비디오 시청자의 감정 예측을 위한 다중 뷰 기반 감독학습 연구

박경화[1], 김병희[2], 장병탁[1,2]
[1]서울대학교 뇌과학 협동과정, [2]서울대학교 컴퓨터공학부

{kwpark, bhkim, btzhang}@bi.snu.ac.kr

Multi-view Approach to Video-Induced Emotion Prediction

Kyung-Wha Park[1], Byoung-Hee Kim[2], Byoung-Tak Zhang[1,2]
[1]Brain Science Program, Seoul National University
[2]Department of Computer Science and Engineering, Seoul National University

요 약

사진, 음악, 동영상 등의 멀티미디어에 대한 감정 반응은 문화적 요인이기에, 감정을 이해하고 인식, 예측을 목표로 하는 학제적 연구가 활발하다. 특히, 난해한 심리학적, 뇌과학적 모델을 대체 또는 보완할 수 있는 대안으로서 데이터 기반의 기계학습 모델을 이용한 감정 인식 연구가 크게 주목받고 있다. 본 논문에서는 다중 뷰 기반의 감독학습이 감정 인식 및 예측 모델로서 다양한 장점과 잠재적 활용성이 있음을 서베이와 실제 데이터 기반 실험을 통해 보인다. 특히, 기존의 휴리스틱에 기반한 단순 뷰 통합 방식보다 최적화 설정 및 커널 기법에 기반한 다중 뷰 학습 기법의 장점에 대해 논의한다.

1. 서 론

최근 MIT의 미디어랩에서는 Jibo라는 이름의 가정용 로봇 보급에 나섰다. 이 작은 로봇은 사람과 상호 작용하며 로봇 자신의 감정 상태를 사용자에게 전달하는 기능을 선보인다고 한다. 이 이전에도 로봇 완구의 영역에서 인공지능이라는 이름으로 아이들의 감정 상태를 예측하려는 시도는 계속 있어왔다. 인공지능은 사람과 감정적으로 교감을 시도하고 있다. 스마트 기술의 다음 목표로 감정, 의도, 선호도와 같은 사용자의 인지 상태 예측 기술에 대한 관심이 증대하고 있으며, 특히 가트너 그룹에서 매년 발표하는 각광받는 신기술(emerging technology) 목록에 그 이름을 올리면서 감성 기반 인지 컴퓨팅(affective computing)의 중요성이 커지고 있다[1]. 유발된 감성을 예측하는 이 분야는 특히 많은 영상 기반의 데이터 셋을 공개하며 연구에 집중하고 있다.

영상을 이용한 감성 기반 인지컴퓨팅 연구에는 크게 두 가지가 있는데, 영상을 분석하여 거기서 유발되는 감정을 예측하는 연구와 영상 시청자의 감정 반응을 분석하여 예측하는 연구로 구분된다[1-2]. 어느 방법이든 유리천장과 새로운 돌파구를 언급할 정도로 매우 어려운 연구이다. 기존의 접근 방식은 다중 모달 퓨전(이하 MMF)을 기반하고 있다[3]. 예를 들어, 모달 데이터가 음성과 텍스트처럼 서로 다른 포맷이거나 음성과 영상의 샘플링 빈도가 다를 경우 등 각각의 모달리티에 대해 수집된 feature를 합치거나 혹은 각각에 대해 분류 결정 결과를 합치는 것을 MMF라고 부른다. 또는 여러 개의 feature sets을 다른 말로 다중 뷰라고 부르며 하나의 데이터 집합 또는 모달리티를 하나의 뷰(view)라고도 한다.

본 논문에서는 기존 MMF에서 휴리스틱 단계에 머물렀던 감정 예측 모델에 기계학습에서 제시하고 있는 보다 체계적인 해법으로서 MKL과 EKP를 도입하여 성능 향상 및 다양한 추가의 장점을 취하고자 한다.

