

EM 알고리즘을 이용한 사물 인터넷 서비스 클러스터링 기법

장준범, 조정훈, 이대원¹
서경대학교 컴퓨터공학과
e-mail : {jjb0726, jojeong3766, daelee}@skuniv.ac.kr

EM Algorithm based Clustering Method for Internet of Things (IoT) Service*

June-Beom Jang, Jeong-Hoon Jo, Daewon Lee
Dept. of Computer Engineering, SeoKyeong University

요 약

다양한 IoT(사물인터넷) 서비스가 등장하고 수요가 많아짐에 따라 이를 통합적으로 관리하고 제어하는 통합 서비스 플랫폼에 관한 연구가 활발하게 진행되고 있다. 하지만 서비스의 표준 부재로 인하여 IoT 서비스 모듈의 재활용 및 이식은 불가능한 상황이다. 이러한 문제를 해결하기 위하여 본 연구에서는 IoT 서비스의 각 동작 단계에 EM 알고리즘을 적용하여 [1]의 동작기반 분류 기법을 확장한다. 제안한 EM 기반 IoT 서비스 분류 알고리즘은 서비스 유사도를 기반으로 분류 함으로 모듈의 재활용성을 높이고 서비스 간의 협업에 있어서 효율성 증대를 기대할 수 있다. 성능 평가를 통하여 평균에 대한 표준편차로 클러스터링되는 것을 확인 할 수 있다.

¹ 교신저자

* 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2016R1C1B1008330)

1. 서론

제 4 차 산업혁명과 관련하여 많은 IoT 서비스들이 제공되고 있다. 다양한 IoT 서비스의 통합 플랫폼 제공을 위하여 다양한 IoT 플랫폼 표준화 연구가 진행 중이다. 현재 삼성전자, 구글, 아마존 등에서 제시하고 있는 다양한 IoT 서비스들은 독립적으로 이루어지고 있으며 그 서비스 간 연동은 불가능한 상황이다. 이러한 다양한 IoT 서비스간의 연동, 기기의 재활용, 데이터 공유를 효율증대등을 위하여 통합 IoT 플랫폼의 제공이 필요하다. 하지만 통합 IoT 플랫폼의 표준 미비로 인하여 비호환성, 비효율성의 문제가 나타난다.

[1]에서는 효율적이고 안정적인 통합 서비스 플랫폼의 제공을 위한 분류 알고리즘을 시스템 환경, 알고리즘 동작 단계를 제시한다. 이에 본 연구는 [1]의 후속 연구로 [1]의 동작 기반 분류 기법에 EM 알고리즘을 적용하여 평균과 표준편차로 클러스터를 설정할 것이다. 또한 제시한 분류 알고리즘을 통하여 서비스 간의 유사도를 측정하고 유사도를 기반한 모듈의 재활용성 및 효율성의 증대를 도모한다.

2. 관련 연구

클러스터링은 기계학습, 데이터 마이닝, 패턴 인식, 영상 분석, 생물정보학을 비롯한 많은 분야에서 사용하는 기술이다. 클러스터링은 하나의 데이터를 다수의 부분집합으로 분할하는 것을 의미하며, 각 부분집합에 있는 데이터는 몇 가지의 공통된 특징을 공유한다. 클러스터링 분석은 동일한 그룹의 개체가 다른 그룹의 개체보다 더 유사하도록 개체 집합을 그룹화하는 작업이다.

데이터 클러스터링 분석은 분할 기법, 계층적 기법, 밀도 기반 기법, 격자 기반 기법, 모델 기반 기법 등으로 분류되고 분석 결과를 기반으로 Euclidean distance, Mahalanobis distance, Lance-Williams distance 등의 유사도 측정법을 이용하여 클러스터간 유사도를 검증한다[3].

대표적인 클러스터링 분석 알고리즘은 K-means, Expectation-Maximization(EM) 알고리즘, DBSCAN 가 있다. EM 알고리즘은 단순성과 용이성이라는 장점이 있다. 프로그램의 단순함으로 인해 쉽게 병렬화 할 수 있고 다른 알고리즘들 보다 메모리를 적게 차지하며 수치적으로 매우 안정적이다[2]. EM 알고리즘은 확률을 사용하여 적합도를 평가하며 확률 기반 군집에서는 하나의 레코드가 여러 개의 모델에 속할 수 있다. 이때 속하는 정도는 가중치(확률)로서 주어진다.

