실시간 센서데이터 분류 모델링

변재석[1] 정연성[2] 황지영[3‡]
[1]성균관대학교 통계학과 [2]성균관대학교 전자전기컴퓨터공학과
[3]성균관대학교 소프트웨어학과

Classification of Real-time Sensor Data

Jaeseok Byun[1] Yeonsung Jung[2] Joyce Jiyoung Whang[3‡]
[1]Department of Statistics, Sungkyunkwan University
[2]Department of Electronic, Electrical & Computer Engineering, Sungkyunkwan University
[3]Department of Software, Sungkyunkwan University

요 약

센서와 네트워크 기술의 발달과 인공지능 기술의 발전으로 주변 환경 정보에 대한 실시간 데이터 수집, 분석 및 활용이 가능해지면서 기존의 기술보다 빠르고 정교한 응용서비스를 제공할 수 있게 되었다. 실시간 센서 데이터 분석에는 데이터 특성과 분석 목적에 따라 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)이 모두 사용되며, 지도학습의 경우 목적 값(target value)의 특성에 따라 회귀(Regression) 혹은 분류(Classification)기법을 이용한다. 본 논문에서는 가스 센서를 통해 수집된 데이터를 이용하여, 사람의 후각으로는 감지하기 어려운 음료의 미세한 신선도 차이를 실시간으로 감지할 수 있는 분류 모델을 고안했다. 실험 결과, 분류 모델의 성능이 100%에 근접했으며, 이를 통해 시간에 따라 변화하는 음료의 신선도를 확연하게 분류해낼 수 있음을 확인했다.

1. 서 론

최근 센서 기술과 네트워크 기술의 발달로 이전에는 감지할 수 없던 물리적 세계의 정보에 대한 실시간 데이터 수집 및 분석과 활용이 가능해지면서, 센서 데이터를 이용한 다양한 응용서비스가 창출되고 있다. 이에 따라 시각과 청각 등 인간의 오감을 인식하고 증강할 수 있는 센서 연구가 활발해지고 있다. 현재 시각과 청각 이외의 오감 센서 연구는 미비한 수준이지만, 후각, 미각, 촉각 관련 센서 기술 분야에서도 많은 연구를 통해 발전 가능성을 보이고 있다.

시각 인식 센서 기술은 소비자의 행동 분석이나 장애인의 기기 조작, 자율주행자동차 개발 등에 사용되고 있으며, 청각 센서 기술 역시 단순한 기기 제어를 넘어 사람과의 소통이 가능한 기술로 활용되고 있다. 후각 인식 기술은 시각과 청각 기술에 비해서는 발전이 미비하지만, 유독가스 같이 인체에 해를 끼치는 냄새를 정교하게 분별해내는 기술이 개발되었고, 가스 센서를 통해 음식과 음료의 완성도를 예측하는 연구가 진행되었다[1].

‡ 교신저자 (Corresponding Author)
본 연구는 한국연구재단을 통해 과학기술정보통신부의 이공분야기초연구사업(2019R1C1C1008956)과 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원(No.2019-0-00421, 인공지능대학원지원), 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(2015-0-00914).

센서 데이터 분석에는 데이터의 특성과 목적에 따라서 다양한 기계학습(Machine Learning) 기법들이 사용된다. Label을 가지고 있는 데이터의 경우, 미래 데이터의 연속적인(Continuous) 목적 값(Target Value)을 예측하기 위해서 회귀(Regression) 분석 기법이 사용되거나, 미래 데이터가 어떤 Class label에 속할지 예측하기 위해 분류(Classification) 기법이 사용된다. 데이터가 특정 Label을 가지고 있지 않을 경우에는 군집화(Clustering) 기법 등 비지도학습(Unsupervised Learning)을 사용하여 데이터를 분석한다[4]. 본 연구에서는 사전 연구 자료를 통해 실시간 센서 데이터에 직접 범주형(Categorical)의 Label을 부여함으로써, 분류(Classification) 모델을 이용하여 실시간으로 센서에서 수집되는 데이터의 Label을 예측했다.

본 논문의 실험에서는 센서 데이터 중 관련 연구와 실험이 다소 미비했던 후각 데이터를 사용하여 분류 모델링을 실시했다. 음식 및 음료는 산소와의 접촉 등 외부의 영향으로 인해 부패가 진행되고, 이 과정에서 많은 가스가 방출된다. 부패가 진행되는 동안에는 시간에 따라 발생하는 가스의 종류 및 양이 달라진다. 따라서, 본 연구에서는 시간에 따라 다르게 발생하는 음료 가스 데이터를 Arduino Multichannel Gas Sensor를 통해 수집하고, 사람의 후각으로는 쉽게 알아차릴 수 없는 음료 신선도의 미세한 차이를 감지할 수 있는 분류 모델을 고안했다. 실험 결과, 사용한 분류 모델에서 대부분 100%에 가까운 성능을 보이며, 시간에 따라 음료에서 분출되는 가스의 양이 확연하게

다름을 실험적으로 확인했다.

