Journal of KIIT. Vol. 21, No. 12, pp. 211-216, Dec. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.12.211

음원 차트의 순위 구간화를 통한 흥행 예측 방법론

김우석*[1], 강준하*[2], 김민제*[3], 정혜진*[4], 이수원*[5], 최상민**

An Approach for Predicting Chart Success by Segmenting Music Chart Rankings

Woo-Seok Kim*[1], Junha Kang*[2], Min-Je Kim*[3], Hye-Jin Jeong*[4], Suwon Lee*[5], and Sang-Min Choi**

본 논문은 교육부와 한국연구재단의 재원으로 지원 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0), 정부(과학기술정보통신부) 재원의 한국연구재단 지원(RS-2022-00165785)과 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다(2021RIS-003)

요 약

최근 음악 시장의 규모가 확대되면서 수많은 음원이 발매되고 있으나, 흥행하는 곡의 비중은 매우 낮다. 이에 따라 발매되는 음원의 흥행을 예측하는 연구가 다양하게 진행되고 있으며, 인기 있는 아티스트의 음원이거나 최고 순위가 높은 음원일수록 오랜 기간 음원 차트에 상주하는 경향을 보인다. 본 연구는 음원 차트를 일정한 크기로 분할하고, 음원을 최고 순위를 기준으로 배치하여 임의의 음원이 어떤 구간에 속할지 예측한다. 실험 방법은 음원 데이터를 최고 순위로 정렬한 후 구간에 따라 라벨링하여 음원의 특성들을 수치 데이터로 전처리하는 과정을 거치고, 음원이 속할 구간을 예측한다. 예측을 위해 9개의 모델을 이용하여 실험을 진행하고, f1-score를 통해 평가한다. 실험 결과로 MLP가 타 모델들에 비해 뛰어난 성능을 보인다.

Abstract

As the music market has expanded in recent years, a large number of songs have been released, but the percentage of successful songs is very low. As a result, various studies have been conducted to predict the success of released songs, and songs by popular artists or songs with the highest rankings tend to stay on the music charts for a long time. This study divides the music chart into bins of a certain size and places the bins based on the highest ranking to predict which bin an arbitrary song will fall into. This allows us to roughly measure the level of a song's success. The experimental method preprocesses the characteristics of the music into numerical data by sorting the music data by the highest ranking and labeling them according to the section, and predicts the section to which the music belongs. Experiments are conducted using 9 models for prediction and evaluated by f1-score. The experimental results show that MLP outperforms other models.

Keywords

machine learning, deep learning, classification, prediction

* 경상국립대학교 컴퓨터과학과
- ORCID[1]: https://orcid.org/0009-0002-8759-3728
- ORCID[2]: https://orcid.org/0009-0002-4975-0294
- ORCID[3]: https://orcid.org/0009-0009-8337-7746
- ORCID[4]: https://orcid.org/0009-0005-1103-5481
- ORCID[5]: http://orcid.org//0000-0003-2603-1385
** 경상국립대학교 컴퓨터과학과 조교수(교신저자)
- ORCID: http://orcid.org//0000-0001-5950-3081

· Received: Nov. 08, 2023, Revised: Dec. 08, 2023, Accepted: Dec. 11, 2023
· Corresponding Author: Sang-Min Choi
Dept. of Computer Science, Gyeongsang National University, Jinju-si, 52828, Korea
Tel.: +82-55-772-1384, Email: jerassi@gnu.ac.kr

Ⅰ. 서 론

2022년 글로벌 음악 시장 규모는 613억 1,7000만 달러로 조사되었다. 이는 전년도 대비 33.3%의 성장을 보이고 있으며, 3년간 꾸준히 성장하고 있다. 또한 음악 감상 시 이용하는 수단 또는 서비스로 '음원 스트리밍'이 두 번째로 높은 순위를 기록했다. 음원 스트리밍은 2021년 63.2% 그리고 2022년 67.0% 그리고 2023년 69.0%로 꾸준한 상승세를 보인다[1]. 이처럼 최근 음원을 소비하는 경향은 대부분 온라인 음원 스트리밍을 통해 이뤄지고 있다.

