HEALTH RISK PREDICTION USING BIG MEDICAL DATA - A COLLABORATIVE FILTERING-ENHANCED DEEP LEARNING APPROACH

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Xin Li

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

May 2018

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

HEALTH RISK PREDICTION USING BIG MEDICAL DATA - A COLLABORATIVE FILTERING-ENHANCED DEEP LEARNING APPROACH

By

Xin Li

The supervisory committee certifies that this paper complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Prof. Juan Li
Chair

Prof. Na Gong

Prof. Simone Ludwig

Approved:

30 May 2018
Date

Prof. Kendall Nygard
Department Chair

ABSTRACT

Deep learning has yielded immense success on many different scenarios. With the success in other real world application, it has been applied into big medical data. However, discovering knowledge from these data can be very challenging because they normally contain large amount of unstructured data, they may have lots of missing values, and they can be highly complex and heterogeneous. In these cases the deep neural network itself is not applicable enough. To solve these problems we propose a Collaborative Filtering-Enhanced Deep Learning Approach. In particular, first we estimate missing values based on the information mining from the structured and unstructured data. Secondly, a deep neural network-based method is applied, which can help us handle complex and multi-modality data. The proposed algorithm is applied to analyze big medical data and make personalized health risk prediction. Extensive experiments on real-world datasets show improvements of our proposed algorithm over the state-of-the-art methods.

ACKNOWLEDGEMENTS

My advisor Dr. Li taught me a lot about research and her support is really appreciated. And also thank my co-advisor Dr Gong for helping me all along the way of research and life.

Thanks for my committee members, Dr Ludwig, for support and guidance. All of you have contributed your ideas, guidance to do better with my paper.

At the same time, I also wish to express my gratitude to all the faculty and staff in the computer science department. You really do your work very well, I have learnt a lot from you.

There are still a lot I wish I could show my thankfulness one by one, my co-workers, friends and families.

For all who have encouraged me to keep going, comforted when in low spirits, provided suggestions when confused or help when it was in need, your help is precious to me.

This work was partially supported by the National Science Foundation (NSF) under Div. of Information and Intelligent Systems (IIS) with award number: 1722913.

DEDICATION

This thesis is dedicated to my father Xiyuan Li and my mother Shaoqing Yuan.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENTS iv

DEDICATION v

LIST OF TABLES vii

LIST OF FIGURES viii

1. INTRODUCTION 1

2. RELATED WORK 4

3. METHODOLOGY 6

3.1. Problem formulation 6

3.2. Identify missing data 6

3.3. Deep neural network training and prediction process 10

3.4. Algorithm design and implementation 11

4. EXPERIMENTS 13

4.1. Datasets 13

4.2. Pre-processing 13

4.3. Training and testing 14

4.4. Result and discussion 15

5. CONCLUSION 17

REFERENCES 18

LIST OF TABLES

Table Page

4.1. List of feature descriptions 14

4.2. Performance measurements for different prediction methods - original dataset 16

4.3. Performance measurements for different prediction methods - 20% missing data 16

LIST OF FIGURES

Figure Page

3.1. The deep neural network. 10

4.1. Confusion matrix 16

1. INTRODUCTION

Deep learning (DL, also known as deep structured learning, hierarchical learning or deep machine learning) is a branch of machine learning based on a set of algorithms that attempts to model high level abstractions in data. Deep neural network (DNN) has been applied to a bunch of prediction and classification applications.

In many real world problems, however, one drawback of the deep neural network is that it requires a full set of input data, and real world data is sometimes incomplete. Missing values in input data are a particular problem in many fields and this phenomenon is of particular consequence in many applications. In this case, the ability to learn from data with uncertain or missing information is a fundamental requirement for DNN.

Collaborative filtering (CF) is a typical way to find out the hidden value in the system. CF is especially used in recommendation system to analyze relationships between users and inter-dependencies among products or items to identify new user-item associations. The fundamental assumption behind CF is that other users preference can be selected and aggregated to provide a reasonable prediction for the active user. CF has different ways like matrix factorization, probability matrix factorization and Bayesian probability matrix factorization.

In this paper, we try to combine the Collaborative filtering and Deep neural network together to form an approach for the applications which have missing data. In the proposed approach, the CF is firstly introduced to mine the hidden value for the missing data. To overcome the data sparsity problem which conventional CF-based methods may suffer, auxiliary information such as the structured data and info stored in the unstructured data (patient's demographic data, social economic data) will be utilized. This information will help us to uncover unexpected relationships. We choose collaborative topic regression (CTR) [16] approach to tightly couple the patients factor and health factor. CTR is a probabilistic model combining topic model and probabilistic matrix factorization (PMF). Since the CTR can utilize the hidden information in the structured data as well as unstructured data and estimate missing data based on patients' similarity, we integrate the CTR into the deep learning to enhance the efficiency and accuracy of deep neural network. By

integrating the combined strength of CRT-based CF with deep learning, we expect to overcome the shortcomings of existing approaches and provide a more accuracy result.

