대규모 감성DB 기반 인지컴퓨팅에서의 편향 요인 진단

곽동현 O1 김병희 2 박태서 3 장병탁 1,2,3
뇌과학 협동과정 1, 컴퓨터공학부 2, 인지과학 협동과정 3, 서울대학교
{dhkwak, bhkim, tspark, btzhang}@bi.snu.ac.kr

Bias Analysis of Large-scale Affective Database for Cognitive Analytics

Dong Hyun Kwak O* Byoung-Hee Kim** Tae-Suh Park*** Byoung-Tak Zhang*,**,***
Brain Science Program*, School of Computer Science & Engineering**,
Cognitive Science Program***, Seoul National University

요 약

최근 감성 기반의 인지컴퓨팅 분야에서는 영상 컨텐츠에 대한 사람들의 공통적인 감정 반응 모델에 대한 연구가 활발하다. 특히, 다수의 사람을 대상으로 한 컨텐츠 시청 실험 결과를 데이터베이스화하고 기계학습을 이용하여 분석 및 모델링하는 과정이 표준적인 프로토콜로 자리잡고 있다. 사람의 감정 반응은 다양한 요인에 영향을 받기 때문에, 이러한 데이터베이스를 기반으로 한 인지컴퓨팅 연구는 다양한 면에서 주의를 요한다. 본 논문에서는 감정 반응 예측을 위한 모델링 과정에서 가장 주의해야 할 두 가지의 편향 요인으로서 감정 반응 수집 방식과 감독학습을 위한 데이터셋 설정 과정을 분석한다. 특히, 최근 공개된 대규모 감성 데이터베이스 분석 결과 발견한 학습데이터와 테스트데이터 간의 편향을 보고하고, 최신 기계학습 연구를 반영한 대안을 제시한다.

1. 서 론

사람의 감정을 모델링하고 컴퓨터 시스템 운용에 접목하고자 하는 감성 컴퓨팅(affective computing)은 인지컴퓨팅 분야의 대표적 연구 주제 중 하나로서 최근 국내외에서 활발한 연구가 진행되고 있다. 데이터를 기반으로 감정 상태 변화의 지표를 파악하고 예측하기 위해서는 사람을 대상으로 실험한 대규모의 데이터가 필요하다. 이에 따라 멀티미디어 자극에 대한 다양한 지표를 측정, 기록하는 실험을 기반으로 대규모 데이터베이스(DB)가 구축되는 추세이다. 대표적인 사례로 LIRIS-ACCEDE[1], FilmStim[2], DEAP[3] 등이 있다. 이들 대규모 감성 데이터베이스에 감독학습 기반의 기계학습 기법을 적용하여 감정 예측 모델을 구축하는 연구가 최근 인지 컴퓨팅 분야에서 활발하다[4-7].

사람의 감성을 정확하게 예측하려는 연구는 학술적, 산업적으로 다양한 활용이 가능하다. 산업적으로는 멀티미디어 컨텐츠로부터 예상되는 하이라이트 지점을 찾아 더 강렬한 효과를 주거나, 유저 성향에 특화된 맞춤형 추천을 해줄 수 있다. 또 사람이 느끼는 감정에 맞춰 적절한 반응과 제스처를 취하는 교감형 로봇도 구현할 수 있다.

이러한 감성 데이터베이스 구축에서 감정 반응 수집 방식에 따른 문제점과 감독학습을 위한 학습데이터와 테스트데이터 설정 과정에 따라 편향이 발생한다.

감정 반응 수집 방법은 크게 설문이나 인터뷰와 같은 직접적인 피드백을 요청하는 방식과 사용자의 암묵적, 무의식적인 반응을 센서를 이용해 기록하는 두 가지 방식이 있다. 두 가지 방법의 장/단점에 대해 살펴보고, 특히 직접적인 피드백 방식의 데이터에서 발생하는 문제를 분석한다.

감독학습 기반의 예측 모델링을 위해서는 수집한 데이터를 학습데이터와 테스트데이터로 구분하게 된다. 보통의 경우 두 데이터 군이 동일한 모집단/환경에서 수집되었다는 가정 하에 모델을 학습한다. 그러나, 실세계 문제에서는 두 데이터가 유래한 모집단이 같지 않거나 시간에 따라 변동하는 상황이 빈번하며, 이에 대해 기계학습 및 인공지능 분야에서는 전이 학습(transfer learning)[8] 또는 개념 유동(concept drift)[9]이라는 주제로 연구가 되고 있다. 감성 컴퓨팅 분야에서는 이러한 문제에 대한 인식이 아직 부족한 것으로 보인다.

본 논문에서는 대표적 감성 데이터베이스인 ACCEDE[1] (그림 1)를 분석하여 이러한 문제를 확인하고, 데이터 구축 단계에서의 해결방안과 모델 학습 단계에서의 해결방안에 대해 논의한다.

