I.J. Information Engineering and Electronic Business, 2019, 6, 24-31
Published Online November 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2019.06.04

Feature Engineering based Approach for Prediction of Movie Ratings

Sathiya Devi S
Assistant Professor, University College of Engineering, BIT Campus, Anna University, India.
Email:sathyadevi.2008@gmail.com

Parthasarathy G
Research Scholar, Anna University, India
Email: parthasarathee.g@gmail.com

Received: 07 June 2019; Accepted: 24 September 2019; Published: 08 November 2019

***Abstract*—**The buying behavior of the consumer is grown nowadays through recommender systems. Though it recommends, still there are limitations to give a recommendation to the users. In order to address data sparsity and scalability, a hybrid approach is developed for the effective recommendation in this paper. It combines the feature engineering attributes and collaborative filtering for prediction. The proposed system implemented using supervised learning algorithms. The results empirically proved that the mean absolute error of prediction was reduced. This approach shows very promising results.

***Index Terms*—**Recommender systems, gradient boost regression, supervised learning, feature engineering

I. Introduction

Due to the large volume of information on the internet, the recommendation of products and services is essential in day to day life. Recommender System (RS) suggests products and services to the users from the information available. RS plays a vital role in promoting the sale of diverse products from different sources through online sales. The popular websites which use RS are Amazon, YouTube, Netflix, Yahoo, last.fm, IMDb and Trip advisor to promote their business. Implicit and explicit information to recommend items is used in RS for various domains like movies, music, shopping, television, books, and news [1].RS is an application which provides suggestions on items to the active user [2]. Commercial RS recommends products to increase the number of items sold and suggests an item to achieve customer satisfaction.

Among the various approaches, the Hybrid approach plays an important role to recommend items to users. Hybrid Recommendation Systems are a combination of two or more recommendation techniques. The hybrid model is made using the power from multiple type machine learning algorithms [3]. A list of methods that are commonly used in building Hybrid Recommendation Systems is (i) Weighted,(ii) Switching, (iii) Mixed (iv) Feature Combination, (v) Cascade, (vi) Feature Augmentation and (vii) Meta-level [4]. In the weighted approach, the recommendation is reached based on a mixture of predictions from different recommendation techniques. Based on the requirement, RS switches between different recommendation techniques, called Switching. In a mixed hybrid approach, a list of results from all recommendations is derived by applying various methods, which are presented as a single list. In feature combination, the results from the collaboration technique are utilized as another feature to build a content-based system over the enlarged feature set. The cascade technique combines the results from different recommendation techniques in a prioritized way. The rating or classification from first stage is used as an additional feature and the subsequent stages called feature augmentation. In Meta-level, a model generated from a recommendation technique acts as an input to the next recommendation technique in the next stages. The major challenges faced by RS are cold start, sparsity, and scalability [5]. The available algorithms and computing power are not enough when the number of existing users or items grows enormously. To address this issue, this paper proposes a recommendation method based on enhanced collaborative filtering with feature engineering. In this paper, the main contributions are

- We apply feature engineering techniques to reduce the prediction error, which increases the accuracy of the model.
- We build three different models using feature engineering techniques and evaluate their efficiency and electiveness in this environment.
- Through these contributions, we can suggest to users who are the most suitable items for them to use at a specific time, and improving the user experience in the domain of movie recommendation.

The rest of paper is organized as follows: Section 2 describes the related literature for Hybrid Recommendation system. The proposed system using collaborative filtering, for improving the accuracy is presented in section 3. Section 4 discusses the experiment and results of our system. Conclusion is arrived in section 5.

II. Related Work

The related literature for model based collaborative approach is as follows.

A traditional memory based collaborative filters recommends similar user's item to the active user. It improves the accuracy by recommending popular items instead of novel items. This work mainly proposed on growing large search space of user's profile to recommend accurate and diverse recommendations on novel items by increasing large user's profile [6].

The user's priorities are changed on items with respect to time. This problem will affect the top n recommendation to the active user. Maryam Khanian Najafabadi et al. proposed a graph based structure model to model user's priorities which give relation between users and items. This model has the user' priorities on times in between current and past time [7].

