기계학습을 이용한 시놉시스 기반 영화장르 분류 기법

이재언*, 홍금원**
*고려대학교 컴퓨터정보통신대학, **고려대학교 컴퓨터 전파통신공학과
e-mail : leeje@onmedia.co.kr

Synopsis-Based Classification of Movie Genres Using Machine Learning Techniques

Jae-Eon Lee*, Gum-Won Hong**
*Graduate School of Computer and Information Technology, Korea University
**Dept of Computer and Radio Communications Engineering, Korea University

요 약

고객의 기호와 요구에 부응하는 서비스의 제공을 위해 영화 요소 중 정확한 장르의 분류는 고객의 선택에 있어 중요한 문제이다. 기존의 수작업에 의한 장르 분류는 시간과 비용, 신뢰성 등에서 비효율적이다. 이러한 문제의 해결을 위해 영화 시놉시스(Synopsis) 기반의 기계학습 방법은 효율적인 대안이 될 수 있다. 본 논문에서는 대다수 영화서비스 주체가 보유하고 있는 시놉시스 정보를 기반으로 하여 기계학습을 이용한 영화장르 분류에 관한 하나의 정형화된 방법을 제시한다. 기계학습 Algorithm 중 LibSVM, RandomComittee, LMT, NaiveBayes, PART Algorithm 을 이용하여 Algorithm 별, 장르별 분류 정확도를 측정하여 비교한다.

1. 서론

최근 문화산업의 중요성에 대한 인식 전환과 이를 실현해 주는 다양한 IT 기술의 발전에 따라 많은 분야에서 새로운 연구와 시도들이 진행되고 있다. 영화 산업에서도 고객의 다양한 기호와 요구에 부합하는 서비스의 제공이 중요해졌다. 이를 위해 많은 영화 요소들에 대한 분석이 진행되고 있으며, 특히 영화장르는 출연배우와 더불어 고객 선택의 가장 중요한 요소로 알려져 있다.

영화장르는 영화의 분류법, 즉 비슷한 종류의 영화를 하나로 묶고 다른 종류의 영화와 구별하는 영화형식에 관한 연구 분야이다. 기존의 장르분류는 영화 홍보제작물을 통해서 혹은 줄거리를 분석해서 수작업으로 구분해 주는 방식이 사용되었으나 이 경우에는 사람의 노력, 비용, 시간 등에서 심각한 어려움을 초래할 뿐만 아니라 각각의 매체마다 서로 다른 분류 기준을 가지고 있어, 동일한 영화가 매체마다 서로 다른 장르로 구분 되는 경우가 발생되어 신뢰성 있는 서비스의 제공에 한계를 가지고 있다.

기존의 영화 장르 분류에 관한 연구들은 Audio 나 영상자료의 Multi-Modal 의 자질들을 이용하여 영화를 관람하지 않고 Trailer 나 Preview 등을 이용한 장르분류를 하였다[1, 2]. 하지만 이러한 영상이나 음성 등의 신호처리를 이용한 분류 방법은 복잡한 신호처리의 기술을 이용해야 한다는 단점이 있다. 영화 시놉시스(Synopsis)는 영화 서비스 주체는 물론 인터넷 등 주변에서 쉽게 자료를 취득하기 용이하고 자연어 처리 방법을 이용하기 때문에 보다 쉽고 편리한 방식으로 영화의 장르 분류가 가능하다.

영화장르는 비슷한 영화들 속에서 반복되는 요소들에 의해 구분되는 것이다. 즉 플롯, 캐릭터, 무대, 배경, 주제, 스타일 등의 기본 요소의 유형화, 반복화된 형식의 관습이다. 그러나 이러한 요소들 만에 의존하여 장르가 구분되는 것은 아니고, 서부영화는 인간의 개척정신을 주제로, 뮤지컬 영화는 노래와 춤이라는 표현방식을 통해서, 코미디 영화는 감정효과가 하나의 장르를 만들어낸다. 이러한 영화장르의 특징을 갖는 요소들을 분석하여 자질을 추출하였으며, 이는 장르 분류의 결과를 결정하는 중요한 요소이다.