In this paper, we apply our proposed methodology to a real-life dataset to predict risk of hospital readmission for diabetic patients. Diabetes is one of the most common and costly chronic diseases. It is estimated that 23.1 million people in the U.S. are diagnosed with diabetes at a cost of more than $245 billion per year[1]. Hospital inpatient care accounts for the largest proportion of medical expenses of diabetes care are(about 43%) [11]. With about 25% patients being readmitted within 30 days of discharge [3], unplanned sessions have become a serious issue. For lots of hospitals, diabetic patients' readmission is becoming one of the biggest concerns. Therefore,it is very important to study the possibility and risks of diabetes patients' readmission. To save hospital cost, diabetic patients' readmission should be given high-priority. This has made healthcare professionals, scientists, and policymakers increasingly focus on the readmission rates to determine the complexity of patient populations, prepare for procedures and interventions, and eventually improve healthcare quality and reduce cost. Although there is increasing interest in hospital readmission rate, not much research effort has been applied to study the readmission rate of diabetes patients.

As diabetes patient readmission rate is becoming one of the major concerns for many national hospitals in the U.S., it is of great significance to study the possibility and risks of diabetes patients readmission. Hospital readmission is a high-priority health care quality measure and target for cost reduction, particularly readmission rate within 30 days of discharge (30-day readmission, aka early readmission). Despite the broad interest in readmission, relatively little research has focused specifically on readmission of patients with diabetes. In recent years, government agencies and healthcare systems have increasingly focused on 30-day readmission rates to determine the complexity of their patient populations and to improve quality.

This paper makes the following five major contributions:

1. It proposes a Collaborative Filtering-Enhanced Deep Learning Approach (CFDL) which can combine the collaborative filtering with deep learning.

2. It provides a way for how to utilize the structured and unstructured data to find the potential value if the input data of deep learning neural network is incomplete.

3. It discusses a new method for the prediction in the diabetes patient readmission rate.

4. With the application of the proposed approach in a real diabetes patient readmission rate, it demonstrate the performance boosting with the introduction of proposed approach. Simulation results show that the proposed algorithm can achieve good performance.

5. The proposed algorithm is compared with other intelligent methods like Support Vector Machine, Decision Tree and Naive Bayes.

Although this paper's experiments focus on diabetes case, we believe that the proposed methodology would be general enough to be used more broadly.

2. RELATED WORK

Nowadays the healthcare sector has generated a huge volume of patient data. Machine learning provides a way to automatically find patterns and make predictions about data. There are many recent works on data mining and data analytics with different types of digital patient data.

To deal with the complex feature selection problem, deep learning has been used as an analysis approach. As pointed out in [18], deep learning has been applied in medical data to automatically select and generate complex features from the input data. For example, Cheng et al. [6] proposed a deep learning-based prediction model using Electronic Health Records (EHRs). In this model, EHRs for patients were represented as a matrix with multiple dimensions including time and event. They employed a convolutional neural network (CNN) of four layers to extract features and assess risks.

In another work, to make healthcare decisions Liang et al. [8] applied a deep learning model to EHRs database to enhance feature representations. They proposed an approach that integrated both unsupervised and supervised learning. Specifically, they applied deep belief network (DBN) to extract complex features, and then performed support vector machine (SVM). In their work, Nie et al. [15] proposed a deep learning scheme to infer people's possible diseases given the questions of that people asks online. The proposed scheme constructed a deep neural network with three hidden layers which are connected sparsely. These three layers were constructed through an iterative process by alternating signature mining and pre-training. At first, medical signatures were discovered from raw features. Then the raw features and their signatures were passed to one layer as input nodes and the next layer as hidden nodes. This process repeated until the model was well tuned. The relations between these two layers can be learned from this process. More applications of deep learning can be found in medical informatics and public health domains [5, 9, 12, 13, 19].

Various methods have been studied to address the data sparsity problem of medical data. For example, to process sparse and non-vector input data, Wang et al. [20] proposed a high order extension of sparse logistic regression model, MulSLR, (for Multilinear Sparse Logistic Regression) to predict clinical risk. Unlike conventional logistic regressions, their approach solved K classification vectors.