Jiangzhou Deng et al. proposed a novel k-medoids clustering algorithm based on probability distribution to address data sparsity problem. The proposed work will use the rating information based on Kullback–Leibler (KL) divergence. The top n recommendation are given based on item filtering by giving user-item rating matrix , the number of cluster centers k and KL distance measure[8]

A. Almuhaimeed et al. have proposed a hybrid recommendation approach based on the combination of the results of both content and collaborative filtering approaches [9]. These results are obtained with semantic and the hidden relationship obtained from multiple resources such as movie ontology and movie night. They claimed that this method has achieved a better result when compared with other methods and also claimed that to improve the accuracy of the recommendation by combining more semantic relations.

A hybrid model based on deep learning neural network utilizes reviews and content-based features for recommendations. These sets of contents and collaborative features are used to create, a model based on the neural network framework and predictions are estimated through this hybrid model [10]. This neural based network recommender system uses a stochastic gradient descent optimization algorithm for minimizing log loss and rating misclassification error. This hybrid model is not applicable for public datasets like Movie Lens and Amazon reviews etc.

A weighted hybrid novel approach [11] in the recommendation, addresses the scalability and improves accuracy. It represents a personalized service recommendation list and provides the most appropriate services to the users. The Agglomerative hierarchical clustering algorithm is used to create clusters. This Weighted Hybrid Recommender System combines a content-based and knowledge-based filtering to generate recommendations using the clusters.

M. A. Ghazanfar et al. proposed a generalized switching hybrid recommendation algorithms that combine machine learning classifiers with item-based collaborative filtering techniques. In general, machine learning classifiers are combined with memory-based collaborative approaches to recommend items. Instead of classifier approaches, regression approaches, feature selection algorithms, and dimensionality reduction technique may improve the accuracy [12].

A hybrid recommendation system for the e-learning environment to choose the right learning resources is proposed by Chen et al. In this hybrid approach, the item based collaborative filtering is used to find the similar items and sequential pattern mining algorithm to find the items based on the learning patterns [13].

M. Chandaka et al. proposed a hybrid approach using collaborative filtering and content-based filtering to recommend books to users. In order to find similar users, the collaborative technique is applied and the results are filtered through demographic features [14]. The users recommended by the collaborative technique are compared with the active user by content-based filtering. Slope one algorithm is used to recommend the items and Min Hash algorithm is used to find similar users.

The related literature for feature engineering is as follows.

Most machine learning performance is heavily dependent on the feature vector. Feature selection is a process to evaluate feature importance and selection of features, whereas feature engineering is the process of creation of features derived from original features [13].Feature engineering is the process of creating features using domain knowledge of the data that improve machine learning algorithms to work more efficiently [15].

Tara Rawat *et al.* proposed tools and techniques used for feature engineering with the purpose of improving classifier accuracy. This work presented the applications of feature engineering in text classification, clinical text classification and link prediction on various domains like social networks, knowledge base construction and fraud detection [16].

A Novel research demonstrated the type of engineering features are suited to the exact machine learning model type. This is accomplished by generating several datasets that are designed to fit from a particular type of engineering feature. The experiment demonstrated that at what extent the machine learning model is capable of synthesizing the needed feature on its own [17].

N.D. Patel et al. proposed a research work, how to select the best feature set from the available dataset through which the customer can classify the review. From the selected best features, a new dataset is created using various feature engineering techniques. The feature set which gives the best decision for analysis was used in

the naïve Bayesian classifier in order to get recommendations [18].

A novel approach is implemented by selecting useful feature combinations to improve performance. Hsiang-Fu Yu combined the results of student sub-teams by regularized linear regression. In this paper, the categorical features were replaced with a numerical value using the correct first attempt rate technique [19].

A. C. Bahnsen et al. created a new set of features based on analyzing the periodic behavior of the time of a transaction using the Von Mises distribution. The various credit card fraud detection models are compared in this paper and results show how the features have an impact on the output [20].