2. 감성DB 기반 인지컴퓨팅 연구 시의 문제점

2.1 감정 반응 수집 방식에 따른 편향 요인

ACCEDE에서 사용된 방식과 같이 피험자의 의견에 대한 직접적인 피드백을 요청하는 실험 방식에는 인지적, 사회 심리학적 요인에 기인한 다양한 편향(bias)이 포함되기 쉽다. 따라서 이러한 데이터를 분석하는 과정에서는 편향에 대한 직접적인 모델링이 필요하다[10].

[가장 낮은 arousal 클립]

[가장 높은 arousal 클립]

그림 1. ACCEDE 는 160개 영화에서 추출한 9800개의 클립(각 8~12초 길이)에 대해 온라인 상에서 피험자의 직접적인 피드백을 받아 arousal(흥분도)와 valence(긍정도)를 기준으로 각각 순위를 매겨 제공하는 데이터베이스이다.

그러나 감정 반응에 관련한 피드백의 경우는 일반적인 설문 조사에 비해 보다 심각한 편향 요소와 노이즈가 개입될 소지가 있다. 예를 들어, 일반적인 설문 조사에서는 객관적인 응답이 가능한 주거 환경, 소득수준, 제품의 사용기간 등과 같은 질의가 많은 반면, 감정 반응에 대한 설문 조사에서는 오직 응답자의 주관적인 느낌과 감정을 수집하기 때문에 개인의 편향이 개입될 요인이 다분하다.

ACCEDE는 온라인 상에서 크라우드소싱을 통해 익명의 유저에게 응답 시 보상을 주는 방식으로 감정 반응을 수집하였다. 이 과정에서 유저는 보상만을 목적으로 무작위 응답을 하는 편향 요인이 추가되었을 가능성이 있다. 또한 감성 DB 수집 시 피험자 모집단과 실험 환경을 일정하게 컨트롤해야 하지만, 크라우드소싱은 이러한 조건을 만족하기 어렵기 때문에, 설문 문항을 접한 순서나 주변의 시간-환경적 요인으로 인한 심리적인 편향이 유저의 응답을 불규칙적으로 만들 수 있다.

2.2 감성 DB 기반 감독학습 모델 구축시의 편향 요인

ACCEDE에서는 제공하는 감독학습을 위한 학습데이터와 테스트데이터는 동일한 영화에서 추출한 클립이 학습데이터 또는 테스트데이터 중 한 곳에만 포함되도록 구성이 되어 있다. 이 경우 각 데이터가 추출된 모집단이 서로 달라 데이터가 충분히 크지 않은 경우 학습데이터와 테스트데이터에서 covariate shift가 발생할 수 있다. 다음 절에서 데이터 분석을 통하여 이러한 편향을 확인한다.

2.3 대규모 감성 DB의 편향 요인 분석

대규모 감성 DB의 대표적인 사례로서 ACCEDE를 선정하고, DB에서 제공하는 학습데이터와 테스트데이터 구성에 편향 요인이 포함되었는지를 확인하기 위한 통계적 분석을 수행하였다. 학습데이터와 테스트데이터 비교를 위한 feature로는 영상과 음성을 구분하였으며, 각 feature 공간을 기준으로 두 데이터 군 간의 차이에 대해 분석한다.

영상 feature로는 4Hz의 주기로 한 스냅샷에 대해 Hue & Tone 기반 130종의 색상을 추출하였으며[5], 음성 feature로는 MFCC와 LPC를 추출하였다(MFCC 20 bands, LPC 12 bands. 기반 볼륨과 Delta 및 ACC 값도 추출). 각 클립 별로 샘플링 된 feature기반의 평균, 중앙값, IQR (inter-quartile range), 최대값 및 최소값을 최종 feature로 선정하였다.

비교를 위해 대표적인 감독학습 성능 비교용 데이터인 Iris와 MNIST를 함께 분석하였다. Iris는 붓꽃의 아종 3가지를 구분하기 위한 고전적인 데이터이며, MNIST는 대표적인 숫자 필기체 데이터베이스로서, 영상처리와 기계학습 분야에서 표준적인 비교 데이터로 널리 쓰이고 있다.

표 1. 전형적인 감독학습용 데이터와 감성DB의 비교실험을 위한 구성. PCA 차원 축소는 학습 데이터에 PCA를 적용하고 원본 분산의 90%를 커버하는 principal component를 선택.