To determine patient acuity using incomplete, sparse and heterogeneous clinical data, Ghassemi et al. [7] proposed an approach that transformed this clinical data into a new latent space using the hyperparameters of multi-task GP (MTGP) models. In this way, patients can be compared based on their similarity in the new hyperparameter space. Information in this hyperparameter space could be viewed as timeseries data, and abstracted features can represent the series dynamics. This approach has been approved to increase classification performance on mortality prediction of ICU patients, however, the computational cost of this approach was very high.

Lipton et al. [10] proposed an approach to model missing clinical data using recurrent neural networks (RNNs). Unlike classical approaches that treat missing data via heuristic imputation, in their approach, the authors modeled missingness as a feature. The proposed RNN can use only simple binary indicators for missingness.

GRU-D [4], a deep learning models, was developed to exploit the missing patterns of missing data for effective imputation and improving prediction performance. GRU-D was developed based on the Gated Recurrent Unit (GRU) which was a recurrent neural network. GRU-D took masking and time interval as two representations of missing patterns, and then integrated them into a deep model architecture. This model can capture the long-term temporal dependencies in time series. In addition, it can estimate the missing value patterns to achieve better prediction results.

Some artificial intelligent techniques have been applied to the identification and prediction of diabetes disease. For example, Mario et al.[14] proposed the modeling analysis of the radial basis function network (RBF Network) to identify possible cases of gestational diabetes that can lead to multiple risks for both the pregnant women and the fetus. The paper in [17] investigated the different roles of pixels in the image-level predictions of ConvNets when applied to the diabetic retinopathy (DR) screening dataset.

In spite of the numerous efforts and achievements of existing research in analyzing digital medical data, analyzing medical data is still an important and challenging task. Accurate and efficient risk prediction has always been an important topic attracting many researchers' interests.

3. METHODOLOGY

In this section we present the detailed methodology of our proposed Collaborative Filtering-enhanced Deep Learning (CFDL) approach.

3.1. Problem formulation

Let $\mathcal{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_P, y_P)\}$ be the data set, where P is the number of patients; $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th instance; and $y_i \in \{1, \ldots, Q\}$ represents its label, where $Q \geqslant 2$ is the number of classes. $(\mathbf{x}_i)_j$ is the j-th feature of patient i; for example, it can be one of a patient's demographic features, or disease symptoms, or vital signs. y_i represent the target label, for example, it can be a particular health risk.

To model the missing data, we divide the data set $\mathcal{X}$ into an observed component $\mathcal{X}^o$ and a missing component $\mathcal{X}^m$. Similarly, for some data vector, $\mathbf{x}_i$ is divided into $(\mathbf{x}_i^o, \mathbf{x}_i^m)$ where these data vector may have different missing components. Note that $\mathbf{x}_i^m$ should not be a subset of $\mathbf{x}_i^o$. For some data vector, the whole data instance is lost, hence $\mathbf{x}_i = \mathbf{x}_i^m$.

To model unstructured data, in our case text data (such as doctor's notes, open-question survey), we can model them as a corpus D of M documents $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ consisting of unstructured data, with a vocabulary of size V.

Given these data, we have two goals. Firstly, we aim to learn the missing data of $\mathcal{X}^m$ from the observed data $\mathcal{X}^o$ and unstructured data of M documents $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$. Secondly, we want to predict the estimated label for the dataset. To realize these two goals, we propose a two-stage methodology presented in the following subsections.

3.2. Identify missing data

The goal of this step is to learn the missing data $\mathcal{X}^m$ from the existing observed and unstructured data. The main idea is that if an important feature is missing for a particular instance, it can be estimated from similar data that are present. Consider we have the unstructured data, we hypothesize that there is rich information within the unstructured data, which can provide additional source of information for the estimation.

We apply the topic model on the unstructured data of M documents. To do this, we assume there exists a fixed number of latent topics that appear across multiple documents. Each topic is

characterized by a multinomial distribution over the vocabulary of the corpus D, drawn from a Dirichlet distribution denoted as $\phi_k \sim Dir(\beta)$. Each document is characterized by a multinomial distribution over the set of topics in the corpus, which is also assumed to have a Dirichlet prior denoted as $\theta_j \sim Dir(\alpha)$. The topic distribution has the following probability density:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{3.1}$$

where the parameter α is a k-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function.

We follow the CF-based recommendation algorithm, the matrix of patient and health-related features can be decomposed by Matrix Factorization (MF), by denoting a set of values correspond to a set of patients with index $i \in \{1, 2, \ldots, N\}$ to a set of health factors with index $j \in \{1, 2, \ldots, M\}$, each entry of the value X_{ij} can be expressed as the inner product of a patient matrix and a health factor matrix:

$$X_{ij} \approx < U_i, V_j > \equiv \sum_{k=1}^{K} U_{ik} V_{kj} \tag{3.2}$$

where U_i , V_j represent the k-dimensional patient-specific and health factor-specific latent feature vectors of patient i and health factor j, respectively.