2. 관련 연구

기존의 가스 센서 데이터 관련 연구에서는 음식이 만들어지는 과정에서 발생하는 가스의 변화량을 통해 음식의 완성도를 예측하는 모델을 개발했다[1]. 음식의 완성도를 3개의 상태(덜 익은 상태, 적절히 익은 상태, 과하게 익은 상태)로 나누어 class label을 부여하고, Random Forest 기법을 이용하여 학습시켰다. 센서 데이터 분석을 위한 분류 모델로는 위의 연구와 같이 하나의 분류(Classification) 기법을 단일 모델로 사용하거나, 센서 데이터에서 발생하는 예측 불가능한 변동을 줄이고 모델의 성능 향상을 위해 여러 모델을 결합한 앙상블(Ensemble) 분류기를 모델로 사용한다[2]. 센서 데이터에 class label을 부여하기 어려운 경우에는 군집화(Clustering) 기법 등 비지도학습(Unsupervised Learning)을 사용한다. 데이터의 변동성이 큰 센서 데이터의 특성상 군집(Cluster)에 속하지 않는 이상치(Outlier)가 존재하거나, object가 하나의 군집이 아닌 여러 군집에 동시에 속하는 경우(Overlapped)가 발생할 가능성이 높다. 따라서 이상치와 여러 군집에 속하는 경우를 동시에 고려할 수 있는 군집화 알고리즘[5]을 사용할 경우, 일반적인 군집화 기법보다 우수한 성능을 보일 것으로 기대된다. 본 논문에서는 class label을 부여하여 분류(Classification) 기법을 통해 모델링을 진행했으며, 다양한 분류 알고리즘을 사용하여 성능 평가를 진행했다.

3. 분류 알고리즘

본 논문에서는 정확한 성능 비교 및 평가를 위해, Scikit-Learn package의 5개의 모델 Support Vector Machines(SVM), Logistic Regression(LR), Random Forest(RF), Artificial Neural Network(ANN)을 사용하여 분류를 실시했다[3].

3.1. Support Vector Machines (SVM)

분류 모델 중 성능이 다른 분류 모델들보다 우수하다고 알려진 Support Vector Machines는 kernel 함수를 이용하여 데이터를 고차원으로 사영시킨다. 사영된 고차원에서 마진(Margin)이 최대가 되는 초평면(Hyperplane)을 찾는 것을 목적함수로 두어, 기존의 단순 분류 모델보다 정밀한 분류가 가능하다. 본 논문에서는 비선형 패턴을 감지할 수 있는 kernel인 Gaussian kernel를 사용했다.

3.2. Logistic Regression (LR)

대표적인 통계학 기반 분류 모델인 Logistic Regression은 반응 변수가 class label이라는 차이점 이외에는 전통적인 선형 회귀식과 유사하다. Log odds(오즈)값을 이용하여 값을 구하며, 함수 형태는 sigmoid 함수와 동일하다. 구현이 간단하며, 다른 모델들과 달리, 모델의 결과를 해석할 수 있다는 장점을 가진다.

3.3. Random Forest (RF)

Random Forest는 앙상블(Ensemble) 학습 분류 기법의 일종으로, 기존 Decision Tree 모델의 과적합(Overfitting) 문제를 완화시킨 모델이다. 여러 Decision Tree를 생성하여, 각 Decision Tree에서 나오는 결과값을 voting 방식으로 종합하여 최종 결과를 도출한다.

3.4. Naïve Bayes (NB)

Naïve Bayes 기법은 베이즈 정리를 이용한 분류기로, 특정 class label에 속할 조건부 확률을 구해 object를 특정 class에 할당하는 모델이다. 독립성을 가정하고 있어, 독립성을 만족하지 않는 데이터의 경우, 분류가 제대로 되지 않을 수 있다.

3.5. Artificial Neural Network (ANN)

Artificial Neural Network는 사람의 신경망 구조를 모사한 분류기법으로, 최근 뛰어난 성능으로 각광받는 모델이다. 입력층, 은닉층, 출력층으로 구성되며, 은닉층에서 입력값에 대한 학습을 거친 뒤 출력값을 내보내는 방식이다. 데이터의 수가 feature 수에 비해 상대적으로 적은 경우, 과적합(overfitting)의 문제가 발생할 수 있다. 따라서, 본 논문에서는 단일 은닉층을 사용하고, 해당 은닉층의 노드 개수를 3~5개로 두고 실험을 진행했다.

4. 실험 결과

실험을 위해 8 종류의 가스(NH3, CO, NO2, C3H8, C4H10, CH4, H2, C2H50H)를 ppm(Parts Per Million) 단위로 수집할 수 있는 Arduino Multichannel Gas Sensor를 사용했으며, 실생활에서 흔히 접하는 음료인 커피에서 분출되는 가스 센서 데이터를 수집했다. 센서 데이터의 특성상 수집된 데이터의 변동이 크기 때문에, 데이터 정규화를 실시한 이후, 분류 모델링을 실시했다.

