Movie Recommendation System Using Clustering and Pattern Recognition Network

Muyeed Ahmed, Mir Tahsin Imtiaz, Raiyan Khan
Department of Electrical and Computer Engineering, North South University
Plot-15, Block-B, Bashundhara, Dhaka, Bangladesh
muyeed.ahmed@northsouth.edu, tahsin.imtiaz@northsouth.edu, raiyan.khan@northsouth.edu

***Abstract*— As the world is going global, people have a lot to choose from when it comes to movies. There are different genres, cultures and languages to choose from in the world of movies. This arises the issue of recommending movies to users in automated systems. So far, a decent number of works has been done in this field. But there is always room for renovation. This paper proposed a machine learning approach to recommend movies to users using K-means clustering algorithm to separate similar users and creating a neural network for each cluster.**

Keywords—Movie Recommendation, K-means, Clustering, Patternnet, Machine

I. INTRODUCTION

Because of revolution in entertainment industry, source of entertainment has been increasing rapidly in today's world. Users have to choose entertainment products from a vast amount of options which can be overwhelming for any user. As a result, recommendation systems for any product have gained popularity for every field in digital systems. As the entertainment world is booming with huge amount of data, automated recommendation systems having different approaches can be useful for recommending products to users. In this paper, we have proposed a machine learning based way to recommend movies using clustering and machine learning approaches.

The objective of this paper is to separate users using clustering algorithm in order to find users with similar taste of movies. Machine learning approaches are used to guess what rating a particular user might give to a particular movie so that this information can be used to recommend movies to viewers.

In this recommendation system, we used the publicly available data from MoveLens [1]. In the MovieLens dataset, there are two files containing information about movies and users. The movie data has been split based on their genre and later outer joined with ratings of movies in order to get user preference, average rating and consumption ratio for each genre of movies in three separate approaches. This resulted in each tuple having 18 attributes in all the approaches since there are 18 genres listed in MovieLens[1] dataset. Clustering was used to separate dissimilar users and the result was compared in the three approaches to choose the best one. Principal Component Analysis [2] (PCA) was used to decrease the dimension for a better clustering result. It was used in such a way that we get don't lose any user data after clustering. Finally, the rating was included in the last column as output column which was later on used for the neural network. We used pattern recognition network as we needed to classify inputs according to 11 target classes. We used 12000 users information for research.

II. RELATED WORK

Davidson, et al. [3] discussed about the recommendation system that YouTube uses to recommend videos to its users based on the previous activity of the users and while doing that, they have also discussed and pointed out some challenges that is faced by the system of YouTube while doing its task. They provided their experiment details and evaluation framework that they used for testing and tuning of new algorithms.

H. Chen and A. Chen [4] designed a music recommendation system. Their workflow consists of analyzing the music objects, determining the representative track, extracting six features from the track. That is how music objects are grouped. In order to understand the interests of the users, their access history is analyzed. Recommendation methods are proposed mainly based on the preferred degrees of the listeners to the music groups.

Ahmed, et al. [5] demonstrated how TV series recommendation can be challenging and different than movie recommendation. In TV series recommendation, time commitment issue needs to be analyzed Other than analyzing genre, which adds some extra work on TV series recommendation and this paper showed a way to achieve that using fuzzy systems.

Park, Hong, and Cho [6] proposed a recommendation system which is personalized where users' preference is reflected by Bayesian Networks. The parameters are learned from a dataset whereas the structure of the Bayesian Network was built by an expert. The system they proposed works by collecting context information such as location, time, and weather condition. It also analyzes user request from the mobile device to infer the most favored item so that it can provide an appropriate facility by showing it in the map.

Huang and Jeng [7] worked on audio recommendation system. Their system takes user assigned rating for songs of his/her song collection and extracts the audio signature. LBG vector quantization is used by the system to rate new audio file.

Baatarjav, Phithakkitnukoon, and Dantu [8] introduced a group recommendation system for the popular social network Facebook [9] understanding the problem users go through to find right groups. They used a combination of hierarchical clustering technique and decision tree. They worked on understanding groups by analyzing the member profiles.

978-1-5386-4649-6/18/$31.00 ©2018 IEEE

Cao and Li [10] addressed the problem of overloaded products data and discussed about the importance of efficient recommendation system to overcome this kind of problem. They suggested a fuzzy based system which works on consumer electronics keeping in mind the idea of personalized recommendation system based on special features of products and users current needs in order to retrieve optimal products. They claimed that their system showed effectiveness and feasibility in experimental results.

