Novel Recommendation Algorithms with Applications to Healthcare Data Analysis

Wenbin Yue

Department of Computer Science
Brunel University London

This dissertation is submitted for the degree of
Doctor of Philosophy

November 2020

I would like to dedicate this thesis to my parents.

Declaration

I, Wenbin Yue, hereby declare that this thesis and the work presented in it are entirely my own. Some of the work has been previously published in journal or conference papers, and this has been mentioned in the thesis. Where I have consulted the work of others, this is always clearly stated.

Wenbin Yue
November 2020

Acknowledgements

I would like to express my gratitude to all those who helped me during the writing for this thesis.

First, I wish to give my sincere gratitude to my supervisors, Prof. Zidong Wang and Prof. Xiaohui Liu for their extraordinary patience, consistent encouragement, professional guidance, and constant support. They have walked me through all the stages of the writing of this thesis. Without their consistent and illuminating instruction, this thesis could not have reached its present form. Prof. Zidong Wang is a great supervisor, and I really appreciate his efforts in these four years. He is my role model and life mentor, who not only let me know what good research is truly about but also taught me a lot of philosophy of life. These will be the precious wealth of my life. My heartfelt thanks also go to Prof. Xiaohui Liu for his help and encouragement when I encounter difficulties in the past four years.

Second, I would like to express my heartfelt gratitude to the following people for useful discussions, suggestions, comments, and supports of my research during the PhD stage: Prof. Xin Luo, Prof. Lifeng Ma, Prof. Bo Shen, Prof. Jun Hu, Prof. Yuan Yuan, Dr. Yan Song, Dr. Licheng Wang, Dr. QinYuan Liu, Dr. Dong Wang, Dr. Hongjian Liu, Dr. Zou lei, Prof. Yurong Liu, Dr. Chuanbo Liu, Dr. Hang Geng, Dr. Wenshuo Li, Dr. Fan Wang, Dr. Wenying Xu, Dr. Yonggang Chen, Dr. Bin Li, Dr. Hongwei Chen, Dr. Song Jun, Dr. Qi Li, Dr. Xiongbo Wang, Prof. Yun Chen, Dr. Shuai Liu, Ms. Di Zhao, Mr. Yuxuan Shen, Mr. Hailong Tan, Prof. Min Wang, Dr. Lei Ma, Mr. Bo Tian, Prof. Weiwei Song, Dr. Dan Liu, Dr. Hua Yang, Ms. Jiahui Li, Dr. Lulu Tian, Ms. Chen Gao, Mr. Junlin Li, Mr. Weihao Song, Dr. Tingli Su, Mr. Kaiqun Zhu, Dr. Nan Hou, Mr. Junyi Li, Ms. Dongying Song, Prof. Chunyan Han, Prof. Wei Wang, Mr. Tongxiang Li, Ms. Jieyu Zhang, Dr. Yinfang Song, Mr. Wei Chen, Ms. Zhiru Cao, Ms. Mincan Li, Mr. Yibo Huang. It is my honor to meet you in these years.

Third, I would like to thank all my lab mates at the Centre for Intelligent Data Analysis for the pleasant and enjoyable working atmosphere: Dr. Weibo Liu, Dr. Khalid Eltayef, Dr. Bashir Dodo, Dr. Navid Dorudian, Ms. Yani Xue. I would also like to thank the Department of Computer Science, Brunel University London for giving me this good opportunity to do my Ph.D. research and meet so many great people.

Last my thanks would go to my beloved family for their financial support, loving considerations, and great confidence in me all through these years. My parents have always been helping me out of difficulties and supporting me without a word of complaint. Their unselfish love and care have been inspiring me and keeping me moving forward.

Abstract

Along with the development of the times, people are paying more and more attention to health issues. The incorporation of machine learning technology has led to unprecedented development in many disease studies that are concentrated on prevalent diseases. Unfortunately, for many rare diseases, there are still many limitations.

Friedreich's ataxia (FRDA) is a rare inherited neurodegenerative disorder that causes progressive damage to nervous systems and performance deterioration of physical movements. European Friedreich's Ataxia Consortium for Translational Studies (EFACTS), which is funded by the European Union project, has integrated disease-related resources and assembled a large pool of experts to promote FRDA research. FRDA baseline data analysis and application play a crucial role in advancing the disease research, but there are many obstacles that prevent EFACTS from collecting patient baseline data:

- Lack of the rare disease awareness (individual): For individuals, the disease can be overlooked by the patients' families due to less severe pre-disease symptoms and lack of relevant medical knowledge.

- Lack of the rare disease awareness (local hospital): For doctors in local hospitals, they might be unable to make correct and effective diagnosis in a timely manner because of the complexity of the clinical manifestations of these diseases and the fact that local hospitals are likely to lack specialists and knowledge in the relevant fields, etc.

- Medical system problems: There is a lack of detailed and effective diagnostic process for rare diseases.

- Economic & physical reasons: Medical resources for rare diseases are concentrated in large cities. Many patients in other regions do not have the financial or physical conditions to go to big cities for diagnosis and treatments.

There are three challenging issues in helping with FRDA baseline data collection from computer science perspective: 1) how to develop appropriate strategies to overcome existing difficulties to help the collection of diseases; 2) how appropriate machine learning methods can be used for effective baseline data collection according to the actual situation of FRDA; and 3) how to develop novel algorithms to ensure the accuracy of data collection based on various scenarios that may occur.

In this thesis, machine learning techniques are used to address the difficulties on the current baseline data collection and missing value prediction in FRDA. Based on the idea of recommendation system (RS) in machine learning, a new collection strategy and some improved algorithms have been proposed to address various possible difficulties in data collection. The main work is as follows:

- To help FRDA baseline data collection, a novel data collection strategy is proposed for the FRDA baseline data by using the collaborative filtering (CF) approaches. This strategy is motivated by the popularity of the nowadays “RS” whose central idea is based on the fact that similar patients have similar symptoms on each test-item. By doing so, instead of having no data at all, the FRDA researchers would be provided with certain predicted baseline data on patients who cannot attend the assessments for physical/psychological reasons, thereby helping with the data analysis from the researchers’ perspective. It is shown that the CF approaches are capable of predicting baseline data based on the similarity in test-items of the patients, where the prediction accuracy is evaluated based on three rating scales selected from the EFACTS database.

- With the aim to facilitate the baseline data collection with improved prediction accuracy, the framework of the proposed algorithm is constructed based on a novel hybrid model combining the merits of model- and memory-based CF methods. The proposed hybrid algorithm exhibits the following two main features: 1) when a patient does not have neighbors sharing similar baseline data, the model-based CF component is activated to employ certain clustering method to find similar neighbors based on their attributes; and 2) in the case that a patient does have neighbors, a novel similarity measure, which accounts for more statistical characteristics by integrating rating habits and degree of co-rated items, is developed in the memory-based component of the algorithm in order to adjust initial similarities between the patients. To evaluate the advantages of the proposed algorithm, the SARA is selected from the database of EFACTS.

- In order to handle cold-start condition during FRDA baseline data collection, a weighted-naive-Bayes based CF (WNBCF) algorithm is proposed by taking into account the patient side-information. To be specific, the patient side-information is treated as weighted attributes in the WNBCF algorithm to facilitate the prediction of the severity of different bodily functions of FRDA patients. An attribute-weighting algorithm is first presented based on the mutual information to support weight selection. To improve the performance of selected weights, the particle swarm optimization algorithm is then exploited to optimize the weights obtained by the attribute-weighting algorithm. In order to assess the superiorities of the proposed WNBCF algorithm, real-world FRDA datasets are chosen from the database provided by EFACTS (the European Friedreich's Ataxia Consortium for Translational Studies).

- A modified collaborative filtering (MCF) algorithm with improved performance is developed for recommendation systems with application in predicting baseline data of FRDA patients. The proposed MCF algorithm combines the individual merits of both the user-based CF method and the item-based CF method, where both the positively and negatively correlated neighbors are taken into account. The weighting parameters are introduced to quantify the degrees of utilizations of the user-based CF and item-based CF methods in the rating prediction, and the particle swarm optimization algorithm is applied to optimize the weighting parameters in order to achieve an adequate tradeoff between the positively and negatively correlated neighbors in terms of predicting the rating values. To demonstrate the prediction performance of the proposed MCF algorithm, the developed MCF algorithm is employed to assist with the baseline data collection for the FRDA patients.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AC	Adjust Cosine
ACM	Association For Computing Machinery
ADL	Activities of Daily Living
AE	Autoencoder
ANN	Artificial Neural Network
CNN	Convolutional neural network
GRU	Gated recurrent unit
MLP	Multilayer perceptron
RNN	Recurrent neural network
CF	Collaborative Filtering
EFACTS	European Friedreich’s Ataxia Consortium for Translational Studies
FRDA	Friedreich’s Ataxia
INAS	Inventory of Non-Ataxia Symptoms
LFM	Latent Factor Model
MAE	Mean Absolute Error
MCF	Modified Collaborative Filtering

PCC	Pearson Correlation Coefficient
PSO	Particle Swarm Optimization
RMSE	Root Mean Square Error
RS	Recommendation System
SAI	SARA-ADL-INAS
SARA	Scale for the Assessment and Rating of Ataxia
SSE	Sum of Squares Error
SVD	Singular Value Decomposition
WNBCF	Weighted-Naive-Bayes based Collaborative Filtering

Chapter 1

Introduction

1.1 Motivation

With the improvement of living standards, people pay more and more attention to their health. Machine learning has great potential in helping clinicians, physicians and researchers discover patterns from existing data sets, thereby improving medical efficiency and quality of care. It is well known that, in the past few decades, prevalent diseases have developed at an unprecedented rate, mainly because of the large number of patients, the high social demand, the strong research base and the ease of data collection. By contrast, there is still a lack of adequate social attention to rare diseases. This thesis focuses on how machine learning techniques can be applied to conduct the treatment for rare diseases such as Friedreich's ataxia (FRDA). It is hoped that our studies can bring more inspirations to researchers in the study of rare diseases.

FRDA is a rare inherited neurodegenerative disorder affecting the multi-system of the body. FRDA occurs more frequently in Europe and America than in other regions of the world with morbidity of 2 to 4 per 100, 000 individuals. To date, there have been no effective treatments for FRDA. To investigate the FRDA in a comprehensive way, a group of experts formed the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) in 2010, aiming to build the first representative international European patient FRDA registry (funded by the European Union project). EFACTS is committed to collecting and analyzing different kinds of baseline data from FRDA patients including demographics, clinical rating scales, quality-of-life measures, etc., which are of significant use in clinical trials. Up to now, more than one thousand FRDA patient baseline data have been collected by EFACTS from

19 study sites in 9 countries. Although the database is continuously updated every year, the speed of the data collection is slowing down. There are several main difficulties for EFACTS in collecting patient baseline data and follow-up data:

- Physical reasons: FRDA is a neurodegenerative disorder, which means FRDA patient's physical condition is not good and will slowly deteriorate so that many patients cannot travel long distances to study sites for testing.
- Economic reasons: EFACTS study sites are all located in large cities. Many patients in other regions may not have enough financial means to go to big cities for testing.
- Psychological reasons: Fear, worry, anxiety and so on.

According to the morbidity rate, Europe may have more than 28,000 FRDA cases. Unfortunately, only about 3.6% of the FRDA patients are recorded in the EFACTS database. Many studies have shown that the clinical sample sizes in many existing FRDA studies are *far too small* to be convincing even if these studies have positive clinical results [108, 129, 109, 31]. In this case, there is an urgent need to introduce some advanced methods to help EFACTS collect more patient data so as to provide better support in clinical trials. More baseline data can not only provide sufficient clinical candidates for clinical trials but also help promote better disease research in terms of effective biostatistical analysis. It should be mentioned that the traditional baseline data collection method is through patient interviews, questionnaires, observations and coordinated tests at the study sites. The doctor will give the patient ratings on the corresponding test-items.

To overcome such challenges, in this thesis, we aim to adopt machine learning techniques to find a quick, convenient, intelligent, and effective way to assist FRDA patient baseline data collection. Inspired by the idea of the nowadays popular recommendation systems (RS), recommendation techniques seem to be a good candidate to assist in collecting FRDA patient baseline data. As one of the most famous recommendation techniques, collaborative filtering (CF) has been widely used in practice due to its great success in modeling the characteristics of users and items. The main idea of CF is to analyze user behaviors in order to seek a group of users who share similar interests in the communities, and then the items are recommended to the target user based on behaviors generated by his/her similar users.

The reasonable assumption is made that similar patients show similar symptoms, just like similar users exhibit similar interests on items in RSs. Note that the main advantage of using RS is that not all ratings are required for all test-items. FRDA patients do not have to go to study sites to take the assessments. As long as the patients/their family can provide some reliable baseline data at home, the unprovided parts are considered as missing values. The prediction of the missing rating values can be considered as a typical RS problem where the patient is treated as a "user", and the test-item is regarded as an "item". By resorting to the RS, the FRDA researchers would be provided with certain predicted baseline data (produced by the proposed RS) on patients who *cannot* attend the assessments for physical/psychological reasons, thereby helping with the data analysis from the researchers' perspective.

In this thesis, the scientific problems can be summarized as follows:

1. Using suitable machine learning algorithms to reveal the relationship between the known data provided by the patients and the unknown data during the data collection process;

2. Conducting in-depth research on how to accurately describe the relationship between patients through known data under different sparse conditions;

3. Studying the influence of the cold-start problem on the RS and exploring the effect of user side-information to solve the cold-start problem;

4. Exploring the feasibility of combing the patients' positive and negative neighbors and test-items' positive and negative neighbors in the RS.

The main purposes of this thesis are to 1) launch a major study on developing a novel strategy using recommendation techniques to assist FRDA patient baseline data collection, and 2) develop advanced algorithms to adapt different situations (sparsity situation, cold-start situation and incomplete information situation) to improve the prediction accuracy of missing values during the data collection process. In Chapter 3, a novel data collection strategy is proposed for the FRDA disease based on the memory-based CF approaches. The proposed strategy makes it possible to construct a new framework for data collection by using popular CF methods. In Chapter 4, a novel hybrid CF framework is introduced, whose idea is to switch between the model-based CF and the memory-based CF according to the actual situation for the comprehensive use of incomplete FRDA baseline data. In Chapter 5, a weighted-naive-Bayes based CF (WNBCF) is proposed to tackle the cold-start problem during the FRDA patient-baseline data collection, which provides valuable support to the clinical trials and further disease research. In Chapter 6, a modified CF (MCF) algorithm is proposed, which not only combines the merits from the UBCF and IBCF methods but also makes full use of the positively and negatively correlated neighbors in predicting the missing values.

1.2 Contribution

The main contributions of the thesis are listed as follows.

- To improve EFACTS baseline data collection, we aim to provide a novel strategy by adopting the idea of the nowadays popular recommendation system (RS) which is based on the fact that similar patients have similar symptoms on each test-item. The proposed strategy makes it possible to construct a new framework for the data collection by using popular CF methods. The proposed strategy does not have any geographic restrictions on the required data and, as a result of the developed framework, the predicted data provides an alternative database for FRDA data analysis that would potentially assist clinical trials.

- We propose a hybrid CF algorithm for baseline data prediction of FRDA patients. The proposed algorithm switches between model-based and memory-based CF techniques according to degrees of the data sparsity and individual differences. More specifically, the model-based CF is used to deal with the situation where a patient does not have similar neighbors because of the sparsity, and the memory-based CF is exploited for a patient who has neighbors but is under uncertainties arising from individual differences. In the former case, the model-based CF is harnessed to find similar neighbors with similar FRDA symptoms by clustering this patient into the class based on his/her attributes. Here, it is quite challenging to choose key attributes for clustering because 1) we need to analyze what kinds of attributes FRDA patients can provide; 2) based on the pathology of FRDA and basic statistical analysis, key attributes are picked out from the results of the previous step to conduct the clustering; and 3) the most suitable number of clusters is determined according to the clustering results. In the case of a patient with similar neighbors, we adopt an advanced memory-based CF algorithm with an improved similarity measure, where both the patient rating habits and the number of co-rated test-items are taken into account from a unified viewpoint.

- A weighted-naive-Bayes based CF (WNBCF) recommendation algorithm is proposed to solve the cold-start problem. The naive Bayes (NB) method is a popular classification method based on the Bayes' theorem and the independence assumption between the features. According to the characteristics of the NB method, in practical applications, the patient side-information (attributes) is adopted to discover the relationship with ratings (classes) on test-items. The patient side-information is the basic and useful information of a patient which include, but are not limited to, patients' age, gender, onset age and so on. The NB algorithm assumes that all the attributes have the same importance for classification. In fact, different attributes may have different influence on classification performance. As such, a weighted NB algorithm is developed whose main idea is to assign different attributes with different weights according to their significance in improving the performance of the classifier. In this situation, a challenging problem is to allocate *proper* weights for each attribute to achieve a superior classification performance. In this context, the weight selection can be treated as an optimization problem, and it becomes a rather challenging task on how to effectively solve such a constrained optimization problem from the perspective of FRDA patient baseline data collection.

To formulate an optimization problem, the utilization of reasonable constraints would definitely help improving both the reliability and the accuracy of the optimization results. In the optimization problem addressed in this thesis for the FRDA patient baseline data collection, the first constraint is an inequality constraint that reflects the relationship between the weights. The actual situation is that different patient side-information has different effects on the ratings of test-items. To describe these effects, mutual information (MI) is introduced in this thesis to investigate the mutual dependence between the attributes and classes. The MI is chosen as an index to reflect the importance of each attribute in the classification. The more important the attributes are, the larger their weights would be. Another equality constraint is that the sum of weights is required to be 1. As such, the problem becomes a constrained optimization one. To facilitate the subsequent development of the optimization algorithm, the penalty-function method is utilized to transform the constrained optimization problem into a series of unconstrained ones.

Evolutionary computation algorithms have shown outstanding performance in solving optimization problems in a wide range of real-world applications such as healthcare, telecommunication, power systems, and so on. As a powerful member of the family of optimization techniques, the particle swarm optimization (PSO) algorithm has been successfully employed to solve the optimization problems owing to its easy implementation and relatively fast convergence towards satisfactory solution. In this context, the PSO algorithm is exploited in this thesis to search for the optimal weights for each attribute in WNBCF, where the selection of weights satisfies the constraints mentioned previously. To test its superiority and effectiveness, our proposed method is compared with some conventional algorithms and applied to the FRDA patient baseline data collection problem under the cold-start condition.

- We propose a modified CF (MCF) algorithm in this thesis by combining the merits of user-based CF and item-based CF methods. Through the utilization of the information from both the positively and negatively correlated neighbors, the proposed algorithm is capable of predicting the missing values in multi-aspects with satisfactory accuracy. In particular, the PSO algorithm is dedicatedly exploited to determine (locally) optimized weights of our proposed MCF algorithm so as to achieve: a) an adequate tradeoff between the user-based and the item-based similarity measures; and b) a proper balance between the positively and negatively correlated neighbors. To illustrate its application potential, our proposed algorithm is applied to assist with the baseline data collection for FRDA patients.

1.3 Publication

The related results in this thesis have been reported in the following publications:

- **W. Yue**, Z. Wang, H. Chen, A. Payne, and X. Liu, Machine learning with applications in breast cancer diagnosis and prognosis, *Designs*, vol. 2, no. 2, art. no. 13, 17 pages, 2018. (Resulting from Chapter 2)

- **W. Yue**, Z. Wang, B. Tian, A. Payne, and X. Liu, A collaborative-filtering-based data collection strategy for Friedreich's ataxia, *Cognitive Computation*, vol. 12, no. 1, pp. 249–260, 2020. (Resulting from Chapter 3)

- **W. Yue**, Z. Wang, B. Tian, M. Pook, and X. Liu, A hybrid model- and memory-based collaborative filtering algorithm for baseline data prediction of Friedreich's ataxia patients, *IEEE Transactions on Industrial Informatics*, 2020, DOI:10.1109/TII.2020.2984540. (Resulting from Chapter 4)

- **W. Yue**, Z. Wang, W. Liu, B. Tian, S. Lauria, and X. Liu, A novel collaborative filtering approach for Friedreich's ataxia baseline data collection under cold-Start condition, under review (submitted to *IEEE Transactions on Cybernetics*). (Resulting from Chapter 5)

- **W. Yue**, Z. Wang, W. Liu, B. Tian, S. Lauria, and X. Liu, An optimally weighted user- and item-based collaborative filtering approach to predicting baseline data for Friedreich's ataxia patients, *Neurocomputing*, 2020, DOI:10.1016/j.neucom.2020.08.031. (Resulting from Chapter 6)

1.4 Thesis structure

This thesis is organised into 7 chapters including the current chapter. The contents of the rest chapters are summarized as follows:

In Chapter 2, we provide background information on knowledge that is closely related to this thesis. This chapter begins with the background knowledge on RS. Following that, we review one of the most successful techniques in the RS, CF. Then, we discuss the development of the variant CF algorithms and their applications in healthcare field.

Starting by the motivation of the work, in Chapter 3, a novel data prediction algorithm for FRDA has been proposed based on the CF methods including both the user-based and item-based ones. By introducing the PCC to evaluate the similarities between different patients and calculating the missing values of the patient-item matrices, the proposed algorithm has been implemented to predict incomplete information of the new patients with an acceptable accuracy. Numerical experiments have been carried out with four different cases (namely, SARA, ADL, INAS, and SAI datasets) in order to demonstrate the effectiveness of the proposed method. This proposed strategy can also be extended to facilitate the collection of data for other diseases. New patient data made available in this way can be used to assist patient selection for clinical trials and data analysts to achieve better management of the underlying disease.

In Chapter 4, a hybrid model- and memory-based algorithm has been presented and successfully applied to improve the prediction performance on FRDA baseline data. By taking model-based CF into account, the drawback of the traditional similarity calculation methods in finding neighbors in the sparse data condition has been overcome. Moreover, an enhanced and more generalized similarity measure has been proposed in memory-based

CF so as to provide a more comprehensive evaluation for the similarity degree between two patients by considering the rating habits and degree of co-rated test-items. Large-scale real-world FRDA experiments have been conducted, showcasing the validity and feasibility of our algorithm.

In Chapter 5, a modified WNBCF algorithm has been proposed to solve the cold-start problem in the FRRS. By employing the WNBCF algorithm, the patient side information is utilized for the missing value prediction. In addition, the PSO algorithm has been applied to automatically select appropriate weights in the WNBCF algorithm. The developed WNBCF algorithm has been successfully applied to the actual FRDA baseline data collection problem with satisfactory performance. In the situation that the patients are unable to provide any rating data, our algorithm can produce reasonable prediction results, which gives a new solution to aid the FRDA patient baseline data collection. Experiment results have shown the feasibility and effectiveness of our proposed algorithm by comparing it with some conventional algorithms.

In Chapter 6, an MCF algorithm has been presented and successfully employed to deal with the data prediction problem of FRDA patient baseline data. The proposed MCF algorithm has combined the merits of both the user-based CF method and the item-based CF method, and has been shown to outperform the user-based CF method alone or the item-based CF method alone. It should be pointed out that the positively and the negatively correlated neighbors have also been taken into account in the MCF algorithm with hope to improve prediction accuracy. In the developed MCF algorithm, the weighting parameters have been employed to balance the usage of 1) the user-based CF method and the item-based CF method; and 2) the positively and the negatively correlated neighbors. The PSO algorithm has been applied to automate the selection of locally optimized weights so as to guarantee the prediction accuracy. The MCF algorithm has been applied to deal with a real-world disease, the FRDA, to justify its application potential. Experiment results have shown that our proposed approach greatly improves the prediction accuracy with better performance than either the user-based CF algorithm or the item-based CF algorithm.

In Chapter 7, we summarize the current work presented in this thesis and discuss several directions for future research in the light of gaps in the work.

Chapter 2

Background

With the advent of the big data era, people enjoy the convenient services brought by information technology, but at the same time, they are bothered by information overload. As an efficient information filtering tool for assisting users in dealing with information overload, recommendation system (RS) has been widely used in various fields, such as E-commerce, movies, music, news and so on (more examples are shown in Fig. 2.1) [29, 83]. Various recommendation algorithms is utilized in RS to learn user's requirements from massive user behavior data so as to recommend goods and services that users are actively interested in [57, 76, 152]. Meanwhile, according to the different user behaviors, the RS adjust the recommended contents with the hope to provide personalized recommendations. At present, state-of-the-art recommendation techniques can be roughly divided into CF-based, content-based and hybrid recommendation techniques [4, 22, 36, 89].

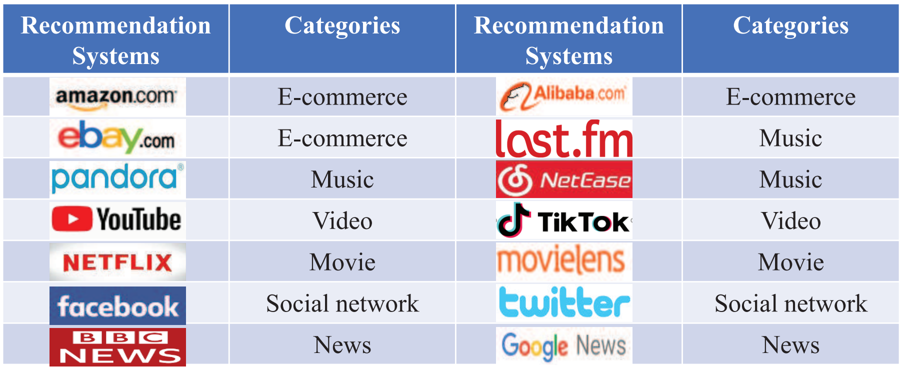
In recent years, the health topic has gained increasing attention from people around the world due to the rapid development of the modern society and the dramatic improvement of living standards. More and more people are eager to live a healthy life and maintain a healthy body state. The modern view of health tells us that health is no longer merely the absence of disease. World Health Organization (WHO) defines health as a state of complete physical, mental and social well-being but not merely the absence of disease or infirmity. On the one hand, health contributes to longevity. On the other hand, health determines the quality of life and career success to a large extent. Non-communicable diseases are currently the main killers that endanger human health. Many common non-communicable diseases, including cardiovascular diseases, cancer, chronic respiratory diseases, diabetes, etc., account for more than 63% of the total deaths worldwide. The main causes of these diseases are occupational and environmental factors, as well as bad life and behavior. Generally, most of the causes can be prevented, and behavior changes can effectively improve the current health state.

In terms of diet, people increasingly pursue green and healthy food. Physically, various health checks are becoming more frequent. People also pay more and more attention to physical exercise. However, due to busy work, bad habits, lack of health knowledge, ambiguous health-related information and other reasons, people cannot maintain a healthy life and may often make wrong decisions that affect their physical and mental health. Generally speaking, although people have different views on a healthy lifestyle, the key factors commonly considered to stay healthy are an optimistic mood, a healthy diet, and regular exercise. However, how to maintain a positive mood? What is a healthy diet for me? And how much exercise should I do every week? Health standards have different requirements for people of different ages and genders. And for different individuals, the health status they pursue is also different. Therefore, how to scientifically provide everyone with suggestions that meet their needs to help them stay/promote/improve the health condition is a huge challenge.

In recent years, the RS for health has become a hot topic in the RS community [130]. Due to the unique advantages of RS and its rapid development in recent years, experts believe that RS can aid the healthcare field by providing valuable and accurate advice, including, but not limited to, disease severity estimation, disease diagnosis and treatment, health management and promotion, behavioral change [59, 65, 189, 32, 188, 47]. At the same time, applications of the RS in the healthcare field also pose huge challenges to the RS community as never before, for example, the accuracy of estimation, reliability of diagnosis, satisfaction and diversity of recommendations, etc [46, 124, 165, 153, 71].

With the continuous development and the deepening applications of the RS in the health field, new application scenarios are emerging one after another, which brings both new opportunities and challenges to the RS community. The purpose of this chapter is to provide a state-of-the-art overview of recommendation techniques with applications to healthcare. It is aimed to provide the readers with a background on different recommendation techniques and how such techniques are applied in different health scenarios. Recalling the existing literature, a rather large number of results have been reported on 'recommendation techniques over health'. Among those results, it is found that the following topics are the most investigated: dietary (or food) recommendation, lifestyle recommendation, training recommendation, decision-making for patients/physicians, and disease-related prediction.

In this chapter, we aim to review some most common techniques in RS, which are content-based methods, CF methods and hybrid methods. We also provide their applications in the healthcare field. First, a brief introduction of RS is given in Section 2.1. After that, the content-based RS is briefly introduced in Section 2.2. Then, the background of the CF is introduced in Section 2.3. We describe two types of methods in CF which are memory-based CF and model-based CF, and then the typical examples of these two methods are shown in Subsections 2.3.1 and 2.3.1. In Section 2.4, hybrid methods are provided and several forms of hybrid systems are given. Finally, the applications of recommendation techniques in healthcare field are presented in Section 2.6.



Recommendation Systems	Categories	Recommendation Systems	Categories
amazon.com	E-commerce	Alibaba.com	E-commerce
ebay.com	E-commerce	last.fm	Music
pandora	Music	NetEase	Music
YouTube	Video	TikTok	Video
NETFLIX	Movie	movielens	Movie
facebook	Social network	twitter	Social network
BBC NEWS	News	Google News	News

Fig. 2.1 Examples of different recommendation systems in reality

2.1 Recommendation system

Given an RS consisting of m users and n items, the user profiles are denoted by a $m \times n$ matrix called the user-item matrix $R^{m \times n}$. Each row in the user-item matrix represents a user's rating for different items, and each column represents an item rated by different users. The sets of users and items are defined as $U = \{u_1, u_2, \ldots, u_m\}$ and $I = \{i_1, i_2, \ldots, i_n\}$, respectively. Each element $r_{u,i}$ in R represents that the user u rates the value r on the item i, where $u \in U$, $i \in I$. If the user u has rated the item i, then $r \in 1, 2, \ldots, \tilde{r}$ ($\tilde{r}$ is the upper bound of the ratings). Furthermore, $r_{u,i} = \emptyset$ if the user u does not rate the item i. Generally speaking, the commonly used rating scale is numerical rating scale. For example, in a 5-star numerical rating scale, the system allows respondents to rank their feedback on a 5-point scale form 1 to 5, where 1 indicates the worst score and 5 indicates the best score.

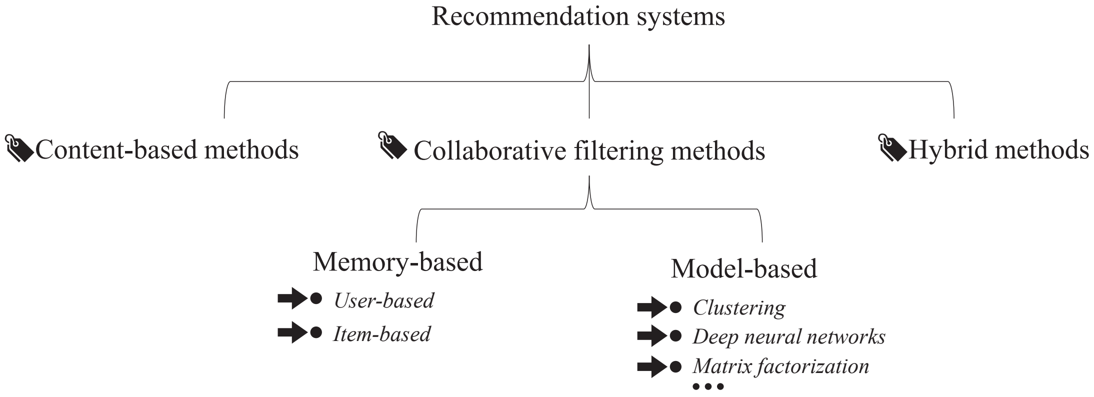


Fig. 2.2 Different types of recommendation algorithms

2.2 Content-based filtering

Content-based filtering is one of the common techniques in building RS. In a content-based RS, new items will be recommended to a user according to their item features similar to the items that this user likes. The way to determine whether a user likes an item is usually based on the user's explicit feedback and implicit feedback. There is a typical example in Fig. 2.3 to show how content-based RS works, for example, if a user has watched an action movie, then the system will suggest other action movies that the user has not watched before.

The high-level architecture of content-based RS mainly has the following three components:

- Content analyzer (feature extraction): For structured information, the content features can be easily extracted. For unstructured information, such as music or text files, pre-processing is needed to extract the content features. The main function of the content analyzer is to utilize feature extraction techniques to extract the item contents from different data sources so as to facilitate the subsequent processing.

- Profile learner (user profile learning): The user preference data are used here, including explicit feedback and implicit feedback, to build a user-specific interest model. Machine learning techniques are often used to analyze preference data to construct an accurate use profile.

- Filtering and recommendation: In this step, the recommendations are given by matching the user profile and item contents.

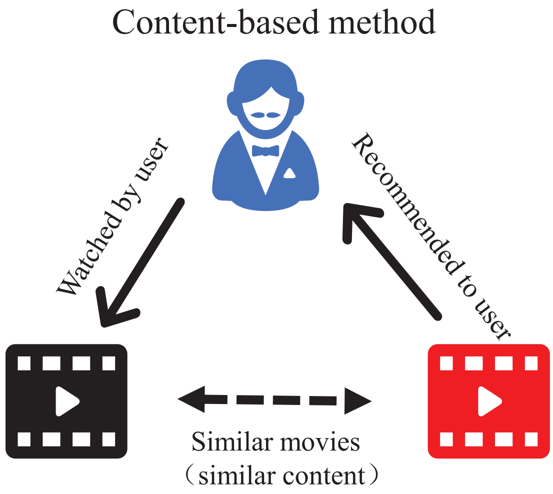


Fig. 2.3 Illustration of content-based method

To sum up, the basic principle of the content-based method is to analyze the characteristics of items that the target user has rated and then recommend new items containing these characteristics to the target user. This recommendation method only depends on content information, so it will be affected by two aspects, namely, one is limited content analysis, and the other is over-specialized. Limited content analysis means that user information and item information in the system are sometimes limited. Over-specialization refers to the lack of differentiated recommendations.

2.3 Collaborative filtering

Since the appearance of the first papers on CF in the mid-1990s, over the decades, CF-based techniques have not only been thoroughly studied in academia but also widely used in industrial circles such as Amazon, YouTube, and Netflix. Unlike the content-based filtering, the CF is to provide recommendations by processing a large amount of user behavior data collaboratively. CF only utilizes the user behavior data and does not consider the content information of the items, so it will not be restricted by limited content. As long as a user has new behaviors for different items, there are certain differences in the recommended content. Therefore, CF can effectively solve the two disadvantages of content-based filtering.