이와 같이 현재 음원 시장이 성장하면서 수많은 새로운 곡이 발매되고 있다. 이 중 약 20%의 곡들의 전체 스트리밍 횟수의 80%를 차지하는 '파레토 법칙(Pareto principle)'이 나타난다[2]. 이처럼 발매되는 곡에 비해 흥행하는 곡은 그 수가 현저히 적다. 또한 음원 발매에는 상당한 비용이 들기 때문에, 음원의 흥행 여부에 따른 경제적인 부담이 크다. 따라서, 음원을 발매하기 전 해당 음원의 수요 예측은 유의미한 시도로 고려할 수 있다.

본 논문에서는 특정 기간 내 음원 플랫폼에서의 곡의 순위 데이터를 분석하여, 새로운 곡의 최고 순위를 예측하는 방법을 제안한다. 이를 위해 전체 순위에 대한 구간을 나누고 특정 곡과 각 구간의 특성을 분석하여 특정 곡이 속할 구간을 예측한다. 예측을 위해 다양한 머신러닝과 딥러닝 모델을 이용하여 실험을 진행한다.

Ⅱ. 관련 연구

음원의 흥행을 예측하는 선행 연구는 크게 두 가지로 나뉜다. 첫째, 음원의 아티스트, 소속사 등과 같은 고유한 정보를 기반으로 흥행을 예측한다. [3]은 아티스트 관련 기사의 개수나 공식 SNS 계정의 팔로워 수와 같은 지표들을 사용하여 음원이 흥행 여부를 예측한다. 이와 유사하게 아티스트의 인지도를 기준으로 흥행 여부를 분석하기도 한다[4]. 인지도가 높은 음원이 낮은 음원과 비교하여 상대적으로 스트리밍을 이용한 음원 소비 비율이 높다. 아티스트 이외에도 음반의 경우 메이저 기획사에서 제작한 음반은 비메이저 기획사에 비해 차트 내 상주 기간이 더 길고, 차트 내 첫 주 진입 순위가 더 높다[5]. 메이저 기획사에서 유통한 음반 또한 비메이저 기획사에 비해 차트 내 첫 주 진입 순위가 더 높다. 그리고 음반의 차트 첫 주 진입 순위가 높을수록 음반의 차트 내 상주 기간이 더 긴 것으로 확인된다. 또한 메이저 기획사나 곡의 최고 순위가 높을 시, 음원이 차트에 상주하는 기간이 더 길다. [6]은 프로듀서의 영향력이 음원 성적에 영향을 미치는지 Graph Centrality를 이용해 분석한다. 모든 분야에서 흥행하는 음원의 경우 함께 작업하는 특정 프로듀서나 아티스트가 존재한다는 공통적인 특징 발견할 수 있다. 마지막으로 전문가 평가라는 새로운 정량적 지표의 활용 하여 음원의 정량적 측정 가능성을 제시한다[7].

둘째, 팬덤의 규모나 기획사 등과 같은 음악 외적인 특성 외에도 음원의 음악 데이터를 활용하여 흥행을 분석한다. [8]에서는 음원이 가지는 음악적 특성만을 고려하여 흥행 여부를 분석한다. 연구에서는 팬덤의 영향이 적은 신인 가수들의 데이터를 모아 추가 실험을 진행하였으며, 실험에서 고려한 여러 요소 중 danceability가 흥행에 가장 큰 영향을 줬다. [9]는 음악 데이터를 분석하여 Bagged, AdaBoost 그리고 Random Forest를 적용하여 음원을 분류하는 실험을 진행한다. 세 가지 방법 중 Random Forest로 선택된 분류 모델이 정확도와 효율성 모두에서 더 나은 성능을 보이는 것을 확인할 수 있다.

