와인 후기 빅 데이터 분석과 딥러닝 기반 추천 시스템 개발

지홍근*, 이태기*
*한국산업기술대학교 컴퓨터공학과
e-mail:hgchi@naver.com, tedli@kpu.ac.kr

Recommender System Development Based on Wine Review Big Data Analysis and Deep Learning

Hong-Geun Ji*, Tae-Ki Lee*
*Dept of Computer Engineering, Korea Polytechnic University

요 약

최근 사람들의 삶의 질이 향상됨에 따라 기호품인 와인의 수요가 늘어나고 있다. 그러나 와인은 생산하는데 길게는 수십 년이 걸리는 고가의 제품이므로 소비자가 와인과 잘못 구매했을 때의 기회비용이 크다. 본 논문에서는 전문 와인 테이스터 들의 후기 빅 데이터를 활용하여 딥러닝 기반 추천시스템을 개발을 다룬다. 테이스터 들의 후기 빅 데이터에 대해 Apache Pig와 자연어 처리를 통한 전 처리 과정을 수행해 리뷰 별로 특징 벡터를 구성하고, 하이퍼 매개변수 최적화와 조기 종료 기법을 사용해 데이터에 대하여 최적의 딥러닝 분류기를 구성하였다. 마지막으로, 구성된 시스템의 신뢰도를 검증하기 위해서 딥러닝의 정확도와 오차율을 확인하였고 시스템이 추천한 와인을 시각화 이미지와 비교하여 성능을 검증하였다.

1. 서론

사람들의 삶의 질이 점점 증가함에 따라 와인도 점차 가까이 할 수 있는 기호품이 되었다. 작년 비네스포에 따르면 한국의 와인 소비는 2021년까지 20만 상자이상이 늘어나면서 1.2%의 성장을 이루고, 스파클링 와인은 그 속도가 더욱 빨라 동일한 기간에 4.3%의 성장률을 기록할거라고 예상했다. 이러한 와인시장의 성장에 맞추어 처음 와인을 접하는 사람들에 대한 기업들의 마케팅은 소비자들의 각각의 취향에 맞추어 갈 필요가 있다. 본 논문에서는 와인 전문가들의 후기 빅 데이터를 분석하여 소비자들의 입맛에 적합한 와인을 추천해주는 시스템 개발을 다룬다.

2. 관련기술

1. 하둡 분산파일 시스템(HDFS)과 피그(Apache Pig)

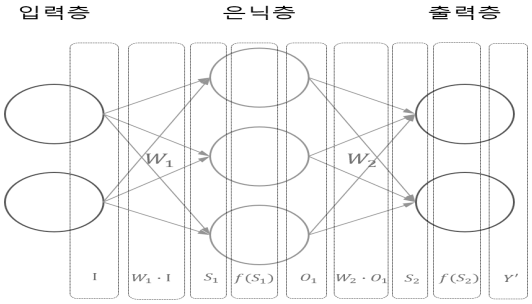
하둡 분산 파일 시스템은 하둡(Hadoop) 프로그램에서 사용되는 기본 데이터 저장소 시스템이다. HDFS는 빅 데이터를 관리하고 응용 프로그램을 지원할 수 있는 인터페이스를 제공한다. 아파치 피그(Pig)는 맵 리듀스를 사용하기 위한 스크립트 언어와 이를 위한 대용량 데이터 분석 플랫폼이다.

2. 맵리듀스 (Mapreduce)

맵 리듀스는 다수의 컴퓨터를 사용해 빅 데이터를 병렬 처리하기 위한 프레임워크이다. 맵 리듀스는 Map, Shuffle, Reduce의 세가지 단계로 구성된다.

