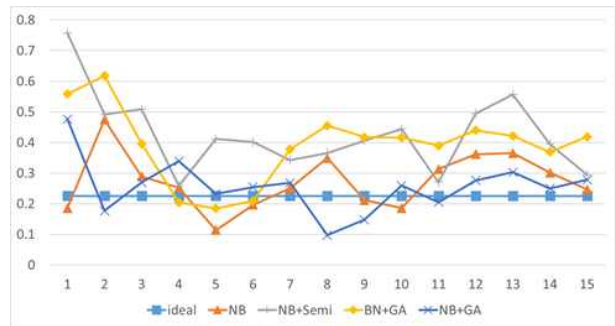
(나) Precision



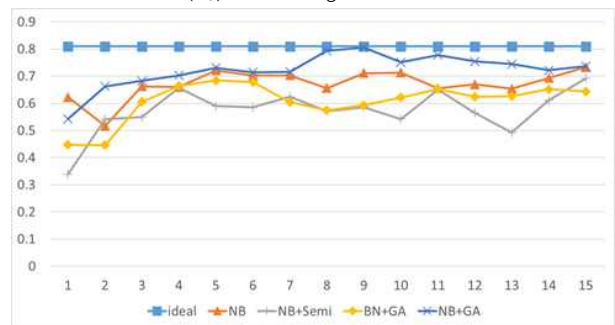
(다) Recall



(라) False Positive Rate



(마) False Negative Rate



(바) F-measure

Fig. 19. Experimental Results Applied to College Student Data

Fig. 19는 대학생 데이터에 각 학습 방안을 적용하여 성능 지표를 측정된 결과 그래프이다. 대학생 데이터를 가공 후 통계를 내면 청소 가능 / 불가능 클래스의 비율이 거의 비슷하다. 때문에 피드백 데이터가 원본 데이터의 비율과 비슷하여 가정주부 데이터와는 다르게 모든 학습 방안의 성능 그래프들이 ideal 그래프의 방향으로 점점 수렴해가는 것을 볼 수 있다. Ideal 그래프를 관찰해보면 (가) accuracy 가 약 0.85 정도의 수치를 가지는 것을 볼 수 있는데 달리 얘기하면 약 15%의 데이터가 잘 학습된 모델로도 제대로 판단을 내릴 수 없는 데이터라는 것이다. 이 15% 데이터 중 사람 행동의 불확실성에서 나오는 노이즈 데이터와 가전기기 사용 조합은 같지만 실제 사용자의 상황이 달라 구분할 수 없는 데이터도 존재한다. 결과적으로 대학생 데이터는 가정주부 데이터에 비해 학습하기 어려운 데이터이며, 수집되는 피드백의 종류에 따라 확연한 성능 차이를 보이게 된다. 실제 모든 방안에서 가전기기 사용 조합은 같으나 청소 상황이 다른 피드백을 받은 것을 확인하였으며, 이러한 피드백이 학습 양상에 큰 영향을 주었다.

가정주부 데이터에 대해 가장 좋은 성능을 보였던 BN+GA 알고리즘은 대학생 데이터에서는 좋은 학습 성능을 보여주지 못하였다. 15회 피드백을 획득하는 동안 (가) accuracy 가 최대 0.71로 저조하며, (다) recall 과 (마) false negative rate 그래프에서 BN+GA 의 성능이 ideal 그래프에 수렴하지 못하고 실제 청소를 해야 하는 상황에서 청소를 잘 하지 않는 모델이 학습됨을 알 수 있다. 분석한 결과 총 10회 실험의 각 세트에서 15개의 피드백 중 가전기기 사용 조합이 같고 청소 상황이 다른 피드

백이 여러 개 존재하는 것을 확인하였다. BN+ GA 에서 이런 피드백들은 해를 평가할 때 다른 방안에 비해 큰 영향을 미치게 된다. 해의 평가 방안으로 탐색한 모델로 각 피드백의 청소 확률 값을 계산하는데, 예를 들어 가전기기 사용 조합이 같고 청소 상황이 다른 피드백이 각각 1개씩 있다고 가정하면 각 피드백에 대한 확률이 0.5가 될 때 평가 결과가 최대가 된다. BN+ GA 는 이러한 해를 찾기 위해 직접 CPT 를 조절하므로, 실제 데이터와는 무관한 확률 값을 학습할 가능성이 있다.