Collaborative topic regression (CTR) additionally molds the health factor vector as the combining of a latent variables that offsets the topic proportion. According to CTR, for each v_j

$$v_j = \varepsilon_j + \theta_j \tag{3.3}$$

where θ_j is the topic proportion computed by latent dirichlet allocation (LDA), therefore the values can be expressed as:

$$X_{ij}^t \approx < U_i, \varepsilon_j + \theta_j > \equiv \sum_{k=1}^{K} U_{ik}(\varepsilon_{kj} + \theta_{kj}) \tag{3.4}$$

CTR places prior distribution on $\mathbf{U}$, $\mathbf{V}$. Specifically zero-mean, independent Gaussian priors are imposed on patient and health factor vector:

$$\begin{aligned} U_i &\sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}),\ \mathrm{i} = 1 \ldots \mathrm{N}, \\ V_j &\sim \mathcal{N}(0, \sigma_v^2 \mathbf{I}),\ \mathrm{j} = 1 \ldots \mathrm{M} \end{aligned} \tag{3.5}$$

where $\mathbf{I}$ is the $D-$by$-D$ identity matrix.

The conditional distribution over the observed ratings and the prior distributions are given by

$$X_{ij}|\mathbf{U}, \mathbf{V} \sim \mathcal{N}(\ U_i,\ \varepsilon_j + \theta_j,\ \alpha^{-1}) \tag{3.6}$$

In this way, the original complex matrix can be decomposed into simpler computations involving the corresponding patient related matrix and health factor related matrix, as well as the topic proportions.

CTR combines the matrix factorization with the topic modeling to collaboratively predict values and to learn topics. After we have trained our LDA implementation on a separate training corpus and learned the model parameters α and β, the whole model can be learned by maximizing the log-posterior distribution, which takes the following form assuming ratings are made independently conditioned on latent factors:

$$\begin{aligned} &\ln p(U, V|X, \sigma_X^2, \sigma_U^2, \sigma_V^2) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij}^t \ln p(X_{ij}^T|U_i, V_j) + \sum_{i=1}^{N} \ln p(U_i) \\ &+ \sum_{j=1}^{m} \ln p(V_j - \theta_j) \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{I_{ij}^t (X_{ij}^T - < U_i, V_j >)^2}{2\sigma_X^2} \\ &- \frac{1}{2} \left(\sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij}^t \right) \ln \sigma_X^2 \\ &- \frac{1}{2\sigma_U^2} \sum_{i=1}^{N} U_i^T U_i - \frac{1}{2\sigma_V^2} \sum_{j=1}^{M} (V_j - \theta_j)^T (V_j - \theta_j) \end{aligned}$$

$$+\sum_{j=1}^{M}\sum_{s=1}^{W(j)}\ln(\sum_{k=1}^{K}\theta_{jk}\beta_{k,w_{js}})-\ln\sigma_0+C \tag{3.7}$$

where C is a constant that does not depend on the parameters. Maximizing the log-posterior over the latent features with hyperparameters (i.e. the observation noise variance and prior variances) kept fixed is equivalent to minimizing the following sum-of-squared-errors objective functions with quadratic regularization terms:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{c_{ij}^{t}(x_{ij}^{t}-<U_i,V_j>)^2}{2}+\frac{\lambda_u}{2}\sum_{i=1}^{N}U_i^TU_i$$
$$+\frac{\lambda_v}{2}\sum_{j=1}^{M}(V_j-\theta_j)^T(V_j-\theta_j) \tag{3.8}$$

where $\lambda_u=\sigma_X^2/\sigma_U^2$, $\lambda_v=\sigma_X^2/\sigma_V^2$, and $\lambda_0=\sigma_X^2/\sigma_0^2$ and Dirichlet prior (α) is set to 1. Note that c_{ij}^t is the confidence parameter for rating x_{ij}^t.