Motivated by the above literature, hybrid approaches and feature engineering techniques play a vital role in recommender systems. This paper proposes a novel hybrid approach based on collaborative filtering with engineering features, which explained in the next section.

III. Proposed System

Collaborative filtering recommends items, based on the ratings given by users and it does not consider item related and user related features for the recommendation. This proposed hybrid approach combines the item and user related features with ratings. It has five steps: (A) Preprocessing, (B) Attribute Selection, (C) Feature Engineering, (D) Model Creation and(E)Prediction and Evaluation. The proposed framework is represented in Fig. 1. In this approach, the preprocessing is performed by data integration, filling missing values and label encoding. After preprocessing, the subset of features is selected with Wrapper based approach using the genetic algorithm and linear regression. Then, four techniques are applied in order to modify the existing features obtained in the previous step based on the feature engineering approach. With these features, the prediction model is constructed with an ensemble technique and the recommendation is performed. Each step in this proposed hybrid approach is explained below.

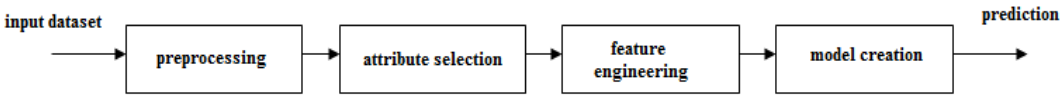

Fig. 1. Proposed Framework

A. Preprocessing

Preprocessing is used to remove inconsistency and noise in the data since it affects the accuracy of the result. These barriers in the dataset should be handled by various preprocessing techniques like data cleaning, data integration, data reduction and data transformations [21]. In the proposed approach, the preprocessing is performed to fill missing values, integrating attributes from files in the referred dataset and encoding the string datatype incompatible formats. For example, consider the dataset presented in Table 1 consisting of three attributes such as (i) User id, (ii) Age and (iii) Occupation. In this dataset, the user id is not interesting and is ignored whereas, the age is continuous and occupation is categorical attributes.

In our experiment, the movie lens 100K dataset is chosen and to be preprocessed. Using data integration preprocessing technique, all the 31 features from various data files are integrated which contain rating, user related and item related. The missing values are taken as zero and the label encoding scheme is applied for all the attributes to make as the same data type. Label encoder can be used to normalize labels and also be used to transform non-numerical labels into numerical labels [22].

To understand the label encoding, take an example dataset in Table1. It has three features user_id, age, and occupation in which occupation has non-numerical labels. Label encoding is applied in the occupation feature and the result is expressed in Table 2. Here, occupation takes the numerical value as 1, the librarian as 2 and the student as 3. In the same way, all the features are modified in order to get the same type. There are a few related features that are not useful in the prediction process. The selection process of related features to predict the rating will be discussed in the next section.

Table 1. Sample dataset before Label Encoding

User_id	Age	Occupation
691	34	educator
704	54	librarian
705	21	student
706	23	student
707	56	librarian

Table 2. Sample dataset after Label Encoding

User_id	Age	Occupation
691	34	1
704	54	2
705	21	3
706	23	3
707	56	2

B. Attribute Selection

The models usually give a fast response when using a low dimensionality dataset, because training time decreases exponentially. Models are having a risk of overfitting with an increasing number of features. There

are three methods, (i) Filter Methods,(ii) Wrapper Methods, (iii) Embedded Methods. Filter Methods considers the relationship between features and the target variable to compute the importance of features (LDA, ANOVA, and Chi-Square). Wrapper Methods generate models with a subset of features and gauge their model performances (forward selection, backward elimination, and recursive feature elimination). The embedded method is a combination of filter and wrapper methods. It is implemented by algorithms that have their own built-in feature selection methods (LASSO, RIDGE) [23]. In this proposed approach, the wrapper method is chosen to select a subset of features from the large dataset. Genetic Algorithms one of the wrapper methods based on heuristic solution search to get the optimized solution, originally motivated by the Darwinian principle of evolution through selection [24]. The various steps of the genetic algorithm are mentioned in Fig. 2.

