International Journal of Computer Sciences and Engineering
Vol.**11**, Issue **2**, pp.**08-11**, February **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

An Efficient Missing Data Prediction Technique using Recursive Reliability-Based Imputation for Book Recommendation System

Thenmozhi Ganesan[1*], Palanisamy Vellaiyan[2]

[1,2]Department of Computer Applications, Alagappa University, Karaikudi, India

**Corresponding Author: thenmozhiganesan23@gmail.com*

Received: 21/Jan/2023; **Accepted**: 04/Feb/2023; **Published**: 28/Feb/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i2.811

Abstract: Collaborative filtering recommender system is utilized as a significant method to suggest products to the users depends on their preferences. It is quite complicate when the user preference and rating data is sparse. Missing value occurs when there are no stored values for the specified dataset. Typical missing data are in three categories such as (i) Missing completely at random, (ii) Missing at random and (iii) Missing not at random. The missing values in dataset affect accuracy and causes deprived prediction outcome. In order to alleviate this issue, data imputation method is exploited. Imputation is the process of reinstating the missing value with substitute to preserve the data in dataset. It involves multiple approaches to evaluate the missing value. In this paper, we reviewed the progression of various imputation techniques and its limitations. Further, we endeavored k-recursive reliability-based imputation (k-RRI) to resolve the boundaries faced in existing approaches. Experimental results evince the studied methodology appreciably improves the prediction accuracy of recommendation system.

Keywords: Sparse Data, Missing Value, Recommendation system, Missing Value Imputation, Recursive Imputation, Prediction

1. Introduction

As a result of the massive volume of online information availability, there arises the prerequisite for information filtering technique grounded on user's curiosity and item combinations which are offered by the recommendation systems [1]. The volume of online information availability in the recent decades has risen dramatically. Owing to this speedy progress, the problem of filtering the information by the recommender system based on the individual user's interest becomes a critical issue for the forth- coming customers [2]. Recommender system techniques are categorized into following categories [17]:

- Content-based Filtering
- Collaborative Filtering
- Hybrid Filtering
- Demographic Filtering

Among the aforementioned techniques, collaborative filtering is considered as a preferred solution to predict the user interest by their explicit data (ratings and reviews). Collaborative filtering technique further classified into two categories as follows [3]:

- Memory-based Collaborative filtering
- Model-based Collaborative filtering

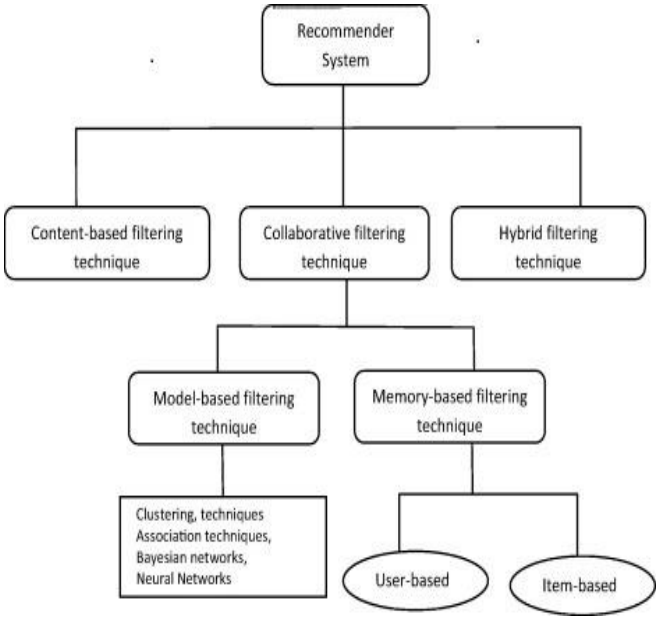

Figure 1: Various recommender system techniques

The benefits of the recommender system are utilized by various fields including e-commerce, e-learning, healthcare, social media and so on [16]. The performance of the recommender system is affected by many factors including insufficient data, arrival of new user or new item, scalability and data redundancy. Dataset with sparse data produces the poor prediction which leads lacking of recommendation accuracy [4].

