다중 인스턴스 학습 기반 사용자 프로파일 식별

송현제○, 김아영, 박성배
경북대학교, 컴퓨터학부
{hjsong, aykim, sbpark}@sejong.knu.ac.kr

Discriminating User Attributes in Social Text based on Multi-Instance Learning

Hyun-Je Song○, A-Yeong Kim, Seong-Bae Park
Kyungpook National University, Department of Computer Science and Engineering

요 약

본 논문에서는 소셜 네트워크 서비스에서 사용자가 작성한 텍스트로부터 그 사용자 프로파일 식별하는 문제를 다룬다. 프로파일 식별 관련 기존 연구에서는 개별 텍스트를 하나의 학습 단위로 간주하고 이를 기반으로 학습 모델을 구축한다. 프로파일을 식별하고자 하는 사용자의 텍스트들이 주어지면 각 텍스트마다 프로파일을 식별하고, 식별된 결과들을 합쳐 최종 프로파일로 선택한다. 하지만 SNS 특성상 프로파일을 식별하는 데에 영향을 끼치지 않는 텍스트들이 다수 존재하며, 기존 연구들은 이 텍스트들을 특별한 처리없이 학습 및 테스트에 사용함으로 인해 프로파일 식별 성능이 저하되는 문제점이 있다. 본 논문에서는 다중 인스턴스 학습(Multi-Instance Learning)을 기반으로 사용자 프로파일을 식별한다. 제안한 방법은 사용자가 작성한 텍스트 전체, 즉 텍스트 집합을 학습 단위로 간주하고 다중 인스턴스 학습 문제로 변환하여 프로파일을 식별한다. 다중 인스턴스 학습을 사용함으로써 프로파일 식별에 유의미한 텍스트들만이 고려되고 그 결과 프로파일 식별에 영향을 끼치지 않는 텍스트로부터의 성능 하락을 최소화할 수 있다. 실험을 통해 제안한 방법이 기존 학습 방법보다 성별, 나이, 결혼/연애 상태를 식별함에 있어서 더 좋은 성능을 보인다.

주제어: 다중 인스턴스 학습, Multi-Instance Learning, 사용자 프로파일 식별, 소셜 네트워크 텍스트 분석

1. 서론

Facebook, Twitter, me2day 등 다양한 소셜 네트워크 서비스(Social Network Service, 이하 SNS)들이 등장하면서 사용자들이 위 서비스들을 이용하여 다양한 주제의 텍스트들을 작성한다. 위 텍스트에는 작성한 사용자의 전기적인 속성(Biographic attribute)에 대한 정보 즉, 성별, 나이, 관심 분야 등을 명시적으로 또는 묵시적으로 포함하고 있다. 이들을 분석하여 사용자 맞춤 서비스에 사용하고자 하는 연구들이 많이 진행되고 있다.

사용자 맞춤 서비스들은 일반적으로 사용자가 직접 입력한 프로파일에 기반을 두고 있다. 하지만, 연애유무, 정치적 성향 등의 속성과 같이 사용자가 특별히 입력하지 않는 경우가 존재할 뿐만 아니라 특정 서비스에서는 사생활 보호를 위해 사용자 프로파일을 직접적으로 수집하지 않는다. 위 상황에서 사용자 맞춤 서비스는 사용자가 생성한 컨텐츠 즉, 사용자와 관련된 문서로부터 그 사용자의 프로파일을 식별하는 것이 선행되어야 한다.

본 논문에서는 사용자가 SNS를 사용하여 작성한 텍스트로부터 그 사용자의 프로파일을 식별하고자 한다. SNS 외 블로그, 전화 대화문서, 영화 리뷰 등으로부터 사용자 프로파일을 식별하고자 하는 연구들은 많이 진행되어 왔다[1-4]. 위 연구들은 사용자 프로파일 식별을 위한 자질들을 정의하는 데에 중점을 두었다. 사용자가 작성한 텍스트를 잘 반영할 수 있는 즉, 주어진 도메인에 따라 도메인의 특징을 잘 반영할 수 있는 부분을 자질로 정의하였다. 이 후, 자질로 표현된 텍스트를 미리 정의된 사용자 프로파일의 속성과 그에 해당하는 클래스로 분류하여 프로파일을 식별한다. 최근 SNS에서부터 프로파일 식별을 위한 연구 또한 활발히 진행되고 있다. 위 연구들은 앞선 기존 연구들과 동일하게 SNS 텍스트 특징과 소셜 네트워크를 반영할 수 있는 자질을 정의하는 데에 중점을 맞추고 있다.