Step 1: Number of Chromosomes, generation and mutation rate and cross over rate

Step 2: number of population, initialization of chromosome with a random value

Step 3: Process steps 4-7 until the number of generation is met setup

Step 4: Evaluate fitness value using objective functions

Step 5: Chromosome selection

Step 6: crossover

Step 7: Mutation

Step 8: Solution (Best Chromosomes)

Fig. 2. Genetic Algorithm

In the proposed approach, the attributes are considered as chromosomes and the binary encoding is applied on the chromosomes at random. The process of encoding is shown below: For example in Table 3. Chromosome encoding is shown. Each row in the table is treated as a chromosome and each attribute is treated as a gene.

Table 3. Chromosome Encoding

User_id	Action	Comedy	Romance	Animation
1	1	0	0	1
1	0	1	0	0
1	0	0	1	1

GA has chromosome encoding, fitness, selection, recombination, and evolution. The preprocessed data is having different attributes related to user and item related and each attribute treated as chromosome encoding. An accuracy score of linear regression is taken as a Fitness function in this work. Tournament Selection initially selects two chromosomes with higher probability and then chooses the one which has the highest fitness. Recombination is performed using one point cross over. After recombination, resultant chromosomes are passed into the successor population. The processes of selection and recombination are then iterated until a complete successor population is produced. At that point, the successor population becomes a new source population, i.e. the next generation.

Table 4. Parameters used in Genetic Algorithm

Parameter Name	Type
Number of Chromosomes	31
Generation	50
Mutation rate	.05
Cross over	One point

The parameters used in this genetic algorithm process are depicted in Table 4. The output of the attribute selection, which has a reduced dataset, given to the feature engineering process.

C. Feature Engineering

Feature engineering is defined as the process of using domain knowledge of data to create features that make machine learning algorithms work more efficiently [25]. Feature engineering is a technique used to modify the existing features or combining two or more relevant features or creating a new feature from the existing features. This feature engineering strategy plays a vital role in the prediction of the user's rating. Mathematical techniques like count, power, square root and polynomial are applied. To perform feature engineering, the above said mathematical techniques are applied in our selected subset of features. The algorithm for our proposed algorithm is shown in Algorithm 1.

Algorithm 1 Proposed Prediction Model

$Let\ u_f, i_f\ \forall\ u$ be the user and item features for all users
1: **begin**
2: Integrate u_f andi_f with rating as dataset F
3: $\boldsymbol{if}(F\ !=numeric)$
4:encode(F)
5: $\boldsymbol{else}$
6: read(F)
7: apply feature selection in {F}
8: $f_s \subseteq \{F\}$
9: Apply feature engineering in $\{f_s\}$
10: $f_e = function(f_s)$
11: Prediction
12: **end**

Table 5. Notations used in Proposed Algorithm

Notation	**Description**
uf	User related features
if	Item related features
F	Total features
fs	Selected features
fe	Feature engineering
rp	Predicted rating

This proposed algorithm reads the user (u_f) and item(i_f) related features for all users(u) from the dataset. It combines all the features a single dataset(F) using data integration. Next, it will check the data type of each attribute and typecast if it is not the same type. By applying the feature selection algorithm on F in order to get the Subset fs. Feature Engineering is the next process which is derived from fs and the resultant as *fe*. The various notations used in the above algorithm are mentioned in Table 5.

In this work, feature engineering is done using count, power, square root and polynomial. In feature engineering using count measure, a threshold value is fixed and each value in the attribute is checked. It replaces the value by 1 if the condition is satisfied otherwise 0. Each attribute's value is multiplied with the same and the result is stored as a new attribute in feature engineering based on power measure. Feature engineering based square root is the same as power instead of taking power, it will take the square root of the attribute value. The second-degree polynomial is used to perform the feature engineering in all the attributes.