Missing value is a common issue that pacts with data and causes data analysis and biased results issues [5]. Missing date can be classified into three categories as follows: Missing Completely at Random (MCAR) – There is no correlation between missing data with observed and unobserved data [6]. Missing at Random (MAR) – Information is missing independently with observed data but not with unobserved data. Missing not at Random (MNAR) – Information is missing independently with unobserved data itself [6]. Missing data issue can resolve by two factors including deletion of instances and missing data imputation. Deletion method or data removal method performed eliminating the entire record which confines one or more missing data. It presents the un- stable outcomes by deducing the statistically analyzed samples [7]. To overcome this problem, data imputation is processed in which statistical value is manipulated to replace the missing data in the record [8]. There is several data imputation techniques are utilized by the researchers. Recursive reliability-based imputation technique is suggested to resolve the missing value in the dataset. It is an iterative process to compute the missing data in k number of steps.

Rest of the paper organized as follows: Section II demonstrates the several works taken in the field. Section III describes the varieties of data imputation techniques. Section IV portrays the studied imputation methodology. Performance evaluation and experimental results are demonstrated in Section V and the paper is concluded in Section VI.

2. Related Work

Caio Ribero et al. studied data-driven missing value imputation approach and tested that approach in 10 various longitudinal datasets. The proposed approach presents feature-wise ranking of a set of missing value imputation techniques and the author estimated the proposed method in two set of experiments [12]. Dieter William Joenssen et al. studied hot deck missing data imputation method and the simulation showed that notable variations of donor usage limitations. The author studied that the limitations of donor usage is better in some situations and unlimited usage of donor is better some other situations according to the circumstances [11].

Rouhia M. Sallam et al. proposed an enhanced collaborative filtering method based recommender system with two approaches: Item-based collaborative filtering (KNN) and Singular Value Decomposition based Collaborative Filtering. The proposed methods evaluated by RMSE and MAE. The proposed memory-based and model-based methods are achieved RMSE and MAE value 1.1969, 0.922 and 1.0187, 0.8077 respectively [3].

3. Data Imputation Techniques

3.1 Mean Substitution Method

Mean substitution is a simple imputation technique that consists in replacing each missing value with the mean of the known values of its respective variable [9]. Formally, if Yi j ϵ Ymiss, its imputed value is calculated as

$$\widetilde{\mathbf{Y}}_{\mathbf{i,j}} = \frac{\sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{R_{k,j}} . \mathbf{Y_{k,j}}}{\sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{R_{k,j}}} \quad (\mathbf{1})$$

where $R_{k,\ j}$ are the respective values of the variable in the matrix of missingness R. One of the most important advantages of this method is the low computational cost it has be- cause it's only necessary to compute each variable's mean [10]. It reduces the variability in the data because of the use of the mean value, as it's repeated several times in each variable. It also weakens the covariance and correlation statistics in the data because this method ignores the relationships between variables.

3.2 Hot-deck imputation Method

Hot deck imputation methods are a group of imputation techniques that aim to infer the unknown values of the samples, or individuals, of the data using information of the dataset from the most similar individuals to the one that's being inferred. Given a specific sample that contains missing values, known as the recipient, the main idea of this group of methods is to select one sample, or a group of samples, within the dataset which will be used to infer the recipient's missing values.

Those samples are known as donors and usually, are selected in terms of similarity with the recipient, i.e. in terms of distance. Each method proposes a specific methodology in selecting the recipient's donors and in inferring the unknowns [11]. Regarding the selection of donors, some of the methodologies that are being used include the use of distance functions or clustering methods, as the k-Nearest Neighbors methods, also known as k-NN, or the k-Means methods. Deterministic hot-deck method and Random hot-deck method are the two types of hot-deck imputation methods.

3.3 Predictive Mean Matching Method

The Predictive Mean Matching method also denoted as PMM method, is a random hot-deck method that consists in retrieving actual dataset values from the nearest observed individuals. In PMM, the recipient's nearest individuals in a specific variable are those individuals that have a known variable value and which linear regression estimated value for the imputed variable is similar to the recipient's linear regression estimated value [12].

Given a dataset with data matrix Yϵ $R^{n\times p}$, let $Y_{i,miss}$ be the value that going to be imputed and, therefore, let X_{miss} be the variable which values are being imputed from. Finally let $\vec{X} = (X_1, X_2, .., X_t)$ be a set of variables of the dataset selected to be used to estimate X_{miss} as a linear combination of them, that is,

$$\boldsymbol{X_{miss} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_t X_t + \varepsilon} \quad (\mathbf{2})$$

where $\vec{\beta} = (\beta_1, \beta_2, \ldots, \beta_t)$ are known as the regression coefficients and where ε is the error term.

