This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

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

Wenbin Yue, Zidong Wang, *Fellow, IEEE*, Bo Tian, Mark Pook and Xiaohui Liu

***Abstract*—Friedreich's ataxia (FRDA) is the most common inherited ataxia that causes progressive damage of nervous systems and performance deterioration of physical movements. FRDA baseline data analysis plays a crucial role in advancing the disease research, where the main obstacle comes from the baseline data collection due primarily to the degenerative symptoms of the FRDA patients. Inspired by the nowadays popular collaborative filtering (CF) method, a new FRDA baseline data collection algorithm is proposed in this paper, with which the patients (or their families) are only required to provide certain reliable baseline data acquired from home and the uncertain/missing parts of the data can then be predicted with acceptable accuracy by utilizing existing patient information. The framework of the proposed algorithm is constructed based on a novel hybrid model combining the merits of model- and memory-based CF methods, thereby facilitating the baseline data collection with improved prediction accuracy. 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 Scale for the Assessment and Rating of Ataxia (SARA) is selected from the European Friedreich's Ataxia Consortium for Translational Studies database. Experimental results demonstrate that our proposed hybrid CF approach is superior to other conventional approaches.**

***Index Terms*—Collaborative filtering, Friedreich's Ataxia, K-means clustering, Shannon entropy, Jaccard index.**

I. INTRODUCTION

FRIEDREICH'S ataxia (FRDA) is a genetic disorder that causes progressive damage to the nervous system and leads to muscle weakness, deep sensory loss, loss of position sense, difficulty in speech or even heart disease [3]. The first symptom for an FRDA patient is usually the difficulty in walking that gives rise to the necessity of using wheelchairs. FRDA was named after Nicolaus Friedreich, a German doctor who first described the condition in 1863. FRDA is the most common hereditary ataxia across most of Europe with the prevalence of 2–4 in every $100,000$ individuals. The symptoms usually first appear around puberty, but in a few cases, symptoms develop in adulthood or early childhood. Though there is currently no effective therapy method to cure FRDA, many of the symptoms and complications of the disease can be treated in order to help patients maintain optimal physical functioning.

To have a comprehensive understanding of FRDA, the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) assembles a group of experts to create the first prospective international European FRDA registry in 2010 with the aim to improve FRDA patients' health status. EFACTS is committed to adopting a translational research strategy by combining basic biology research with clinical trials to solve practical problems for FRDA patients [18], [19]. One of the most important tasks is to collect and analyze different kinds of baseline data which are of significant use in clinical trials and fundamental research. Until the end of 2018, EFACTS has collected 989 patients' baseline data from 14 study sites distributed in 9 European countries. This number only accounts for less than 7% of the total FRDA patients in the European Union, and the potential FRDA patient data size is considerably large.

Many existing clinical studies suffer from small sample sizes that cause the results to be insignificant [16]. Clearly, more baseline data can help promote better disease research in terms of sufficient sample selection, effective biostatistical analysis and extensive clinical studies [6], [17]. By the traditional data collection method, the patients take part in the corresponding tests at nearby study sites where the organization (EFACTS) can collect accurate and detailed baseline data. Unfortunately, such a traditional method has many drawbacks such as unaffordable cost and low efficiency. Furthermore, the specific pathology of the FRDA is likely to cause inconvenience for some patients to be physically present at the study sites due to poor physical conditions. Thus, EFACTS is faced with many challenges which result in slow/inefficient data collection. The FRDA patient baseline data is collected at different local EFACTS study sites through patient interviews, questionnaires, observations and coordinated tests. It was considered whether an alternative way can be designed for patients who are unable to attend the assessments in the study sites.

This work was supported in part by the Seventh Framework Programme of the European Union under Grant 242193 (EFACTS), the Royal Society of the U.K., and the Alexander von Humboldt Foundation of Germany. (*Corresponding author: Zidong Wang*)

W. Yue, Z. Wang and X. Liu are with the Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, U.K. (e-mail: Zidong.Wang@brunel.ac.uk).

B. Tian is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China.

M. Pook is with the Department of Life Sciences, Brunel University London, Uxbridge UB8 3PH, U.K.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

In here, we make a reasonable assumption that most FRDA patients or their families can provide some accurate baseline data after long-term observation and care. In this case, there is an urgent need to develop effective/efficient algorithms with particular aim to handle the data imperfection/incompleteness issues.