The Feature engineering technique using count is explained below. The following Table 6 shows the sample data before applying feature engineering. It has two features user_id and age, taken from the movie lens dataset. In count based feature engineering, a threshold level is selected for each feature and it is checked with the existing value. Once the threshold condition is satisfied, it puts the entry as 1, otherwise 0. For our experiment, we have taken the mean value as the threshold value. Similarly, other techniques are also applied in the available features. Table 6 shows the sample dataset after feature engineering.

Table 6. Sample dataset before FE

user_id	age
46	27
47	53
48	45
49	23
50	21

Table 6. Sample dataset after FE

user_id	age
46	0
47	1
48	1
49	0
50	1

After completing the feature engineering techniques in the selected subset of features, the resultant dataset is given to model creation which will be discussed in the next section.

D. Model and Prediction

Collaborative Filtering is mainly classified into memory and model- based. In memo-y based collaborative filtering, scalability is a big issue. Among the various supervised learning algorithms, classification and regression models are important types, where we know the target variable. GNB is one of the simplest classification algorithms, which assigns the label of the class that maximizes the posterior probability of each sample which is chosen for our experiment [31]. This algorithm predicts the dependent variable using the independent class variables. It is a probabilistic classifier that uses the properties of Bayes theorem assuming the strong independence between the features shown in equation 1[26].

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{1}$$

where

$P(A/B)$ - posterior probability
$P(B/A)$ - prior probability
$P(A)$ - maximum likelihood
$P(B)$ - evidence

Next, the linear regression model is chosen for its simplicity and used for continuous variables. In our experiment, multiple linear regression is selected to predict the ratings [28]. Regression is a statistical technique, used to formulate the linear relationship between independent and dependent variables shown in equation 2.

$$Y = a + bX \tag{2}$$

Linear regression (LR) is a way to model the relationship between two variables. Here, Y is the dependent variable, X is the independent variable, b is the slope of the line and a is the intercept.

The important algorithm comes under ensemble learning, which is viewed as a collection of individual systems in order to obtain stronger generalization ability than the individual systems [27]. There are two types of ensemble methods; (i) Gradient bagging (ii) Gradient Boosting. Gradient bagging regression creates different decision trees and takes the average of different models as a result. In Gradient Boosting regression, the decision tree solution is given as input of the next decision tree. Gradient boosting involves three elements, a loss function to be optimized, a weak learner to make predictions and an additive model to add weak learners to minimize the loss function. Decision trees are used as the weak learner in gradient boosting and trees are added one at a time, and existing trees in the model are not changed. The developed model is evaluated in the next section.

E. Performance Measures

The developed models are evaluated with the available prediction measures. Different prediction measures are available such as mean absolute error, mean squared error, root mean squared error and normalized root mean squared error. Among the prediction error metrics, the mean absolute error is the common technique which is represented in equation 3[28]. The mean absolute is the difference between the prediction of a rating of user u on item i($p_{u,i}$) and the real rating of user u on item i ($r_{u,i}$).

$$MAE = \frac{1}{N}\sum_{u,i}\left|p_{u,i} - r_{u,i}\right| \quad (3)$$

In our work, the above prediction measures and recommendation measures are used. The experiment details and results are discussed in the next section.

IV. Experiments and Result

The proposed method is tested with Movie Lens 100K [29] dataset provided by the Group Lens Project. A test set up is created using the Windows operating system and Python3.5 software. In the movie lens dataset consists of 943 users and 1642 movies. Each user at least rated 20 movies and the ratings are between 1 to 5. This dataset has 14 files among which this proposed approach takes three files namely (i) user (ii) item and (iii) rating. These files are integrated into one file consisting of 31 attributes for experiment and evaluation. The experiment starts with the preprocessing of the dataset. The integrated data set contains sparsity (93.7%) [32] and inconsistent attributes. In the proposed approach preprocessing is performed by filling of missing value by zero and resolve inconsistency by encoding. The label encoding schemes are applied in order to make the same type of features which was discussed in section 3. A genetic algorithm is applied in the preprocessed data. Subsets of 19 attributes are selected among 31 attributes in order to reduce the large dimension of the dataset.