1) http://www.gartner.com/newsroom/id/2819918

2. 기존 연구

MMF에는 고려해야 될 네 가지 사항이 있다. 퓨전의 단계, 어떻게, 언제, 무엇을 퓨전할 것인가에 대한 결론이 필요하다. 퓨전의 정도란 feature 단계에서의 퓨전(early-fusion)을 할 것인가 아니면 decision 결과 단계에서의 퓨전(late-fusion)을 할 것인가에 대한 결론이다.

퓨전을 언제 할 것인지가 큰 문제가 되는데, 다양한 capture rate가 존재하는 데이터를 어떻게 동기화 시킬 것이며 언제까지 수집할 것인지 등 많은 결정을 요구하기에 어려운 단계라고 할 수 있다. 무엇을 퓨전할 것인지 결정할 때, 퓨전 시 모달리티 간의 상호보완 관계이거나 서로 모순되는 정보가 존재할 수 있기 때문에 어떻게 선택할지를 결정하는 단계이다. 어떻게 퓨전 하느냐에 대해선 아주 다양한 방식이 있다. rule-based 방식으로 majority voting을 통한 퓨전, Estimation-based 방식으로 Kalman Filter를 통한 퓨전, classification-based 방식으로 SVM을 통한 퓨전 등 매우 다양하다[1].

이렇게 MMF에는 고려할 점이 많고 휴리스틱이 개입할 여지가 많아 이러한 연구에서는 견고한 체계(framework)가 중요한 연구 소재가 된다. 특히 앞서 말한 성능이 뛰어난 커널 머신인 SVM을 실험에 사용하는 연구는 많이 있지만, 그 연장선상에 있는 커널을 여러 개 사용하는 다중 커널 학습(이하 MKL)을 이용한 연구는 드물었다. 다중 커널 학습은 Ensemble Kernel Predictors와 더불어 대표적인 다중 뷰 기반 학습 중 하나이다.

다중 모달리티 데이터의 classifier를 결정할 때 heuristic에 의존하거나 크로스 확인을 이용하는데, 모든 것을 아우르는 체계적인 방법으로서 다중 뷰 기반 학습이 좋은 선택이라는 것을 이 논문에서 보여주고자 한다.

Mid-level의 시각 features를 통해 계산 모형을 설계하고, 각각의 비디오의 같은 장면에서 사람들 마다 다른 감정을 일으킨다는 결과의 연구가 있었다[4]. 또는 비디오 시청자로부터 EEG, 홍채 반응, 응시 거리 수집하여 MMF를 한 분류 결과가 설문 조사보다 낫고 late-fusion이 제일 낫다는 것을 보여주는 연구가 있었다[5]. 기존 연구에

서 다중 모달에 대한 연구는 많이 있었지만 이것을 퓨전하는 데 있어서는 굉장히 휴리스틱에 기반을 둔 접근을 했었다.

3. 다중 뷰 기반 감독학습

3.1 다중 뷰 학습

다중 뷰 판별기를 학습할 때 서로 다른 views로 학습한 판별기를 통해 성능을 향상시키는 것이 다중 뷰 학습에서 집중하는 문제이다[6]. 여러 개의 서로 다른 feature sets을 통해 학습하는 방식은 감성 기반 인지컴퓨팅과 멀티미디어 분야에서 다루는 MMF와 문제 설정 면에서 유사한 점이 있다. 그러나, MMF는 앞서 말했 듯이, 체계적이지 않고 휴리스틱이 개입하기 때문에 데이터 일부가 비어있거나 서로 상반된 판별기 결과를 내놨을 경우 등에서 한계를 보인다.

다중 뷰 학습에는 대표적으로 두 가지 접근 방법이 있다. 하나는 앙상블 학습 기반 방식으로, 각 뷰에 대한 판별기 학습이 독립적으로 이뤄지기 때문에 해당 뷰의 특성을 최대한 반영이 가능한 방식이다. 각 뷰 별로 서로 다른 기계학습 파라미터 등과 같이 특화된 방법을 적용할 수 있다. 예를 들어, 서로 다른 감독 학습을 뷰 별로 적용하거나, 감독 학습과 반감독 학습을 뷰 마다 적용할 수도 있다. 특정 뷰에서 발생 가능한 부정적 효과(negative effect)를 앙상블 과정에서 상쇄할 수도 있다.