After the parameters have been learned, we can do the prediction for the missing data. For one data instance if only some parts of the feature data are missing, the *point estimate* can be used to approximate their expectations as:

$$\mathrm{E}[x_{ij}^t|X]\approx\mathrm{E}[u_i^t|X](\mathrm{E}[\theta_j^t|X]+\mathrm{E}[\varepsilon_j^t|X]) \tag{3.9}$$

$$x_{ij}^{t\ *}\approx(u_i^*)^Tv_j^* \tag{3.10}$$

If the whole data instance is lost, $\mathrm{E}[\varepsilon_j^t|X]=0$ and the missing values can be predicted as:

$$\mathrm{E}[x_{ij}^t|X]\approx\mathrm{E}[u_i^t|X](\mathrm{E}[\theta_j^t|X]) \tag{3.11}$$

$$x_{ij}^{t\ *}\approx(u_i^*)^T\theta_j^* \tag{3.12}$$

Using this approach, the missing data can be estimated and added back to the dataset for further processing.

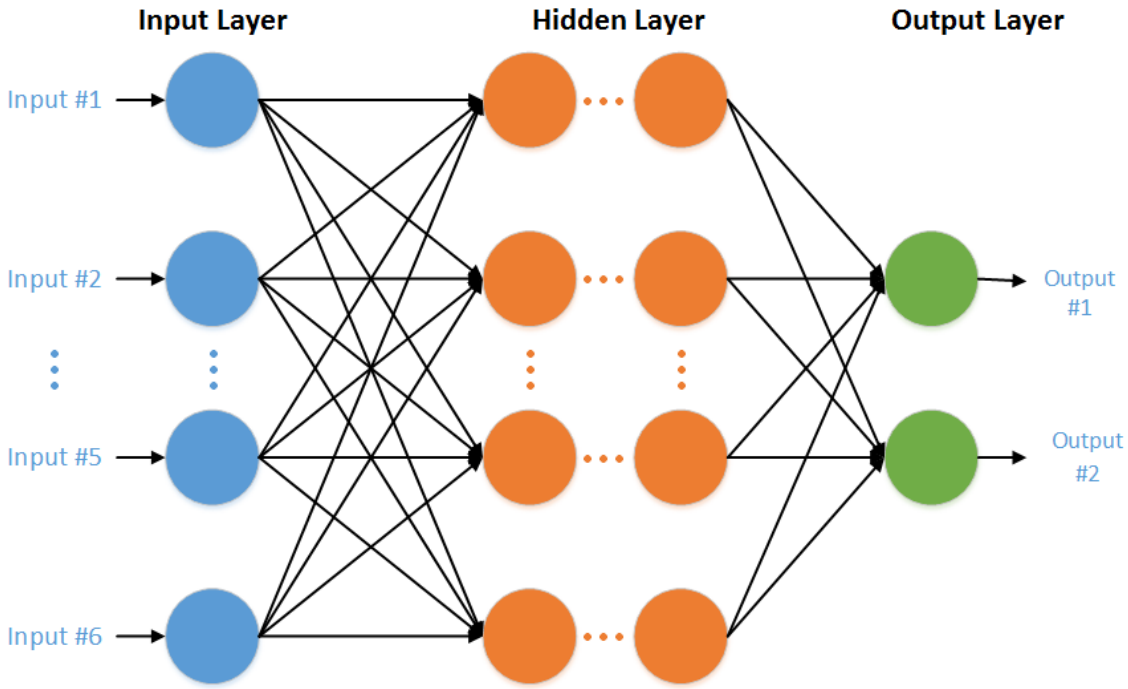

Figure 3.1. The deep neural network.

3.3. Deep neural network training and prediction process

After the missing values have been filled by CTR, deep neural network can be used for the model prediction. Figure 3.1 describes the architecture of the deep neural network that we used for classification. The system does not need any complicated syntactic or semantic preprocessing. The feature vector is fed into the input nodes of the network. Each node generates an output with an activation function, and the linear combinations of the outputs are linked to the next hidden layers. The activation functions among different layers are different. The training data is defined as $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_P, y_P)\}$ of P samples. $\mathbf{x}$ is the input feature vector, y is the class label information. In succession, the features are respectively extracted and then directly concatenated to form the final feature vector. Finally, to compute the confidence of each relation, the feature vector is fed into a *softmax* classifier. The output of the classifier is a vector, the dimension of which is equal to the number of predefined classification types. The value of each dimension is the confidence score of the corresponding classification.

In the training process, the input feature goes through the input nodes at the bottom of the deep learning network, where the weights are initialized with random values. Thereafter the weight

vectors are fine-tuned in sequence. The training goal of this process is to minimize a comprehensive cost function given as the mean squared error function between the prediction value and the real output value:

$$C(\mathbf{w}; \mathbf{x}, y) = \|h_w(\mathbf{x}) - y\| \tag{3.13}$$

where $\mathbf{w}$ is the set of the weights in the deep learning network, which needs to be trained in this phase, y is the label, $h_w(\mathbf{x})$ is a hypothesis function which will yield an estimated output, and $\|\bullet\|$ denotes the Frobenius norm.

