텐서 기반 데이터 생성 모델을 이용한 생체데이터 분류

윤동우*, 박혜영**
*경북대학교 컴퓨터과학과, **경북대학교 전자전기컴퓨터학부
e-mail: *dongwoo0926@hamail.net, **hypark@knu.ac.kr

Bio-Data Classification Using Tensor-based Data Generation Model

Dongwoo Yoon*, Hyeyoung Park**
*Dept of Computer Science, Kyungpook National University
**School of EECS, Kyungpook National University

요 약

생체데이터란 인간개체로부터 얻을 수 있는 고유의 생체신호를 통틀어 일컫는 것이다. 본 연구에서는 생체데이터를 위한 팩터 분석 모델에 텐서 개념을 적용하여, 2차 텐서로 표현된 데이터를 위한 생성 모델을 제안한다. 이 모델을 바탕으로 데이터로부터 분류에 핵심이 되는 정보를 안정적으로 추출하여 유사도 함수를 만들고 분류를 수행하는 방법을 제안한다. 실험을 통해 제안하는 방법이 기존의 벡터 형태의 데이터에 대한 생성 모델을 사용한 경우보다 우수한 성능을 가짐을 확인할 수 있었다.

1. 서론

생체데이터가 가지는 특성 중 개개인의 클래스로부터 얻을 수 있는 데이터의 수가 제한적이고 차원이 높은 점은 시스템의 성능을 떨어뜨리는 주요한 원인이 된다. 이러한 특성을 고려하여 분류에 핵심적인 정보를 안정적으로 추출할 수 있는 데이터 생성 모델을 개발하는 연구가 수행되어 왔다[1,3]. 한편, 최근 영상 데이터와 같은 고차원 데이터에 대한 새로운 접근 방법으로 벡터의 개념을 확장한 텐서 개념을 적용한 연구가 활발히 수행되고 있다. 대표적으로 텐서 분석[4], 텐서 PCA[2], 텐서 LDA[2] 등이 있다.

본 연구에서는 데이터 생성 모델에 텐서 개념을 적용하여 2차 텐서로 표현된 생체 데이터의 생성 모델을 제안한다. 이를 바탕으로 분류에 핵심적인 정보를 추출하여 유사도 함수를 만들고, 이를 이용하여 생체 영상데이터의 분류를 효과적으로 수행하는 분류기를 개발하고자 한다.

2. 텐서 기반 데이터 생성 모델

본 논문에서는 [1,3]에서 개발된 환경요인과 클래스 요인을 가지는 데이터 생성 모델을 텐서 형태로 확장한다. 텐서는 벡터를 일반화한 기하학적인 물리량을 말한다. 1차 텐서는 벡터, 2차 텐서는 행렬이 되고 그 이상의 차원을 n차 텐서라고 부른다. 본 논문에서는 영상데이터가 가진 지역적 특성을 표현하기 위해 2차 텐서를 사용한다.

[3]에서 개발된 데이터 생성 모델을 2차 텐서 모델로 확장하여 표현하면 다음 식과 같다.

$$\boldsymbol{X_k} = \boldsymbol{U^T \Xi_k V} + \boldsymbol{\Delta}$$

여기서 $\boldsymbol{X}_k$는 클래스 C_k에 속하는 하나의 영상 데이터를 2차원 입력 형태 그대로 두고 2차 텐서로 본 것이며, 이에 맞게 클래스 요인 $\boldsymbol{\Xi}_k$와 환경요인 $\boldsymbol{\Delta}$도 2차원 텐서로 주어진다. 클래스 요인 $\boldsymbol{\Xi}_k$는 클래스에 의존하여 고유하게 정해지는 값이며, 환경요인 $\boldsymbol{\Delta}$는 클래스에 의존하지 않고, 데이터를 획득할 때 가해지는 다양한 노이즈를 결정하는 것으로, 따라서 모든 클래스에 공통적인 분포 특성을 가진다고 가정할 수 있다.