또 하나는 앞서 말한 다중 커널 학습 기반 방식이 있다. 다중 커널 학습은 뷰 별로 기반 커널을 각각 구성한 후에 가중치를 부여하여 합하는 방식이다. 가중치도 데이터에서 convex-formulation을 통해 학습하며 특정 뷰에서 발생하는 오류를 다른 뷰를 통해 보정이 가능하다.

학습 데이터는 다음과 같이 구성된다. 입력 공간 X는 다중 뷰로 구성되어 있으며, 각 뷰마다 고유의 feature를 가진다.

$$S=\left\{(\mathbf{x}_1^i,\mathbf{x}_2^i,...,\mathbf{x}_p^i,\mathrm{y}^i)\right\}_{i=1}^{l},\quad \text{where } y_i\in Y \text{ and } \mathbf{x}_v^i\in \mathrm{X}_\mathrm{v} \text{ for } v\in\{1,2,...p\}. \tag{1}$$

각 뷰 별로 각각 feature selection/transformation g를 적용하여 뷰별 최적의 feature를 선별하는 작업이 가능하다:

$$S=\left\{\left(g_1(x_1^i),g_2(x_2^i),...,g_k(x_p^i),y^i\right)\right\}_{i=1}^{l}. \tag{2}$$

3.2 커널 머신 기반의 다중 뷰 학습

다중커널 학습 기법(MKL)과 커널 머신 앙상블(EKP)를 결합한 다중 뷰 학습 기법을 제안한다.

각 뷰 별로 다중 커널 학습 기법(MKL)을 적용하여 타겟 y를 예측한다. 이 때, Basis 커널은 뷰마다 동일하게 적용하는 것이 일반적이다.

$$f(x)=\sum_{i=1}^{l}\alpha_i^* K(x,x_i)+b^*,$$
$$K(x,x')=\sum_{m=1}^{M}d_m K_m(x,x'),\ d_m\geq 0,\ \sum d_m=1 \tag{3}$$

MKL을 위한 다양한 해법 중 본 논문에서는 simpleMKL[7]을 적용한다. 각 뷰별 MKL로 학습한 f값을 EKP (ensemble of kernel predictors)[8]를 통해 취합한다.

표 1. 동영상의 내용 기반 감정 예측을 위한 다중 뷰 구성

요인	뷰 및 raw feature 구성	window 모드	
		전체	3분할-병치
시각	색상 (Hue&Tone 130[4] - 4Hz, contrast 2, 장면전환요인 2)	670	3350
소리	MFCC (delta, acc 포함 21*3=63, 80Hz)	315	1575
	LPC (delta, acc 포함 13*3=39, 80Hz)	195	975

EKP는 L_q regularization의 일반화된 형태이다.

$$E_p^q=\sum_{v=1}^{p}\mu_k f_k,\ \ \boldsymbol{\mu}\in\Delta_\mathrm{q},\ \Delta_\mathrm{q}=\left\{\mu:\mu\geq 0,\ \sum_{v=1}^{p}\mu_v^q=1\right\}. \tag{4}$$

보통 q=1(lasso)과 q=2(ridge) 설정을 적용한다.

4. 실험 및 분석

4.1 데이터 및 다중 뷰 feature set 구성

다중 뷰 기반 감정 예측을 위한 데이터로 LIRIS-ACCEDE[2][9]를 선정하였다. ACCEDE는 160편의 영화에서 추출한 9800개의 클립(8~12초 길이)에 대해 크라우드소싱 결과 정리한 시청자 감정 반응의 세기(arousal)와 긍정·부정도(valence) 순위를 제공하는 데이터베이스이다. 4900개씩의 클립을 각각 학습 및 테스트 데이터로 지정하여 제공한다.