Feature engineering techniques explained in section 3 are applied in the selected features and modified. The output of the feature engineered dataset is divided into training and test set in the 80/20 ratio using a 10 fold cross-validation technique. At first Gaussian naïve Bayesian model(GNB) is trained and results are predicted. Similarly, the above experiment procedure is repeated for the linear regression model (LR) and gradient boost regression tree model(GBT). Fig. 3 shows the MAE values of three models in the initial dataset. Among these three models, the ensemble method gives the minimum error rate. In the second stage of the experiment, the feature selection technique is applied and the same experimental procedure followed. Fig. 4 shows the MAE values after the feature selection technique. The results show that the ensemble method gives better performance when compared with other models.

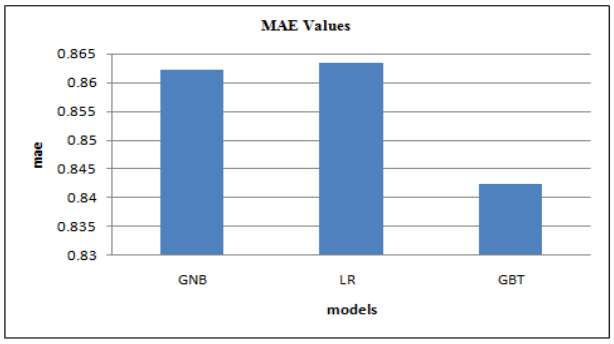

Fig. 3. Comparison of MAE values before feature selection

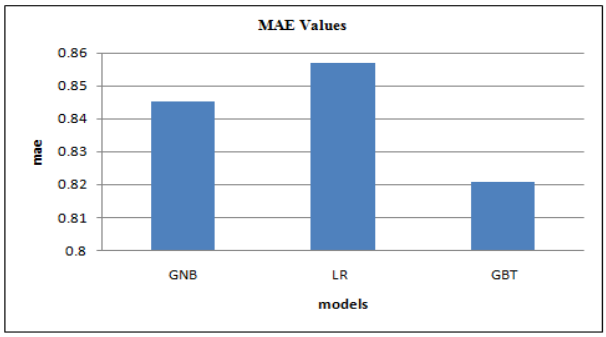

Fig. 4. Comparison of MAE values after feature selection

The prediction measure after applying featuring engineering techniques are given in Table 7. It shows that the ensemble methods give better performance in all the feature engineering techniques. Among the different feature engineering techniques, polynomial-based feature engineering performs better.

Table 7. Prediction measures after feature engineering

Feature Engineering	MAE		
	GNB	LR	GBT
Count	0.9261	0.9034	0.8910
Power	0.8914	0.9018	0.8477
Square root	0.9133	0.8926	0.8486
Polynomial	0.8978	0.8964	0.8472

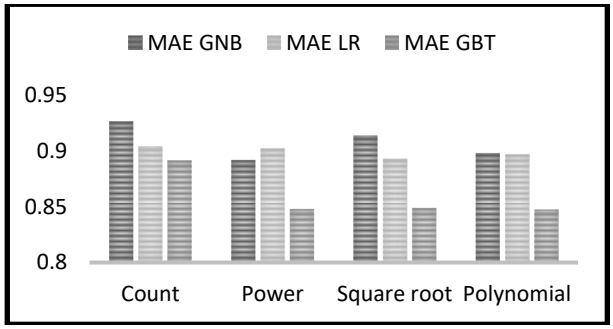

Fig. 5. Comparison MAE values of Different Models after Feature Engineering

Fig. 5 shows the three models of performance after applying selected feature engineering techniques. The gradient boost regression techniques which give 0.8472 for polynomial based feature engineering which is 5.06%

improvement while comparing with Gaussian naïve Bayesian model results. A 4.9% improvement in gradient boost regression tree while comparing with linear regression. Finally, we have obtained low error values for the gradient boost regression ensemble method. The results empirically show that the gradient boost regression tree model with feature engineering yields better results when comparing with Gaussian naïve Bayesian classification and linear regression techniques. Our proposed feature engineering technique with a gradient boost regression tree yields a 2.6% decrease in prediction error value (MAE) while comparing with the existing recommender systems [30].

