An Approach to Developing and Implementing a Recommendation System

SIRANUSH SARGSYAN
Department of Programming and Information Technologies
YerevanState University
1, A. Manoogian, Yerevan
ARMENIA

ANNA HOVAKIMYAN
Department of Programming and Information Technologies
YerevanState University
1, A. Manoogian, Yerevan
ARMENIA

VARDITER KEROPYAN
Department of Programming and Information Technologies
YerevanState University
1, A. Manoogian, Yerevan
ARMENIA

Abstract: - The modern internet space provides the user with a huge amount of information, which is becoming more and more difficult to search. Recommendation systems are designed to provide information that best meets the needs of the user. Significant progress has been made in recent years in the development of proposal systems.

The authors of the article have developed principles that use machine learning models to solve a number of problems in different areas. Methods are proposed, by means of which, having the objective characteristics of the subject domain, the solutions to some of the required problems are implemented. One of the tasks is to determine the quality criterion of the object and to give a proposal scenario for the creation of a high-quality object or product. The proposed set of models was used to predict the IMDB Movie Quality Assessment. Model experiments yielded relatively better results than [1].

Key-Words: - Recommendation system, AI, teaching sequence prediction, classifier, regression MSE և R^2 metrics, IMDB Movie Quality Assessment

Received: June 9, 2021. Revised: April 20, 2022. Accepted: May 23, 2022. Published: June 22, 2022.

1 Introduction

The System of Recommendation in accordance with the preferences and interests of the user provides additional information.

Recommendation systems are used in various areas of human activity. Algorithms developed and implemented by the authors of the article, use a group of machine learning models GMML (A group of models of machine learning) to solve a number of practical problems in different fields. GMML enables the implementation of the following problems by having the object description of the subject domain.

1. Creating a proposal script using object attributes.
2. Determining the quality criterion of an object through its attributes.

For example, if we look at the field of medicine, in one case the patient can be viewed as an object. In this case, by classifying patients into different classes, GMML can be used to create a patient-oriented treatment scenario. On the other hand, if we look at the doctor as an object, then by classifying doctors into different classes, the GMML can determine the standard of a doctor's professional quality. Solutions to problems in that field can be found using objects in different fields using GMML.

For example, in the field of education: a well-progressed pupil, student, high-quality teacher, supervisor, etc. Identify the products that are most in-demand in the trade sector. Cultural events (concerts, movie watching, sports, etc.) identify the high-profile event as a quality characteristic. Effective performance scenarios can be defined using GMML for the above areas.

2 Description of the GMML System

GMML is a set of algorithms, programs, and services that help build a system of recommendations.

Required to use the GMML system

1. Object description of the subject domain,
2. Subject data collection,
3. Determining the teaching sequence.

An information vector structure has been developed for the object domain object, which we will call the object attribute vector VOA (Vector of Object Attributes). In the data collection process, the set of VOA vectors of the subject domain is determined. A subset of labeled VOA vectors is selected from this set, the elements of which form the training sequence. Building a sequence that teaches applied problems is often challenged by the lack of object labels. Solving the problems posed in this article through the GMML system requires two different approaches. They solve the problems of determining the quality criterion of the proposal scenario and object quality, according to the types of data collected (labeled or not).

Approach 1 (In this case we have labeled objects)

1. Model learning through a teaching sequence
2. Division of classes by classification
3. Determining the quality criterion of an object belonging to a given class
4. Creating a proposal scenario using VOA vectors
5. Determine what attributes an object must have in order to belong to a high-quality class

Approach 2 (In this case we do not have labeled objects)

1. Formatting a set of VOA vectors of a subject domain
2. Class division by clustering
3. Development of VOA vector attributes to change attribute type and quantity
4. Organize clustering according to required attributes:
5. Assessment of object quality criteria

To perform the above steps, the classification, regression, and clustering problems in the GMML system are solved.

3 Using GMML for IMDB Movies to Predict Quality Assessment

Now let's describe how to predict the success of the film before the screening, having information about it. The IMDB score determines the success of the film. The VOA vector in this subject area is the film, and one of its attributes is the director's name, year, genre, line, number of views, etc. The following tasks have been set

1. Predict its IMDB rating for the movie;
2. Having an IMDB rating for the film, give a script for the filmmakers to make the film a success in the future.

The principles of GMML modeling were used to predict the score of IMDB movies. Like having a VOA vector from a successful movie class to create a new movie, suggest a script for a new movie to be successful

Data for IMDB movies were taken from Kaggle to perform the tasks. MovieGenre.csv[5], movies_metadata.csv, keywords.csv, and credits.csv databases[6] were also used to create VOA vectors for IMDB movies, which were also taken from Kaggle. The movieGenre.csv database object attributes have been adapted to solve the problem using software tools developed in the GMML system. The data taken from the movies_metadata.csv database was scanned, and keywords.csv, and credits.csv, identified important attributes for VOA vectors to solve the problem.

In the next step of data processing, attributes were reduced, added, and changed for all VOA vectors of the subject domain. Reduction and Addition were made using Python's Pandas, NumPy, Re, Os, and other libraries. Attribute modification was done using the CascadeClassifier of the Python CV2 library. The movie poster was designed to get an extra attribute for VOA.