Valence과 arousal 중앙값(median)을 원점으로 하는 2차원의 Valence-arousal 감정 평면[10] 상에서 [11-12]와 같이 세 영역(1: calm, 2: positive excited, 3: negative excited)을 구분하는 문제로서 감정 예측 문제를 설정한다. 즉, $Y=\{1,2,3\}$의 3-레이블 분류 문제로 설정한다. 레이블 간 균형을 맞추기 위해 가장 사례가 적은 2번 레이블을 기준으로 1 및 3번 레이블에 해당하는 영화의 수를 조정하였다. 학습데이터의 경우 1083*3=3249, 테스트데이터의 경우 961*3=2883 개의 클립을 랜덤하게 선택하여 데이터를 구성하였다.

동영상 자극에 대한 직접적인 감정 반응을 예측하기 위해, 영상 및 소리 정보를 추출하였다(표 1). 윈도우 내 샘플의 대표값으로 평균, 표준편차, 중간값, IQR (inter-quartile range) 및 최대값의 다섯 가지 통계량을 적용한다. 3분할-병치 모드의 경우, 하나의 클립을 3등분한 크기의 window를 50% 중복 모드로 이동시켜 얻은 5개의 윈도우 각각에서 정보를 추출하고 순서대로 병치하여 하나의 벡터로 구성한다.

세 요인 각각을 별개의 뷰로 설정하고 다중 뷰 감독학습을 다음과 같이 적용한다. 각 뷰 별로 전체 feature군 및 information gain이 0인 feature를 제거한 선별군을 구성 후 다양한 분류 알고리즘을 이용하여 뷰별 감정 예측 모델을 구성한다.

4.2 MKL을 이용한 뷰별 학습

MKL 연구자들 사이의 표준 커널을 그대로 적용한다[5-6]. MKL 표준 커널은 10개의 가우시안 커널(폭(gamma)$\in\{2^{-3},\ 2^{-2},\ 2^{-1},\ 2^{0},\ ...\ 2^{6}\}$과 세 개의 다항 커널(차

2) http://liris-accede.ec-lyon.fr/database.php

표 2. 세 가지 모달리티 데이터를 whole, D&C, Delta 방식으로 처리하고 거기서 또 information gain (IG)을 기준으로 feature selection한 데이터의 분류 비교 결과 (accuracy). 38% 이상의 결과는 진하게 표시.

Classifier	Data		Vision	Mfcc	Lpc
Naive Bayes	whole	full	35.62	35.9	37.53
		IG	35.06	35.69	37.43
	D&C	full	36.05	34.56	37.92
		IG	35.62	36.66	37.84
	Delta	full	33.92	35.24	35.76
		IG	34.44	34.06	36.66
TAN (BN)	whole	full	42.8	**38.05**	**39.13**
		IG	41.03	37.57	35.62
	D&C	full	**42.75**	**39.45**	**39.56**
		IG	**42.84**	**38.05**	**39.13**
	Delta	full	40.31	33.37	34.86
		IG	40.31	33.37	34.86
Random Forest	whole	full	41.66	36.56	36.52
		IG	41.69	36.42	37.08
	D&C	full	**43.82**	**38.06**	**39.45**
		IG	42.14	37.91	37.63
	Delta	full	40.65	35.83	35.38
		IG	39.26	33.19	35.59
simple MKL	D&C	IG	44.89	40.13	41.2

수∈{1,2,3})으로 구성된다. 비교를 위해 세 가지 대표적인 분류 알고리즘으로서 나이브베이즈, 트리보완 나이브베이즈(TAN) 및 랜덤 포리스트를 추가로 적용한다.

