ATLANTIS PRESS
Advances in Computer Science Research, volume 63
International Conference on Artificial Intelligence and Engineering Applications (AIEA 2016)

Medical health data analysis based on Spark Mllib

Tong Xiao [a], Yijie Shi [b]

Beijing University of Posts and Telecommunications, Beijing 100876, China

[a]xt1993124@163.com, [b]yijieshi@bupt.edu.cn

Abstract. In recent years, health and disease prediction has become an important part of medical wisdom, and has attracted more and more attention. At this stage, the prediction of health care mainly relies on the medical health records data. For predicting results, it is only in view of the disease or not. At present, for the lack of adaptability and limitations of the data feature selection, in this paper we use the existing health records data and available life habit data, combined with the current popular Spark machine learning computing platform, and establish a multi-classification model, which can provide a reasonable prediction and evaluation. This design has a certain degree of accuracy and efficiency and it has a certain use value.

Keywords: Machine learning; Medical wisdom; Health records; Classification.

1. Introduction

With the reform of the health care industry, the application of machine learning in medical health greatly promotes the development of the medical wisdom. How to quickly and accurately predict people's health status helps to improve work efficiency and quality of medical workers, and thereby reduce the incidence of the diseases.

At present, in view of the medical health prediction, it mainly focuses on the binary classification of disease or not [1]. Nowadays, as people work pressure increasing, more and more people are in a state of sub-health. Sub-health is a kind of critical state, and although there is no clear disease, a spiritual vitality, adaptability and ability to respond to drop, and if not corrected in time, it is easily to cause mental and physical ill health and disease.

Health data can be obtained through the analysis of the existing electronic health records, and the main parameters include gender, age, height, weight, blood pressure, and etc. At present, most of the medical health prediction only focuses on the prediction disease or not, and at the same time, ignores some of the life habits of data, which makes the model has a certain of imperfection and limitations [2].

2. Prepare Data

In this paper we aim to build two forecasting models: two-classification disease prediction model based on logistic regression algorithm and multiple-classification sub-health prediction based on Random Forests algorithm. By selecting some appropriate features from the medical health records data and life habits data as the model input, we train the input data into training models by using logistic regression algorithm and random forests algorithm to obtain the classification accuracy.

2.1 Electronic health records

EHR (Electronic health records) is the electronic records of people directly formed in health care activities [3].In this paper, the electronic health record data mainly includes sex, age, height, weight and other basic information and heart rate, blood pressure, blood sugar, family history and other biological information [4].At the same time, exercise and sleep health or other health data from the pedometer bracelet are also uploaded as an important feature selection. As shown in the table, 1 presents male, and 2 presents female:

Table1 EHR Info

Sex	Age	Height	Weight	...
1	456	456	123	...
2	789	213	644	...

2.2 Life habit data

In this paper, we select habit data, including diet status, mood status and number of sport per week. The diet status level is divided into 1: regular, 2: occasionally irregular, and 3: irregular; mood status level is divided into 1: never anxious irritability, 2: occasionally anxiety and irritability and3: regular anxiety and irritability. All of those are as shown in the following figure:

Table2 Life Habit Data

Diet	Mood	Number Of Sports	...
1	2	1	...
2	2	3	...

2.3 Division of health data

According to the diagnostic result of electronic health records, people's health conditions are divided into the two levels: disease condition and non-disease condition. Moreover, the condition of non-health include disease condition and sub-health condition. Experts suggest that according to the characteristics of life and physical condition, sub-health condition can be divided into light, middle and severe three levels. Sub-health condition, is typically characterized by frequent colds, mouth ulcers, cervical discomfort, feeling tired all day long, poor memory ,and strong obsessive compulsive disorder.

3. Classification Model

3.1 Logistic based on time window

In this paper, in order to improve the prediction accuracy and robustness, we introduce the Time window model. As shown in Table 3, each time window contains a characteristic average value of a different time interval. In this paper we use the time window model on the two features sleep time and blood pressure. It increases the accuracy of the classification [5].