반면 NB+ GA 역시 해의 평가는 BN+ GA 와 동일하게 피드백 데이터의 청소 확률을 곱하는 방법이지만, NB+ GA 은 unlabeled 데이터의 label 을 찾아 모델을 만들기 때문에 실제 데이터에 기반을 두어 최적 해를 찾는다. 따라서 중복되는 데이터가 들어오더라도 전체 학습 데이터에 비해 소량이므로, BN+ GA 처럼 큰 영향을 받지 않는다. NB+ GA 에서 피드백이 9개일 때 (가) accuracy 와 (바) f-measure 가 가장 높은 성능을 보였다가 이후에 성능이 점차 내려가는 현상을 보이는데, 이는 시나리오의 흐름 상 좋은 모델을 학습하게 되면 사용자 피드백으로 노이즈 데이터가 들어올 가능성이 커지며, 이런 노이즈 데이터에 대해 학습하기 때문에 성능이 내려가는 것을 확인하였다.

NB 와 NB+ Semi 를 이용하여 학생 데이터를 학습 시킨 결과, 두 알고리즘으로 학습한 모델은 성능 수치가 수렴하지 않고 발산하는 것을 확인하였다. (다) recall, (라) false positive rate, (마) false negative rate 성능 지표를 보면 두 방안의 성능이 계속해서 진동하는 양상을 보이는데, 두 학습 방안 역시 가전기기 사용 조합이 같고 청소 상황이 다른 피드백이 여러 개 들어오며 다양한 학습데이터를 수집하지 못하였다. 결과적으로 (가) accuracy 와 (마) f-measure 그래프를 보았을 때 제안 방안에 비해 성능 수치가 낮음을 확인하였다.

Table 3. Results of Measuring CPU process time

Num of Feedback	Searching CPT Method	Proposed Method
1	689	10
2	1347	16
3	1997	22
...		
15	10352	98

마지막으로 [7] 의 BN+ GA 와 NB+ GA 의 학습에 걸리는 시간을 비교하기 위해 각 방안의 실험을 동일한 환경에서 실시하고 피드백의 개수에 따라 유전 알고리즘으로 해를 한번 탐색하기 위해 걸리는 CPU process time (단위: 초)를 비교하였다.

Table 3에서 볼 수 있듯이 BN+ GA 는 NB+ GA 에 비해 탐색 소요 시간이 매우 길다. 주로 시간이 많이 걸리는 부분은 해를 평가할 때 BN+ GA 에서 총 164개의 CPT entry 를 이용해 베이지안 네트워크를 구성하여 피드백에 대해 확률을 추론하는 부분이다. 베이지안 네트워크 모델은 매우 복잡하기 때문에 CPT entry 로 모델을 갱신하기 위한 시간이 많이 걸리는 반면, NB+ GA 는 데이터를 탐색하고 학습을 함에도 불구하고 간단

한 모델인 Naive Bayes 를 이용하기 때문에 모델을 갱신하는 시간이 비교적 짧다. 근래 들어 하드웨어의 성능 증가로 인해 학습에 걸리는 시간이 중요하지 않으나, 본 논문에서는 로봇 청소기 단말에서의 학습이라는 가정을 두었기 때문에 제한된 자원 내에서 빠르게 학습할 수 있는 알고리즘이 필요하다. 이런 면에서 NB+ GA 와 BN+ GA 의 성능을 비교하였을 때, 실험환경에서 NB+ GA 가 최소 68배 이상의 속도 차이로 우수함을 확인하였다.

IV. Conclusions

본 논문은 스마트 홈 환경에서 자동화 로봇 청소기의 청소 상황 판단 모델을 학습하는 방안을 제안하였다. 학습 데이터를 얻기 위해 사용자의 피드백을 받아 수집하였고, 학습 데이터를 기반으로 Naive Bayes 모델을 학습하였다. 제안 방안에서는 피드백의 영향을 크게 받았던 기존 연구들의 문제점을 보완하기 위하여 유전 알고리즘을 이용하여 unlabeled 데이터에 label 을 부여하는 방법을 이용하여 데이터의 활용도를 높였다. 학습 알고리즘의 성능은 가정주부와 대학생에 대상으로 가상의 시뮬레이션 환경에서 수집한 데이터로 기존 CPT 탐색, semi-supervised learning, 피드백 데이터만을 이용하여 relative frequency 를 계산하는 학습 방법과 비교 실험을 통해 평가하였고, 제안하는 알고리즘이 총 15회 피드백을 획득하는 동안 대학생 데이터에 대해 accuracy 와 f-measure 측면에서 좋은 성능을 보임을 확인하였다.