III. Dataset

For our system, we used the publicly available MovieLens [1] dataset. The MovieLens [1] dataset mainly has two files. The first file contains information about movies. It has movie id, movie title and its list of genre. The MovieLens [1] dataset has a movie list from 18 genres. The genre list includes genres such as "Action", "Comedy", and "Adventure" etc. The other file contains information about the users. It has multiple rows of user id, movie id and rating. Each row represents that a particular user has given a particular rating to a particular movie. These two files have been preprocessed and manipulated in order to build our system.

IV. System Design

In our system, the MovieLens [1] data has been used to calculate average rating, preference and consumption ratio of each user for all the 18 genres. These two files are outer-joined in order to calculate average rating and consumption ratio for 12000 users for 18 genres of movies. Later preference was calculated by multiplying average rating and ratio. Table 1 and table 2 shows Average rating and consumption ratio.

TABLE I. Average Rating

User Id	Action	Comedy	War
12	4.25	3.39	4.12

TABLE II. Consumption Ratio

User Id	Action	Comedy	War
12	0.35	0.46	0.21

A. Data Preprocessing

At first, the movie data is split into the genres to find their genre memberships. Then from the users file, average rating for each genre for each user is calculated. This gives us 18 attributes for each user representing 18 genres. These two files are outer-joined in order to average rating, preference and consumption ratio for 12000 users for 18 genres of movies.

B. Principal Component Analysis (PCA)

Since in our system, attribute number is really high, before clustering, principal component analysis was used to reduce the dimension from 18 to 6 after analyzing the latent graph in fig 1. Fig.1 shows the latent vs attribute number for average rating data.

