processes

MDPI

Article

CDL4CDRP: A Collaborative Deep Learning Approach for Clinical Decision and Risk Prediction

Mingrui Sun *, Tengfei Min, Tianyi Zang * and Yadong Wang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; hitvitamin@163.com (T.M.); ydwang@hit.edu.cn (Y.W.)

* Correspondence: mingrui.sun@gmail.com (M.S.); tianyi.zang@gmail.com (T.Z.)

Received: 19 February 2019; Accepted: 3 May 2019; Published: 7 May 2019

Abstract: (1) Background: Recommendation algorithms have played a vital role in the prediction of personalized recommendation for clinical decision support systems (CDSSs). Machine learning methods are powerful tools for disease diagnosis. Unfortunately, they must deal with missing data, as this will result in data error and limit the potential patterns and features associated with obtaining a clinical decision; (2) Methods: Recent years, collaborative filtering (CF) have proven to be a valuable means of coping with missing data prediction. In order to address the challenge of missing data prediction and latent feature extraction, neighbor-based and latent features-based CF methods are presented for clinical disease diagnosis. The novel discriminative restricted Boltzmann machine (DRBM) model is proposed to extract the latent features, where the deep learning technique is adopted to analyze the clinical data; (3) Results: Proposed methods were compared to machine learning models, using two different publicly available clinical datasets, which has various types of inputs and different quantity of missing. We also evaluated the performance of our algorithm, using clinical datasets that were missing at random (MAR), which were missing at various degrees; and (4) Conclusions: The experimental results demonstrate that DRBM can effectively capture the latent features of real clinical data and exhibits excellent performance for predicting missing values and result classification.

Keywords: clinical diagnosis; recommender systems; prediction algorithms; machine learning; decision support systems

1. Introduction

The clinical decision support systems as a practical tool is designed to take the full advantage of patient medical record data, thereby influencing the clinical decisions of doctors in key areas via data mining, and assist doctors in providing timely and correct decisions for patients [1,2]. How to utilize patient medical record data effectively, helping doctors reduce the rates of misdiagnosis and missed diagnosis, has been one of the focuses clinical decision support systems (CDSSs) research [3]. Remarkable progress has been made in clinical disease auxiliary diagnosis using methods for natural language processing and machine learning, such as multilayer perceptron (MLP) [4], support vector machines (SVMs) [5], logistic regression (LR) [6], and random forests (RFs) [7]. However, a large number of diagnostic test results are at various degrees of missing in patient medical record datasets. Clinical diagnostic datasets, accompanied by such data features, need various interpolations for missing data, which lead to data standard errors caused by the data filling [8]. These data standard errors limit the ability to obtain potential patterns and implicit features of clinical decisions for traditional machine learning methods.

Recommender systems have been widely used to provide data-driven suggestions for individuals in all fields [9,10]. The prediction of recommendations based on historical data from an individual

is similar to buying behaviors or preferences. Recommender systems can be broadly classified as either content-based or based on collaborative filtering (CF) [11]. Content-based methods deduce a preference structure of the individual based on detailed attributes of their personal preferences. Content-based approaches utilize a series of discrete characteristics of an item to recommend additional items with similar properties. CF relies on the concept that, individuals who agree on ratings of items are likely to also agree on ratings of other items, but are probably not aware of those items. CF can be used to predict item ratings for an individual and collectively develop a personalized ranking of items that may be of interest to them. In clinical diagnosis, the objectives are similar to predicting ratings or classifying missing items. In the context of clinical diagnosis, items may represent clinical variables or diagnostic codes. However, distinct from classical marketing applications, clinical data are based on factors, such as medical examinations, clinical measurements, and professional experience, that cannot be changed according to an individual's preferences. In other words, patient preferences are irrelevant in clinical application. Additionally, the lack of diagnostic data may indicate that a patient has not yet been diagnosed, but not necessarily that he does not have a disease. In general, recommender systems use a Likert scale to different degrees to unify all item attributes. In contrast to recommender systems, clinical diagnostic data consist of variable types (e.g., continuous, discrete, and taxonomic), and how to model and fuse data from different source databases, is a considerable challenge. In clinical diagnosis, recommender systems, based on CF are mainly applied to predict the probability of disease occurrence from clinical data, consisting of diagnostic codes. It is clear that predicting the probability of disease from clinical data is perfectly reasonable, just as the recommender systems are often used with massive and sparse databases. The features of the data in clinical diagnosis are similar to those in a recommender system; thus, the processing of clinical data, using technology in the field of recommender systems, is a further direction of exploration.

Compared with shallow machine learning, deep learning has an excellent ability to automatically extract abstract features [12]. The essence of intelligent recommendation systems is abstracting an individual interest factor from a jumble of raw streaming data to discover user preferences. Thus, combining deep learning with recommender systems, has become a new focus in recent years. As a foundation and representative of the deep learning field, the restricted Boltzmann machine (RBM) has generated wide interest and has been developed for a large variety of learning problems [13]. The RBM is a generative latent variable model that models the joint distribution of a set of input variables [14], and it has recently been extended for representational learning [15], document modelling [16], multi-label learning [17,18], dimensionality reduction [19], CF [20], and computational biology [21], as well as many other tasks [11]. Recently, the predictive power of RBMs was also demonstrated in a Netflix prize contest [20,22], a public competition to develop the best CF algorithm in order to predict user ratings for movies. However, RBMs are usually used for feature extraction for another learning algorithm, or to provide a good initialization for supervised feed-forward neural network classifiers; they are not considered a standalone solution to classification problems. The problem that needs to be settled is whether RBMs can provide a self-contained framework for a recommender system and derive classifiers and predictors.

In this article, we propose an effective deep-learning approach, named CDL4CDRP, a collaborative deep learning approach for clinical decisions and risk prediction. Specifically, the problem of predicting the values of an unknown outcome, and missing predictor variables, was addressed by leveraging implicit feedback CF algorithm, based on discriminative RBM (DRBM). For verifying the validity of the algorithm. We also compared our algorithm with SVM, RF, LR, MLP, and user-based CF with different types of imputation. These algorithms were compared by two different publicly available clinical datasets: University of California Irvine (UCI) chronic kidney disease (CKD) and dermatology data. Our experimental results demonstrate that the DRBM-based method has a superior overall performance for every dataset across varying levels of missing data (10%, 20%, and 30%). In particular, compared with other algorithms, the accuracy of classification was still over 90% for two datasets, with a degree of missing data as high as 30%. Moreover, the DRBM-based method can maintain good

prediction performance under the conditions of severe missing data by deeply studying the complex associations between physiological information and preliminary information. The recommender algorithm, based on DRBM, can be used as a stand-alone non-linear classifier and an acceptable alternative for clinical risk prediction. The main contributions of this work include:

(1) A collaborative deep-learning approach is proposed for clinical decisions and risk prediction, named CDL4CDRP;

(2) Proposed DRBM-based CF as a stand-alone non-linear classifier for missing value prediction;

(3) Analysis latent features of the DRBM model.

The structure of the article is as follows: Section 2 states related work; Section 3 described materials and proposed formulates classification methods; Section 4 gives the experimental results; Section 5 discusses related issues, and Section 6 concludes this work.

2. Related Work

Because of the popularity of user-based CF for generating recommendations based on similarities, the researchers aspired to investigate user-based CF techniques for clinical prediction and diagnosis in clinical data [23]. In the biomedical field, most applications that are based on CF have focused on the prediction of comorbidities from clinical data composed of diagnostic disease codes [24,25]. These methods combine the diagnostic codes with extra information, such as clinical data, patient preferences and history, and environmental factors, in order to improve forecasting accuracy.

In recent years, the implicit feedback recommendation algorithm has attracted increasing academic attention [26]. Compared with the traditional machine learning methods, the biggest problems for implicit feedback is the lack of negative feedback, and due to the data features missing and data noise problems, it unable to use the supervised learning method [27]. Pan et al. [28] presented one-class collaborative filtering (OCCF) methods based on positive examples (explicit feedback), used to solve the recommendation problem of the implicit feedback for binary data [29]. Paque et al. [30] presented to combine Bayesian generative model with random graphs for implicit collaborative filtering, and formulated a probabilistic model to prediction. Meanwhile, a lot of implicit feedback methods based on model are also appeared, such as derivative model based on matrix factorization: Singular value decomposition (SVD++) [31], probability matrix decomposition (PMF) [32] and HeteroMF [33]. One-class collaborative filtering include the confidence model and probability theory model. The advantage of the model is simple and easy to implement, and the disadvantage is the integration problem of the sampling training results. Karatzoglou et al. [34] proposed three recommendation algorithms, based on the idea of "learn to rank": point-wise [35], pair-wise [36] and list-wise [37]. The advantage is that the time complexity of the algorithm is low, and recommended list results are reasonable. But ignore the association between the samples. Researchers also consider to introduce auxiliary information, combined with implicit feedback for recommendation, such as music recommendation [38], and graph-based model for context-aware recommendation [39]. Algorithms that introduce auxiliary information can combine different scenarios to mine meaningful feature attributes. However, it is necessary to pay attention to the mutual influence of multi-domain knowledge. These methods above provide theory and method for implicit feedback recommendation.