4.1. 실험 데이터

커피 신선도 측정을 위해, Arduino Multichannel Gas Sensor를 사용하여 같은 시간, 같은 장소에서 하루 간격으로 데이터 수집을 실시했다. 3일간 가스 데이터를 수집했으며, 하루당 1분 동안 측정했다. Arduino Gas Sensor에서 데이터는 0.2초에서 0.8초 간격으로 수집되며, 1분 동안 데이터를 측정할 경우, 100~150개의 데이터가 수집된다. 본 실험에서 수집된 총 데이터의 개수는 329개이다. 일별 가스 분출량의 차이를 측정하기 위한 실험이기 때문에 수집된 데이터에 3개의 class label(Day1, Day2, Day3)을 부여했다. 데이터의 분포는 아래 <표1>, <표2>와 같다.

<표 1. Class Label 분포>

Label	Day1	Day2	Day3
데이터 수	121	63	145

<표 2. Feature 분포>

Gas	Range
NH3 (암모니아)	0.0 ~ 0.22
CO (일산화탄소)	0.89 ~ 2.86
NO2 (이산화질소)	1.23 ~ 2.72
C3H8 (프로판)	0.03 ~ 100.93
C4H10 (뷰테인)	26.7 ~ 91.52
CH4 (메테인)	1.75 ~ 130.56
H2 (수소)	0.06 ~ 0.38
C2H50H (에탄올)	0.2 ~ 0.93

4.2. 데이터 전처리

위 <표2>에서 확인할 수 있듯이, 수집된 데이터의 feature 변수 간에 변동이 크고, 개별 feature 안에서의 편차 역시 크기 때문에, 데이터 정규화를 진행했다. 본 실험에서 사용한 정규화 기법은 Min-Max Normalization 기법으로, 수식은 다음과 같다.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-Max Normalization을 통해 변수 값의 범위를 0~1로 조정함으로써, 변수 간의 변동을 줄이고, 개별 feature의 편차를 줄였다.

4.3. 모델 성능 평가

음료 신선도를 실시간으로 분류하기 위해서는, 미리 학습된 모델을 사용하여 신선도를 예측해야 한다. 따라서, 본 논문에서는 실제 환경과 유사한 환경에서의 분류 모델 성능을 측정하기 위해, 일별로 시간상 먼저 수집된 데이터의 80%를 훈련 데이터(Training data)로, 남은 20%의 데이터를 테스트 데이터(Test data)로 두고 실험을 진행했다. 실험 결과는 아래 <표3>과 <표4>와 같다.

<표 3. SVM 모델 사용했을 시 Confusion Matrix>

	Day1	Day2	Day3
Day1	19	0	0
Day2	0	16	0
Day3	0	0	31

<표 4. 모델 성능 평가>

	SVM	LR	RF	NB	ANN
Accuracy	100%	100%	100%	98.48%	100%
Precision	100%	100%	100%	98.77%	100%
Recall	100%	100%	100%	98.81%	100%
Micro-F1	100%	100%	100%	98.77%	100%
Macro-F1	100%	100%	100%	98.48%	100%

<표3>에서 SVM을 사용했을 때, 여러 개의 Label을 모두 정확하게 분류하고 있음을 confusion matrix를 통해 확인할 수 있다. 또한 표<4>에서도 알 수 있듯이, SVM 뿐만 아니라 사용한 모든 분류 모델에서 대부분의 지표가 100%에 근접함을 알 수 있다. 이를 통해 일별로 발생하는 커피 센서 데이터를 효과적으로 분류해내고 있음을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 센서에서 실시간으로 수집되는 데이터를 다양한 기계학습(Machine Learning) 알고리즘을 이용하여 분류했다. 실험결과 사용한 대부분의 모델에서 모든 지표에서 100%에 가까운 성능을 보였고, 이를 통해 본 논문에서 사용한 분류 모델이 특정 시간 간격에 따른 음료의 신선도 차이를 효과적으로 분류해내고 있음을 확인했다. 향후 연구에서는 데이터 수집 간격을 24시간에서 3시간으로 줄여, 보다 짧은 시기에서 발생하는 음료 신선도의 미세한 차이를 분류할 수 있는 실험을 진행할 계획이며, 가스 센서 데이터 이외에도 촉각 센서와 미각 센서 등 보다 다양한 센서 데이터를 이용하여 분류 실험을 진행할 계획이다.

참고문헌

[1] Sen H. Hirano, Gillian R. Hayes, Khai N. Truong, “uSmell: exploring the potential gas sensors to classify odors in ubicomp applications relative to airflow and distance”, *Personal and Ubiquitous Computing*, Vol. 19, 2015.

[2] A. Vergara, T. Ayhan, S. Vembu, R. Huerta, M. Ryan, M. Homer, “Gas Sensor Drift Mitigation using Classifier Ensembles”, *SensorKDD*, 2011.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrotm and E. Duchesnay, “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, Vol. 12, 2011.

[4] J. J. Whang, Y. Hou, D. F. Gleich, and I. S. Dhillon, “Non-exhaustive, Overlapping Clustering”, *TPAMI*, 2018.

[5] J. J. Whang and I. S. Dhillon, “Non-exhaustive, Overlapping Co-Clustering”, *CIKM*, 2017.