The overall cost function for a batch training is defined as:

$$J(\mathbf{w}) = \frac{1}{P} \sum_{p} C(\mathbf{w}; \mathbf{x}_p, y_p) \tag{3.14}$$

We want to obtain the optimal parameter set to achieve the minimization of the objective function as:

$$\mathbf{W}^* = \arg\min_{\mathbf{w}} J(\mathbf{w}) \tag{3.15}$$

This can be achieved by the back propagation algorithm. In the back propagation algorithm, we use the stochastic gradient method to update the weight vectors from the top layer to the bottom layer as

$$w_{ji}^n = w_{ji}^{n-1} + \eta \frac{\partial}{\partial w_{ji}^{n-1}} J(\mathbf{w}) \tag{3.16}$$

where η is an adaption parameter.

After the neural network structure and the weighted parameter are determined, the deep neural network can make prediction directly.

3.4. Algorithm design and implementation

Given i for the number of patients, j for health-related items/factors and K for topics, the proposed algorithm is summarized as:

Algorithm 1 Algorithm for CFDL

Input: The training dataset in

Output: The architecture and weights for DNN out

Initialization : Initialize the network weights to random number

for each patient i **do**
  Draw patient latent vector $u_i \sim \mathcal{N}(0,\ \lambda_u^{-1} I_K)$
end for
for each health-related item j **do**
  Draw topic proportions $\theta_j \sim Dirichlet(\alpha)$
  Draw item latent offset vector $\varepsilon_j \sim \mathcal{N}(0,\ \lambda_v^{-1} I)$, and set the item latent vector as $v_j = \varepsilon_j + \theta_j$
  for each word w_{jn} **do**
    Draw topic assignment $z_{jn} \sim Mult(\theta)$
    Draw word assignment $w_{jn} \sim Mult(\beta_{z_{jn}})$
  end for
end for
for each patient-health-related item pair (i, j) **do**
  Draw the rating $x_{ij} \sim \mathcal{N}(\mathrm{u_i}^T \mathrm{v_j},\ \mathrm{c}_{ij}^{-1})$, where c_{ij} is a confidence parameter for rating x_{ij}, $a > b$. $c_{ij} = a$. (higher confidence), if $x_{ij} = 1$ and $c_{ij} = b$, if $x_{ij} = 0$.
end for

while the iteration number is greater than 0 **do**
  for each of the input data
  $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_P, y_P)\}$ **do**
    Form the vector of input, run the network forward with the input data to get the network output
  end for
  for each output node **do**
    Compute $\|h_w(\mathbf{x}) - y\|$
  end for
  for each of the layer **do**
    Update the weight as $w_{ji}^n = w_{ji}^{n-1} + \eta \frac{\partial}{\partial w_{ji}^{n-1}} J(\mathbf{w})$
  end for
end while
return the network structure and weights

4. EXPERIMENTS

In this section, we present our initial efforts on evaluating the proposed methodologies.

4.1. Datasets

We have been working on a project studying diabetes disease and its risk factors among American Indian men and women. Data collected from this project include patients' vital signs from their body sensors, their daily life style information including diet, workout and social activities, and other demographical data. The data format is heterogeneous including unstructured data, noise and missing values, and the large number of features are complex. However as our data collection is in the initial stage, currently we do not have statistically enough data for analyzing. Therefore, in this paper we have to turn to public available data to test our methodology.

We applied our proposed methodology on the UCI dataset [2]. This dataset contains real medical records of patients diagnosed with diabetes, collected over a period of 10 years (1999-2008) from 130 hospitals in USA. The medical records of each patient included 55 different attributes and a label indicating whether the patient was readmitted to the hospital within 30 days, after 30 days, or never readmitted. The distribution is as follows - 11% of patients were readmitted within 30 days, 35% after 30 days and 54% patients were never readmitted. In total, there were 101,765 encounters available for analysis that satisfy these criteria. Each encounter was labeled with one of three classes ('<30', '>30', 'NO') based on whether the patient was readmitted within 30 days ('<30'), readmitted in more than 30 days ('>=30'), or did not have a recorded readmission ('NO').

4.2. Pre-processing

For the 10173 patients, they have 24 medicine features, which have four kinds of discrete values, *No*, *Up*, *Down*, and *Steady*. we map these values to numerical values from 1 to 4 respectively. To verify the performance of our proposed method on dealing with missing values, we intentionally removed 20% of these medicine values. We molded the medicine feature matrix as a collaborative filtering problem with 10173 users and 24 items. Therefore, in this step, 1,791,064 points of data were used for training and 447,766 points of data were used for prediction of the CTR model. As there is no unstructured data in our dataset, our model has been degenerated to a PMF (probability matrix factorization) problem.