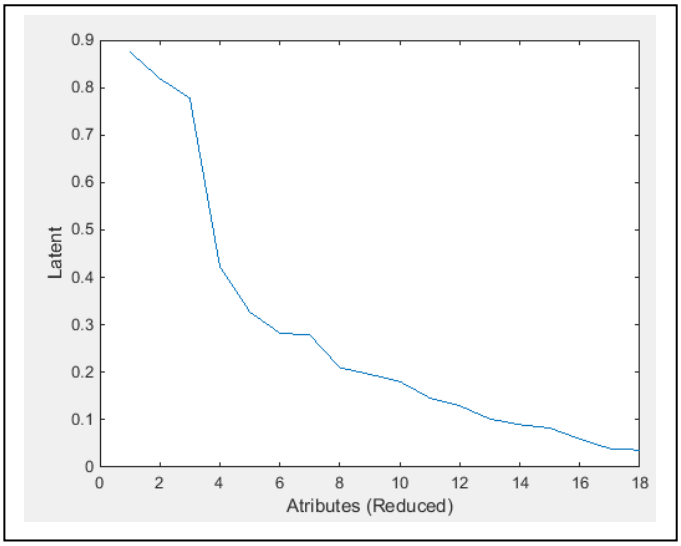

Fig. 1. Intra-Cluster Distance

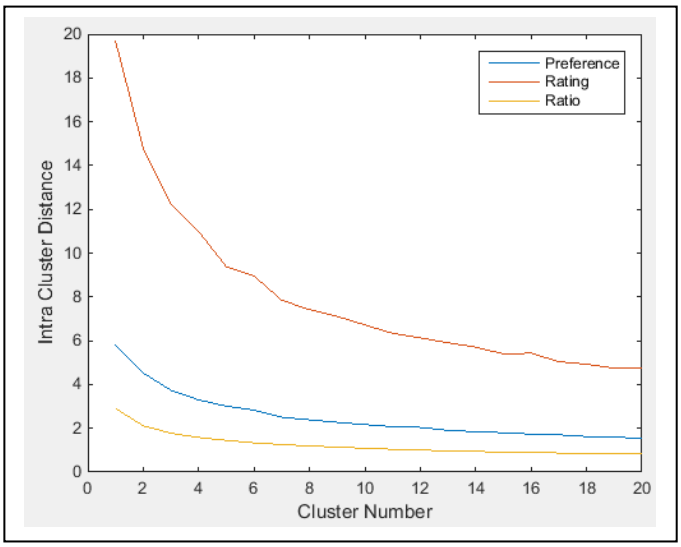

Fig. 2. Intra-Cluster Distance in Different Approaches

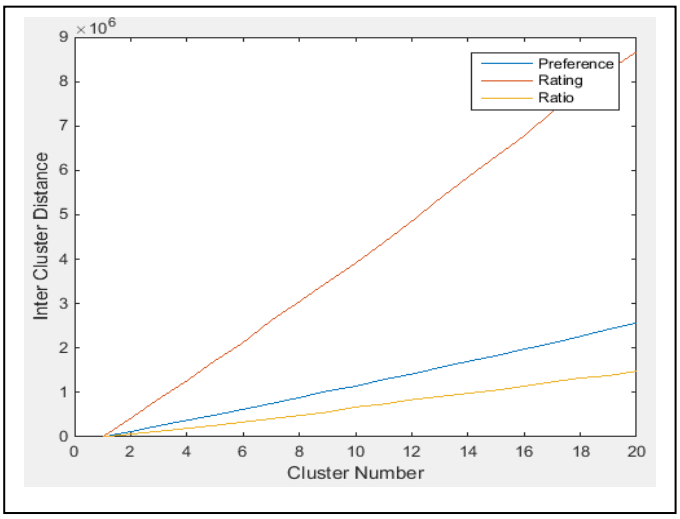

Fig. 3. Inter-Cluster Distances in Different Approaches.

C. Clustering

Different approaches have been taken to see which approach gives us the best result while clustering the users using k means clustering algorithm. Consumption ratio, user genre preference and user rating has been considered while checking the validity of clustering. Fig.2 and fig.3 shows intra cluster difference and inter clustering difference respectively for each of the approaches.

We decided to go with rating since it has a far superior inter cluster difference and a better curve for intra cluster difference. Also, after analyzing the cluster number validation graph represented in fig.4, we have decided to go with seven clusters.

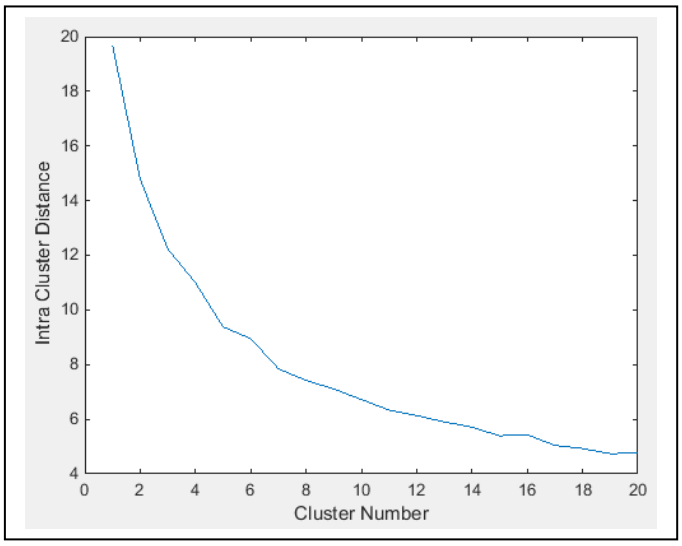

Fig. 4. Cluster Number Validation

D. Data Preprocessing for Neural Network

For the training of the neural network, we modified the original dataset differently. In the initial dataset, the data looked like table 3.

TABLE III. ACTUAL MOVIE DATA

User Id	Rating	Genre
12	4	'Action |Comedy'

This information was split and was made to a 20 column matrix where first column is user id, second column was rating, and 3-20th columns were genre memberships (0 or 1) as illustrated in table 4.

TABLE IV. SPLIT MOVIE DATA

User Id	Rating	Action	Comedy	…	History
12	4	1	1	0	0

After that, this data is inner-joined with a particular cluster and user average rating and genre membership of that particular movie is multiplied. For example, data of Table 5 will turn into data of table 6.

TABLE V. DATA BEFORE JOINING

User Id	AvgRating Action	AvgRating Adventure	…	Action	Adventure	…
12	3.15	2.33		1	0	

TABLE VI. DATA AFTER MULTIPLICATION

User Id	Action	Adventure	…	Rating
12	3.15	0		4

We took the transpose of all genre values as input and transpose of all rating values as label/target/output to build out network for particular cluster. But since we used pattern recognition network, it required us to divide our outputs/targets. For example, table 7 will become table 8.

TABLE VII. ORIGINAL LABEL

Label#1	Label#2	Label#3	Label#4	Label#5	Label#6
4	3	3.5	5	1.5	4

TABLE VIII. MODIFIED FORMAT OF LABEL FOR NETWORK

Label#1	Label#2	Label#3	Label#4	Label#5	Label#6
0	0	0	1	0	0
0	0	0	0	0	0
1	0	0	0	0	1
0	0	1	0	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	1	0
0	0	0	0	0	0
0	0	0	0	0	0

E. Neural Network

We used pattern recognition network in order to build the network, as there are more than two but limited number of target classes. For each cluster 70% of the data was used for training whereas 15% was used for validation and the other 15% was used for testing. Fig.4 shows the network of this system. Input has 18 nodes representation 18 genre.