Recent developments have demonstrated that deep learning is a powerful generative models, which are able to extract latent features automatically and obtain high predictive performance [18]. Tomczak et al. [40] focused on a variant of RBM adopted to the classification setting, and showed how to obtain a sparse representation in RBM, by adding a regularization term to the learning objective which enforces sparse solution. Eickholt et al. [21] presented DNCON, a new sequence-based residue–residue contact predictor, using RBMs to learn patterns in the data and initialized parameters, and then fine-tune them with back propagation. Maxwell et al. [18] developed a multi-label classification method, using deep-learning architecture for the purposes of predicting chronic disease, such as hypertension in patients for physicians. Due to the lack of practicality in industrial applications, Luo et al. [41] presented a non-negative matrix-factorization (NMF), based CF model with a single-element-based

approach. It is suitable for solving CF problems subject to the constraints of non-negativity. Suffering from problems of high computational and storage complexity, as well as slow convergence rate, Luo et al. [42,43] also built non-negative latent factor (NLF) models, with two variants, INLF and ANLF, with large-scale sparse matrices and high performance. In the domain of service computing, for missing QoS data of web services, diversified NLF model and DNLF ensemble model [44], and second-order optimization-based LF model [45] are present to accurate predictions missing QoS value.

Our research concerns are similar to Hao et al. [46], who conducted risk prediction, using user-based CF to find similarities in clinical features, in order to predict similar diseases among patients. They compared user-based CF recommenders with LR and RF algorithms, using various imputations on four clinical datasets and found that user-based CF was inferior to the other two algorithms. With the advent of the era of big data, recommender systems are natural tendency to combine with CF and other extensible deep learning methods for clinic data, which are growing in both size and complexity. In addition, RBM model usually used as feature extraction and provided initialization for other deep learning algorithms, we explore if RBM model can be considered as a standalone solution to address the classification problems.

3. Materials and Methods

3.1. Materials

Two different publicly available clinical datasets were surveyed, namely, CKD data and dermatology data, obtained through the UCI machine learning repository. There are missing data and heterogeneity in the population for many of the measured predictors. Each of two datasets has different levels of missing data within predictor variables. In this context, many of the features revealed by Big Data are present on a smaller scale in the two datasets, but our experiment increases the missing rate for the purpose of simulating a real medical environment where there is often a high degree of missing clinical data. Each data under investigation has a category outcome, can therefore be modelled as a classification problem.

3.1.1. Chronic Kidney Disease Dataset

The CKD dataset was collected from the hospital over a period of nearly two months and can be used to predict CKD through a set of 24 feature properties, including age and 23 physiological measurements, and 1 class label (ckd, notckd). Of the 25 attributes, 11 are continuous numeric, and 14 are nominal class attributes. There are 400 observations in the dataset. The response variable is a binary indicator for CKD. The predictor variables that have continuous values are discretized to categorical values. This dataset is available through the UCI machine learning repository [47] and is summarized in Table 1.

Table 1. The chronic kidney disease dataset used in the experiments [47].

f_{25}: Classes	Features		
	Numerical (Numeric Assigned Values of 0, 1, ... and 6)		
C_1: ckd C_2: notckd	f_1. (2, 34], (34, 46], (46, 54], (54, 60], (60, 67], (67, 90] f_2. (50, 60], (60, 70], (70, 76], (76, 80], (80, 90], (90, 180] f_{10}. (22, 94], (94, 108], (108, 125], (125, 148.2], (148.2, 203], (203, 490] f_{11}. (1.5, 23], (23, 32], (32, 44], (44, 53], (53, 85], (85, 391] f_{12}. (0.4, 0.8], (0.8, 1.1], (1.1, 1.3], (1.3, 2.2], (2.2, 3.9], (3.9, 76] f_{13}. (4.5, 135], (135, 137], (137, 137.5], (137.5, 139], (139, 142], (142, 163] f_{14}. (2.5, 3.7], (3.7, 4.2], (4.2, 4.6], (4.6, 4.62], (4.62, 4.9], (4.9, 47] f_{15}. (3.1, 9.8], (9.8, 11.4], (11.4, 12.5], (12.5, 13.7], (13.7, 15.2], (15.2, 17.8] f_{16}. (9, 31], (31, 37], (37, 38.9], (38.9, 42], (42, 47], (47, 54] f_{17}. (2200, 6300], (6300, 7700], (7700, 8406], (8406, 8406.2], (8406.2, 9800], (9800, 26,400] f_{18}. (2.1, 3.9], (3.9, 4.7], (4.7, 4.71], (4.71, 4.8], (4.8, 5.4], (5.4, 8]		
	Nominal Class Attributes		
	f_3. 1.005, 1.01, 1.015, 1.02, 1.025 f_6. abnormal, normal f_9. notpresent, present f_{21}. no, yes f_{24}. no, yes	f_4. 0, 1, 2, 3, 4, 5 f_7. abnormal, normal f_{19}. no, yes f_{22}. poor, good	f_5. 0, 1, 2, 3, 4, 5 f_8. notpresent, present f_{20}. no, yes f_{23}. no, yes

3.1.2. Dermatology Dataset

The dermatology dataset was collected to classify erythemato-squamous diseases among six possible disease types. Such differential diagnosis of erythemato-squamous diseases has been a challenge in dermatology. Firstly, these diseases all share the clinical features of erythema and scaling, with very few differences. The diseases in this group are psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis and pityriasis rubra pilaris. In addition, some patients can be diagnosed by these clinical features alone; however, a biopsy is usually necessary for a correct and definite diagnosis. Furthermore, the difficulty for differential diagnosis is that disease may show the histopathological features of another disease at the beginning stage and may have different features in later stages. There are 366 instances in the dataset and a set of 34 attributes, 33 of which are numeric and 1 of which is nominal. This dataset is available through the UCI machine learning repository [47], and is summarized in Table 2.

Table 2. The dermatology dataset used in the experiments [47].

Classes (Class Label)	Features						
	Clinical (Numeric Assigned Values of 0, 1, 2, and 3)			**Histopathological (Numeric Assigned Values of 0, 1, 2, and 3)**			
C_1: Psoriasis	f_1.	f_2.	f_3.	f_{12}.	f_{13}.	f_{14}.	f_{15}.
C_2: Seborrheic dermatitis	f_4.	f_5.	f_6.	f_{16}.	f_{17}.	f_{18}.	f_{19}.
C_3: Lichen planus	f_7.	f_8.	f_9.	f_{20}.	f_{21}.	f_{22}.	f_{23}.
C_4: Pityriasis rosea	f_{10}.	f_{11} (0 or 1)		f_{24}.	f_{25}.	f_{26}.	f_{27}.
C_5: Cronic dermatitis	f_{34}: (0, 21], (21, 29], (29, 36], (36,			f_{28}.	f_{29}.	f_{30}.	f_{31}.
C_6: Pityriasis rubra pilaris	43], (43, 52], (52, 75]			f_{32}	f_{33}.		

3.2. *Classification Methods*

Each of the two datasets used for prediction and classification requires recasting a supervised learning task as unsupervised learning problem. The purpose is, not only to compare the performance of the various classification methods, but also to evaluate the CF-based recommender system, whether it can be appropriate for clinic data imputation and classification. We focused on the contrast recommender system, based on SVMs [48], RFs [49], LR [50], and MLP [51]. The models of RBM, DRBM, and user-based CF were implemented, based on a deep learning framework, named TensorFlow [52].

3.2.1. User-Based Collaborative Filtering

The recommender system, based on CF, is the most widely used recommender algorithm in personalized recommendation systems. The essence of CF can be summarized as "birds of a feather flock together". The algorithm relies on individual rating data to deduce the missing values for other users and items. Personalized ranked lists of items are created, based on other users/items, with similar attributes of ratings. To make personalized recommendations, the algorithm can first look for neighbours with similar interests within a certain range, and then analyse the items that the neighbours liked, finally recommending the new items to the target users.

In clinical applications, the users are patients, and items are come from clinical and histopathological features of the patients. Several of these features may not exist or be random missing. We define the patients as $P = \{P_1, P_2, \ldots, P_n\}$ and the clinical and histopathological features of these patients as $F = \{F_1, F_2, \ldots, F_P\}$. The classification task of predicting target users is to calculate the k most similar users, and the calculation is then weighted by similarity. The patients similarity between P_i and P_j is defined as the cosine distance between their features F_i [46]:

$$sim_{\cos}\left(P_i\left(F^i\right), P_j\left(F^j\right)\right) = \frac{\left\langle P_i\left(F^i\right),\ P_j\left(F^j\right)\right\rangle}{||P_i(F^i)||||P_j\left(F^j\right)||} \tag{1}$$

where $\langle\cdot\rangle$ is the inner product and $\|\cdot\|$ is the Euclidean norm. With regard to the evaluation score of the target patients, we chose the formula $r_{c,s} = k \sum_{\dot{c}\in\hat{C}} sim\left(c, \dot{c}\right) \times r_{\dot{c},s}$ because the more similar the patient features, the better the weights can predict the missing values. It would not be suitable to choose the method of preference filtering for evaluation because differences already exist in the diagnostic data of patients. Feature F_i of patient P_j is estimated as:

$$\hat{F}_i^{P_j} = \frac{1}{\sum_{h\in N(P_j)} sim_{cos}\left(P_j, h\right)} \sum_{h\in N(P_j)} sim_{cos}\left(P_j, h\right) \cdot F_{h,i} \tag{2}$$

where $h \in N\left(P_j\right)$ is the neighbourhood centred on patient P_j.

The steps of user-based CF algorithm is described below.

Step 1: This step to get inversion matrix of patients and features;

Step 2: This step to get patient similarity matrix through a similarity computation method (cosine distance) and by statistics the number of similar features between two patients.

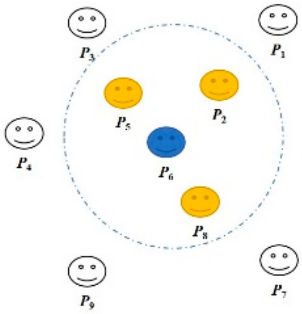
Step 3: This step to get the patients set of high similarity for target patients according to patient similarity matrix. And predict feature missing value of target patient by k most similar neighborhoods patients.