하지만, 기존의 다른 문서들과 달리 SNS 텍스트들은 짧은 길이와 더불어 다양한 주제를 다루고 있다. 이에, 사용자들이 작성한 텍스트들 중에는 프로파일을 식별하는 데에 영향을 끼치지 않거나 잘못 식별하게 만드는 텍스트들이 존재한다. 하지만, 기존 연구들은 이들 모두를 사용하거나 간단한 방법으로 필터링하여 프로파일 식별 모델을 구축한다. 위 방법들은 다음과 같은 문제점들이 있다. 첫 번째로 주어진 텍스트들 모두 학습에 사용할 경우, 주어진 프로파일과는 관련 없거나 잘못 태깅된 학습 데이터로 인해 프로파일 식별 성능이 떨어질 수 있다. 두 번째로 주어진 텍스트가 프로파일을 식별하는 데 도움이 되는지, 아닌지를 파악하는 것은 쉽지 않으며, 앞서 잘못 식별된 결과가 다음 단계로 전파될 수 있다. 또한, 이를 위한 학습 데이터를 구축하는 것은 많은 비용이 든다.

본 논문에서는 앞선 문제점들을 해결하는 사용자 프로파일 식별 방법을 제안한다. 제안한 방법은 다중 인스턴스 학습(Multi-Instance Learning)에 기반하여 사용자 프로파일을 식별한다. 다중 인스턴스 학습이란 인스턴스 하나를 학습 단위로 간주하는 개별 인스턴스 학습

(Single-Instance Learning)과는 달리 인스턴스 집합을 하나의 학습 단위로 수행하는 학습 방법이다. 학습시 인스턴스 집합에서 "도움이 되는" 인스턴스만을 찾아 학습을 수행하므로 관련 없는 또는 잘못 태깅된 인스턴스들로부터의 성능 하락을 최소화할 수 있다. 위 학습 방법은 사용자들이 작성한 텍스트들을 개별로 태깅하는 것이 아닌 텍스트 집합에 태깅을 수행하기 때문에 개별 결과를 합치는 모듈이 필요하지 않으며 학습 데이터를 쉽게 구축할 수 있는 장점을 가진다. 실험을 통해 제안한 방법이 기존의 방법보다 프로파일 식별에 도움이 됨을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자 프로파일 식별과 관련된 기존 연구들을 살펴본다. 3장에서는 다중 인스턴스 학습에 대한 설명과 다중 인스턴스 학습을 고려하는 여러 알고리즘 중 Multi-Instance Support Vector Machine에 대해 간략히 다룬다. 4장에서는 다중 인스턴스 학습 기반 사용자 프로파일 식별에 관해 설명한다. 5장에서는 실험 및 결과를 분석하고 마지막 6장에서는 결론 및 향후 연구를 다룬다.

2. 관련연구

블로그, 이메일, 방문한 웹 페이지 및 다른 사용자와의 대화문서 등 사용자가 생성 및 작성한 문서로부터 사용자들의 성별, 나이 등을 식별하고자 하는 연구는 오래 전부터 많이 진행되어 왔다. Boulis and Ostendorf(2005)는 전화로 주고받은 대화 문서에서 사용자의 성별을 식별하고자 하였다[1]. 이들은 전화 대화 문서에서 성별 간의 단어 사용이 다름을 파악하고, 이를 반영하기 위해 n-gram모델로 대화를 표현하고 기계학습을 사용하여 성별을 식별하였다. Garera and Yarowsky(2009)는 Boulis and Ostendorf에서 사용한 전화 대화문서와 이메일에서 사용자의 성별, 나이와 모국어로 말하는지(Native speaker) 여부를 식별하였다[2]. 이들은 사회 언어학(Sociolinguistic) 자질과 담화(Discourse) 자질을 정의하고 이로부터 사용자의 속성을 식별하였다. Otterbacher(2010)과 Hemphill and Otterbacher(2012)는 영화 리뷰 데이터에서 성별을 추정하고자 하였다[3,4]. 이들은 IMDb에서 성별에 따라 언어 사용에 있어 다른 점을 반영하고자 하였다. 즉, 여성이 남성에 비해 사교적인 스타일(Social style)로 리뷰를 쓰는 반면 남성은 여성에 비해 제 3자가 다른 사람에게 알리고자 하는 스타일(Broadcast style)로 리뷰를 쓴다는 것을 발견하였다. 이를 위해 단어와 문장의 Complexity[5]와 대명사 사용 빈도, 관용구(Hedging phrase)등을 자질로 사용하였다.