3.4 Multiple Imputation Method

Several imputation methods that we've explained up to this point, as the Predictive Mean Matching or the Random Hot-deck methods, include a random selection from the pool of donors [7]. The fact of choosing randomly carries ambiguity about the imputed values. Multiple imputation method provides a constructive strategy based on three steps.

- **Imputation** - Given a dataset which contains missing values and which is stored in a matrix Y, the multiple imputation method generates a set of m complete datasets Y1, Y2,..., Ym by using the same random imputation method. Usually, m takes value from 3 to 5.
- **Analysis** - Each of the m datasets extracted from the previous step are analyzed by using standard procedures, such as the computation of some statistics.
- **Pooling** - After the conclusions extracted from the analysis of the datasets, the obtained results are combined from the different datasets into a final one.

However, an important disadvantage of multiple imputations is, it's a time consuming method due to the multiple iterations in the datasets and their respective analysis. It's also required a statistical expertise to understand the analysis step.

4. Studied Imputation Methodology

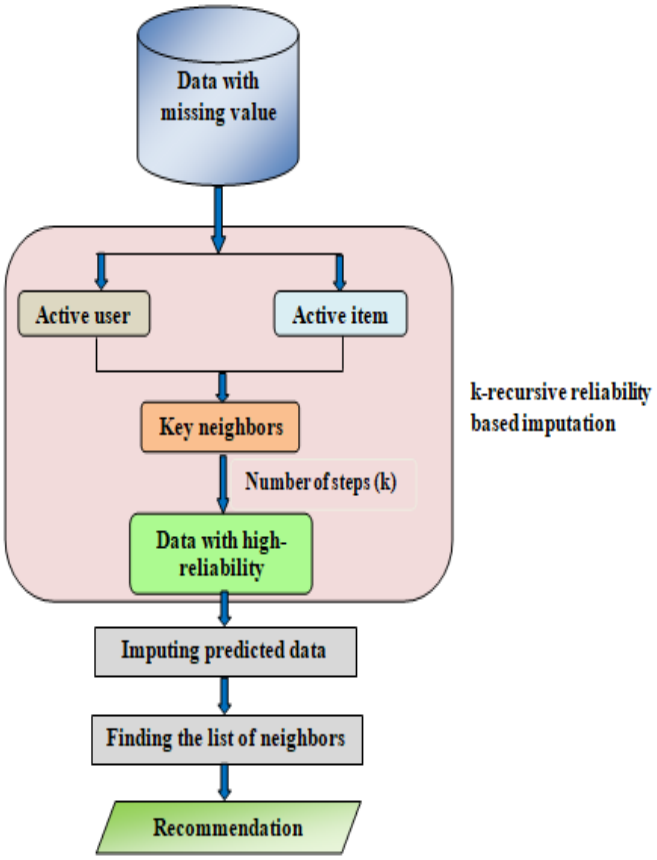

Figure 2. Working principles of k-recursive reliability-based imputation

In k-recursive reliability-based imputation method, missing values are imputed recursively. High reliability data are identified and selected at the early stage. Then the selected reliability data are imputed recursively. Whenever lower reliability criteria are occurred, additional selection progress is performed in an iterative manner.

Unlike existing approaches, to resolve the data sparsity issue, our proposed k-RRI restored the real and reliable virtual data [13]. In accordance with the actual rating histories of the users, predicted data are assigned in a provided step and the virtual data are assigned in the previous step. The threshold cut-off is employed to set the user-item similarity index.

4.1 k-Recursive Reliability-based Imputation Method Steps

Step 1: Initialize the input values i.e.) user-item rating matrix, similarity set of user-item, active user and item prediction, recursive count and the threshold cut-off.
Step 2: Initialize the imputed matrix.
Step 3: Calculate and impute the missing value based on the threshold cut-off by executing the k recursive steps**.**
Step 4: Diminish the cutoff value as per the progress of algorithm.
Step 5: Set the similarity value for current user and item.
Step 6: Select the set of key neighbours and merge the active user and item.
Step 7: Impute the missing data in the imputed matrix.