A conceptual schematic, describing the idea of a neighbor for patient P_6, is shown in Figure 1a. The missing data and classification results are predicted by aggregating across 3 neighbors (Figure 1b). In order to calculate the similarity between the target patient and all n patients, we first need to compare m features and find patients with similar features of target patient. Thus the time complexity for calculating the similarity between the target patient and all n patients is $O(n \times m)$. Since m and n are the same order of magnitude, the time complexity is $O(n^2)$. Then, we list n patients order by similarity and sort k nearest neighbor, the time complexity is $O(n\log n)$. Recommendation list by k nearest neighbor are given and time complexity is $O(k \times n)$. Where k is a constant, it is much less than n, so the time complexity is $O(n)$. In general, when the number of patients is large, the algorithm for calculate the similarity is unacceptable.

In contrast to traditional machine learning, where clinical prediction problems are treated as a supervised learning task, CF recasts a supervised learning task as an unsupervised problem. In the execution of CF using clinical data, the response variable Y is treated as another predictor, F_{P+1}, with patient similarity as the prediction being made. There are disadvantages in methods, based on

neighborhood similarity. First, such methods will lead to a larger error of similarity calculation and be incomputable. Second, the classifier tends to output multicategory results when the disease category itself is unbalanced in the dataset. In our applications, the selection of the number of neighbors, k, was made based on a 3-fold cross-validation. The implementation of CF was performed using recommender lab in the *R* programming language (https://www.r-project.org) [53].

a. Local neighborhood of size k=3 b. Collaborative filtering for P_9 on neighborhood of size k=3



	F_1	F_2	F_3	F_4	F_5	.....	F_p	Y
P_1	2	NA	3	1	3	.....		0
P_2	2	4	3	1	NA	.....		1
P_3	4	2	1	3	4	.....		1
P_4	NA	1	NA	2	NA	.....		0
P_5	2	2	1	NA	1	.....		1
P_6	1	3	NA	3	NA	.....		NA
P_7	1	NA	1	1	4	.....		0
P_8	2	4	NA	3	2	.....		0
P_9	1	NA	NA	NA	3	.....		1
P_6 (predicted)	2	1	2.5	3	2			1

Figure 1. User-based collaborative filtering: (**a**) Conceptual schematic describing the idea of a neighbor for patient P_6 for collaborative filtering of size k = 3. The distance between neighborhoods and patient P_6 is quantified by the cosine distance; (**b**) The predicted recommended outcomes for patient P_6 are aggregate estimates over the three neighborhoods.

3.2.2. RBM-Based Collaborative Filtering for Clinical Diagnosis

The RBM [13,14] is an undirected generative graphical model, with two layers that use a layer of hidden variables to model a distribution over visible variables. It is derived from the Boltzmann machine, a type of depth probability graph model, where the vertex represents random variables and the edge represents the interdependence of the variables. The RBM has two layers: A visible layer that consists of input variables and a hidden layer, which consists of hidden variables. In the visible layer node-set V, v represents visible units $v = \{v_1, \ldots, v_n\} \in V$, n represents the number of neurons, and v_i indicates the value of the ith neuron. Similarly, in the hidden layer node-set H, h represents hidden units $h = \{h_1, \ldots, h_m\} \in H$, m represents the number of neurons, and h_j indicates the value of the jth neuron. The two layers are fully interconnected, but no connections exist between any two hidden units or any two visible units, consequently forming a complete dichotomy, as shown in Figure 2. This provides a good hypothetical condition for our later training, that is the neurons in the same layer are independent of each other.

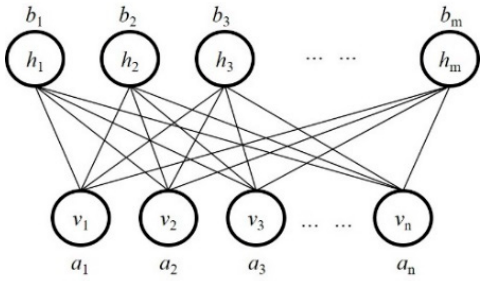


Figure 2. Network graph of an restricted Boltzmann machine (RBM) with n visible units and m hidden units.

In the RBM, the visual layer V is also the input layer, and its value can either be a real number or binary (0 or 1). The hidden layer H is used to extract the features of the visible layer with the binary value of 0 or 1 and is subject to a Bernoulli distribution. A neuron equal to 1 indicates activation; otherwise, it is inactive or inhibitory. In this paper, we assign the hidden layer and visual layer of the

RBM to a binary neural network model. The edge between the ith visible unit and the jth hidden unit is associated with a weight w_{ij}. Together, the weights are represented as a weight matrix W. The weights of connections between visible units and the bias unit are contained in a visible bias vector $a = (a_1, a_2, \cdots, a_n)^T$. Likewise, for the hidden units, there is a hidden bias vector $b = (b_1, b_2, \cdots, b_m)^T$. The RBM is fully characterized by the parameters W, a and b.

The RBM is an energy-based probabilistic graphical model that gives the joint probability of every possible pair of visible and hidden vectors via an energy function [13,14]:

$$E(v,h) = -a^T \cdot v - b^T \cdot h - v^T \cdot W \cdot h \tag{3}$$

$$E(v,h) = -\sum_{i=1}^{n} a_i \times v_i - \sum_{j=1}^{m} b_j \times h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i \times w_{i,j} \times h_j \tag{4}$$

The maximum clique set Q is constituted through any visual layer unit v_i and any hidden layer unit h_j. We define the potential function of the maximum clique set Q, which describes the interrelationship between variable sets, as:

$$\psi_Q\left(v_i, h_j\right) = e^{-E(v_i,\, h_j)} = e^{a_i \times v_i + b_j \times h_j + v_i \times w_{i\ j} \times h_j} \tag{5}$$

In a Markov network, the model joint probability distribution $P(v, h)$ of multiple variables is based on the product of the maximum clique decomposed into multiple potential functions:

$$P(v,h) = \frac{1}{Z} \prod_{Q=(v_i,h_j)\epsilon C} \psi_Q\left(v_i, h_j\right) = \frac{1}{Z} e^{-E(v,h)} \tag{6}$$

where C denotes all of the maximum clique set, and the "partition function" Z denotes the normalization factor obtained by summing over all possible pairs of visible and hidden vectors:

$$Z = \sum_{v}\sum_{h} e^{-E(v,h)} \tag{7}$$

and ensures that $P(v, h)$ is a probability range from 0 to 1. The marginal probability $P(v)$ assigned by the model to the elements of a visible vector v is obtained by summing (marginalizing) over all possible hidden vectors:

$$P(v) = \sum_{h\in\{0.1\}^m} \frac{e^{-E(v,h)}}{Z} = \frac{e^{-F(v)}}{Z} \tag{8}$$

where $F(v) = -a^T \cdot v - \sum_{j=1}^{m} \ln\left(1 + e^{b_j + v^T \cdot w_{*,j}}\right)$ denotes free energy and $\ln\left(1 + e^{b_j + v^T \cdot w_{*,j}}\right)$ denotes the softplus function. Likewise, the marginal probability $P(h)$ assigned by the model to the elements of a hidden vector h is obtained by summing (marginalizing) over all possible visible vectors:

$$P(h) = \sum_{v\in\{0.1\}^n} \frac{e^{-E(v,h)}}{Z} = \frac{e^{-F(h)}}{Z} \tag{9}$$

where $F(h) = -b^T \cdot h - \sum_{i=1}^{n} \ln\left(1 + e^{b_j + w_{i,*} \cdot h}\right)$. In its original form, the RBM models the Bernoulli distribution in its visible and hidden layers. As there is no intra-layer connection between any pair of visible or hidden units, we can define the following conditional probabilities [13,14]:

$$P(v_i = 1|h) = \frac{1}{1 + e^{-a_i - w_{i,*} \cdot h}} \tag{10}$$

$$P\left(h_j = 1 \middle| v\right) = \frac{1}{1 + e^{-b_j - v^T \cdot w_{*,j}}} \tag{11}$$

These conditional probabilities are important for the iterative updates between hidden and visible layers when training an RBM model.