이에 본 논문에서는 영화 시놉시스 데이터를 이용해서 사전기반의 장르분류를 실험해 보고, 이와 비교하여 기계학습을 이용한 시놉시스 기반의 장르 분류라는 하나의 정형화 된 방법을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 시놉시스 기반의 영화장르 분류를 위한 전처리 작업과 자질(Feature) 선택, 그리고 선택된 자질의 문서표현과 가중치 부여에 대해서 기술한다. 3 장에서는 실험 내용과 평가에 대해서 기술한다. 마지막으로 4 장에서는 결론과 향후 연구를 기술하는 것으로 본 논문을 맺는다.

2. 시놉시스 기반 영화장르 분류

시놉시스(Synopsis)는 간단한 줄거리, 영화의 개요를 이르는 말로 작가가 생각하는 주제를 다른 사람에게

알리기 위해 쉽고 간단하게 적어놓은 글을 말한다.

본 논문에서 영화장르 분류 실험을 위해 사용된 말뭉치(corpus)는 케이블 TV 채널 브랜드인 OCN 의 영화 Database 를 사용하였으며, 분류의 대상으로 삼은 영화장르는 OCN 과 영화관 메가박스(MegaBox), 그리고 온라인서비스 네이버(naver) 에서 공통으로 사용되고 있는 공포, 성인(에로), 액션, 전쟁, 코미디의 5 개 장르를 분류 실험 장르로 한정하였다. 내용어의 사이즈(size)는 공포, 성인(에로), 액션, 전쟁, 코미디 장르에서 균등하게 100 편씩, 총 500 편의 영화 시놉시스 데이터를 가지고 실험하였다. 분류 실험에 사용된 시놉시스 데이터는 OCN 과 네이버(naver)에서 동일한 장르로 분류된 데이터 만을 대상으로 하였고, 같은 영화임에도 서로 다른 장르로 분류하고 있는 경우에는 실험 데이터에서 제외하였다.

기계학습 장르 분류의 테스트를 위해서 기본 학습 데이터와 테스트데이터의 비율을 9:1 로 하는 10-fold cross validation 기법을 적용하였다.

2.1 전처리 작업

시놉시스 데이터로부터 장르의 내용이나 특징을 잘 반영하고 있는 내용어(content word)를 추출하기 위해 형태소 분석을 수행하였다.

<표 1>은 장르별 고빈도 출현을 보이는 내용어의 예제이다.

<표 1> 장르별 고빈도 출현 내용어

	공포	성인	액션	전쟁	코미디
1	공포	유혹	복수	세계대전	코미디
2	악몽	섹스	살해	전쟁	소동
3	살인마	욕망	조직	포로	사사건건
4	악마	외도	암흑가	독일군	설상가상
5	연쇄살인	정사	음모	병사	실수

본 논문에서는 개념 도입을 가장 잘 설명해 주는 품사인 명사만을 내용어로 선택해서 기계학습 실험을 해 보고, 명사와 함께 사물의 성질이나 상태 또는 존재를 나타내는 품사인 형용사를 내용어로 선택해서 기계학습 시켰을 때의 결과를 비교 실험 하였다. 명사와 더불어 형용사를 내용어의 품사로 사용한 이유는 코미디 장르의 경우, 코믹한 내용에 대한 표현이 어떤 개념적인 설명보다는 멍청한, 어설픈 등의 형용사가 많이 나타날 것이라고 가정하였다.

2.2 자질(Feature) 선택

전처리 단계에서 도출된 내용어는 그 숫자가 많을 뿐더러, 좀 더 좋은 실험 결과를 기대하기 위해서는 유용한 자질(Feature)의 선별이 필수적이다.