EM 알고리즘은 관측되지 않는 잠재변수에 의존하는 확률 모델에서 최대 가능도나 최대 사후 확률을 갖는 매개변수를 찾는 반복적인 알고리즘이다. EM 알고리즘은 <표 1>과 같다.

Observed data : X	Unobserved latent data : Z
Parameters : θ	Expected value : Q
1 : Estimate of the parameter θ	
2 : E-Step	
Calculate the expected value	
$Q(\theta \theta^{(t)}) = \sum_z p(z X, \theta^{(t)}) \log L(\theta; X, Z)$	
3 : M-Step	
Find the parameter that maximizes quantity	
$\theta^{(t+1)} = \text{argmax} Q(\theta \theta^{(t)})$	
4 : Iterate steps 2 and 3 until convergence	

<표 1> EM 알고리즘

3. 시스템 환경

본 연구에서 기반하는 시스템 환경은 (그림 1)과 같다. 센서는 소리, 온도, 명암 등의 데이터를 수집하는 장치이다. 서비스 환경은 센서를 비롯한 다양한 디바이스로 구성되어 하나의 서비스를 제공한다. 서비스 그룹은 유사한 기능을 하는 서비스의 집합이다. 플랫폼은 다양한 서비스 그룹들을 제어한다.



(그림 1) 시스템 환경

4. EM 기반 IoT 서비스 분류 알고리즘

본 연구에서 제안하는 EM 기반 IoT 서비스 분류 알고리즘은 3 단계로 구성된다. 1 단계에서는 각 서비스 구성 모듈의 독립변수화를 위하여 Sensing 과 Data management 단계에 유사도 별 분류를 후 EM 알고리즘을 적용하고 그 결과와 기존 Sensing 과 Data management 를 다시 통합하여 EM 알고리즘을 적용한다. 2 단계에서는 Processing type 과 Output 에 EM 알고리즘을 각각 적용한다. 3 단계에서는 Sensing, Data management, Processing type, Output 에서 나온 클러스터링 결과에 다시 EM 알고리즘을 적용한다. 계층적 3 단계 EM 알고리즘 적용으로 보다 세밀한 유사도 추출과 높은 클러스터링 적합도를 기대할 수 있다.

표 2 는 제안하는 EM 기반 IoT 서비스 분류 알고리즘이다.

<표 2> EM 기반 IoT 서비스 분류 알고리즘

```

Clustering of EM algorithm is
input :
  Sensing_type_power S1
  Sensing_type_IP S2
  Sensing_all S
  Data_manage_type_pre D1
  Data_manage_type_volatile D2
  Data_manage_type_all D
  Processing_type P
  Output O
output :
  Result_Clustering X

  Apply_to_EM_1
  input :
    S1, S2, S, D1, D2, D, P, O
  output :
    ClusterS1, ClusterS2, ClusterS, ClusterD1,
    ClusterD2, ClusterD, ClusterP, ClusterO

  Bind_to_similarity
  GroupA = ClusterS1 and ClusterS2 and ClusterS
  GroupB = ClusterD1 and ClusterD2 and ClusterD

  Apply_to_EM_2
  input :
    GroupA, GroupB, ClusterP, ClusterO
  output :
    ClusterA, ClusterB, ClusterC, ClusterI

  Bind_to_remainder
  GroupX = ClusterA and ClusterB
  and ClusterC and ClusterD

  Apply_to_EM_3
  input :
    GroupX
  output :
    ClusterX

return X
    
```

5. 실험 및 분석

본 연구에서는 성능평가를 위하여 Waikato 대학에서 개발한 데이터 마이닝 프로그램 WEKA 3.8.1 [4]을 사용했다. 상용화된 IoT 서비스의 구성 및 동작을 조사하여 37 개의 서비스를 수집하여 데이터 셋으로 구축하였고 수집한 37 개의 서비스에 대하여 제안한 EM 기반 IoT 서비스 분류 알고리즘을 수행하였다. 그리고 적합한 클러스터 개수를 측정하기 위해 클러스터의 개수를 3 개에서 10 개까지 조정하며 실험을 진행하였다. <표 3>은 K = 4, 5, 6 의 결과이다.