Table 3 Logistic Classification Data

30 Days average sleep	60 Days average sleep	......	180 Days average sleep
7.2	6.8		7.5
6.9	7.1		6.6

The prediction steps are as follows:

1) Establish a set of formatted data files and read data from files.
2) Parse each index: value, and convert it into a double value.
3) Divide the data set into training and testing sets.
4) Use the model training data set.
5) Calculate the training error.

3.2 Random forests multi-classification

In this paper, we choose random forests algorithm and select the sample data including medical health records and people life habits data [6].

We extract these features from the data set: Sex, Age, BMI index, Steps, Sleep hours, Diet status, Mood Status, Number of sports, and Family History Disease.

The prediction steps are basically consistent with the above prediction model. The prediction steps are as follows:

1) Establish a set of formatted data files and read data from files.
2) Parse each index: value, and convert it into a double value.
3) Divide data set into training and testing sets.
4) Set parameters.
5) Use the model training data set.
6) Model evaluation.

4. Experience

Apache Spark is a quick cluster computing system. Spark Mllib is a machine learning algorithms library, including the relevant test and data generator. Its goal is to realize the simplicity and integration of machine learning. In this paper, we use the Spark MLlib machine learning classification accuracy, and classification error the two standards to measure the effect of the experiment.

4.1 Logistic regression binary classification

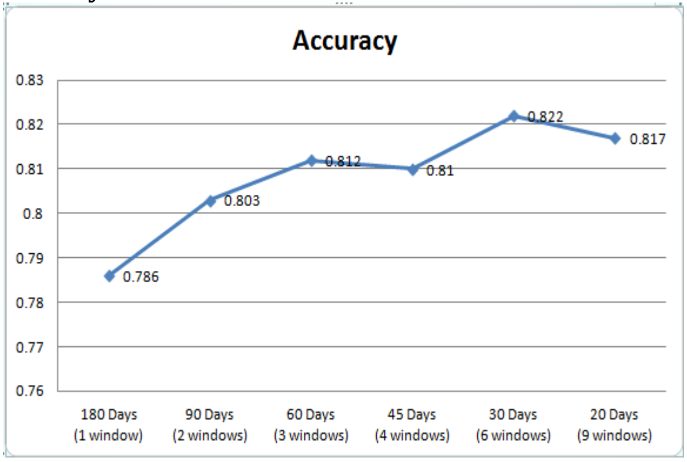

Fig. 1 Logistic binary classification

As is shown above, the abscissa represents the number of divided time window in which the first 180 days as a representative of a time window, the last 15 days as a time window [7]. The ordinate represents the accuracy of logistic regression algorithm classification. In this paper, the experiments shows that with the time window numbers increases, the prediction accuracy increases. When the time window number is 6, the prediction accuracy reach 82%.As the time window number continue to increase, the curve shows a downward trend. In this paper, we select 30 days as a time window, the result is the best.

4.2 Random forests multi-classification

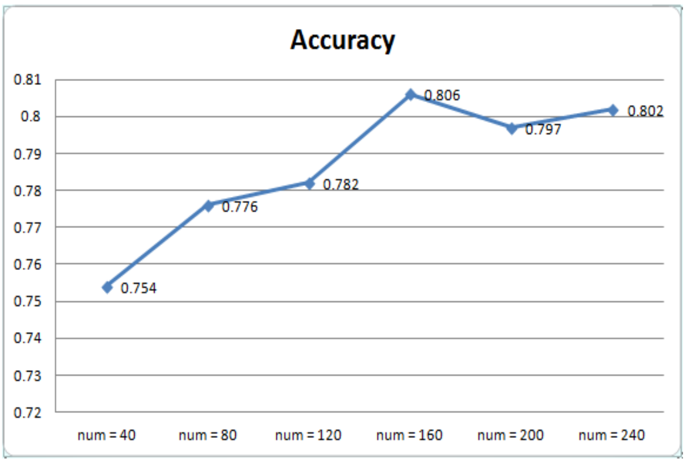

Fig. 2 Random forests multi-classification

The non-health condition can be divided into light sub-health condition, middle sub-health condition, severe sub-health condition and disease condition.