본 논문에서는 기존에 연구되어 발표된 자질 선택의 방법 중 가장 좋은 성능을 보이는 것 중 하나로 평가되고 있는 카이 제곱 통계량(χ^2 statistics)을 사용하여 자질을 선택하였다 [6].

카이 제곱 통계량은 용어 t와 범주 c와의 의존성을 측정하는 것으로 그 계산식은 식(1)과 같다.

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

t : 내용어(content word)
c : 범주
N : 전체 문서의 수
A : c에 속해있는 문서 중 t를 포함하는 문서 수
B : c 외의 문서 중 t를 포함하는 문서 수
C : c에 속해있는 문서 중 t를 포함하지 않는 문서 수
D : c 외의 문서 중 t를 포함하지 않는 문서 수

2.3 문서표현 및 가중치 부여

(Document Representation and Term Weighting)

선택된 자질을 어떻게 사용해서 문서를 표현할 것인가에 대한 방법에 관한 것으로, 가장 일반적인 문서 표현 방법은 벡터 공간 모델이다. 이것은 문서 전체에 나타난 각 자질의 출현 빈도(Term Frequency)를 이용하여 문서를 하나의 벡터로 표현 하는 방식으로 자질의 출현 빈도와 역문헌빈도(Inverse Document Frequency) 혹은 역범주빈도(Inverse Category Frequency)를 이용하여 가중치를 줘서 표현하는 방식이다[3, 7].

본 논문에서는 정보 검색 분야에서 가장 기본적으로 사용하는 방법인 TF-IDF(Term Frequency - Inverse Document Frequency) 가중치 방법을 적용하여 문서를 표현하였다. TF-IDF 가중치 방법은 문서에서 각 자질의 가중치는 해당 문서에서 각 자질의 빈도와 역문헌빈도의 곱으로 나타낸다[8].

i 번째 문서에서 나타나는 자질 k 의 가중치는 식(2)와 같다.

$$a_{ik} = f_{ik} \times \log(N / n_k) \quad (2)$$

N : 전체 문서의 수
n_k : 자질 k 가 출현한 문서의 수

3. 실험 및 평가

기계학습 장르분류의 타당성 비교평가를 위해 사전기반의 장르분류 실험을 진행하였다. 총 500 편의 영화 시놉시스 중 450 편을 학습데이터로 활용하였고, 50 편의 데이터를 테스트데이터로 9:1 의 비율로 실험하였다. 학습데이터를 대상으로 장르별로 내용어를 추출하였고, 장르별 단어사전을 구성하였다. 구성된 장르별 단어사전과 시놉시스의 내용과 일치하는 Match 율 기준으로 가장 높은 Match 율을 보인 장르를 결과값으로 선택하였다.

사전기반 장르 실험과 기계학습 실험에서 도출된 분류 정확도를 표현하기 위해서 정확율(Precision)과

재현율(Recall)을 반영한 평균값(F-Measure)을 사용하였으며 그 계산식은 식(3)과 같다.

$$\text{Precision} = \frac{\text{시스템이 추출한 정답 장르 계수}}{\text{시스템이 추출한 장르 계수}}$$

$$\text{Recall} = \frac{\text{시스템이 추출한 정답 장르 계수}}{\text{전체 장르 계수}}$$

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

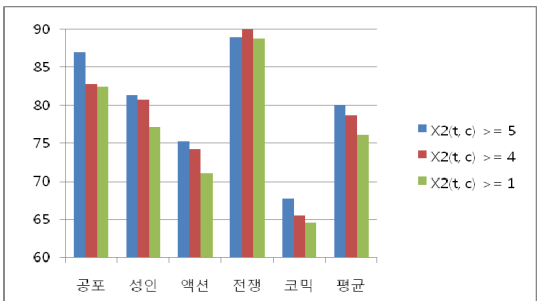
사전기반 장르실험의 결과는 <표 2>와 같았고, 계산식(3)을 이용하여 F-Measure 를 계산한 결과 공포 63.3%, 성인 66.9%, 액션 56.8%, 전쟁 94.7%, 코미디 46.2%로 사전기반의 장르 분류는 평균 65.6%의 F-Measure 를 보였다. 전쟁 장르가 94.7%의 F-Measure 로 예외적으로 좋은 분류 결과를 보였고, 코미디 장르가 46.2%의 F-Measure 로 가장 좋지 않은 분류 결과를 보였다.