In search of suitable approaches for overcoming the sparsity problem in the baseline data collection for FRDA patients, the collaborative filtering (CF) based algorithms appear to be a competitive candidate. CF is one of the most popular and successful techniques for recommendation systems [20] that have received considerable attention since the mid-1990s with wide application in a variety of fields such as E-commerce, media, entertainment, government, education and other fields [14], [15], [21], [28]. The main idea of CF-based algorithms is to analyze the active users' interests through user behaviors so as to find similar users in the communities. Based on these similar users' comments on a piece of information, CF-based algorithms predict the preference degree of active user to determine whether to make a recommendation. In this sense, the main idea of the CF-based algorithms is particularly suitable for predicting the missing/uncertain/incomplete FRDA baseline data.

In this paper, 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 [27]. 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 paper, 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 paper are outlined in three-fold 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, and the memory-based component takes the individual differences between patients into the calculation of similarities quantified by Shannon information entropy and Jaccard index. 3) Comprehensive experiments are carried out to show that our proposed hybrid CF algorithm improves the prediction accuracy of the FRDA baseline data.

II. The literature review

The CF algorithm is used to design recommendation systems and this algorithm was first introduced in 1992 by Goldberg et al [8]. In this section, we review some major approaches of CF that will be used in this paper.

A. Memory-based CF approaches

The memory-based CF approaches (also called neighborhood-based CF approaches) are among the most

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

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 [10]. 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 [1], [22], PCC similarity measure can be easily implemented and can achieve a better overall performance than others.

B. 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 recent years, different artificial neural network (ANN) models [9] (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 [11], convolutional neural network [24], autoencoder [23] and so on [5]. Other well-known model-based approaches are latent factor model and probabilistic model which involves probabilistic semantic analysis, aspect modeling and probabilistic matrix factorization.

C. 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 [2]. 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 paper, 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.

III. The Methodology

A. 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 [25], 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 I 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.

B. Hybrid collaborative filtering framework

In this paper, 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.

C. The model-based CF component

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 paper, 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 [26].

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 [4].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

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:

$$P(r_{a,i}) = \frac{\sum_{u \in \hat{K}_a} r_{u,i}}{|\hat{K}_a|}, \quad (1)$$

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.

D. 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 [29]. The similarity degree between patients u and a is calculated by

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

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 (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:

$$J(u,a) = \frac{|I_u \cap I_a|}{|I_u \cup I_a|} = \frac{|I_u \cap I_a|}{|I_u| + |I_a| - |I_u \cap I_a|}, \quad (3)$$

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):

$$\mathrm{Sim}^J(u,a) = J(u,a) \times \mathrm{Sim}(u,a), \quad (4)$$

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

Remark 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) 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) 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. [7], [12]), is defined as:

$$H_u = -\sum_{r \in RD} \mathcal{P}_{u,r} \log_2 \mathcal{P}_{u,r}, \quad (5)$$

where H_u denotes the entropy of patient u, $\mathcal{P}_{u,r}$ 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 [13] as follows:

$$\mathrm{Sim}^E(u,a) = \frac{1}{1 + |H_u - H_a|}\mathrm{Sim}(u,a). \quad (6)$$

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:

$$\mathrm{Sim}^{EJ}(u,a) = \frac{1}{1 + |H_u - H_a|} \times J(u,a) \times \mathrm{Sim}(u,a) \quad (7)$$

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 2: Our proposed enhanced similarity measure (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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

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]
$0 \sim 8$	$0 \sim 6$	$0 \sim 4$	$0 \sim 6$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$

Fast alternating hand movements			Heel-shin slide			SARA Total [b]
right	left	mean [a]	right	left	mean [a]	
$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 4$	$0 \sim 40$

[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.

differences in rating scores. As such, the enhanced PCC (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 (7) constitutes the second contribution of this paper.

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 [29] are added to the top-n algorithm as follows:

$$\hat{S}(a) = \{a_u | a_u \in T(a), \mathrm{Sim}(a_u, a) > 0, a_u \neq a\}, \quad (8)$$

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 [1]:

$$P(r_{a,i}) = \bar{a} + \frac{\sum_{a_u \in \hat{S}(a)} \mathrm{Sim}^{EJ}(a_u, a)(r_{a_u,i} - \bar{a}_u)}{\sum_{a_u \in \hat{S}(a)} \mathrm{Sim}^{EJ}(a_u, a)}, \quad (9)$$

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 3: In this paper, 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;)
  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;)
  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 III-D)
  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.

