Analysis of Anime Rating Preferences and Clustering of Both Shows and Viewers

Yuhao Zhou, Haopeng Huang, Yuanzheng Zhao, Kaihao Liu

Introduction

Anime, an emerging sector of Japanese culture after WWII, has spread throughout the world in the past several decades and become a symbol for modern pop-culture.

Although anime is already a significant part of youth life, fewer people considered studying its impact on both our society and culture.

Today, since technology has made data collection easier than ever, we want to utilize methods of machine learning to analyze anime shows and its viewing population. By doing so, we want to cluster the shows and viewers, an essential step for future construction of recommendation system; additionally, we also want to create a model to predict an anime's rating based on multiple criteria.

In our study, we used K-means clustering to cluster anime and its viewing population. The shows were divided into three clusters, while the viewers were divided into two.

Additionally, we utilized multilinear regression and random forest to analyze and predict which kinds of anime attract the most people.

Data

We obtained two datasets from Kaggle (https://www.kaggle.com/CooperUnion/anime-recommendations-database). Of the two, the anime.csv file contains 12292 different anime with detailed information such as genre, type, number of episodes, and number of people who are in the anime's 'group'. The second dataset, namely the 'ratings.csv', contains information of ratings about different anime from 73.5k individuals.

Data Preprocessing

In the "anime.csv" data, we found out that some entries are not TV series, but rather, movies or music. Since the number of episodes for movies or music is always one, in order for the number of episodes to make sense, we removed the entries of movies and music. In addition, we also removed the entries with "Unknown" episodes.

For the anime rating prediction, we further removed any data with an empty rating.

For the "ratings.csv" data, we found that some ratings are -1, meaning that the users watched the show but did not rate it. We decided to remove such entries.

Anime Clustering

Our future goal is to create an intelligent anime recommendation system. To achieve our goal, we decided to divide both the anime and the views into different clusters, so that we will know what kind of shows should be recommended to what kind of users. For this model, we utilized the clustering technique separately on both the shows and the viewers.

We clustered the shows based on three metrics, the number of episodes (denoted as “episodes” below), number of community members that are in this anime's

"group" (“members”), and the rating of the show (“rating”). In essence, the data for clustering are three-dimensional vectors.

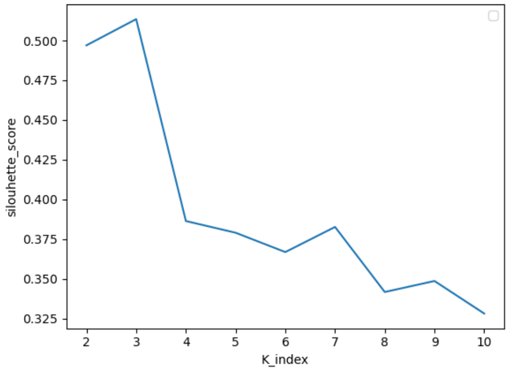
Having cleaned up the data, we did a basic survey over the data, and we found:

	Mean	Variance
episodes	35.977578	6.519568e+03
members	44635.572029	8.198475e+09
rating	6.929487	6.901854e-01

As the mean and the variance of the three metrics of anime data vary by an exceptionally large margin, we standardized the data before clustering. To select the best number of clusters, we did the silhouette analysis.

Silhouette analysis

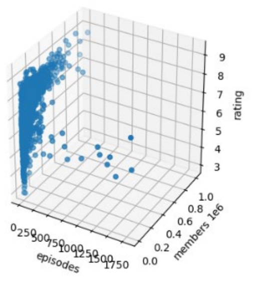
A silhouette score for a sample i measures how well sample i “matches the clustering at hand (that is, how well it has been classified)” (Rousseeuw 1987). The range for the silhouette score is between -1 and 1, and when the score for sample i is at its largest, we can say that “i is ‘well-clustered’” (Rousseeuw 1987). To get a general picture of the whole data, we calculated the average of the silhouette scores for all samples in the dataset and chose the best K based on the largest average silhouette scores for different K (Rousseeuw 1987). To implement the silhouette score analysis, we used the silhouette.score module of sklearn.metrics.

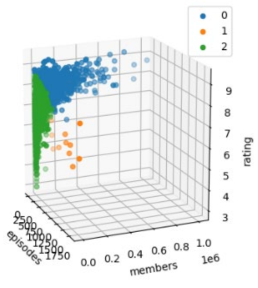


For K ranging from 2 to 10, the list of average silhouette score is: [0.4982, 0.5135, 0.3864, 0.3757, 0.3666, 0.3794, 0.3398, 0.3272, 0.3627]. As shown by both the figure and the numerical result, the silhouette score is the greatest when K=3. Therefore, we chose 3 as the number of clusters.

Clustering result

Using the K-means module from sklearn.cluster, we calculated the labels for each of the anime. Each anime is distributed into one of 3 clusters. Anime in cluster 0 have varying number of members, high ratings, and relatively fewer episodes. Similar to those in cluster 0, anime in cluster 2 also tend to have fewer episodes, but the numbers of members are less varied, and





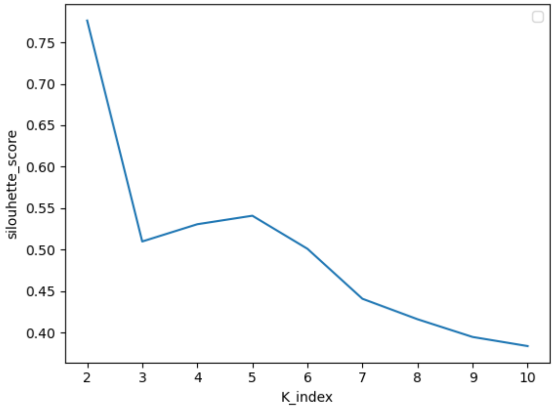
their ratings also span in a wide range. Lastly, cluster 1 tend to contain anime which have a lot of episodes, fewer members, and a comparably consistent rating.