<표 2> 사전기반 장르분류 실험 (Confusion Matrix)

정답 \ 시스템결과	공포	성인	액션	전쟁	코미디
공포	6	0	4	0	0
성인	0	7	3	0	0
액션	1	1	8	0	0
전쟁	0	0	1	9	0
코미디	2	3	2	0	3

본 논문에서는 기계학습에 사용될 우수한 자질의 집단을 추출하기 위해서 카이 제곱 통계량의 계산 결과값을 이용하여 비교 실험하였다. 자질의 결과값이 5 이상인 자질 집단, 결과값이 4 이상인 자질 집단, 결과값이 1 이상인 자질 집단을 선택하여 기계학습 실험을 진행하였다. 자질 선택을 위한 내용어는 명사만을 실험 대상으로 하였다.

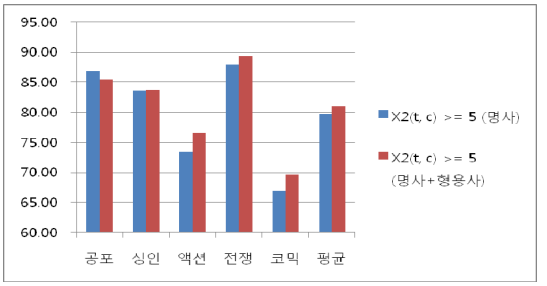
실험 결과는 (그림 2)에서 보는 바와 같이 전쟁 장르를 제외한 모든 장르에서 카이 제곱 통계량 5 이상의 자질 집단을 선택했을 때 가장 좋은 결과를 나타냈으며, 전쟁 장르는 전체적으로 90% 수준의 비슷한 결과가 도출 되었다. 자질 집단의 비교 실험 결과로 볼 때, 단순히 많은 수의 자질을 선택하기 보다는 적절한 수량의 유용한 자질 집단을 선택하였을 경우 더 좋은 결과를 얻을 수 있다는 것을 보여주고 있다.



(그림 2) 자질 선택을 위한 장르 분류 실험

자질 선택 실험에서 가장 우수한 성능을 보인 카이 제곱 통계량 결과값이 5 이상인 자질 집단을 대상으로 내용어의 품사가 명사와 명사+형용사의 비교 실험을 하였다. 결과적으로 코미디 장르를 비롯해서 액션 장르 등에서 형용사가 많이 쓰이고 있었지만, 이러한 형용사는 명사에 비해 출현 빈도수가 적었고, 카이 제곱 통계량의 결과값이 5 이상이 되어 자질로 선택되어지는 경우는 예상보다 많지는 않았다.

실험 결과는 (그림 3)에서 보여지는 것처럼 공포 장르를 제외한 모든 장르에서 명사와 형용사를 함께 내용어로 사용했을 경우 더 좋은 결과를 얻을 수 있었다. 내용어의 품사로 명사만을 선택하여 사용했을 때 평균 79.74%의 F-Measure 를 보였고, 명사와 형용사를 함께 선택해서 사용했을 경우 평균 80.97%의 F-Measure 를 보였다. 공포 장르의 경우 명사만을 내용어로 사용했을 경우 86.78%의 F-Measure 를 보인데 비해, 명사와 형용사를 내용어로 사용했을 경우 85.5%의 F-Measure 로 좋지 않은 예외적인 결과가 나타났다.