Ⅲ. 데이터 수집 및 전처리

3.1 데이터 수집

표 1. 데이터 수집 조건
Table 1. Conditions for data crawling

Conditions	Contents
Collection period	2021. 01. 01 ~ 2023. 09. 20
Data feature types	Daily ranking, singer, composer, lyricist, arranger, agency, genre, singer type
Collection constraints	Songs released before 2021. 01. 01
	Songs remaining on the chart as of 2023. 09. 20

데이터 수집은 온라인 음원 플랫폼인 '지니 뮤직'에서 진행한다[10]. 일간 차트를 기준으로 하며, 데이터 수집 조건은 표 1과 같다.

3.2 데이터 라벨링

데이터 라벨링을 위해 각 음원을 대상으로 수집 기간 내 최고 순위를 기준으로 오름차순 정렬한다. 실험에 사용할 구간 개수에 따라 전체 음원 데이터를 같은 크기의 구간으로 나눈 뒤, 순차적으로 구간에 번호를 부여한다. 이후 구간별로 속해있는 음원 데이터들에 해당 구간의 번호를 라벨로 부여한다. 구간의 개수는 2개부터 16개까지이며, 전체 구간을 동일한 크기로 나눈다. 그 중 Figure 1과 같이 구간의 개수가 4개인 경우의 라벨링은 다음 설명과 같다. 우선 전체 데이터를 4개의 동일한 크기의 구간으로 나눈다. 이후 최고 순위가 1위부터 50위인 데이터에 1, 51위부터 100위에 2, 101위부터 150위에 3 그리고 151위부터 200위에 4의 라벨을 부여한다.

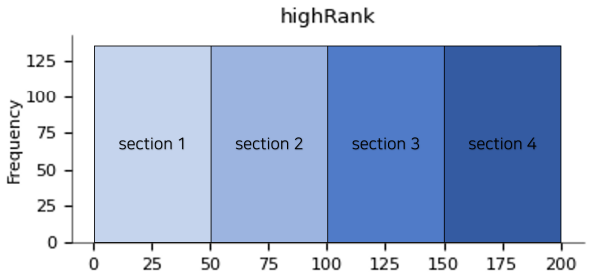


그림 1. 전체 데이터를 4개 구간으로 나눈 라벨링 결과
Fig. 1. Result of labelling the entire data into four sections

3.3 데이터 전처리

데이터 라벨링이 끝난 뒤, 모든 음원 데이터의 각 특성값을 수치 데이터로 변경하는 전처리를 진행한다. 전처리 전의 음원 데이터의 특성 종류는 최고 순위, 가수, 작곡가, 작사가, 편곡가, 소속사, 장르 그리고 가수 유형으로 구성되어 있다. 이 중, 가수, 작곡가, 작사가, 편곡가, 소속사 특성은 아래의 정의들과 식들을 활용하여 전처리를 진행한다.

$$K = \{S_1, S_2, \ldots, S_n\} \tag{1}$$

$$S = \{S_i \mid S_i\text{는 } C_i\text{를 특성으로 갖는 어떤 값}\} \tag{2}$$

$$C = \{C_i \mid C_i\text{는 음원 데이터의 어떤 특성}\} \tag{3}$$

$$T = \{T_i \mid T_i\text{는 특성 값이 } s_i\text{인 음원 데이터}\} \tag{4}$$

$$P = \{P_i \mid P_i\text{는 집합 } T_i\text{의 파티션}\} \tag{5}$$

$$\bigcup P_i = T \tag{6}$$

$$|T_i| = \sum_{j=1}^{n} |P_j| \tag{7}$$

$$P'_i = \frac{|P_i|}{|T|} \tag{8}$$

먼저 식 (1)에 정의된 K는 식 (2)의 집합 S를 원소로 갖는 집합이다. 식 (2)에 정의된 집합 S의 원소는 식 (3)에 정의된 C를 특성으로 갖는 어떤 값, 즉 하나의 데이터를 의미한다. 그리고 C의 원소는 음원의 특성으로 장르와 가수 유형을 제외한 5개로 구성된다. 집합 K는 모든 집합 S를 원소로 가지므로, 전체 데이터의 특성 값을 포함하는 집합이 된다. 식 (4)의 집합 T는 전체 음원 데이터 중 특성 값이 s_i인 음원 데이터이다. 또한 식 (5)의 집합 P는 집합 T의 파티션으로 집합 P의 개수는 구간의 개수와 동일하다. 식 (8)의 P'_i는 집합 P_i의 카디널리티를 집합 T의 카디널리티도 나눈 값으로 최종적으로 전처리를 하는 값이다.