The CF is to make recommendations by analyzing a large amount of user behavior data. It is well known that in reality users tend to rate only a small number of items, so the rating matrix is often sparse. Take the example of the movie review site, many users may only rate a small number of movies, and these known ratings are referred to as observed ratings. Unknown ratings are referred to as missing ratings or unobserved ratings. More generally, the CF is to estimate unknown ratings from known ratings. According to the different data analysis methods, the well-known CF-based methods can be categorized as memory-based CF algorithms and model-based CF algorithms.

2.3.1 Memory-based CF algorithms

Memory-based CF algorithms are also known as neighborhood-based CF algorithms. Depending on the object, memory-based CF algorithms can be divided into the user-based CF and item-based CF approaches. The main idea of user-based CF is to analyze user behaviors to find a subset of users (named as neighbors) who are sharing similar tastes. Then, the items will be recommended to a target user based on his/her neighbors' tastes. Similar to the user-based CF, the item-based CF considers the similarity between items rather than users, and then recommends to a target user those items similar to the ones the active user preferred in the past.

User-based CF

Specifically, the user-based CF approach is to utilize the neighbors who are similar to user u to predict the rating $r_{u,i}$ that the user u is likely to give on item i by observing his/her neighbors' ratings on that item. Based on the previous explanation, the user-based CF approach can be divided into four parts: similarity computation, neighbor selection, rating prediction and recommendation.

- **Similarity computation**: For any user, once he interacts with the platform, his behavior data are stored in the system and described as a vector. The similarity between users can be described by the distance relationship between the vectors. This part shows how to calculate the similarity between two users by different similarity measures. Some commonly used similarity measures include:

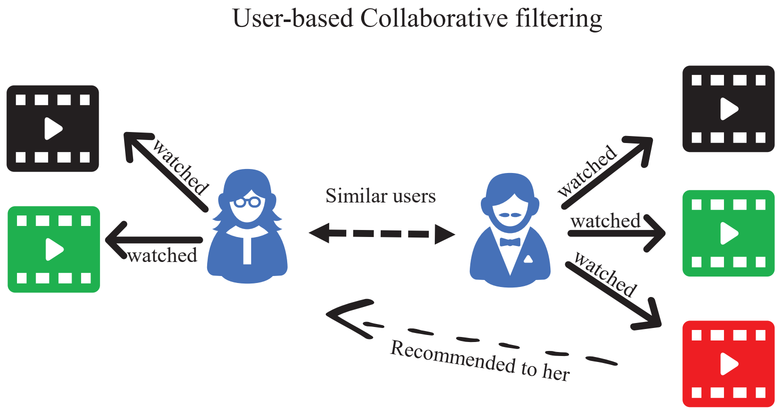


Fig. 2.4 Illustration of user-based CF method

Cosine:

$$\mathrm{Sim}_{\mathrm{COS}}(u,a) = \frac{\sum_{i\in I}(r_{u,i})(r_{a,i})}{\sqrt{\sum_{i\in I}(r_{u,i})^2}\sqrt{\sum_{i\in I}(r_{a,i})^2}} \tag{2.1}$$

where $\mathrm{Sim}_{\mathrm{COS}}(u,a)$ is the cosine similarity between users u and a; $I = I_u \cap I_a$ is the subset of items that both user u and user a have rated together, with I_u (respectively, I_a) representing all items that user u (respectively, a) has rated. $r_{a,i}$ means the rating value that user a has rated on items i.

Adjusted Cosine:

$$\mathrm{Sim}_{\mathrm{AC}}(u,a) = \frac{\sum_{i\in I}(r_{u,i}-\bar{r}_i)(r_{a,i}-\bar{r}_i)}{\sqrt{\sum_{i\in I}(r_{u,i}-\bar{r}_i)^2}\sqrt{\sum_{i\in I}(r_{a,i}-\bar{r}_i)^2}}, \tag{2.2}$$

where $\bar{r}_i$ denotes the average value of all user ratings for item i.

Pearson Correlation Coefficient:

$$\mathrm{Sim}_{\mathrm{PCC}}(u,a) = \frac{\sum_{i\in I}(r_{u,i}-\bar{r}_u)(r_{a,i}-\bar{r}_a)}{\sqrt{\sum_{i\in I}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\in I}(r_{a,i}-\bar{r}_a)^2}}, \tag{2.3}$$

$\bar{r}_u$ and $\bar{r}_a$ denote the average values of different items that users u and a have rated, respectively.

Compared to cosine similarity, adjusted cosine (AC) and Pearson correlation coefficient (PCC) do the data centralization. The difference between AC and PCC lies in the way of centralization. More specifically, AC is to subtract the average rating of all users for this item, while PCC is to subtract the average rating of all items given by the current user. Although many literatures have proved that using PCC similarity measure can bring higher prediction accuracy in most cases, which similarity measure to choose needs to be determined according to the actual situation [20, 52].

- **Neighbor selection**: This part selects the nearest k neighbors based on the similarity degree to make subsequent predictions. Neighbors are a group of users that similar to the target user. In a typical scenario, the number of k will be selected from the following three main options:

  1) *Experience-based option*: the number of neighbors is usually chosen between 20 and 50;

  2) *Experiment-based option*: the optimal number of neighbors is selected by cross-validation;

  3) *Rule-based option*: This option usually sets a rule to help select, for example, the similarity between users should be greater than 0.5 and joint rated items are more than 3.

  Rating prediction: The target user's rating of an unrated item depends on the ratings of this item by his neighbors. Assume that u_a is a set of neighbors of user u. The most commonly used rating prediction formula is:

$$\hat{r}_{u,i} = \bar{u} + \frac{\sum_{u_a} \mathrm{Sim}(u_a, u)(r_{u_a,i} - \bar{u_a})}{\sum_{u_a} \mathrm{Sim}(u_a, u)}, \tag{2.4}$$

  where $\hat{r}_{u,i}$ is the predicted value of $r_{u,i}$; $\bar{u}$ is the mean value of different items provided by user u; and $\bar{a}$ is the mean value of items provided by the user a.

  Compare to the formula that takes an average of all neighbors' ratings, (2.4) takes into account the effect of similarity between users on the results. Similarity can be seen as the weight of a neighbor's rating. The greater the weight, the more important this neighbor's rating is to the final result. The first part of the formula reflects the overall rating habit of the target user, and we can also understand this part as the initial expectation for the item. Then, the second part of the formula is to revise previous expectation by using the neighbors' ratings for this item.

- **Recommendation**: This part ranks the items according to the size of the predicted rating and then recommends the items with the highest ratings to the target user.

Item-based CF

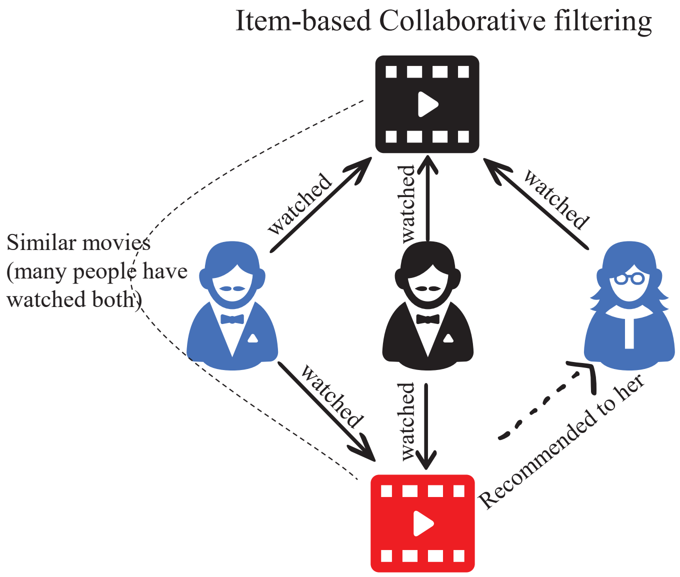


Fig. 2.5 Illustration of item-based CF method

In 1998, the concept of item-based CF was proposed and used by *Amazon.com* [88]. Item-based CF is similar to user-based CF in terms of steps and purpose, but the starting points of their hypothesis are different. The main idea of item-based CF is that you like something similar to what you like before. An example is shown in the Fig. 2.5, if many users have watched a movie A and a movie B at the same time, the item-based CF deems that movies A and B are similar. Then, if the target user has watched movie A, the item-based CF will recommend movie B to this user. To be specific, the first step of item-based CF is to determine the similarity between items. Then, when the target user behaves toward an item, the item-based CF method will recommend similar items to him.

Item-based CF and user-based CF have the same procedure. There are just some differences in similarity computation and rating prediction parts. Item-based CF considers the similarity between items i and j rather than between users u and a. In the rating prediction part, in order to predict user u's rating on item i, item-based CF is to correct the average value of all users' ratings on item *i* by using the ratings of other similar items that user *u* likes. Here, we only list similarity computation and rating prediction parts:

- **Similarity computation**: This part calculates the similarity between items and three commonly used similarity measures are as follows:

  Cosine:

  $$\mathrm{Sim}_{\mathrm{COS}}(i,j) = \frac{\sum_{u \in U}(r_{u,i})(r_{u,j})}{\sqrt{\sum_{i \in I}(r_{u,i})^2}\sqrt{\sum_{i \in I}(r_{u,j})^2}}, \tag{2.5}$$

  where $U = U_i \cap U_j$ is the subset of users who have rated both items i and j, where U_i (respectively, U_j) denotes the users who have rated item i (respectively, j).

  AC:

  $$\mathrm{Sim}_{\mathrm{AC}}(i,j) = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_u)(r_{u,j} - \bar{r}_u)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in I}(r_{u,j} - \bar{r}_u)^2}}, \tag{2.6}$$

  PCC:

  $$\mathrm{Sim}_{\mathrm{PCC}}(i,j) = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{i \in I}(r_{u,j} - \bar{r}_j)^2}}, \tag{2.7}$$

- **Rating prediction**: The target user's rating of an unrated item is based on his ratings on items that similar to this unrated item. Assume that i_j is a set of neighbors of item i. The rating prediction formula is:

  $$\hat{r}_{u,i} = \bar{i} + \frac{\sum_{i_j} \mathrm{Sim}(i_j, i)(r_{u,i_j} - \bar{i_j})}{\sum_{i_j} \mathrm{Sim}(i_j, i)}, \tag{2.8}$$

User-based CF vs. item-based CF

Although both user-based CF and item-based CF predict how user *u* will rate the item *i*, the emphasis of the two methods is different according to the previous explanation. Item-based CF usually recommends items that are similar to what have been purchased or viewed. Similarity as defined here refers to the similarity of items being purchased or rated by the users rather than the similarity of the item contents, and similarity calculations are based on either explicit feedback (rating) or implicit feedback (viewing, purchase) of users. Specifically, item-based CF does not take product attributes into account but rather calculates the similarity between items by analyzing the user behavior history data. For example, historical information shows that users who give high marks to the movie "*Justice League*" tend to like the movie "*Frozen*" even though both movies are very different in content.

Even so, if a user gave a high mark to *Justice League*, the item-based CF algorithm would recommend *Frozen* to him.

User-based CF considers the hotspots in the community. There will be many similarities between similar users, while at the same time there will be some differences, so user-based CF will recommend something novel. Item-based CF and user-based CF each have their own advantages and shortcomings. In terms of accuracy, item-based CF is usually superior to user-based CF, while in terms of variety and novelty, user-based CF is usually better than item-based CF. Therefore, user-based CF is more suitable for news recommendations, while item-based CF is more suitable for e-commerce recommendations. In terms of time complexity, the offline time complexity of UBCF and IBCF is $O(m^2 \cdot n)$ and $O(m \cdot n^2)$, respectively, where m denotes the number of users and n denotes the number of items. If there are too many users in the user-item matrix, then user-based CF will be very time-consuming in calculating the similarity between users. Correspondingly, if there are too many items, then item-based CF will be very time-consuming in calculating the similarity between items. In the actual application, which algorithm should be used will depend on the specific situations of the target requirements.

2.3.2 Model-based CF algorithms

With the advent of the big data era, the amount of data becomes larger and larger. This leads to a very large size of the user-item rating matrix. In this case, the memory-based CF methods will consume a lot of computing resources, resulting in a decrease of system performance. The model-based CF recommendation algorithms can provide faster training speed, occupy less memory and obtain better accuracy in most cases. First, the model-based CF recommendation algorithms build a model and then predict the unknown information by training the known information in the user-item rating matrix. This is similar to the traditional classification methods in machine learning, so we can find that many classification models are generalized to CF scenarios, such as decision tree, Bayesian classifier, support vector machine, neural network, and so on. However, it is very difficult for these models to complete the unknown information in the matrix when the matrix is very sparse. The matrix factorization technique has always been favored and paid attention to by researchers because of its advantages and scalability in dealing with sparse problems. Next, the basic matrix factorization technique in RS is briefly introduced.

Matrix factorization-based CF

Matrix factorization is one of the most commonly used algorithms in RSs [99, 104, 101, 102]. In the "Netflix price" contest, a large number of recommendation algorithms based on matrix factorization technology have emerged. Because of its excellent accuracy and scalability, matrix factorization technology has attracted more and more researchers' attention.

When talking about matrix factorization techniques, singular value decomposition (SVD) is the first thing that comes to mind because it has been widely used in the field of mathematics for a long time. A fatal flaw in SVD is the requirement that the user-item matrix must be dense, which is far away from the practice applications. Although some methods can be used to simply fill in missing values, it is often not effective in the face of extreme sparseness of user rating data. At the same time, when the number of users and products is very large, the computation of traditional SVD is also huge.

In 2006, Simon Funk proposed the Funk-SVD, also called latent factor model (LFM) in his blog, which has had an important impact on the RS community. Given an RS consisting of m users and n items, the user profiles are represented by a $m \times n$ matrix called the user-item matrix $R^{m \times n}$. The user-item matrix is a sparse matrix containing the part of known elements and the part of unknown elements. The LFM is a form in which user-item matrix $R^{m \times n}$ is decomposed into user matrix $P^{m \times k}$ times item matrix $Q^{k \times n}$ in order to obtain a new user-item matrix $\hat{R}$ with no missing values, where k is the dimension of the latent factor. Please note that k determines the expression ability of the hidden vector. The greater the value of k, the stronger the expression ability. In practical applications, the number of k is determined by experiments.

By adjusting the elements in user matrix P and item matrix Q to minimize the difference between the known elements in user-item matrix R and their corresponding elements in new user-item matrix $\hat{R}$, the predicted value of the unknown elements in R can be got in the new matrix. The specific mathematical expression is as follows:

$$R \simeq P \times Q = \hat{R}, \tag{2.9}$$

for the known value $r_{u,i}$ in the matrix R, its corresponding value in the matrix $\hat{R}$ is:

$$\hat{r}_{u,i} = \sum_{k=1}^{K} p_{u,k} q_{k,i}, \tag{2.10}$$

where K is the number of latent factor, $p_{u,k}$ represents the value at the uth row and kth column in P and $q_{k,i}$ represents the value at the kth row and ith column in Q.

The loss function is to minimize the SSE between true and predicted values of rated positions in R:

$$\text{Loss} = \text{argmin} \sum_{u,i} (r_{u,i} - \hat{r}_{u,i})^2 , \tag{2.11}$$

The suitable user matrix P and item matrix Q are obtained by minimizing the SSE between the known rating values in matrix R and the predicted value in the corresponding position in $\hat{R}$. After obtaining the user matrix P and the item matrix Q, the full rank matrix $\hat{R}$ is obtained by multiplying these two matrices. From the $\hat{R}$, the predicted ratings are obtained. Next, the ratings can be sorted and then recommended to users.

Improvements to the LFM include the addition of regularization terms to prevent overfitting due to oversize of an element inside user matrix P and item matrix Q. The equation is shown as follows:

$$\text{Loss} = \text{argmin} \sum_{u,i} (r_{u,i} - \hat{r}_{u,i})^2 + \lambda_p \sum_{u,k} ||p_{u,k}||^2 + \lambda_q \sum_{k,i} ||q_{k,i}||^2, \tag{2.12}$$

where λ_1 and λ_2 are the positive constants denoting the regularizing coefficients for user matrix P and item matrix Q respectively.

Considering that many properties are unique to users or items, it is necessary to consider the characteristics of users and items themselves. For example, some users have harsh ratings and some users have loose ratings, or some items have good quality and some items have good quality. Therefore, another improvement is the addition of bias terms, and (2.9) is rewritten as follows:

$$\hat{r}_{u,i} = \sigma + b_u + b_i + \sum_{k=1}^{K} p_{u,k} q_{k,i}, \tag{2.13}$$

where σ is the average value of the matrix R, b_u is the bias related to user u and b_i is the bias related to item i. The updated loss function is:

$$\text{Loss} = \text{argmin} \sum_{u,i} (r_{u,i} - \hat{r}_{u,i})^2 + \lambda_1 \sum_{u,k} ||p_{u,k}||^2 + \lambda_2 \sum_{u,k} ||q_{k,i}||^2 + \lambda_3 \sum_{u} ||b_u||^2 + \lambda_4 \sum_{i} ||b_i||^2, \tag{2.14}$$

where λ_3 and λ_4 are the positive constants denoting the regularizing coefficients for b_u and b_i.

Other advanced LFM model can be found in [103, 104, 169, 170, 136, 105, 106].

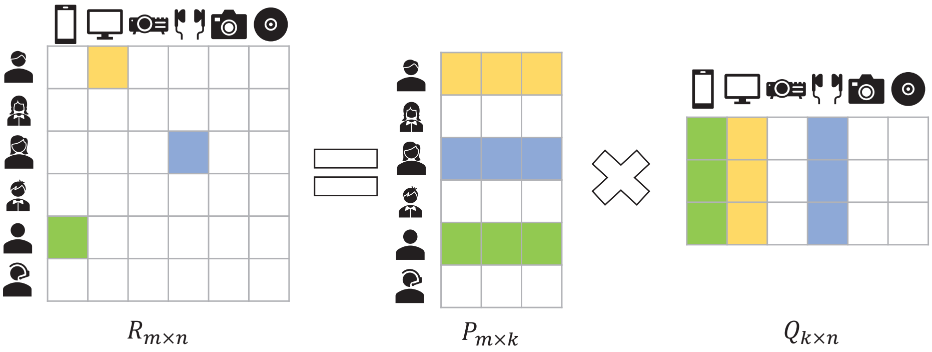


Fig. 2.6 Illustration of matrix factorization

Other model-based CF

Deep learning has flourished in areas of computer vision, pattern recognition, speech recognition and so on. In recent years, the application of deep learning in RSs has also been very widespread. A large number of deep learning models have been applied to the RSs, which provide many novel recommendation frameworks to the recommendation community and improve the performance of the RSs [190]. Deep learning can effectively mine the non-linear relationship between users and items and learn the hidden features of users and items. Some of the most commonly used deep learning models in RSs will be briefly introduced, which are multilayer perceptron (MLP), autoencoder (AE), convolutional neural network (CNN) and recurrent neural network (RNN). A comprehensive summary of the research on RSs using deep learning techniques can be found in [190].

MLP is a feed-forward artificial neural network, which contains at least one hidden layer between the input layer and the output layer. MLP is a model that represents the nonlinear mapping between input and output vectors. Most of the existing recommendation algorithms are linear methods, so the knowledge of MLP is able to provide nonlinear transformation to existing methods. Neural CF [51] is the representative work of the MLP model in RSs. Neural CF can achieve a significant result by using a neural architecture to replace the inner production on the user- and item-latent vectors in the MF model. By leveraging the neural architecture of MLP, neural CF can learn the nonlinear interaction between the users and the items. Other well-known works can be found in [86, 163, 85].

AE is a kind of artificial neural network that is used for unsupervised learning. By taking the input data as the learning target in the output layer, AE is used for dimensionality reduction or feature learning. In general, the data in the bottleneck layer represent the low dimensional form of the input data. Representative AE models are denoising AE, sparse AE, contractive AE and variational AE [74, 111, 122, 161]. AE is one of the most successful and extensive deep learning models applied in the RSs. AutoRec [126] trains the AE model so that the observed ratings in the output layer are as good as possible as the input layer. In this way, the bottleneck layer can learn the user features through the observed ratings to predict unobserved ratings. In the input layer, the ratings of unobserved part are set as 3. The loss function only considers the observed ratings without taking into account the unobserved ratings. The extension of AutoRec can be found in [139], which introduces side-information to alleviate sparsity and overcome cold start problems. Other well-known works of different AE models in RSs are [84, 150, 174].

CNN is a feed-forward neural network with deep structure. In general, CNN consists of three structures which are convolution, activation and pooling. CNN can effectively capture features from a large amount of data, so it has achieved great success in many research fields, such as speech recognition, image recognition, natural language processing, etc. The main application of CNN in the RS focuses on feature extraction of data from multiple sources, which can help the RS effectively extract the features of users and items to improve the recommendation accuracy. In the process of fashion consumption, consumers' preference for products has a great relationship with the visual appearance of products and will evolve over time. In order to provide users with more accurate recommendations, He and McAuley [49] have adopted the CNN model to extract visual features from the product images and identify evolving trends to evaluate the complex and evolving visual elements considered by the users in purchasing the products. A Deep Cooperative Neural Networks model has been proposed in [194] which consists of two parallel neural networks. One neural network consists of learning user behavior through the user comments, and the other one consists of analyzing what the user has commented to determine the features of the product. These two parallel neural networks first use the word embedding technique to obtain semantic information in comments. Then, the CNN model is used to discover multilevel features for users and items from semantic information. Finally, the two networks are coupled together, and factorization machine techniques are utilized to interact with the latent factors learned by CNN to complete the final prediction.

RNN is a kind of neural network with short-term memory capabilities to describe the relationship between the current output of a sequence and previous information. In RNN, a neuron can receive not only information from other neurons but also its own information. These neurons form a network structure with loops and memories. In general, RNN is used for sequential data processing. In session-based recommendation, RNN can be used to integrate current browsing history and browsing order to effectively model the dynamics of user preferences to provide more accurate recommendations. GRU (gated recurrent unit, a variant of RNN) has been used in [55] to model short session-based data. Other RNN methods used in session-based Recommendation can be found in [90, 171, 142].

2.4 Hybrid algorithms

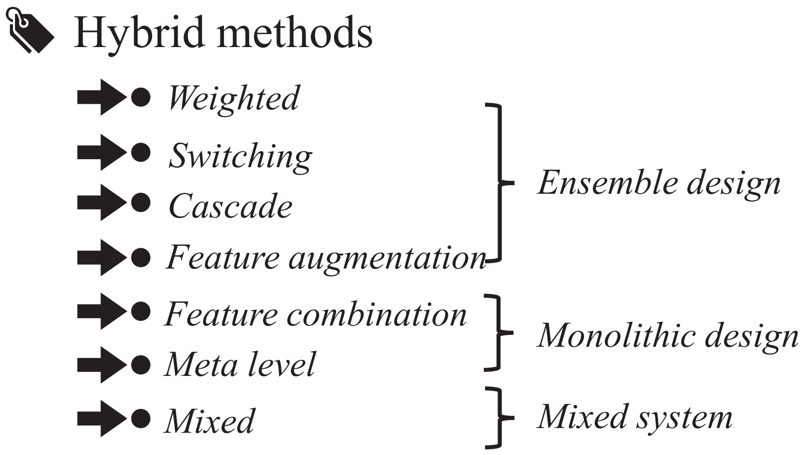


Fig. 2.7 Illustration of hybrid algorithms

Hybrid algorithms are a class of algorithms that combine the advantages of content-based and CF-based algorithms to process different data sources in order to improve prediction accuracy [155, 72]. Three primary ways of creating hybrid RSs can be found in Fig. 7, which are ensemble design, monolithic design and mixed system [6]. These three ways include 7 commonly used methods, which are weighted, switching, cascade, feature augmentation, feature combination, metal-level and mixed [22, 63, 6].

Ensemble design combines the results of different recommendation algorithms into a single output by some rules. The commonly used methods are:

- Weighted: The weighted linear combination of the prediction results from different recommendation techniques is used to obtain the final prediction result.
- Switching: According to current needs, this method switches between various recommendation techniques.
- Cascade: This is a multistage method using the sequential design that the subsequent recommendation techniques will optimize the results of the previous one in order of priority.‘
- Feature augmentation: This is another multistage method using the sequential design that the output of the previous recommendation technique is used as the input features to the subsequent recommendation technique.

Monolithic design integrates multiple recommendation strategies into one algorithm. The representative methods are:

- Feature combination: The features from different data sources, for example, the user ratings and content features, are combined together and used to do the recommendation.
- Meta-level: The model generated by the previous recommendation technique becomes the input of the subsequent recommendation technique.

Mixed system is to show a list of all the recommendation results obtained by different recommendation techniques to users.

2.5 Evaluation of RSs

Typical metrics used in RSs can be divided into three groups on the basis of their particular purposes:

- Accuracy-based metrics: MAE (mean absolute error) [19] and RMSE (root mean squared error) [14] are the most representative metrics in this group. The purpose of accuracy-based metrics is to measure the average error between the true and predicted values [44]. RMSE is more sensitive to large errors than MAE, therefore, RMSE is more useful for the system where large errors are particularly undesirable. The mathematical expression of MAE and RMSE are:

$$\text{MAE} = \frac{1}{|\mathscr{T}|} \sum_{(u,i)\in\mathscr{T}} |r_{u,i} - \hat{r}_{u,i}| \tag{2.15}$$

  and

$$\text{RMSE} = \sqrt{\frac{1}{|\mathscr{T}|} \sum_{(u,i)\in\mathscr{T}} (r_{u,i} - \hat{r}_{u,i})^2} \tag{2.16}$$

  where $|\mathscr{T}|$ represents the total number of predicted values in the testing set, $r_{u,i}$ is the true value and $\hat{r}_{u,i}$ is the predicted value.

- Decision-based metrics: The most popular metrics among these are Precision [79] and Recall [28]. The purpose of decision support metrics is to distinguish right predictions from those wrong predictions. Precision represents how many selected items are relevant and recall represents how many items are selected:

$$\text{Precision} = \frac{\text{\# of relevant recommendations}}{\text{\# of recommended items}} \tag{2.17}$$

  and

$$\text{Recall} = \frac{\text{\# of relevant recommendations}}{\text{\#of all possible relevant recommendations}} \tag{2.18}$$

  Precision (respectively, recall) takes all recommended items (respectively, all possible relevant recommendations) into account. If only top N recommendations are considered, P@N (precision at cutoff N) and R@N (recall at cutoff N) are used to represent.

- Rank-based metrics: The RS generates a recommendation list for user by analyzing his preference, and then rank-based metrics are used to evaluate the effectiveness and accuracy of this recommendation list. The most representative rank-based metrics are

MRR (mean reciprocal rank) and nDCG (normalized discounted cumulative gain) [64]. MRR measures the mean of the reciprocal ranks of multiple relevant items:

$$\mathrm{MRR} = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{\mathrm{rank}_i} \tag{2.19}$$

where $rank_i$ represents the rank position of i-th item when it first appears. And nDCG is used to measure the ranking quality compared to the ideal situation:

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p} \tag{2.20}$$

where DCG_p is the discounted cumulative gain accumulated at rank position p,

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.21}$$

and $IDCG_p$ is the ideal discounted cumulative gain accumulated at rank position p,

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.22}$$

and rel_i denotes the graded relevance of the i-th item the and REL_p denotes a list of the first p items sorted by their relevance from big to small.

2.6 Applications of recommendation techniques in Healthcare

The rapid development of the times has made people's life rhythm faster and faster. Busy work, life pressure, irregular diet, and bad habits make more and more people in a sub-health state. If the sub-health state is not improved in time, it will cause various diseases. Experts believe that the recommendation technology can help people improve their health by providing constructive personalized suggestions [159, 168]. COVID-19 pandemic in 2020 has swept across the world causing tremendous changes in people's daily life and having lasting impacts on the economy and society. The pandemic has also affected the entire medical field and promoted the rapid development of health-oriented systems, services, and solutions, among which health RS is able to play an important role in assisting professionals and individuals in clinical and non-clinical applications [160].

Before 2015, although there is not much work related to health RS, researchers have noticed the great research and application potential of health RS [127]. The successful holding of the first health RS workshop co-located with the 10th ACM Conference on RS in 2016 has provided a very good platform for communication and cooperation of researchers, which promoted the spread and development of health RS. The experts have focused on how to use the RS techniques to help people adopt a healthier lifestyle and improve their own health, including the improvement of cognition, the deepening of understanding, and the improvement of behavior.

After years of development, there is a lot of work to show that recommendation techniques have been successfully applied to disease prediction, disease prevention, medical diagnosis, and so on [81, 191, 115, 58, 116, 60, 144, 181]. Moreover, recent work has indicated that health RS has developed from the applications of basic recommendation techniques to the algorithm improvement and model innovation [62, 148]. Here, the applications of recommendation techniques in healthcare are mainly divided into the following topics:

Dietary recommendations: The choice of healthy food is affected by many aspects including culture, preferences, personality goals, and economic conditions. Improper choices will not only affect physical and mental health but also pay high economic and time costs. In general, the way that people choose healthy food is basically through active methods rather than passive recommendations. Dietary (or food) recommendation is aimed at utilizing recommendation techniques to provide users with personalized dietary recommendations based on their needs, including healthier foods, correct diet combinations, and reasonable eating methods.

In order to help people improve their health by providing healthier dietary recommendations, a method of substitution among different foods has been proposed in [2], which has used positive pointwise mutual information and truncated SVD to analyze the attributes and contextual relationships among foods with the aim to find a set of similar alternative foods that are healthier on the premise of satisfying the user needs. In [45], the content-based recommendation technique has been adopted to design and implement a personal health augmented reality assistant that can help people choose healthier alternative products in daily life. A context-aware RS has been proposed in [8]. By analyzing whether two foods have been consumed in similar situations to estimate the substitutability, a personalized dietary recommendation has been established to help people improve their eating habits instead of simply providing general dietary guidelines.

Knowledge of eating habits is the cornerstone of the personalized dietary RS. In order to explore the eating habits of users, Akkoyunlu et al. [9] have proposed a novel meal based method. This method has used Doc2Vec technology to learn the similarity of meals between users in the embedding space to determine the similarity between users and then analyze the user eating habits through the clustering method. Food RS called DIETOS (DIET Organizer System) has been mentioned in [5] for health profiling and diet management in chronic diseases by using content-based filtering technique. DIETOS has provided personalized dietary recommendations by analyzing the consumption data of healthy people and diet-related chronic disease patients. In [24], multi-objective optimization technology has been introduced into the RS to do healthy menus recommendation by considering healthy nutrients, harmonization, and coverage of ingredients in the pantry. For different practical situations, the objectives can be replaced, increased, or decreased. The knowledge-based RS technique has been used in [70] to provide personalized health advice by using medical claims, demographics, and symptoms. A mobile nutrition assistance system has been designed to provide personalized persuasion for nutritional intake [82].

Ensemble topic modeling (EnsTM) based feature identification technique has been studied in [75] to achieve effective user modeling and recipe recommendation. This technique has taken into account not only food tastes, demographics, and costs, but also user nutritional preferences, which has helped users find recipes under different nutritional categories. Alcaraz-Herrera and Palomares[10] have introduced an evolutionary RS, which has recommended diet plans and training packages to users based on their preferences and goals, with the aim to provide users with a more comprehensive experience. As the user's current preferences may conflict with the new dietary goals and lead to a decline in recommendation quality, Starke [138] has discussed how the Rasch model can be used to obtain changes in user habits to help CF-based approaches. In order to provide valuable suggestions and bring new thinking to the RS community, a series of CF and content-based algorithms have been tested on food recommendations in a large online recipe dataset with the aim to give a comprehensive analysis of different recommendation algorithms' advantages and limitations [148].

Healthy lifestyle recommendations: A healthy lifestyle can effectively improve people's health. Many experts and professional organizations have given suggestions and standards for healthy lifestyles. However, many people are unwilling to adhere to a healthy lifestyle. Even if some people make a series of health plans, many of them may give up halfway due to boredom. The personalized recommendation is an effective way to promote their lifestyles.

A cyber-physical RS has been proposed in order to allow people to actively participate in exergames rather than to immerse themselves in electronic media (smartphone and Internet) in their spare time [7]. Exergaming is a form of physical exercise that can combine sports and games to achieve the ultimate workout. This cyber-physical RS first has collected users' measurable and implicit indicators through smartphone sensing technology, then analyzed their preferences through the RS technology, and finally recommended the appropriate exergames.

Siriaraya et al.[137] have introduced the project they are developing, which is a mobile app used to record three happy things that are happened to users every day. The CF technique has been used to analyze the historical behavior of a large number of users to find neighbors with similar interests for the target users. Then, interesting activities and places nearby will be recommended to the users according to their neighbors' interests via mobile app, thus helping users increase their sense of well-being.

Training recommendations: In order to achieve the ideal physical state, many people may sign up for training programs or make individual exercise programs. However, users may suddenly abandon the training programs because of decreased motivation and lack of enthusiasm, which will make the training results fall short of success. In [157], some model-based CF methods have been used to predict whether the users will give up their training plan by analyzing the user behavior changes. If any abnormal situation has been found, the RS will remind the coach in time. For the actual remote fitness platform *e4fit*, the coach sometimes has to be responsible for multiple students and may not be able to help the students in a timely manner. Boratto et al. [18] have adopted the RS technique to find the problem in time by analyzing the student behavior data. Then, any problems have found that will be notified to the coach, which not only effectively improves the user experience but also reduce the burden on the coach.

The basic neighborhood-based CF method has been adopted to analyze the training plans of a large number of runners to recommend appropriate elite training plans and competition strategies to target users with the hope to make significant progress in a short time [15]. The impact of two dimensions of visual aesthetics (classical and expressive) has been discussed in

[151] on perceived credibility in fitness applications, which has provided valuable suggestions for the design of health RS. A hybrid RS combining content-based filtering technique and neighborhood-based CF technique has been proposed in [34]. By analyzing user preferences, historical viewing information, and like-minded users' information, the proposed hybrid RS has recommended tailored fitness videos for users.

Decision-making for patients/physicians: *For patients*: An argumentation-based RS called ArgoRec has been proposed to provide complex chronic patients with personalized recommendations to effectively support their daily activities. ArgoRec has utilized argumentation for leveraging explanatory power and natural language interaction to improve the patient experience and recommendation quality [35]. In [42], the neighborhood-based CF has been applied to clinical decision support systems with the aim to provide the best-personalized treatment plan for psoriasis patients.