As mentioned above, RBM is an energy-based model; if the combination of random variables is reasonable, the smaller the corresponding energy function value, the larger the distribution probability. Learning in energy-based models can be carried out generatively by determining the weights and biases that minimize the overall energy of the system, with respect to the training data. According to the concept of maximum likelihood, all the samples in the training data are real and reasonable, which should have larger probability. The learning goal of RBM can be formalized into maximizing the log-likelihood function L over the training data V (containing N examples), which is given by [13,14]:

$$L = \frac{1}{N} \sum_{k=1}^{N} ln\left(P\left(V^k\right)\right) \tag{12}$$

where V^k denotes the kth sample. When solving the above optimization problem using gradient descent-based optimization, the maximization problem can be transformed into a minimization problem by negating the above equation. We further analyse $\ln(P(V^k))$ and obtain:

$$\ln(P(V)) = -F(V) - \ln(Z) \tag{13}$$

The gradient of the log-likelihood function with respect to the θ parameters of RBM needs to be calculated first. We take the derivative of the last formula and obtain [13,14]:

$$\frac{\partial \ln(P(V))}{\partial \theta} = \frac{\partial(-F(V))}{\partial \theta} - \frac{\partial(Z)}{\partial \theta} \tag{14}$$

RBM-based CF for clinical diagnosis is mainly divided into two parts: Training model parameters and the reconstruction input layer. In the previous data pre-processing, we discretized continuous attributes and recast them into category attributes using the characteristics of medical datasets with a variety of features. For the traditional RBM model, all units are binary neurons, and the characteristics of datasets are all transformed into numerical category attributes; therefore, we considered the use of the common one-hot coding method. We processed all features with one-hot coding, which transforms each sample into a binary vector. The vector is the input of the visual layer, in which each sample with a missing value is given a default value of 0 and retains the mask matrix with the same dimension of the transformed binary vector. The mask matrix used to record the one-hot coding location of missing features is shown in Figure 3:

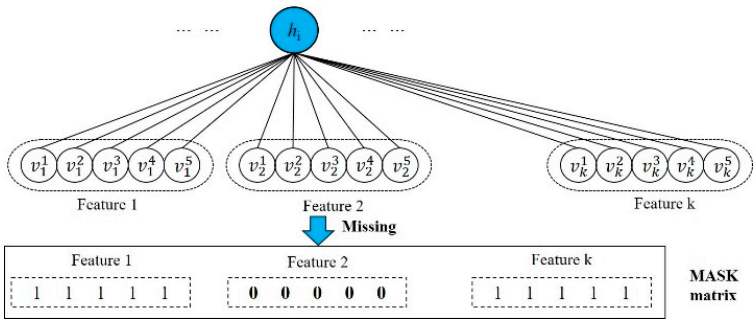


Figure 3. Data preprocessing based on the restricted Boltzmann machine (RBM) model.

For formula (14), the first term is the calculation of the positive gradient, and depends on the sample data. The latter term is the calculation of the negative gradient, it need to sum over all values

of v. Because of all values of the visual layer neurons v is binary, $v = \{0, 1\}^n$, a total of 2^n possibility, time complexity for the calculation of the negative gradient is $O(2^n)$. It obviously that exponential time complexity is not suitable for the value n is very large.

The derivation of negative gradient is actually solving expectations, we can consider to use distribution $P(v)$ to sampling v, and get k different sets of sample data, which is approximately equal to all possible values of v. Based on this idea, Larochelle et al. [13,54] put forward contrastive divergence (CD) learning algorithm, on the basis of the Gibbs sampling further reduces the time complexity of RBM training. It mainly based on two assumptions: (1) Sampling can be directly from training data, because it has been very close to the real distribution; (2) CD algorithm does not need to be repeated iteration, until meet the convergence conditions. Generally, a step number of sampling $k = 1$ (CD − 1) can achieve the desired effect. The CD algorithm pseudo-code, used to train and update parameters for RBM, is described according to Algorithm 1. The algorithm complexity only depends on the value of n, m, n_step and cd_k. Generally speaking, cd_k is usually set to 1, and n_step almost equal to constant c. The time complexity for RBM training is $O(c(m + n))$.

Algorithm 1 Training update for RBM over (h, v) based on CD

Input: training set D, number of iterations n_step, step number of sampling cd_k
Output: updated parameters, a^*, b^*, W^*
Step 1: Input training dataset D, number of iterations n_step, step number of Gibbs sampling cd_k;
Step 2: Initialization RBM parameters a, b, W;
Step 3: Gibbs sampling is performed by controlling the number of iterations cd_k, update h^k and v^{k+1} according to Equations (10) and (11) respectively;
Step 4: Update parameter a, b, W according to the following formula,

$$w_{ij} \leftarrow w_{ij} + \left(p\left(h_j = 1 \middle| v\right) \times v_i - p\left(h_j = 1 \middle| v^{cd_k}\right) \times v_i^{cd_k}\right)$$

$$a_i \leftarrow a_i + \left(v_i - v_i^{cd\_k}\right)$$

$$b_j \leftarrow b_j + \left(p\left(h_j = 1 \middle| v\right) - p\left(h_j = 1 \middle| v^{cd\_k}\right)\right)$$

Step 5: Output parameters a^*, b^*, W^* as model parameters

After learning the parameters of the neural network by the CD algorithm, we use RBM to reconstruct the input layer, so that we can obtain the prediction results of the sample category attribute, and other missing attributes that need to be predicted. There are two main parts of the reconstruction input: Encoding and decoding. In the process of encoding, we can obtain the state H^* of the hidden layer neuron according to conditional distribution $P(h_j = 1|v)$ sampling, by using the original input sample. In contrast to encoding, the decoding process reconstructs the state V^* of the visible layer through the state H^* of the hidden layer neuron and the conditional distribution $P(v_i = 1|H^*)$. To reconstruct each feature (including the category attribute of the sample), we build a Softmax module for each neuron that restores category attribute of the feature.

$$P\left(v_i^k = 1 \middle| h\right) = \frac{e^{a_i^k + \sum_{j=1}^{m} w_{i,j}^k \times h_j}}{\sum_{t=1}^{l} e^{a_i^t + \sum_{j=1}^{m} w_{i,j}^t \times h_j}} \tag{15}$$

where l denotes the number of neurons. By taking $\max_k\left(p\left(v_i^k = 1 \middle| h\right)\right)$, we reconstruct the category of the sample and the values of other missing category attributes.

In fact, recommender system based on RBM model applied to CDSSs were required to transform a supervised learning task into an unsupervised problem. In RBM, we treated the predictive sample category by the same process as that used for the other features as the input data of the visible layer, which is an unsupervised learning method that underutilizes the sample category information.

3.2.3. DRBM-Based Collaborative Filtering for Clinical Diagnosis

The DRBM is a variant of RBM with the difference that the DRBM adds a layer of category, tag y, connected to the hidden layer [13,14], as shown in Figure 4.

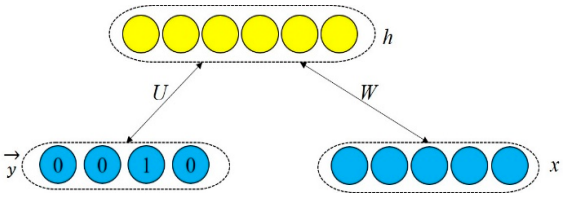


Figure 4. DRBM modelling the joint distribution of inputs and target classes.

We can obtain a posteriori probability [13,14]:

$$P(y|x) = \frac{e^{-E_{free}(x,y)}}{\sum_{y^*} e^{-E_{free}(x,y)}} \tag{16}$$

where $x = (x_1, x_2, \ldots, x_D)$ denotes the input matrix of the visible layer, $y \in \{1,2, \ldots, C\}$ denotes the one-hot coding matrix of the sample category label, and y^* denotes all possible category labels. Compared with the original RBM model described in the previous section, it can be seen that both x and y are visible layers in the DRBM model. This conditional distribution can be computed exactly and efficiently by writing it as follows:

$$P(y|x) = \frac{e^{d_y} \prod_j \left(1 + e^{c_j + U_{j,y} + \sum_i W_{j,i} \cdot x_i}\right)}{\sum_{y^*} e^{d_{y^*}} \prod_j \left(1 + e^{c_j + U_{j,y^*} + \sum_i W_{j,i} \cdot x_i}\right)} \tag{17}$$

The parameter learning of the DRBM model can refer to the learning of the original RBM model. Similarly, gradient estimation can be carried out by the Gibbs sampling method to enable rapid parameter learning.

Unlike the RBM model, we put the sample labels into the vector y of the visible layer. By maximizing all samples based on the product of the conditional probability $P(x|y)$, we can learn the probability distribution model that represents the interrelationship between the sample features and the sample category labels. We can make the correct category judgement for the samples through the probability distribution, even though the characteristic information of the sample is completely missing. Precomputing the terms $c_j + \sum_i W_{j,i} \cdot x_i$ and reusing them when computing $c_j + U_{j,y^*} + \sum_i W_{j,i} \cdot x_i$ for all classes y^* permits to comput the conditional probability in time $O(nD + nC)$ [13,54].

3.3. Simulation

We designed an experiment for manipulating the datasets, in order to contain a percentage of missing values. The CKD and dermatology data correspond to 8%, and 5% of the missing data at baseline, respectively. In addition to the baseline missing data, a percentage of the data is deleted at random with low (10%), moderate (20%), and severe (30%) levels of missing data. Data pre-processing in six steps: (1) data cleaning; (2) data division into 3 folds; (3) create missing data; (4) data discretization; (5) model fitting and imputation; and (6) data visualization and evaluations. Detailed steps are below:

1. Data Statistics and Cleaning

The datasets were input into the simulation pipeline: Chronic kidney and dermatology. Data cleaning mainly includes format content cleaning (we convert data features into numerical attributes) and abnormal value checking and processing (by drawing box diagrams, the value distribution of one-dimensional numerical attributes is displayed, and outliers are found and corrected).

2. Data Division into Three Folds for Rotation Estimation

Each dataset is divided into $k = 3$ folds for rotation estimation. The analyses were performed through a 3-fold rotation estimation on each dataset. Therefore, two-thirds of the data were used as training data in each fold, and the rest were used as test data. In our experiments of real and simulated data, we repeated the rotation estimation process 30 times. For each run, the folds were fixed throughout the simulation of the different levels of missing data to achieve a cumulative effect.

3. Missing Data Creation at Random

For the real datasets, a fixed percentage of values of predictor variables were randomly deleted with the goal of simulating MAR settings [55]. The MAR rate of missing data was matched to the MCAR rate of missing data, in order to enable fair comparisons. Since the deletion was random across all predictor variables from the original data, each was affected to a comparable extent. In all settings, the pattern of missing data was deleted at random and cumulative across the varying levels of severity, with low (10%), moderate (20%), and severe (30%) levels of missing data.