4.4 EKP를 이용한 다중 뷰 통합 학습

MKL을 이용한 각 뷰별 예측 결과를 EKP로 취합하고, 기존의 단순 majority voting 등과 결과를 비교한 결과, 단순 majority voting의 정확도는 44.3%인 반면, 제안한 MKL+EKP 프로토콜의 정확도는 48.5%로 측정되었다.

5. 분석 및 논의

실험 결과, 다중 뷰(모달리티) 데이터에 대해 MKL을 사용한 결과가 일반 단일 커널 머신인 SVM 사용한 결과보다 좋았다. 이는 여러 커널을 이용하기에 데이터에 대한 표현력을 높였기 때문에, 딥 러닝과 유사한 representation 학습 효과가 있는 것으로 추정된다. 또한 모달리티 간 상충되는 상황도 자연스럽게 해결하므로 감정 기반 인지 컴퓨팅 연구에 있어서 좀 더 체계적이고 표준화된 방식으로서 제안할 수 있다. 체계적인 방식이지만 여전히 성능 면에서는 아직까지 연구가 많이 남았는데 이것은 감성 기반 인지컴퓨팅의 근본적인 문제인 피처 문제에 있다. 이 논문에서 다룬 ACCEDE set을 비롯한 이 분야의 많은 데이터들은 피처 추출 방식에 대해 아직 명확한 표준이 없고 많은 연구가 이뤄지고 있다. 우리가 한 것처럼 색깔이나 명암 대비 같은 low-level feature를 기반으로 한 mid-level feature를 사용한다고 하지만 여전히 유리천장을 넘지는 못하고 있다.

6. 결론

이 논문에서 우리는 최근 회자되고 있는 감성 기반 인지 컴퓨팅 연구에 있어서 좀 더 이론적으로 정립된 다중 커널 머신 기반의 다중 뷰 체계를 제안했고 이를 통해 다중 뷰 체계가 기존 단일 커널 머신인 SVM에 비해 낫다는 것을 실험적으로 증명했다. 이 연구를 통해 표준화되지 않고 진행중인 감성 기반 인지 컴퓨팅 분야에 새로운 대안을 제시함과 더불어 새로운 피처 발굴이라는 과제를 던지고 있다. 물체 인식과 딥 러닝을 통한 새로운 피처 발굴을 통해 향후 연구를 계속할 예정이며, MKL과 관련해서도 최근 비모수적 베이지안 기반 EKP-사전 확률을 MKL에 적용하는[13] 방식에 대해서도 추가 분석을 하고 있다.

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] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: A survey,” *Multimed. Syst.*, vol. 16, pp. 345–379, 2010.

[2] M. Barthet, G. Fazekas, and M. Sandler, “Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models,” in *9th International Symposium on Computer Music Modelling and Retrieval*, pp. 492–507, 2012.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: audio, visual, and spontaneous expressions.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[4] 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 (CogSci 2013)*, 2013.

[5] 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, Apr. 2012.

[6] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, “Multiple kernel learning, conic duality, and the SMO algorithm,” in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, p. 6, 2004.

[7] A. Rakotomamonjy and F. Bach, “SimpleMKL,” J. Mach. Learn Res., vol. 9, pp. 2491–2521, 2008.

[8] C. Cortes, M. Mohri, and A. Rostamizadeh, “Ensembles of kernel predictors,” in *Proc. Uncertainty Artif. Intell. (UAI)*, pp. 145–152, 2011.

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

[10] J. A. Russell, “A circumplex model of affect.,” *J. Pers. Soc. Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[11] M. Soleymani, G. Chanel, and T. Pun, “A Bayesian Framework for Video Affective Representation,” in *International Conference on Affective Computing and Intelligent Interaction*, pp. 1–7, 2009.

[12] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, “A supervised approach to movie emotion tracking,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference* on, 2011.

[13] Q. Mao, I. W. Tsang, S. Gao, and L. Wang, “Generalized Multiple Kernel Learning With Data-Dependent Priors.,” *IEEE Trans. neural networks Learn. Syst.*, pp. 1–15, 2014.