이어서 같은 클래스로 부터 얻어지는 두개의 데이터 텐서의 차로 만들어지는 새로운 랜덤 변수 $\boldsymbol{Y}$를 정의하면, 이는 다음 식과 같이 나타낼 수 있다.

$$\boldsymbol{Y} = \boldsymbol{X_k} - \boldsymbol{X_k'} = \boldsymbol{\Delta} - \boldsymbol{\Delta'}$$

이 때 같은 클래스로부터 얻어진 데이터의 차를 구하게 되면 같은 클래스의 경우 하나의 고유한 값으로 정해진다고 가정할 수 있는 클래스 요인이 상쇄되어 환경 요인만 남게 됨을 알 수 있다. 이렇게 얻어진 새로운 데이터 집합 $\boldsymbol{Y}$를 이용하여 환경요인 $\boldsymbol{\Delta}$의 분포를 추정함으로써 유사도 함수를 얻어낼 수 있다.

3. 유사도 함수

본 논문에서는 [3]에서와 마찬가지로 $\boldsymbol{Y}$의 확률 밀도 함수 $p(\boldsymbol{Y})$를 추정하여 유사도 함수를 정의한다. 추정된 $p(\boldsymbol{Y})$는 같은 클래스에 속하는 데이터 쌍의 차에 대한 분포를 나타내므로, $p(\boldsymbol{X}-\boldsymbol{X'})$의 값이 클수록 $\boldsymbol{X}$ 와 $\boldsymbol{X'}$가 같은 클래스에 속할 확률이 높다고 볼 수 있다. 따라서 두 데이터의 유사도 함수 $S(\boldsymbol{X},\boldsymbol{X'})$는 $p(\boldsymbol{X}-\boldsymbol{X'})$에 비례하는 값

으로 정의 할 수 있다.

[3]에서는 환경 요인 Δ가 정규분포를 가지는 경우의 최우도 추정에 의한 유사도 함수 정의 방법과, SVM을 이용하여 유사도 함수를 직접 학습하는 방법을 제안하였다. 여기서는 간단히 Δ가 정규분포를 따른다고 가정한 경우의 유사도 함수를 제안하다. Δ가 정규분포를 따르는 경우, 그 차로 이루어진 확률변수 Y도 정규분포를 따른다. 그러나 이 경우에는 [3]과 달리 2차 텐서 형태의 확률 변수 Y에 대한 정규분포를 사용하므로 다음과 같이 행렬 정규분포(matrix normal distribution)로 나타낼 수 있다.

$$G(\boldsymbol{Y}|\boldsymbol{M},\boldsymbol{\Omega},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Omega}|^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}tr[\boldsymbol{\Omega}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^T\Sigma^{-1}(\boldsymbol{Y}-\boldsymbol{M})]\right)$$

이때 M는 Y의 평균 행렬이고, Ω는 Y의 각 행을 하나의 데이터로 보았을 때의 공분산행렬, Σ는 Y의 각 열을 하나의 데이터로 보았을 때의 공분산 행렬이다. 여기서 우리가 추정해 주어야 하는 파라미터는 M, Ω, Σ이다.

평균 행렬 M은 단순히 Y의 데이터 집합으로부터 계산되는 표본평균으로 추정할 수 있다. 공분산 행렬에 대해서는 먼저 Y의 각 열 y_i을 하나의 데이터로 보았을 때의 표본 분산과 각 행 y^i을 하나의 데이터로 보았을 때의 표본 분산은 다음과 같이 계산할 수 있다.

$$\boldsymbol{\Sigma_y} = cov(\boldsymbol{y_i},\boldsymbol{y_j}) = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{Y_k}-\boldsymbol{M})(\boldsymbol{Y_k}-\boldsymbol{M})^T$$

$$\boldsymbol{\Sigma^y} = cov(\boldsymbol{y^i},\boldsymbol{y^j}) = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{Y_k}-\boldsymbol{M})^T(\boldsymbol{Y_k}-\boldsymbol{M})$$

본 논문에서는 이를 그대로 사용하는 대신, [2]에서 제안한 방법을 이용하여, Y의에 대한 직교변환 행렬 U_y 와 V_y를 각각 구하여 이에 의해 변환된 값 $Z = U_y^T Y V_y$를 얻어 사용한다. 이렇게 하면 Z의 공분산 행렬 Ω_Z와 Σ_Z는 각각 공분산 행렬 Σ_y와 Σ^y의 고유치를 대각 원소로 가지는 대각 행렬로 얻어지므로, 역행렬의 계산이 간단해지고, 또한 추정 파라미터의 수가 줄어들어 데이터에 대한 오버피팅을 피하는 효과를 얻는다. 이렇게 추정된 파라미터를 이용하면 두 데이터 X와 X'의 유사도 값은 다음과 같이 정의될 수 있다.