4. Discretization of Continuous Variables

Recommender algorithm models were designed to utilize ratings, which are generally categorical or ordinal values. The datasets under experimentation contained a mixture of variable types. However, to facilitate fair comparisons, the predictor variables that have continuous values are discretized to categorical values. Specifically, for each of the two sets, the maximal levels taken by categorical or ordinal variables were used to discretize the continuous variables. The threshold of discretization is determined according to the quantile, which ensures the balance of the categorical variables after discretization. Based on this theory, the CKD and dermatology data were subject to 5- and 4-level discretization, respectively.

5. Model Fitting and Imputation

CF and traditional machine learning methods were applied to each dataset, as mentioned above. Imputation mainly includes the following methods: (1) The mean imputation was used for each predictor variable in the training data. Subsequently, this mean value replaced all of the missing values for the corresponding variable. The interpolations were the mean values of all non-null values of each feature. (2) The *k*-nearest neighbours algorithm (*k*-NN) is a non-parametric method used for classification. The nearest *k* tuples for missing-value data are determined according to the Euclidean distance and Markov distance functions, and the weighted mean of the *k* values is then used to estimate the missing value. (3) Multiple imputation by chained equations (MICE) [56] uses Gibbs sampling to complete a multivariate dataset by iterating over a set of conditional densities representing the variables in the dataset. This paper mainly adopts following MICE imputation methods: Classification and regression tree (cart) imputation, predictive mean matching (pmm) imputation, and random forest (rf) imputation. These method were applied only in traditional machine learning. For methods based on CF, the algorithm directly predicts the results of the missing value.

6. Data Visualization and Performance Evaluations

Performance was based on the mean misclassification rate (0–1 loss) across all the folds, and the standard error of this mean estimate was calculated as the standard deviation across the folds. Matrix diagrams are often used to visualize missing datasets. Data visualization helps us better understand the characteristics of the experimental data.

4. Results

Missing data for CKD and dermatology datasets are created with low (10%), moderate (20%), and severe (30%) levels of missing data. For each scenario, data are divided into $k = 3$ folds for rotation estimation. The dataset experienced discretization and interpolation. Classification algorithms, with various imputations, were performed. Performance is evaluated using the precision rate and F1-score. F1-score used to assist in evaluating overall performance.

4.1. Data Visualization

The matrix graph of the dermatology dataset and CKD, with varying degrees of missing data, are drawn in Figure 5. The matrix graphs with the original and missing values show that each row represents instance data, and the numerical data are rescaled to the interval from 0 to 1; the size is represented by grey scale. The deeper colours indicate that the value of the instance related to the value of the feature is large, and vice versa. Red indicates missing values. Only one attribute of the baseline data has several missing values. As seen from the missing datasets of 10%, 20% and 30%,

the red area is gradually enlarged and is evenly distributed, which indicates that the created random missing dataset is of good quality. The matrix graph of the CKD dataset (as Figure 5b), the degree of missing data of the real datasets is nearly 8%. Almost every instance has a missing value in the datasets, with levels of missing data of 20% or more.

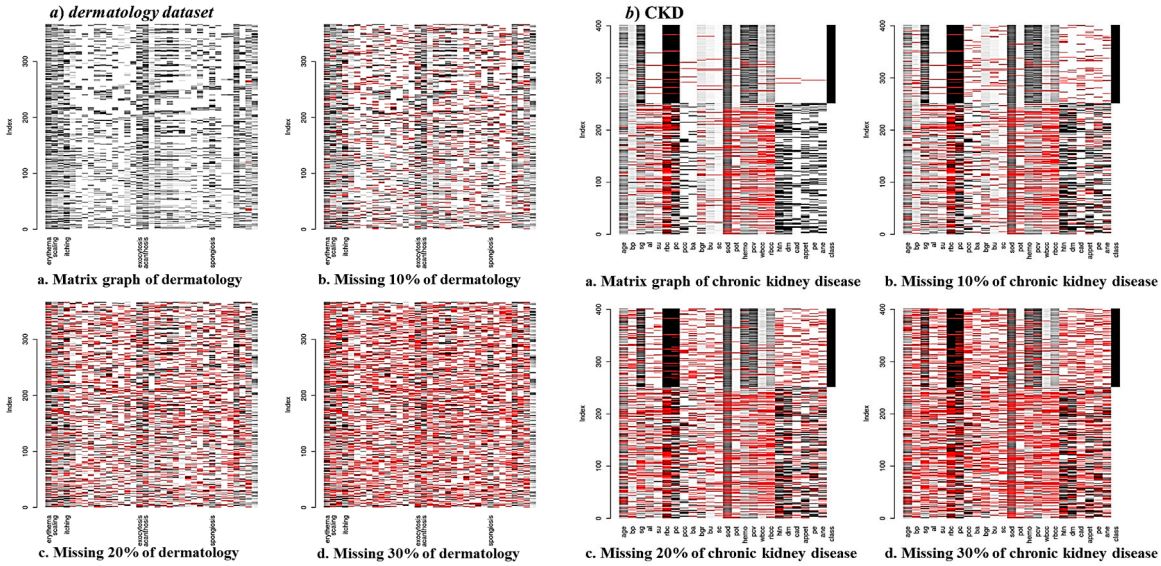


Figure 5. Matrix graph of dataset with varying degrees of missing: (**a**) dermatology dataset; (**b**) chronic kidney disease (CKD).

4.2. User-based CF for Clinical Disease Diagnosis

In this section, we apply the user-based CF model to conduct experiments, with the dermatology and CKD datasets, and the corresponding missing dataset created from them. The experimental results are shown in Figure 6.

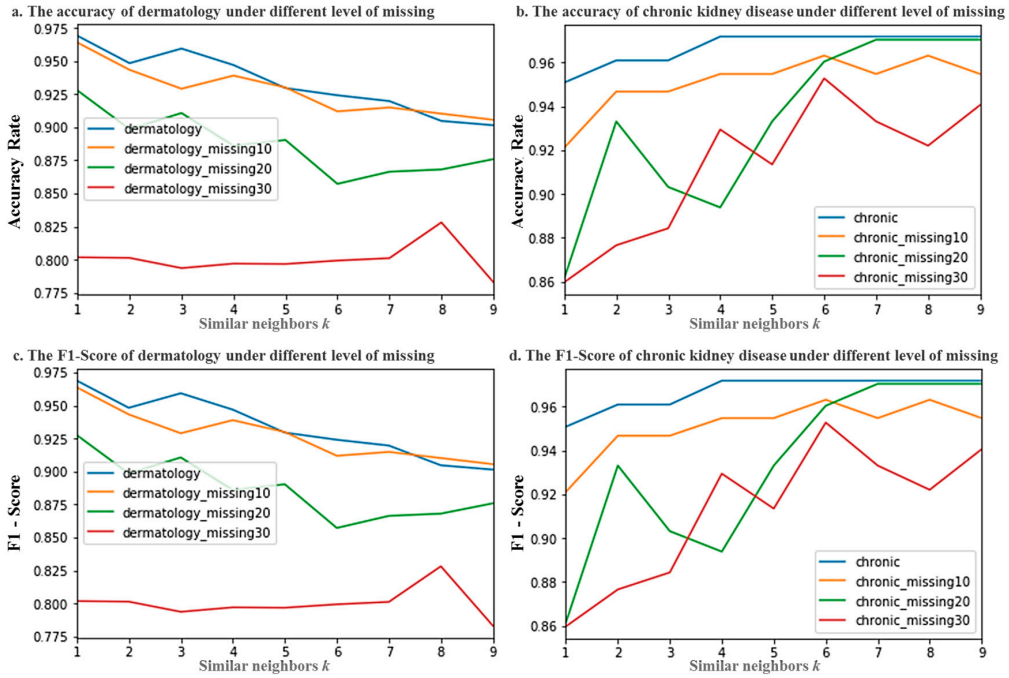


Figure 6. Experimental results of a user-based collaborative filtering model.

The user-based CF method has an accuracy of more than 80% for both dermatology and CKD, with a maximum accuracy of approximately 97.5%, and the F1-scores are similar. Moreover, the variation trend of the F1-score is identical to the accuracy, which indicates that when the accuracy of the k value is high, its F1-score is also high. This result shows that the recall rate of the model is also good. In addition, it can be observed that when the missing degree of dataset is severe, the prediction accuracy of the model and the F1-score is lower, which indicates that information loss, has hindered model prediction. We collate the best results of the model in each dataset, as shown in Table 3.

Table 3. Best results of user-based collaborative filtering (CF) model for each dataset.

Missingness	Baseline	10%	20%	30%
Accuracy of dermatology	0.96894 ($k = 1$)	0.96382 ($k = 1$)	0.92764 ($k = 1$)	0.82814 ($k = 8$)
F1-score of dermatology	0.96787 ($k = 1$)	0.96313 ($k = 1$)	0.8937 ($k = 1$)	0.72510 ($k = 8$)
Accuracy of CKD	0.97183 ($k = 4$)	0.96894 ($k = 6$)	0.97041 ($k = 7$)	0.95275 ($k = 6$)
F1-score of CKD	0.97019 ($k = 4$)	0.96894 ($k = 6$)	0.96894 ($k = 7$)	0.95077 ($k = 6$)

As demonstrated in the statistics in Table 3, with regard to the diagnosis of the dermatology datasets, when there is a lower level of missing data and fewer referred similar patients (small k value), the results are more accurate. When the level of missing data reaches more than 30%, more patient information (high k value) needs to be referred for a more accurate diagnosis. In the diagnosis of CKD, the number of patients referred increase, with an increase in the missing data; even when the level of missing data is 30%, the model still has a diagnostic accuracy of 95.275% and an F1-score of 0.95077. In summary, the models learned about the different features of the two datasets. The severe level of missing data (30%) of the dermatology dataset is accompanied by an F1-score of only approximately 0.73, but other deficiencies (baseline, 10%, and 20%) remain at approximately 0.9 or greater. In particular, even though the original data are missing by 10%, the model still accurately predicts the category based on the existing information. For the CKD dataset, the decline is not significant for the prediction accuracy and the F1-score of the model, even if the level of missing data is more serious, which indicates that the model can make a correct diagnostic prediction by using other information even if certain data are missing. Overall, model of user-based CF effectively solves the problem of medical data prediction with a large degree of missing data.

