International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13349-13354

Feature Engineering Incorporated Enhanced Hybrid Model based Collaborative Filtering

K.Reka[1], T.N.Ravi[2]

[1]Department of Computer Science, Cauvery College for Women, Trichy, Tamilnadu, India.

[2]Department of Computer Science, Periyar E.V.R College, Trichy, Tamilnadu, India.

Abstract

Information enormity has led to the requirement of an automated model to provide recommendations to the users. This paper presents a model based collaborative filtering approach to provide appropriate recommendations to the users. The prediction model has been hybridized with multiple learning algorithms such that shortcomings from one model can be overcome by the other model. An aggregation based combiner enables to effectively combine the results for the final predictions. Feature engineering is incorporated into the proposed model such that the training data for the hybrid model is composed not only of the normal features but also the engineered features. These engineered features aid in effectively incorporating hidden patterns into the prediction process. Experimental evaluations were performed with Movie lens data and comparisons were performed with state-of-the-art models. Results indicate improved performances with respect to of MAE and RMSE.

Keywords: Collaborative Filtering; Hybrid prediction model; Decision Tree; Stochastic Gradient Descent; Feature Engineering

INTRODUCTION

Identifying information from huge amounts of data is one of the major issues faced by the online community. This has been the result of the huge growth levels associated with information available online. This has made information search a time consuming and tedious process. The process has moved beyond human cognition, hence artificial intelligence based systems are on the raise in this domain to aid humans in making better decisions. Such AI based models are called recommendation systems, which analyses the available data and the user profiles to provide item recommendations to the users depending on their interests [1, 2]. Such models are highly valuable in domains like e-commerce, where a huge collection of items are to be analyzed to provide recommendations to a much larger collection of users. The major necessity for such systems is that too many choices are available for a single user and the users tend to miss their important choices due to the highly noisy data. Although such models were developed for product recommendations, they are currently more prominent in other varied domains. This includes tourism [3], business decision making processes [4], forecasting [5], etc.

Collaborative Filtering(CF) is one of the major models used for such operations. Collaborative filtering systems are information retrieval models, which selects information based on similarity with the items the same user has liked in the past. Collaborative filtering systems are of two major categories; memory based systems [6] and model based systems [7]. Memory based system uses a heuristic based modelling architecture, where information similarity is identified by methods like correlation analysis and vector similarities. These models tend to be highly accurate, however, they require the entire correlation matrix to be available in memory. Hence the memory complexity of these methods are huge. Model based collaborative filtering approaches utilizes machine learning models to provide recommendations. These models have gained huge prominence due to the increase in the number of users and items available for analysis.

This paper presents a model based collaborative filtering approach that utilizes a hybrid model to perform the recommendations. Feature engineering has been incorporated into the model to uncover hidden patterns in the transaction data.

RELATED WORKS

Recommender systems have been a major requirement for users since the advent of e-commerce architecture. Since then several recommendation systems have been developed. This section discusses some of the most recent recommender system architectures.

A genre based recommender system was proposed by Fremal et al. in [8]. This model uses a metadata based clustering model to provide effective correlations for effective recommendations. Clustering is based on categories or genres.

Since a single item can correspond to multiple categories, this model also enables an item to be placed in multiple category slots for effective evaluation. A heuristic based recommender system was proposed by Katarya et al. in [9]. This model utilizes Cuckoo search to perform recommendations. This is also a clustering based approach. Optimization of the clustering process is performed using the Cuckoo search algorithm. Another similar evolutionary based approach was proposed by Silva et al. in [10]. This is an evolutionary based combination approach, aims at combining results from multiple recommenders to provide optimal best results. A cascaded hybrid recommendation model, combined with one-class classification model was proposed by Lampropoulos et al. in [11]. The problem of recommendation is divided into two levels and each level imposes a classifier model to perform predictions. Another clustering based recommender model for multi domain operation was proposed by Liu et al. in [12]. An ANN based hybrid recommender was proposed by Paradarami et al. in [13]. Other recent recommenders include ontology based models by Nilashi et al. [14] and a social evolutionary based model by Pascoal et al. [15].