Table 4.1. List of feature descriptions

Feature name	Type
Gender	Nominal
Race	Nominal
Age	Nominal
Weight	Numeric
Medical specialty	Nominal
Time in hospital	Numeric
Number of lab procedures	Numeric
Number of procedures	Numeric
Number of medications	Numeric
Number of outpatient visits	Numeric
Number of emergency visits	Numeric
Number of inpatient visits	Numeric
Diagnosis 1	Nominal
Diagnosis 2	Nominal
Diagnosis 3	Nominal
Number of diagnoses	Numeric
Glucose serum test result	Nominal
A1c test result	Nominal
Change of medications	Nominal
Diabetes medications	Nominal
24 features for medications	Nominal

To avoid over-fitting the model, we removed some of the attributes in the original dataset such as a patient's Encounter ID and patient number. The features we used are summarized in Table 4.1 Some of the features were processed. For example, age is divided into 10 ranges and is calculated as the mean of the interval. The target variable (class label) is readmission or not. It has three levels: '<30' (patient is readmitted within 30 days), '>30' (patient is readmitted after 30 days) and 'no' (patient was not readmitted). In this paper, the target variable has been re-coded to '1' (patient is readmitted) and 0 (patient is not readmitted) as a binary variable.

4.3. Training and testing

We train a five-layer deep neural network with 1 input layer, 1 output layer and 3 hidden layers. The output layer's active function is *softmax*. Each layer consists of some neurons. The number of neurons of the input layer is the same as the input feature dimension, and the number of neurons of the output layer is the same as the output classes. For the neurons in the input layer, they receive a single value on their input and are sent to all of the hidden nodes. The nodes of the hidden and output layers are active, and each layer is fully interconnected. The output layer is

responsible for producing and presenting the final network outputs, which are generated from the procedure performed by neurons in the previous layers.

In the deep learning prediction step, we split the dataset into two parts: a training dataset (70%) and a testing dataset (30%). To conduct the experiment, we take an average of 10 runs by shuffling the dataset. We have used 10 folds cross validation by subdividing the original dataset. For all these folds we got similar results which indicates the stability of the score.

4.4. Result and discussion

To measure the performance of the proposed algorithm, CFDL, we compare it with the state-of-the-art classification approaches, namely Support Vector Machine (SVM), Decision Tree, and Naive Bayes. The result is represented based on the most commonly used metrics: *accuracy*, *precision*, *recall*, *micro-averaged F1*, *macro-averaged F1*, which are defined as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$micro - avg.F1 = \frac{TP}{num} \tag{4.4}$$

$$macro - avg.F1 = \frac{2 \times recall \times precision}{recall + precision} \tag{4.5}$$

All these measurements are being calculated using the confusion matrix in Figure 4.1 based on two possible outcomes: positive (p) and negative (n).

Summaries of these performance measurements for the above-mentioned approaches can be found in Table 4.2 and 4.3. Table 4.2 shows the prediction results using the original UCI dataset, while Table 4.3 presents the prediction results using a dataset that were intentionally deleted 20% of the random selected medication data from the UCI dataset. For both scenarios, CFDL is the

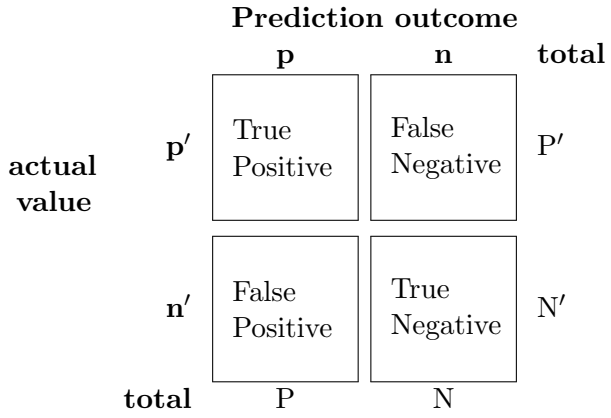

Figure 4.1. Confusion matrix

winner among all of these comparisons. The findings in these two tables clearly indicate that the proposed approach CFDL outperforms other approaches in almost all of the five metrics.

Although the experiments performed on the UCI dataset demonstrate the good performance of our proposed approach, we expect that the proposed method would performs even better compared with other approaches on more complex, heterogeneous, and unstructured data, as our algorithm was designed to overcome the challenges of such dataset.