For medical staffs: By analyzing the patient behavior, the most suitable patient ranking list has been proposed to nursing staffs for increasing the number of closed care-gaps of patients [147]. In order to provide consumers with timely and personalized suggestions to improve the consumer's medical experience, a mixed technology considering probabilistic graphical models (PGM), random forest (RF), and CF techniques, has been proposed in [62] to obtain a vector of recommendations. Then, an ensembler has been used to combine the results and decide which results will be recommended to users.

Disease-related prediction: The essence of the RS can be seen as predicting unknown data through the analysis of known data, so a large number of recommendation algorithms are used for disease-related prediction work [11]. Most of these papers are based on the assumption of "Similar users will have similar preferences for items" in which users and items will be replaced by different patients and disease-related items. CircRNA, as a marker of many diseases, is often used to identify the correlation with diseases. Lei et al. [81] have employed a CF-based recommendation algorithm to predict circRNA–disease associations. Based on the assumption that similar cell lines and similar drugs exhibit similar drug responses, a hybrid interpolation weighted CF method has been adopted to predict the missing drug response [191].

Other aspects: In addition to the five main aspects mentioned above, some work has been proposed on the health RS improvement, sleep improvement, smoking cessation, and so on. For improving user trust and overall experience of health RS, the prediction uncertainty has been fully discussed in [54], which has made the recommendations of user health-related behaviors more transparent. In order to increase the understanding of health RS, Torkamaan and Ziegler[146] have discussed multi-criteria grading with the aim to analyze the criteria that users should consider when evaluating health promotion recommendations.

Context-aware lifestyle RS has been proposed to improve sleep [154]. In [56], a hybrid RS has been used in the smoking cessation app. This app can push personalized information to users at the right time to help them strengthen their confidence in quitting smoking. There has been also a study designing hybrid RS through merging trust with health-sensitive semantic information in a complex environment to accurately discover and recommend the great potential collaborators to help medical product development [21]. Adaji et al. [3] have discussed how hedonistic and meritocracy values affect their healthy shopping habits among people of different ages. Pasta et al. [156] have extended the application to hearing aids. By analyzing user preferences, the personalized hearing aid parameters have been configured for users. More than 85 percent of participants have shown that their user experience has been improved.

Chapter 3

A Collaborative-Filtering-Based Data Collection Strategy for Friedreich's Ataxia

3.1 Motivation

Friedreich's ataxia (FRDA) is the most common hereditary ataxia and was first identified by a German pathologist and neurologist, Nikolaus Friedreich, in 1863 [38]. FRDA is an autosomal recessive disease affecting the multisystem of the body regarding both neurological and non-neurological cases, which consists of degenerative symptoms characterized by the loss of position sense, muscle weakness, deep sensory loss, impaired coordination, dysarthria, etc, see [25, 48]. To date, there have been no effective treatments for FRDA. To investigate the FRDA in a comprehensive way, a group of experts formed the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) in 2010 with aim to build the first representative international European patient FRDA registry. EFACTS is committed to collecting and analyzing different kinds of baseline data from FRDA patients including demographics, clinical rating scales, quality-of-life measures, etc., which are of significant use in clinical trials. Most baseline data is collected through interviewing and observing the patients without invasiveness except the blood sample for genetic testing.

A cross-sectional analysis in [118] has been reported, followed by 2-year study in [119] using the FRDA baseline data from EFACTS. According to these earlier results, only 592 patients have been included in the database during the period from 15 September 2010 to 30

April 2013, and the number of patients has increased to 605 by November 2013. Although the database has been continuously updated, the speed of the data collection remains low, which gives rise to the difficulties in making sense of data analysis. For example, only 949 FRDA patients have been involved in the database as of 31 May 2018, and the collected data is insufficient in the statistical sense. Note that the current data collection relies on the physical presence of the patients at the study sites who can take the tests, answer disease-related questions and fill the clinical rating scales. Although EFACTS has 19 study sites in 9 countries now, the coverage is still far from adequate. What's worse, the FRDA symptoms are degenerating and the FRDA patients are usually progressively in poor physical conditions. As in the experts' opinions, the dilatory increase of the FRDA data size can be mainly attributed to physical reasons such as mobility, psychological, economic ones. Furthermore, according to the morbidity rate, there are as many as 28,000 FRDA patients across Europe, but there are only less than 3.5% of them in the EFACTS database. As such, there is an urgent need to improve the data collection for both new patients and follow-up assessments with hope to enrich the baseline data required by the FRDA researchers to make meaningful research.

FRDA is a rare disease, in addition there is currently no effective therapy method to cure this disease. For FRDA researchers, more baseline data would provide better bases for research into this disease. Unfortunately, existing clinical studies suffer mostly from the fact that sample sizes that are too small to be significant [107]. Many existing clinical trials have shown tantalizing positive improvements in small numbers of FRDA patients, while the true therapeutic potential of drug candidates still needs to be assessed on larger sample numbers (including longitudinal studies and different onset ages) to lead to approval of a therapeutic agent [108, 129, 109, 31]. In statistical analysis, large FRDA baseline data can be used for discovering underlying patterns. Clearly, more significant statistical results can help identify the health problems (occurring throughout the disease course), predict disease progression and prognosis in a more reliable way [114, 131]. Motivated by the above discussions, it is of vital importance to construct a practical and effective data collection strategy for enlarging the sample size that facilitates the subsequent clinical analysis.

In this chapter, to improve EFACTS baseline data collection, we aim to provide a novel strategy by adopting the idea of the nowadays popular recommendation system (RS) which is based on the fact that similar patients have similar symptoms on each test-item. Note that the main advantages of using RS is that not all ratings are required for all testing-items, which means that the patients only need to provide some basic ratings in case they are unable to go to the EFACTS' study sites. By resorting to the RS, instead of having very limited data, the FRDA researchers would be provided with certain predicted baseline data (produced by the proposed RS) on patients who *cannot* attend the assessments for physical/psychological reasons, thereby helping with the data analysis from the researchers' perspective. When using RS, those important data from unattended patients are regarded as missing, and these missing data can be predicted from RS with help from the testing ratings of the "similar" patients.

A typical recommendation consists of users, items, and users' ratings for certain items [143]. The role of the recommendation system is to estimate the user preferences and model the connection between users and items, thereby predicting ratings of the test-items. In the context of the FRDA date collection, the patient is treated as a "user", the test-item is regarded as an "item", and therefore the proposed prediction strategy is in the structure of a recommendation system [191], which is named as the FRDA Rating Recommendation System (FRRS). Collaborative filtering (CF) is a well-known technique used by recommendation systems [121] and has been widely applied in a number of commercial companies such as Amazon, YouTube, Netflix, etc.

Generally speaking, there are two common neighborhood-based CF approaches used in the recommendation systems, namely, the user-based CF and item-based CF. The user-based CF is one of the neighborhood-based CF [53] which can make automatic value prediction of a current user by using the information collected from other similar users. In contrast, the item-based CF approach focuses on the similarity among items (rather than the similarity among users). Once the neighborhood-based CF method is selected, the patients are only asked to provide certain/essential values of testing items, and then the similarities among patients or test-items in the database can be calculated. According to the obtained similarities, similar patients or test-items can be determined corresponding to the patients or test-items with incomplete information. By doing so, the missing values can be adequately predicted by using the neighborhood-based CF.

In this chapter, we use the CF approaches to predict baseline data according to the similarity in test-items of the patients, where the prediction accuracy is evaluated based on three rating scales selected from the EFACTS database. Experimental results demonstrate the validity and efficiency of the proposed strategy. The main contributions of this chapter are summarized as follows. 1) a novel data collection strategy is proposed for the FRDA disease based on the neighborhood-based CF approaches. The proposed strategy makes it possible to construct a new framework for the data collection by using popular CF methods. 2) the proposed strategy does not have any geographic restrictions on the required data and, as a result of the developed framework, the predicted data provides an alternative database for FRDA data analysis that would potentially assist clinical trials. 3) extensive experiments are conducted to verify the feasibility and the effectiveness of our strategy using neighborhood-based CF approach..

The remainder of this chapter is organized as follows. In Section 3.2, an FRDA baseline data collection mechanism is introduced. Section 3.3 presents CF algorithms including the similarity computation, the similar neighbors selection and the FRDA score prediction. In Section 3.4, the experimental results are discussed and, finally, concluding remarks are given in Section 3.5.

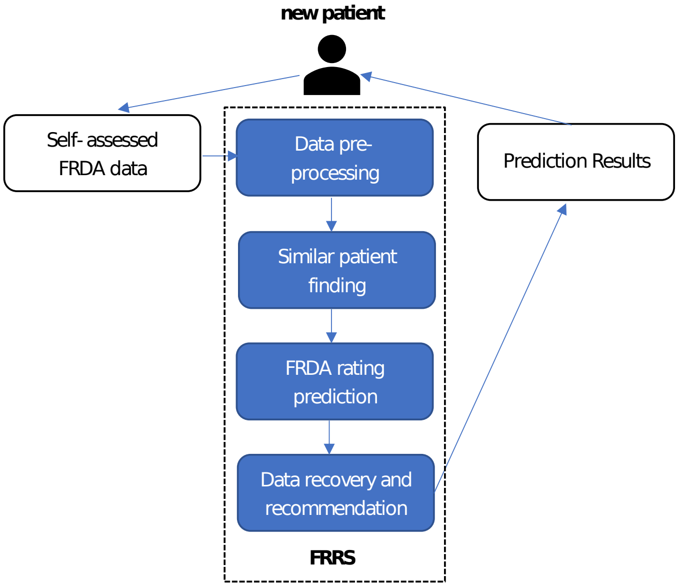


Fig. 3.1 Mechanism of the FRRS.

3.2 FRDA baseline data collection

The data for the EFACTS patient study has been collected through patient interviews, questionnaires, observations and coordinated tests at the study sites. Nevertheless, most FRDA patients (and their families) can be relied upon to provide accurate information on some features of their disease by long-term observation and dedicated caring. Patients (and their families) recognize patients' physical conditions and the collected data is this accurate enough for analysis purpose. For example, for the test-item “swallowing” in the ADL form, it is divided into 5 levels, namely, normal, rare choking, frequent choking, soft food required and feeding tube/gastrostomy. After a long period of self-cognition and care, patients (and their families) are best placed to assess their level of impairment in such kind of test-items. As such, it is natural in our strategy to assume that those patients (and their families) who contribute more of the required data will obtain more accurate predictions on the test-item values. Fig. 3.1 shows the FRDA baseline data collection mechanism, which is elaborated as follows:

Table 3.1 Rating data formulation in SARA

Gait	Stance	Sitting	Speech disturbance	Finger chase			Nose-finger test		
				right	left	mean	right	left	mean
0 ∼ 8	0 ∼ 6	0 ∼ 4	0 ∼ 6	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4

Fast alternating hand movements			Heel-shin slide			SARA Total
right	left	mean	right	left	mean	
0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 40

Table 3.2 Rating data formulation in ADL

Speech	Swallowing	Cutting food&Use of Cutlery	Dressing	Personal hygiene	Falls	Walking
0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4	0 ∼ 4

Sitting	Bladder function	ADL Total
0 ∼ 4	0 ∼ 4	0 ∼ 36

1. New patients or follow-up patients contribute some certain FRDA baseline data to FRRS by some ways provided by EFACTS. Unfilled parts regard as missing values.

2. FRRS pre-processes the data into the same interval automatically (details will be shown in Section 3.4.A).

3. FRRS compares added baseline data from new/follow-up patients with existing ones to find the most similar neighbors. These similar neighbors are used for making missing value predictions for the new/follow-up patients (technique details will be introduced in Section 3.3.A and 3.3.B).

4. FRRS predicts the missing values by nowadays popular CF methods (see Section 3.3.C for technique details).

5. FRRS feeds the predicted data back into the original data state. These processed data are employed for FRDA rating recommendation (see Section 3.4.C for more details).

6. New patients use the predicted values from FRRS as well as the recommendation results to assist their decision making on missing parts. Once the values are determined, these values will be stored in the database and used to assess clinical trial results, select samples and identify the health problems.

Remark 3.1 *The data collection strategy we propose exhibits the following distinctive features: 1) as opposed to going to study site physically which is extremely inconvenient for FRDA patients, they are now given more channels (for instance, through internet and telephone) to submit their information without geographic restrictions; 2) the proposed strategy is also applicable to other diseases baseline data collection, especially in rating scales, for example, some famous scales like International Cooperative Ataxia Rating Scale, Hamilton Rating Scale for Depression and National Institutes of Health Stroke Scale.*

Table 3.3 Rating data formulation in INAS

Reflexes				Reported Abnormalities		
Biceps(BTR)	Patellar(PTR)	Achilles(ATR)	Extensor plantar reflex	Double vision	Urinary dysfunction	Cognitive impairment
$0 \sim 3$	$0 \sim 3$	$0 \sim 3$	$0 \sim 3$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$

Ophthalmological findings			Sensory symptoms	
Broken up smooth pursuit	Square wave jerks on fixation	Downbeat-nystagmus on fixation	Impaired vibration sense	
$0 \sim 1$	$0 \sim 1$	$0 \sim 1$	$0 \sim 4$ (Right foot)	$0 \sim 4$ (Left foot)

Motor symptoms							
	Spasticity		Paresis				
Gait	Upper Limbs	Lower Limbs	Face/tongue	UL proximal	UL distal	LL proximal	LL distal
$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$
Fasciculations			Muscle atrophy				
Face/tongue	Upper Limbs	Lower Limbs	Face/tongue	UL proximal	UL distal	LL proximal	LL distal
$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$

3.3 CF algorithms

In this section, the user-based and item-based CF approaches are introduced to establish our strategy, which can be divided into three steps, namely, the computation of the similarity, the selection of the similar neighbors, and the prediction of the missing value. The FRRS consists of p FRDA patients and q test-items, where the relationship between patients and test-items is denoted by a matrix of the dimension $p \times q$, which is referred to as the patient-item matrix R. Define the sets of patients as $U = \{u_1, u_2, ..., u_p\}$, and sets of test-items as $I = \{i_1, i_2, ..., i_q\}$. Each element $r_{u,i}$ in this matrix represents rating value r of the patient u on the test-item i, where $u \in U$, $i \in I$. $r \in 1, ..., |r|$ if the item has been rated by the patient, and $r_{u,i} = \emptyset$ if the item has not been rated. Note that it is unavoidable to have randomly missing values in the patient-item matrix because the patients (or their families) might have different levels of awareness on different test-items.

3.3.1 Computation of the similarity

In this section, the similarity between the *u*th patient and the *a*th patient is measured by the Pearson Correlation Coefficient (PCC). The PCC is a popular index for similarity computing in CF that has been attracting an ongoing research interest, see [20, 52] for some representative works. A typical user-based PCC is exploited to reflect the similarity between the *u*-th patient and the *a*-th patient based on their co-ratings on the test-items by the following definition:

$$\mathrm{Sim}(u,a) \triangleq \frac{\sum_{i\in I}(r_{u,i}-\bar{r}_u)(r_{a,i}-\bar{r}_a)}{\sqrt{\sum_{i\in I}(r_{u,a}-\bar{r}_u)^2}\sqrt{\sum_{i\in I}(r_{a,i}-\bar{r}_a)^2}} \tag{3.1}$$

where $\mathrm{Sim}(u,a)$ is the similarity between FRDA patients u and a; $I = I_u \cap I_K$ is the subset of test-items on which both patients u and a have rated, with I_u (respectively, I_k) denoting all test-items that patient u (respectively, a) has evaluated; $r_{u,i}$ (respectively, $r_{a,i}$) is a vector of ratings of test-item i provided by patient u (respectively, a); and $\bar{r}_u$ (respectively, $\bar{r}_a$) denotes the average scores on different test-items that patient u (respectively, a) has rated. It follows from (3.1) that the similarity of two patients is in the interval of $[-1,1]$, where a higher similarity indicates that patients u and a are more similar.

Another neighborhood-based CF method, known as item-based CF, focuses on the similarity between test-items i and j with definition given as follows:

$$\mathrm{Sim}(i,j) \triangleq \frac{\sum_{u\in U}(r_{u,i}-\bar{r}_i)(r_{u,j}-\bar{r}_j)}{\sqrt{\sum_{u\in U}(r_{u,i}-\bar{r}_i)^2}\sqrt{\sum_{u\in U}(r_{u,j}-\bar{r}_j)^2}} \tag{3.2}$$

where $\mathrm{Sim}(i,j)$ represents the similarity between test-items i and j; $U = U_i \cap U_j$ is the subset of patients who have provided ratings on both test-items i and j; $r_{u,i}$ is a vector of ratings of test-item i provided by patient u and $r_{u,j}$ represents a vector of ratings of test-item j provided by patient u; and $\bar{r}_i$ and $\bar{r}_j$ are the average scores of test-items i and j provided by all patients, respectively. It can be seen from (3.2) that, as with the user-based PCC, the similarity of two test-items is also in the interval of $[-1,1]$.

3.3.2 Selection of similar neighbors

The basic idea of the user-based CF is that similar patients have similar ratings, and it is therefore of great importance to identify the "similar" neighbors for a certain new patient whose rating information is incomplete [30, 31]. For the purpose of choosing the similar patients, we employ a top-n algorithm [197], in which we rank the similarities between users in our database from high to low and then select the n patient with highest ratings. In case that the patients do not have enough similar neighbors, the top-n algorithm will start to use dissimilar neighbors to fill the vacancies which would unavoidably induce the inaccuracy. In order to avoid such a problem, we propose some conditions on the user-based top-n algorithm as follows. Let

$$\hat{S}(a) = \{a_u | a_u \in T(a), \mathrm{Sim}(a_u, a) > 0, a_u \neq a\}, \tag{3.3}$$

where $\hat{S}(a)$ denotes the set of similar patients for patient a that we choose to use in the following experiments and $T(a)$ is the set of top n similar patients of patient a. Similarly, the condition on the item-based top-n selection is given by

$$\hat{S}(j) = \{j_i | j_i \in T(j), \mathrm{Sim}(j_i, j) > 0, j_i \neq j\}, \tag{3.4}$$

where $\hat{S}(j)$ represents the set of similar test-items for test-item j that we select to employ in the following experiments and $T(j)$ is the set of top n similar test-items for test-item j. The neighbors that are not among the top n or whose correlation coefficients are below 0 will be removed from the similar neighbor sets.

3.3.3 Prediction of the FRDA score

Prediction constitutes the most important step in CF algorithm [125]. After the top n neighbors of the patient are determined, the values of unfilled test-items can be predicted by combining similar patients $\hat{S}(a)$ according to the following [20]:

$$\hat{r}_{a,j} = \bar{a} + \frac{\sum_{a_u \in \hat{S}(a)} \mathrm{Sim}(a_u, a)(r_{a_u,j} - \bar{a}_u)}{\sum_{a_u \in \hat{S}(a)} \mathrm{Sim}(a_u, a)}, \tag{3.5}$$

where $P(r_{a,j})$ denotes a vector of predicted values of the missing value $r_{a,j}$ in the patient-item matrix; $\bar{a}$ is a vector of average values of different test-items provided by patient a, and $\bar{a}_u$ is a vector of average values of test-items provided by the similar patient a_u.

The missing value prediction of item-based CF can be calculated as follows:

$$\hat{r}_{a,j} = \bar{j} + \frac{\sum_{j_i \in \hat{S}(j)} \mathrm{Sim}(j_i, j)(r_{a,j_i} - \bar{j}_a)}{\sum_{j_i \in \hat{S}(j)} \mathrm{Sim}(j_i, j)}, \tag{3.6}$$

where $\bar{j}$ denotes a vector of average values of test-item j provided by different patients; r_{a,j_i} is a vector of rating values of similar test-item j_i provided by the patient a; and $\bar{j_i}$ is a vector of average values of test-items provided by the similar test-item j_i.

Table 3.4 MAE and RMSE of User-based CF and Item-based CF from density 20% to 80%.

	User-based CF				Item-based CF			
	SARA	ADL	INAS	SAI	SARA	ADL	INAS	SAI
20%								
MAE	0.2002	0.2182	0.2207	0.2348	0.2051	0.2292	0.1857	0.1841
RMSE	0.2714	0.3049	0.3313	0.3235	0.2637	0.2623	0.2880	0.2741
30%								
MAE	0.1906	0.2063	0.1990	0.2111	0.1842	0.1962	0.1662	0.1654
RMSE	0.2548	0.2853	0.3091	0.3030	0.2463	0.2392	0.2706	0.2599
40%								
MAE	0.1811	0.1938	0.1809	0.1901	0.1640	0.1877	0.1534	0.1509
RMSE	0.2402	0.2653	0.2897	0.2860	0.2219	0.2319	0.2551	0.2427
50%								
MAE	0.1715	0.1829	0.1621	0.1620	0.1493	0.1696	0.1453	0.1346
RMSE	0.2269	0.2481	0.2712	0.2626	0.2037	0.2190	0.2509	0.2225
60%								
MAE	0.1627	0.1734	0.1434	0.1426	0.1401	0.1607	0.1386	0.1283
RMSE	0.2147	0.2331	0.2543	0.2416	0.1867	0.2057	0.2390	0.2156
70%								
MAE	0.1478	0.1664	0.1317	0.1314	0.1315	0.1601	0.1251	0.1261
RMSE	0.1976	0.2226	0.2422	0.2281	0.1811	0.2055	0.2252	0.2109
80%								
MAE	0.1282	0.1569	0.1235	0.1229	0.1192	0.1549	0.1231	0.1225
RMSE	0.1750	0.2107	0.2321	0.2197	0.1643	0.2018	0.2203	0.2060

3.4 Implementation and experiments

In order to evaluate the performance of the proposed algorithms, three datasets have been chosen from EFACTS database, which are the datasets for Scale for the Assessment and Rating of Ataxia (SARA), Activities of Daily Living (ADL) and Inventory of Non-Ataxia Symptoms (INAS) with respect to clinical rating scales and quality-of-life measures. These three datasets include 80 test-items and 949 patients, and the details of each dataset are explained as follows:

- SARA is an important clinical rating scale reflecting the severity of ataxia symptoms. The reliability and validity of SARA have been confirmed in [175] and [119] on both ataxia and FRDA severity. As shown in Table I, there are 8 test-items, namely, gait, stance, sitting, speech disturbance, finger chase, nose-finger test, fast alternating hand movements and heel-shin slide. The last four test-items have three sub-items: left part, right part and the mean of both parts. The total SARA scores range from 0 to 40 and are calculated by summing scores on 8 test-items, where a larger score indicates a more severe state of the patient. In our experiments, we use the left and right parts of last four test-items, which mean that 12 test-items are considered.

- ADL is another important rating scale used for assessing the impairment of patients' activities in daily living. ADL is a good complement to SARA and they two to provide the insight on how the severity of FRDA symptoms interferes with patients' daily activities and quality of life. The 9 ADL test-items (e.g., speech, dressing, and sitting) are shown in Table II with the total scores ranging from 0 to 36.
- INAS provides a checklist of non-ataxia signs in FRDA which is used to find corresponding non-ataxia symptoms with different orders of FRDA severity. Non-ataxia symptoms have 5 main features (i.e., reflexes, motor symptoms, sensory symptoms, ophthalmological findings and reported abnormalities) which contain a total of 63 test-items. Some of the major items are displayed in Table III.

Based on the three selected datasets, we carry out the experiment as detailed in the following subsections to show the effectiveness and merits of the proposed neighborhood-based CF algorithms.

3.4.1 Data preprocessing

The chosen three datasets are mutually independent as they are reported on three separate rating forms. Nevertheless, there seems to have potential relationships between the test-items of different datasets and, with this concern in mind, we decide to carry out experiments not only on each dataset separately but also on a combination of the three datasets. As shown in Tables I, II and III, the test-items have different rating intervals, so we normalize the rating values to a notionally common scale in the interval of $[0,1]$ by using min-max normalization as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{3.7}$$

where x' is a vector of normalized values; x is the original value from each test-item; and x_{min} and x_{max} are the minimum and maximum values in x given its range.

3.4.2 Experiment setup

There are totally four patient-item matrices in our experiments, namely, SARA patient-item matrix of dimension 949×12, ADL patient-item matrix of dimension 949×9, INAS patient-item matrix of dimension 949×63, and SARA-ADL-INAS (SAI) patient-item matrix of dimension 949×80. In order to verify the feasibility of the algorithm, we randomly remove different numbers of entries and form the testing data. As such, the patient-item matrices with different sparse densities (e.g. 70% and 80%) are constructed during our experiment.

For the propose of measuring prediction accuracy of the algorithm, the indices of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are taken into account, where MAE is defined as

$$\mathrm{MAE} = \frac{1}{N} \sum_{a \in A_d} \sum_{j \in J_d} |r_{a,j} - \hat{r}_{a,j}|, \tag{3.8}$$

and RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{a \in A_d} \sum_{j \in J_d} (r_{a,j} - \hat{r}_{a,j})^2}, \tag{3.9}$$

where N denotes the total number of predicted values, A_d is the user set of the testing data and J_d is the test-item set of the testing data, $r_{a,j}$ is the actual value of test-item j provided by patient a, and $\hat{r}_{a,j}$ denotes the predicted value from the user-based CF.

3.4.3 Results and discussion

To demonstrate the validity of the proposed algorithms, user-based CF and item-based CF have been tested in four patient-item matrices with densities from 80% to 20% incremented by the interval of 10%, where the density refers to the ratio of number of entries presented to the total number of the entries in the patient-item matrix. For example, the way that we generate the density of 80% is to randomly keep 80% of the entries in the patient-item matrix.

In reality, unlike the recommendation system with abundant users/items, FRDA data are quite limited in a quantitative at the moment, so we assume that there is indeed the sparsity (e.g., densities of 20% and 30%) with the dataset which would result in few co-rating

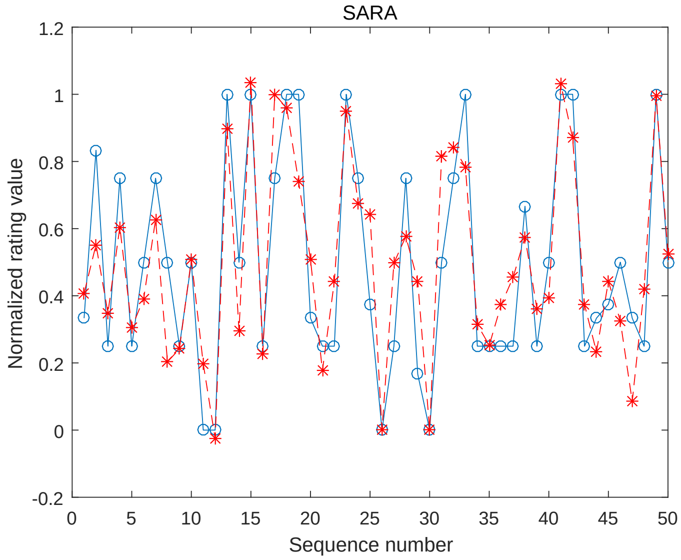


Fig. 3.2 Prediction for partial testing data (SARA dataset).

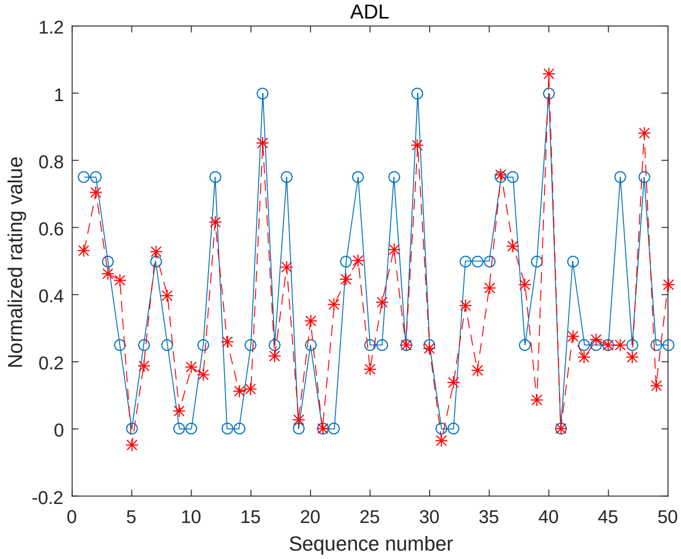


Fig. 3.3 Prediction for partial testing data (ADL dataset).

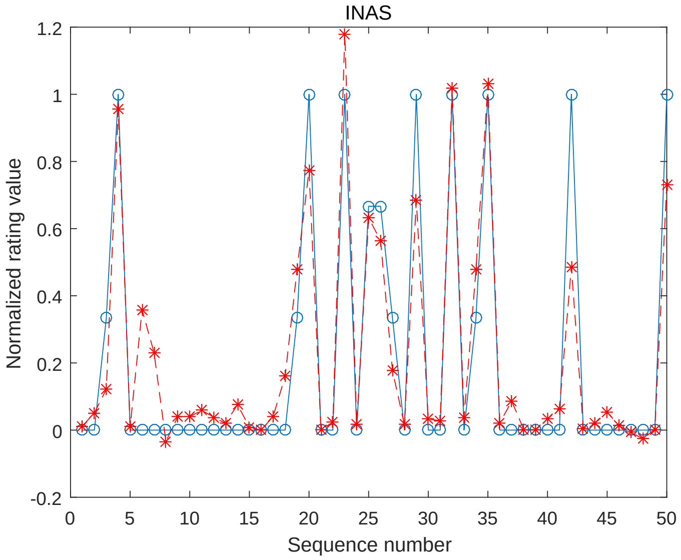


Fig. 3.4 Prediction for partial testing data (INAS dataset).

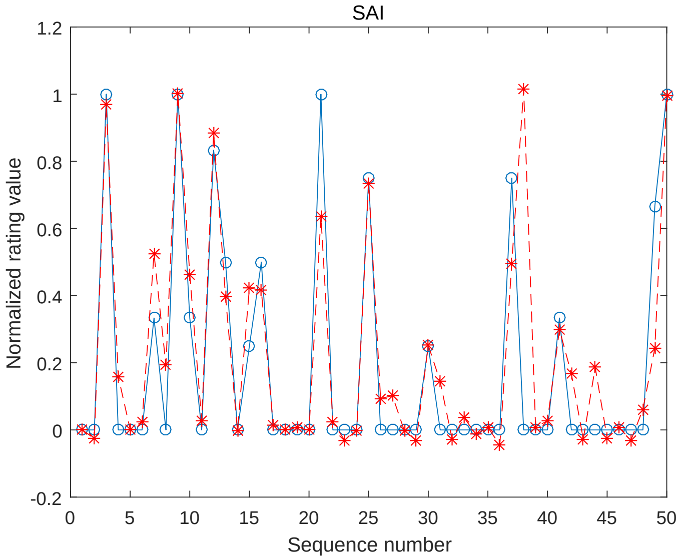


Fig. 3.5 Prediction for partial testing data (All datasets).

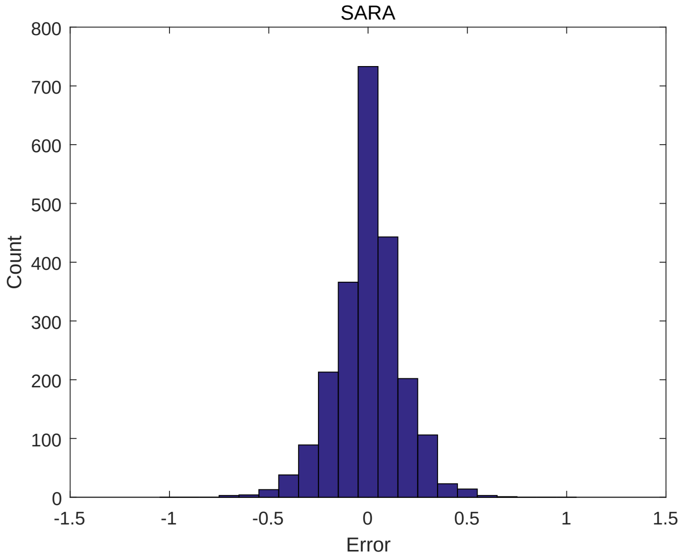


Fig. 3.6 Histograms of prediction error (SARA dataset).

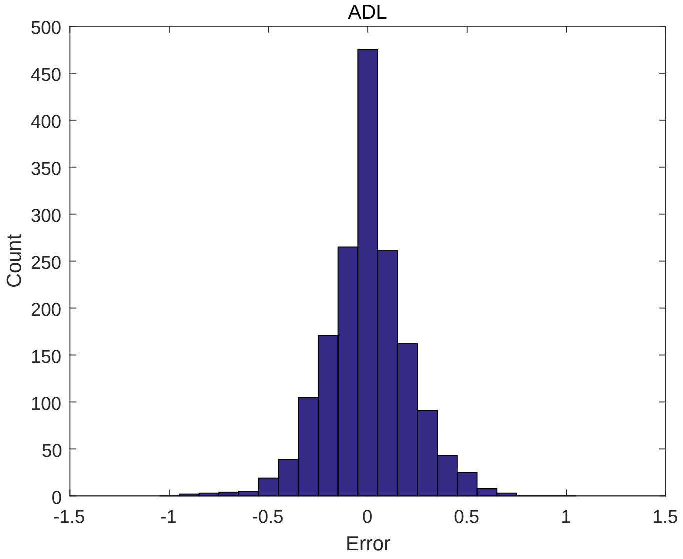


Fig. 3.7 Histograms of prediction error (ADL dataset).

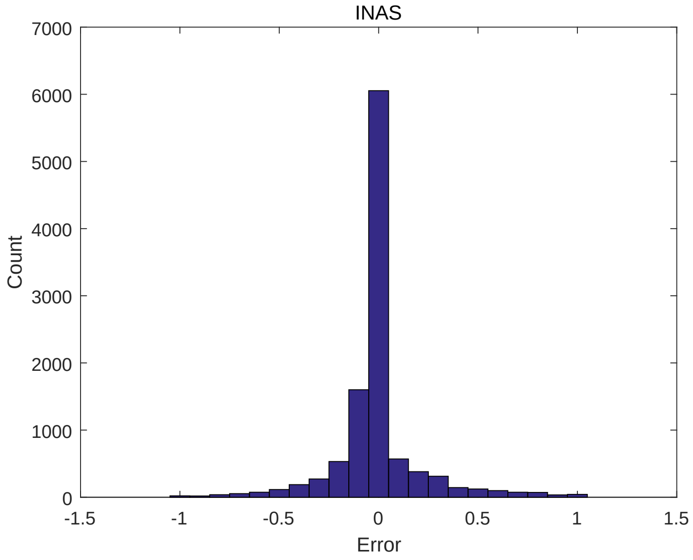


Fig. 3.8 Histograms of prediction error (INAS dataset).