$$S(\boldsymbol{X},\boldsymbol{X'}) = G(\boldsymbol{U_y^T(X-X')V_y}|\boldsymbol{M},\boldsymbol{\Omega_Z},\boldsymbol{\Sigma_Z})$$

정의된 유사도 함수를 이용하여 분류를 수행하기 위해, 본 논문에서는 [1]에서 사용한 방법과 유사하게 영상 데이터를 그대로 사용하지 않고 먼저 텐서PCA를 사용하여 데이터 행렬(텐서)의 크기를 줄인 후에 그에 대하여 제안하는 방법의 유사도 함수 계산법을 적용한다. 유사도 함수를 이용한 분류를 위해서는 K-근접이웃 분류기를 사용한다.

4.실험 및 결과

제안한 방법의 성능을 평가하기 위해 실제 생체인식 데이터에 대해 기존의 PCA 및 텐서PCA 방법, 그리고 [1]에서 사용된 방법 (DataPair 방법)과의 비교 실험 하였다. 실험데이터는 두 가지를 사용하였다. 첫 번째 FERET 데이터[5]는 50명의 서로 다른 사람으로부터 각각 서로 다른 각도의 9개의 데이터로 이루어지며, 영상의 크기는 70 x 50이다. 이를 이용하여 사람과 포즈를 인식하였는데, 사람 인식을 위해서는 개인당 3개의 데이터를 학습데이터로 사용하고 나머지 데이터로 테스트하였다. 포즈인식은 25명에 대한 데이터로 학습하고 나머지 데이터로 테스트하였다. 두 번째 PICS 데이터[6]는 69명의 서로 다른 사람의 4가지 표정에 대한 영상으로, 하나의 데이터는 90 x 80의 이미지이다. 이 데이터로 표정인식을 수행하였는데 20명의 대한 데이터(80개)로 학습하고 나머지 49명에 대한 데이터로 테스트하였다.

표 1에 실험 결과를 나타내었다. 괄호 안의 값은 최적의 인식률을 보인 차원을 나타낸다. 전체적으로 텐서를 사용함으로써 기존의 방법에 비해 성능 향상을 효과를 가져옴을 확인할 수 있다. 특히 분류 대상이 되는 클래스에 나타난 특징이 지역적인 특성을 가지는 포즈인식과 표정인식의 경우 보다 성능 향상의 효과가 두드러짐을 볼 수 있다.

표 1 인식결과 (괄호 안은 축소된 차원)

	PCA	DataPair	텐서PCA	Proposed
FERET person	97.0(117)	96.7(14)	99.0(14x3)	99.7(15x1)
FERET pose	36.4(65)	40.0(27)	36.4(21x4)	48.0(7x11)
PICS expression	35.7(65)	37.6(79)	37.8(15x1)	56.6(25x1)

5.결론

최근 영상데이터의 분류 및 특징추출 분야에서 관심을 모으고 있는 텐서 개념과 데이터 생성 모델을 결합하여 새로운 생체 데이터 분류 방법을 제안하였다. 실험을 통해 제안하는 방법의 가능성을 확인하였으며, 이후 보다 정교한 분포 추정 및 유사도 함수 개발에 대한 연구가 수행될 것이다.

참고문헌

[1] 조민국, 박혜영, "변형된 팩터 분석 모델을 이용한 생체데이타 분류 시스템" 정보과학회논문지 : 소프트웨어 및 응용 34(7), 667-680, 2007.

[2] D. Cai, X. He and J. Han, "Subspace Learning Based on Tensor Analysis", UIUCDCS-R-2005-2572 2005.

[3] K. Lee and H. Park, "A New Similarity Measure Based on Intraclass Statistics for Biometric Systems" Etri Journal, 25(5), 401-406, 2003.

[4] M. A. O. Vasilescu and D Terzopoulos "Multilinear Image Analysis for Facial Recognition" In proc. ICPR, 511-514, 2002.

[5] http://www.itl.nist.gov/iad/humanid/feret/

[6] http://pics.psych.stir.ac.uk/