Clustering the viewers

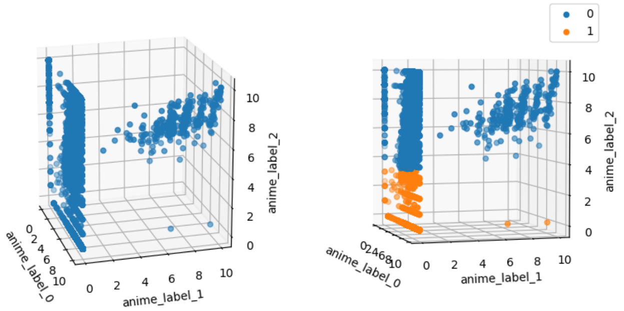
In addition to clustering the anime, we also clustered the viewers based on the average rating made by each viewer for different clusters of shows. Since the result of the user clustering depends on the three generated anime clusters, we pre-processed the user data again to better fit those clusters. Like the anime clusters, each data point for the viewers is also represented by a 3d vector

Some users did not rate anime from some clusters, so we set their average rating to those clusters to 0. Because each element in the rating is of the same dimension, ranging from 0 to 10, we chose not to standardize the data this time.

Similar to the previous model, we conducted the silhouette analysis for the ratings. The resulting highest average silhouette score comes from K = 2.:



The result of the clustering is as follows: users in cluster 0 generally rate cluster-2 anime higher, and their counterparts in cluster 1 tend to rate cluster-2 anime lower. In terms of cluster-0 anime, both groups of people rated from a very wide range. Lastly, cluster-1 anime generally received higher rating from viewers in cluster-0 then viewers from cluster-1. We think that there are possibly two reasons behind our last observation: either cluster-0 viewers prefer cluster-1 anime to a much larger extent, or cluster-1 viewers simply did not rate the shows.



After completing the two K-means clusters, we can reasonably infer a random anime fan’s preference by finding out what user cluster he/she belongs to and what kind of anime that cluster of users enjoy watching.

Anime Rating Prediction

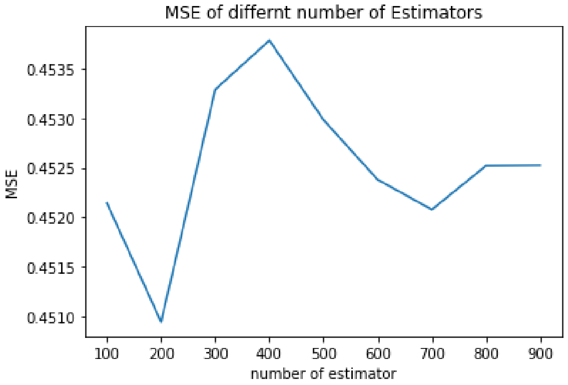
To further analyze the data, we tried to create a model that predicts the rating of an anime from its type, variability of genre (how many different genres are contained in one work), number of people in its group, and number of episodes.

For comparison, we applied both the multilinear regression model and the random forest model.

After applying our test data on the trained models, we calculated MSE (Mean Squared Error) to evaluate the quality of our models. For the multilinear regression model, we got a MSE of 0.6877. The resulting coefficients for variables: number of episodes, number of people in group, and number of genres included are 9.97e-04, 4.91e-06, and 1.71e-01 respectively, with an intercept of 5.933.

The random forest model is one we learned outside the class during our exploration for sklearn packages. Random forest utilizes a multitude of decision trees to give a result base on existing data with similar attributes. We chose to test this model because we think that anime with similar attributes (especially number of genre and membership population) can have similar ratings.

For the random forest model, we tried to different numbers of estimators from 100 to 1000. The resulting MSE graph is shown below.



Since 200 yields the lowest MSE, we picked it as the final input value; consequently, the final MSE we got from random forest model is 0.4514.

Considering that the ratings are given in a scale between 1 and 10, both MSEs are rather significant; however, we still believe that our models can give some insight into people's anime preferences and give a reasonable prediction for a new anime's rating.

Conclusion

In conclusion, we developed two machine learning models to achieve two goals. Firstly, we clustered the anime shows into three clusters and viewers into two. By doing so, we can obtain a general idea of a user's preference in anime.

In addition, we constructed two different training model to predict the possible ratings of an anime based on its type, genre, number of people in the group, and the number of episodes.

Direction for future work includes finding a way to further categorize the anime and the user for a more specific preference type and make the rating prediction model more accurate. We hope to use our knowledge to find a better machine learning model that leads to the productions of more interesting and intriguing anime.

GitHub Link

https://github.com/lkh9908/Anime_ML_Project

References

Mahendru, Khyati. 2019. *How to Determine the Optimal K for K-Means?* June 17. Accessed May 10, 2021. https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb#:~:text=on%20this%20dataset.-,The%20Elbow%20Method,becomes%20first%20starts%20to%20diminish.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics* 20: 53-65. Accessed May 10, 2021. doi:10.1016/0377-0427(87)90125-7.