<표 3> 클러스터링 결과(K = 4, 5, 6)

K = 4

	Cluster0	Cluster1	Cluster2	Cluster3
Mean	28.1425	2.1425	6.8788	77.6465
Std. Dev	2.2315	0.5308	3.2366	4.2787
Instances	7	22	6	2

K = 5

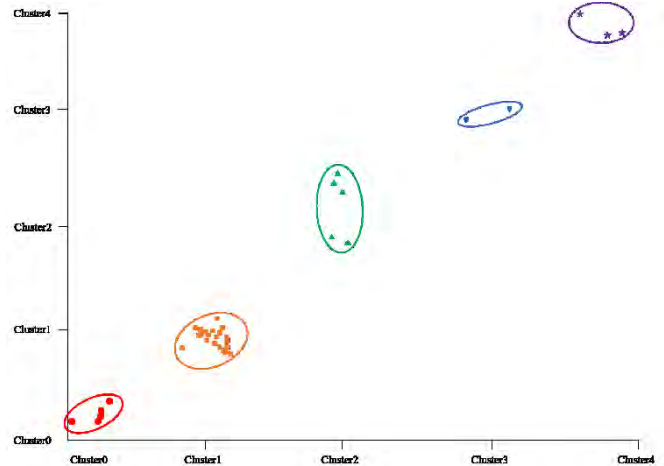
K = 5	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
Mean	26.5066	2.2692	5.8433	77.9994	25.6283
Std. Dev	6.787	0.5276	2.1838	0.1752	0.9526
Instances	5	22	5	2	3

K = 6

K = 6	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Mean	2.2727	25.6667	30	78	14	6
Std. Dev	0.5505	1.1547	18.8891	18.8891	0	0.7071
Instances	22	3	4	2	1	5

적합한 클러스터 개수를 파악하는 기준은 각 클러스터 표준 편차의 평균이 가장 작은 것을 기준으로 했다. 실험 결과 K= 5 인 경우 기준점에 대한 표준 편차의 평균이 가장 작고 유사 서비스 간 클러스터링이 이루어짐을 알 수 있다.

(그림 2)는 클러스터링이 완료 된 것을 그래프로 나타낸 것이다. 클러스터 0 에 5 개, 클러스터 1 에 22 개, 클러스터 2 에 5 개, 클러스터 3 에 2 개, 클러스터 4 에 3 개의 IoT 서비스가 클러스터링 된 것을 확인할 수 있다.



(그림 2) EM 클러스터링 적용 그래프

6. 결론

통합 IoT 플랫폼 서비스 모듈 재활용성과 효율성 증대를 위해 서비스 동작 기반 분류 알고리즘의 필요성이 제안되었다. 본 논문에서는 EM 알고리즘을 적용하여 [1]의 서비스 동작 기반 분류 기법을 확장하였다. 제시한 EM 기반 IoT 서비스 분류 알고리즘을 이용하여 서비스 간 유사도기반으로 클러스터를 구성하였다. 실험을 통하여 구성된 클러스터(K=5)는 기준점에 대한 표준 편차의 평균이 가장 작고 유사 서비스간의 클러스터가 구축됨을 증명하였다.

참고문헌

[1] Lee, Daewon, and Kinam Park. "Development of IoT Service Classification Algorithm for Integrated Service Platform." *International Journal on Advanced Science, Engineering and Information Technology* 7.4 (2017): 1206-1212.

[2] Couvreur, Christophe. "The EM algorithm: A guided tour." *Computer Intensive Methods in Control and Signal Processing* (1997): 209-222.

[3] Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." *IEEE Transactions on neural networks* 16.3 (2005): 645-678.

[4] Weka 3: Data Mining Software in Java, <http://www.cs.waikato.ac.nz/~ml/weka/>