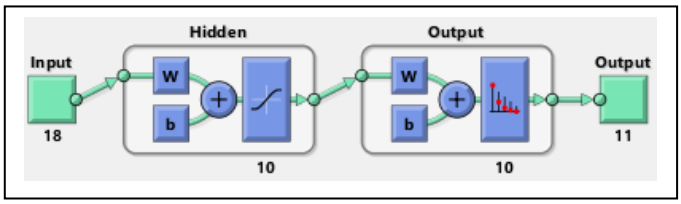

Fig. 5. Network

Output has 11 layers as it has 11 target classes. Because, a user only has 11 options from 0 to 5 to rate a movie.

V. RESULT

The result of our system showed on average 95% accuracy depending on the cluster. For example, performance graph and histogram for cluster 1 and 4 is given in fig.6, 7, 8 and 9 respectively.

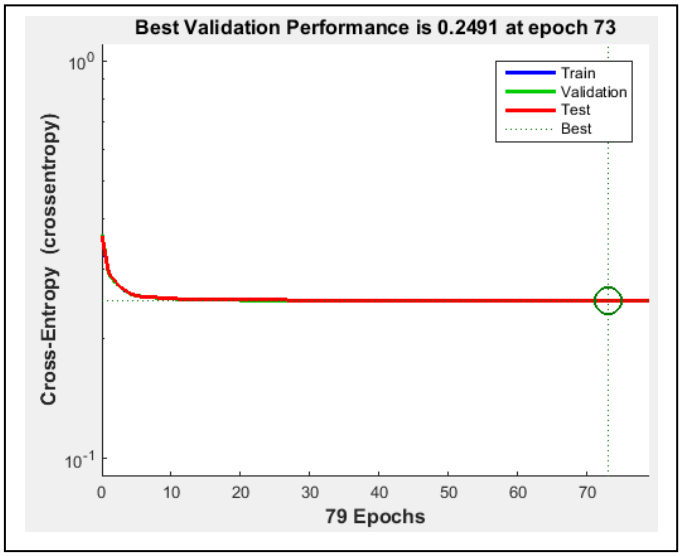

Fig. 6. Performance Graph of Cluster 1

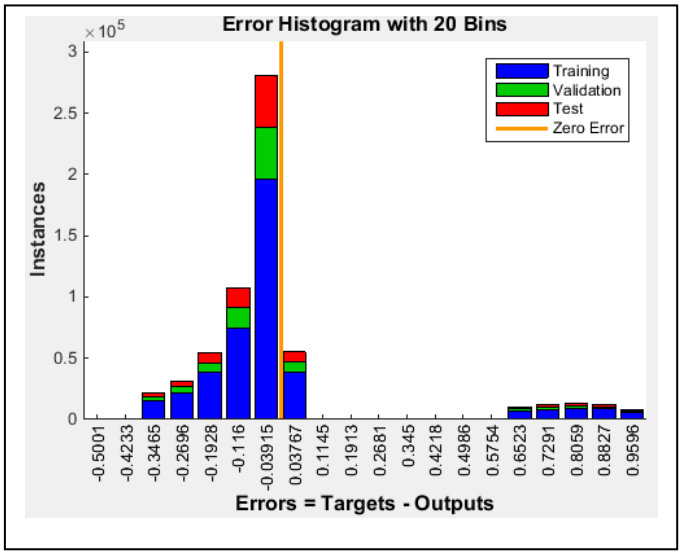

Fig. 7. Error Histogram of Cluster 1

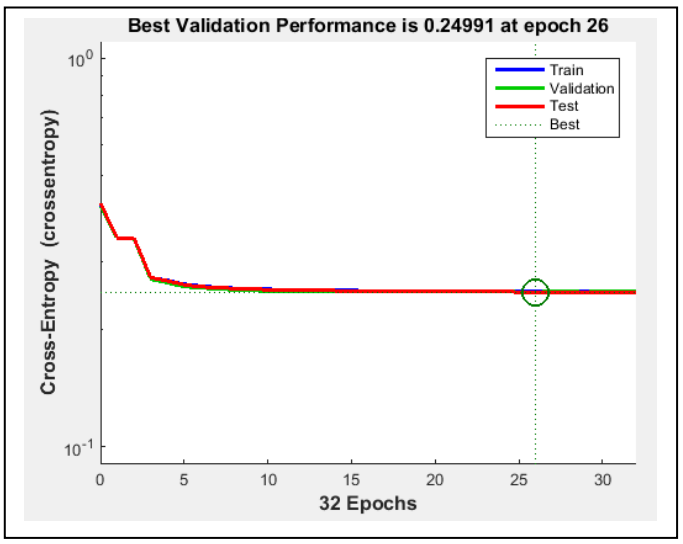

Fig. 8. Performance Graph of Cluster 4

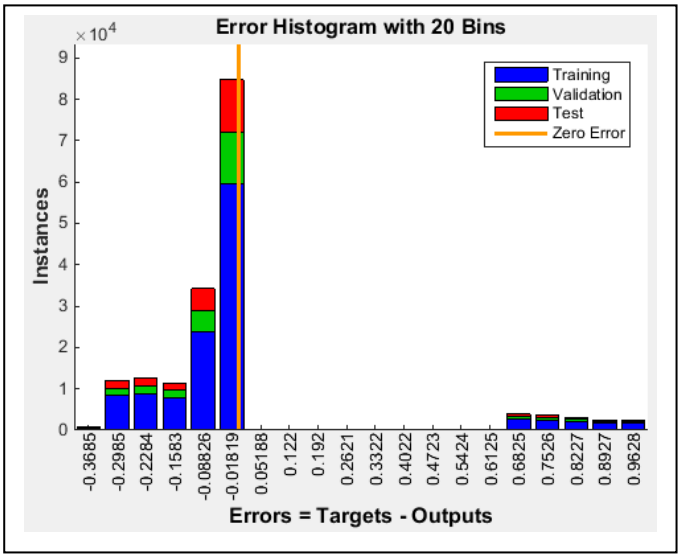

Fig. 9. Error Histogram of Cluster 4

VI. CONCLUSION

So far, a lot of work has been done in the field of recommending products like movies. But since the recommendation process of any product is not a static issue, the efficiency and accuracy can be improved over time with more and more research.

In this paper, we worked on different approaches to make movie recommendation as good as possible. User rating, user consumption ratio and user preference have been considered while building the system. K-means clustering has been used to separate users with similar taste in movies. Separate neural networks have been built to predict rating value of movies given by new user by analyzing their behavior. Depending on the cluster, our system showed 95% accuracy on average in predicting rating from new user data which can be used to

analyze which movie should be recommended to new users. This proves that our system is a valid one for prediction in the field of movies. Moreover, the dataset we used for our system has a huge user base of 12000 regular users. This ensures that, our system can deal with different types of users with diverse attitude towards movies.