(그림 3) 내용어(contend word) 품사 비교 실험

기계학습 Algorithm 별 성능 평가를 위해서 각각의 학습 Algorithm 기법을 선택하였고, 어떠한 학습기법에서 우수한 결과를 보이는지 평가하고자 하였다. 신경망 학습기법인 LibSVM, Meta 학습기법인 RandomComittee, 결정트리 학습기법인 LMT, 베이지안 학습기법 NaiveBayes, 규칙기반 학습기법 PART 5 가지의 기계학습 Algorithm 을 이용하여 장르 분류 실험을 진행하였다.

기계학습 Algorithm 별 비교 실험의 결과는 <표 3>에서 보여지는 것처럼 신경망 학습기법인 LibSVM Algorithm 이 전체적으로 평균 84.5%의 F-Measure 로 가장 좋은 결과를 보였다. LibSVM Algorithm 은 공포 장르를 제외한 성인, 액션, 전쟁, 코미디 장르에서 가장 좋은 F-Measure 를 보였다. 공포 장르에서는 베이지안 학습기법인 NaiveBayes Algorithm 이 86.7%의 F-Measure 로 가장 우수한 결과를 도출하였고, LibSVM Algorithm 은 83.5%의 F-Measure 를 보여 5 개 Algorithm 중 좋지 않은 결과가 나타났다.

기계학습 실험에 대한 결과를 장르별로 분석해 보

면, 실험대상으로 한 5 개의 장르 중 전쟁 장르가 사용된 5 개의 Algorithm 평균 89.2%의 F-Measure 를 보여 가장 좋은 분류 결과를 보였고, 공포 83.5%, 성인 80.0% 장르도 비교적 좋은 분류 결과가 도출되었다. 액션 73.1% 장르는 평균 이하의 결과가 도출되었고, 코미디 장르가 68.2%의 F-Measure 로 가장 좋지 않은 분류 결과를 보였다.

<표 3> Algorithm 별/장르별 기계학습 실험 결과

	X2(t, c) >= 5					
	공포	성인	액션	전쟁	코믹	평균
LibSVM	83.5	89.9	79.9	92.5	76.8	84.5
RandomCommittee	82.6	83.5	78.4	91.6	70.1	81.2
LMT	86.3	82.7	76.5	86.2	66	79.5
NaiveBayes	86.7	78.1	70.2	91.6	63.8	78.1
PART	78.5	65.6	60.6	84.2	64.4	70.7
평균	83.5	80.0	73.1	89.2	68.2	78.8

사전기반 장르분류 실험 결과와 기계학습 실험을 비교하여 보았을 때, 사전기반 장르분류 실험이 평균 65.6%의 F-Measure 를 보였고, 기계학습 Algorithm 중 가장 좋은 성능을 보인 LibSVM Algorithm 이 평균 84.5%의 F-Measure 를 보였다.

장르별로 비교해 보았을 때, 대부분의 장르에서 기계학습 분류 실험의 결과가 사전기반의 분류 실험의 결과보다 23~31% 우수한 결과가 도출되었다. 전쟁 장르는 예외적으로 사전기반 분류를 하였을 때와 기계학습 실험을 하였을 때의 오차가 거의 나타나지 않았다. 전쟁 장르는 LibSVM Algorithm 에서 92.5%의 F-Measure 를 보였고 사전기반의 분류 실험에서는 94.7%의 F-Measure 를 보여 근소하지만 더 우수한 결과가 나타났다. 이러한 결과가 나타난 것은 시놉시스 데이터에 전쟁 장르를 특징짓는 유용한 자질들이 다른 장르에 비해 충분히 내포되어 있음을 의미한다. 코미디 장르는 사전기반 장르분류 실험에서 46.2%의 F-Measure 를 보였고, 기계학습 실험에서는 LibSVM Algorithm 이 76.8%의 F-Measure 를 보였다. 코미디 장르는 사전기반 분류실험과 기계학습 실험에서 모두 전체적으로 좋지 않은 분류 결과를 보였다. 이러한 결과로 볼 때, 시놉시스 내에서 코미디 장르를 분류할 수 있는 내용어가 다른 장르에 비해 제한적이라는 것을 알 수 있으며, 코미디 장르에서 좀 더 좋은 분류 결과를 기대하기 위해서는 코미디 장르를 특징으로 하는 제작사, 감독, 배우 등에 대해서도 추가적으로 연구해서 보완할 수 있을 것이다.