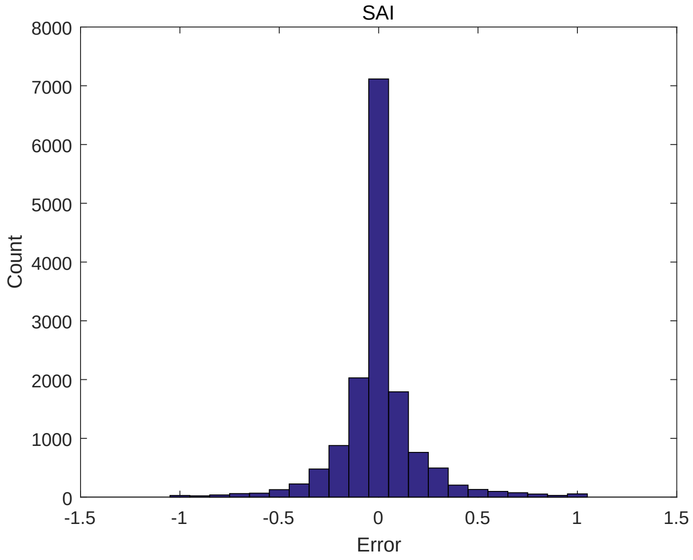


Fig. 3.9 Histograms of prediction error (ALL datasets).

test-items and therefore certain inaccuracy of the prediction. Nonetheless, for the benefits of comprehensive analysis, it would be interesting to see how the MAE/RMSE behaves in the case of different degrees of sparsity.

We set the value of top-n as 5 in user-based CF, which means that the top 5 similar patients will be considered in missing value prediction. We set top-n as 5 in item-based CF, which means the top 5 similar test-items will be considered. Each experiment is repeated 100 times, and the average MAE and RMSE are used to show the estimation error. The MAE and RMSE of the experimental results are displayed in Table 3.4, from which we observe that

1. The user-based CF and item-based CF can provide satisfactory prediction accuracy with relatively small MAE (around 0.12) and RMSE (around 0.16).

2. According to the trend of MAE/RMSE with the density varying from $80\% - 20\%$, we can draw a conclusion that, with more values of test-items provided, a greater prediction accuracy can be expected.

3. In the four datasets under investigation, SAI shows a relatively smaller MAE in most density cases. This indicates that a larger amount of data would lead to a higher accuracy of the algorithm.

4. SARA has the smallest RMSE, which implies that SARA has least predicted outliers and the overall predicted values of SARA are closest to the actual values among the four datasets.

Remark 3.2 *Clearly, a large dataset in terms of both density and quantity would improve the prediction accuracy of user-based/item-based CF algorithms because more co-rating test-items give rise to more accurate similarity scores by using the PCC, thereby avoiding some undesirable situations, such as the case where the PCC overrates the similarities of some patients that are actually dissimilar but have similar ratings on a few co-rating test-items. Consequently, as the density decreases, less co-rating test-items will lead to the increase of the MAE. All these observations are illustrated in Table IV.*

Variations of the MAEs with respect to the changes of data density in the different four datasets are shown from Tab. 3.4. Results show that, during the density reduction, the user-based CF is more suitable for the matrices with fewer test-items and item-based CF is more suitable for the matrices with more test-items. This is because low density means fewer co-ratings, thereby affecting the prediction accuracy of (3.6) due to insufficient neighbors. When the density exceeds 50%, these two neighborhood-based CF algorithms have similar performance in terms of MAE.

In order to illustrate the prediction performance of neighborhood-based CF algorithms, two other kinds of figures are presented to show the prediction error and the corresponding distribution. As shown from Fig.3.2 to Fig.3.5, line graphs display 50 randomly selected actual values and their predictions, where the blue "o" points represent actual values and the red "*" points represent the corresponding predictions. The histograms from Fig.3.6 to Fig.3.9 show the statistical distribution of the prediction errors.

Remark 3.3 *The MAE index reflects the prediction error as compared with the actual values in the same interval. As we can see from Fig.3.2 to Fig.3.5, some predicted values (red "*") are very close to the true values (blue "o") but without overlapping. After converting the data back to their original rating intervals/forms, the predicted values would be nearest integer. It is shown in [141] that, if the rating system is integer-based, then rounding the prediction results will reduce the error of MAE.*

Having gone through the computation of the similarity, the selection of the similar neighbors and the prediction of the missing value, we come to the last step of the CF-based data analysis for FRDA database, which is to convert the data to their original intervals/forms, where the detailed results are presented in Table V. For illustration purpose, we present part of the user-based CF with densities 80%, 50% and 20%. From Table V, we can calculate that 93.0% predicted outcomes in SARA dataset, 93.6% predicted outcomes in ADL dataset, 98.2% predicted outcomes in INAS dataset and 97.1% predicted outcomes in SAI dataset are all within a margin of $\pm 1\%$ error at 80% density. When the density decreases to 50%, the prediction accuracy is reduced to 87.1% (SARA), 91.6% (ADL), 96.7% (INAS) and 96.0% (SAI), respectively. Furthermore, in the case of 20% density, the user-based CF can still manage to have prediction accuracies of 76.4% (SARA), 85.4% (ADL), 94.3% (INAS) and 91.8% (SAI) within a margin of error of $\pm 1\%$. Since the main purpose of our strategy is to provide rating assistance and selecting samples of clinical trials, the prediction error at the level of $\pm 1\%$ is considered to be reasonable as can be seen from Tables IV and V. Overall, we believe that the proposed FRRS can be very helpful in assisting medical institutions to gather new patient data, which further shows a great application potential in the area of clinical trials.

Table 3.5 Examples of predicted values within original intervals and forms.

	Total	0	±1	±2	±3	±4	±5
80%							
SARA							
$0 \sim 4$	1693	1049	576	65	3	*	*
$0 \sim 6$	372	152	170	43	7	*	*
$0 \sim 8$	186	68	78	32	8	*	*
Overall	2251	1269	824	140	18	*	*
ADL							
$0 \sim 4$	1681	855	718	98	10	*	*
INAS							
$0 \sim 1$	1945	1606	339	*	*	*	*
$0 \sim 2$	720	535	149	36	*	*	*
$0 \sim 3$	8159	6814	1186	142	17	*	*
Overall	10824	8955	1674	178	17	*	*
SAI							
Overall	14756	11177	3160	387	32	*	*
50%							
SARA							
Overall	5628	3004	1895	577	121	20	11
ADL							
Overall	4203	1964	1886	305	44	4	*
INAS							
Overall	27059	20926	5244	782	105	2	*
SAI							
Overall	36890	26912	8487	1295	180	15	1
20%							
SARA							
Overall	9005	3481	3401	1498	558	34	33
ADL							
Overall	6725	2469	3272	846	129	9	*
INAS							
Overall	43291	29541	11302	2213	227	11	*
SAI							
Overall	59024	36286	17891	3903	866	52	26

3.5 Conclusion

In this chapter, a novel data prediction algorithm for FRDA has been proposed based on the CF methods including both the user-based and item-based ones. By introducing the PCC to evaluate the similarities between different patients and calculating the missing values of the patient-item matrices, the proposed algorithm has been implemented to predict incomplete information of the new patients with an acceptable accuracy. Numerical experiments have been carried out with four different cases (namely, SARA, ADL, INAS, and SAI datasets) in order to demonstrate the effectiveness of the proposed method. This proposed strategy can also be extended to facilitate the collection of data for other diseases. New patient data made available in this way can be used to assist patient selection for clinical trials and data analysts to achieve better management of the underlying disease. Our future research will focus on improving prediction accuracy by considering demographic information and statistical information.

Chapter 4

A Hybrid Model- and Memory-based Collaborative Filtering Algorithm for Baseline Data Prediction of Friedreich's Ataxia Patients

4.1 Motivation

In this chapter, we make one of the first few attempts to view the FRDA baseline data prediction as a recommendation problem where patients correspond to users and test-items on symptoms correspond to items. Intuitively, similar patients should exhibit similar symptoms under reasonable conditions where the severity degree of symptoms can be reflected by different rating values, and therefore the ratings between similar patients should be similar as well [191]. For FRDA patients, it is often the case that they can only provide a moderate amount of auxiliary baseline data of test-items, and there might be unfilled parts of the data that can be regarded as *missing* values. The prediction of missing rating values can be naturally considered as a typical design problem of the recommendation system, which is referred to as the FRDA Rating Recommendation System (FRRS). The FRRS consists of U FRDA patients and I test-items, the relationship between patients and test-items is denoted by a $U \times I$ matrix, which is called as a patient-item matrix. FRRS can predict unfilled part through retrieving the similarities between patients in EFACTS database.

Due to its nature of recommendation system, the proposed FRRS should be capable of achieving good prediction accuracy on FRDA missing value. Nevertheless, two possible drawbacks with the FRRS are identified with the first one being the sparsity problem of the database. In the progress of collecting new patient data, the CF algorithm generates predictions by calculating the similarities between patients, and the corresponding accuracy might not be guaranteed when the self-assessed data is very sparse. The second drawback is that the commonly used similarity measures only consider the ratings on test-items but largely overlook the uncertainty issue arising from individual differences (e.g. different rating habits from different users in recommendation system). These individual differences stem mainly from different physical condition, autognosis, treatment method and environment, onset age, disease duration, and so on. For example, an adolescent patient and an adult patient might have similar disease levels but with different specific symptoms. In this case, the traditional CF algorithm might lead to the so-called overestimation problem of the patient similarities. To this end, it is theoretically necessary and practically significant to improve CF algorithms in FRRS by overcoming the emerging drawbacks, thereby achieving satisfactory performance in a wider environment.

Motivated by the above discussions, in this chapter, we propose a hybrid CF algorithm for baseline data prediction of FRDA patients. The proposed algorithm switches between model-based and memory-based CF techniques according to degrees of the data sparsity and individual differences. More specifically, the model-based CF is used to deal with the situation where a patient does not have similar neighbors because of the sparsity, and the memory-based CF is exploited for a patient who has neighbors but is under uncertainties arising from individual differences. In the former case, the model-based CF is harnessed to find similar neighbors with similar FRDA symptoms by clustering this patient into the class based on his/her attributes. Here, it is quite challenging to choose key attributes for clustering because 1) we need to analyze what kinds of attributes that FRDA patients can provide; 2) based on the pathology of FRDA and basic statistical analysis, key attributes are picked out from the results of the previous step to conduct the clustering; and 3) the most suitable number of clusters is determined according to the clustering results. In the case of a patient with similar neighbors, we adopt an advanced memory-based CF algorithm with an improved similarity measure, where both the patient rating habits and the number of co-rated test-items are taken into account from a unified viewpoint.

The main contributions of this chapter are outlined in threefold as follows:

1. A novel hybrid CF framework is introduced whose idea to switch between model-based CF and memory-based CF according to the actual situation for a comprehensive use of incomplete FRDA baseline data.

2. By analyzing different attributes of the patients, the model-based component of the hybrid CF framework deals with the situation for patients who cannot find neighbors due to data sparsity. In the memory-based component, the Shannon information entropy and Jaccard index are first combined together to describe the rating habits and degree of co-rated test-items between patients, which provide a more comprehensive evaluation for the similarity degree than basic PCC.

3. Comprehensive experiments are carried out to show that our proposed hybrid CF algorithm improves the prediction accuracy of the FRDA baseline data, where the optimal number of clusters for FRDA is determined in order to provide an appropriate categorization to assist disease research.

The remainder of this chapter is organized as follows. In Section 4.2, the literature review of memory-based CF, model-based CF and hybrid CF are discussed. Our proposed hybrid model- and user-based CF algorithm is introduced in Section 4.3. In Section 4.4, the implementation and experiments results are presented and, finally, concluding remarks are given in Section 4.5.

4.2 Literature review

The CF algorithm is used to design recommendation systems and this algorithm was first introduced in 1992 by Goldberg et al [41]. In this section, we review some major approaches of CF that will be used in this chapter.

Memory-based CF approaches

The memory-based CF approaches (also called neighborhood-based CF approaches) are among the most popular prediction techniques in the family of CF methods. In general, the memory-based CF approaches can be classified into user-based and item-based CF approaches according to the performance specifications [53]. The basic idea of the user-based CF approach is to make interest prediction of a target user on an item by analyzing the collective taste information of similar users. First, a user-based CF approach calculates the similarity between a target user and other existing users. It then chooses the *n* most similar users as the nearest neighbors and their similarity values are regarded as weights. Finally, a weighted average is employed to predict the rating of the target user. The only difference between user-based and item-based CF approaches is that item-based CF approach focuses on the similarity between items instead of users. Some commonly used similarity measures include the Cosine, the adjusted Cosine, the Pearson Correlation Coefficient (PCC) and the Spearman's Rank Correlation Coefficient. As described in [20, 125], PCC similarity measure can be easily implemented and can achieve a better overall performance than others.

Model-based CF approaches

The model-based CF approaches utilize different data mining and machine learning algorithms to learn an appropriate model from the collection of ratings, which is then used to predict users' ratings on unrated items. The commonly used techniques are clustering, Bayesian classifiers, probabilistic models, latent factor model, artificial neural networks and so on. Clustering models work by clustering like-minded users into classes. The unrated ratings of a target user can be predicted by averaging the ratings of other users in the same cluster. In Bayesian classifiers, each node in a Bayesian network represents a class of items, and the status of each node corresponds to the possible rating value for each item. In [13, 77, 145], some clustering models indicate that each user could belong to multiple clusters with different levels of participation which are expressed by degree of membership. The prediction is given by averaging the ratings across the clusters that are weighted by the degree of membership.

In recent years, different artificial neural network (ANN) models [51] (including deep neural network models) have been widely applied in recommendation systems. Some rather popular ANN models include, but are not limited to, restricted Boltzmann machine [66], convolutional neural network [149], autoencoder [126] and so on [27]. Other well-known model-based approaches are latent factor model and probabilistic model which involves probabilistic semantic analysis, aspect modeling and probabilistic matrix factorization.

Compared to memory-based CF methods, model-based CF methods can better address the problems of scalability and sparsity. Also, model-based CF methods can improve the prediction performance and prediction speed. At the same time, model-based CF methods have some disadvantages which are 1) model-building is a time- and resource-consuming process, and 2) model-based CF methods have trade-off between scalability and prediction performance.

Hybrid CF approaches

In certain circumstances, memory- and model-based CF techniques have been combined together to yield the so-called hybrid ones that would help the performance improvement [22]. Based on different cases, a hybrid CF approach can include two or more techniques, thereby achieving a better overall performance than any individual one, and this is particularly true when dealing with the data imperfection issues such as sparsity, individual differences and loss of information. In this chapter, a hybrid CF approach is proposed, which combines clustering-based and modified user-based CF methods, in order to achieve satisfactory results on FRDA patient baseline data prediction.

4.3 Methodology

4.3.1 Data description

To implement the method for the addressed data collection problem, three data sets have been chosen from the EFACTS database, which are Scale for the Assessment and Rating of Ataxia (SARA), Demographics and Onset data sets.

SARA data set. SARA, first introduced in 2006 [175], is an effective assessment tool for assessing the severity and treatment effectiveness of ataxia symptoms. SARA has fewer assessment items than other well-known scales like International Cooperative Ataxia Rating Scale (ICARS), thereby possessing the advantage of easier daily assessment of ataxia symptoms. For a decade or so, many researchers have demonstrated the validity and reliability of SARA in handling different kinds of ataxia, and EFACTS has thus used SARA to evaluate the severity of FRDA. It can be seen from Table 4.1 that SARA contains 16 features in 8 categories reflecting neurologic manifestations of ataxia which are gait, stance, sitting, speech disturbance, finger chase, nose-finger test, fast alternating hand movements and heel-shin slide. A scale of 0 to n ($n \in \{4,6,8\}$) is created for each test-item to describe the order of severity of FRDA, where 0 means the normal condition and n implies the most serious situation. The total SARA scores reflect overall severity degree which is calculated by adding scores of eight categories.

Demographics and Onset data sets. The Demographics data set includes demographic information of the FRDA patient such as year of birth, country of birth, age and sex. Onset data set contains onset information of FRDA patient, which includes age of first FRDA symptoms, symptoms at onset and problems during neonatal period. After preliminary data analysis, two pieces of crucial yet essential information, namely, onset age and disease duration, are extracted from the demographics and onset data sets.

4.3.2 Hybrid collaborative filtering framework

In this chapter, a hybrid CF framework is proposed in Algorithm 1, which is fairly general to include the model-based CF and memory-based CF components and is particularly suitable to solve the baseline data prediction problem for FRDA patients. Based on the circumstances, the model-based CF and memory-based CF can switch back and forth between them over the course of the execution of the algorithm.

Model-based CF component

In the model-based CF component, the clustering method is first used to divide existing patients into different groups by using their side-information and SARA scores. When a new patient who cannot find neighbors by using memory-based CF appears, K-NN is used to identify which cluster this new patient belongs to. In this way, the missing values of this new patient can be predicted by using the known values of patients in this cluster.

Clustering is a method to divide a set of data into a specific number of groups through a form of association. There are many algorithms that can be used to do clustering. In this chapter, we use K-means algorithm as the basic clustering algorithm with the aim of evaluating the intrinsic nature and regularity of data by using unlabeled training samples [176].

The main steps of K-means algorithm and how it is implemented in SARA dataset are illustrated in Algorithm 1.

Algorithm 1 K-means algorithm

```
1. Load the data and set the value of $K$ as the desired number of clusters
2. Choose $K$ patients randomly as the initial cluster centroids
3. Allocate the remaining patients to the closest cluster centroids
4. Update the cluster centroids and re-evaluated the cluster-membership
5. Repeat step 3 and 4 until the cluster centroids do not update
```

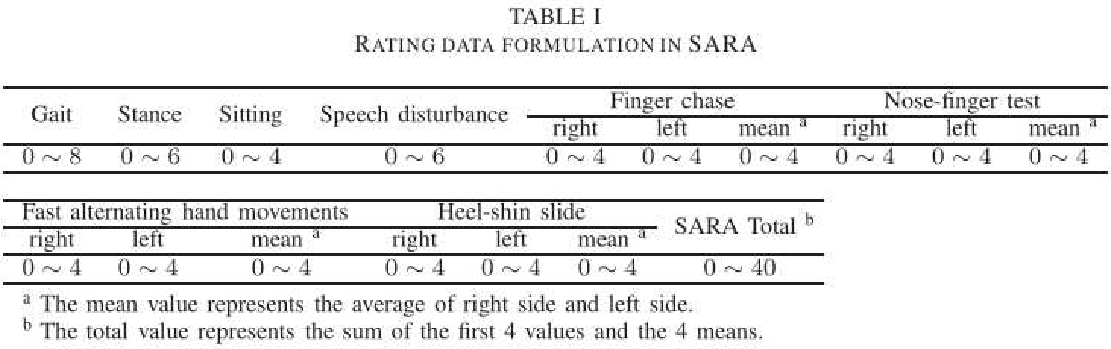


TABLE I
RATING DATA FORMULATION IN SARA

Gait	Stance	Sitting	Speech disturbance	Finger chase			Nose-finger test		
				right	left	mean [a]	right	left	mean [a]
$0\sim8$	$0\sim6$	$0\sim4$	$0\sim6$	$0\sim4$	$0\sim4$	$0\sim4$	$0\sim4$	$0\sim4$	$0\sim4$

Fast alternating hand movements			Heel-shin slide			SARA Total [b]
right	left	mean [a]	right	left	mean [a]	
$0\sim4$	$0\sim4$	$0\sim4$	$0\sim4$	$0\sim4$	$0\sim4$	$0\sim40$

[a] The mean value represents the average of right side and left side.
[b] The total value represents the sum of the first 4 values and the 4 means.

Fig. 4.1 Rating data formulation in SARA

Following the operation on existing patients based on the above clustering algorithm, some traditional machine learning algorithms can be further applied to solve the sparsity problem for the new patients. Here, K-NN is used for obtaining precise classification when the new patients provide sparse data [26].

The K-NN algorithm is illustrated in Algorithm 2.

Algorithm 2 K-NN algorithm

```
1. Load the data and select the number of $\check{k}$
2. Calculate the distance between new patient $a$ and all existing patients $u$
3. Sort the calculated distances in ascending order
4. Get top $\check{k}$ patients and choose the most frequent class
5. Return the predicted class as new patient $a$'s class
```

After determining the class of new patient (specified as a), we retrieve the similar neighbors who have same ratings on overlapped test-items within the same class. The missing values of test-items can be predicted by the following equation:

$$P(r_{a,i}) = \frac{\sum_{u\in\hat{K}_a} r_{u,i}}{|\hat{K}_a|}, \tag{4.1}$$

where $\hat{K}_a$ is a set of existing patients who have the same ratings with new patient a on overlapped test-items in the same class, and $|\hat{K}_a|$ denotes the number of matched patients.

Table 4.1 Rating data formulation in SARA

Gait	Stance	Sitting	Speech disturbance	Finger chase			Nose-finger test			Fast alternating hand movements			Heel-shin slide			SARA $Total^b$
				right	left	$mean^a$	right	left	$mean^a$	right	left	$mean^a$	right	left	$mean^a$	
$0 \sim 8$	$0 \sim 6$	$0 \sim 4$	$0 \sim 6$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 40$

a: The mean value represents the average of right side and left side.

b: The total value represents the sum of the first 4 values and the 4 means.

Memory-based CF component

There are two kinds of methods for memory-based CF, which are user-based CF and item-based CF method. In this subsection, we present the user-based CF method with an *enhanced* similarity measure.

Let us start with the PCC, which is a popular similarity computation method in CF and has been widely used in a number of recommendation systems owing to its capability of achieving a high accuracy [197]. The similarity degree between patients u and a is calculated by

$$\mathrm{Sim}(u,a) = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_u)(r_{a,i} - \bar{r}_a)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in I}(r_{a,i} - \bar{r}_a)^2}}, \tag{4.2}$$

where $\mathrm{Sim}(u,a)$ is the similarity degree between FRDA patients u and a; $I = I_u \cap I_a$ is the subset of test-items that both patients u and a have rated, with I_u (respectively, I_a) denoting all test-items that patient u (respectively, a) has evaluated; $r_{u,i}$ (respectively, $r_{a,i}$) is the rating value of test-item i provided by patient u (respectively, a); $\bar{r}_u$ and $\bar{r}_a$ denote average ratings of different test-items that patients u and a have rated, respectively. It can be easily seen from (4.2) that the similarity of two patients takes value in the interval of $[-1,1]$. Clearly, a larger similarity indicates that patients u and a are more similar.

The PCC index, though widely used, might suffer from the issue of overestimating the similarities of patients who are actually dissimilar but happen to have similar symptoms on a few co-rated test-items. In order to avoid such an issue, one can make use of the so-called *Jaccard index* which is a sample statistic measuring the similarity and diversity of sample sets as defined as follows:

$$J(u,a) = \frac{|I_u \cap I_a|}{|I_u \cup I_a|} = \frac{|I_u \cap I_a|}{|I_u| + |I_a| - |I_u \cap I_a|}, \tag{4.3}$$

where $|I_u \cap I_a|$ represents the number of co-rated test-items of patients u and a; $|I_u \cup I_a|$ denotes the number of total test-items that patients u and a have rated; and $|I_u|$ and $|I_a|$ are the numbers of test-items rated by patient u and patient a, respectively.

Taking advantage of the diversity reflected in the Jaccard index, one can define the following *modified PCC* (with hope to get rid of the overestimating issue):

$$\mathrm{Sim}^J(u,a) = J(u,a) \times \mathrm{Sim}(u,a), \tag{4.4}$$

where $\mathrm{Sim}^J(u,a)$ is a modified similarity measure.

Remark 4.1 *When the number of co-rated test-items (i.e. $|I_u \cap I_a|$) is small, the introduction of Jaccard index $J(u,a)$ in (4.4) helps reduce the similarity between patients u and a, thereby mitigating the overestimating issue. Since the Jaccard index $J(u,a)$ takes value in interval of* $[0,1]$ *and the PCC similarity varies in the interval of* $[-1,1]$*, the new index* $\mathrm{Sim}^J(u,a)$ *is still in the interval of* $[-1,1]$.

Apparently, the similarity measure defined in (4.4) serves as a modified version of the PCC index by taking into account the patients' differences via the consideration of the co-rated test-items. This modified similarity measure is, however, not sufficiently comprehensive yet and there is still a room for further improvement. Specifically, we need to further examine the individual rating distribution of the patients. In fact, the occurrence, the development and the cure of a disease are influenced by various factors (e.g. climate, geographical environment, constitution, sex and age) that give rise to the individual differences on the ratings through filled forms. More specifically, there might be the case that two patients have similar order of disease severity (i.e., similar overall ratings) but their score on the same test-items might be significantly different, and such kind of differences needs to be reflected in the similarity measurement. To this end, we introduce Shannon's entropy concept to describe the individual differences of the patients' ratings through considering the degree of uncertainty/disorder of the scores.

Shannon entropy, which has been applied on CF algorithms (see e.g. [40, 69]), is defined as:

$$H_u = -\sum_{r \in RD} \mathscr{P}_{u,r} \log_2 \mathscr{P}_{u,r}, \tag{4.5}$$

where H_u denotes the entropy of patient u, $\mathscr{P}_{u,r}$ represents the frequency of value r which has been rated by patient u on test-items, and RD denotes the rating domain which contains a finite number of discrete values. The PCC with entropy weighting has been defined in [78] as follows:

$$\mathrm{Sim}^{E}(u,a) = \frac{1}{1+|H_u - H_a|}\mathrm{Sim}(u,a). \tag{4.6}$$

Clearly, when the values of H_u and H_a differ greatly, the similarity degree between patients u and a is reduced accordingly. Also, it is easy to see that the value of $\mathrm{Sim}^{E}(u,a)$ remains in the interval of $[-1,1]$.

Having gone through the discussions on the PCC, the modified PCC and the PCC with entropy weighting, we are now ready to present our proposed *enhanced similarity measure* as follows:

$$\mathrm{Sim}^{EJ}(u,a) = \frac{1}{1+|H_u - H_a|} \times J(u,a) \times \mathrm{Sim}(u,a) \tag{4.7}$$

where $\mathrm{Sim}^{EJ}(u,a)$ is an enhanced similarity index, and the value of $\mathrm{Sim}^{EJ}(u,a)$ is clearly within the interval of $[-1,1]$.

Remark 4.2 *Our proposed enhanced similarity measure (4.7) has the remarkable advantages of 1) retaining the merits of the PCC such as clear practical insights and neat mathematical property (i.e., invariance under location and scale changes in the two variables); 2) accounting for the impact from the diversity of the patients; and 3) reflecting the individual differences in rating scores. As such, the enhanced PCC (4.7) provides a unified basis to quantify the similarity between the patients, which is more comprehensive than existing ones. In fact, in addition to the establishment of the hybrid CF algorithm for FRDA baseline data collection, this enhanced similarity measure (4.7) constitutes the second contribution of this chapter.*

The basic yet natural assumption for the CF algorithms is that similar patients should have similar ratings on test-times and, therefore, appropriate selection of similar neighbors is vitally important in improving the prediction accuracy. For this purpose, we employ a top-n algorithm by which we first arrange the similarities between patients in the descending order and then select the top n patients as the similar neighbors. In order to avoid using dissimilar neighbors, some conditions [197] are added to the top-n algorithm as follows:

$$\hat{S}(a) = \{a_u | a_u \in T(a), \mathrm{Sim}(a_u,a) > 0, a_u \neq a\}, \tag{4.8}$$

where $\hat{S}(a)$ denotes a set of similar patients of patient a that is chosen to use in the following rating prediction, and $T(a)$ is a set of top n similar patients of patient a.

After the top n similar neighbors of the patient are selected, the missing values of test-items can be predicted by the following equation[20]:

$$P(r_{a,i}) = \bar{a} + \frac{\sum_{a_u \in \hat{S}(a)} \mathrm{Sim}^{EJ}(a_u, a)(r_{a_u,i} - \bar{a}_u)}{\sum_{a_u \in \hat{S}(a)} \mathrm{Sim}^{EJ}(a_u, a)}, \tag{4.9}$$

where $P(r_{a,i})$ denotes the predicted value of the missing value $r_{a,i}$ in the patient-item matrix, $\bar{a}$ is the average value of different test-items provided by patient a, and $\bar{a}_u$ is the average value of test-items provided by the similar patient a_u.

Remark 4.3 *In this chapter, a new FRDA baseline data collection scheme is put forward based on a combination of the merits of model- and memory-based CF methods with a much enhanced similarity measure. The new data collection scheme exhibits the following three distinctive characteristics: 1) it switches between model-based CF and memory-based CF according to when a certain patient has neighbors sharing similar baseline data; 2) a new yet comprehensive similarity index is proposed to take into account the individual differences between patients by employing the Shannon information entropy and the Jaccard index; and 3) extensive experiments are conducted in the next section to show the superiority of the proposed scheme with the determination of optimal number of clusters for FRDA. The proposed FRDA baseline data collection scheme is believed to be effective in assisting disease research.*

Algorithm 1: Hybrid CF framework

```
• Given a new patient $a$ with $I$ rating test-items, onset age and disease duration;
• Analyze $I$ rating test-items;
• If the new patient only provides single rating or multiple ratings with same values (system switches to model-based CF);
  The model-based CF component
    1. Create $K$ patient clusters by using the attributes: onset ages, disease durations and total SARA scores;
       (K-means algorithm is applied;) see 4.3.2)
    2. Find $n$ neighbors in the database with same rating test-items, then averaging total scores of $n$ neighbors as the initial SARA score $S_a$ of new patient $a$;
    3. Classify the new patient $a$ into cluster $K_a$ by using the attributes of onset age, disease duration and initial SARA score $S_a$;
       (K-NN algorithm is applied;) see 4.3.2)
    4. Retrieve $n'$ similar neighbors who have same rating test-items in cluster $K_a$;
    5. Predict the rating on the target test-item $i$ for $a$ by averaging the corresponding values rated by $n'$ similar neighbors on the test-item $i$;
• else (system switches to memory-based CF)
  The memory-based CF component
    1. Calculate similarity $\mathrm{Sim}^{EJ}(u,a)$ between each existing $u$ and new patient $a$ by considering their PCC, Jaccard index and Shannon entropy;
       (Technique details will be introduced in Section 4.3.2)
    2. Select top-$\tilde{n}$ similar users as the nearest neighbors of new patient $a$;
    3. Predict the rating of the target test-item $i$ for $a$ by the behaviors of the $\tilde{n}$ nearest neighbors.
```

Table 4.2 Baseline demographic characteristics

	Age (years)	Male	Female	Age of onset (years)	Disease duration (years)	Education (years)	SARA
Aachen, Germany (n=56[6%]	29(6-62)	29(52%)	27(48%)	13(0-25)	14(2-54)	14(0-49)	19(2-40)
Athens, Greece (n=20[2%])	25(8-42)	12(60%)	8(40%)	12(3-21)	10(4-22)	15(3-31)	23(7.5-40)
Bonn, Germany (n=23[3%])	39(20-59)	11(48%)	12(52%)	13(0-19)	20(9-50)	19(5-44)	20(3.5-31.5)
Brussels, Belgium (n=52[6%])	25(7-69)	26(50%)	26(50%)	12(9-21)	14(3-60)	11(1-38)	18(3-34)
Dublin, Ireland (n= 8[1%])	25(7-69)	6(75%)	2(25%)	15(10-19)	19(3-42)	11(4-19)	16(8.5-26)
Innsbruck, Austria (n=57[6%])	31(8-62)	31(54%)	26(46%)	11(2-18)	17(1.5-47)	13(2-35)	20(6-38)
Kassel, Germany (n= 6[1%])	44(23-73)	3(50%)	3(50%)	13(9-15)	19(10-40)	25(11-37)	23(8.5-40)
london, UK (n=205[23%])	33(15-77)	94(46%)[a]	110(54%)[a]	15(0-30)	14(1-55)	20(11-37)	22(1.5-40)
Madrid, Spain (n=78[9%])	32(6-65)	34(44%)	44(56%)	14(0-24)	14(2-44)	17(1-44)	21(5-37)
Milano, Italy (n=195[22%]	34(7-70)	94(48%)	101(52%)	12(0-22)	16(3-61)	18(1-46)	22(3-39)
Munich, Germany (n=66[8%])	33(12-60)	35(53%)	31(47%)	12(0-22)	16(2-56)	17(2-45)	19(2-40)
Paris, France (n=60[7%])	37(19-76)	28(47%)	32(53%)	13(0-23)	20(3-65)	17(0-36)	23(5-39)
Rome, Italy (n=17[2%])	24(9-61)	7(41%)	10(59%)	9(3-21)	14(1-40)	10(2-22)	15(7-36)
Tübingen, Germany (n=35[4%])	35(14-74)	16(46%)	19(54%)	11(5-19)	18(0-46)	17(5-39)	22(7.5-39)

a: Data for sex was missing for one patient in London.

4.4 Implementation and experiments

4.4.1 Data preprocessing

The data set SARA is constantly updated. Until 31*st* December 2018, the SARA data set has contained the information of 989 patients. It should be mentioned that the SARA data includes missing values and redundant information. Hence, 111 patients are deleted because their data is null, missing or abnormal. A total of 878 patients have been selected for the follow-up experiments. The details of these patients are displayed in Table 4.2. “Aachen, Germany(n=56[6%])” means the baseline data of 56 patients were collected in Aachen (Germany) and accounted for 6% of the total patients. “29(6-62)” means average age is 29 and spread between 6 to 62 years old.

4.4.2 Experimental setup

We divide the 878 patients into two parts, with the first part consisting of existing patients and the second part containing the new patients. As mentioned in Section 4.1, there are two situations that we need to consider. The first situation is that the new patients cannot find neighbors from existing patients and the other situation is that new patients can find neighbors. In the first situation, we randomly keep one rating value of new patients and set other values as testing data. In the second situation, we randomly remove different number of elements to make the patient-item matrix sparser with different density (e.g., 50%, 60%, etc.), where the density refers to the ratio of number of entries presented to the total number of the entries in the patient-item matrix. The developed hybrid model- and memory-based CF algorithm is employed for predicting the rating values of new patients' unfilled parts.

For the propose of evaluating the prediction accuracy of the algorithm, the criteria of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are taken into account, which are defined as follows:

$$MAE = \frac{1}{N}\sum_{a\in A_d}\sum_{i\in I_d}|r_{a,i}-P(r_{a,i})|, \tag{4.10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{a\in A_d}\sum_{i\in I_d}(r_{a,i}-P(r_{a,i}))^2} \tag{4.11}$$

where N denotes the total number of predicted values, A_d is the user set of the testing data and I_d is the test-item set of the testing data, $r_{a,i}$ is the actual value of test-item i provided by patient a, and $P(r_{a,i})$ denotes the predicted value from the developed CF method.