Because VOA vectors are labeled, a training sequence is created from them. It is divided into Train, Validation, and Test Sets. Linear Regression, Decision Tree, Random Forest, XGBoost, Support

Vector Regressor, and LightGBM machine learning models were used. The algorithms used for model training performed K-Folds cross-validation and hyperparameter adjustment.

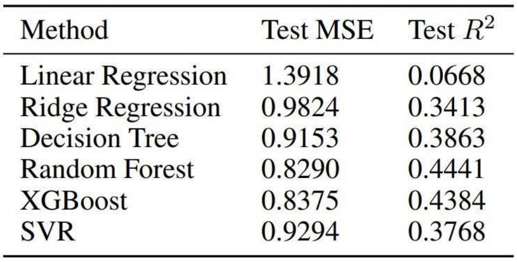
The quality of the models was assessed by MSE and R^2 metrics obtained as a result of software implementation. For each model, hyper parametric tuning was performed to improve its quality to reduce errors.

The grid search method was used for hyper parametric tuning. In the model, various hyper parametric values are given as input. For each possible combination of hyper parameter values, the method calculates the error of the loss function and selects the combination of hyper parameters with the least error. So all GMML models are configured. As a result of the tests, the quality criteria for the GMML models were obtained and the model with the least error was selected. Problems with IMDB rating and recommendation script were solved with all models, but the best results were obtained with Random Forest and LightGBM models..

3 Description of Test Results Assessment

All GMML models were tested for MSE և R2 metric values (Table 1). The analysis shows that Random Forest is the best, as its MSE is the smallest, and R2 is the largest. The use of the Random Forest model provided a better result than the Random Forest result shown in [1] (Table 1).

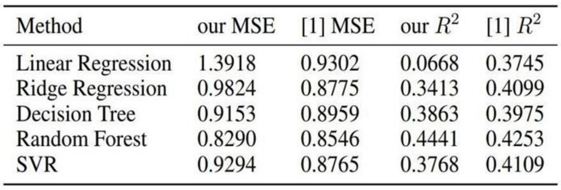
Table 1: Train Validation Test Results



Method	Test MSE	Test R^2
Linear Regression	1.3918	0.0668
Ridge Regression	0.9824	0.3413
Decision Tree	0.9153	0.3863
Random Forest	0.8290	0.4441
XGBoost	0.8375	0.4384
SVR	0.9294	0.3768

Note that in the Random Forest Regressor model, the best result is obtained only if the VOA vector attributes are preserved. Table 2 compares the quality criteria obtained in the tested models with the results obtained in [1]. As we can see, we also got a better result here in Random Forest.

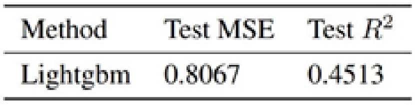
Table 2: Comparison with [1]



Method	our MSE	[1] MSE	our R^2	[1] R^2
Linear Regression	1.3918	0.9302	0.0668	0.3745
Ridge Regression	0.9824	0.8775	0.3413	0.4099
Decision Tree	0.9153	0.8959	0.3863	0.3975
Random Forest	0.8290	0.8546	0.4441	0.4253
SVR	0.9294	0.8765	0.3768	0.4109

For the implementation of the LightGBM model, software changes were made to both the quantity and type of VOA vector attributes, which significantly reduced the loss function. The resulting values for MSE և R2 were obtained (Table 3).

Table 3: LightGBM Result



Method	Test MSE	Test R^2
Lightgbm	0.8067	0.4513

4 Conclusion and Future Work

The GMML system was developed, and the concept of the VOA vector of object information was used, which is used to estimate the subject domain. To determine the IMDB score, the GMML system models for the IMDB movie data collection were tested. In order to select the best of the models, various works were performed with the object attributes (modification, reduction, addition, tuning, etc.) to improve the behavior of the model. The best models were selected from different areas to solve similar problems. Motivating is the fact that the experiments yielded relatively better results compared to [1].

As the authors of this article are related to the field of education, it was decided to use the GMML system for that field. We intend to find the best lecturer in the department, the student in the group, the required subject in the semester, etc.

Further work will be done to increase the efficiency of education at the faculty.

References:

[1] Yichen Yang et al., “Predicting Movie Rating with Multimodal Data”,

http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26260680.pdf

[2] Andreas C. Müller, Sarah Guido, *Introduction to Machine Learning with Python. A Guide for Data Scientists*,O'Reilly Media, 2016.

[3] Prateek Joshi, *Artificial Intelligence with Python*, Packt, 2017.

[4] Y. J. Lim and Y. W. Teh, Variational Bayesian approach to movie rating prediction, *Proceedings of KDD Cup and Workshop*, vol. 7, 2007, pp. 15–21.

[5] KaggleInc, Movie genre from its poster, https://www.kaggle.com/datasets/neha1703/movie-genre-from-its-poster

[6] Kaggle, The movies dataset, https://www.kaggle.com/rounakbanik/the-movies-dataset

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Siranush Sargsyan has proposed the idea to use of a group of models of machine leaning and the concept regarding the object vectors

Anna Hovakimyan has proposed the idea to approbate and estimate the models on movie datasets by preprocessing the data

Varditer Kerobyan has implemented the models in Python and provided experiments.