최근 SNS에서 사용자 프로파일을 식별하고자 하는 연구들이 많이 진행되고 있다. 기존의 연구되었던 방법에 추가로 SNS에서 얻을 수 있는 정보 및 특징을 반영하는 연구가 주를 이뤄왔다. Rao et al.(2010)는 Twitter에서 성별, 나이, 지역출신, 정치적 성향을 식별하고자 하였다[6]. 이들을 딸림벗(Follwer), 따름벗(Follwing)을 통한 네트워크 관계 및 트윗(Tweet), 리트윗(Retweet) 등으로 얻어지는 사용자간의 커뮤니케이션 횟수에 기존 연구의 사회 언어학 자질을 합쳐 사용자 속성을 식별하였다. Pennacchiotti and Popescu(2011)는 Twitter에서 기계학습 기반으로 사용자를 분류하였다[7]. Bio 필드에서 추출하는 프로파일 자질과 트윗 작성 비율 및 리트윗 비율과 작성한 트윗이 URL을 포함하는지 여부를 사용하였다. 뿐만 아니라 토픽 모델을 사용하여 트윗을 분석한 자질과 소셜 네트워크 관계를 자질로 사용하여 사용자의 정치적 성향, 스타벅스 팬인지를 식별하였다. Burger et al.(2011)는 Twitter에서 성별을 식별하고자 하였다[8]. 이들은 다량의 Twitter 사용자와 트윗을 수집하였으며, 언어와 독립적인 모델을 구축하기 위해 Bio 필드, 닉네임과 작성한 트윗으로부터 단어 레벨 n-gram 모델과 음절 레벨의 n-gram 모델을 사용하여 자질을 추출하고 이를 Winnow 알고리즘을 사용하여 성별을 식별하였다.

앞선 연구들은 사용자가 작성한 정보들을 잘 반영할 수 있는 자질에 대해 주로 연구하였다. 하지만, 사용자들이 작성한 정보들 중에는 프로파일 식별에 직접적으로 또는 간접적으로 영향을 끼치는 것이 있는 반면 그렇지 않는 것들도 존재한다. 이들 모두를 프로파일 식별에 사용하면 영향을 끼치지 않는 다수의 텍스트들로 인해 성능 하락이 발생할 수 있다. 본 논문에서는 기존 학습 방법이 아닌 새로운 학습 방법으로 사용자의 프로파일을 식별하고자 한다.

3. 다중 인스턴스 학습과 Multi-Instance Support Vector Machines

3.1 다중 인스턴스 학습

다중 인스턴스 학습은 Dietterich et al.(1997)이 *Drug Activity Prediction* (DAP)을 풀기 위해 제안된 학습 방법이다[9]. DAP란 새로운 약이 개발되었을 때, 이 약이 적합한지, 적합하지 않는지를 예측하는 문제이다. 세부적으로 약은 여러 분자들로 구성되어 있으며 이들 중 하나라도 주어진 단백질에 반응을 하면 약이 적합하다고 하고, 모두가 그 단백질에 반응을 하지 않으면 약으로 적합하지 않는다고 한다.

Dietterich et al.(1997)은 기존 학습을 통해 DAP 문제를 해결할 경우 학습시 다수의 긍정 오류(False Positive)가 포함하게 되어 성능이 떨어지는 것을 보였다. 즉, 특정 단백질에 반응을 한 약이 주어질 때, 기존의 학습은 이들을 구성하고 있는 분자들이 모두 특정 단백질에 적합한 것으로 여긴다. 이 후, 이들로부터 학습을 수행하면 실질적으로 단백질과 관련없는 다수의 분자들로 인해 잘못 예측을 하게 된다.