4.4.3 Performance comparison

This section is divided into two parts which describe two situations, one is that the new patient does not have neighbors and the other one is that the new patient does have neighbors. In these two situations, we compare our approach with other well-known approaches. In the experiment, we set the value $k = 7$ during the K-means clustering. Fig. 4.2 shows the experimental results. The clustering performance is evaluated by the silhouette coefficient which is shown in Fig. 4.3.

The patient without neighbors

In this part, the single regression (SREGR) imputation and expectation-maximization (EM) algorithms as well as four different mean imputation methods for missing value prediction are employed to compare with our method. These mean imputation methods include the rating-mean (RMEAN) imputation, user-mean (UMEAN) imputation, centered user-mean (CUMEAN) imputation and adjusted user-mean (AUMEAN) imputation methods, where the RMEAN approach employs the average of filled ratings of the current new patient, the UMEAN approach utilizes the average SARA ratings of the existing patients in database to

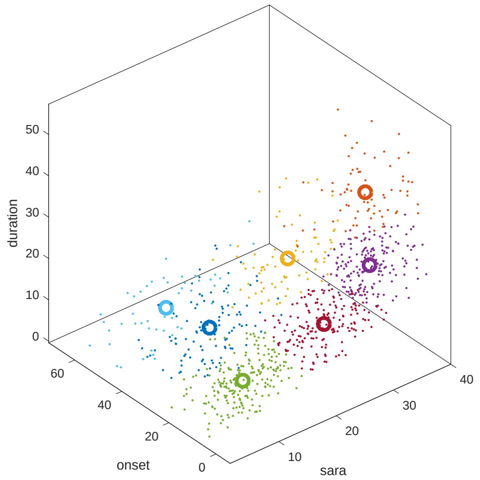


Fig. 4.2 *K*-means clustering diagram

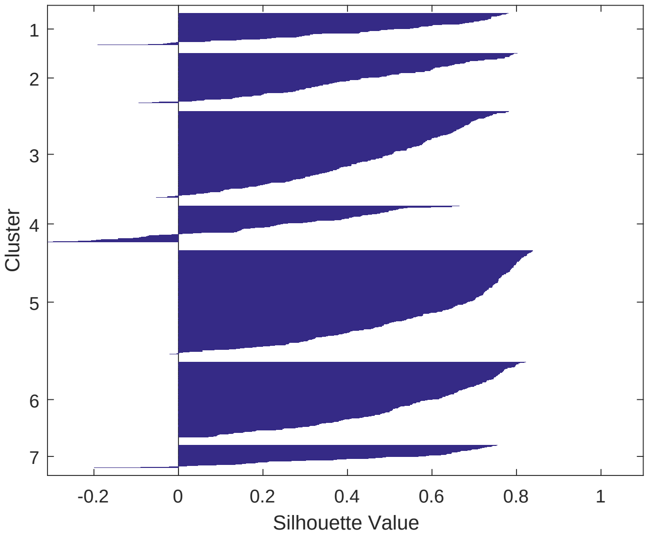


Fig. 4.3 Silhouette coefficient of *K*-means algorithm

predict the new patient unfilled ratings, the CUMEAN approach considers the rating bias by subtracting the mean value of each existing patient, and the AUMEAN approach uses the average SARA ratings of the existing patients who have the same ratings on overlapped test-items. The mathematical expressions for these four approaches are displayed as follows:

$$\text{RMEAN}: \; P(r_{a,i}) = \bar{r}_a \tag{4.12}$$

where $\bar{r}_a$ is the average rating value of different test-items rated by the new patient a.

$$\text{UMEAN}: \; P(r_{a,i}) = \frac{\sum_{u=1}^{n} r_{u,i}}{n} \tag{4.13}$$

where $r_{u,i}$ represents the rating value of test-item i rated by existing patient u in the database.

$$\text{CUMEAN}: \; P(r_{a,i}) = \bar{r}_i + \frac{\sum_{u=1}^{n}(r_{u,i} - \bar{r}_u)}{n} \tag{4.14}$$

where $\bar{r}_u$ represents average rating value of different test-items rated by the existing patient u in the database.

$$\text{AUMEAN}: \; P(r_{a,i}) = \frac{\sum_{u \in \hat{S}} r_{u,i}}{|\hat{S}|} \tag{4.15}$$

where $\hat{S}$ is a set of existing patients who have the same ratings with new patient a on overlapped test-items, $|\hat{S}|$ denotes the number of matched patients.

The detailed information of performance comparison of different approaches is displayed in Table 4.3. To demonstrate the validity and the superiority of the proposed algorithm, we randomly choose 10% of total patients and set them as new patients during each experiment. In order to facilitate the situation of no neighbors, we only keep one rating value and remove all remaining rating values by setting them as the unfilled part. Each experiment is repeated 50 times, and the average MAE and RMSE values are reported in Table 4.3. From the experiment results, we conclude that:

1. Under all experimental settings, our approach obtains the smallest MAE and RMSE values consistently, which indicates the best prediction accuracy.

2. Relative to AUMEAN (the best of the four different mean imputation methods) which considering the all patients with same ratings on overlapped test-items, our approach only consider the patients within the same class. Experimental results demonstrate

the MAE of our approach is 15.6% better and the RMSE is 7.5% better than those produced by AUMEAN.

Table 4.3 MAE and RMSE Comparison with four Basic Approaches

Metric	Methods	New patients(10%)
MAE	RMEAN	0.1999
	UMEAN	0.2880
	CUMEAN	0.2038
	AUMEAN	0.1644
	SREGR	0.2025
	EM	0.1498
	Our approach	**0.1388**
RMSE	RMEAN	0.2964
	UMEAN	0.3442
	CUMEAN	0.2421
	AUMEAN	0.2163
	SREGR	0.2765
	EM	0.2097
	Our approach	**0.2001**

The patient with neighbors

To evaluate the prediction performance on a new patient who has neighbors, we compare our approach with four other approaches: MEAN imputation, SREGR imputation, user-based CF using PCC (UPCC), user-based CF using PCC with entropy (UPCCE) and user-based CF using PCC with Jaccard index (UPCCJ). UPCC only considers the performance of similar patients to make the prediction according to (4.2). UPCCE considers the disorder degree of the data and UPCCJ considers the overlapped part of the data. To study the impact of our approach that combines the information entropy and Jaccard index, we implement our approach on SARA dataset by employing the density decrementing from 90% to 50% with the interval of 10%.

The results of the performance comparison with our proposed algorithm are shown in Tab. 4.4, where the vertical coordinate represents the value of MAE/RMSE, and the horizontal coordinate denotes the different degrees of density of the test data. Additionally, the detailed experimental results are displayed in Table 4.4 from which we conclude that:

1. The proposed algorithm demonstrates its superiority over MEAN, SREGR, UPCC, UPCCE and UPCCJ in terms of evaluation indices including the MAE and RMSE.

Especially, compared to the most basic approach UPCC, our approach achieves a vast improvement.

2. Experimental results show that it is better to consider both information entropy and Jaccard index at the same time than anyone individually.

3. According to the changes of MAE and RMSE with the density varying from 90% – 50%, we can see that as the density of test data decreases, the superiority of our algorithm can be reflected more significantly.

Table 4.4 MAE and RMSE comparison with basic approaches from density 50% to 90%.

	MEAN	SREGR	UPCC	UPCCE	UPCCJ	Our approach
50%						
MAE	0.2639	0.1585	0.1650	0.1458	0.1444	**0.1397**
RMSE	0.3116	0.2307	0.2201	0.2007	0.1957	**0.1936**
60%						
MAE	0.2632	0.1515	0.1579	0.1383	0.1332	**0.1290**
RMSE	0.3094	0.2240	0.2100	0.1912	0.1795	**0.1792**
70%						
MAE	0.2661	0.1341	0.1417	0.1276	0.1215	**0.1189**
RMSE	0.3132	0.1879	0.1887	0.1768	0.1656	**0.1649**
80%						
MAE	0.2628	0.1259	0.1260	0.1153	0.1130	**0.1111**
RMSE	0.3110	0.1806	0.1670	0.1598	0.1533	**0.1531**
90%						
MAE	0.2698	0.1202	0.1039	0.1028	0.1031	**0.1016**
RMSE	0.3178	0.1668	0.1438	0.1456	0.1435	**0.1426**

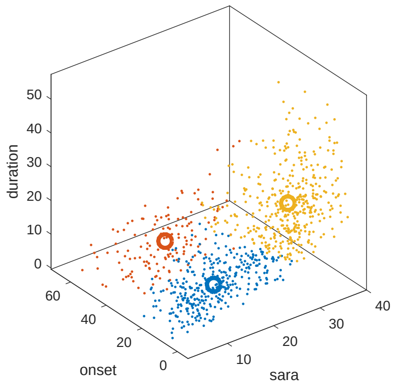


Fig. 4.4 *k*-means clustering(*k*:3).

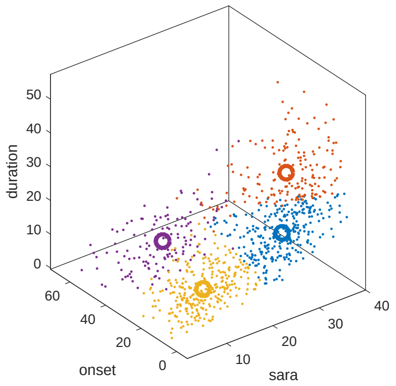


Fig. 4.5 *k*-means clustering(*k*:4).

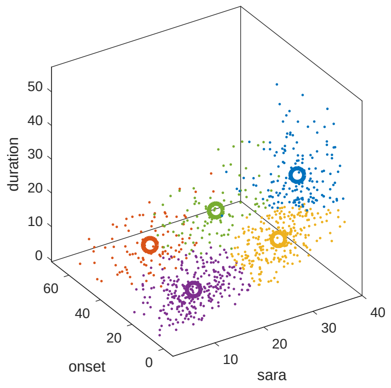


Fig. 4.6 *k*-means clustering(*k*:5).

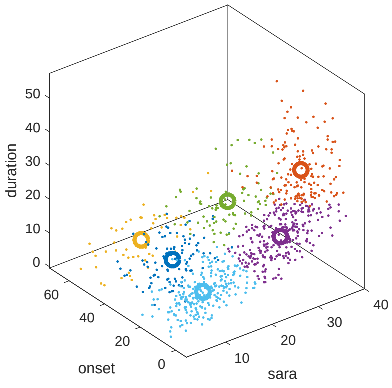


Fig. 4.7 *k*-means clustering(*k*:6).

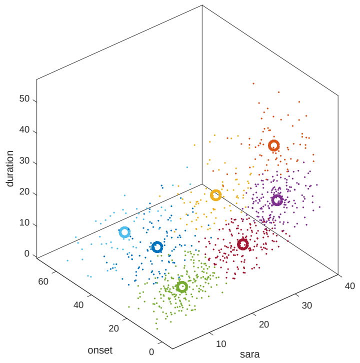


Fig. 4.8 *k*-means clustering(*k*:7).

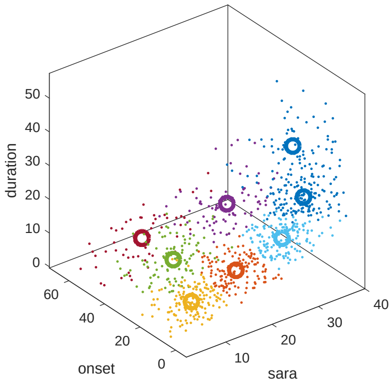


Fig. 4.9 *k*-means clustering(*k*:8).

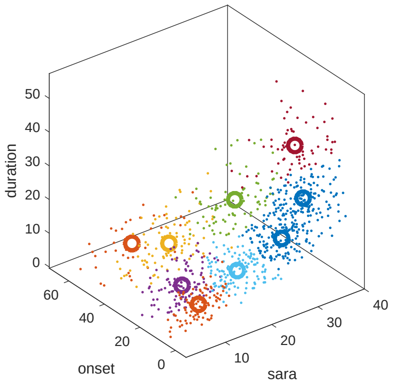


Fig. 4.10 *k*-means clustering(*k*:9).

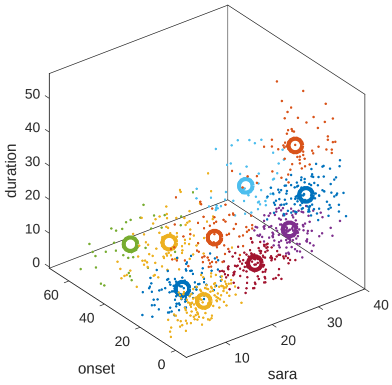


Fig. 4.11 *k*-means clustering(*k*:10).

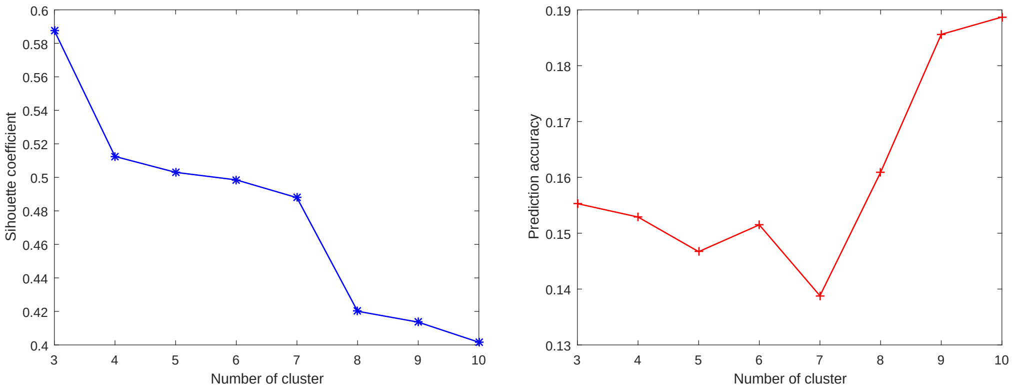


Fig. 4.12 clustering performance and prediction accuracy of *k*-means clustering with different number of *k*

4.4.4 Impact from the number of the clusters

To examine the impact of the number of the clusters, we study three aspects of performance with k taken from 2 to 10, which are 1) prediction accuracy, 2) interpretability of results at the disease level, and 3) clustering performance. Then, we select the most suitable k based on the comprehensive evaluation of these three aspects, where the prediction accuracy is evaluated by (4.10), the interpretability of results is based on FRDA pathology and expert advice, and the clustering performance is evaluated by comparing the silhouette coefficients according to

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4.16}$$

where $b(i)$ denotes the smallest average distance of vector i to all points in other clusters, $a(i)$ represents the average distance of vector i to all points in cluster with i being a member. The silhouette coefficient is in the interval of $[-1, 1]$. Silhouette coefficient with the lager value means the better clustering performance.

Figs. 4.4 to 4.11 show the clustering diagrams with different number of k. As shown in Fig. 4.12, three clusters can provide the largest silhouette coefficient, but we can see from Figs. 4.4 to 4.11 that they are rather difficult to explain the practical insights of the clustering results whereas the prediction accuracy is not very good as well.

When the number of clusters takes the value from 4 to 7 , the silhouette coefficients are similar. However, the 7 clusters can provide better prediction accuracy while, at the same time, the clustering results can be well explained as follows: 1) "mild patients with early&intermediate onset and short duration"; 2) "moderate patients with early&intermediate

onset and middle duration"; 3) "severe patients with early&intermediate onset and long duration"; 4) "mild patients with late onset and short duration"; 5) "Moderate patients with late onset and middle duration"; 6) "severe patients with late onset and long duration"; and 7) "mild patients with very late onset and short duration". Based on the comprehensive comparison, the number of $k = 7$ is regarded as the best one for the clustering.

4.4.5 Discussion

Many existing clinical studies suffer from small sample sizes that cause the results to be insignificant. In our research, the output of the algorithm is to assist in the collection of baseline data for patients who cannot attend the assessments, thereby helping with the clinical sample collection and data analysis from the researchers' perspective. Once more and more patient baseline data are collected, our follow-up plan will be carried out in the following two ways.

- The first way is to increase the interpretability of the imputed data. The most common view of the interpretability in recommendation system is to increase the algorithm transparency, and this is particularly true in our research where reliable explanations can largely increase the confidence of the end users (patients and/or doctors) in the imputed data. Also, with a satisfactory interpretability of the imputed data, the end users could evaluate the predicted ratings and make appropriate adjustments in real-time based on the explanations, thereby providing us with more reliable data. Such a cycle would help improve the performance of the developed recommendation system with hope to have more accurate predictive information.

- The second way is to combine adequate machine learning algorithms with our proposed method to classify patients accurately. As discussed in the introduction, we are committed to helping EFACTS in collecting more patient data and assisting clinical sample collection. In our chapter, we have divided patients into 7 categories by clustering. In practical application, we could consider different patient-side information and adjust our method according to the complicated actual situations. In this case, the latest deep learning algorithms can be employed to classify the patients in a more accurate way with hope to help doctors/researchers in the selection of clinical samples.

On the other hand, it is predictable that the future analysis will use longitudinal data rather than only the baseline data. We are pleased that some FRDA patients are taking follow-up assessments every year for many years (leading to longitudinal data) but, unfortunately, we are also aware that the number of return visits is decreasing every year, which is inevitable because FRDA symptoms are degenerating and the FRDA patients are usually progressively in poor physical conditions.

- In the context of FRDA patients, the existing longitudinal data do have certain limitations and need to be further improved because 1) the number of patients in the EFACTS database is very limited; and 2) these patients have different disease durations and onset ages with different numbers of follow-up assessments. In order to make more sense of the longitudinal analysis, we need to expand the number of patients to find enough suitable samples in order to observe/study their disease progression, drug reaction and so on. In this sense, our presented method can not only effectively increase the number of potential clinical samples but also help the missing value prediction in longitudinal data, which provides the expected assistance for future longitudinal data analysis.
- As potential disease-modifying therapies in FRDA are emerging, there is indeed an urgent need to conduct longitudinal studies to identify and validate robust measures of clinical progression so as to guide the design of future clinical trials. Our future work will include the adoption of the advanced dynamic models for the time series analysis of the disease progression, for which our purpose is to determine the long-term trends and also consider the seasonal changes, cyclic fluctuations and irregular changes in the time series, with the ultimate goal of making reliable statistical predictions. The above analysis requires high quality of longitudinal data and we believe that our presented method will definitely help EFACTS in improving the quality of longitudinal data.

4.5 Conclusion

In this chapter, a hybrid model- and memory-based algorithm has been presented and successfully applied to improve the prediction performance on FRDA baseline data. By taking model-based CF into account, the drawback of the traditional similarity calculation methods in finding neighbors in the sparse data condition has been overcome. Moreover, an enhanced and more generalized similarity measure has been proposed in memory-based CF so as to provide a more comprehensive evaluation for the similarity degree between two patients by considering the rating habits and degree of co-rated test-items. Large-scale real-world FRDA experiments have been conducted and the comprehensive experimental results have shown the validity and feasibility of our algorithm. Finally, we have obtained the best number of clusters which provides an important reference for disease research.

Future work can be summarized into three aspects: (1) how to further improve the prediction performance of the FRDA baseline data by considering matrix factorization, deep learning techniques and dynamics analysis; (2) how to extend our algorithm to other disease baseline data collection problems and the wider health systems; and (3) how to provide explanations for the recommended results. The explainable recommendation is our key research direction because the effectiveness and persuasiveness of the recommended results can be greatly improved if the system uses the easy-to-understand explanation to let the patients know why the results are recommended to them. Interpretation of prediction results can also assist doctors and patients to make the accurate decision about whether to accept predicted results or to make reasonable adjustments.

Chapter 5

A Novel Collaborative Filtering Approach for Friedreich's Ataxia Baseline Data Collection under Cold-Start Condition

5.1 Motivation

Due to the physical condition of patients and other unexpected issues, it is impossible to guarantee that all patients are physically capable of attending the assessments in the study sites. Nevertheless, it is very likely that patients' families can provide accurate baseline data on some test-items at home after long-term daily care, and the unprovided parts can then be treated as *missing* rating values. Based on the fact that similar patients show similar symptoms, just like similar users exhibit similar interests in items in RSs, the prediction of the missing rating values can be considered as a typical RS problem. In this context, it seems natural to apply the CF algorithms to deal with the FRDA baseline data collection problem, where patients amount to users and test-items for symptoms amount to items. As such, the severity degree of different symptoms can be quantified based on the ratings on the test-items.

The FRDA Rating Recommendation System (FRRS) was established in [179] to assist FRDA patient baseline data collection. The FRRS analyzes the known ratings on test-items of new patients and those of other existing patients in the EFACTS database. Then, the missing ratings of these new patients are predicted by FRRS based on the ratings of their similar patients (neighbors). FRRS can achieve a good prediction accuracy on the assumption that the patients or their families can provide some partially accurate rating data on test-items [179, 180]. However, in reality, there are some cases where the newly recorded patients *do not* have the ability to provide any baseline rating data. Such a "no-data" situation can be seen as a cold-start problem as discussed in [199], and it is really difficult for the FRRS to make valid predictions under this situation. Consequently, it is of practical importance to design an advanced algorithm to provide relatively satisfactory prediction ratings under the cold-start condition.

In this chapter, a weighted-naive-Bayes based CF (WNBCF) recommendation algorithm is proposed to solve the cold-start problem. The naive Bayes (NB) method is a popular classification method based on the Bayes' theorem and the independence assumption between the features. According to the characteristics of the NB method, in practical applications, the patient side-information (attributes) is adopted to discover the relationship with ratings (classes) on test-items. The patient side-information is the basic and useful information of a patient which include, but are not limited to, patients' age, gender, onset age and so on. The NB algorithm assumes that all the attributes have the same importance for classification. In fact, different attributes may have different influence on classification performance. As such, a weighted NB algorithm has been developed whose main idea is to assign different attributes with different weights according to their significance in improving the performance of the classifier [184, 37]. In this situation, a challenging problem is to allocate *proper* weights for each attribute to achieve a superior classification performance. In this context, the weight selection can be treated as an optimization problem, and it becomes a rather challenging task as how to effectively solve such a constrained optimization problem from the perspective of FRDA patient baseline data collection.

To formulate an optimization problem, the utilization of reasonable constraints would definitely help improving both the reliability and the accuracy of the optimization results. In the optimization problem addressed in this chapter for the FRDA patient baseline data collection, the first constraint is an inequality constraint that reflects the relationship between the weights. The actual situation is that different patient side-information has different effects on the ratings of test-items. To describe these effects, mutual information (MI) is introduced in this chapter to investigate the mutual dependence between the attributes and classes. The MI is chosen as an index to reflect the importance of each attribute in the classification. The more important the attributes are, the larger their weights would be. Another equality constraint is that the sum of weights is required to be 1. As such, the problem becomes a constrained optimization problem. To facilitate the subsequent development of the optimization algorithm, the penalty-function method is utilized to transform the constrained optimization problem into a series of unconstrained ones.

Evolutionary computation algorithms have shown outstanding performance in solving optimization problems in a wide range of real-world applications such as healthcare, telecommunication, power systems, and so on [183, 185]. As a powerful member of the family of optimization techniques, the particle swarm optimization (PSO) algorithm has been successfully employed to solve the optimization problems owing to its easy implementation and relatively fast convergence towards satisfactory solution [91, 92]. In this context, the PSO algorithm is exploited in this chapter to search for the optimal weights for each attribute in WNBCF, where the selection of weights satisfies the constraints mentioned previously. To test its superiority and effectiveness, our proposed method is compared with some conventional algorithms and applied to the FRDA patient baseline data collection problem under the cold-start condition. To this end, the main contributions of this chapter can be briefly summarized as follows:

1. A modified CF algorithm is proposed to tackle the cold-start problem during the FRDA patient-baseline-data collection, which provides valuable support to the clinical trials and further disease research.
2. By combining MI with PSO algorithm, a novel computational framework is established to fine tune the weights of our WNBCF method.
3. Comprehensive experiments are carried out to show that our proposed WNBCF algorithm indeed provides satisfactory prediction accuracy under the cold-start condition on the FRDA patient baseline data.

5.2 Methodology

Given an FRRS consisting of $\bar{m}$ patients and $\bar{n}$ test-items, the patients' profiles are represented by a $\bar{m} \times \bar{n}$ matrix called the patient-item matrix $R^{\bar{m}\times\bar{n}}$. The set of patients and test-items are defined as $U = \{u_1, u_2, ..., u_{\bar{m}}\}$ and $I = \{i_1, i_2, ..., i_{\bar{n}}\}$, respectively. Each element $r_{u,i}$ in this matrix represents the case that the patient u gives a rating value r on the test-item i, where

$u \in U$, $i \in I$. Each patient has his/her own specific side-information which is described as individual attributes. Some new patients might be incapable of providing any ratings on test-items, which gives rise to a kind of cold-start situation. In this situation, the FRRS is no longer applicable in finding similar patients in the patient-item matrix R, and it is thus impossible to accurately predict the unfilled ratings of these patients. As mentioned in Section 5.1, the main idea of our proposed method is based on the weighted NB, where the classification results calculated by WNBCF correspond to the ratings on test-items. In what follows, we give a brief introduction to basic weighted naive Bayes.

5.2.1 Weighted naive Bayes

Assume that $A_1, A_2, \ldots, A_n$ are n attributes in patient side-information. A data sample is represented by a vector $E = (a_1, a_2, \ldots, a_n)$, where a_k is the value of A_k ($k \in \{1, 2, \ldots, n\}$). All possible class labels are expressed as C. For any $c \in C$, based on Bayes theorem, the posterior probability $P(c|E)$ is written as:

$$P(c|E) = \frac{P(c)P(E|c)}{P(E)} \tag{5.1}$$

where $P(c|E)$ is the conditional probability of c given E; $P(E|c)$ is the conditional probability of E given c; $P(c)$ is the probability of c occurring and $P(E)$ is the probability of E occurring. In the NB, based on the conditional independence assumption, all attributes are independent of each other with respect to class variables. In this case, $P(E|c)$ is rewritten as:

$$P(E|c) = P(a_1, a_2, ..., a_n|c) = \prod_{k=1}^{n} P(a_k|c) \tag{5.2}$$

and (5.1) is rewritten as:

$$P(c|E) = \frac{P(c)P(E|c)}{P(E)} = \frac{P(c)}{P(E)} \prod_{k=1}^{n} P(a_k|c) \tag{5.3}$$

Since $P(E)$ is the same for all classes, the NB classifier is expressed by:

$$h_{nb}(E) = \arg\max_{c \in C} P(c) \prod_{k=1}^{n} P(a_k|c) \tag{5.4}$$

where the data sample E is classified to the class with the maximum posterior probability, $h_{nb}(E)$ represents the classification given by the NB algorithm and the conditional probability $\prod_{k=1}^{n} P(a_k|c)$ is directly calculated from the training set.

The NB classifier assumes that all conditional attributes are independent of each other on classification. In response to the influence of different attributes, the weighted NB classifier has been proposed to achieve a better prediction accuracy than the NB classifier. A weighted NB classifier is defined as:

$$h_{wnb}(E) = \arg\max_{c \in C} P(c) \prod_{k=1}^{n} P(a_k|c)^{w_k} \tag{5.5}$$

where $h_{wnb}(E)$ represents the class given by the weighted NB algorithm and w_k is the weight of the attribute value a_k. The larger the weight, the greater the impact. In addition, the Laplace smoothing method is adopted in this chapter to prevent the problem of zero probability, for example, if an attribute value appears with a class that is not in the training set, the multiplication calculation of equation (5.5) is equal to 0. The Laplace smoothing method is defined as follows:

$$P(c) = \frac{|T_c| + 1}{|T| + |C|} \tag{5.6}$$

and

$$P(a_k|c) = \frac{|T_{c,a_k}| + 1}{|T| + |C_k|} \tag{5.7}$$

where T is the training set; T_c denotes a set of samples belonging to class c in training set; T_{c,a_k} represents a set of samples with the value of a_k on the k-th attribute in T_c; and C_k represents possible classes to which the k-th attribute belongs.

5.2.2 Mutual information

The selection of weights plays an important role in the weighted NB algorithm. Because of the good capability of MI on quantifying the interdependency of two random variables, MI is utilized here to measure the mutual dependence between attribute variable and class variable in this chapter. A larger MI value indicates a stronger association, so the importance of attributes can be determined by the MI values.

For any attribute variable A_k and class variable C, the MI can be defined as:

$$MI(A_k, C) = H(A_k) + H(C) - H(A_k, C) \tag{5.8}$$

where $H(A_k)$ is the information entropy of attribute A_k, $H(C)$ is the information entropy of class C, and $H(A_k,C)$ is the joint entropy of attribute A_k and class C. $H(A_k)$ is defined as follows:

$$H(A_k) = -\sum_{a_k \in A_k} \mathscr{P}(a_k) \log_2 \mathscr{P}(a_k) \tag{5.9}$$

where $\mathscr{P}(a_k)$ is the frequency of occurrence of value a_k in attribute A_k, and $H(C)$ is defined as:

$$H(C) = -\sum_{c \in C} \mathscr{P}(c) \log_2 \mathscr{P}(c) \tag{5.10}$$

where $\mathscr{P}(c)$ is the frequency of occurrence of value c in variable C, and $H(A_k,C)$ is defined as:

$$H(A_k,C) = -\sum_{a_k \in A_k} \sum_{c \in C} \mathscr{P}(a_k,c) \log_2 \mathscr{P}(a_k,c) \tag{5.11}$$

where $\mathscr{P}(a_k,c)$ is the frequency of simultaneous occurrence of values a_k and c.

The MI value of each attribute is normalized to a notionally common scale with interval of $[0,1]$. Let the sum of the MI values be one. The formulation is given by:

$$\hat{w}_k = \frac{u(A_k,C)}{\sum_{k=1}^{n} u(A_k,C)} \tag{5.12}$$

where

$$u(A_k,C) = \frac{2 * MI(A_k,C)}{H(A_k) + H(C)} \tag{5.13}$$

Based on (5.12), the weights of the attributes can be obtained. Then, one of the powerful evolutionary computation algorithms, the PSO algorithm, is employed in our work to further optimize the weights. In a PSO algorithm, the initial positions of the particles are randomly chosen, and a good initial position could comprehensively improve the search performance of the optimizer. In this case, the convergence speed and the probability of finding the optimal solution are improved. In this chapter, the initial position of each particle satisfies the multivariate Gaussian distribution $\mathscr{N}(\hat{W},\Sigma)$, where

$$\hat{W} = \begin{bmatrix} \hat{w}_1 & \hat{w}_2 & \cdots & \hat{w}_n \end{bmatrix}^T$$

and Σ is a covariance matrix. Additionally, MI values are used to determine the relationship between attributes, and the correlation between the attributes is treated as a constraint in the PSO algorithm.

5.2.3 PSO-based parameter design

The PSO algorithm, which is inspired by the simulation of social behavior of fish-schooling/birds-flocking, is a well-known heuristic intelligent optimization algorithm. Given a set of data: $(E_1, c_1), (E_2, c_2), \ldots, (E_s, c_s)$, where $E_j = (a_{j1}, a_{j2}, \ldots, a_{jn})$ represents the values of attributes in the j-th sample, $c_j \in C$ represents the class of j-th sample.

The WNBCF method is a probability-based method, for example, if a data sample E_j appears in one class with the largest probability, then the sample E_j belongs to this class. Inspired by [87], the fitness function is designed based on the idea of maximizing the probability that the data sample E_j belongs to class c_j and minimizing the sum of probability that the data sample E_j belongs to other classes. We modify the fitness function in [87] and consider two constraints to achieve more reasonable and superior optimized results. We define an n-dimensional weight parameter vector as $W = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^T$. The fitness function of our work is constructed as follows:

$$
\begin{aligned}
f(W) = \sum_{j=1}^{s} \Big(& \sum_{\substack{c_i \in C \\ c_i \neq c_j}} P(c_i) \prod_{k=1}^{n} P\left(a_{jk} | c_i\right)^{w_k} \\
& - P(c_j) \prod_{k=1}^{n} P\left(a_{jk} | c_j\right)^{w_k} \Big)
\end{aligned}
\tag{5.14}
$$

where $w_k \in W$ represents the weight of k-th attribute, $\sum_{\substack{c_i \in C \\ c_i \neq c_j}} P(c_i) \prod_{k=1}^{n} P\left(a_{jk} | c_i\right)^{w_k}$ represents the sum of probabilities of the data sample a_j belonging to the classes except c_j.

It should be noted that the selection of weights is required to satisfy two constraints: 1) the relationship between weights satisfies the descending order according to the values of (5.12), and 2) the sum of the weights is equal to one.

The constrained optimization problem is defined as:

$$\begin{aligned} &\min f(W) \\ \text{s.t. } &g_k(W) = w_k - w_{k+1} \geq 0, k = 1,2,\ldots,n-1 \\ &h(W) = \sum_{k=1}^{n} w_k - 1 = 0 \end{aligned} \tag{5.15}$$

To convert the addressed constrained minimization problem into a series of unconstrained minimization problems, two penalty functions are introduced to (5.15) which leads to

$$\begin{aligned} &\min \mathrm{F}(W,\sigma,\phi) \\ &= f(W) + \sigma \sum_{k=1}^{n-1} [\max\{0, -g_k(W)\}]^2 + \phi[h(W)]^2 \\ &= f(W) + \sigma \sum_{k=1}^{n-1} [\max\{0, -(w_k - w_{k+1})\}]^2 + \phi \left| \sum_{k=1}^{n} w_k - 1 \right|^2 \end{aligned} \tag{5.16}$$

where $\sigma \sum_{k=1}^{n-1} [\max\{0, -g_k(W)\}]^2$ and $\phi[h(W)]^2$ are two exterior penalty functions, and σ and ϕ are penalty coefficients that are adjustable.