REFERENCES

[1] "MovieLens." GroupLens, 18 Oct. 2016, grouplens.org/datasets/movielens/.

[2] "Principal component analysis." Wikipedia, Wikimedia Foundation, 2 Sept. 2017, en.wikipedia.org/wiki/Principal_component_analysis.

[3] Davidson, James, et al. "The YouTube Video Recommendation System."

[4] Chen, Hung-Chen, and Arbee L. P. Chen. "A music recommendation system based on music data grouping and user interests." Proceedings of the tenth international conference on Information and knowledge management - CIKM01, 2001, doi:10.1145/502585.502625.

[5] Ahmed, Muyeed, et al. "TV Series Recommendation Using Fuzzy Inference System, K-Means Clustering and Adaptive Neuro Fuzzy Inference System." 2017, pp. 1512–1519.

[6] Park, Moon-Hee, et al. "Location-Based Recommendation System Using Bayesian User"s Preference Model in Mobile Devices." Ubiquitous Intelligence and Computing Lecture Notes in Computer Science, pp. 1130–1139., doi:10.1007/978-3-540-73549-6_110.

[7] Huang, Yao-Chang, and Shyh-Kang Jenor. "An audio recommendation system based on audio signature description scheme in MPEG-7 Audio." 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), doi:10.1109/icme.2004.1394273.

[8] Baatarjav, Enkh-Amgalan, et al. "Group Recommendation System for Facebook." On the Move to Meaningful Internet Systems: OTM 2008 Workshops Lecture Notes in Computer Science, 2008, pp. 211–219., doi:10.1007/978-3-540-88875-8_41.

[9] Facebook, www.facebook.com/. Accessed 5 Sept. 2017.

[10] Cao, Yukun, and Yunfeng Li. "An intelligent fuzzy-Based recommendation system for consumer electronic products." Expert Systems with Applications, vol. 33, no. 1, 2007, pp. 230–240., doi:10.1016/j.eswa.2006.04.012.