이를 해결하기 위해 Dietterich et al.(1997)은 다중 인스턴스 학습을 제안하였다. 인스턴스 하나를 학습 단위로 사용하는 개별 인스턴스 학습과는 달리 다중 인스턴스 학습에서는 인스턴스의 집합인 Bag을 학습 단위로 간주한다. 따라서 개별 인스턴스 학습에서는 인스턴스 하나에 레이블을 다는 반면, 다중 인스턴스 학습에서는 Bag에 대해 레이블을 달게 된다. 다중 인스턴스 학습은 개별 인스턴스 학습과 학습 단위를 제외하고 유사하지만 다음 제약사항을 추가로 가진다. 이진 분류를 수행함에 있어 Bag이 긍정(Positive)으로 레이블이 되면 적어도 Bag 안에는 하나 이상의 긍정 인스턴스를 포함하고 있음을, Bag이 부정(Negative)로 레이블링이 되면 Bag 안에 모든 인스턴스들이 부정임을 가정한다.

Dietterich et al.는 DAP 문제를 해결하기 위해 축-평행 사각형(axis-parallel rectangle) 알고리즘을 제안하였다[9]. 위 알고리즘은 자질 공간에서 긍정 Bag을 다 포함하는 사각형을 먼저 찾고, 이 사각형에 부정 Bag이 포함되지 않도록 점차적으로 줄여나간다. 실험을 통해 Musk 데이터[10]에서 가장 좋은 성능을 보였다. 이후, 이미지 분류[11] 및 내용기반 이미지 검색[12] 등에서 다중 인스턴스 학습이 많이 사용되었으며, Diverse Density (DD)[11], EM Diverse Density (EM-DD)[13], Multi-Instance Neural Network (MI-NN)[14] 등의 많은 알고리즘들이 다중 인스턴스 학습을 고려하기 위해 제안되었다. 본 논문에서는 분류 문제를 해결하는데 있어 강력하고 널리 쓰이는 모델인 Support Vector Machines에 다중 인스턴스 학습을 고려한 Multi-Instance Support Vector Machines을 사용하였다. 이에 대해서는 3.2절에서 자세히 다룬다.

3.2 Multi-Instance Support Vector Machines

Multi-Instance Support Vector Machines(MI-SVMs)은 기존의 SVMs에서 다중 인스턴스 학습을 고려하도록 Andrews et al.(2002)가 제안한 알고리즘이다[15]. 기존의 SVMs의 경우 모든 인스턴스를 고려하여 마진(Margin)이 최대화되는 결정 경계를 찾는다. MI-SVMs에서는 인스턴스 간의 마진을 최대화 하는 것이 아니라 Bag 간의 마진을 최대화하는 결정 경계를 찾는다.

Bag 간의 마진은 다중 인스턴스 학습에서의 가정을 고려하여 결정된다. 긍정 Bag의 관점에서 마진은 Bag 안에 인스턴스들 중에 "가장 긍정"인 인스턴스로 결정된다. 반면 부정 Bag의 관점에서 마진은 Bag 안에 인스턴스들 중 긍정에 "가장 가까운" 인스턴스로 결정된다. 그림 1는 Bag 간의 마진을 고려할 때, 어떻게 결정 경계가 그려지는 것을 보여주고 있다.

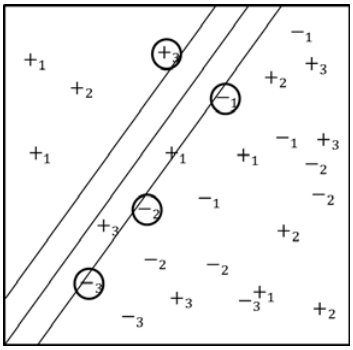

그림 1 MI-SVMs에서의 결정 경계

그림에서 +는 긍정 인스턴스를, -는 부정 인스턴스를 의미한다. 인스턴스에 부착된 숫자는 Bag을 나타낸다. 즉, 같은 숫자로 표시된 인스턴스들은 같은 Bag에 속해 있음을 의미한다. 부정 Bag에서는 긍정에 "가장 가까운" 인스턴스들이 Bag의 마진을 결정하는데 사용되므로 그림에서 부정 Bag 1,2,3에 속해있는 여러 인스턴스들 중 동그라미로 표시된 인스턴스들이 결정 경계를 선택하는데 사용되었다. 긍정 Bag에서는 적어도 하나의 긍정 인스턴스가 포함되어야 하므로 왼쪽 상단부의 긍정 Bag 1의 인스턴스가 아닌 긍정 Bag 3의 인스턴스 중 동그라미로 표시된 인스턴스가 결정 경계를 선택하는데 사용되었다. 앞서 살펴본 Bag 간의 마진을 고려한 soft-margin SVM은 아래의 수식과 같이 표현될 수 있다.