V. Conclusion

We have implemented a recommendation model based on hybrid features, in which user-related and item related features are combined with ratings given by the user. The results show that among the chosen models, the gradient boost regression model outperformed when compared with the Gaussian naïve Bayesian, linear regression model and the existing work. The next finding is the feature engineering techniques improve the model performance which is indicated in our test results. The combination of gradient boost regression model and feature engineering technique gives a 2.6% (MAE) decrease in mean absolute error while comparing with an existing model. The hybridization of density-based clustering and ensemble method which uses gradient boosting is our proposed future work, which addresses the scalability issue.

References

[1] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierre, "Recommender systems survey" ,*Journal of Knowledge Based Systems*, 2013, 103-132.

[2] Francesco Ricci, Lior Rokach ,Bracha Shapira and Paul B. Kantor, "Recommender Systems Handbook", *Springer*, e-ISBN 978-0-387-85820-3,2011, 1-35.

[3] C.C Aggarwal," Recommender Systems: The Textbook", Springer International Publishing Switzerland 2016.

[4] Robin Burke,"Hybrid Recommender Systems: Survey and Experiments",*User Modelinga nd User-Adapted Interaction* 12: 331^370, 2002..Burke, R. User Model User-Adap Inter (2002) 12: 331. https://doi.org/10.1023/A:1021240730564

[5] Xiaoyuan Su and Taghi M. Khoshgoftaar "A Survey of Collaborative Filtering Techniques". *Hindawi Publishing Corporation,Advances in Artificial Intelligence* Volume 2009, Article ID 421425, 19 pages

[6] Nour El Islem Karabadji, Samia Beldjoudi, Hassina Seridi, Sabeur Aridhi, Wajdi Dhifli, Improving Memory-Based User Collaborative Filtering with Evolutionary Multi-Objective Optimization, *Expert Systems With Applications* (2018), doi: 10.1016/j.eswa.2018.01.015

[7] Maryam Khanian Najafabadi, Azlinah Mohamed, Choo Wou Onn, "An impact of time and item influencer in collaborative filtering recommendations using graph-based model", *Journal of Information and Processing and Management*, 2019,pp 526-540

[8] J. Deng, J. Guo and Y. Wang, A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.03.009

[9] Abdullah Almuhaimeed and Maria Fasli, "A Hybrid Semantic Method for Enhancing Movie Recommendations", *IEEE Transactions* on ,2017

[10] Tulasi K. Paradarami, Nathaniel D. Bastian, Jennifer L. Wightman, A Hybrid Recommender System Using Artificial Neural Networks, Expert Systems With Applications (2017), doi:10.1016/j.eswa.2017.04.046

[11] Avadhut D.Wagavkar and S.S.Vairagar, "Weighted Hybrid Approach in Recommendation Method", *International Journal of Computer Science Trends and Technology (IJCST)* – Volume 5 Issue 2, Mar – Apr 2017, 5 pages

[12] Mustansar Ali Ghazanfar and Adam Prügel-Bennett, "Building Switching Hybrid Recommender System Using Machine Learning Classifiers and Collaborative Filtering", *IAENG International Journal of Computer Science*, 37:3, IJCS_37_3_09

[13] Chen, W., Niu, Z., Zhao, X. et al., World Wide Web (2014) 17: 271. https://doi.org/10.1007/s11280-012-0187-z

[14] Manisha Chandaka, Sheetal Giraseb, Debajyoti Mukhopadhyayc, "Introducing Hybrid Technique for Optimization of Book Recommender System", *Procedia Computer Science* 45 (2015) 23 – 31

[15] Felipe F. Bocca, Luiz Henrique Antunes Rodrigues, "The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling", *Journal of Computers and Electronics in Agriculture* 128 (2016) 67–76.