HYBRID DECISION TREE STOCHASTIC GRADIENT (HDTSG) MODEL BASED ON ENGINEERED FEATURES

Collaborative filtering operates by identifying the user's interests to perform recommendations. However, the available attributes might provide limited scope for identifying user's interests. The proposed model solves this issue by extracting features to create engineered additional features to aid in effective prediction process. The proposed model consists of three major phases namely; modeling the training data, feature engineering attributes for the training data and the prediction phase. Algorithm of the proposed model is shown below.

Training Data Modelling

Recommendation systems are developed to provide predictions for users of a system involving several items and transaction records related to their purchase or selection patterns. However, this is not a group prediction process, hence cannot be performed for the entire set of users. Predictions or recommendations are to be performed individually for each user. Every user is different and so is their requirements. Hence in-order to effectively identify a user individually, the proposed model builds a user profile. The user profile contains details about the user's interests. Further, the proposed model also incorporates profile data from similar users, in-order to avoid the cold start problem. Further, feature engineering is built into the architecture to enhance the prediction process.

Pseudo code

1. ***Modelling the training data***
   - ***1.1 User based item selection from transaction repository***
   - ***1.2 Item based similar user identification***
   - ***1.3 Combined item grouping to create the final item list***
2. ***Feature Engineering attributes for the training data***
   - ***2.1 Derive independent category for each item in the item list***
   - ***2.2 Generate a category matrix to identify categories pertaining to each item***
   - ***2.3 Identify frequency level of single category items and category pairs/groups pertaining to items***
   - ***2.4 Threshold based category and category group selection***
   - ***2.5 Training matrix creation using the filtered category values***
3. ***Prediction Phase***
   - ***3.1 Prediction using Decision Tree and Stochastic Gradient Descent***
   - ***3.2 Prediction averaging to identify the final prediction***

User based Item Selection

The initial process in preparing the training data is to identify items or products the user under analysis is interested. This can be identified by using the transaction data to determine the items of interest. The transaction data is analyzed and transactions pertaining to the current user are filtered. This data is further fused with the user ratings. Not all items contained in the user transaction data might be of interesting to the customer. Sometimes, the user might even have provided low ratings to an item indicating their lack of interest towards the item. Hence elimination of such items are mandatory to provide effective predictions. This incorporates an initial level filter to avoid unnecessary data for the training process. Selection of items of highest interest to the user forms the Level 1 data for the model.

Similar User based Item Incorporation

The Level 1 data exhibits the associations the customer exhibits towards the items selected. However, sometimes this data might not be sufficient for the training model. The performance of a machine learning model is directly proportional to the quality of data provided to it. Data insufficiency might deteriorate the performance of the model. Data insufficiency is usually handled by oversampling. Oversampling generates additional records such that quantitative data insufficiency is handled. However, the

created records depict the initial records, hence creating reinforcement in the data. In order to avoid this, the proposed model identifies data selected from similar users, rather than moving towards data replication.

Users similar to the current user are selected for data addition. User similarity is identified by finding users who provided similar ratings to the items contained in the Level 1 data. Items involved in transactions performed by the similar users are retrieved to obtain the Level 2 data. The Level 1 data has strong and direct implications to the current user, while the Level 2 data has moderate and indirect implications to the current user. Both Level 1 and Level 2 data are integrated to form the training data.

Enhanced Feature Identification using Feature Engineering

The training data is composed of items interesting to the current user and user's similar to the current user. The rating levels pertaining to the items contained in the training data are integrated to form the final training data. However, the item-rating data alone is not sufficient for training the model. The properties pertaining to the items exhibit the type of the item. Hence item properties are to be incorporated into the training data. Item properties pertains to the category or sub-category corresponding to the item. However, a single item need not necessarily belong to one single category. It can also belong to several categories. HDTSG model engineers features based on the categories and category combinations.

Single and multiple categories corresponding to all the items are identified. The item properties corresponds to the category of the items. A significance of each category is identified by finding the frequency levels of the occurrence of the category or the category group. Category group corresponds to items belonging to more than one categories. Significance based reduction is performed in terms of each category and the category features satisfying the threshold frequency are filtered as the final features. The thresholds corresponding to category features are user defined. The actual values for the engineered features are identified by incorporating numerical values for the features. The value is set to one if the given item pertains to the category, zero otherwise. The rating values pertaining to the given items are incorporated to create the training data for the proposed model.

