*Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*
*A. Ugon et al. (Eds.)*


*doi:10.3233/978-1-61499-852-5-111*

# Using Machine Learning Approaches for Emergency Room Visit Prediction Based on Electronic Health Record Data

Zhi QIAO[1], Ning SUN, Xiang LI, Eryu XIA, Shiwan ZHAO, Yong QIN
*IBM Research Lab - China*

**Abstract.** Emergency room(ER) visit prediction, especially whether visit ER or not and ER visit count, is crucial for hospitals to reasonably adapt resource allocation and` for patients to know future health state. Some existing studies have explored to use machine learning methods especially kinds of general linear model to settle down the task. But, in the clinical problems, there exist complex correlation between targets and features. Generally, liner model is difficult to model complex correlation to make better prediction. Hence, in this paper, we propose to use two non-linear models to settle the problem, which are XGBoost and Recurrent Neural Network. Experimental results show both methods have better performance.

**Keywords.** ER visit prediction, EHR, Machine Learning

## 1. Introduction

One of the central goals of the Affordable Care Act is to improve access to health care, which was expected to decrease the number of emergency room (ER) visits as patients relied on their primary physicians rather than visiting the ER for non-emergency conditions [1]. Hence, ER visit prediction, especially whether visit ER or not and ER visit count, is crucial for hospitals to reasonably adapt resource allocation and for patients to know future health state. Some studies have explored to use machine learning methods to settle down the task. [2] proposed to use logistic regression to make whether ER visit prediction. [3] proposed to use linear regression to model the correlation between ER visit count and clinical features to make a prediction. Both logistic regression and linear regression models can be considered as the generalized linear model, which means they work better when the data has a linear shape. But, in the clinical problems, there exists complex correlation between objects and features. Generally, liner model is difficult to model complex correlation to make better prediction. Hence, in this paper, we propose to use two non-linear models to settle the problem, which are Extreme Gradient Boosting (XGBoost)[4] and Recurrent Neural Network (RNN)[5]. XGBoost is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. Recurrent neural network (RNN) is a class of artificial neural network which can exhibit dynamic temporal behavior. Both models have more complex model structure and stronger fitting ability.

[1] Corresponding author. E-mail: qzbj@cn.ibm.com.

## 2. Problem Definition

In this paper, there are two ER-related prediction tasks, whether visit ER prediction (task 1) and ER visit count prediction (task 2). Based on observed HER data in this year, both tasks make ER related value prediction in the next year.

Assuming there are n patients, there are two label sets $Y=\{y_i, i=1,\ldots,n\}$ and $P=\{p_i, i=1,\ldots,n\}$, where $y_i \in N$ represents ER visit count of i-th patient in the next year and $p_i \in \{0,1\}$ represents whether visit ER of i-th patient in the next year.

There are basis feature set $F=\{f_i, i=1,\ldots,n\}$, where $f_i$ is denoted as feature vector of i-th patient extracted from current year data. The features contain demography features and historical summary features. The demography features contain sex, age and so on. The historical summary features contain cost statistic, hospital visit statistic, ER visit statistic, disease statistic and so on.

Additionally, for furthermore recurrent neural network exploration, we also extract visit logs of patients to organize visit sequence features for each patient, $V=\{v_i, i=1,...,n\}$ where $v_i$ presents the visit sequence of i-th patient. For i-th patient visit sequence, $v_i=\{v_{i1}, v_{i2}, ..., v_{iT(i)}\}$, where T(i) indicates the length of the visit sequence of i-th patient in the EHR data. We denote all the unique diagnose codes from the EHR data as $C = \{c_1,c_2,\cdots,c_{|C|}\}$, where $|C|$ is the number of unique diagnose codes. Thus, in visit sequence $v_i$, each visit $v_{it} \in \{0, 1\}^{|C|}$ is denoted by a binary vector where the j-th element is 1 if the visit contains the code $c_j$. Each diagnosis code can be mapped to a node of the International Classification of Diseases (ICD-9)[2].

Figure 1 shows the prediction problems.