4.3. RBM-Based CF for Clinical Disease Diagnosis

The experimental results, that are based on the unsupervised disease diagnosis model of RBM, are shown in Table 4. As seen in this table, the overall classification performance of the RBM-based model is poor. We analyze the training process as shown in Figure 7.

Table 4. Results of restricted Boltzmann machine (RBM)-based model for each dataset.

Missingness	Baseline	10%	20%	30%
F1-score of dermatology	0.441	0.382	0.291	0.152
F1-score of CKD	0.730	0.686	0.619	0.530

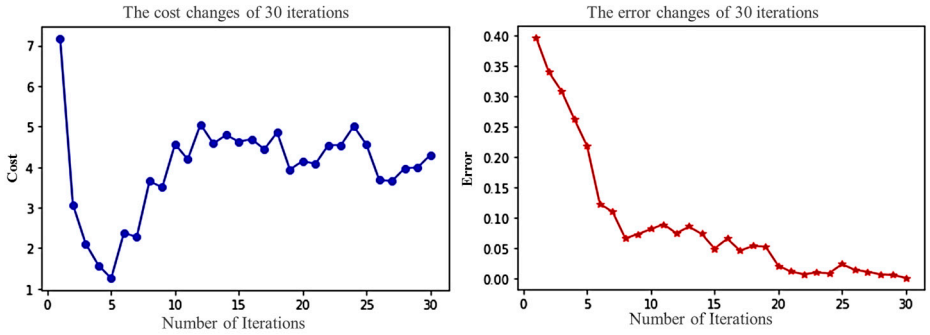


Figure 7. Variations in cost and error in RBM training.

As seen from the figure, the cost function of the RBM model is optimized to a very small value, and only 5 iterations are needed. However, the average error of the visible layer is still high, and the RBM model has a large prediction error for missing values. This error is mostly due to the small datasets. The datasets are too small to provide enough information; therefore, it is difficult to extract enough feature information from the visible layer simply through multiple iterations to construct a reasonable joint probability distribution. Additionally, it is not appropriate to treat the diagnosis target as the same feature as other diseases: We pay the most attention to the diagnosis target, while other diseases features form the basis of our judgement.

4.4. DRBM Based CF for Clinical Disease Diagnosis

The DRBM-based CF model, with the following (model and optimization) hyper-parameters settings. For both of the two datasets, the number of iterations *n_step* for model training of DRBM is set to 30-time, 10 hidden layers and an L1 regularization model, step number of sampling *cd_k* is set to 1, and learn rate (*lr*) is set to 0.05. The experimental results based on the unsupervised disease diagnosis model of DRBM are shown in Table 5. The DRBM-based CF model can predict CKD, with a prediction accuracy of more than 0.968, even in the case of severe missing data. The DRBM-based CF correctly learns the given features and can synthetically prognosticate disease. The prediction accuracy of the DRBM-based model for the dermatology dataset decreases to a small extent with an increase in missing data, indicating that the model has learned the correlations of various disease information well, and can still make accurate predictions based on other features, even without critical disease features.

Table 5. Best results of the DRBM-based CF model for each dataset.

Missingness	Baseline	10%	20%	30%
F1-score of dermatology	0.964	0.953	0.941	0.903
F1-score of CKD	0.996	0.982	0.978	0.968

4.5. Experimental Result and Comparative Analysis

For the traditional supervised learning model, we first use the *R* language package VIM and MICE to achieve the missing value imputation and then standardize the data with sklearn and use each model in Python to make predictions. The parameters we set are *R* language package defaults for traditional machine learning. Uppercase means machine learning method, lowercase means interpolation (See Appendix A for details). The results are shown in Table 6. As shown in the table, the prediction performance of the traditional machine-learning model for dermatology datasets is very poor. For multiple categories of dermatology datasets, combining the methods of machine learning with the technology of missing value imputation does not result in better predictive performance. However, for the diagnosis of CKD, the traditional machine learning method combined with the missing value imputation method can achieve a better predictive performance.

Table 6. Best results of datasets based on traditional machine learning.

Missingness	Baseline	10%	20%	30%
Accuracy of dermatology	0.027 (SVM-cart)	0.108 (RF-cart)	0.064 (RF-rf)	0.109 (LR-rf)
F1-score of dermatology	0.035 (MLP-pmm)	0.066 (LR-cart)	0.053 (RF-rf)	0.087 (SVM-rf)
Accuracy of CKD	0.987 (RF-mean)	0.964 (LR-cart)	0.975 (RF-cart)	0.966 (LR-pmm)
F1-score of CKD	0.987 (RF-mean)	0.963 (RF-Mean)	0.975 (RF-cart)	0.964 (LR-pmm)

All the experimental results were analysed, the comparative analysis of various classification methods, with a baseline data and various level of missing data, are shown in Figure 8. For dermatology data, the classification methods of user-based CF and CF, based on DRBM, are obviously superior to other machine learning methods in terms of accuracy rate. For chronic kidney disease dataset, the above two methods are equally outperform other methods. Whereas, for the problem of classification and discrimination discussed in this paper, CF based on RBM is slightly inferior than DRBM. Compared with the disease diagnosis performance of the traditional machine-learning model, we conclude that both the user-based CF diagnosis model and the DRBM-based CF diagnosis model can accurately predict multiple complex diseases in the case of severe missing data. Compared with the user-based CF method, the DRBM model deeply learns the complex associations between the various features can maintain its good prediction performance in certain situations of severe missing. Our results demonstrate that the DRBM-based CF method is consistently superior to traditional machine learning with different imputations of real data, and it is a preferable means for solving the auxiliary diagnosis problem of clinical diseases, to some extent, with a large amount of missing data.

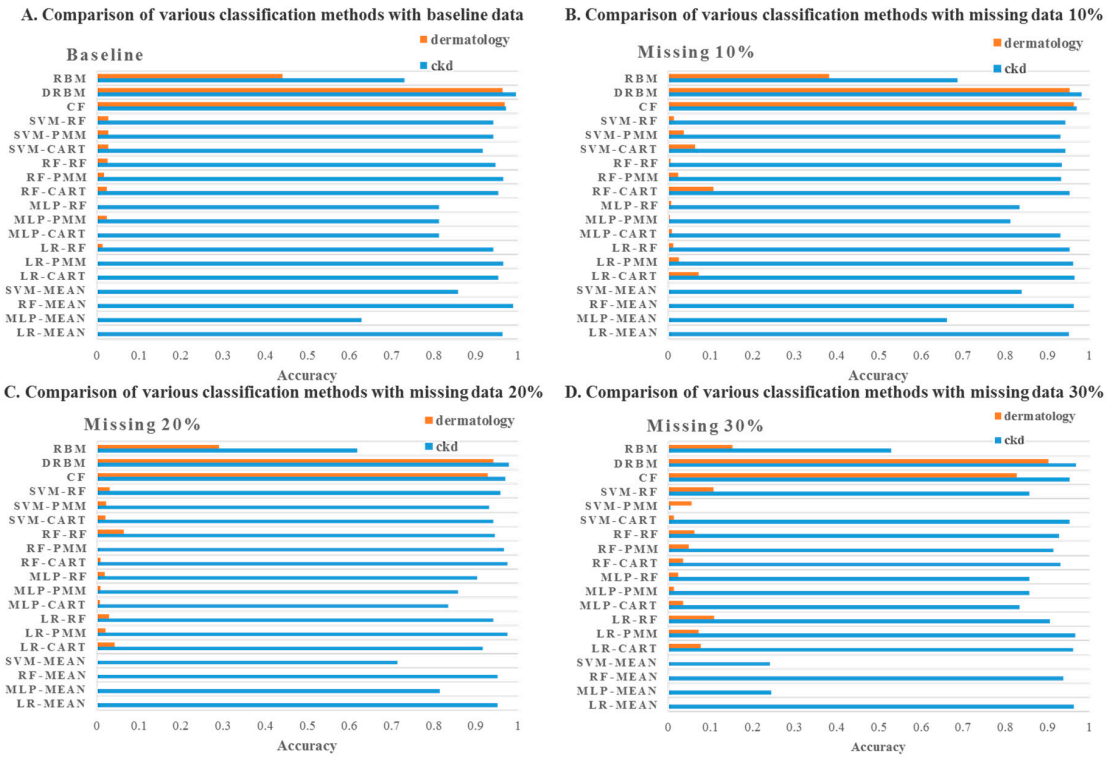


Figure 8. Comparison of traditional machine-learning classification methods with baseline data and various level of missing data.

We also conducted comparative experiments, with up-to-date classification methods on baseline data, such as a principal components analysis (PCA) [57], Naïve Bayes (NB) [58] and non-negative matrix factorization (NMF) [59] , and adopted *sklearn* machine learning package to implement these algorithms. The experimental results are shown in Figure 9. As it can be seen from the figure, for dermatology data, the highest performance figures were obtained by using user-based CF. Overall accuracy rate of user-based CF is 0.969. DRBM model also resulted in high accuracy rate of 0.964. The overall classification accuracy rate of NB, PCA, and NMF are 0.962, 0.953, and 0.960, respectively. For the CKD datasets, the DRBM models outperform others, with an accuracy rate of 0.996. The overall classification accuracy rate of user-based CF, NB, PCA, and NMF are 0.972, 0.983, 0.975, and 0.989, respectively. Our results also demonstrate that DRBM model is consistently superior to the other classification methods, with real clinical datasets, and it is a preferable means for solving the auxiliary diagnosis problem of clinical diseases.