예를 들어, C_i의 값이 '가수'이고 s_i가 '가수 A'라고 하자. 집합 T의 요소는 전체 음원 데이터 중 가수 특성 값이 가수 A인 음원들이다. 구간의 개수가 4개일 때, 집합 T는 4개의 집합 P_i로 구성된다. 식 (6)에 따라 모든 집합 P_i의 합집합은 T가 된다. 따라서 식 (7)과 같이 집합 T의 카디널리티는 모든 집합 P_i의 카디널리티의 합과 같다. $|T|$의 값을 10이라 가정했을 때, 구간에 따른 P'_i값은 식 (8)에 의해 Figure 2와 같이 구할 수 있다.

따라서 전처리 전엔 같은 값이라도 전처리가 끝난 뒤엔 구간 별로 다른 값이 들어가게 된다. 수식에 따라 P'_i은 전체 구간의 k번째 구간에 P_i값이 많이 분포할수록 큰 값을 가진다. 즉, 특정 구간에 해당 값이 있을 확률이다. 이후 장르 및 가수 유형은 범주형 값이므로 원-핫 인코딩(one-hot encoding)을 사용하여 전처리한다.

$number\ of\ sections = 4$
$|T| = 10$

	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$\|P_i\|$	5	2	2	1
P'_i	$\frac{5}{10} = 0.5$	$\frac{2}{10} = 0.2$	$\frac{2}{10} = 0.1$	$\frac{1}{10} = 0.1$

그림 2. 전처리 과정 예시
Fig. 2. Example of preprocessing

IV. 모델 및 처리방법

표 2는 실험에 사용된 9개의 모델 종류와 각 모델에 대한 간단한 설명이다. 또한 머신 러닝 기반의 방법들은 4개의 구간으로 나눈 다음, 최적의 파라미터를 구하는 도구로 GridsearchCV를 통해 파라미터를 구한다. 이후 해당 파라미터를 활용하여 실험을 진행한다.

표 2. 실험에 사용한 모델
Table 2. Models used in the experiment

Model		Description
ML	Random Forest (RF)	Using multiple decision trees, each tree performs individual predictions, and the most frequently predicted result is selected as the final prediction
	Gaussian Naive Bayes (NB)	Assuming that the characteristics of the given data are independent of each other, Gaussian distribution is used to handle continuous data
	Adaptive Boosting (AB)	It is a method of increasing the weight of samples that the previous model misclassified to improve the learning of the next model
	Gradient Boosting (GB)	It is a method of improving the model by reducing the error of the previous model using the gradient of the loss function
	Decision Tree (DT)	It learns decision rules based on rule-based methods and performs predictions on inputs using those rules
	Support Vector Machine (SVM)	It operates by maximizing the distance between decision boundaries of different classes and the nearest data points
	K-Nearest Neighbors (KNN)	It is a method of classifying new data points based on the information of the 'K' nearest neighbors among the existing data points
DL	Multi-Layer Perceptron (MLP)	It is a type of artificial neural network consisting of multiple hidden layers and an output layer
	Convolutional Neural Network (CNN)	It extracts local features of data through convolutional layers and pooling layers, and learns based on them

V. 실험 및 결과

5.1 실험 계획

전처리가 끝난 1,881곡의 8개의 항목을 이용하여 각각의 곡의 라벨을 예측하는 실험을 진행한다. 전체 데이터의 90%를 학습 데이터로 사용하고 10%를 테스트 데이터로 사용한다. 사용한 모델의 종류는 6개의 머신러닝 모델, Multi-Layer Perceptron(MLP) 그리고 Convolutional Neural Network(CNN)이다. 실험 결과에 대한 평가 메트릭은 f1-score를 이용한다[11].