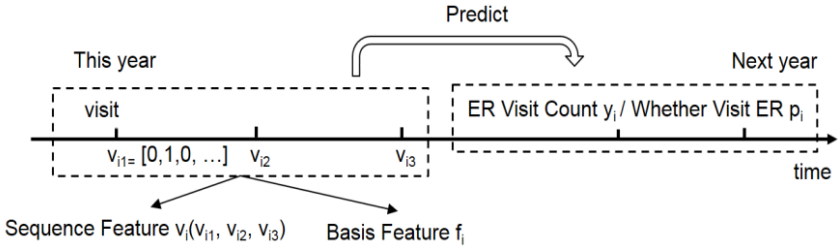


**Figure 1.** Problem definition (Take data of i-th patient for example. For i-th patient, $v_i$ is the extracted sequence feature. $f_i$ is the extracted basis feature and task target is to predict value $p_i$ or $y_i$).

## 3. Model Description

### *3.1. XGBoost*

XGBoost[4] is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. It is an implementation of gradient boosted decision trees designed for speed and performance. It is used for supervised learning problems, where we use the training data $X_i$ (with multiple features) as input to predict a target variable $Y_i$ which is considered as label. The model usually refers to the mathematical structure of how to make the prediction $Y_i$ given $X_i$. The prediction value

[2] https://www.cdc.gov/nchs/icd/icd9cm.htm

can have different interpretations, depending on the task, i.e., regression or classification.

In this paper, in order to apply XGBoost method, we use extracted basis features as input and use separately ER visit count and Whether visit ER as label.

### *3.2. Recurrent Neural Network.*

Recurrent neural network (RNN)[5] is a class of artificial neural network where connections between units form a directed cycle. This allows it to exhibit dynamic temporal behavior. Unlike feedforward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs. Basic RNNs are a network of neuron-like nodes, each with a directed (one-way) connection to every other node. Each node (neuron) has a time-varying real-valued activation. Each connection (synapse) has a modifiable real-valued weight. Nodes are either input (receiving data from outside the network), output nodes (yielding results) or hidden nodes (that modify the data route from input to output). For supervised learning in discrete time settings, sequences of real-valued input vectors arrive at the input nodes, one vector at a time. At any given time step, each non-input unit computes its current activation (result) as a nonlinear function of the weighted sum of the activations of all units that connect to it. Supervisor-given target activations can be supplied for some output units at certain time steps.

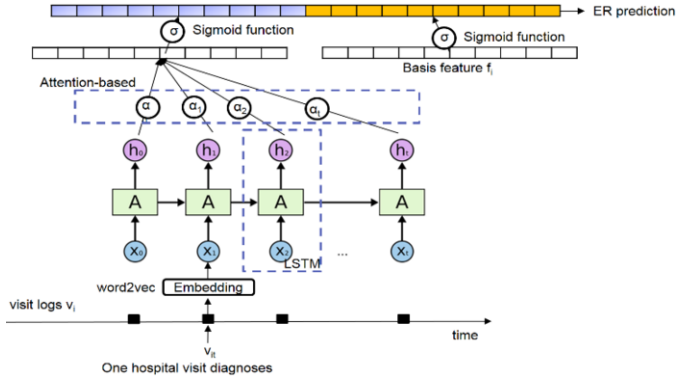


Figure 2. Recurrent neural network model

In this paper, we adapt basic recurrent neural network to model sequence feature data and basis feature data. Figure 2 shows proposed recurrent neural network model. In the model, we use typical Long short-term memory (LSTM)[6] as recurrent unit, which are a special kind of RNN, capable of learning long-term dependencies. As introduced above, we use visit sequence of patients as inputs of RNN. While each visit vector $v_{it}$ of each patient visit set $v_i$ is sparse where a small part of values is 1 else 0, the sparse input will impact parameters fully learning. Hence, we firstly embed the visit vector as dense vector using typical embedding method word2vec[7]. Through RNN learning, each step will generate out value $h_t$, which can be seen as distillation feature of current step. Then, we apply attention-based learning to ensemble all outs ($h_0$, $h_1$, …, $h_{T(i)}$) to generate new vector. Then, we use sigmoid function to map the generated feature vector to generate new feature vector $G_L$ as blue stick in Figure 2. Similarly, for basis feature vector $f_i$, we also use sigmoid function to map the basis feature vector $f_i$ to generate new feature vector $G_R$ as orange stick in Figure 2. Then we concatenate vectors $G_L$ and $G_R$ to construct final feature vector G. In the end, we use least square as

loss function to model correlation between feature vector G and ER visit count for task 1 ER visit count prediction; we use logistic cross entropy as loss function to model correlation between feature vector G and the target whether visit ER for task 2 whether visit ER prediction.