데이터	Instance 수	Feature 수	PCA 차원축소 (분산 범위 90%)
Iris	학습 : 100 테스트 : 50	4	1
MNIST (숫자 0,1)	학습 : 12,000 테스트 : 2,000	768	49
ACCEDE	**학습 : 3,638** **테스트 : 3,662**	**Vision : 670** **MFCC : 315** **LPC : 195**	**Vision : 191** **MFCC : 141** **LPC : 40**

[표 1]과 같이 구성한 데이터에 대해 학습데이터와 테스트 데이터의 평균과 분산이 동일한지에 대한 통계적 검정을 수행하였다. 평균 비교는 각 feature 별로 이분산을 가정하여 T-검정을 수행하고[1], 분산 비교는 F-검정을 수행한다. 다중 변수를 갖는 집단 간의 비교를 위해 다중 가설 검정을 추가로 수행한다. 다중 가설 검정의 기준으로는 FDR (false discovery rate)를 선택하여, 두 집단이 동일할 확률 값으로서 π_0를 보고한다[2].

통계적 검정 수행 결과 [표 2]에서와 같이 Iris와 MNIST 데이터의 경우 학습데이터와 테스트데이터가 큰 차이를 보이지 않았지만, ACCEDE는 평균과 분산 모두 차이가 있다는 결과가 나타났다. 즉, 기존의 표준적인 기계학습 모델 비교용 데이터와는 달리 ACCEDE에서는 학습데이터와 테스트데이터가 동일한 모집단에서 추출되었다는 가정이 유효하지 않다고 판단할 수 있다.

3. 감성 DB 편향 요인에 대한 대안

표 2. 학습 데이터와 테스트 데이터간 차이에 대한 통계적 검정 결과(H0: 두 집단간 통계량 동일함). Feature별 t-test 수행 후 다중 가설 검정을 FDR로 수행한 결과 π0값, 즉, **두 데이터의 통계량이 동일할 확률 값**을 표기함.
(X: 전체 feature 적용, PCA(X): 학습 데이터에서 추출한(표 1) principal component를 기준으로 비교, (*): Iris 데이터의 유효 PC 1개를 기준으로 t-test 수행한 결과 p-value, (**): feature별 t-test 수행 결과 모든 경우에서 p-value≪0.001로 관측됨)

데이터 셋	평균 비교		분산 비교	
	X	PCA(X)	X	PCA(X)
Iris	1.0	0.857(*)	1.0	0.921(*)
MNIST	0.689	0.475	0.582	0.813
ACCEDE-Visual	0.394	0.158	0.015	0.005
ACCEDE-MFCC	0.12	0.455	7.06e-4	0.006
ACCEDE-LPC	0.009	0.257	0.058	(**)

[1] Feature 전체를 기준으로 Hotelling's T^2 기반 검정도 수행하였으나 표2 결과 표기의 통일성을 위해 제외함

[2] Tool은 Matlab Bioinformatics Toolbox의 mafdr 적용

3.1 감정 반응 수집 방식에 따른 편향 요인 처리

감성 DB에서의 편향을 처리하기 위한 방법으로 먼저 편향 요인을 명시적으로 모델에 포함시키는 방법이 있다. 감독학습 시 유저의 성향, 성별, 나이, 응답 시간대 등을 편향 요소로써 고려하여 모델 구축 단계에 적절한 처리 방법을 포함시켜 예측 성능을 높인 사례는 추천 시스템 연구에서 찾아볼 수 있다[11]. 또는 데이터의 전처리 단계에서 이런 편향 요소를 먼저 제거하고 사용할 수도 있다[12].

그밖에 감정 DB를 구축하고 수집하는 방법으로 직접적 피드백 요청 방식 대신, 암묵적 반응 수집 방식으로 대체하는 방법이 있다. 이는 유저의 무의식적인 반응을 기록하므로 직접적인 피드백에 비해 편향이 적다. 그러나 이 경우 수집된 데이터를 분석하고 모델링하여 해석하는 추가적인 과정이 필요하다.

표 3. 두 가지 감정 반응 수집 방식 비교

방식	장점	단점
직접적 피드백 요청	크라우드소싱 등의 방법을 통한 대규모 데이터 수집 가능	실험 데이터에서 편향 요소에 대한 고려가 필요
암묵적 반응 기록	피험자의 편향되지 않은 데이터 수집이 가능	수집된 데이터에 정교한 분석과 모델링을 통한 해석이 과정이 필요

3.2 감정 반응 예측 모델 구축을 위한 대안

2절의 분석 결과처럼 유사한 환경에서 수집한 학습데이터와 테스트데이터가 유의미한 차이를 보이는 가장 큰 이유는, 데이터를 나눌 때 특정 패턴이 한쪽에만 포함되기 때문이다. 따라서 이를 해결하는 가장 직관적인 방법은 특정 패턴을 보이는 데이터를 학습데이터와 테스트데이터에 적절히 분배하는 것이다. 예를 들어 ACCEDE의 경우, 현재는 특정 영화의 클립이 학습 또는 테스트 데이터 한 쪽에만 포함되어 있는데, 이를 양쪽에 분배하여 이러한 편향을 줄일 수 있다.