Hybrid Learner based Predictor

The constructed training data is passed to the hybrid learner for prediction. The proposed hybrid learner is composed of multiple learning models, whose results are combined using an aggregator. The proposed architecture for hybrid learner is flexible and can incorporate any number and type of base learners. This work uses two learners for prediction, namely; decision tree learner and stochastic gradient descent.

Decision Tree (DT) [16, 17] is a tree based classifier that generates prediction rules based on the node conditions and branches. Division in every node is performed based on a condition. The condition is formulated using the entropy of attributes. Every leaf node contains a class label, representing the end of the condition sequence.

Stochastic Gradient Descent (SGD) [18, 19] is a stochastic optimization based model, which operates based on the incremental learning principle. The model begins with a minimal constraint based solution and builds incrementally over the already built solution to obtain the complete rule set.

Several literature works under the domain of recommendations has exhibited the efficiency of tree based models as effective recommenders. However, the entropy based divisions in tree based models show slight fluctuations in efficiency levels due to cold start issues. Hence to overcome this problem, SGD, an incremental learning model has been incorporated into the processing architecture. Combining the results of both the models was observed to overcome this issue to a large extent.

The numerical training data is passed to both the models and trained models of DT and SGD are obtained. The data to be predicted is feature engineered and predicted using the trained models of DT and SGD. The hybrid model results in two prediction sets. The final prediction set is obtained by imposing an aggregation operation on both the sets to obtain a single prediction result.

A final filtering phase is applied to the prediction result by selecting the predictions depicting threshold levels higher than the defined prediction threshold. These filtered results form the final recommendations for the user recommenders. However, the entropy based divisions in tree based models show slight fluctuations in efficiency levels due to cold start issues. Hence to overcome this problem, SGD, an incremental learning model has been incorporated into the processing architecture. Combining the results of both the models was observed to overcome this issue to a large extent.

RESULTS AND DISCUSSION

Our HDTSG model has implemented using Python and uses MovieLens data [20, 21] is used to measure the prediction performance. MovieLens is a benchmark dataset, containing 1

Million reviews and is a standard data used in recommendation systems. It contains details about movies, genres, users and the ratings provided by users for the movies. Ratings are provided on a 5 point scale.

Performance of recommendation models are measured using Root-Mean-Square Error (RMSE) and Mean Absolute Error [22, 23]. Mean Absolute Error measures the effectiveness of the predictions, and is given by

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - y'_i| \quad (1)$$

Root Mean Square Error indicates the variability of the predictions, and is given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - y'_i)^2} \quad (2)$$

Where y_i and y_i' are the actual and the predicted ratings for the N test reviews.

The proposed model and its results are compared with the weighted strategy based model (SW I, MLR and CM II) proposed by Fremal et al. [8] and K-Means and Cuckoo Search based model proposed by Katarya et al. [9].

A comparison of the MAE values obtained is shown in figure 1. The model exhibiting lowest MAE value is considered to be the best. It could be observed that the proposed model exhibits the second best MAE values compared to the other algorithms. The category based grouping mechanism enables effective grouping, hence low MAE.

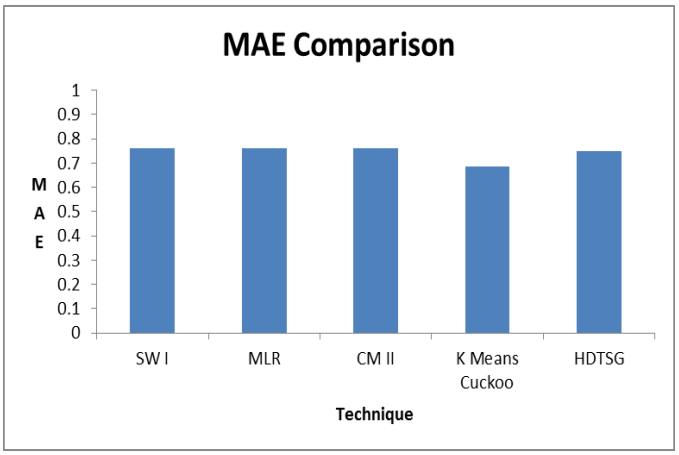

Figure 1: MAE Comparison

A comparison of the RMSE values obtained from the models is shown in figure 2. It could be observed that the proposed model exhibits the least RMSE values compared to all the other models. Even though K Means Cuckoo search model exhibits low MAE values, the RMSE levels were observed to be very high exceeding. Where as our proposed HDTSG model performs better in both cases and found to be most promising than other methods. A tabulated view of the obtained performances are shown in table 1.