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_I \xi_I$$
$$\text{s.t.} \quad \forall I: \; Y_I \max_{i\in I}(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_I, \; \xi_I \geq 0$$

위 수식에서 X는 인스턴스를 나타내고 Y는 Bag을 나타낸다. Y_I는 I번째 Bag의 레이블을 나타내며, X_i는 I번째 Bag에 존재하는 인스턴스를 나타낸다. 최적화(optimization)를 위해 위 수식에서 제약조건 부분의 max 연산자를 각 클래스별로 다시 작성하면, 부정 Bag의 경우 Bag에 존재하는 모든 인스턴스들이 마진을 결정하는데 고려되어야 한다. 하지만, 긍정 Bag의 경우, Bag에 존재하는 인스턴스들 중에 마진을 결정하는데 도움이 되는 인스턴스 즉, "근거 인스턴스(witness instance)"만이 적어도 하나 이상 고려되어야 한다. 이를 위해 근거 인스턴스를 반환하는 선택 변수(selector variable) s(·)을 도입한다. 변경된 제약사항으로 위 수식을 다시 작성하면 아래 수식과 같이 표현된다.

$$\min_s \min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_I \xi_I$$
$$\text{s.t.} \quad \forall I: \; Y_I = -1 \;\wedge\; -\langle \mathbf{w}, \mathbf{x}_i\rangle - b \geq 1 - \xi_I, \; \forall i \in I,$$
$$\text{or} \;\; Y_I = 1 \;\wedge\; \langle \mathbf{w}, \mathbf{x}_{s(I)}\rangle + b \geq 1 - \xi_I, \;\; \text{and} \; \xi_I \geq 0$$

위 수식은 혼합 정수 계획법(mixed integer programming)으로 변환하여 최적화할 수 있다. 이와 관련한 자세한 내용은 [15]의 5장에서 설명하고 있다.

4. 다중 인스턴스 학습 기반 사용자 프로파일 식별

사용자 프로파일 속성과 클래스가 정의되어 있다면 사용자가 작성한 텍스트로부터 그 사용자의 프로파일 식별은 주어진 텍스트들을 앞서 정의된 속성과 클래스로 분류하는 문제로 볼 수 있다. 개별 인스턴스 학습 기반 관점에서 프로파일 식별 문제를 살펴보면, 먼저 사용자가 작성한 데이터 $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$가 주어진다고 가정한다. 이 때, $y_i \in \{-1, +1\}$[1]이며, $y_i = +1$은 텍스트 x_i가 클래스 +1에 속함을 의미하고, $y_i = -1$은 x_i가 클래스 -1에 속함을 의미한다. 프로파일 식별 모델은 데이터 D 로부터 함수 $f: X \rightarrow Y$를 추정함으로써 구축된다. 이렇게 구축된 프로파일 식별 모델은 텍스트 전체가 아닌 텍스트 하나에 대해 클래스를 추정하게 된다.

1) 두 개의 클래스만 있다고 가정한다.

따라서 사용자가 작성한 새로운 텍스트들이 주어지면, 함수 f를 통해 텍스트마다 클래스를 추정하고, 추정된 클래스 중 가장 많이 분류된(Majority Voting) 클래스로 사용자 프로파일을 식별한다.

개별 인스턴스 학습에서는 사용자가 작성한 텍스트들 모두 동일한 중요도를 가지고 함수 f를 추정하는데 사용된다. 하지만, 사용자가 작성한 텍스트에는 프로파일 추정 함에 있어 영향을 끼치지 않거나 성능을 하락시키는 것들이 존재한다. 이들 모두를 데이터로 사용하여 $f: X \rightarrow Y$를 추정하면 잘못된 모델 파라미터가 선택되고 프로파일 식별 모델의 성능을 하락시킬 수 있다.