The position of the m-th particle ($m \in \{1,2,...,N\}$ with N being the size of swarm in PSO algorithm) at the $\hat{k}$-th iteration is expressed by a D-dimensional vector

$$X_m(\hat{k}) = \begin{bmatrix} x_{m1}(\hat{k}) & x_{m2}(\hat{k}) & \cdots & x_{mD}(\hat{k}) \end{bmatrix}^T.$$

The velocity of m-th particle at the $\hat{k}$-th iteration is represented by a D-dimensional vector

$$V_m(\hat{k}) = \begin{bmatrix} v_{m1}(\hat{k}) & v_{m2}(\hat{k}) & \cdots & v_{mD}(\hat{k}) \end{bmatrix}^T.$$

The historical best position of the m-th particle at the $\hat{k}$-th iteration is represented as

$$P_m(\hat{k}) = \begin{bmatrix} p_{m1}(\hat{k}) & p_{m2}(\hat{k}) & \cdots & p_{mD}(\hat{k}) \end{bmatrix}^T,$$

and the best particle of the swarm is denoted by $P_g(\hat{k})$. The velocity and the position updating equations of each particle are presented by the following two equations:

$$\begin{aligned} V_m(\hat{k}+1) = \bar{w}V_m(\hat{k}) + c_1 r_1 (P_m(\hat{k}) - X_m(\hat{k})) \\ + c_2 r_2 (P_g(\hat{k}) - X_m(\hat{k})) \end{aligned} \tag{5.17}$$

$$X_m(\hat{k}+1) = X_m(\hat{k}) + V_m(\hat{k}+1) \tag{5.18}$$

where $\bar{w}$ indicates the inertia weight factor; c_1 is a positive constant called the cognitive parameter and c_2 is another positive constant called the social parameter; r_1 and r_2 are two uniformly distributed random numbers in the interval of [0,1]; and $\hat{k}$ is the number of current iteration. (5.17) is applied to determine the velocity of the m-th particle at $(\hat{k}+1)$-th iteration, and (5.18) updates the position based on the velocity of m-th particle at $(\hat{k}+1)$-th iteration.

To improve the convergence speed and optimization performance of the PSO algorithm, a large collection of variant PSO algorithms have been proposed, see in [91, 92]. In this chapter, $\bar{w}$ is formulated according to the relationship between current and maximum iterations number as mentioned in [133, 134]. $\bar{w}$ is given as follows:

$$\bar{w}(\hat{k}) = \frac{\hat{k}_{\max} - \hat{k}}{\hat{k}_{\max}} \times (\bar{w}_i - \bar{w}_f) + \bar{w}_f \tag{5.19}$$

where $\hat{k}$ indicates the number of current iteration; $\hat{k}_{max}$ represents the number of maximum iteration in the experiment; $\bar{w}_i$ is the initial value of the inertia weight when $\hat{k} = 0$; $\bar{w}_f$ is the final value of the inertia weight when $\hat{k} = \hat{k}_{max}$. In general, $\bar{w}_i$ is set as 0.9, and $\bar{w}_f$ is set as 0.4. It is worth mentioning that a large value of $\bar{w}$ will contribute to the global exploration, and a small value of $\bar{w}$ will benefit the local exploitation.

The acceleration coefficients are important parameters to adjust the particle's own experience and group experience to influence the particle's motion trajectory. If the value of c_1 is small, the group experience will have a large impact on the particles. In this case, the algorithm will converge quickly but it may converge to the local optima. A small c_2 will reduce the group interaction, and it is difficult for the particles to converge to the optimal solution. In this chapter, c_1 and c_2 are chosen according to [117]:

$$c_1 = \frac{\hat{k}_{\max} - \hat{k}}{\hat{k}_{\max}} \times (c_{1i} - c_{1f}) + c_{1f} \tag{5.20}$$

$$c_2 = \frac{\hat{k}_{\max} - \hat{k}}{\hat{k}_{\max}} \times (c_{2i} - c_{2f}) + c_{2f} \tag{5.21}$$

where c_{1i} and c_{1f} denote the initial and final value of the acceleration coefficient c_1, respectively; c_{2i} and c_{2f} are the initial and final value of the acceleration coefficient c_2, respectively. The values of c_{1i}, c_{1f}, c_{2i} and c_{2f} are set to be $2.5, 0.5, 0.5$ and 2.5, respectively. The optimal parameter vector $W = P_g(\hat{k}_{\max})$ is obtained when the PSO algorithm terminates. Once the optimal weights are determined, we can use the (5.5) to calculate the ratings on test-items based on patient side-information.

Table 5.1 Key attributes in patient side-information

	1	2	3	4	5	6	7
Disease duration (year)	0 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 and above
Onset age (year)	0 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 and above
ISCED education level	ISCED 0	ISCED 1	ISCED 2	ISCED 3	ISCED 4	ISCED 5	ISCED 6
Sex	Male	Female	*	*	*	*	*
Children	Yes	No	*	*	*	*	
Years of education	0 to 5	6 to 10	11 to 15	16 to 20	21 and above	*	*
Marital status	Single	Married	Widowed	Divorced	Separated	In a relationship	*
Employment	Yes	No	*	*	*	*	*

To sum up, the proposed WNBCF algorithm is designed to solve the cold-start problem of the FRRS under the condition that the new/follow-up patients cannot provide any ratings on test-items. The pseudocode of the proposed WNBCF algorithm is shown in Algorithm 5.2.3.

Algorithm 1: Hybrid CF framework

```
• Input: Patient side-information data, rating data on test-items, some parameters in the PSO method;

1.  Divide the known data into the training set, the validation set and the testing set with a certain proportion;
2.  Calculate the class prior probability $P(c)$ ($c \in C$) and the class conditional probability $P(a_k|c)$ between each attribute and class in training set;
3.  Determine the $\hat{W}$ by (5.12);
4.  Employ the PSO technique to select the optimal weight parameter vector $W$ (presented in Steps 5-16):
5.  Initialize velocity for each particle;
6.  Set the position for each particle that satisfies the multivariate normal distribution $\mathcal{N}(\hat{W}, \Sigma)$;
7.  for For $\hat{k} = 0$ to $\hat{k}_{\max}$
8.      for$p = 1$ to $N$
9.          Calculate the probability of occurrence of each class by using the weighted NB method based on (5.3);
10.         Calculate the fitness value based on (5.14);
11.         Obtain $P_m(\hat{k})$;
12.     end for
13.     Obtain $P_g(\hat{k})$;
14.     Update each particle's velocity and position based on (5.18);
15. end for
16. Set $W = P_g(\hat{k}_{\max})$;
17. Calculate the posterior probability of each class with the corresponding weights on attributes;
18. Predict the unfilled rating values in the testing set according to (5.5);
```

5.3 Implementation and experiments

5.3.1 Data description

In this chapter, the datasets of Demographics, Onset and Scale for the Assessment and Rating of Ataxia (SARA) have been selected from the database provided by the EFACTS. Here, we use the data of 874 patients in the experiments.

Side-information datasets: The side-information is extracted from the Demographics dataset and Onset dataset. The Demographics dataset includes the basic demographic information of FRDA patients. The Onset dataset includes information on the patient's first symptoms and basic diagnosis of FRDA. The main attributes in these two datasets for algorithmic applications are listed in Table 5.1.

5.3.2 Data pre-processing

As we mentioned in the introduction section, different rating values of each test-item in the SARA dataset correspond to the classes, and the side-information from Demographics dataset and Onset dataset corresponds to attributes.

The rating values of test-items in the SARA dataset are from "0" to "4, 6 or 8". In order to facilitate the subsequent experiments, "0" sets as "class 1", "1" sets as "class 2", and so on. The SARA overall score (0 to 40) is divided into 8 classes where the score ($0 \sim 5$) is set as "class 1", ($6 \sim 10$) is set as "class 2", and so on.

Patient side-information is listed in Table 5.1, which are disease duration, onset age, ISCED education level, and so on. Due to the reason that patient side-information has different expressional forms, we unify them into discrete values from 1 to up to 7 respectively.

To sum up, we will use patient side-information which includes 8 attributes to predict rating values (classes) of each test-item and SARA overall score.

5.3.3 Experiment setting

In our simulation, 874 patients are divided into the training set (70%), the validation set (15%), and the testing set (15%). The training set is used to train parameters in our WNBCF method. The initial positions of the particles satisfy the multivariate Gaussian distribution $\mathcal{N}(\hat{W}, \Sigma)$ in Section 5.2.3. In this chapter, the elements on the diagonal of the covariance matrix Σ are set to be 0.1. After that, the data in the training set are used to select the weights

Table 5.2 Experimental results under different densities

Metrics	Methods		Name of test-items						
			Gait	Stance	Sitting	Speech disturbance	Right finger chase	Left finger chase	Right nose-finger
MAE	IWNB-CF	Minimum	1.0920	0.8514	0.7257	0.6857	0.5371	0.5486	0.6171
		Mean	1.3065	1.0571	0.8771	0.8421	0.6274	0.6440	0.7217
	NB	Minimum	1.1552	0.8743	0.7543	0.7143	0.5314	0.5543	0.6629
		Mean	1.3880	1.0934	0.9048	0.8477	0.6369	0.6672	0.7422
	Weighted NB	Minimum	1.1264	0.8914	0.7143	0.6914	0.5486	0.5714	0.6800
		Mean	1.3144	1.0560	0.9069	0.8562	0.6341	0.6649	0.7629
RMSE	IWNB-CF	Minimum	1.8352	1.4323	1.2259	1.0198	0.9196	0.8751	0.9681
		Mean	2.0996	1.7136	1.3886	1.2375	1.0033	0.9960	1.0664
	NB	Minimum	1.9238	1.4501	1.2513	1.0850	0.8281	0.8685	0.9798
		Mean	2.1792	1.7156	1.4035	1.2495	1.0057	1.0215	1.0841
	Weighted NB	Minimum	1.9579	1.4462	1.1735	1.0610	0.8816	0.8685	1.0226
		Mean	2.1193	1.7295	1.4074	1.2663	0.9912	1.0172	1.1104

Metrics	Methods		Left nose-finger	Right alternating hand movements	Left alternating hand movements	Right heel-shin slide	Left heel-shin slide	Sara total
MAE	WNBCF	Minimum	0.6914	0.6629	0.6571	0.6514	0.6686	1.0743
		Mean	0.7640	0.7916	0.7937	0.7629	0.7486	1.1625
	NB	Minimum	0.6514	0.6857	0.7086	0.6629	0.6057	1.0629
		Mean	0.7795	0.8232	0.8013	0.7597	0.7467	1.1976
	Weighted NB	Minimum	0.6514	0.6514	0.6800	0.6286	0.6229	1.0114
		Mean	0.7712	0.7930	0.7872	0.7472	0.7455	1.1981
RMSE	WNBCF	Minimum	1.0057	0.9562	0.9289	1.0797	1.0823	1.4794
		Mean	1.0997	1.1016	1.0852	1.2057	1.1722	1.6337
	NB	Minimum	0.8944	0.9739	1.0029	1.0744	1.0392	1.5062
		Mean	1.1130	1.1311	1.1065	1.1967	1.1634	1.6669
	Weighted NB	Minimum	0.9562	0.9681	0.9651	0.9592	0.9914	1.4283
		Mean	1.1110	1.1007	1.0905	1.1720	1.1682	1.6713

to minimize the fitness function by using the PSO algorithm. The validation set is used to validate the effect of weights. In the testing set, the patient side information is regarded as known, and all the ratings on test-items are unknown and need to be predicted. For any set of patient side information, it corresponds to 13 test items. Two conventional algorithms which are basic NB algorithm and basic weighted NB algorithm are employed to compare with our WNBCF for the rating prediction accuracy.

To measure the prediction accuracy of algorithms, two most popular evaluation indicators which are Mean absolute error (MAE) and root mean square error (RMSE) are introduced. MAE and RMSE are two representative metrics of accuracy-based metrics. The accuracy-based metrics is utilized to measure the average error between the ground truth and predicted values. RMSE is more sensitive to large errors than MAE, therefore, RMSE is more useful for the system where large errors are particularly undesirable. The MAE and RMSE between actual ratings and predicted ratings are given as follows:

$$\mathrm{MAE} = \frac{1}{|\mathscr{T}|} \sum_{(u,i)\in\mathscr{T}} |r_{u,i} - \hat{r}_{u,i}| \tag{5.22}$$

and

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathscr{T}|} \sum_{(u,i)\in\mathscr{T}} (r_{u,i} - \hat{r}_{u,i})^2} \tag{5.23}$$

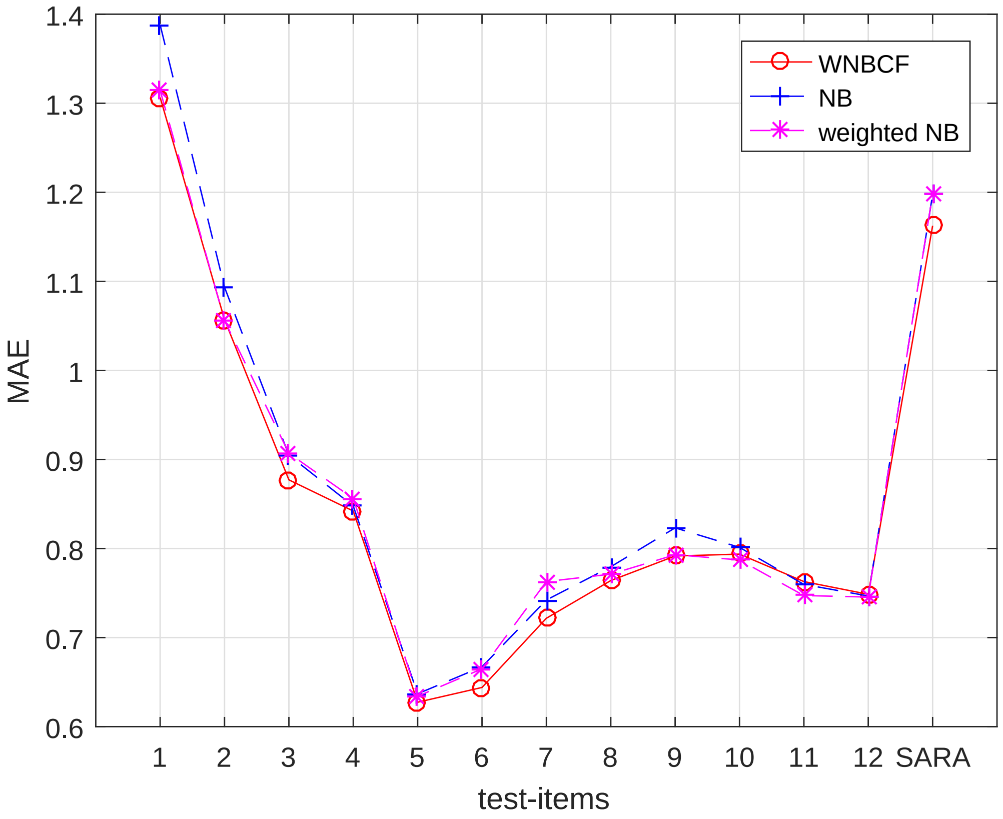


Fig. 5.1 MAE metric under different densities.

where $|\mathscr{T}|$ represents the total number of predicted values in the testing set; $\hat{r}_{u,i}$ denotes the predicted value generated in the testing set $\mathscr{T}$.

In our simulation, experimental parameters are as follows: 1) the dimension of each particle is $D = 8$; 2) the size of the swarm s is set to be 20; 3) the maximum iteration is set to be $K = 3000$; 4) the search space is in the interval of (0, 1]; and 5) penalty coefficients σ and ϕ are set to be 50.

5.3.4 Results and discussion

To comprehensively evaluate its effectiveness, the proposed WNBCF algorithm has been applied to the FRDA patient baseline data prediction problem. The performance of the WNBCF algorithm is evaluated by comparing it with the performance of other conventional algorithms, including the basic NB algorithm and the basic weighted NB algorithm. In the basic weighted NB algorithm, the weights are determined directly by using MI without using

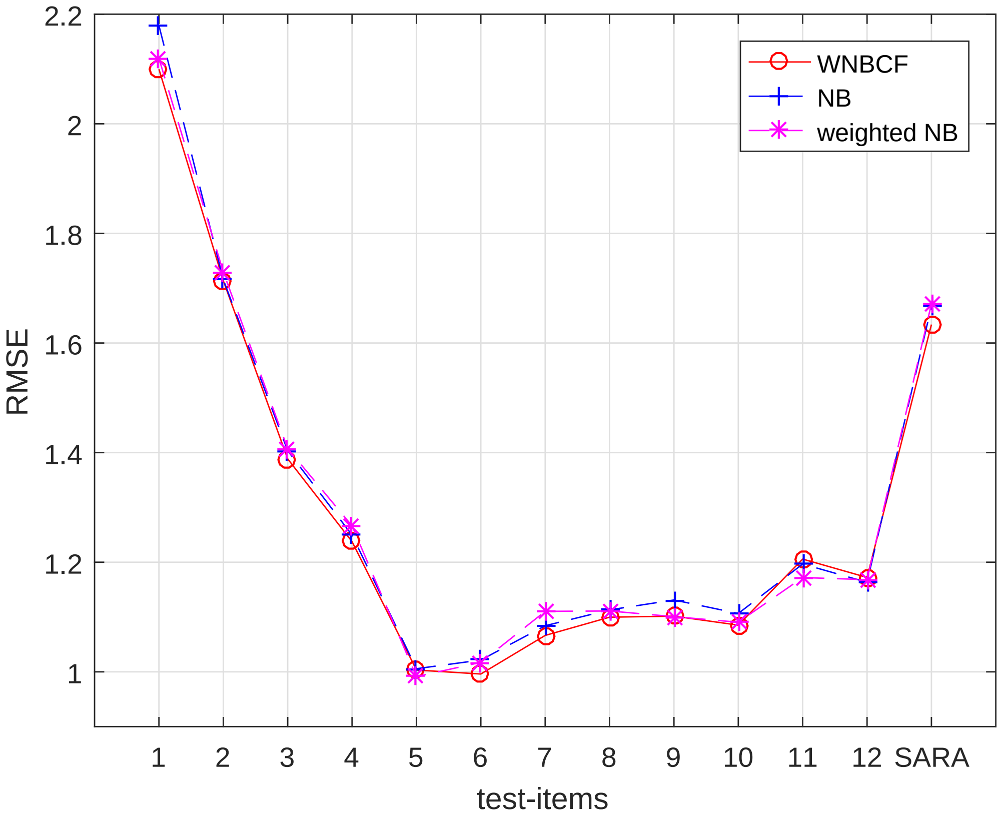


Fig. 5.2 RMSE metric under different densities.

the PSO algorithm to optimize. The patient side information is utilized to predict the ratings on all the 12 test-items and the SARA total score. Each experiment is repeated 50 times to avoid random influence. The minimum and average values of the MAE and RMSE on each test-items are recorded.

Experiment results of the WNBCF algorithm, the NB algorithm and the weighted NB algorithm are shown in Table 5.3. It can be seen that the proposed algorithm demonstrates its superiority over the NB algorithm and the weighted NB algorithm in evaluation indices of the MAE and RMSE. Compared to the basic weighted NB, the introduction of PSO has the advantage of choosing suitable weights. In addition, most results of the weighted than that of NB have the lower MAE and RMSE than NB which indicates that different attributes have different degrees of importance to the result. Compared with other test-items in the same interval, MAE and RMSE of the “Right finger chase” and “Left finger chase” are much lower than others, which means that patient side-information has a greater impact on these two test-items than other test-items. To sum up, experiment results have shown the effectiveness of the proposed WNBCF algorithm on the FRDA patient baseline data.

5.4 Conclusion

In this chapter, a modified WNBCF algorithm has been proposed to solve the cold-start problem in the FRRS. By employing the WNBCF algorithm, the patient side information is utilized for the missing value prediction. In addition, the PSO algorithm has been applied to automatically select appropriate weights in the WNBCF algorithm. The developed WNBCF algorithm has been successfully applied to the actual FRDA baseline data collection problem with satisfactory performance. In the situation that the patients are unable to provide any rating data, our algorithm can produce reasonable prediction results, which gives a new solution to aid the FRDA patient baseline data collection. Experiment results have shown the feasibility and effectiveness of our proposed algorithm by comparing it with some conventional algorithms.

Our future work aims to 1) adopt the popular deep learning techniques to dig deep latent factors of the FRDA patients and test-items from patient side information and disease-related information; and 2) study the time-series modeling of disease progression of FRDA patients.

Chapter 6

An Optimally Weighted User- and Item-based Collaborative Filtering Approach to Predicting Baseline Data for Friedreich's Ataxia Patients

6.1 Motivation

During the past few decades, the recommendation systems (RSs) have received an ever-increasing interest from various communities such as computer science, engineering research and medical applications [192, 198, 93]. Owing to their outstanding performance in providing users with product or service recommendations, the RSs have found successful applications in a variety of domains including e-commerce, music, movies, news and so on [178, 95, 94]. In order to recommend goods and services that users are interested in, the RSs mainly employ information filtering technology to analyze users' requirements by mining user behavior data.

Collaborative filtering (CF), as one of the most successful recommendation techniques, has been receiving considerable attention ever since the mid-1990s with fruitful applications in the development of various RSs by Amazon, YouTube, Netflix and so on [17]. Generally speaking, the well-known CF-based recommendation algorithms (RAs) include the user-based CF (UBCF) algorithms and the item-based CF (IBCF) algorithms. The main idea of the UBCF algorithms is to analyze the user behaviors to find similar users (named as neighbors) in the communities. In this case, the items are recommended to a target user based on his/her neighbors' interested items. Similarly, the IBCF algorithms make use of the similarity between the items rather than users. The items that are similar to those in which the target user is interested are recommended to the concerned user.

It should be noticed that the similarity measures play a critical role in the CF-based RAs. Some commonly used similarity measures in the UBCF and IBCF algorithms include the adjusted cosine (AC), cosine, and Pearson correlation coefficient (PCC) measures. Nevertheless, in the case that the user behaviors are complicated, the performance of the CF-based RAs which use the PCC, cosine or AC as the similarity measure cannot be always guaranteed. As such, tremendous efforts have been devoted to the design of more comprehensive similarity measures [187, 40, 69, 78, 68, 197]. For example, the Shannon entropy has been employed to quantify the users' rating habits [69, 78], where the difference of entropy between users has been utilized as the weight to adjust the result of similarity.

While the state-of-the-art similarity measures have helped improving the prediction accuracy of the RAs, most of the measures take *either* users *or* items to predict the missing values. It has been shown in some literature that the combination of the UBCF method and the IBCF method could effectively improve the performance of the RSs [197, 198, 16, 113, 191]. In [197], the confidence weights, which use the degree of similarity of the neighbors as a reference, have been utilized to balance the predictions obtained by the UBCF method and the IBCF method. In the typical RAs, only positively correlated neighbors are utilized to compute the similarity between the users/items. Nevertheless, the negatively correlated neighbors are also useful in predicting the missing values from another perspective [73]. In this context, a seemingly natural idea is to combine the UBCF and IBCF methods by developing a new prediction model where the positively and negatively correlated neighbors in both methods are taken into account.

To balance the impacts from the UBCF method and the IBCF method, a typical approach is to introduce the weighting parameters to predict the missing values, where the weighting parameters are utilized to make an adequate tradeoff between the positively and negatively correlated neighbors in the UBCF/IBCF methods. It is worth mentioning that, in the literature, such weighting parameters have been manually selected according to engineering practice by means of certain rules on an ad-hoc basis [197, 198]. Clearly, manual selection of the weighting parameters requires in-depth domain knowledge and specific fine-tuning techniques, which is not always possible in practice. As such, it makes practical sense to *automate* the parameter selection algorithm with locally optimized performance.

In search of an effective algorithm capable of locating optimally weighted parameters in terms of improving the prediction performance, the Evolutionary computation (EC) algorithms appear to be an ideal candidate. EC algorithms have shown distinguished advantages in solving optimization problems in a diverse range of real-world applications including telecommunication, signal processing, system science and so on [183, 185]. An effective yet popular EC algorithm is the so-called particle swarm optimization (PSO) algorithm that owns the distinctive advantages of easy implementation, quick convergence and great competence in effectively searching the global optimum. So far, the PSO algorithm has gained much attention from both academia and industry with successful applications in solving various multi-objective optimization problems, see e.g. [91, 92]. Owing to its particular suitability, the PSO algorithm is exploited in this chapter to optimize the weighting parameters in order to achieve an adequate tradeoff between the positively and negatively correlated neighbors in terms of predicting the rating values.

Motivated by the above discussions, we propose a modified CF (MCF) algorithm in this chapter by combining the merits of UBCF and IBCF methods. Through the utilization of the information from both the positively and negatively correlated neighbors, the proposed algorithm is capable of predicting the missing values in multi-aspects with satisfactory accuracy. In particular, the PSO algorithm is dedicatedly exploited to determine (locally) optimized weights of our proposed MCF algorithm so as to further improve the prediction accuracy. To illustrate its application potential, our proposed algorithm is applied to assist with the baseline data collection for Friedreich's ataxia (FRDA) patients. The main contributions are summarized as follows:

1. An MCF algorithm is proposed which not only combines the merits from the UBCF and IBCF methods but also makes full use of the positively and negatively correlated neighbors in predicting the missing values.

2. The PSO algorithm is utilized to optimize the weights in the MCF algorithm so as to achieve a) an adequate tradeoff between the user-based and the item-based similarity measures; and b) a proper balance between the positively and negatively correlated neighbors.

3. The developed algorithm is successfully applied to the FRDA assessment system to assist clinical sample collection for FRDA patients who are unable to attend the tests in the study sites.

The remainder of this chapter is structured as follows. The detailed introduction of the proposed MCF approach is presented in Section 6.2. The performance of our proposed MCF approach is evaluated in the case of a real-world neurological disease in Section 6.3. Finally, conclusions are drawn in 6.4.

6.2 Main results

Given an RS consisting of m users and n items, the user profiles are denoted by a $m \times n$ matrix called the user-item matrix $R^{m \times n}$. The sets of users and items are defined as $U = \{u_1, u_2, \ldots, u_m\}$ and $I = \{i_1, i_2, \ldots, i_n\}$, respectively. Each element $r_{u,i}$ in R represents that the user u rates the value r on the item i, where $u \in U$, $i \in I$. If the user u has rated the item i, then $r \in 1, 2, \ldots, \tilde{r}$ ($\tilde{r}$ is the upper bound of the ratings). Furthermore, $r_{u,i} = \emptyset$ if the user u does not rate the item i.

6.2.1 Computation of similarity

The PCC is one of the most well-known similarity measures in RSs due to its high prediction accuracy and easy implementation [20, 197]. In the UBCF algorithm, the PCC similarity degree between user u and user a is calculated according to the following formula:

$$\mathrm{Sim}_{u,a}^{PCC} = \frac{\sum_{i \in I_{u,a}} (r_{u,i} - \bar{r}_u)(r_{a,i} - \bar{r}_a)}{\sqrt{\sum_{i \in I_{u,a}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{u,a}} (r_{a,i} - \bar{r}_a)^2}} \tag{6.1}$$

where $\mathrm{Sim}_{u,a}^{PCC}$ is the PCC similarity degree between users u and a; $I_{u,a} = I_u \cap I_a$ is the subset of items on which both users u and a have rated, where I_u denotes all the items that have been evaluated by user u and I_a denotes all the items that have been evaluated by user a; $r_{u,i}$ indicates the rating value of item i rated by user u and $r_{a,i}$ indicates the rating value of item i rated by user a; $\bar{r}_u$ is the mean rating value of items that user u has rated; and $\bar{r}_a$ is the mean rating value of items that user a has rated. The values calculated by (6.1) are in the range of -1 to 1. A larger value of $\mathrm{Sim}_{u,a}^{PCC}$ means that the user u and user a are more similar.

In the IBCF algorithm, the AC method is introduced to evaluate the degree of similarity between the item i and item j by the following formula [125]:

$$\mathrm{Sim}_{i,j}^{AC} = \frac{\sum_{u \in U_{i,j}} (r_{u,i} - \bar{r}_u)(r_{u,j} - \bar{r}_u)}{\sqrt{\sum_{u \in U_{i,j}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{u \in U_{i,j}} (r_{u,j} - \bar{r}_u)^2}} \tag{6.2}$$

where $\mathrm{Sim}_{i,j}^{AC}$ is the AC similarity between items i and j; $U_{i,j} = U_i \cap U_j$ is the subset of users who have rated both item i and item j, where U_i denotes the users who have rated item i and U_j denotes the users who have rated item j; and $r_{u,j}$ denotes the rating value provided by user u on item j. Notice that the values calculated by AC are in the range of -1 to 1.

6.2.2 Neighbor selection

Traditionally, the top-k algorithm is used to rank the neighbors based on their similarity degrees in the descending order, and then the top k neighbors are chosen to predict the missing values. As mentioned previously, the values of $\mathrm{Sim}_{u,a}^{PCC}$ and $\mathrm{Sim}_{i,j}^{AC}$ lie in the range of [-1, 1]. The closer that similarity of PCC/AC is to 1, the more similar the users/items are. Users with positive correlations can undoubtedly be used to make predictions. On the contrary, negative correlation also expresses the relationship between two users from the negative side. The closer that similarity of PCC/AC is to -1, the more dissimilar the users/items are. For example, if users u and a have the similarity of -1, it means when user u rates an item with a high value then user a will definitely give a low value on that item, and vice versa. To sum up, the neighbors with both positive and negative correlations should be utilized to forecast the missing values from different perspectives. The neighbor selection has always been a key yet hot topic in RSs. A large number of neighbor selection strategies have been designed with hope to improve the RS's performance. Based on the neighbor selection strategy suggested by Breese [20], the neighbors with high correlations are more valuable than those with low correlations. Therefore, the positive and negative neighbor sets of user u and item i are formed by:

$$\mathrm{Pos}_u = \{a^+ | \mathrm{Sim}_{u,a^+}^{PCC} > 0.5, a^+ \neq u\} \tag{6.3}$$

$$\mathrm{Neg}_u = \{a^- | \mathrm{Sim}_{u,a^-}^{PCC} < -0.5, a^- \neq u\} \tag{6.4}$$

$$\mathrm{Pos}_i = \{j^+ | \mathrm{Sim}_{i,j^+}^{PCC} > 0.5, j^+ \neq i\} \tag{6.5}$$

$$\mathrm{Neg}_i = \{j^- | \mathrm{Sim}_{i,j^-}^{PCC} < -0.5, j^- \neq i\} \tag{6.6}$$

where Pos_u represents the set of similar users having positive correlation with user u; Neg_u represents the set of similar users having a negative correlation with user u; Pos_i indicates the set of similar items having positive correlation with item i; and Neg_i indicates the set of similar items having negative correlation with item i.

6.2.3 Prediction of missing values

In the UBCF methods, the missing values on items are predicted by utilizing positively correlated neighbors of users according to the following formula [20]:

$$\hat{r}_{u,i} = \bar{u} + \frac{\sum_{a^+ \in \text{Pos}_u} \text{Sim}^{PCC}_{u,a^+}(r_{a^+,i} - \bar{a}^+)}{\sum_{a^+ \in \text{Pos}_u} \text{Sim}^{PCC}_{u,a^+}} \tag{6.7}$$

where $\hat{r}_{u,i}$ is the predicted value of $r_{u,i}$; $\bar{u}$ is the mean value of different items provided by user u; and $\bar{a}^+$ is the mean value of items provided by the user a^+ who has the positive similarity degree with the target user u. For the UBCF methods that utilize the negative correlation neighbors, the missing values of the test-item are predicted by the following formula:

$$\hat{r}_{u,i} = \bar{u} - \frac{\sum_{a^- \in \text{Neg}_u} \text{Sim}^{PCC}_{u,a^-}(r_{a^-,i} - \bar{a}^-)}{\sum_{a^- \in \text{Neg}_u} \text{Sim}^{PCC}_{u,a^-}} \tag{6.8}$$

where $\bar{a}^-$ represents the mean value of items rated by the user a^- who has the negative similarity degree with target user u.

In the IBCF methods employing the positive neighbors, the missing values of the test-items are determined based on

$$\hat{r}_{u,i} = \bar{i} + \frac{\sum_{j^+ \in \text{Pos}_i} \text{Sim}^{AC}_{i,j^+}(r_{u,j^+} - \bar{j}^+)}{\sum_{j^+ \in \text{Pos}_i} \text{Sim}^{AC}_{i,j^+}} \tag{6.9}$$

where $\bar{i}$ represents the average values of item i rated by users, and $\bar{j}^+$ is the average value of item j^+ which has the positive similarity degree with the target item i. To be specific, the missing values on items are predicted by utilizing the negatively correlated neighbors according to the following formula:

$$\hat{r}_{u,i} = \bar{i} - \frac{\sum_{j^- \in \text{Neg}_i} \text{Sim}^{AC}_{i,j^-}(r_{u,j^-} - \bar{j}^-)}{\sum_{j^- \in \text{Neg}_i} \text{Sim}^{AC}_{i,j^-}} \tag{6.10}$$

where $\bar{j}^{-}$ is the average value of the item j^{-} which has the negative similarity degree with the target item i.

In our work, the UBCF method and the IBCF method are combined where both the positively and the negatively correlated neighbors are taken into account to predict the missing values. Three weighting parameters are employed in the developed MCF algorithm in order to achieve 1) a proper balance between the UBCF method and the IBCF method, 2) an adequate tradeoff between the positively and negatively correlated neighbors in UBCF method, and 3) an adequate tradeoff between the positively and negatively correlated neighbors in IBCF method. The formula for prediction is shown as follows:

$$\begin{aligned}
&\hat{r}_{u,i}\\
=&\alpha \times \left(\bar{u}+\lambda \times \frac{\sum_{a^{+}\in \mathrm{Pos}_u} \mathrm{Sim}^{PCC}_{u,a^{+}}(r_{a^{+},i}-\bar{a}^{+})}{\sum_{a^{+}\in \mathrm{Pos}_u} \mathrm{Sim}^{PCC}_{u,a^{+}}}\right.\\
&\left.-(1-\lambda)\times \frac{\sum_{a^{-}\in \mathrm{Neg}_u} \mathrm{Sim}^{PCC}_{u,a^{-}}(r_{a^{-},i}-\bar{a}^{-})}{\sum_{a^{-}\in \mathrm{Neg}_u} \mathrm{Sim}^{PCC}_{u,a^{-}}}\right)\\
&+(1-\alpha)\times \left(\bar{i}+\beta \times \frac{\sum_{j^{+}\in \mathrm{Pos}_i} \mathrm{Sim}^{AC}_{i,j^{+}}(r_{u,j^{+}}-\bar{j}^{+})}{\sum_{j^{+}\in \mathrm{Pos}_i} \mathrm{Sim}^{AC}_{i,j^{+}}}\right.\\
&\left.-(1-\beta)\times \frac{\sum_{j^{-}\in \mathrm{Neg}_i} \mathrm{Sim}^{AC}_{i,j^{-}}(r_{u,j^{-}}-\bar{j}^{-})}{\sum_{j^{-}\in \mathrm{Neg}_i} \mathrm{Sim}^{AC}_{i,j^{-}}}\right)
\end{aligned} \tag{6.11}$$

where α denotes the weight for the UBCF method; $(1-\alpha)$ denotes the weight for the IBCF method; λ and $(1-\lambda)$ represent the weights of the positively correlated neighbors and negatively correlated neighbors in the UBCF method, respectively; β and $(1-\beta)$ denote the weights of the positively correlated neighbors and negatively correlated neighbors in the IBCF method, respectively.