IV. IMPLEMENTATION AND EXPERIMENTS

A. Data preprocessing

The data set SARA is constantly updated. Until $31st$ December 2018, the SARA data set has contained the information of 989 patients. As shown in Table I, the rating intervals for test-items are not identical and, therefore, we adopt a feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

TABLE II
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.

scaling method, known as unity-based normalization, to bring all rating values into the range of $[0, 1]$ according to

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{10}$$

where x' represents a normalized value, $x_{\min}$ represents the minimum value in x given its range, and $x_{\max}$ describes the maximum value of in x given its range.

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 II. "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.

B. 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 I, 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:

$$MAE = \frac{1}{N} \sum_{a \in A_d} \sum_{i \in I_d} |r_{a,i} - P(r_{a,i})|, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{a \in A_d} \sum_{i \in I_d} (r_{a,i} - P(r_{a,i}))^2} \tag{12}$$

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.

C. 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. 1 shows the experimental results.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

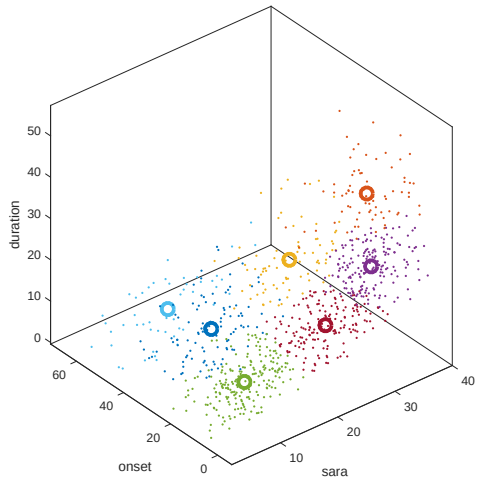

Fig. 1. K-means clustering diagram

1) 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 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:

$$\text{RMEAN}: \quad P(r_{a,i}) = \bar{r}_a \tag{13}$$

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

$$\text{UMEAN}: \quad P(r_{a,i}) = \frac{\sum_{u=1}^{n} r_{u,i}}{n} \tag{14}$$

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

$$\text{CUMEAN}: \quad P(r_{a,i}) = \bar{r}_i + \frac{\sum_{u=1}^{n}(r_{u,i} - \bar{r}_u)}{n} \tag{15}$$

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

$$\text{AUMEAN}: \quad P(r_{a,i}) = \frac{\sum_{u \in \hat{S}} r_{u,i}}{|\hat{S}|} \tag{16}$$

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 III. 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 III. 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 III
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**

2) 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 (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 Fig. 2 and Fig. 3, 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 IV from which we conclude that:

1) The proposed algorithm demonstrates its superiority over MEAN, SREGR, UPCC, UPCCE and UPCCJ in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

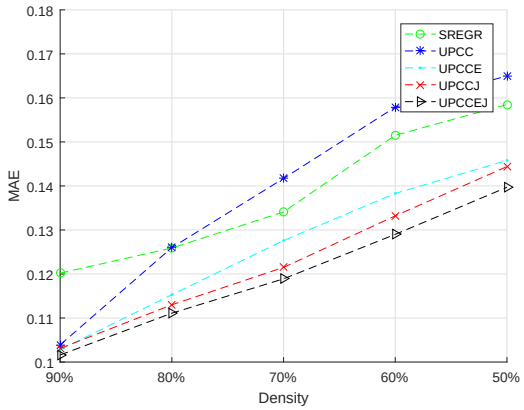

Fig. 2. Line graphs of MAE.

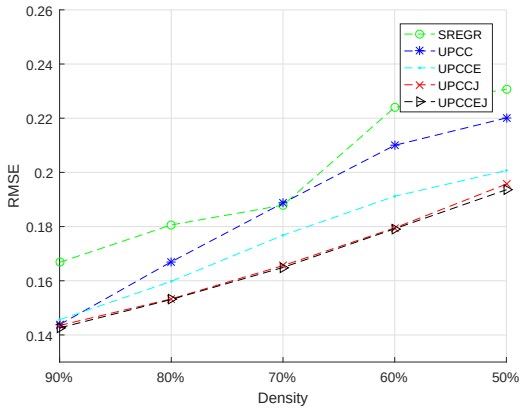

Fig. 3. Line graphs of RMSE.