[16] Chuan Zhang, Liwei Cao, Alessandro Romagnoli, On the feature engineering of building energy data mining, <![CDATA[Sustainable Cities and Society]]>(2018), https://doi.org/10.1016/j.scs.2018.02.016

[17] Tara Rawat and Vineeta Khemchandani, "Feature Engineering (FE) Tools and Techniques for Better Classification Performance", *International Journal of Innovations in Engineering and Technology (IJIET)*,http://dx.doi.org/10.21172/ijiet.82.024,2017.

[18] Jeff Heaton, "An Empirical Analysis of Feature Engineering for Predictive Modeling'arXiv:1701.07852v1 [cs.LG] 26 Jan 2017.

[19] Nikita D. Patel , Chetana Chand, "Selecting Best Features Using Combined Approach in POS Tagging for Sentiment Analysis", *IJCSMC*, Vol. 3, Issue. 3, March 2014, 425 – 430

[20] Hsiang-Fu Yu, "Feature Engineering and Classifier Ensemble for KDD Cup 2010",JMLR: Workshop and Conference Proceedings 1: 1-16

[21] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic,Björn Ottersten,"Feature engineering strategies forcre ditcard fraud detection",Expert Systems With Applications 51(2016)134–142

[22] JiaweiHan Micheline Kamber ,JianPei, "Data Preprocessing-Data Mining (Third Edition) The Morgan Kaufmann Series in Data Management Systems 2012, 83-124

[23] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, (2011), 2825-2830.

[24] Senthil Kumar P., Daphne Lopez, "A Review on Feature Selection Methods for High Dimensional Data", *International Journal of Engineering and Technology*,e-ISSN : 0975-4024,1-4

[25] John McCall, "Genetic algorithms for modelling and optimisation", *Journal of Computational and Applied Mathematics* 184 (2005) 205–222
[26] Chuan Zhang, Liwei Cao, Alessandro Romagnoli, On the feature engineering of building energy data mining, Journal of Sustainable Cities and Society,(2018), https://doi.org/10.1016/j.scs.2018.02.016
[27] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques', Procedia Computer Science 57 (2015) 821 – 829
[28] Gulden Kaya Uyanik,Nese Guler,"A study on Multiple linear regression analysis",Procedia - Social and Behavioral Sciences 106 (2013) 234 – 240
[29] Tong Xiao , Jingbo Zhu , Tongran Liu, "Bagging and Boosting statistical machine translation systems",*Journal of Artificial Intelligence* 195 (2013) 496–527
[30] J. Bobadilla , F. Ortega, A. Hernando, A. Gutiérrez, "Recommender systems survey",*Knowledge-Based Systems* 46 (2013) 109–132
[31] https://grouplens.org/datasets/movielens/100k/last accessed on 12.12.2017
[32] Yousef Kilani ,Ahmed Fawzi Otoom ,Ayoub Alsarhan and Manal Almaayah, "Genetic Algorithms-Based Hybrid Recommender System of Matrix Factorization and Neighborhood-Based",*Journal of Computational Science (2018)*, https://doi.org/10.1016/j.jocs.2018.08.007
[33] Marlis Ontivero-Ortega a, Agustin Lage-Castellanos a,c, Giancarlo Valente c, Rainer Goebel c, Mitchell Valdes-Sosa, "Fast Gaussian Naïve Bayes for searchlight classification analysis",Journal of NeuroImage,2017, 1-9
[34] Feng J, Feng X, Zhang N, Peng J (2018) An improved collaborative filtering method based on similarity. PLoS ONE13(9):e0204003. https://doi.org/10.1371/journal.pone.0204003

Authors' Profiles

Dr. S.Sathiya Devi Ph.D., working as Assistant Professor in University College of Engineering, BIT Campus,Anna University,India. Having vast experience in the field of Computer Science and published many research articles

Mr.G.Parthasarathy M.E., doing his research in Anna University Chennai in the field of Information Communication and Engineering

How to cite this paper: Sathiya Devi S, Parthasarathy G, "Feature Engineering based Approach for Prediction of Movie Ratings", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.11, No.6, pp. 24-31, 2019. DOI: 10.5815/ijieeb.2019.06.04