Table 4.2. Performance measurements for different prediction methods - original dataset

Approach	Micro-avg. F1	Macro-avg. F1	Accuracy	Precision	Recall
CFDL	0.107	0.64	0.8894	0.5018	0.9637
SVM	0.077	0.593	0.8706	0.448	0.875
Decision Tree	0.087	0.35	0.6774	0.2258	0.7788
Naive Bayes	0.077	0.263	0.5693	0.1629	0.6908

Table 4.3. Performance measurements for different prediction methods - 20% missing data

Approach	Micro-avg. F1	Macro-avg. F1	Accuracy	Precision	Recall
CFDL	0.087	0.568	0.869	0.448	0.7788
SVM	0.098	0.477	0.78	0.328	0.875
Decision Tree	0.087	0.292	0.5756	0.1796	0.787
Naive Bayes	0.096	0.327	0.634	0.216	0.867

5. CONCLUSION

A novel collaborative filtering-enhanced deep learning approach was proposed in this paper. In the proposed approach, the collaborative topic regression (CTR) was first utilized to mine the hidden value for the missing data from the structured and unstructured data. Thereafter the deep neural network was employed and trained for the prediction. The proposed approach was applied to a big complex medical problem, specifically, a real-world diabetes patient readmission dataset.

Our experiment and result showed that our algorithm can effectively utilize both structured and unstructured data to find hidden relationship from these data; it can accurately estimate missing values of the dataset, it can also integrate feature learning for prediction in the complex data. We also analyzed and compared with other intelligent algorithm such as Support Vetector Machine (SVM), Decision Tree and Naive Bayes. It is observed that the proposed method outperformed these approaches, the comprehensive comparative experiments results demonstrated the superiority.

REFERENCES

[1] National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

[2] Uci machine learning repository: Diabetes 130-us hospitals for years 1999-2008 data set. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008.

[3] K. E. Bergethon, C. Ju, A. D. Devore, N. C. Hardy, G. C. Fonarow, C. W. Yancy, P. A. Heidenreich, D. L. Bhatt, E. D. Peterson, and A. F. Hernandez. Trends in 30-Day Readmission Rates for Patients Hospitalized with Heart Failure: Findings from the Get with the Guidelines-Heart Failure Registry. *Circulation: Heart Failure*, 9(6), 2016.

[4] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, pages 1–14, 2018.

[5] Z. Che, S. Purushotham, and Y. Liu. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. In *NIPS Workshop on Machine Learning for Healthcare*, pages 1–13, 2015.

[6] CY. Cheng, F. Wang, P. Zhang, and J. Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *SIAM International Conference on Data Mining*, pages 432–440, 2016.

[7] M. Ghassemi, M. A. F. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 446–453, 2015.

[8] Z. Liang, G. Zhang, J. Huang, and Q. Hu. Deep learning for healthcare decision making with EMRs. In *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*, number Cm, pages 556–559, 2014.

[9] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell. Learning to Diagnose with LSTM Recurrent Neural Networks. In *International Conference on Learning Representations*, pages 1–18, 2016.

[10] Z. C. Lipton, D. C. Kale, and R. Wetzell. Modeling Missing Data in Clinical Time Series with RNNs. In *Proceedings of Machine Learning for Healthcare*, volume 56, 2016.

[11] Petersen M. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care*, 39(7):1033–1046, 2016.

[12] S. Mehrabi, S. Sohn, D. Li, J. Pankratz, T. Therneau, J. Sauver, H. Liu, and M. Palakal. Temporal Pattern and Association Discovery of Diagnosis Codes Using Deep Learning. In *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, pages 408–416, 2015.

[13] R. Miotto, L. Li, B. Kidd, and J. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6(January):1–10, 2016.

[14] M. Moreira, J. Rodrigues, N. Kumar, J. Al-Muhtadi, and V. Korotaev. Evolutionary radial basis function network for gestational diabetes data analytics. *Journal of Computational Science*, (October), 2017.

[15] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. Chua. Disease Inference from Health-Related Questions via Sparse Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2107–2119, 2015.

[16] S. Purushotham, Y. Liu, and C.-C. J. Kuo. Collaborative topic regression with social matrix factorization for recommendation systems. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 691–698, USA, 2012. Omnipress.

[17] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard. Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, pages 178–193, 2017.

[18] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Yang. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[19] H. Shin, Le Lu, L. Kim, A. Seff, J. Yao, and R. Summers. Interleaved Text / Image Deep Mining on a Large-Scale Radiology Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 17, pages 1090–1099, 2015.

[20] F. Wang, P. Zhang, B. Qian, X. Wang, and I. Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 145–154, 2014.