본 논문에서는 사용자 프로파일 식별을 위해 사용자가 작성한 데이터는 아래와 같은 형식으로 주어진다고 가정한다.

$$D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$$

$X_i = \{x_{i1}, x_{i2}, .., x_{in}\}$는 i번째 사용자가 작성한 텍스트 집합을, $Y_i \in \{-1, +1\}$는 i번째 사용자의 클래스를 나타낸다. 제안한 모델은 사용자가 작성한 모든 텍스트를 하나의 학습 단위로 사용하므로 추정된 결과를 결합하는 모듈 없이 바로 사용자 프로파일을 식별할 수 있다. 또한, 학습 시 텍스트 집합에서 식별에 "도움이 되는" 텍스트만을 찾아 학습을 수행하므로 관련 없는 또는 잘못 태깅된 텍스트들로부터의 성능 하락을 최소화한다.

함수 f를 추정하기에 앞서 SNS에서 작성된 텍스트를 자질로 표현할 필요가 있다. SNS 텍스트는 기존 문서와 달리 길이가 짧고, 이모티콘, 알파벳 반복 등이 자주 사용된다. 본 논문에서는 이러한 특징을 반영하기 위한 자질을 정의하였으며 표 2는 본 논문에서 사용한 자질을 나열한 것이다.

표 2 프로파일 식별을 위해 사용한 자질들

자질	설명
Bag-of-Word	명사, 미등록어, 기호
이모티콘 유무	ㅡㅡ, ㅠㅠ, ㅜㅜ, T_T, ㅇㅁㅇ, ㅇㅅㅇ, ^^, ^_^, :), :(, ♡, ... 등이 포함되면 1, 아니면 0
같은 문자, 기호 반복 유무	[ㅋ|ㅎ|ㄷ|.|!|;|~]+ 으로 매치되면 1, 아니면 0

문서 분류에서 널리 사용되는 Bag-of-Word 자질에 소셜 텍스트의 특성인 이모티콘과 알파벳 반복 유무를 사용한다. Bag-of-Word를 추출하기 위해 본 논문에서는 주어진 텍스트를 형태소 분석한 후, 명사, 미등록어 및 기호만을 사용하였다.

데이터 D 로부터 함수 $f: X \rightarrow Y$를 추정하여 프로파일 식별 모델을 구축하며, 이는 3.2에서 살펴본 MI-SVMs에 바로 적용하여 추정할 수 있다.

5. 실험

5.1 실험 데이터

실험을 위해 본 논문에서는 Facebook, Twitter, me2day에서 사용자와 그 사용자의 프로파일을 수집하였다. 사용자의 프로파일은 Facebook을 통해 수집한 사용자의 경우 명시적으로 얻을 수 있는 정보를 사용하였으며, Twitter나 me2day의 경우 사용자가 작성한 텍스트와 약력 등으로부터 수작업으로 구축하였다. 표 3에서는 실험에 사용한 데이터의 통계 정보를 보여준다.

표 3 실험 데이터에 대한 간단한 통계(statistics)

정보	값
사용자	270
작성한 텍스트	3,968
사용자가 작성한 텍스트의 평균 개수	14.70

총 270명의 사용자를 수집하였으며, 수집한 사용자들이 작성한 전체 텍스트 개수는 3,968개이다. 한 사용자당 평균적으로 14.70개의 텍스트를 작성하였다.

본 논문에서 식별하고자 하는 사용자 프로파일의 속성으로 총 3가지(성별, 나이, 결혼/연애 상태)로 정의하였다. 클래스는 성별의 경우, 남성, 여성으로, 나이는 10대, 20대, 30대, 40대, 50대로, 결혼/연애 상태는 솔로, 연애중, 기혼으로 정의하였다. 표 4는 각 속성과 해당 클래스에 대한 간단한 통계 정보를 보여준다.