5. Performance Evaluation and Experimental Results

We tested our method on BookCrossing (BX) dataset which were collected by the Cai-Nicolas Ziegler. It contains 1,149,780 ratings from 278,858 users on 271,379 books [14]. The dataset was divided as testing model and training model for our convenient. In test data model, rated items are limited with values labelled as T1, T2, T3 and T4. The performance of the recommender system is evaluated by performance matrices. We evaluated the performance accuracy by mean absolute error (MAE). It is written as [15],

$$\mathbf{MAE} = \frac{1}{|\hat{R}|} \sum_{\hat{r}_{ui} \in R} |r_{ui} - \hat{r}_{ui}| \qquad (3)$$

where R is the test dataset.

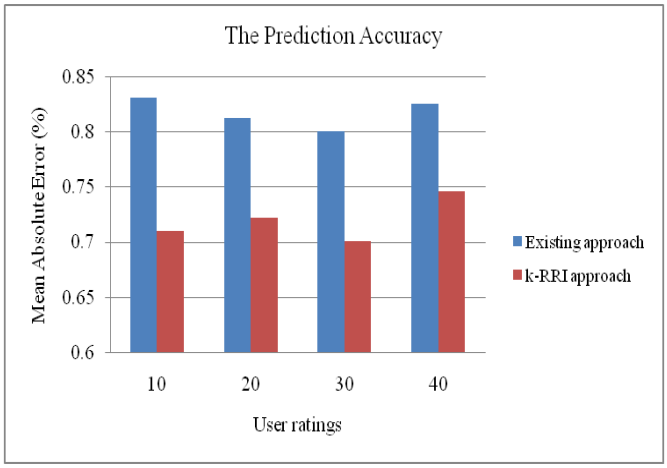

Figure 3. Data imputation for missing data with prediction accuracy

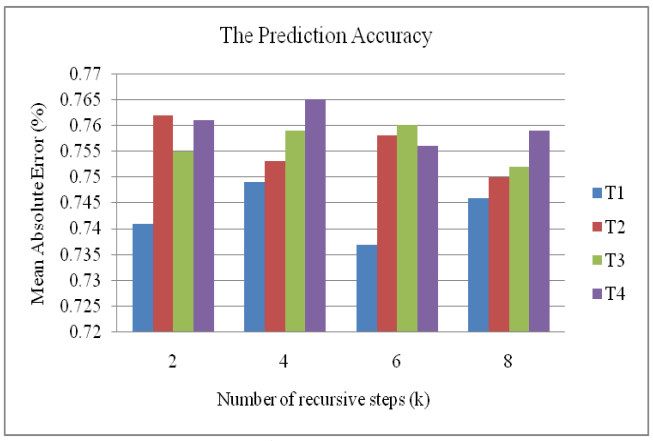

Figure 4. Prediction accuracy for various number recursive values (k) in k-RRI

6. Conclusion

Recommender system is utilized by researchers to overcome the information overloaded problem. Yet, Sparseness in dataset affects the performance and accuracy of the recommender system. To alleviate the aforementioned problem, an effective missing value handling technique is obligated. In this paper, we discussed the missing data issue which leads data sparsity in datasets and various approaches utilized for the data imputation techniques. Further, we studied k-recursive reliability-based imputation method to impute the missing data to effectively alleviate the data sparsity problem. k-RRI method is based on the recursive process and the optimized result is obtained by lowering the threshold cutoff value. The efficiency of the algorithm is measured by mean absolute error and the studied approach relatively increased the prediction accuracy. The experimental results demonstrated the effectiveness of the algorithm compared with the existing imputation measures.

References

[1]. B. Alhijawi, G. Al-Naymat, N. Obeid and A. Awajan, "Novel predictive model to improve the accuracy of collaborative filtering recommender systems", Inf. Syst., **vol. 96, p. 101670**. **2021.**

[2]. S. Chinchanachokchai, P. Thontirawong and P.Chinchanachokchai, "A tale of two recommender systems: The moderating role of consumer expertise on artificial intelligence based product recommendations", J. Retail. Consum. Serv. vol. **61**, January. p**. 102528. 2021.**

[3]. RM. Sallam, M. Hussein, and HM. Mousa, "An Enhanced Collaborative Filtering-based Approach for Recommender Systems", Int. J. Comput. Appl. **vol. 176**. no. **41. p. 9–15. 2020**.