3. 딥러닝(Deep Learning)

딥러닝(Deep Learning)은 퍼셉트론 선형 분류기를 다층으로 쌓아 구성한 인공 신경망에 기반 한 알고리즘이다.[1] 인공 신경망의 학습을 위해 오차 역전파 학습 알고리즘이 사용되는데 이는 경사 하강법(Gradient Descent)을 신경망의 여러 은닉층에 적용하는 알고리즘이다. [2]



(그림 1) 신경망 개념도

입력 벡터를 I, 입력층과 은닉층 사이의 가중치 행렬을 W_1, 은닉층과 출력층 사이의 가중치 행렬을 W_2, 활성화 함수를 f라고 하면 신경망의 출력 벡터 Y'는 다음과 같이 나타낼 수 있다.

$$S_1 = W_1 I, \quad O_1 = f(S_1)$$

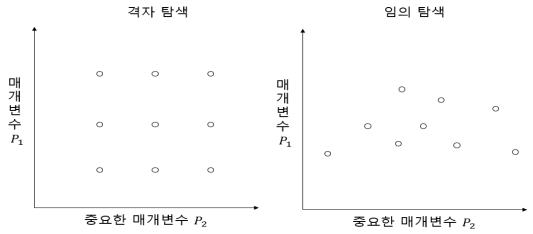
$$S_2 = W_2 O_1, \quad Y' = f(S_2)$$

손실함수를 MSE(mean square error)를 사용하였을 때 손실함수 L과 은닉층의 가중치 행렬 W_1과 W_2의 갱신은 다음과 같다

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{'})^2$$

$$W_1 = W_1 - \eta\frac{\partial L}{\partial W_1},\ W_2 = W_2 - \eta\frac{\partial L}{\partial W_2}$$

3.1. 하이퍼 매개변수 최적화와 조기종료



(그림 2) 하이퍼 매개변수 최적화 개념도

신경망의 모델은 은닉층의 개수, 뉴런 개수, 활성화 함수, 학습률 등 다양한 하이퍼 매개변수가 있다. 최적의 하이퍼 매개변수 조합을 결정하는 방법에는 격자 탐색(grid search), 임의 탐색(random search)이 있다. 격자 탐색은 일정 간격을 두어 가능한 모든 하이퍼 매개변수의 조합을 시도해 보는 방법이다. 임의 탐색은 가능한 모든 하이퍼 매개변수에서 임의로 매개변수 조합을 선택하여 탐색하는 방법이다. [4]

조기 종료는 신경망이 훈련 데이터에 과도하게 학습이 되기 이전에 손실함수의 값이 최소인 세대에서 학습을 종료하여 과적합을 방지하는 알고리즘이다.[5]

4. 텐서플로(TensorFlow)와 사이킷런(Scikit-learn)

텐서플로는 머신러닝을 위한 수치계산 오픈소스 라이브러리로 구글이 주도하여 개발하고 있다. 대부분의 머신러닝과 신경망을 모델링 할 수 있다.

사이킷런은 기계학습용 라이브러리 이다. 모델을 학습시키고 평가, 사용하기 위한 다양한 기능을 제공한다.

5. 자연어 처리와 nltk (Natural Language Tool Kit)

자연어 처리(natural language processing)는 사람들이 사용하는 자연어를 컴퓨터와 같은 기계가 이용 할 수 있도록 하는 인공지능의 연구 분야이다. nltk(Natural Language Toll Kit)는 토큰화(tokenizing), 형태소 분석(morphological analysis), 품사 태깅(tagging)기능을 제공한다.

6. 시각화

빅 데이터는 대량의 데이터를 사용하므로 이를 시각화 함으로 데이터에 내재된 특성을 파악할 수 있다. 이는 이해하기 쉽고 데이터 기반 의사결정을 내리는데 효과적이다. 워드 클라우드는 시각화의 일종으로 빈출된 단어의 빈도를 계산하여 표현한다.

3. 관련연구

1. 와인과 선호도와 관련된 연구

2005년에 와인 소비자 분류에 따른 와인 선호도에 관한 연구에 따르면 와인 소비자를 4단계로 분류하고 와인 선호 항목을 18가지로 정한 뒤 조사 대상자들의 특성을 임의로 분류하여 특정 소비자층이 선호하는 와인의 특징에 대해 조사한 사례가 있다. 하지만 이는 소비자들의 특성을 하나의 데이터 집합으로 묶어서 기업의 마케팅 전략으로 보았기 때문에 소비자 개개인의 성향을 간과한 문제점이 존재한다.

4. 구현 및 실험

1. 데이터 수집 및 하둡 분산 환경에 저장

캐글에서 와인 후기 데이터를 총 16757개 와인에 대해서 147963개의 후기 데이터를 수집하였다. 각 열은 country, description, designation, points, price, promince, region, taster name, title, variety, winery로 구성되어 있다. 이를 하둡 HDFS(Hadoop Distributed File System)에 저장하였다.