A comparison of the variation levels from the best solution is shown in figure 3. It could be observed that the proposed model exhibits the best solution with a variance level of zero. The K Means Cuckoo search model was observed to exhibit maximum and the highest variance, with all the other models exhibiting moderate variance levels. The proposed model was observed to exhibit lowest RMSE, hence exhibits zero variance. The SW I and CM II models exhibit slight variations, while MLR and K Means Cuckoo exhibits high variation levels, exhibiting that the proposed model enables effective predictions in comparison with the other models.

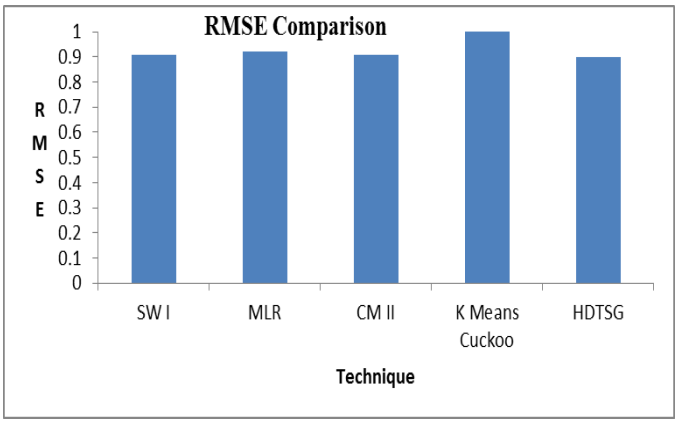

Figure 2: RMSE Comparison

Table 1: Performance Comparison

Model	MAE	RMSE
SW I	0.7616	0.9096
MLR	0.7611	0.9212
CM II	0.7615	0.9086
K Means Cuckoo	0.6842	1.231
HDTSG	0.75	0.9

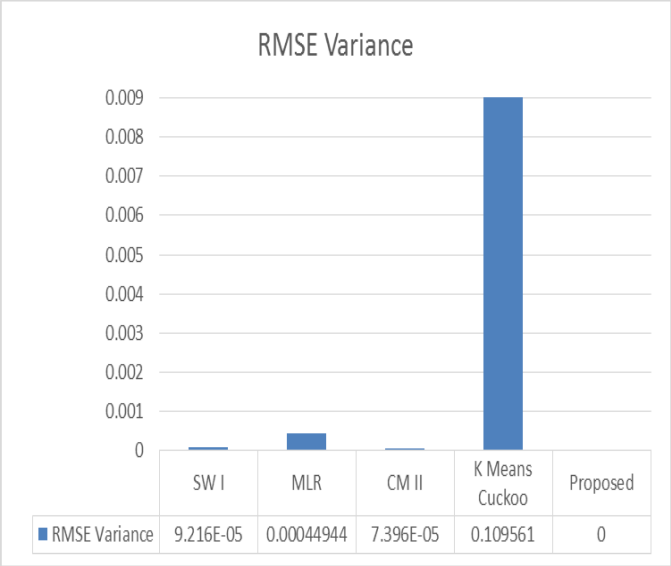

Figure 3: Variance Comparison

CONCLUSION

This paper proposes a hybrid collaborative filtering model incorporated with Feature engineering to provide enhanced predictions. Experimental evaluations were performed with MovieLens data and the results indicate low MAE and RMSE values. Comparisons with state-of-the-art models indicate effective performances of the proposed model exhibiting very low variance levels from the best solution. The major advantages of this model is that, it uses data from the current user and similar users to build the training data. This avoids the problem of data insufficiency and further handles the cold-start issue. Further, training data is composed of only a part of the data, hence leads to reduced computational complexity. Further, feature engineering operates to provide enhanced insights into the data under analysis. Limitations of this model is that it exhibits slightly increased MAE levels compared to the K-Means Cuckoo search model. Future enhancements will operate on providing lowered MAE values.

REFERENCES

[1] Jugovac, M., Jannach D., Lerche, L., "Efficient optimization of multiple recommendation quality factors according to individual user tendencies", Expert Systems with Applications, 81 , 321–331 , 2017.