[4]. FT. Abdul Hussien AM. Rahma and HB. Abdul Wahab, "Recommendation Systems for E-commerce Systems An Overview", J. Phys. Conf. Ser. **vol. 1897, no.1. 2021**.

[5]. L. Li, Y. Zhang and L. Chen, "EXTRA: Explanation Ranking Datasets for Explainable Recommendation", SIGIR 2021 - Proc. *44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr*. pp. **2463–2469. 2021.**

[6]. BM. Marlin and RS. Zemel, "Collaborative prediction and ranking with non-random missing data", RecSys'09 - Proc. 3rd *ACM Conf. Recomm. Syst*. pp. **5–12. 2009**.

[7]. T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago and O. Tabona, "A survey on missing data in machine learning", **vol. 8. no. 1**. Springer International Publishing. **2021.**

[8]. PV. Kumar and MV. Gopalachari, "A Review on Prediction of Missing Data in Multivariable Time Series", Int. J. Comput. Appl. **vol. 7. no. 6.** pp**. 84–91. 2017**.

[9]. C. Ribeiro and AA. Freitas, "A data-driven missing value imputation approach for longitudinal datasets", **vol. 54. no. 8**. Springer Netherlands. **2021**.

[10]. G. Özbal, H. Karaman and FN. Alpaslan, "A content-boosted collaborative filtering approach for movie recommendation based on local and global similarity and missing data prediction", Comput. J. **vol. 54, no. 9**. pp. **1535–1546. 2011**.

[11]. Joenssen DW. and Bankhofer U. Hot deck methods for imputing missing data: The effects of limiting donor usage. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics).* **vol. 7376 LNAI. no. September. pp. 63–75. 2012.**

[12]. Isinkaye FO. Folajimi YO. and Ojokoh BA. Recommendation systems: Principles, methods and evaluation. *Egypt. Informatics J.* **vol. 16. no. 3. pp. 261–273. 2015.**

[13]. Ihm SY. Lee SE. Park YH. Nasridinov A. Kim M and Park SH. A technique of recursive reliability-based missing data imputation for collaborative filtering. *Appl. Sci.* **vol. 11. no. 8. 2021.**

[14]. Kondo K. Recommendation system for CINEMA. no. January. 2008.

[15]. Berisha F. Quality of the predictions : mean absolute error. accuracy and coverage. no. September **2018.**

[16]. M. Murugeswari, S. Vimala, "The Recommender System for Smart E-Learning System Using Big Data: A Survey", *International Journal of Computer Sciences and Engineering*, **Vol.8, Issue.10, pp.94-99, 2020.**

[17]. Prachi Dahiya, Neelam Duhan, "Comparative Analysis of Various Collaborative Filtering Algorithms", *International Journal of Computer Sciences and Engineering*, **Vol.7, Issue.8, pp.347-351, 2019.**

AUTHORS PROFILE

Ms.Thenmozhi Ganesan is pursuing Doctor of Philosophy in Machine Learning in the Department of Computer Applications, Alagappa University, Karaikudi, Tamilnadu, India. She has completed Master of Philosophy in 2019 and Master of Computer Application in 2018 at Alagappa University. Her main research involved in thrust areas such as Machine Learning, Data Mining, Image Processing and Biometrics based authentication. She has published 3 research articles in reputed journals and attended more than 4 International Conferences. E-Mail:thenmozhiganesan23@gmail.com

Dr.Palanisamy Vellaiyan obtained his B.Sc degree in Mathematics from Bharathidasan University in 1987. He also received M.C.A. and Ph.D. Degree from Alagappa University in 1990 and 2005 respectively. After working as Lecturer in AVVM Sri Pushpam College, Poondi Thanjavur from 1990 to 1995, He joined Alagappa University as Lecturer in 1995. He is currently working as Professor and Head of the Department of Computer Applications, Member of the Syndicate and Dean Student Affairs of the Alagappa University. He also received M.Tech. degree from Bharathidasan University in 2009. He has published more than 130 international journals and he has attended 30 national conferences and 60 international conferences and his research interest includes Computer Networks & Security, Data Mining & Warehousing, Mobile Communications, Computer Algorithms ,Biometrics and Machine Learning. E-Mail: palanisamyv@alagappauniversity.ac.in.