2. Pig로 데이터 정제

빅 데이터 처리용 언어인 Pig를 이용하여 대규모의 데이터에서 사용자가 작성한 와인 후기 description과 와인의 이름 winery을 추출하였다. 이중에서 와인 후기의 일반화 성능을 높이기 위해서 와인 당 후기가 130개 이상인 와인들만 선별하였다. 이러한 와인은 모두 38개를 선별하였고 와인 후기 개수는 도합 6584개를 선별하였다. 이러한 와인에는 chateau ste. michelle, bodega norton, dfj vinhos 이 있다.

3. 단어 토큰화와 맵리듀스로 단어 카운트

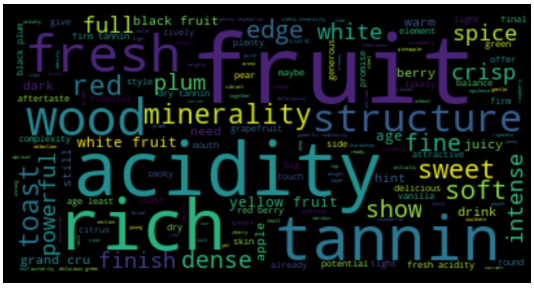
각 와인마다 리뷰에 대해서 토큰화 하고 이중에서 와인의 특징을 나타내도록 명사, 형용사를 추출하였다. 자주 사용되지만 의미가 없는 단어들인 불 용어(stop word)들은 와인의 특징을 나타내는 데에는 의미가 없으므로 제외하였다. 맵 리듀스 기반 워드 카운트로 와인마다 단어의 빈출 비율을 측정하여 내림차순으로 정렬하였다.

<표 1> albert bichotwords 와인 단어 빈도

albert bichotwords 와인 단어 빈도
fruit : 1.0
acidity : 0.5625
rich : 0.515625
tannin : 0.453125
fresh : 0.296875
wood : 0.28125
structure : 0.265625
minerality : 0.25
(생략) ...

4. 워드 클라우드로 시각화 및 특징 단어 추출

와인 대해서 워드 클라우드를 만들어 시각적으로 이 와인의 주요 특징이 무엇인지를 알 수 있다.



(그림 3) albert bichotwords 워드 클라우드

워드 클라우드로 확인 후에 대부분의 와인 리뷰에서 많이 사용되는 단어들은 개별 와인의 특징을 나타내지 못하므로 제외하였다. 이러한 단어에는 flavor, well, good, year, note, best, texture, show, next, great, nice, vineyard 등이 있다.

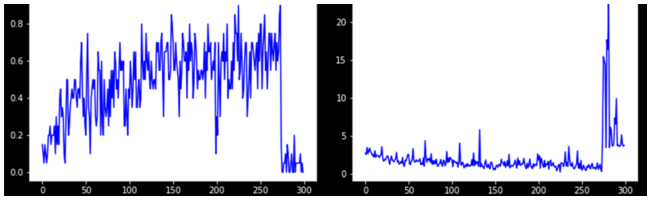
와인 리뷰에서 등장한 단어 총 188개의 와인 특징 단어들을 선별하였다. 이렇게 선별된 단어에는 fruit, acidity, rich, tannin, fresh, structure, wood, toast, minerality, sweet 등이 있다. 단어를 188차원의 one-hot 벡터로 각 단어를 나타내었다.

5. 리뷰를 특징 벡터로 변환

하나의 와인 후기에 대해 리뷰에 포함된 단어들의 값만 1로 설정하여 희소행렬(sparse matrix)을 구성하여 특징 벡터를 구성하였다. 예를 들어, 특정 와인 리뷰에서 등장한 단어가 fruit, red, soft일 때, fruit이 [1, 0, 0, ... ,0]이고 red이 [0, 0, 1, 0, 0, ... ,0] 이고 soft이 [0, 0, 0, 1, 0, 0, ... ,0] 이라면 이 와인 리뷰는 [1, 0, 1, 1, 0, ... ,0]으로 표현한다.