## 4. Experimental Setting

In this section, we evaluate the performance of both methods in real-world datasets. Firstly, we describe the datasets. Then we describe evaluation methods. In the end, we present the experiment results with discussions.

### *4.1. Dataset description*

We conducted experiments on real-world EHR dataset. The data is an actual collection of logs of patient hospital-visit which contain visit time, diagnoses, cost information and corresponding ER flag.

Firstly, we extract the first-year data from the original data to construct new feature dataset which contains 6 thousand patients, 0.1 million visits and hundreds of unique diagnose codes. For each chose patient, we extract corresponding second-year ER information as label data from original data.

Secondly, we construct basis features from the new feature dataset. Each patient has a corresponding basis feature vector. The basis feature vector contain gender (Male or Female, filled with 1 or 0), age, 27 chronic conditions[3] (if patient meet some conditions, the corresponding values is filled with 1 else 0), patient hospital-visit count and total cost, ER visit count, cost & count in the top-5 frequent and top-5 cost diagnoses.

Then, we reorganize the visit log according to sequence feature construction of section 2 to adapt to RNN learning.

We set the proportion of validation set, training set and testing set 1:8:1.

### *4.2. Evaluation*

To evaluate ER visit count prediction (task 1) results, we adopt standard regression evaluation metrics R-squared value. **R-squared value** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

To evaluate Whether visit ER prediction (task 2), we adopt standard classification evaluation metric, the area under the curve(AUC). **AUC value** is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

### *4.3. Experimental Results*

In this part, we conduct separately Logistic Regression, XGBoost and RNN on the generated experimental data for task 1, and Linear Regression, XGBoost and RNN for

[3] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/MCC_Main.html

task 2. The basis features are applied in all of three methods. The sequence features just are used in RNN. For RNN, we conduct RNN-100, RNN-50 and RNN-10 with different embedding sizes 100, 50 and 10.

#### *4.3.1. Task 2 Whether visit ER prediction*

We show the experimental results in Table 1. We can find that both XGBoost and RNN have better performance than traditional linear logistic regression method. Compared with logistic regression, XGBoost and RNN have more complex model structure and stronger fitting ability. Additionally, XGBoost has best performance among three methods. Moreover, with growth of embedding size, performance gets better.

**Table 1**. Whether visit ER prediction

| | **Linear Regression** | **XGBoost** | **RNN-100** | **RNN-50** | **RNN-10** |
|---|---|---|---|---|---|
| AUC | 0.6455 | 0.6974 | 0.6881 | 0.6804 | 0.6766 |

#### *4.3.2. Task 1 ER visit count prediction*

We show the experimental results in Table 2. We can find that both XGBoost and RNN have better performance than traditional linear regression method. Compared with logistic regression. Additionally, RNN has best performance among three methods. Compared with XGBoost, RNN uses additional sequence features besides of basis features to learn disease state evolving features of patients. Different from finding of task 1, experimental results show using sequence features can improve prediction performance.

**Table 2**. ER visit count prediction

| | **Logistic Regression** | **XGBoost** | **RNN-100** | **RNN-50** | **RNN-10** |
|---|---|---|---|---|---|
| R-squared | 0.1784 | 0.3354 | 0.3421 | 0.3405 | 0.3367 |

## 5. Conclusion

In this paper, we propose to use two non-linear models to settle the problem, which are XGBoost and Recurrent Neural Network. Experimental results show both methods have better performance and using sequence features can improve performance of ER visit count prediction.

## References

[1] Patient Protection and Affordable Care Act, 42 U.S.C., Section 18001 et seq. United States of America; 2010.
[2] Sarah Poole, Shaun Grannis, and Nigam H. Shah, MBBS. AMIA Jt Summits Transl Sci Proc. 2016; 2016: 438–445.
[3] Suffoletto B., Miller T., Shah R., Callaway C., Yealy D.M. Predicting older adults who return to the hospital or die within 30 days of emergency department care using the ISAR tool: subjective versus objective risk factors. Emerg Med J. 2015
[4] Schmidhuber, Jürgen. Habilitation thesis: System modeling and optimization. 1993
[5] Hochreiter, Sepp; Schmidhuber, Jürgen. Long Short-Term Memory. Neural Computation. 9 (8): 1735-1780. 1997
[6] Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781.