It is worth mentioning that the formula (6.11) would be degenerated into that for the traditional UBCF algorithm when α and λ are equal to 1, and into that of the traditional IBCF algorithm when $\alpha = 0$ and $\beta = 1$.

6.2.4 PSO-based parameter selection strategy

The PSO algorithm, which is a popular evolutionary computation algorithm inspired by the simulation of the social behavior of fish-schooling/birds-flocking, is applied in this chapter to dispose of the parameter optimization problem because of its competitive strength in seeking a relatively satisfactory solution as well as its easy-to-implement feature [91]. Here, each particle in the swarm indicates a candidate solution to the research problem.

In the proposed MCF algorithm, we select three appropriate weighting parameters to guarantee the prediction performance. The weights are expressed by a 3-dimensional vector as follows:

$$\omega \triangleq \begin{bmatrix} \alpha & \beta & \lambda \end{bmatrix}^T .$$

Without loss of generality, we divide the user-item matrix R into the training set (with 60 percent of the data), the validation set (with 20 percent of the data) and the testing set (with 20 percent of the data). The training set is applied to train the weighting parameters, and the validation set is utilized to validate the predicted results by using the trained weighting parameters. As the prediction accuracy reaches the desired threshold, the trained weighting parameters are applied to predict the results in the testing set.

The fitness function of the PSO algorithm is shown as follows:

$$fitness = \frac{1}{|V|} \sum_{r_{u,i} \in V} |r_{u,i} - \hat{r}_{u,i}| \tag{6.12}$$

where V represents the validation set, $|V|$ denotes the number of ratings in the validation set and $\hat{r}_{u,i}$ is calculated by formula (6.11).

Our attention is focused on choosing suitable ω so as to minimize the fitness function of the PSO algorithm. The optimization problem in our work is defined by:

$$\omega^* = \arg\min fitness \tag{6.13}$$

In this chapter, the particles move at a certain speed in a 3-dimensional search space. Denote

$$v_m(k) = \begin{bmatrix} v_{m1}(k) & v_{m2}(k) & v_{m3}(k) \end{bmatrix}^T ,$$
$$\omega_m(k) = \begin{bmatrix} \omega_{m1}(k) & \omega_{m2}(k) & \omega_{m3}(k) \end{bmatrix}^T$$

as the velocity and position of the m-th particle at the k-th iteration, respectively. The historical best position of the m-th particle ($m = 1, 2, \ldots, N$) at the k-th iteration and the global best position detected by the entire swarm are, respectively, denoted by

$$p_m(k) = \begin{bmatrix} p_{m1}(k) & p_{m2}(k) & p_{m3}(k) \end{bmatrix}^T,$$
$$g(k) = \begin{bmatrix} g_1(k) & g_2(k) & g_3(k) \end{bmatrix}^T.$$

The velocity and the position of the m-th particle are updated by the following equation:

$$\begin{aligned} v_m(k+1) &= wv_m(k) + c_1 r_1 (p_m(k) - \omega_m(k)) \\ &\quad + c_2 r_2 (g(k) - \omega_m(k)) \\ \omega_m(k+1) &= \omega_m(k) + v_m(k+1) \end{aligned} \tag{6.14}$$

where w is the inertia weight factor; c_1 is the acceleration coefficient called the cognitive parameter, and c_2 is another acceleration coefficient called the social parameter; r_1 and r_2 are two random numbers that satisfy the uniform distribution in the range of 0 to 1; k is the number of current iteration.

In order to enhance the search ability and reduce the possibility of getting trapped into local optima, lots of improved algorithms have been proposed to adjust the parameters in PSO algorithm. In this chapter, w is formulated according to the relationship between current iteration and maximum iteration number as mentioned in [133, 134], which is given as follows:

$$w(k) = w_f + (w_i - w_f) \times \frac{k_{\max} - k}{k_{\max}} \tag{6.15}$$

where k and $k_{\max}$ are the number of current iteration and maximum iteration, respectively; w_i is the initial inertia weight value when $k = 0$, and w_f indicates the final value of the inertia weight when $k = k_{\max}$.

In this chapter, the initial and final inertia weights values are set as $w_i = 0.9$ and $w_f = 0.4$, respectively. In general, a large inertia weight will benefit the global exploration at the early stage and a small inertia weight will help the local exploitation at the later stage. In addition, the acceleration coefficients c_1 and c_2 are calculated by the following equations [117]:

$$c_1 = c_{1f} + (c_{1i} - c_{1f}) \times \frac{k_{\max} - k}{k_{\max}} \tag{6.16}$$

$$c_2 = c_{2f} + (c_{2i} - c_{2f}) \times \frac{k_{\max} - k}{k_{\max}} \tag{6.17}$$

where c_{1i} denotes the initial value of cognitive acceleration coefficient c_1 and c_{1f} denotes the final value of cognitive acceleration coefficient c_1, c_{2i} denotes the initial value of cognitive acceleration coefficient c_2 and c_{2f} denotes the final value of cognitive acceleration coefficient c_2. According to experiment experience, the values of c_{1i}, c_{1f}, c_{2i} and c_{2f} are set to be 2.5, 0.5, 0.5, and 2.5, respectively. Finally, when the PSO algorithm terminates, we can obtain the optimal parameter vector as $\omega^* = g(k_{\max})$, where $k_{\max}$ represents the number of maximum iteration.

Table 6.1 Description of SARA dataset

Gait	Stance	Sitting	Speech disturbance	Finger chase			Nose-finger test		
				right	left	*mean*[a]	right	left	*mean*[a]
0 to 8	0 to 6	0 to 4	0 to 6	0 to 4	0 to 4	0 to 4	0 to 4	0 to 4	0 to 4

Fast alternating hand movements			Heel-shin slide			*SARATotal*[b]
right	left	*mean*[a]	right	left	*mean*[a]	
0 to 4	0 to 4	0 to 4	0 to 4	0 to 4	0 to 4	0 to 40

a: The mean indicates the average value of right and left sides.

b: The SARA Total indicates the sum of the values on first 4 test-items and the mean values on last 4 test-items.

The pseudocode of the MCF algorithm is shown in Algorithm 6.2.4 on next page.

The MCF Algorithm

- **Input:** User-item rating matrix R, k in top-k method, parameters in the PSO algorithm

```
1. Divide all the known data in $R$ into the training set and the validation set with a certain proportion;
2. Calculate the PCC similarity between users and the AC similarity between items by utilizing the data in the training set;
3. Employ the PSO technique to select the optimal parameter vector $\omega^*$ (presented in Steps 4-14) on the validation set:
4. Initialize velocity and position for each particle;
5. for$k = 0$ to $k_{\max}$
6.     for$p = 1$ to $N$
7.         Predict the rating values on the validation set based on equation (6.11);
8.         Calculate the fitness value based on equation (6.12);
9.         Obtain $p_m(k)$;
10.    end for
11.    Obtain $g(k)$;
12.    Update velocity and position for each particle based on equation (6.14);
13. end for
14. Set $\omega^* = g(k_{\max})$;
15. Calculate the PCC similarity between users and the AC similarity between items by utilizing all the known data;
16. Predict the missing values in $R$ by equation (6.11) according to the values of $\alpha, \beta, \lambda$ in $\omega^*$.
```

- **Output:** The predictions of missing values in rating matrix R

6.3 Application in friedreich's ataxia assessment system

6.3.1 FRDA assessment with the help of CF method

Friedreich's ataxia (FRDA), which is defined by a German neurologist in 1863, is an inherited neurodegenerative disorder that affects the nervous system and the heart with symptoms of deep sensory loss, muscle weakness, kyphoscoliosis, dysarthria, heart disease and difficulty in speech [25]. FRDA is the most common hereditary ataxia with 1-2 cases in every 50,000 white people. To comprehensively study FRDA, the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) has assembled a body of expertise to adopt a translational research strategy for FRDA [118, 119].

EFACTS has been devoted to collecting and analyzing FRDA patient baseline data since 2010. Up to now, EFACTS has collected more than one thousand patients' baseline data from nearly twenty study sites in nine European counties, but the coverage is still far from enough. According to the morbidity rate, the potential FRDA patients are huge. Due to the limitations of physical, psychological or economic reasons, many patients may not be able to go to the study sites for the FRDA medical assessment.

Note that most baseline data are collected through interviews, questionnaires, observations and coordinated tests at the study sites without using any medical instruments. Here, the detailed test methods and rating rules have been provided by EFACTS. Therefore, we make a reasonable assumption that patients who are not able to go to the study sites can be assessed at home and let their families (or themselves) act as examiners. The examiners can be relied upon in providing certain reliable ratings in the portion of test-items during long-term observation and care.

Intuitively, similar FRDA patients exhibit similar symptoms. The unfilled parts in test-items are regarded as missing values. The prediction of missing values can be considered as a typical RS problem, where the patients correspond to the users, and FRDA test-items correspond to the items. Inspired by the idea of CF, the missing values can be predicted by utilizing the certain values provided by the examiners and the data collected by EFACTS. Therefore, the application of our proposed MCF algorithm in FRDA provides an alternative way to assist patient baseline data collection. In this way, many more patient samples can be exploited in clinical trials, which will provide better bases for FRDA research [179, 180].

6.3.2 Data pre-processing

In this chapter, the scale for the assessment and rating of ataxia (SARA) dataset has been selected from the database provided by the EFACTS. SARA is a new clinical scale that is utilized to evaluate the treatment effectiveness and severity of different types of cerebellar ataxia such as Friedreich's, spinocerebellar and sporadic ataxia [175]. As shown in Tab. 6.1, there are 12 test-items in 8 categories to assess a range of different impairments. The categories are gait, stance, sitting, speech disturbance, finger chase, nose-finger test, fast alternating hand movements and heel-shin slide. SARA has an accumulative score ranging from 0 to 40 where 0 means no ataxia and 40 means most severe ataxia.

The number of patients in the SARA dataset is continuously updated. Up to now, the SARA dataset has included the baseline data of 1029 patients. The user-item matrix R is a 1029×12 matrix, where each row denotes an FRDA patient, and each column denotes a test-item. As shown in Tab. 6.1, the rating intervals are different. Therefore, we normalize the rating values into the 0-1 range based on

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{6.18}$$

where x' is the normalized value, $x_{\min}$ and $x_{\max}$ are, respectively, the minimum and maximum values of x which give the range of x.

Table 6.2 Experimental results under different densities

Metrics	Methods	density of matrix						
		90%	80%	70%	60%	50%	40%	30%
MAE	MCF	0.1132	0.1166	0.1302	0.1348	0.1449	0.1535	0.1684
	UBCF	0.1198	0.1231	0.1356	0.1471	0.1543	0.1627	0.1740
	(Improve)	(5.51%)	(5.28%)	(3.98%)	(8.36%)	(6.09%)	(5.65%)	(3.22%)
	IBCF	0.1157	0.1183	0.1314	0.1391	0.1486	0.1634	0.1841
	(Improve)	(2.16%)	(1.44%)	(0.91%)	(3.09%)	(2.49%)	(6.06%)	(8.53%)
RMSE	MCF	0.1583	0.1592	0.1722	0.1769	0.1907	0.1977	0.2265
	UBCF	0.1634	0.1643	0.1811	0.1945	0.2065	0.2209	0.2389
	(Improve)	(3.12%)	(3.10%)	(4.91%)	(9.05%)	(7.65%)	(10.50%)	(5.19%)
	IBCF	0.1601	0.1627	0.1802	0.1845	0.2012	0.2156	0.2431
	(Improve)	(1.12%)	(2.15%)	(4.44%)	(4.12%)	(5.22%)	(8.30%)	(6.83%)

6.3.3 Experiment setting

In our simulation, 1029 patients have been divided into the training set (70%), validation set (15%) and testing set (15%). The training set and validation set are used for selecting the

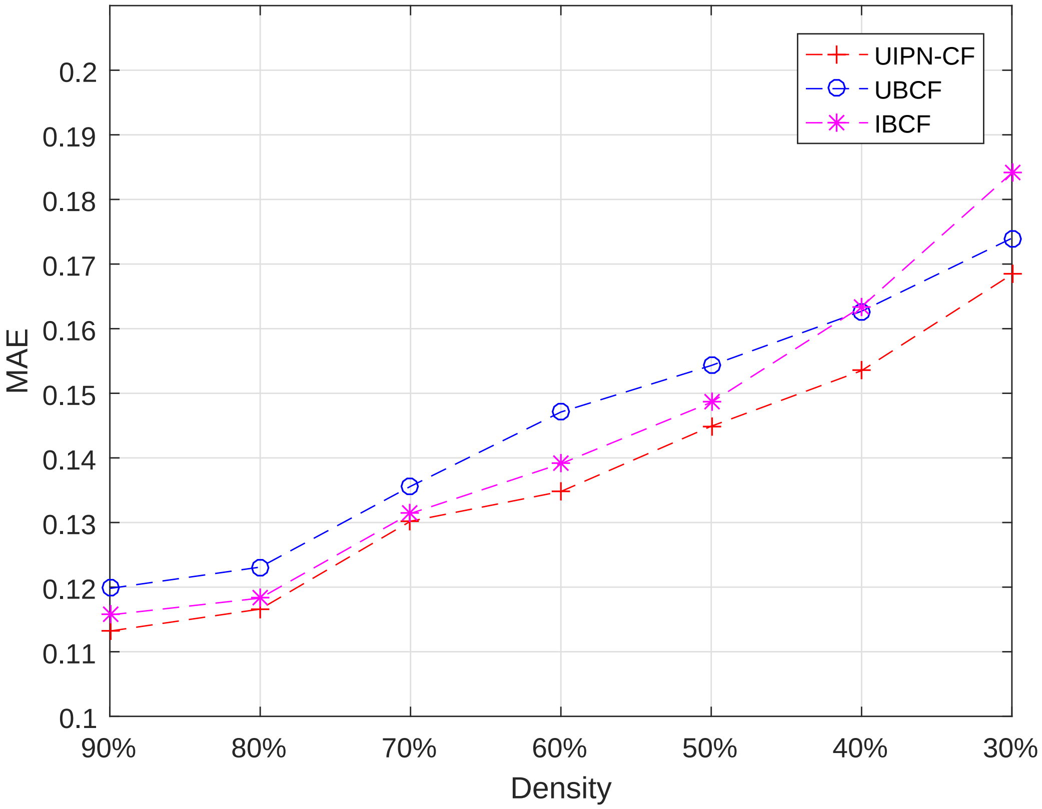


Fig. 6.1 MAE metric under different densities.

parameter vector ω to minimize the error. The data in the testing set is regarded as patients who cannot take the tests in any study site. In this case, the patients in the testing set only provide ratings on the portion of test-items. The proposed MCF method is utilized to predict the rating values on patients' unfilled parts.

To evaluate the prediction quality of the algorithm, the mean absolute error (MAE) and the root mean square error (RMSE) used in our experiments are given as follows:

$$\mathrm{MAE} = \frac{1}{N} \sum_{u \in U_d} \sum_{i \in I_d} |r_{u,i} - \hat{r}_{u,i}| \tag{6.19}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{u \in U_d} \sum_{i \in I_d} (r_{u,i} - \hat{r}_{u,i})^2} \tag{6.20}$$

where N represents the total number of predicted values in the testing set; U_d and I_d represent the user set and test-item set in the testing set, respectively; $r_{u,i}$ is the true rating value in the testing set; and $\hat{r}_{u,i}$ is the predicted value provided by our proposed CF algorithm.

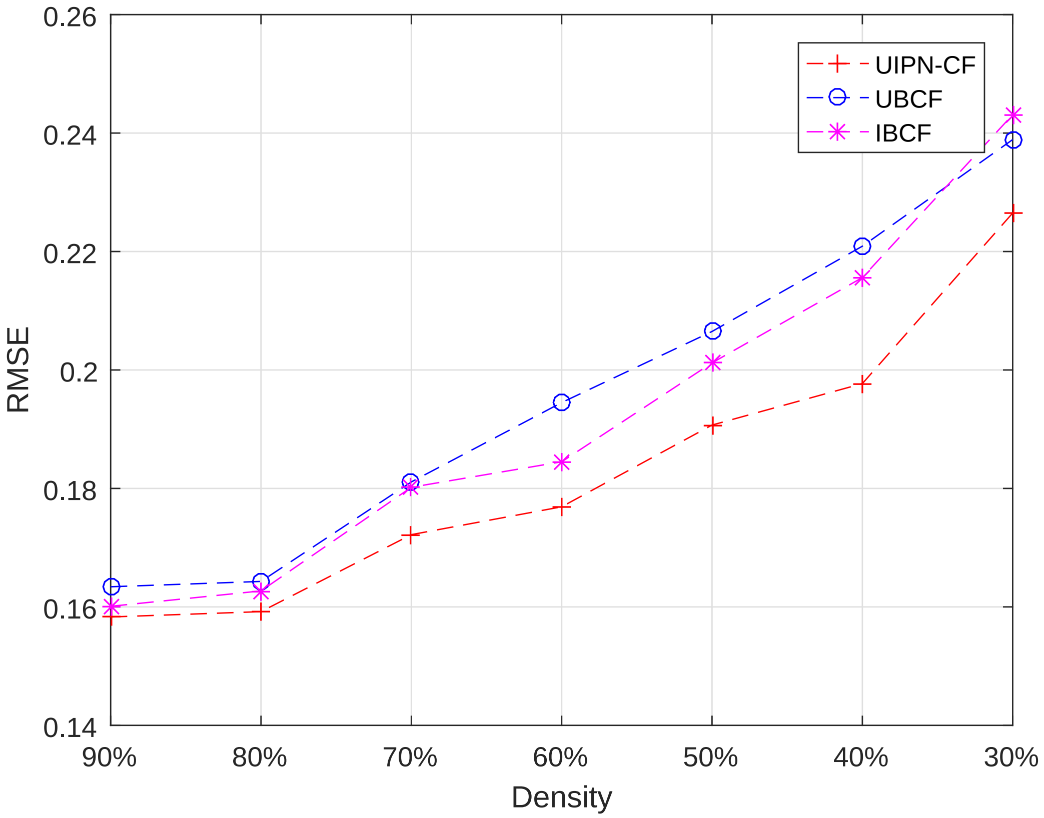


Fig. 6.2 RMSE metric under different densities.

The parameters of the PSO algorithm in the simulation are given as follows. The dimension of each particle is 3; the population of the swarm is 20; the maximum iteration number is set to be 1000; and the search space of α, β, λ is in the interval of [0, 1].

6.3.4 Results and discussion

In this chapter, we implement our approach on the SARA dataset provided by EFACTS to evaluate the effectiveness of our algorithm by employing the density of the testing set from 90% to 30% with a step size of 10%. We repeat each experiment 100 times to avoid random influence, and the average values of MAE and RMSE have been recorded. To demonstrate the superiority of our proposed MCF algorithm, we make a comparison of the UBCF and IBCF methods with our proposed MCF method on the MAE and RMSE metrics.

Experiment results of the UBCF, IBCF and MCF methods are shown in Figs. 6.1 and 6.2. The vertical coordinate denotes the values of MAE or RMSE, and the horizontal coordinate represents the different densities of the user-item matrix. The MAE and RMSE of different CF-based algorithms are displayed in Tab. 6.2. The results indicate that our MCF algorithm has better MAE and RMSE values than the UBCF and the IBCF algorithms under different densities. To sum up, the proposed MCF algorithm has shown satisfactory prediction accuracy in the FRDA baseline data.

6.3.5 Complexity analysis

Classic UBCF (IBCF) algorithm involves the calculation of user-user (item-item) similarity matrix in an offline way, which is computationally expensive. For both UBCF and IBCF, the offline computation of similarity matrices is very time-consuming. The offline time complexity of UBCF and IBCF is $O(m^2 \cdot n)$ and $O(m \cdot n^2)$, respectively, where m denotes the number of users and n denotes the number of items. In MCF, the offline computation is even more expensive because our proposed algorithm needs to compute both user-user and item-item similarity matrices. The offline time complexity of the MCF is $O(m^2 \cdot n + m \cdot n^2)$.

In the online phase, the time complexity of MCF method in the prediction part is the same as that of UBCF/IBCF method, which is $O(k)$ where k is the size of the neighbors of the target user and item. To sum up, our proposed method improves the prediction accuracy at the expense of extra offline computation.

6.4 Conclusion

In this chapter, an MCF algorithm has been presented and successfully employed to deal with the data prediction problem of FRDA patient baseline data. The proposed MCF algorithm has combined the merits of both the UNCF method and the IBCF method, and has been shown to outperform the UNCF method alone or the IBCF method alone. It should be pointed out that the positively and the negatively correlated neighbors have also been taken into account in the MCF algorithm with hope to improve prediction accuracy. In the developed MCF algorithm, the weighting parameters have been employed to balance the usage of 1) the UBCF method and the IBCF method; and 2) the positively and the negatively correlated neighbors. The PSO algorithm has been applied to automate the selection of locally optimized weights so as to guarantee the prediction accuracy. The MCF algorithm has been applied to deal with a real-world disease, the FRDA, to justify its application potential. Experiment results have shown that our proposed approach greatly improves the prediction accuracy with better performance than either the UBCF algorithm or the IBCF algorithm. In the future, we aim to investigate the application and improvement of different deep neural network models in RSs.

Chapter 7

Conclusions and Future Research

In this chapter, we first summarize our work in this thesis and then point out several possible directions in further research.

7.1 Concluding remarks

For FRDA (Friedreich's ataxia), insufficient clinical samples have led to uncertainty about the effectiveness of existing clinical studies, even if many existing clinical trials have shown positive effects. Traditional collection methods are not only slow, but also expensive. In addition to taking into account the characteristics of this disease, the poor health of many patients also makes data collection face unprecedented difficulties.

In this thesis, we propose a novel FRDA patient data collection strategy which is inspired by the popularity of the nowadays "Recommendation System (RS)" and a series of advanced collaborative filtering (CF) based methods. Specifically, to overcome physical/psychological difficulties in recruiting new patients and collecting follow-up assessment data, a novel data collection strategy for the FRDA baseline data by using the CF approaches is presented (Chapter 3); a novel hybrid method combining the merits of model- and memory-based CF methods is proposed for addressing the situations of patients that neighbors and patients do not have neighbors (Chapter 4); a weighted naive Bayes based CF (WNBCF) algorithm is proposed to assist the FRDA baseline data collection under the cold-start condition by taking into account the patient side-information (Chapter 5); a modified collaborative filtering (MCF) algorithm with improved performance is developed, which combines the individual merits of both the user-based CF method and the item-based CF method, where both the positively and negatively correlated neighbors are taken into account (Chapter 6). Next, we summarise the research results presented in each of these chapters.

Chapter 3 has presented a data collection strategy for the FRDA baseline data by using the CF approaches. This strategy adopts the idea of the nowadays popular RS which is based on the fact that similar patients have similar symptoms on each test-item. Note that the main advantages of using RS is that not all ratings are required for all testing-items, which means that the patients only need to provide some sure ratings at home instead of going to the EFACTS' study sites. The unfilled parts will be predicted by using one of the most successful RS techniques, memory-based CF methods. It is shown that the CF approaches are capable of predicting baseline data based on the similarity in test-items of the patients, where the prediction accuracy is evaluated based on three rating scales selected from the EFACTS database. Experimental results demonstrate the validity and efficiency of the proposed strategy. The limitation in Chapter 3 is that based on what may happen in reality, basic memory-based CF methods are able to overcome the problems of sparsity, cold-start, and low prediction accuracy.

In Chapter 4, a hybrid model- and memory-based CF algorithm has been proposed in order to improve the prediction accuracy of unfilled values when dealing with the situations of patients who have neighbors and do not have neighbors. If patients have similar neighbors, the enhanced memory-based CF method is adopted with an improved similarity measure, where both the patient rating habits and the number of co-rated test-items are taken into account from a unified viewpoint. If patients do not have any similar neighbors, the model-based CF is harnessed to find similar neighbors with similar FRDA symptoms by clustering this patient into the class based on his/her attributes. To evaluate the advantages of the proposed algorithm, the Scale for the Assessment and Rating of Ataxia (SARA) is selected from the EFACTS database. Experimental results demonstrate that our proposed hybrid CF approach is superior to other conventional approaches. The limitation in Chapter 4 is that we must assume that new patient must provide at least one rating value. If not, our proposed method cannot handle this kind of situation.

In Chapter 5, a WNBCF algorithm has been proposed to tackle the cold-start problem during the FRDA patient-baselinedata collection. In practical applications, the patient side-information (attributes) with different weights according to their significance is adopted to discover the relationship with ratings (classes) on test-items. By combining MI with PSO algorithm, a novel computational framework is established to fine tune the weights of attributes. The superiorities of the proposed WNBCF algorithm is demonstrated over some conventional algorithms in real-world FRDA datasets from the database provided by EFACTS. The limitation in Chapter 5 is that the conditional probability is normally very small because the equation for calculating the probability is continuous multiplication. This can lead to unstable predictions.

In Chapter 6, an MCF has been developed by combining the merits of uer-based CF and item-based CF methods. Through the utilization of the information from both the positively and negatively correlated neighbors, the proposed algorithm is capable of predicting the missing values in multi-aspects with satisfactory accuracy. The PSO algorithm is utilized to optimize the weights in the MCF algorithm so as to achieve a) an adequate tradeoff between the user-based and the item-based similarity measures; and b) a proper balance between the positively and negatively correlated neighbors. The effectiveness of the proposed MCF algorithm is confirmed by extensive experiments and, furthermore, it is shown that our algorithm outperforms some conventional approaches. The limitation in Chapter 6 is that the fitness function is used to minimize the overall mean absolute error, however, for each patient, these weight parameters may not be necessarily good.

7.2 Recommendations for future research

In this thesis, we have presented a novel FRDA baseline data collection strategy that are capable to assist EFACTS in overcoming existing obstacles and improving collection speed. And we have also proposed several enhanced recommendation algorithms to improve the prediction accuracy of missing values during data collection. Next, the work may be further researched in a number of ways:

- Interpretability: The explainable recommendation is very important in today's applications [1, 195], especially in medical and healthcare fields. If the designed RS can come up with easy-to-understand explanation, the patients would know why the results are recommended to them and, therefore, both the effectiveness and the persuasiveness of the recommended results are greatly improved. Interpretation of the prediction results can also assist doctors/patients to make the accurate decision about whether to accept predicted results or make adjustments. As such, one of our future research directions is to increase the interpretability of our designed RS.

- Algorithm development: The algorithms we are improving now are mainly focused on memory-based CF methods. Some model-based CF methods have been shown to work very well in RSs, for example, matrix factorization [170, 67], deep learning [126, 51, 50, 166] and so on. In particular, the recommendation algorithms based on deep learning models have ushered in explosive growth in recent years [39, 80, 162, 190]. In the future, we aim to focus on developing appropriate deep learning models to help improve the prediction accuracy by extract useful patient characteristics.

- Other application scenarios: As we discussed in Section 2.6, RS techniques can be applied to lots of different scenarios. For FRDA patients, our future work is to help design an RS to provide personalized recommendations to effectively support their daily activities, rehabilitation programmes and so on.

- Other disease applications: RS technology has been widely used in the medical and healthcare fields. The proposed strategy in Chapter 3 is also applicable to other diseases baseline data collection, especially in rating scales, for example, some famous scales like International Cooperative Ataxia Rating Scale, Hamilton Rating Scale for Depression and National Institutes of Health Stroke Scale, etc. How to extend our proposed strategy and algorithms to other application scenarios will be the next topic in our future research.

- Privacy preserving: Privacy preserving is very important, and it is also one of the key directions of our next research [158, 12, 110, 135]. For example, the collected patient data not only contains a large amount of personal information of users, but also includes disease conditions, medication, and family information. Leakage of these data will bring potential threats to patients and their families. Some simple encryption methods sometimes do not guarantee data security, so our following work will design a robust RS to defend against different types of attacks.

References

[1] B. Abdollahi, O. Nasraoui, Using explainability for constrained matrix factorization, In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*, Como, Italy, pp. 79–83, Aug. 27–31, 2017

[2] P. Achananuparp, and I. Weber, Extracting food substitutes from food diary via distributional similarity, In: *Proceedings of the ACM workshop on engendering health with recommender systems*, Boston, MA, USA, Sep. 15–19, 2016.

[3] I. Adaji, K. Oyibo, and J. Vassileva, Shopping value and its influence on healthy shopping habits in e-commerce, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 36–39, Oct. 2–7, 2018.

[4] G. Adomavicius, and A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[5] G. Agapito, M. Simeoni, B. Calabrese, et al, DIETOS: a recommender system for health pro

ling and diet management in chronic diseases, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.32–35, Aug. 27–31, 2017.

[6] C. C. Aggarwal, Ensemble-based and hybrid recommender systems, *in Recommender Systems*, Boston: Springer, 2016, pp. 199–224.

[7] E. Agu, and M. Claypool, Cypress: A cyber-physical recommender system to discover smartphone exergame enjoyment, In: *Proceedings of the ACM workshop on engendering health with recommender systems*, Boston, MA, USA, Sep. 15–19, 2016.

[8] S. Akkoyunlu, C. Manfredotti, A. Cornuéjols, et al, Investigating substitutability of food items in consumption data, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.27–31, Aug. 27–31, 2017.

[9] S. Akkoyunlu, C. Manfredotti, A. Cornuéjols, et al, Exploring eating behaviours modelling for user clustering, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 46–51, Oct. 2–7, 2018.

[10] H. Alcaraz-Herrera, and I. Palomares, Evolutionary approach for ’healthy bundle’ wellbeing recommendations, In: *Proceedings of the 4th International Workshop on Health Recommender Systems co-located with the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 19–23, Sept. 16–20, 2019.

[11] F. Alvarez, M. Popa, V. Solachidis, et al, Behavior analysis through multimodal sensing for care of Parkinson’s and Alzheimer’s patients, *IEEE Multimedia*, vol. 25, no. 1, pp. 14–25, 2018.

[12] S. Badsha, X. Yi, and I. Khalil, A practical privacy-preserving recommender system, *Data Science and Engineering*, vol. 1, no. 3, pp. 161–177, 2016.

[13] C. Birtolo, and D. Ronca, Advances in clustering collaborative filtering by means of fuzzy C-means and trust, *Expert Systems with Applications*, vol. 40, no. 17, pp. 6997–7009, 2013.

[14] J. Bennett, and S. Lanning, The netflix prize, In: *Proceedings of KDD cup and workshop at the the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, USA, Aug. 12–15, 2007.

[15] J. Berndsen, A. Lawlor, and B. Smyth, Running with Recommendation, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located*

with the 11th ACM Conference on Recommender Systems, Como, Italy, pp.18–21, Aug. 27–31, 2017.

[16] J. Bobadilla, A. Hernondo, F. Ortega, and A. Gutiérrez, Collaborative filtering based on significances, *Information Sciences*, vol. 185, no. 1, pp. 1–17, 2012.

[17] J. Bobadilla, F. Ortega, A. Hernondo, and A. Gutiérrez, Recommender systems survey, *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[18] L. Boratto, S. Carta, W. Iguider, F. Mulas, and P. Pilloni, Predicting workout quality to help coaches support sportspeople, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 8–12, Oct. 2–7, 2018.

[19] J. S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, *arXiv preprint arXiv:1301.7363*, 2013.

[20] J. S. Brerse, D. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, *in Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, USA, pp. 43–52, July. 24–26, 1998.

[21] M. Bukowski, A. C. Valdez, M. Ziefle, et al, Hybrid collaboration recommendation from bibliometric data, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.36–38, Aug. 27–31, 2017.

[22] R. Burke, Hybrid recommender systems: Survey and experiments, *User modeling and user-adapted interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[23] K. Bürk, U. Mälzig, S. Wolf, et al, Comparison of three clinical rating scales in Friedreich ataxia (FRDA), *Movement Disorders*, vol. 24, no. 12, pp. 1779–1784, 2009.

[24] J. Caldeira, R. S. Oliveira, L. Marinho, et al, Healthy menus recommendation: optimizing the use of the pantry, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 2–7, Oct. 2–7, 2018.

[25] V. Campuzano, L. Montermini, M. D. Moltò, et al, Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion, *Science*, vol. 271, no. 5254, pp. 1423–1427, Mar. 1996.

[26] X. Chen, X. Liu, Z. Huang, and H. Sun, Regionknn: A scalable hybrid collaborative filtering algorithm for personalized web service recommendation, in *Proceedings of the 2010 IEEE International Conference on Web Services*, Miami, USA, pp. 9–16, July 5–10, 2010.

[27] P. Covington, J. Adams, and E. Sargin, Deep neural networks for youtube recommendations, in *Proceedings of the 10th ACM Conference on Recommender Systems*, Boston, USA, pp. 191–198, September 15–19, 2016.

[28] P. Cremonesi, Y. Koren, and R. Turrin, Performance of recommender algorithms on top-n recommendation tasks, In: *Proceedings of the 14th ACM conference on Recommender systems*, Barcelona, Spain, pp. 39–46, Sept. 26–30, 2010.

[29] J. Davidson, B. Liebald, J. Liu, et al, The YouTube video recommendation system, In: *Proceedings of the 4th ACM conference on Recommender systems*, Barcelona, Spain, pp. 293–296, Spet. 26–30, 2010.

[30] M. Deshpande, G. Karypis, Item-based top-n recommendation algorithms, *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 143–177, Jan. 2004.

[31] N. A. Di Prospero, A. Baker, N. Jeffries, and K. H. Fischbeck, Neurological effects of high-dose idebenone in patients with Friedreich's ataxia: a randomised, placebo-controlled trial, *The Lancet Neurology*, vol. 6, no. 10, pp. 878–886, Oct. 2007.

[32] L. Duan, W. N. Street, E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, *Enterprise Information Systems*, vol. 5, no. 2, pp. 169–181, 2011.

[33] J. D. Ekstrand, M. D. Ekstrand, First do no harm: Considering and minimizing harm in recommender systems designed for engendering health, In: *Proceedings of the ACM workshop on engendering health with recommender systems*, Boston, MA, USA, Sep. 15–19, 2016.

[34] E. Ezin, E. Kim, and I. Palomares, 'Fitness that fits': A prototype model for workout video recommendation, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 40–45, Oct. 2–7, 2018.

[35] J. M. Fernández, M. Mamei, S. Mariani, et al, Towards argumentation-based recommendations for personalised patient empowerment, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.1–5, Aug. 27–31, 2017.

[36] I. Fernández-Tobías, I. Cantador, M. Kaminskas, and F. Ricci, Cross-domain recommender systems: A survey of the state of the art, In: *Proceedings of the Spanish conference on information retrieval 2012*, Valencia, Spain, pp. 1–12, Jun. 18–19, 2012.