6. 사이킷런과 텐서플로 기반 딥러닝 분류기 구성

텐서플로를 활용하여 사이킷런과 호환되는 딥러닝 모델을 구성하였다. elu 활성함수와 100개의 뉴런이 있는 은닉층을 5층 구성하였다.[6] 가중치 초기화 방법은 He 초기화 방법[7]을 사용하였고 아담(Adam)을 옵티마이저로 사용하였다.[8] 손실함수로는 교차 엔트로피 함수를 사용하였고 배치사이즈 20개로 한 세대마다 모든 훈련 데이터를 사용하여 학습하였다. 분류 문제로서 softmax 함수의 출력과 one-hot 인코딩 된 와인의 이름을 목표 값으로 설정하여 학습을 수행하였다.



(그림 4) 정확도(좌)와 오차 값(우) 그래프

초기 정확도는 10~ 20% 정도에 머물러 있었으나 학습 세대 273번째에서 90%의 정확도를 최고점으로 정확도가 5~15 % 정도로 감소하였고 손실함수의 오차 값도 크게 증가하는 문제점이 발생했다.

7. 조기 종료와 하이퍼 매개변수 최적화

학습 중에 오차 값이 최소인 세대 k에 도달하면 n 세대만큼 더 학습을 진행해보고 새로운 최소 오차값에 도달하지 않으면 k세대의 가중치를 사용하는 방법으로 조기 종료를 수행하였다. 그리고 최적의 하이퍼 매개변수를 탐색하기 위해서 하이퍼 매개변수 최적화 방법으로 임의 탐색 방법을 사용하였다. 임의 탐색 방법에서 탐색할 하이퍼 매개변수는 배치 정규화와 elu, relu등의 활성함수를 포함하여 <표 2>와 같다.[3] 테스트 환경 Intel Xeon, RAM 12GB, Tesla K80에서 수행하였다. 최적의 하이퍼 매개변수 조합은 <표 3> 와 같다.

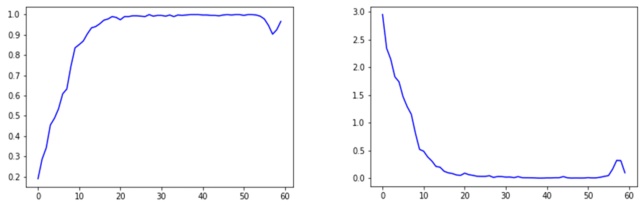
<표 2> 하이퍼 매개변수

하이퍼 매개변수	매개변수의 값
층의 뉴런 개수	70, 90, 100, 120, 140, 160
배치 사이즈	50, 100, 500
학습률	0.01, 0.02, 0.05, 0.1
활성화 함수	relu, elu, leaky relu
은닉층의 개수	3, 4, 5, 6, 7, 8, 9, 10
배치정규화모멘텀	0.9, 0.95, 0.98, 0.99, 0.999

<표 3> 최적의 하이퍼 매개변수

하이퍼 매개변수	매개변수의 값
층의 뉴런 개수	140
배치 사이즈	500
학습률	0.01
활성화 함수	elu
은닉층의 개수	4
배치정규화모멘텀	0.98

최적의 하이퍼 매개변수로 학습을 진행하였을 때의 정확도 그래프와 오차값 그래프는 다음과 같다. 학습이 진행되고 대략 40세대 만에 최적 점에 도달한 것을 확인하였다.



(그림 5) 최적 매개변수 신경망의 정확도, 오차 값 그래프

8. 와인 추천 시스템 동작 예

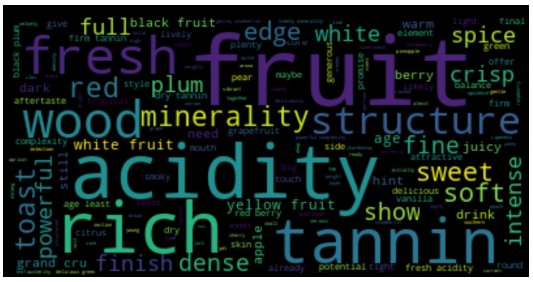
사용자로부터 <표 4>와 같은 키워드를 입력으로 받았다고 가정하고 테스트를 진행하였다. 와인 추천 확률에는 구성한 모델의 softmax 함수 출력을 사용하였다. 와인 추천 시스템에 따르면 사용자에게 추천해줄 상위 3개의 와인은 <표 5>와 같다. 시스템의 성능을 검증하기 위해서 albert bichot 와인 리뷰에서 많이 등장한 단어로 구성된 워드 클라우드(word cloud)를 살펴보면 (그림 6)과 같다.