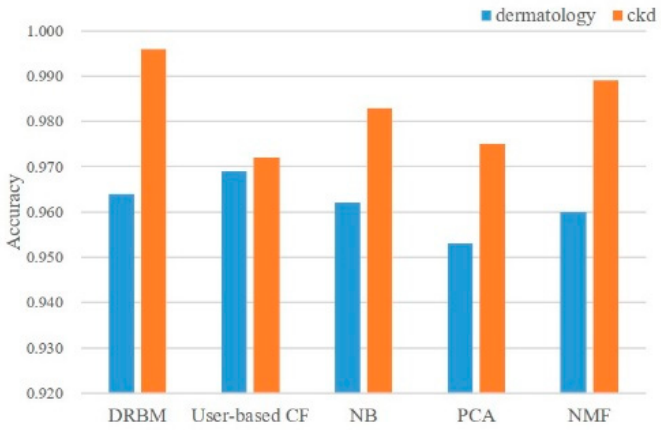


Figure 9. Comparative analysis of different classification methods with DRBM and user-based CF in terms of accuracy on baseline data.

5. Discussion

With the advent of the medical era of big data, ubiquitous recommender systems, based on CF, are widely used to solve forecasting and recommendation problems in clinical medical research. However, many open research challenges remain in the area of clinical medicine, especially clinical decisions and classification prediction. Data integration is also a challenge for Big Data and the translation of data to knowledge [60]. Existing research methods and Web-based frameworks for medical decision-making has led the future of research [61]. Compared with existing methods in related work, we are limited in the matter of data size and paradigm from clinical diagnosis and medical research. For our study, this owes to the lack of accessibility to medical databases, most of which are generally not publicly available. Although the present data exploration is on a small scale, according to the experimental results, we have demonstrated that the categorical outcome of clinical diagnostic datasets can be modelled as classification problems, and could be dealt with CF methods, which is superior to the traditional machine learning methods.

The research purpose of this article was to examine DRBM-based CF and user-based CF, compared with traditional machine-learning methods, using real clinical datasets with category outcomes. Our theory of intrinsic motivation concern on more comprehensive understanding of how recommender systems apply to clinical diagnosis and clinical data, and how the gradually increased missing data influences performance. To solve these problems, we present CF, based on latent features and user-based CF, using default real clinical data with various simulation of missing data. We employed two different publicly available datasets, one for CKD and one for dermatology. Both of them have different background and property, but each had an outcome and could be viewed as a discrimination and classification problem. Our simulation experiment consists of cleaning, division, creation of

missing data, discretization, imputation, and performance evaluation. Our simulation approach has demonstrated that user-based CF recommender systems can be seamlessly applied in clinical diagnosis and possess good prediction performance. Generally, there is no solution appropriate for all classification problems. In this paper, we selected commonly used classification methods from machine learning, such as SVMs, MLP, LR, RFs, PCA, NB, and NMF. Our primary purpose is not only to compare the performance of the various classification methods, but also to evaluate the traditional CF-based recommender system, and whether it is appropriate for clinical data. We also proposed a classification solution, based on latent features, namely RBM. The RBM is a generative latent variable model that models the joint distribution of a set of input variables. It has gained popularity over the past decade in many applications, including feature learning, CF, high-dimensional sequence modelling, and pre-training deep neural network classifiers. One of its applications is as a standalone classifier, referred to as the DRBM. As the name might suggest, the DRBM is a classifier obtained by carrying out discriminative learning in the RBM, and it directly models the conditional distribution of interest for predictions. Through the experiments, we came to the conclusion that clinical data forecasting and classification, using DRBM-based methods, will achieve a better predictive accuracy, compared with traditional machine learning classification methods.

Our experimental method exposed a weakness in DRBM-based CF recommender systems are limited to a narrow range for prediction missing data in the clinical environment, but is not without limitations. We only adopted support vector machines, random forest, logistic regression, and multi-layer perceptron, but there are also some machine learning methods suitable for a medical scenario. Except for the mean value imputation method, we also used multiple imputations by chained equations method for missing values. Different imputation methods, in various environments, have different effects. Our target is to evaluate and optimize the performance of recommender algorithms in the area of clinical disease diagnosis, rather than to compare classifier and interpolation methods. Another limitation of our work is the missing degree and pattern of data. The data creation in our simulation are missing as random, however in the reality, this may not be the case, especially in clinical medical database. The algorithm proposed in this paper is equally applicable to similar problems in other fields. The algorithm can solve the problem of missing features in different fields.

6. Conclusions

In this paper, our experiments consistently demonstrated that CF can be effectively used in clinical decisions and risk prediction. We observed empirical studies and proved that the proposed DRBM-based CF method performs much better than traditional machine learning methods in real clinical datasets, with various levels of missing data situations, and clinical data, with different imputations, and exhibit excellent performance for predicting missing values and result classifications. Scalability and universality for recommender systems are currently practical challenges, we proved that the DRBM-based CF as a preferred solution for classification problems, when the data size can be suitable for the traditional machine learning classification method. The DRBM-based CF method proposed in this paper has a sound basis and should be used as a stand-alone non-linear classifier in any domain, rather than being considered a simple feature extractor, especially in clinical medical field, which demands stable solutions. In clinical medical decision-making, by reconstructing patients' medical records and accurately forecasting patients' category outcomes, doctors can reduce the rate of misdiagnosis by utilizing patients' complete information; thereby improving the efficiency of patient treatment. The method proposed in this paper is equally applicable to similar problems in other fields. In future work, we would like to investigate the use of variations versions of RBMs in more challenging settings, such as in multitask or structured output problems. In order to promote precision medical application, the next step will continue to be combined with scientific research and clinical application. Likewise, we also consider introducing massive heterogeneous clinical medical datasets for large-scale data processing and classification recommendation.

Author Contributions: Conceptualization, M.S. and T.M.; data curation, T.M.; formal analysis, M.S.; funding acquisition, T.Z. and Y.W.; investigation, M.S.; methodology, M.S. and T.M.; project administration, T.Z. and Y.W.; resources, T.M.; supervision, T.Z. and Y.W.; validation, M.S.; visualization, T.M.; writing—original draft, M.S.; writing—review and editing, T.Z.

Funding: This work was supported in part by the National Key Research and Development Programs of China under Grants 2016YFC0901605 and 2016YFC1201702-01 and in part by the National High-Tech Research and Development Program of China (863 Program) under Grants 2012AA02A601, 2015AA020101, and 2015AA020108.

Acknowledgments: The experimental datasets used in this paper are derived from UCI Machine Learning Repository. Thanks for the assistance of the repository. See the reference below for details. (Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.)

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

RBM	Restricted Boltzmann Machine
DRBM	Discriminative Restricted Boltzmann machine
SVM	Support Vector Machines
RF	Random Forest
LR	Logistic Regression
MLP	Multi-layer Perceptron
Mean	Mean Imputation
kNN	k-nearest neighbors
cart	Classification and Regression Tree
pmm	Predictive Mean Matching Imputation
rf	Random Forest Imputation
MICE	Multivariate imputation by Chained equations
MAR	Missing at random
MCAR	Missing completely at random
RecSys	Recommender systems
RF-Mean	Random forest with Mean imputation
RF-kNN	Random forest with kNN imputation
RF-cart	Random forest with cart imputation
RF-pmm	Random forest with pmm imputation
RF-rf	Random forest with rf imputation

References

1. Musen, M.A.; Middleton, B.; Greenes, R.A. Clinical Decision-Support Systems. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*; Shortliffe, E.H., Cimino, J.J., Eds.; Springer: London, UK, 2014; pp. 643–674.
2. Berner, E.S.; La Lande, T.J. Overview of Clinical Decision Support Systems. In *Clinical Decision Support Systems: Theory and Practice*; Berner, E.S., Ed.; Springer: New York, NY, USA, 2007; pp. 3–22.
3. Newman-Toker, D.E.; Pronovost, P.J. Diagnostic errors—The next frontier for patient safety. *JAMA* **2009**, *301*, 1060–1062. [CrossRef]
4. Isa, I.S.; Saad, Z.; Omar, S.; Osman, M.K.; Ahmad, K.A.; Sakim, H.A.M. Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection. In Proceedings of the 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, Tuban, Indonesia, 28–30 September 2010; pp. 39–44.
5. Shen, L.; Chen, H.; Yu, Z.; Kang, W.; Zhang, B.; Li, H.; Yang, B.; Liu, D. Evolving support vector machines using fruit fly optimization for medical data classification. *Knowl. Based Syst.* **2016**, *96*, 61–75. [CrossRef]
6. Cawley, G.C.; Talbot, N.L.C. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* **2006**, *22*, 2348–2355. [CrossRef]
7. Chen, X.; Ishwaran, H. Random forests for genomic data analysis. *Genomics* **2012**, *99*, 323–329. [CrossRef]