가정주부 데이터에 대해서는 CPT 탐색 방안에 비해 f-measure 수치가 낮았으나, CPT 탐색 방안은 제안 방안에 비해 최소 68배 이상 학습 시간이 소요되어 실제로 사용하기에는 큰 어려움이 있다. CPT 탐색을 제외한 다른 방안들은 학습 시간이 오래 걸리지 않았기 때문에 피드백 15개 이후로도 학습 경향을 관찰할 수 있었고, 충분한 피드백 데이터가 쌓이면 이상적인 성능에 수렴한다는 것을 확인하였다. 실험 결과에서 확인할 수 있듯이 제안하는 온라인 학습 방안의 문제점은 성능이 좋은 모델을 학습 시키면 사용자의 피드백으로 노이즈 데이터를 받을 확률이 매우 커진다. 이런 노이즈 데이터에 대해 모델을 학습하면 성능이 떨어지는 현상이 발생하게 된다. 따라서 향후 연구로는 적당한 성능을 내는 모델을 학습한 시점을 계산하여 학습을 그만두는 방안이 필요하다. 또한 현재는 로봇 청소기의 청소 가능 / 불가능의 이분화 된 상황만 판단하지만 향후에는 사용자의 다양한 상황에 대해 인지하는 방안 연구가 필요하다.

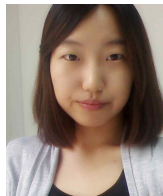
REFERENCES

- [1] Brdiczka, Oliver, James L. Crowley, and Patrick Reignier. "Learning situation models in a smart home." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.1 pp. 56-63, Feb. 2009.
- [2] Jalal, Ahmad, et al. "Robust human activity recognition from depth video using spatiotemporal multi-fused features." *Pattern recognition* 61 pp. 295-308, 2017.
- [3] Wongpatikaseree, Konlakorn, et al. "Activity recognition using context-aware infrastructure ontology in smart home domain." *Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on. IEEE*, pp. 50-57, 2012.
- [4] Guo, Lei, and Hongqing Fang. "Real-time human activity recognition in smart home with binary tree SVM." *Control Conference (CCC), 2016 35th Chinese. IEEE*, pp. 6977-6981, 2016.
- [5] Chahuaara, Pedro, François Portet, and Michel Vacher. "Context aware decision system in a smart home: knowledge representation and decision making using uncertain contextual information." *The 4th International Workshop on Acquisition, Representation and Reasoning with Contextualized Knowledge (ARCOE-12)*. pp. 52-64, 2012.
- [6] Ha, Eu Tteum, et al. "Learning a Bayesian Network for Situation-Aware Smart Home Service: A Case Study with a Robot Vacuum Cleaner." *International Science Index Computer and Information Engineering* 2.2 pp. 877-884, Feb. 2015.
- [7] Kwon, Seongcheol, Jeongmin Kim, and Kwang Ryel Ryu. "Interactive Learning of Bayesian Network for Situation-Aware Smart Home Service." *Korean Institute of Information Scientists and Engineers* pp. 486-488, Jul. 2015.
- [8] Kwon, Seongcheol, Jeongmin Kim, and Kwang Ryel Ryu. "Performance Comparison of Situation-Aware Models for Activating Robot Vacuum Cleaner in a Smart Home." *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 10.2 pp. 312-316, 2016.
- [9] Zhu, Xiaojin. "Semi-supervised learning literature survey." 2005.
- [10] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." *Machine learning* 29.2-3 pp. 131-163, 1997.
- [11] Murphy, Kevin P. "Naive bayes classifiers." *University of British Columbia*, 2006.
- [12] Umbarkar, A. J., and P. D. Sheth. "Crossover Operators in Genetic Algorithms: a review." *ICTACT Journal on Soft Computing* 6.1 pp. 1083-1092, 2015.
- [13] Soni, Nitasha, and T. Kumar. "Study of various mutation operators in genetic algorithms." *IJCSIT International Journal of Computer Science and Information Technologies* 5.3 pp. 4519-4521, Jul. 2014.
- [14] Blicke, Tobias, and Lothar Thiele. "A comparison of selection schemes used in genetic algorithms.", 1995.

Authors



Seongcheol Kwon received the B.S. degrees in Computer Science and Engineering from Pusan National University, Korea, in 2015, respectively. He is currently a M.S. candidate in the Department of Computer Science and Engineering at Pusan National University, Busan, Korea. He is interested in artificial intelligence, machine learning and data mining.



Seyoung Kim received the B.S. degrees in Mathematics and M.S. degrees in Computer Science and Engineering from Pusan National University, Korea, in 2014 and 2016, respectively. She is currently a Ph.D candidate in the Department of Computer Science and Engineering at Pusan National University, Busan, Korea. She is interested in artificial intelligence, machine learning and data mining.



Kwang Ryel Ryu received the B.S. and M.S. degrees in Electronic Engineering from Seoul National University, Korea, in 1979 and 1981, respectively, and received the Ph.D. degrees in Computer Science and Engineering from the University of Michigan, U.S, in 1992. Dr. Ryu joined the faculty of the Department of Computer Science and Engineering at Pusan National University, Busan, Korea, in 1993. He is currently a Professor in the Department of Computer Science and Engineering, Pusan National University. He is interested in artificial intelligence, machine learning and data mining.