<표 4> 와인 추천 시스템의 입력(1)

fruit	신선한 과실맛과 과실향
tannin	포도의 씨, 껍질, 줄기, 오크통의 페놀 성분.
fresh	신선함을 주는 느낌
wood	오래 숙성된 와인에서 나는 나무 향과 맛
sweet	단맛
aroma	원산지에 따른 포도의 향기

<표 5> 와인 추천 시스템의 출력(1)

74.58%	albert bichot
12.22%	louis latour
5.41%	joseph drouhin



(그림 6) albert bichot의 워드 클라우드

워드 클라우드를 구성한 단어를 살펴보면, fruit, tannin, fresh, sweet 등이 등장한 것을 알 수 있다. 따라서 albert bichot와인을 적절하게 추천해 준 것을 확인 할 수 있었다.

이번에는 <표 6>과 같은 키워드를 입력받았다고 가정하고 테스트를 진행하였다.

<표 6> 와인 추천 시스템의 입력(2)

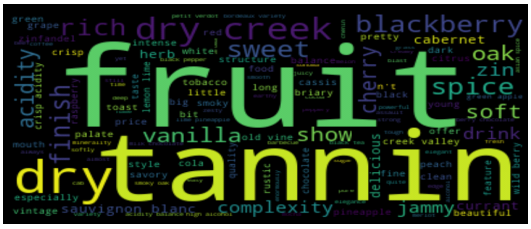
rich	와인의 감미성 정도
finish	뒷맛으로 얻어지는 즐거움을 의미
herb	허브향
oak	참나무로 만든통 에서 발효된 와인의 향
vanilla	바닐라 향

시스템이 추천해준 상위 3개의 와인과 추천 확률은 <표 7>과 같다.

<표 7> 와인 추천 시스템의 출력(2)

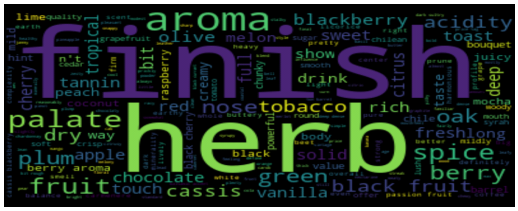
68.74%	dry creek vineyard
27.70%	errazuriz
2.04%	yalumba

시스템이 추천해준 최적의 와인인 dry creek vineyard의 워드 클라우드를 살펴보면 다음과 같다.



(그림 7) dry creek vineyard의 워드 클라우드

추천해준 와인의 워드 클라우드를 살펴보면 vanilla, oak, finish등의 단어가 상위에 분포되어 있다. 또한 이번 실험에서는 두 번째로 추천해준 와인의 추천도도 25%가 넘었다. 두 번째 와인의 워드 클라우드는 다음과 같다. finish와 herb등의 특징 단어가 상위에 분포하고 있는 와인임을 확인 할 수 있다. 따라서 시스템이 사용자가 원하는 적절한 와인을 추천해주고 있음을 알 수 있다.



(그림 8) errazuriz의 워드 클라우드

5. 결론

본 논문에서는 전문 테이스터 들의 와인 후기 빅 데이터를 자연어 처리를 통한 전 처리를 수행한 후 최적화된 딥러닝을 활용해 사용자 후기 기반 와인 추천 시스템의 구현과 작동 예시를 다루었다.

참고문헌

[1] Marvin Minsky and Seymour Papert "Perceptrons: an introduction to computational geometry"

[2] Rumelhart, D.E. and Mcclelland, J.L. "Parallel distributed processing: explorations in the microstructure of cognition"

[3] Sergey Ioffe and Christian Szegedy "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift"

[4] James Bergstra and Yoshua Bengio "Random Search for Hyper-Parameter Optimization", Journal of Machine Learning Research 13

[5] Lutz Prechelt "Early Stopping - but when?"

[6] Djork-Arné Clevert, Thomas Unterthiner, Sepp Hochreiter "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)"

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", Microsoft Research

[8] Diederik P. Kingma, and Jimmy Lei Ba "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION", conference paper at ICLR