[2] Nilashi, M., Jannach, D. , bin Ibrahim, O. , Esfahani, M. D. , & Ahmadi, H., " Recommendation quality, transparency, and website quality for trust-building in recommendation agents", Electronic Commerce Research and Applications, 19 ,70–84 , 2016.

[3] Kabassi, K., "Personalizing recommendations for tourists", Telematics and Informatics, 27 (1), 51–66, 2010.

[4] Lilien, G. L. , Kotler, P. , & Moorthy, K. S., "Marketing models",. Prentice-Hall Englewood Cliffs , 1992.

[5] Armstrong, J. S.," *Principles of forecasting: a handbook for researchers and practitioners*", Vol. 30. Boston: Kluwer Academic Publishers, 2001.

[6] Sarwar, B. , Karypis, G. , Konstan, J. , & Riedl, J., "Item-based collaborative filtering recommendation algorithms", *In Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.

[7] Goldberg, K. , Roeder, T. , Gupta, D. , & Perkins, C., "Eigentaste: A constant time collaborative filtering algorithm", Information Retrieval, 4 (2), 133–151, 2001.

[8] Frémal, S. and Lecron, F., "Weighting strategies for a recommender system using item clustering based on genres", Expert Systems with Applications, vol.77, pp.105-113, 2017.

[9] Katarya, R. and Verma, O.P., "An effective collaborative movie recommender system with cuckoo search", Egyptian Informatics Journal, 2016.

[10] da Silva, E.Q., Camilo-Junior, C.G., Pascoal, L.M.L. and Rosa, T.C., "An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering", Expert Systems with Applications, 53, pp.204-218, 2016.

[11] Lampropoulos, A.S., Sotiropoulos, D.N. and Tsihrintzis, G.A., "Evaluation of a cascade hybrid recommendation as a combination of one-class classification and collaborative filtering In Tools with Artificial Intelligence", IEEE *24th International Conference on (Vol. 1, pp. 674-681). (ICTAI),* 2012.

[12] Liu, L., Koutrika, G. and Wu, S., "April. Learningassistant: A novel learning resource recommendation system", *In Data Engineering (ICDE), IEEE 31st International Conference on (pp. 1424-1427). IEEE, 2015.*

[13] Paradarami, T.K., Bastian, N.D. and Wightman, J.L.," A hybrid recommender system using artificial neural networks", Expert Systems with Applications, 83, pp.300-313, 2017.

[14] Nilashi, M., Ibrahim, O. and Bagherifard, K., "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques", Expert Systems with Applications, 92, pp.507-520, 2018.

[15] Pascoal, L.M.L., Camilo, C.G., da Silva, E.Q. and Rosa, T.C., "A social-evolutionary approach to compose a similarity function used on event recommendation. In Evolutionary Computation (CEC)", IEEE *Congress on (pp. 1512-1519)*, 2014.

[16] J.R. Quinlan, "Simplifying decision trees," International Journal of Man-Machine Studies, doi:10.1016/S0020-7373.80053-6,Vol.27(3), pp.221-34, 1987.

[17] B. Kamiński, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees", Central European Journal of Operations Research. doi:10.1007/s10100-017-0479-6,pp.1-25, 2017.

[18] L. Bottou, and Léon, "Online Algorithms and Stochastic Approximations," Online Learning and Neural Networks. Cambridge University Press. ISBN 978-0-521-65263-6,Vol.17(9),pp.142, 1998.

[19] L. Bottou, Léon, O. Bousquet, and Olivier, "The Tradeoffs of Large Scale Learning," Advances in Neural Information Processing Systems, Vol.20, pp. 161–168, 2008.

[20] *https://grouplens.org/datasets/movielens/*

[21] Harper, F.M., and Konstan, J.A., "The MovieLens datasets: History and context", ACM Transactions on Interactive Intelligent Systems, doi: 10.1145/2827872 .vol. 5(4), pp.19:1-19:19, 2015.

[22] Dooms, S., De Pessemier, T., & Martens, L., "Offline optimization for user-specific hybrid recommender systems", Multimedia Tools and Applications, 74(9), 3053-3076, 2015.

[23] Ge, X., Liu, J., Qi, Q., & Chen, Z., "A new prediction approach based on linear regression for collaborative filtering", *IEEE Eighth International Conference In Fuzzy Systems and Knowledge Discovery (FSKD),* Vol. 4, pp. 2586-2590, 2011.