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 IV
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**

V. DISCUSSIONS AND CONCLUSION

A. Discussions

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 paper, 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TII.2020.2984540, IEEE Transactions on Industrial Informatics

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.

B. Conclusions

In this paper, 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.

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.

REFERENCES

[1] 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.

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

[3] 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, 1996.

[4] 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.

[5] 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.

[6] 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, 2007.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] X. Luo, M. Zhou, S. Li, and M. 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.

[15] 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.

[16] D. R. Lynch, and E. Kichula, "Challenges ahead for trials in Friedreichs ataxia," *The Lancet Neurology*, vol. 15, no. 13, pp. 1300–1301, 2016.

[17] 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, 2012.

[18] 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, 2015.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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, 2011.

Wenbin Yue received his B.S degree in 2013 and M.S. degree in 2015, both in information technology from the Queensland University of Technology, Brisbane, Australia. He is currently pursuing the Ph.D. degree in computer science at Brunel University London, London, UK. His research interests include machine learning, recommendation system and big data analysis.

Zidong Wang (SM'03-F'14) was born in Jiangsu, China, in 1966. He received the B.Sc. degree in mathematics in 1986 from Suzhou University, Suzhou, China, and the M.Sc. degree in applied mathematics in 1990 and the Ph.D. degree in electrical engineering in 1994, both from Nanjing University of Science and Technology, Nanjing, China.

He is currently a Professor of Dynamical Systems and Computing in the Department of Computer Science, Brunel University London, UK. From 1990 to 2002, he held teaching and research appointments in universities in China, Germany and the UK. His research interests include dynamical systems, signal processing, bioinformatics, control theory and applications. He has published 220+ papers in IEEE Transactions and 60+ papers in Automatica. He is a holder of the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, William Mong Visiting Research Fellowship of Hong Kong.

Prof. Wang serves (or has served) as the Editor-in-Chief for Neurocomputing, the Deputy Editor-in-Chief for International Journal of Systems Science, and an Associate Editor for 12 international journals including IEEE Transactions on Automatic Control, IEEE Transactions on Control System Technology, IEEE Transactions on Neural Networks, IEEE Transactions on Signal Processing, and IEEE Transactions on Systems, Man, and Cybernetics-Systems. He is a Fellow of the IEEE, a Fellow of the Royal Statistical Society and a member of program committee for many international conferences.

Bo Tian received the B.Eng. degree in automation from Beihang University, Beijing, China, in 2015, and is now pursuing his Ph.D. degree in control science and engineering in Beihang University, Beijing, China. From Mar. 2018 to May 2018, he was a visiting student in the Department of Computer Science, Brunel University London, UK. His research interests include stochastic control and estimation, information theory, and intelligent data analysis. He is an active reviewer for some international journals.

Mark Pook received the B.Eng. degree in genetics from University of Leeds, Leeds, U.K., in 1985 and the Ph.D. degree in medical genetics from University of Manchester, Manchester, U.K., in 1990.

He is currently a reader in the Division of Biosciences at Brunel University, U.K. He has a longstanding interest in cellular and molecular genetic studies of the inherited neurodegenerative disorder, Friedreich ataxia (FRDA), contributing over the last 24 years to the mapping of the disease locus, the isolation of candidate genes and the identification of novel FXN mutations. He has also amassed 20 years of experience in transgenic technology, culminating in the development of GAA repeat-based FRDA transgenic mouse models for the investigation of FRDA disease pathogenesis and therapy. He currently heads the Ataxia Research Group at Brunel University London, where he has recently performed several FRDA mouse model drug studies and published the results in peer-reviewed journals. He is a member of a number of national and international FRDA research collaborations that has the ultimate goal of developing novel FRDA therapeutic strategies. In particular, He is committed to the generation and distribution of FRDA mouse models to investigators throughout the world to promote the development of FRDA therapeutic interventions.

Xiaohui Liu received the B.Eng. degree in computing from Hohai University, Nanjing, China, in 1982 and the Ph.D. degree in computer science from Heriot-Watt University, Edinburg, U.K., in 1988.

He is currently a Professor of Computing at Brunel University. He leads the Intelligent Data Analysis (IDA) Group, performing interdisciplinary research involving artificial intelligence, dynamic systems, image and signal processing, and statistics, particularly for applications in biology, engineering and medicine. Professor Liu serves on editorial boards of four computing journals, founded the biennial international conference series on IDA in 1995, and has given numerous invited talks in bioinformatics, data mining and statistics conferences.