8. King, G.; Honaker, J.; Joseph, A.; Scheve, K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *Am. Polit. Sci. Rev.* **2002**, *95*, 49–69.
9. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749. [CrossRef]
10. Yang, X.; Guo, Y.; Liu, Y.; Steck, H. A survey of collaborative filtering based social recommender systems. *Comput. Commun.* **2014**, *41*, 1–10. [CrossRef]
11. Zhang, N.; Ding, S.; Zhang, J.; Xue, Y. An overview on Restricted Boltzmann Machines. *Neurocomputing* **2018**, *275*, 1186–1199. [CrossRef]
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef]
13. Larochelle, H.; Bengio, Y. Classification using discriminative restricted Boltzmann machines. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 536–543.
14. Cherla, S.; Tran, S.N.; d'Avila Garcez, A.; Weyde, T. Generalising the Discriminative Restricted Boltzmann Machines. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2017; pp. 111–119.
15. Srivastava, N.; Salakhutdinov, R. Multimodal learning with deep Boltzmann machines. *J. Mach. Learn. Res.* **2014**, *15*, 2949–2980.
16. Srivastava, N.; Salakhutdinov, R.; Hinton, G. Modeling documents with a Deep Boltzmann Machine. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*; AUAI Press: Bellevue, WA, USA, 2013; pp. 616–624.
17. Li, X.; Zhao, F.; Guo, Y. Conditional Restricted Boltzmann Machines for Multi-label Learning with Incomplete Labels. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 635–643.
18. Maxwell, A.; Li, R.; Yang, B.; Weng, H.; Ou, A.; Hong, H.; Zhou, Z.; Gong, P.; Zhang, C. Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinform.* **2017**, *18*, 523. [CrossRef]
19. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504. [CrossRef]
20. Salakhutdinov, R.; Mnih, A.; Hinton, G. Restricted Boltzmann machines for collaborative filtering. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, OR, USA, 20–24 June 2007; pp. 791–798.
21. Eickholt, J.; Cheng, J. Predicting protein residue–residue contacts using deep networks and boosting. *Bioinformatics* **2012**, *28*, 3066–3072. [CrossRef]
22. Bell, R.M.; Koren, Y. Lessons from the Netflix prize challenge. *SIGKDD Explor. Newsl.* **2007**, *9*, 75–79. [CrossRef]
23. John, A.; Muhammed Ilyas, H.; Vasudevan, V. Medication recommendation system based on clinical documents. In Proceedings of the 2016 International Conference on Information Science (ICIS), Kochi, India, 12–13 August 2016; pp. 180–184.
24. Felix, G.; Stefanie, B.; Denise, K.; Jochen, S.; Susanne, A.; Hagen, M.; Sebastian, Z. Therapy Decision Support Based on Recommender System Methods. *J. Healthc. Eng.* **2017**, *2017*. [CrossRef]
25. Folino, F.; Pizzuti, C. A recommendation engine for disease prediction. *Inf. Syst. e-Bus. Manag.* **2015**, *13*, 609–628. [CrossRef]
26. Hu, Y.; Koren, Y.; Volinsky, C. Collaborative Filtering for Implicit Feedback Datasets. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 263–272.
27. Jannach, D.; Lerche, L.; Zanker, M. Recommending Based on Implicit Feedback. In *Social Information Access: Systems and Technologies*; Brusilovsky, P., He, D., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 510–569.
28. Pan, R.; Zhou, Y.; Cao, B.; Liu, N.N.; Lukose, R.; Scholz, M.; Yang, Q. One-Class Collaborative Filtering. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 502–511.
29. Pan, R.; Scholz, M. Mind the gaps: Weighting the unknown in large-scale one-class collaborative filtering. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 667–676.

30. Paquet, U.; Koenigstein, N. One-class collaborative filtering with random graphs. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 999–1008.
31. Koren, Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, 24–27 August 2008; pp. 426–434.
32. Mnih, A.; Salakhutdinov, R.R. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*; Mit Press: Cambridge, MA, USA, 2008; pp. 1257–1264.
33. Jamali, M.; Lakshmanan, L. HeteroMF: Recommendation in heterogeneous information networks using context dependent factor models. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 643–654.
34. Karatzoglou, A.; Baltrunas, L.; Shi, Y. Learning to rank for recommender systems. In Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, China, 12–16 October 2013; pp. 493–494.
35. Cao, B.; Hou, C.; Peng, H.; Fan, J.; Yang, J.; Yin, J.; Deng, S. Predicting e-book ranking based on the implicit user feedback. *World Wide Web* **2019**, *22*, 637–655. [CrossRef]
36. Huang, J.; Wang, J.; Yao, Y.; Zhong, N. Cost-sensitive three-way recommendations by learning pair-wise preferences. *INT J. Approx. Reason.* **2017**, *86*, 28–40. [CrossRef]
37. Shi, Y.; Karatzoglou, A.; Baltrunas, L.; Larson, M.; Hanjalic, A.; Oliver, N. TFMAP: Optimizing MAP for top-n context-aware recommendation. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, Oregon, USA, 12–16 August 2012; pp. 155–164.
38. Oord, R.V.D.; Dieleman, S.; Schrauwen, B. Deep content-based music recommendation. In Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2, Lake Tahoe, Nevada, 5–10 December 2013; Curran Associates Inc.: New York, NY, USA; pp. 2643–2651.
39. Yao, W.; He, J.; Huang, G.; Cao, J.; Zhang, Y. A Graph-based model for context-aware recommendation using implicit feedback data. *World Wide Web* **2015**, *18*, 1351–1371. [CrossRef]
40. Tomczak, J. Application of classification restricted boltzmann machine to medical domains. *World Appl. Sci. J.* **2014**, *31*, 69–75.
41. Luo, X.; Zhou, M.; Xia, Y.; Zhu, Q. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems. *IEEE Trans. Ind. Inform.* **2014**, *10*, 1273–1284.
42. Luo, X.; Zhou, M.; Li, S.; Shang, M. An Inherently Nonnegative Latent Factor Model for High-Dimensional and Sparse Matrices from Industrial Applications. *IEEE Trans. Ind. Inform.* **2018**, *14*, 2011–2022. [CrossRef]
43. Luo, X.; Zhou, M.; Li, S.; You, Z.; Xia, Y.; Zhu, Q. A Nonnegative Latent Factor Model for Large-Scale Sparse Matrices in Recommender Systems via Alternating Direction Method. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 579–592. [CrossRef]
44. Luo, X.; Zhou, M.; Xia, Y.; Zhu, Q.; Ammari, A.C.; Alabdulwahab, A. Generating Highly Accurate Predictions for Missing QoS Data via Aggregating Nonnegative Latent Factor Models. *IEEE Trans. Neural Netw. Learn.* **2016**, *27*, 524–537. [CrossRef]
45. Luo, X.; Zhou, M.; Li, S.; Xia, Y.; You, Z.; Zhu, Q.; Leung, H. Incorporation of Efficient Second-Order Solvers Into Latent Factor Models for Accurate Prediction of Missing QoS Data. *IEEE Trans. Cybern.* **2018**, *48*, 1216–1228. [CrossRef]
46. Hao, F.; Blair, R.H. A comparative study: Classification vs. user-based collaborative filtering for clinical prediction. *BMC Med. Res. Methodol.* **2016**, *16*, 172. [CrossRef]
47. Dua, D.A.K.T. Efi {UCI} Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml (accessed on 14 February 2019).
48. Ravindra, B.; Sriraam, N.; Geetha, M. Classification of non-chronic and chronic kidney disease using SVM neural networks. *Int. J. Eng. Technol.* **2018**, *7*, 191–194.
49. Subasi, A.; Alickovic, E.; Kevric, J. Diagnosis of Chronic Kidney Disease by Using Random Fores. In *CMBEBIH 2017. IFMBE Proceedings, Singapore, 2017*; Badnjevic, A., Ed.; Springer: Singapore, 2017; pp. 589–594.
50. Khanna, D.; Sahu, R.; Baths, V.; Deshpande, B. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. *Int. J. Mach. Learn. Comput.* **2015**, *5*, 414. [CrossRef]
51. Yildirim, P. Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction. In Proceedings of the 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, Italy, 4–8 July 2017; pp. 193–198.

52. Avram, A. TensorFlow: Google Open Sources Their Machine Learning Tool. Available online: https://www.infoq.com/news/2015/11/tensorflow (accessed on 14 February 2019).
53. John Chambers: The R Project for Statistical Computing. Available online: https://www.r-project.org (accessed on 14 February 2019).
54. Larochelle, H.; Mandel, M.; Pascanu, R.; Bengio, Y. Learning algorithms for the classification restricted Boltzmann machine. *J. Mach. Learn. Res.* **2012**, *13*, 643–669.
55. Little, R.J.A. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J. Am. Stat. Assoc.* **1988**, *83*, 1198–1202. [CrossRef]
56. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–68. [CrossRef]
57. Luukka, P. A New Nonlinear Fuzzy Robust PCA Algorithm and Similarity Classifier in Classification of Medical Data Sets. *Int J. Fuzzy Syst.* **2011**, *13*, 153–162.
58. Dulhare, U.N.; Ayesha, M. Extraction of action rules for chronic kidney disease using Naïve bayes classifier. In Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, India, 15–17 December 2016; pp. 1–5.
59. Li, G.; Ou, W. Pairwise probabilistic matrix factorization for implicit feedback collaborative filtering. *Neurocomputing* **2016**, *204*, 17–25. [CrossRef]
60. Margolis, R.; Derr, L.; Dunn, M.; Huerta, M.; Larkin, J.; Sheehan, J.; Guyer, M.; Green, E.D. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: Capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 957–958. [CrossRef]
61. Yao, J.; Azam, N. Web-Based Medical Decision Support Systems for Three-Way Medical Decision Making With Game-Theoretic Rough Sets. *IEEE T Fuzzy Syst.* **2015**, *23*, 3–15. [CrossRef]