안정적 모델링을 위한 보다 적극적인 방법으로서, 학습데이터에 유효한 변형을 가하여 추가의 학습데이터를 생성함으로써 일반화 성능을 높이려는 시도가 가능하다. 대표적인 사례로 무한 MNIST가 있다[3]. 감정을 설명하는 요인에 대한 유효한 변형 탐색은 이러한 점에서 의미 있는 연구 주제이다.

또 다른 근본적인 대안으로 전이 학습(transfer learning)이 있다[8]. 전이 학습은 여러 가지 관점에서 응용이 가능하여 어떤 문제에 적용된 모델을 다른 문제에서 활용하거나, 또는 서로 다른 모집단에서 학습과 테스트 데이터를 수집한 경우 covariate shift를 해결하기 위해 적용 가능하다. 최근 많은 연구가 이루어지는 중이다.

4. 논의 및 결론

본 논문에서는 감성 컴퓨팅 분야에서 자리잡고 있는 대규모 DB 구축 및 기계학습을 이용한 감정 반응 예측 모델링 과정에 두 가지 큰 장애 요소가 있음을 밝히고, 그에 대한 대안을 제시하였다. 감정 반응에 대한 모델링은 암묵적 피드백 정보를 기초로 데이터 기반 접근법을 취하는 것이 적절하다는 결론을 내렸으며, 기계학습의 감독학습 모델링을 적용하는 경우 학습데이터와 테스트데이터 구성에 유의해야 함을 살펴보았다. 특히, 최근 공개된 ACCEDE 감성 DB의 경우, DB에서 제공하는 학습데이터와 테스트데이터가 동일한 모집단에서 나왔다는 가정을 취할 수 없음을 통계적 검정을 통해 보였다.

인간의 감정 반응을 예측하는 문제는 개별적 차이와 정적이지 않은 특성으로 인해 난제로 남아 있다. 현재 감성 컴퓨팅 연구자들이 집중하고 있는 감정 반응 유발 요인 탐색과 함께 기계학습을 중심으로 한 데이터 과학의 첨단 기법이 함께 적용되면 의미 있는 발전이 이루어질 것으로 기대한다.

감사의 글

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734-Videome,), 정부(미래창조과학부 및 정보통신기술진흥센터)의 정보통신·방송 연구개발사업 지원 (10035348-mLife, 14-824-09-014, 10044009-HRI.MESSI)을 일부 받았음.

참고문헌

[1] Y. Baveye, J.-N.Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, “A Large Video Data Base for Computational Models of Induced Emotion,” *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 13–18, 2013.

[2] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, “Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,” *Cogn. Emot.*, vol. 24, no. 7, pp. 1153–1172, 2010.

[3] S. Koelstra, M. Soleymani, J. Lee, A. Yazdani, T. Pun, A. Nijholt, and I. Patras, “DEAP : A Database for Emotion Analysis Using Physiological Signals,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[4] 김은솔, 김지섭, 이대근, 장병탁, 맞춤형 추천을 위한 다중센서기반 사용자 인지 지능형 TV 플랫폼, *한국정보과학회 가을학술발표 논문집*, 제39권 2(B), pp. 259-261, 2012.

[5] W. Choe, H.-S.Chun, J. Noh, S. Lee, and B.-T. Zhang, “Estimating multiple evoked emotions from videos,” in *Proceedings of Annual Meeting of the Cognitive Science Society*, pp. 2046–2051, 2013.

[6] T.-S.Park, B.-H.Kim, and B.-T. Zhang, “A viewer preference model based on physiological feedback”, *Journal of the Korean Institute of Intelligent Systems*, vol. 24, no. 3, pp. 316-322, 2014.

[7] M. Soleymani, M. Pantic, and T. Pun, “Multimodal Emotion Recognition in Response to Videos,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, 2012.

[8] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[9] N. Lu, G. Zhang, and J. Lu, “Concept drift detection via competence models,” *Artif.Intell.*,vol. 209, pp. 11–28, 2014.

[10] M. Hilbert, “Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making,” *Psychol. Bull.*, vol. 138, no. 2, pp. 211–237, 2012.

[11] N. Koenigstein, G. Dror, and Y. Koren, “Yahoo! Music Recommendations: Modeling Music Ratings with Temporal Dynamics and Item Taxonomy,” in *ACM Conference on Recommender Systems*, pp. 165–172, 2011.

[12] Uncovering response biases in recommendation, K.-W. Park, B.-H. Kim, T.-S. Park, and B.-T. Zhang, *AAAI 2014 Multidisciplinary Workshop on Advances in Preference Handling (M-PREF)*, pp. 73-78, 2014

[3] http://leon.bottou.org/projects/infimnist