5.2 실험 결과

구간 개수 별 각 모델의 f1-score는 표 3과 같다. NB를 제외한 8개의 모델이 구간 개수가 늘어남에 따라 f1-score가 감소하는 추세를 보인다. 즉, 구간의 개수가 늘어날수록 대부분의 모델에서 부정확한 예측결과가 도출됨을 관찰할 수 있다. 대표적으로 DT와 KNN은 구간의 개수 증가에 따른 f1-score의 감소 폭이 타 모델보다 크다. 전체 모델 중 MLP가 타 모델들과 비교하였을 때, 모든 구간 개수에서 제일 높은 f1-score를 보인다. 따라서 논문에 서 제시하는 음원 수요 예측 방법에서는 MLP 모델을 사용하는 것이 평균적으로 f1-score가 0.928인 성능을 얻을 수 있다.

VI. 결 론

본 논문에서는 머신러닝과 딥러닝 모델들을 활용하여 음원의 최고 순위가 어느 구간에 속하는지를 예측하였다. 구간 개수에 따라 전체 음원을 라벨링한 뒤, 해당 라벨을 예측하는 실험을 진행했다.

표 3. 구간 개수 별 각 모델의 f1-score
Table 3. f1-score for each model by number of sections

number of sections	RF	NB	AB	GB	DT	SVM	KNN	MLP	CNN
2	0.955	0.863	0.934	0.904	0.848	0.890	0.918	**0.957**	0.924
3	0.926	0.841	0.876	0.910	0.758	0.874	0.851	**0.943**	0.928
4	0.923	0.844	0.837	0.895	0.665	0.849	0.804	**0.937**	0.892
5	0.899	8.856	0.850	0.884	0.628	0.850	0.759	**0.927**	0.889
6	0.881	0.864	0.843	0.874	0.535	0.840	0.722	**0.925**	0.862
7	0.887	0.864	0.841	0.875	0.472	0.798	0.662	**0.930**	0.874
8	0.858	0.864	0.839	0.858	0.399	0.773	0.651	**0.922**	0.838
9	0.842	0.876	0.828	0.844	0.354	0.751	0.652	**0.926**	0.849
10	0.837	0.883	0.796	0.841	0.310	0.774	0.630	**0.921**	0.819
11	0.816	0.879	0.776	0.862	0.293	0.765	0.599	**0.917**	0.847
12	0.803	0.857	0.767	0.850	0.284	0.741	0.600	**0.922**	0.828
13	0.813	0.886	0.757	0.838	0.250	0.718	0.593	**0.927**	0.842
14	0.821	0.879	0.718	0.847	0.240	0.715	0.553	**0.923**	0.803
15	0.809	0.873	0.718	0.852	0.220	0.715	0.550	**0.925**	0.844
16	0.792	0.867	0.709	0.852	0.204	0.697	0.506	**0.921**	0.816

실험 결과로 MLP가 모든 구간 개수에 대해 평균 f1-score 0.928로 가장 뛰어난 성능을 보였다. 이를 통해 딥러닝 기반의 모델이 해당 문제를 해결하는 데 조금 더 적합한 모델임을 알 수 있다. 추후 연구로는 MLP가 타 모델들보다 뛰어난 이유와 MLP과 다른 딥러닝 모델들을 비교하는 연구를 진행할 수 있다.

이를 통해 새로운 음원이 발매되기 전, 해당 음원이 어느 정도의 순위를 기록할지에 대한 대략적인 정량적 판단이 가능할 것이다. 이를 통해 손익계산이 가능할 것이며 음원 발매에 있어 보조 지표로 사용될 수 있다.

References

[1] "2023 music industry white paper", https://welcon.kocca.kr/ko/info/trend/1953253 [accessed: Dec. 06, 2023]

[2] J. B. Jung "Digital music market analysis : A study on streaming count", Unpublished master's thesis, Department of Statistics Graduate School of Keimyung University, Daegu, Feb. 2019.