표 4. 속성 및 클래스 별 사용자 및 작성한 텍스트에 대한 정보

속성	클래스	사용자	작성한 텍스트
성별	남성	142	2,157
	여성	128	1,811
나이	10대	22	397
	20대	91	1,352
	30대	50	841
	40대	70	971
	50대	37	407
결혼/연애 상태	솔로	125	2,147
	연애 중	50	708
	기혼	95	1,113

모든 실험은 5회 교차 검증(cross validation)을 통해 평가하였다. 프로파일 식별에 대한 성능 평가는 정확도(Accuracy)를 사용하였다.

본 논문에서는 해당 속성에 대해 가장 많은 사용자를 포함하고 있는 클래스로 분류하는 모델(Base-Model)과 텍스트 하나를 인스턴스로 사용하는 Single-Instance SVMs(SI-SVMs)을 비교 모델(Baseline)로 사용하였다. SI-SVMs의 분류기로는 LIBSVM[16]을 사용하였으며 제안한 방법과 LIBSVM에서 모두 선형 커널(Linear Kernel)을 사용하였다. 다중 클래스를 처리하기 위해 본 논문에서는 One-Versus-All[17] 방법을 사용하였다.

5.2. 프로파일 식별에 대한 실험 결과

그림 2는 성별, 나이, 결혼/연애 상태에 대해 비교 모델들과 제안한 방법의 정확도를 보여준다. 분산의 경우 SI-SVMs과 제안한 방법 모두다 모든 속성에서 0.005이하의 값이 나와 그림에서 제외하였다.

SI-SVMs에서는 사용자가 작성한 텍스트 각각을 클래스 정보를 포함하는 개별 인스턴스로 만들어 학습하였다. 사용자의 클래스를 추정하는 과정에서는 사용자가 작성한 텍스트 각각에 대해서 클래스를 추정하였고, 추정된 클래스 중 가장 많이 분류된 클래스로 사용자 프로파일을 식별하였다.

제안한 방법에서는 사용자 별로 그 사용자가 작성한 모든 텍스트를 포함하는 Bag을 만들었고, Bag 단위로 학습과 클래스 추정이 이루어졌다.

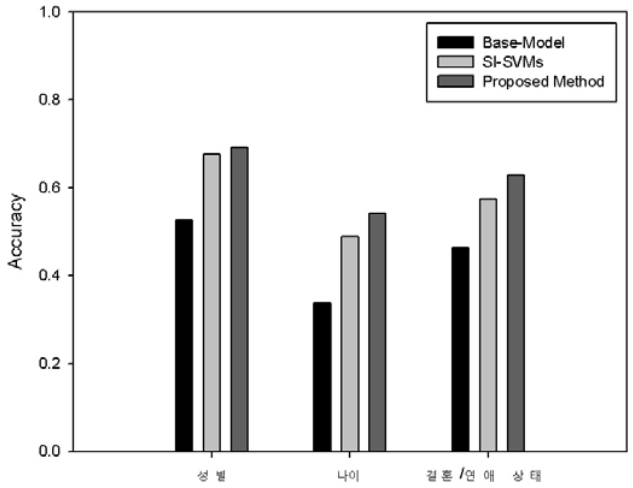

그림 2 사용자 프로파일 식별 실험 결과

그림에서 보여 주듯이 제안한 방법이 비교 모델들보다 모든 속성에서 좋은 성능을 보였다. 특히, 제안한 방법은 SI-SVMs에 비해 성별에서는 0.02, 나이와 결혼/연애 상태에서는 0.05 정도의 차이를 보였다. 위 결과는 개별 인스턴스 학습 방법으로 사용자 프로파일 식별할 시, 프로파일을 식별하는 데에 영향을 끼치지 않거나 잘못 식별하게 만드는 텍스트들이 학습에 고려되어 결정 경계를 잘못 선택하였음을 의미한다. 제안한 방법은 다중 인스턴스 학습을 사용하여 사용자들이 작성한 텍스트 중에 프로파일 식별에 도움이 되는 텍스트들이 결정 경계를 찾는데 사용되었고, 이로 인해 기존의 방법들보다 더 좋은 성능을 보였다. 이를 통해 제안한 방법 즉, 다중 인스턴스 기반 학습 방법이 SNS 텍스트로부터 사용자의 프로파일 식별에 유의미함을 보여준다.

6. 결론 및 향후 연구

사용자 프로파일 식별은 사용자와 관련된 정보로부터 성별, 나이 등의 사용자의 전기적인 속성을 찾는 것이다. 사용자 맞춤형 서비스에서 사용자에게 적합한 서비스를 제공하기 위해 사용자 프로파일 식별을 위한 연구들이 많이 이뤄지고 있다.