[37] E. Frank, M. Hall, and B. Pfahringer, Locally weighted naive Bayes, In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, Acapulco, Mexico, pp. 249–256, Aug. 7–10, 2003.

[38] N. Friedreich, Ueber degenerative Atrophie der spinalen Hinterstränge, *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, vol. 26, no. 3–4, pp. 391–419, May. 1863.

[39] M. Fu, H. Qu, Z. Yi, et al, A novel deep learning-based collaborative filtering model for recommendation system, *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1084–1096, 2018.

[40] F. S. Gohari, F. S. Aliee, and H. Haghighi, A new confidence-based recommendation approach: combining trust and certainty, *Information Sciences*, vol. 422, pp. 21–50, 2018.

[41] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, vol. 35, no. 12, pp. 61–71, 1992.

[42] F. Gräßer, S. Beckert, D. Küster, et al, Neighborhood-based Collaborative Filtering for Therapy Decision Support, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.22–26, Aug. 27–31, 2017.

[43] A. Gunawardana, and C. Meek, Tied boltzmann machines for cold start recommendations, in *Proceedings of the 2008 ACM conference on Recommender systems*, Lausanne, Switzerland, pp. 19–26, October 23–25, 2008.

[44] A. Gunawardana, and G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, *Journal of Machine Learning Research*, vol. 10, no. 12, pp. 2935–2962, 2009.

[45] F. Gutierrez, B. Cardoso, and K. Verbert, PHARA: a Personal Health Augmented Reality Assistant to Support Decision-Making at Grocery Stores, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.10–13, Aug. 27–31, 2017.

[46] H. M. Habeeb, A. Al-Azawei, and N. Al-A'araji, Developing a healthcare recommender system using an enhanced symptoms-based collaborative filtering technique, *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 3, pp. 920–926, 2019.

[47] Q. Han, Í. M. R. de Troya, M. Ji, M. Gaur, and L. Zejnilovic, A collaborative filtering recommender system in primary care: Towards a trusting patient-doctor relationship, In: *Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, USA, pp. 377–379, Jun. 4–7, 2018.

[48] A. E. Harding, Friedreich's ataxia: a clinical and genetic study of 90 families with an analysis of early diagnostic criteria and intrafamilial clustering of clinical features, *Brain*, vol. 104, no. 3, pp. 589–620, Sep. 1981.

[49] R. He, and J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, In: *Proceedings of the 25th international conference on world wide web*, Quebec, Canada, pp. 507–517, Apr. 11–15, 2016.

[50] X. He and T. S. Chua, Neural factorization machines for sparse predictive analytics, In: *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, Tokyo, Japan, pp. 355–364, Aug. 07–11, 2017.

[51] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. S. Chua, Neural collaborative filtering, in *Proceedings of the 26th International Conference on World Wide Web*, Perth, Australia, pp. 173–182, April 03–07, 2017.

[52] J. L. Herlocker, Understanding and Improving Automated Collaborative Filtering Systems, PhD thesis, University of Minnesota, 2000.

[53] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, An algorithmic framework for performing collaborative filtering, *in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, USA, pp. 230–237, Aug. 15–19, 1999.

[54] K. Herrmanny, and A. Dogangün, The Impact of Prediction Uncertainty in Recommendations for Health-Related Behavior, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.14–17, Aug. 27–31, 2017.

[55] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, Session-based recommendations with recurrent neural networks, *arXiv preprint arXiv:1511.06939*, 2015.

[56] S. Hors-Fraile, F. J. N. Benjumea, L. C. Hernández, F. O. Ruiz, and L. Fernandez-Luque, Design of two combined health recommender systems for tailoring messages in a smoking cessation app, In: *Proceedings of the ACM workshop on engendering health with recommender systems*, Boston, MA, USA, Sep. 15–19, 2016.

[57] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, In: *Proceedings of the 2008 8th IEEE International Conference on Data Mining*, Pisa, Italy, pp. 263–272, Dec. 15–19, 2008.

[58] A. S. Hussein, W. M. Omar, X. Li, M. Ati, Efficient chronic disease diagnosis prediction and recommendation system, In: *Proceedings of the 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences* , Langkawi, Malaysia, pp. 209–214, Dec. 17-–19, 2012.

[59] A. S. Hussein, W. M. Omar, X. Li, M. Amer Hatem, Smart collaboration framework for managing chronic disease using recommender system, *Health Systems*, vol. 3, no. 1, pp. 12–17, 2014.

[60] F. Jabeen, M. Maqsood, M. A. Ghazanfar, et al, An IoT based efficient hybrid recommender system for cardiovascular disease, *Enterprise Information Systems*, vol. 12, no. 5, pp. 1263–1276, 2019.

[61] A. Jameson, A Tool That Supports the Psychologically Based Design of Health-Related Interventions, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.39–42, Aug. 27–31, 2017.

[62] S. Jamshidi, A. Torkamani, J. Mellen, A hybrid health journey recommender system using electronic medical records, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 57–62, Oct. 2–7, 2018.

[63] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender systems: An introduction, Cambridge, UK: Cambridge University Press, 2010.

[64] K. Järvelin, and J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[65] X. Ji, S. A. Chun, and J. Geller, A collaborative filtering approach to assess individual condition risk based on patients' social network data, In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, Newport Beach, USA, pp. 639–640, Sept. 20–23, 2014.

[66] X. Jia, X. Li, K. Li, V. Gopalakrishnan, G. Xun, and A. Zhang, Collaborative restricted Boltzmann machine for social event recommendation, in *Proceedings of the 2016*

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, San Francisco, USA, pp. 402–405, August 18–21, 2016.

[67] J. Jiang, W. Li, A. Dong, et al, A fast deep autoencoder for high-dimensional and sparse matrices in recommender systems, *Neurocomputing*, 2020, DOI: 10.1016/j.neucom.2020.06.109.

[68] Y. Jiang, J. Liu, M. Tang, and X. Liu, An effective web service recommendation method based on personalized collaborative filtering, In: *Proceedings of the 2011 IEEE International Conference on Web Services*, Washington, USA, pp. 211–218, July 4–9, 2011.

[69] C. Kaleli, An entropy-based neighbor selection approach for collaborative filtering, *Knowledge-Based Systems* vol. 56, pp. 273–280, 2014.

[70] S. Kafle, P. Pan, A. Torkamani, S. Halley, J. Powers, and H. Kardes, Personalized symptom checker using medical claims, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 13–17, Oct. 2–7, 2018.

[71] H. Kaur, N. Kumar, S. Batra, An efficient multi-party scheme for privacy preserving collaborative filtering for healthcare recommender system, *Future Generation Computer Systems*, vol. 86, pp. 297–307, 2018.

[72] N. R. Kermany, and S. H. Alizadeh, A hybrid multi-criteria recommender system using ontology and neuro-fuzzy techniques, *Electronic Commerce Research and Applications*, vol. 21, pp. 50–64, 2017.

[73] H. N. Kim, A. Alkhaldi, A. E. Saddik, and G. S. Jo, Collaborative user modeling with user-generated tags for social recommender systems, *Expert Systems with Applications*, vol. 38, no. 7, pp. 8488–8496, 2011.

[74] D. P. Kingma, M. Welling, Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114*, 2013.

[75] M. A. Khan, E. Rushe, B. Smyth, and D. Coyle, Personalized, health-aware recipe recommendation: an ensemble topic modeling based approach, In: *Proceedings of the 4th International Workshop on Health Recommender Systems co-located with the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 4–10, Sept. 16–20, 2019.

[76] Y. Koren, Collaborative filtering with temporal dynamics, In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, pp. 447–456, Jun. 28—Jul. 1, 2009.

[77] H. Koohi, and K. Kiani, User based collaborative filtering using fuzzy C-means, *Measurement*, vol. 91, pp. 134–139, 2016.

[78] H.-J. Kwon, T.-H. Lee, J.-H. Kim, and K.-S. Hong, Improving Prediction accuracy using entropy weighting in collaborative filtering, in *IEEE 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, Brisbane, Australia, 2009, pp. 40–45.

[79] Q. Le, and A. Smola, Direct optimization of ranking measures, *arXiv preprint arXiv:0704.3359*, 2007.

[80] J. Lee, S. Abu-El-Haija, B. Varadarajan, et al, Collaborative deep metric learning for video understanding, In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, London, UK, pp. 481–490, Aug. 19–23, 2018.

[81] X. Lei, Z. Fang, and L. Guo, Predicting circRNA-disease associations based on improved collaboration filtering recommendation system with multiple data, *Frontiers in Genetics*, vol. 10, art. no. 897, 11 pages, 2019.

[82] N. Leipold, M. Madenach, H. Schäfer, et al, Nutrilize a personalized nutrition recommender system: an enable study, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 24–29, Oct. 2–7, 2018.

[83] L. Li, W. Chu, J. Langford, and R. E. Schapire, A contextual-bandit approach to personalized news article recommendation, In: *Proceedings of the 19th international conference on World wide web*, Raleigh, USA, pp. 661–670, Apr. 26-–30, 2010.

[84] X. Li, and J. She, Collaborative variational autoencoder for recommender systems, *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, Halifax, Canada, pp. 305–314, Aug. 13–17, 2017.

[85] Y. Li, S. Wang, Q. Pan, H. Peng, T. Yang, and E. Cambria, Learning binary codes with neural collaborative filtering for efficient recommendation systems, *Knowledge-Based Systems*, vol. 172, pp. 64–75, 2019.

[86] J. Lian, F. Zhang, X. Xie, and G. Sun, CCCFNet: a content-boosted collaborative filtering neural network for cross domain recommender systems, In: *Proceedings of the 26th international conference on World Wide Web companion*, Perth, Australia, pp. 817—818, Apr. 03—07, 2017.

[87] J. Lin, and J. Yu, Weighted naive Bayes classification algorithm based on particle swarm optimization, In: *Proceedings of the 2011 IEEE 3rd International Conference on Communication Software and Networkse*, Xi'an, China, pp. 444–447, May 27–29, 2011.

[88] G. D. Linden, J. A. Jacobi, and E. A. Benson, Collaborative recommendations using item-to-item similarity mappings, U. S. Patent 6266649B1, Sept. 18, 1998.

[89] J. Liu, and C. Wu, Deep Learning Based Recommendation: A Survey, In: *Proceedings of the International Conference on Information Science and Applications 2017*, Macau, China, pp. 451–458, Mar. 20–23, 2017.

[90] Q. Liu, Y. Zeng, R. Mokhosi, and H. Zhang, STAMP: short-term attention/memory priority model for session-based recommendation, In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, UK, Aug. 19–23, 2018.

[91] W. Liu, Z. Wang, X. Liu, N. Zeng, and D. Bell, A novel particle swarm optimization approach for patient clustering from emergency departments, *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 4, pp. 632–644, 2018.

[92] W. Liu, Z. Wang, Y. Yuan, N. Zeng, K. Hone, and X. Liu, A novel sigmoid-function-based adaptive weighted particle swarm optimizer, *IEEE Transactions on Cybernetics*, vol. 23, no. 4, pp. 632–644, 2019.

[93] F. Luo, G. Ranzi, X. Wang, and Z. Y. Dong, Social information filtering-based electricity retail plan recommender system for smart grid end users, *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 95–104, 2017.

[94] X. Luo, M. Zhou, S. Li, D. Wu, Z. Liu, and M.-S. Shang, Algorithms of unconstrained non-negative latent factor analysis for recommender systems, *IEEE Transactions on Big Data*, 2019, DOI: 10.1109/TBDATA.2019.2916868.

[95] X. Luo, M. Zhou, S. Li, and M.-S. Shang, An inherently nonnegative latent factor model for high-dimensional and sparse matrices from industrial applications, *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 2011–2022, 2017.

[96] X. Luo, M. Zhou, S. Li, Z. You, Y. Xia and Q. Zhu, A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 579–592, 2016.

[97] X. Luo, M. Zhou, Y. Xia, Q. Zhu, “An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems,” *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1273–1284, 2014.

[98] X. Luo, M. Zhou, Y. Xia, Q. Zhu, A. C. Ammari and A. Alabdulwahab, Generating highly accurate predictions for missing QoS data via aggregating nonnegative latent factor models, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 524–537, 2016.

[99] X. Luo, H. Wu, H. Yuan, M. C. Zhou, Temporal pattern-aware qos prediction via biased non-negative latent factorization of tensors, *IEEE Transactions on Cybernetics*, vol. 50, no. 5, pp. 1798–1809, 2019.

[100] X. Luo, M. C. Zhou, S. Li, L. Hu, and M. Shang, Non-negativity constrained missing data estimation for high-dimensional and sparse matrices from industrial applications, *IEEE Transactions on Cybernetics*, vol. 50, no. 5, pp. 1844–1855, 2019.

[101] X . Luo, D. Wang, M. C. Zhou, and H. Yuan, Latent factor-based recommenders relying on extended stochastic gradient descent algorithms, *IEEE Transactions on System Man Cybernetics: Systems*, 2018, DOI: 10.1109/TSMC.2018.2884191.

[102] X. Luo, Z. Liu, S. Li, M. Shang, Z. Wang, A fast non-negative latent factor model based on generalized momentum method, *IEEE Transactions on System Man Cybernetics: Systems*, 2018, DOI: 10.1109/TSMC.2018.2875452.

[103] X. Luo, Z. Wang, M. Shang, An instance-frequency-weighted regularization scheme for non-negative latent factor analysis on high dimensional and sparse data, *IEEE Transactions on System Man Cybernetics: Systems*, 2019, DOI: 10.1109/TSMC.2019.2930525.

[104] X. Luo, Y. Yuan, M. C. Zhou, Z. Liu, and M. Shang, Non-negative latent factor model based on β-divergence for recommender systems, *IEEE Transactions on System Man Cybernetics: Systems*, 2019, DOI: 10.1109/TSMC.2019.2931468.

[105] X. Luo, M. C. Zhou, S. Li, Y. Xia, Z. H. You, Q. Zhu, and H. Leung, Incorporation of efficient second-order solvers into latent factor models for accurate prediction of missing qos data, *IEEE Transactions on Cybernetics*, vol. 48, no. 4, pp. 1216–1228, 2017.

[106] X. Luo, M. C. Zhou, Z. Wang, Y. Xia, and Q. Zhu, An effective qos estimating scheme via alternating direction method-based matrix factorization, *IEEE Transactions on Services Computing*, vol. 12, no. 4, pp. 503–518, 2019.

[107] D. R. Lynch, and E. Kichula, Challenges ahead for trials in Friedreich's ataxia, *The Lancet Neurology*, vol. 15, no. 13, pp. 1300–1301, Dec. 2016.

[108] D. R. Lynch, S. L. Perlman, and T. Meier, A phase 3, double-blind, placebo-controlled trial of idebenone in Friedreich ataxia, *Archives of Neurology*, vol. 67, no. 8, pp. 941–947, Aug. 2010.

[109] D. R. Lynch, S. M. Willi, R. B. Wilson, et al, A0001 in Friedreich ataxia: biochemical characterization and effects in a clinical trial, *Movement Disorders*, vol. 27, no. 8, pp. 1026–1033, Jul. 2012.

[110] X. Ma, H. Li, J. Ma, et al, APPLET: A privacy-preserving framework for location-aware recommender system, *Science China Information Sciences*, vol. 60. no. 9, art. no. 092101, 16 pages, 2017.

[111] A. Makhzani, and B. Frey, K-sparse autoencoders, *arXiv preprint arXiv:1312.5663*, 2013.

[112] M. Mao, J. Lu, G. Zhang and J. Zhang, Multirelational social recommendations via multigraph ranking, *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4049–4061, 2017.

[113] P. Melville, R. J. Mooney, and R. Nagarajan, Content-boosted collaborative filtering for improved recommendations, In: *Proceedings of the 18th National Conference on Artificial Intelligence*, Edmonton, Canada, pp. 187–192, July 28–August 1, 2002.

[114] A. Mondal, E. Cambria, D. Das, A. Hussain, and S. Bandyopadhyay, Relation extraction of medical concepts using categorization and sentiment analysis, *Cognitive Computation*, vol. 10, no. 4, pp. 670–685, Aug. 2018.

[115] A. Mustaqeem, S. M. Anwar, and M. Majid, A modular cluster based collaborative recommender system for cardiac patients, *Artificial Intelligence in Medicine*, vol. 102, art. no. 101761, 12 pages, 2020.

[116] M. Nasiri, B. Minaei, and A. Kiani, Dynamic recommendation: Disease prediction and prevention using recommender system, *International Journal of Basic Science in Medicine*, vol. 1, no. 1, pp. 13–17, 2016.

[117] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients, *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 240–255, 2004.

[118] K. Reetz, I. Dogan, A. S. Costa, et al, Biological and clinical characteristics of the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) cohort: a cross-sectional analysis of baseline data, *The Lancet Neurology*, vol. 14, no. 2, pp. 174–182, Feb. 2015.

[119] K. Reetz, I. Dogan, R.-D. Hilgers, et al, Progression characteristics of the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS): a 2 year cohort study, *The Lancet Neurology*, vol. 15, no. 13, pp. 1346–1354, Dec. 2016.

[120] U. Reimer, E. Maier, T. Ulmer, Automatic user adaptation for behavior change support, In: *Proceedings of the ACM workshop on engendering health with recommender systems*, Boston, MA, USA, Sep. 15–19, 2016.

[121] F. Ricci, L. Rokach, and B. Shapira, Recommender systems: introduction and challenges, *in Recommender systems handbook*, Boston: Springer, 2015, pp. 1–34.

[122] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, Contractive auto-encoders: Explicit invariance during feature extraction, *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, USA, pp. 833—840, Jun. 28–Jul. 2, 2011.

[123] R. L. Rose, G. M. Schwartz, and W. V. Ruggiero, "A Knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2124–2135, 2018.

[124] A. K. Sahoo, C. Pradhan, R. K. Barik, H. Dubey, DeepReco: deep learning based health recommender system using collaborative filtering, *Computation*, vol. 7, no. 2, art. no. 25, 18 pages, 2019.

[125] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, Item-based collaborative filtering recommendation algorithms, In: *Proceedings of the 10th international conference on World Wide Web*, Hong Kong, China, pp. 285–295, May 1–5, 2001.

[126] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie, Autorec: Autoencoders meet collaborative filtering, In: *Proceedings of the 24th international conference on World Wide Web*, Florence, Italy, pp. 111–112, May 18–22, 2015.

[127] H. Schäfer, S Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, and C. Trattner, Towards health (aware) recommender systems, In: *Proceedings of the 2017 international conference on digital health*, London, UK, pp. 157–161, Jul. 02-05, 2017.

[128] T. Schmitz-Hübsch, S. T. Du Montcel, L. Baliko, et al, Scale for the assessment and rating of ataxia: development of a new clinical scale, *Neurology*, vol. 66, no. 11, pp. 1717–1720, Jun. 2006.

[129] L. Seyer, N. Greeley, D. Foerster, et al, Open-label pilot study of interferon gamma-1b in Friedreich ataxia, *Acta Neurologica Scandinavica*, vol. 132, no. 1, pp. 7–15, Jul. 2015.

[130] D. Sharma, G. Singh Aujla, R. Bajaj, Evolution from ancient medication to human-centered Healthcare 4.0: A review on health care recommender systems, *International Journal of Communication Systems*, 2019, DOI:10.1002/dac.4058.

[131] F. Shi, N. Dey, A. S. Ashour, D. Sifaki-Pistolla, and R. S. Sherratt, Meta-KANSEI modeling with Valence-Arousal fMRI dataset of brain, *Cognitive Computation*, vol. 11, no. 2, pp. 227–240, Apr. 2019.

[132] X. Shi, Q. He, X. Luo, Y. Bai, M. Shang, Large-scale and scalable latent factor analysis via distributed alternative stochastic gradient descent for recommender systems, *IEEE Transactions on Big Data*, 2020, DOI: 10.1109/TBDATA.2020.2973141.

[133] Y. Shi and R. C. Eberhart, Empirical study of particle swarm optimization, In: *Proceedings of the 1999 IEEE Congress on Evolutionary Computation*, Washington, USA, pp. 1945–1950, July 6–9, 1999.

[134] Y. Shi and R. C. Eberhart, Parameter selection in particle swarm optimization, In: *Proceedings of the 7th International Conference on Evolutionary Programming*, San Diego, USA, pp. 591–600, March 25–27, 1998.

[135] H. Shin, S. Kim, J. Shin, and X. Xiao, Privacy enhanced matrix factorization for recommendation with local differential privacy, *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1770–1782, 2018.

[136] X. Shi, Q. He, X. Luo, Y. Bai, M. Shang, Large-scale and scalable latent factor analysis via distributed alternative stochastic gradient descent for recommender systems, *IEEE Transactions on Big Data*, 2020, DOI: 10.1109/TBDATA.2020.2973141.

[137] P. Siriaraya, K. Suzuki, and S. Nakajima, Utilizing collaborative filtering to recommend opportunities for positive affect in daily life, In: *Proceedings of the 4th International Workshop on Health Recommender Systems co-located with the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 2–3, Sept. 16–20, 2019.

[138] A. Starke, RecSys challenges in achieving sustainable eating habits, In: *Proceedings of the 4th International Workshop on Health Recommender Systems co-located with the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 29–30, Sept. 16–20, 2019.

[139] F. Strub, R. Gaudel, and J. Mary, Hybrid recommender system based on autoencoders, *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, Boston, USA, pp. 11–16, Sept. 15–19, 2016.

[140] S. H. Subramony, SARA-a new clinical scale for the assessment and rating of ataxia, *Nature Clinical Practice Neurology*, vol. 3, no. 3, pp. 136–137, 2007.

[141] G. Takács, I. Gábor, B. Németh, and D. Tikk, Major components of the gravity recommendation system, *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 80–83, Dec. 2007.

[142] Y. K. Tan, X. Xu, Y. Liu, Improved recurrent neural networks for session-based recommendations, In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, Boston, USA, Sept. 15–19, 2016.

[143] N. D. Thanh, M. Ali, and L. H. Son, A novel clustering algorithm in a neutrosophic recommender system for medical diagnosis, *Cognitive Computation*, vol. 9, no. 4, pp. 526–544, Aug. 2017.

[144] N. D. Thanh, M. Ali, Neutrosophic recommender system for medical diagnosis based on algebraic similarity measure and clustering, In: *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Naples, Italy, pp. 1–6, Jul. 9–12, 2017.

[145] N. T. Thong, and L. H. Son, HIFCF: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis, *Expert Systems with Applications*, vol. 42, no. 7, pp. 3682–3701, 2015.

[146] H. Torkamaan, and J. Ziegler, Multi-criteria rating-based preference elicitation in health recommender systems, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 18–23, Oct. 2–7, 2018.

[147] M. A. Torkamani, M. Jhaveri, J. Mellen, Engagement scoring for care-gap intervention optimization, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 52–56, Oct. 2–7, 2018.

[148] C. Trattner, and D. Elsweiler, An evaluation of recommendation algorithms for online recipe portals, In: *Proceedings of the 4th International Workshop on Health Recommender Systems co-located with the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 24–28, Sept. 16–20, 2019.

[149] A. Van den Oord, S. Dieleman, and B. Schrauwen, Deep content-based music recommendation, in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp. 2643–2651, December 5–8, 2013.

[150] Y. Ouyang, W. Liu, W. Rong, Z. Xiong, Autoencoder-based collaborative filtering, *Proceedings of the 21st International Conference on Neural Information Processing*, Kuching, Malaysia, pp. 284–291, Nov. 03–06, 2014.

[151] K. Oyibo, I. Adaji, and J. Vassileva, What drives the perceived credibility of health apps: Classical or expressive aesthetics?, In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, pp. 30–35, Oct. 2–7, 2018.

[152] W. Pan, E. W. Xiang, N. N. Liu, Q. Yang, Transfer learning in collaborative filtering for sparsity reduction, In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, USA, vol. 10, pp. 230–235, Jul. 11-–15, 2010.

[153] X. Pan, J. Song, and F. Zhang, Dynamic Recommendation of Physician Assortment With Patient Preference Learning, *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 1, pp. 115–126, 2018.

[154] V. Pandey, D. D. Upadhyay, N. Nag, and R. Jain, Personalized user modelling for context-aware lifestyle recommendations to improve sleep, In: *Proceedings of the 5th International Workshop on Health Recommender Systems co-located with 14th ACM Conference on Recommender Systems (HealthRecSys'20)*, Online, pp. 8–14, Sept. 26, 2020.

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

[156] A. Pasta, M. K. Petersen, K. J. Jensen, and J. E. Larsen, Rethinking hearing aids as recommender systems, In: *Proceedings of the 4th International Workshop on Health Recommender Systems co-located with the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 11–18, Sept. 16–20, 2019.

[157] P. Pilloni, L. Piras, L. Boratto, et al, Recommendation in persuasive eHealth systems: An effective strategy to spot users' losing motivation to exercise, In: *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th ACM Conference on Recommender Systems*, Como, Italy, pp.6–9, Aug. 27–31, 2017.

[158] L. Qi, X. Zhang, W. Dou, and Q. Ni, A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data, *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2616–2624, 2017.

[159] A. C. Valdez, M. Ziefle, K. Verbert, A. Felfernig, and A. Holzinger, Recommender systems for health informatics: Stateof-the-art and future perspectives, *Machine Learning for Health Informatics*, Springer, New York, NY, USA, 2016.

[160] Y. Varatharajah, H. Chen, A. Trotter, and R. Iyer, A Dynamic Human-in-the-loop Recommender System for Evidence-based Clinical Staging of COVID-19, In: *Proceedings of the 5th International Workshop on Health Recommender Systems co-located with 14th ACM Conference on Recommender Systems (HealthRecSys'20)*, Online, pp. 21–22, Sept. 26, 2020.

[161] P. Vincent, H. Larochelle, Y. Bengio, Extracting and composing robust features with denoising autoencoders, *Proceedings of the 25th international conference on Machine learning*, Helsinki, Finland, pp. 1096–1103, Jul. 5–9, 2008.

[162] L. Wang, W. Zhang, X. He, and H. Zha, Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation, In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, London, UK, pp. 2447–2456, Aug. 19–23, 2018.

[163] X. Wang, X. He, L. Nie, T. S. Chua, Item silk road: Recommending items from information domains to social users, In: *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, Tokyo, Japan, pp. 185-–194, Aug. 7—11, 2017.

[164] Q.-X. Wang, X. Luo, Y. Li, X.-Y. Shi, L. Gu and M.-S. Shang, Incremental slope-one recommenders, *Neurocomputing*, vol. 272, pp. 606–618, 2018.

[165] M. Waqar, N. Majeed, H. Dawood, A. Daud, N. R. Aljohani, An adaptive doctor-recommender system, *Behaviour & Information Technology*, vol. 38, no. 9, pp. 959–973, 2019.

[166] J. Wei, J. He, K. Chen, et al, Collaborative filtering and deep learning based recommendation system for cold start items, *Expert Systems with Applications*, vol. 69, pp. 29–39, 2017.

[167] A. Weyer, M. Abele, T. Schmitz-Hübsch, et al, Reliability and validity of the scale for the assessment and rating of ataxia: A study in 64 ataxia patients, *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 22, no. 11, pp. 1633–1637, 2007.

[168] M. Wiesner, D. Pfeifer, Health recommender systems: Concepts, requirements, technical basics and challenges, *International journal of environmental research and public health* , vol. 11, no. 3, pp. 2580–2607, 2014.

[169] D. Wu, Q. He, X. Luo, M. Shang, Y. He, and G. Wang, A posterior-neighborhood-regularized latent factor model for highly accurate web service qos prediction, *IEEE Transactions on Services Computing*, 2019, DOI: 10.1109/TSC.2019.2961895.

[170] D. Wu, X. Luo, M. Shang, Y. He, G. Wang, and M. C. Zhou, A deep latent factor model for high-dimensional and sparse matrices in recommender systems, *IEEE Transactions on System Man Cybernetics: Systems*, 2019, DOI: 10.1109/TSMC.2019.2931393.

[171] S. Wu, W. Ren, C. Yu, G. Chen, D. Zhang, and J. Zhu, Personal recommendation using deep recurrent neural networks in NetEase, In: *Proceedings of the 32nd IEEE International Conference on Data Engineering*, Helsinki, Finland, May 16–20, 2016.

[172] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, Collaborative denoising auto-encoders for top-n recommender systems, *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, San Francisco, USA, pp. 153–162, Feb. 22–25, 2016.

[173] Z. Wu, J. Cao, Y. Wang, Y. Wang, L. Zhang and J. Wu, hPSD: A hybrid PU-learning-based spammer detection model for product reviews, *IEEE Transactions on Cybernetics*, in press, 2019. DOI:10.1109/TCYB.2018.2877161.

[174] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, Collaborative denoising auto-encoders for top-n recommender systems, in *Proceedings of the 9th ACM International Con-*

ference on Web Search and Data Mining, San Francisco, USA, pp. 153–162, February 22–25, 2016.

[175] T. Schmitz-Hübsch, S. T. Du Montcel, L. Baliko, et al, Scale for the assessment and rating of ataxia: development of a new clinical scale, *Neurology*, vol. 66, no. 11, pp. 1717–1720, Jun. 2006.

[176] G. R. Xue, C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu, and Z. Chen, Scalable collaborative filtering using cluster-based smoothing, in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, pp. 114–121, August 15–19, 2005.

[177] I. Yabe, M. Matsushima, H. Some, et al, Usefulness of the scale for assessment and rating of ataxia (SARA), *Journal of the Neurological Sciences*, vol. 266, no. 1–2, pp. 164–166, 2007.

[178] Y. Yang, Z. Zheng, X. Niu, M. Tang, Y. Lu, and X. Liao, A location-based factorization machine model for web service QoS prediction, *IEEE Transactions on Services Computing*, DOI: 10.1109/TSC.2018.2876532.

[179] W. Yue, Z. Wang, B. Tian, A. Payne, and X. Liu, A collaborative-filtering-based data collection strategy for Friedreich's ataxia, *Cognitive Computation*, vol. 12, pp. 249–260, 2020.

[180] W. Yue, Z. Wang, B. Tian, M. Pook, and X. Liu, A hybrid model- and memory-based collaborative filtering algorithm for baseline data prediction of Friedreich's ataxia patients, *IEEE Transactions on Industrial Informatics*, 2020, DOI:10.1109/TII.2020.2984540.

[181] W. Yue, Z. Wang, W. Liu, B. Tian, S. Lauria, and X. Liu, An optimally weighted user- and item-based collaborative filtering approach to predicting baseline data for Friedreich's ataxia patients, *Neurocomputing*, 2020, DOI:10.1016/j.neucom.2020.08.031.

[182] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim Y. Li, and X. Liu, An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochro-

matographic strips, *IEEE Transactions on Nanotechnology*, vol. 18, pp. 819–829, 2019.

[183] Z.-H. Zhan, Z.-J. Wang, H. Jin, and J. Zhang, Adaptive distributed differential evolution, *IEEE Transactions on Cybernetics*, 2019, DOI:10.1109/TCYB.2019.2944873.

[184] H. Zhang, and S. Sheng, Learning weighted naive Bayes with accurate ranking, In: *Proceedings of the Fourth IEEE International Conference on Data Mining*, Brighton, UK, pp. 567–570, Nov. 1–4, 2004.

[185] H. Zhang, D. Yue, C. Dou, K. Li, and X. Xie, Event-triggered multi-agent optimization for two-layered model of hybrid energy system with price bidding based demand response, *IEEE Transactions on Cybernetics*, 2019, DOI:10.1109/TCYB.2019.2931706.

[186] L. Zhang, X. Chen, N-N. Guan, H. Liu, and J-Q. Li, A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction, *Frontiers in Pharmacology*, DOI:10.3389/fphar.2018.01017.

[187] J. Zhang, Y. Lin, M. Lin, and J. Liu, An effective collaborative filtering algorithm based on user preference clustering, *Applied Intelligence*, vol. 45, no. 2, pp. 230–240, 2016.

[188] Q. Zhang, G. Zhang, J. Lu, D. Wu, A framework of hybrid recommender system for personalized clinical prescription, In: *Proceedings of the 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Taipei, Taiwan, pp. 189–195, Nov. 24–27, 2015.

[189] Q. Zhang, G. Zhang, J. Lu, D. Wu, Semantics-enhanced recommendation system for social healthcare, In: *Proceedings of the 2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, Victoria, Canada, pp. 765–770, May 13–16, 2014.

[190] S. Zhang, L. Yao, A. Sun, and Y. Tay, Deep learning based recommender system: A survey and new perspectives, *ACM Computing Surveys (CSUR)*, vol. 52. no. 1, art. no. 5, 38 pages, 2019.

[191] X. L. Zhang, X. Chen, N. N. Guan, H. Liu, and J. Q. Li, A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction, *Frontiers in Pharmacology*, vol. 9, art. no. 1017, 11 pages, 2018.

[192] Y. Zhang, M. Chen, D. Huang, D. Wu, and Y. Li, iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization, *Future Generation Computer Systems*, vol. 66, pp. 30–35, 2017.

[193] Y. Zhang, K. Meng, and W. Kong, "Collaborative filtering-based electricity plan recommender system," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1393–1404, 2019.

[194] L. Zheng, V. Noroozi, and P. S. Yu, Joint Deep Modeling of Users and Items Using Reviews for Recommendation, In: *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, Cambridge, UK, pp. 425–434 Feb. 06–10, 2017.

[195] L. Zheng, C. T. Lu, L. He, et al, MARS: Memory attention-aware recommender system, In: *Proceedings of the 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Washington DC, USA, pp. 11–20, Oct. 05–08, 2019.

[196] Z. Zheng, H. Ma, M. R. Lyu, and I. King, Collaborative web service QoS prediction via neighborhood integrated matrix factorization, *IEEE Transactions on Services Computing*, vol. 6, no. 3, pp. 289–299, Jul. 2013.

[197] Z. Zheng, H. Ma, M. R. Lyu, and I. King, Qos-aware web service recommendation by collaborative filtering, *IEEE Transactions on Services Computing*, vol. 4, no. 2, pp. 140–152, Apr. 2011.

[198] Z. Zheng, H. Ma, M. R. Lyu, and I. King, WSRec: A collaborative filtering based web service recommender system, In: *Proceedings of the 2009 IEEE International Conference on Web Services*, Los Angeles, USA, pp. 437–444, July 6–10, 2009.

[199] Y. Zhu, J. Lin, S. H, B. Wang, Z. Guan, H. Liu, and D. Cai, Adressing the item cold-start problem by attribute-driven active learning, *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 631–644, 2020.