[3] G. Y. Kim and M. J Kim, "A Study on the Prediction Index for Chart Success of Digital Music Contents based on Analysis of Social Data", Journal of Digital Contents Society, Vol. 19, No. 6, pp. 1105-1114, Jun. 2018. http://doi.org/10.9728/dcs.2018.19.6.1105.

[4] J. W. Yoo, J. H. Hyeong, and J. Y. Lee, "What to Download and What to Stream? : Investigating Music Characteristics Driving Preferences in Digital Music Consumption Modes", Korean Journal of Marketing, Vol. 33, No. 1, pp. 1-21, Feb. 2018. http://doi.org/10.15830/kmr.2018.33.1.1.

[5] H. M. Yim, "Who can be a deus ex machina in the industry ?", Unpublished master's thesis, Department of Business Administration The Graduate School of Chung-Ang University, Seoul, Feb. 2015.

[6] G. H. Kim and C. G. Han, "Analysis of between producer and artists' influence and music chart using graph centrality and Recommend New Songs", in Korean Institute of Information Scientists and Engineers, Vol. 49, No. 2, pp. 1863-1865, Dec. 2022.

[7] T. H. Jun, Y. R. Lee, and C. W. Kim, "Quantitative Sound Quality Classification System using Expert Evaluation", Journal of Digital

Contents Society, Vol. 23, No. 3, pp. 423-431, May 2022. http://doi.org/10.9728/dcs.2022.23.3.423.

[8] D. H. Kim, "Analysis of Success Factors in the K-POP Music Market according to Musical Features", in Proc. of KIIT Conference, Jeju, Korea, pp. 887-890, Jun. 2023.

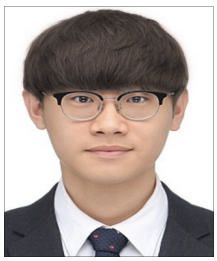
[9] Y. Zhang and D. LV, "Selected features for classifying environmental audio data with random forest", The Open Automation and Control Systems Journal, pp. 135-142, Jul. 2015. http://dx.doi.org/10.2174/1874444301507010135.

[10] Genie music Daily Chart, https://www.genie.co.kr/chart/top200. [accessed: Sep. 20, 2023]

[11] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", arXiv:2010.16061, Oct. 2020. https://doi.org/10.48550/arXiv.2010.16061.

저자소개

김 우 석 (Woo-Seok Kim)



2023년 2월 : 경상국립대학교 컴퓨터과학과(공학사)
2023년 3월 ~ 현재 : 경상국립대학교 컴퓨터과학과 석사과정
관심분야 : 추천 시스템, 인공지능, 자연어처리

강 준 하 (Junha Kang)

2020년 3월 ~ 현재: 경상국립대학교 컴퓨터과학과 학부과정
관심분야 : 인공지능, 딥러닝

김 민 제 (Min-Je Kim)

2018년 3월 ~ 현재 : 경상국립대학교 컴퓨터과학과 학부과정
관심분야 : 인공지능, 시계열 분석

정 혜 진 (Hey-Jin Jeong)

2018년 3월 ~ 현재: 경상국립대학교 컴퓨터과학과 학부과정
관심분야 : 추천시스템, 알고리즘

이 수 원 (Suwon Lee)

2012년 2월 : 한국과학기술원 전산학과(공학석사)
2017년 2월 : 한국과학기술원 전산학과(공학박사)
2018년 3월 ~ 현재 : 경상국립대학교 컴퓨터과학과 부교수
관심분야 : 증강현실, 컴퓨터비전

최 상 민 (Sang-Min Choi)

2015년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2018년 3월 ~ 8월 : 연세대학교 컴퓨터과학과 박사후연구원
2022년 3월 ~ 현재 : 경상국립대학교 컴퓨터과학과 조교수
관심분야 : 추천 시스템, 알고리즘