4. 결론 및 향후 연구

본 논문에서는 기계학습 기법을 이용해서 영화 시놉시스에 기반한 장르 분류의 방법을 제안하였다. 사전기반 분류방식과 비교할 때 기계학습을 이용한 장르 분류는 분류 결과에서 우수한 성능이 도출되었다. 아울러 시간과 비용 등을 감안할 때 그 효용성에서 기계학습을 이용한 장르 분류는 충분히 가치가 있음이 실험 결과로 나타났다.

기계학습을 통해 보다 좋은 분류의 결과를 얻기 위해서는 유용한 자질을 선택하는 것이 중요하였다. 실험 결과, 기계학습 Algorithm 중에서 신경망 학습기법인 LibSVM Algorithm 이 전체적으로 좋은 분류 성능을 보였고, 규칙기반의 학습방법인 PART Algorithm 이 가장 좋지 않은 분류 결과를 보였다. 영화 장르별로는 전쟁, 공포, 성인 장르가 LibSVM Algorithm 을 포함한 전체 기계학습 Algorithm 에서 우수한 분류 결과가 나타났다. 코미디 장르는 모든 Algorithm 에서 가장 좋지 않은 분류 결과가 나타났는데, 향후 연구에 있어서 내용어에 대한 추가적인 분석과 자질 선택에 있어서 새로운 고려사항의 적용이 필요하겠다.

향후 연구에 있어서는 시놉시스 보다 더 충분한 내용을 보유하고 있는 대본이나 전체 줄거리 자료를 가지고 기계학습 실험을 통한 장르 분류를 진행해 볼 필요가 있겠다.

참고문헌

[1] Wengang Cheng, Chang'an Liu, Xingbo Wang, " A Rough Set Approach to Video Genre Classification", 2003.

[2] Rasheed. Z, Shah. M, "Movie Genre Classification by Exploiting Audio-Visual Features of Previews", Proc. Of 16th ICPR, 2002.

[3]고영중, 서정연, "문서관리를 위한 자동문서범주화에 대한 이론 및 기법", 정보관리연구, Vol.33, no.2, pp. 19-32., 2002.

[4] 송진석, "지능형 개인화 EPG 를 위한 프로그램 정보 장르 분류", 고려대학교 컴퓨터정보통신대학 공학석사 학위논문, 2007.

[5] Lewis David D. and Marc Ringuette, " A Comparison of Two Learning Algorithms for Text Categorization", Proceeding of The 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[6] Yang. Y, J. O. Pederson, "A Comparative Study on Feature Selection in Text Categorization", Proceedings of The 14th International Conference on Machine Learning, 1997.

[7] Salton. G, E. A. Fox, H.Wu, "Extended Boolean Information Retrieval", Communications of The ACM, pp. 1022-1036, 1983.

[8] Salton G, M. J. McGill, "An Introduction to Modern Information Retrieval", McGraw-Hill, 1983.

[9] 박상준, "기계 학습을 이용한 내용 기반의 음악 장르 분류", 서울대학교 컴퓨터공학과 공학석사학위논문, 2002.

[10] 한경수, "질의분해를 이용한 적합성 피드백 기반 자동 문서요약", 고려대학교 컴퓨터학과 공학석사 학위논문, 2000.

[11] 김명윤, 김명섭, "애플리케이션 트래픽 분류를 위한 머신러닝 알고리즘 성능 분석", 한국정보처리학회, 제 15 권, 제 1 호, 2008.