Random forests algorithm needs to set the number of trees .Literature indicates that when the tree number N is large enough， the model accuracy will tend to be stable, but it will cost more. In this paper, we choose n = 40, 80, 120, 160, 200, 240 to build a random forests. The classification accuracy is as shown below. It can be seen that when N is 160 and the accuracy is the best.

In this paper we select logistic regression algorithm and CART decision tree algorithm as the compared algorithms. The data set was classified by the three algorithms and the result is shown as below:

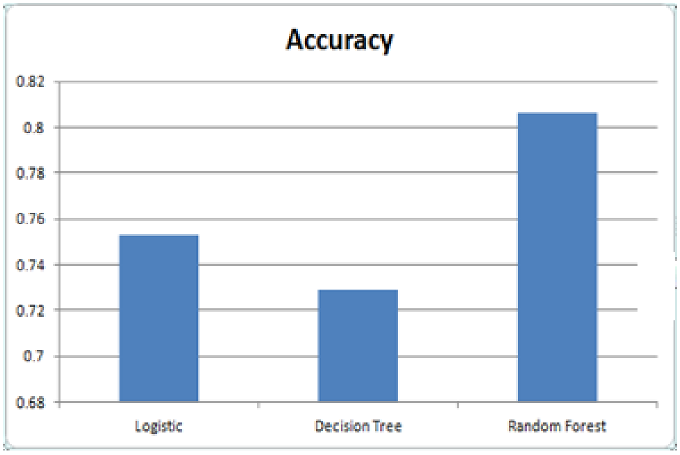

Fig. 3 Multi-classifications comparation

As is shown above, Random forests algorithm used in this paper reach the highest accuracy 80.4%, the second is logistic algorithm, and the CART decision tree algorithm is relatively worse. The CART decision tree, if there is no pruning so that reduce the classification accuracy.

5. Summary

Combined with the characteristics of medical health records and life habits characteristic data, in this paper we propose a logistic regression binary classification algorithm based on time window and a random forests health condition multi-classification prediction algorithm. The experiments show that the proposed algorithm compared with other classifier algorithm, has a certain improvement in accuracy. At the same time, through the multi-classification of non-health condition, we grasp the health level of people's living condition better, effectively prevent the happening of the disease. The next step we will dynamically adjust the parameters according to the result of prediction experiment, and enhance the performance of the algorithm and expand the scale of the experiment.

References

[1] Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding physiological state: mortality modelling in intensive care units[C]// ACM Knowledge Discovery and Data Mining. KDD, 2014:75-84.

[2] Asiimwe S B, Abdallah A, Ssekitoleko R. A simple prognostic index based on admission vital signs data among patients with sepsis in a resource-limited setting [J]. Critical Care, 2015, 19(1):1-8.

[3] Pollettini J T, Panico S R G, Daneluzzi J C, et al. Using machine learning classifiers to assist healthcare-related decisions: classification of electronic patient records.[J]. Journal of Medical Systems, 2012, 36(6):3861-3874.

[4] Seera M, Lim C P, Wei S L, et al. Classification of electrocardiogram and auscultatory blood pressure signals using machine learning models [J]. Expert Systems with Applications, 2015, 42(7):3643–3652.

[5] Golino H F, Amaral L S, Duarte S F, et al. Predicting increased blood pressure using machine learning.[J]. Journal of Obesity, 2014, 2014(5):637635-637635.

[6] Rodriguez J D, Perez A, Arteta D, et al. Using Multi-Dimensional Bayesian Network Classifiers to Assist the Treatment of Multiple Sclerosis[J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2012, 42(6):1705-1715.

[7] Schrom J. Machine Learning for Healthcare [J]. Oreilly, 2015.