본 논문은 SNS에서 사용자들이 작성한 짧은 텍스트들로부터 사용자의 프로파일 식별하는 문제를 다루었다. 제안한 방법은 사용자가 작성한 텍스트 전부를 하나의 학습 단위로 삼는 다중 인스턴스 학습을 기반으로 프로파일 식별하였다. 텍스트 집합을 하나의 학습 단위로 간주하기 때문에 개별 텍스트를 학습 단위로 여기는 기존의 학습 방법에서 발생할 수 있는 오류를 최소화할 수 있었다.

제안한 방법은 SNS 텍스트의 특징을 반영할 수 있는 자질과 분류 문제를 해결하는데 좋은 성능을 보이고 있는 SVMs에 다중 인스턴스 학습을 고려한 MI-SVMs을 이용하였으며, 실험을 통해 기존 학습 방법보다 성별, 나이, 결혼/연애 상태를 식별함에 있어 더 좋은 성능을 보였다.

본 논문에서는 소셜 네트워크 특징을 반영할 수 있는 자질로 Bag-of-Word 및 이모티콘과 같은 문자, 기호 반복 유무만을 사용하였다. 기존 연구에서 n-gram모델과 감탄용어 및 프로파일 식별에 있어 빈번히 사용되는 특정 단어들이 성능을 향상시키는 것으로 알려져 있다. 향후 연구로 이들을 사용하여 프로파일 식별의 성능을 높일 예정이다.

감사의 글

본 논문은 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발)의 지원으로 수행되었습니다.

참고문헌

[1] C. Boulis and M. Ostendorf. “A quantitative analysis of lexical differences between genders in telephone conversations,” In *Proceedings of ACL,* pp. 435-442, 2005.

[2] N. Garera and D. Yarosky, “Modeling latent biographic attributes in conversational genres,” In *Proceedings of ACL and IJCNLP*, pp. 710-718, 2009.

[3] J. Otterbacher, “Inferring gender of movie reviewers: exploiting writing style, content and metadata,” In *Proceedings of CIKM*, pp. 369-378, 2010.

[4] L. Hemphill and J. Otterbacher, “Learning the Lingo? Gender, Prestige and Linguistic Adaptation in Review Communities,” In *Proceedings of CSCW*, pp. 305-314, 2012.

[5] P. W. Foltz, D. Laham, and T.K Landauer, “Automated Essay Scoring: Applications to Educational Technology,” In *Proceedings of EdMedia*, 1999.

[6] D. Rao, D. Yarosky, A. Shreevats, and M. Gupta, “Classifying Latent User Attributes in Twitter,” In *Proceedings of SMUC*, pp. 37-44, 2010.

[7] M. Pennacchiotti and A.-M. Popescu, “A Machine Learning Approach to Twitter User Classification,” In *Proceedings of the Fifth International AAAI*

Conference on Weblogs and Social Media, pp. 281-288, 2011.
[8] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating Gender on Twitter," In *Proceedings of the EMNLP*, pp. 1301-1309, 2011.
[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. "Solving the multiple-instance problem with axis-parallet rectangles," *Artificial Intelligence*, vol, 89, no. 1-2, pp. 31-71, 1997
[10] C. Blake, E. Keogh, and C. J. Merz. UCI repository of machine learning databases. available at "http://www.ics.uci.edu/~mlearn/MLRepository.html"
[11] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," In *Proceedings of ICML*, pp. 341-349, 1998.
[12] C. Yang and T. Lozano-Perez, "Image database retrieval with multiple-instance learning techniques," In *Proceedings of ICDE*, pp. 233-243, 2000.
[13] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," In *Advances in Neural Information Processing Systems,* pp. 1073-1080, 2002.
[14] J. Ramon and L. De Raedt, "Multi instance neural networks," In *Proceedings of ICML*, 2000.
[15] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multi-Instance Learning", In *Advances in Neural Information Processing Systems*, pp. 577-584, 2002.
[16] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1-27:27, 2011.
[17] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," *Journal of Machine Learning Research*, Vol. 5, pp. 101-141, 2004.