International Journal of Advanced Smart Convergence Vol.9 No.2 1-7 (2020)
http://dx.doi.org/10.7236/IJASC.2020.9.2.1

IJASC 20-2-1

Adaptive Recommendation System for Health Screening based on Machine Learning

Namyun Kim, Sung-Dong Kim

Professor, School of Computer Engineering, Hansung University, Korea
nykim@hansung.ac.kr, sdkim@hansung.ac.kr

Abstract

As the demand for health screening increases, there is a need for efficient design of screening items. We build machine learning models for health screening and recommend screening items to provide personalized health care service. When offline, a synthetic data set is generated based on guidelines and clinical results from institutions, and a machine learning model for each screening item is generated. When online, the recommendation server provides a recommendation list of screening items in real time using the customer's health condition and machine learning models. As a result of the performance analysis, the accuracy of the learning model was close to 100%, and server response time was less than 1 second to serve 1,000 users simultaneously. This paper provides an adaptive and automatic recommendation in response to changes in the new screening environment.

Keywords: *Machine Learning, Health Screening, Recommendation System, Cancers and Cardiovascular*

1. Introduction

As more than 15 million people conduct health screening annually in Korea, the health screening industry is gradually expanding. Screening is an important component of health promotion program to identify those individuals who have a disease but do not yet have symptoms. In Korea, various programs are being conducted, including national screening programs [1]. In screening program, it is necessary to examine screening items with high risk of disease rather than quantitative expansion of screening items. However, it is difficult to choose screening items because the probability of disease occurrence varies according to an individual's age, gender, and risk factors, etc. Therefore, there is a need for research that provides personalized health screening optimized for individual characteristics to contribute to the prolongation of people's health. The selection of screening items so far has been insufficient because the screening items are simply listed according to age and gender [2-3].

Recently, studies on recommending systems using big data have been actively conducted [4-5]. In relation to health screening industry, institutions and hospitals provide guidelines for health screening and various

Manuscript Received: March. 12, 2020 / Revised: March. 17, 2020 / Accepted: March. 21, 2020
Corresponding Author: nykim@hansung.ac.kr
Tel: +82-2-760-4348, Fax: +82-2-760-4488
Professor, School of Computer Engineering, Hansung University, Korea

clinical results [6]. This paper proposes a method that recommends screening items based on screening information. To do this, we generate a set of health data using screening guideline and clinical results, build machine learning models based on the data set, and predict screening items based on the machine learning models. For predicting desired screening items, we use supervised learning model which is the most common branch of machine learning [7]. The goal of a supervised learning is to predict the output variable for newly presented input data. We can use many different supervised learning algorithms such as decision tree, random forest and support vector machines [8-9]. Finally, the recommendation server provides a final recommendation list after predicting whether or not the screening is needed under the user's health condition. The method proposed by this paper has the advantage of generating a flexible and automatic prediction model in response to changes in the new screening environment compared to the existing rule-based system.

This paper is organized as follows. Section 2 describes the system architecture for recommending health screening items. Section 3 analyzes the experiment results in terms of accuracy and efficiency of proposed system. Finally, Section 4 discusses our conclusions and proposes future research topics.

2. System Architecture

The recommendation system in this paper builds machine learning models based on the data set and predicts the desired screening items. Figure 1 shows the overall structure of recommendation system for health screening items. In offline mode, the data generator generates the synthetic data set which contains user's medical features and labels using guidelines issued by government and hospitals. Then, we split data set into training set and test set. The training set will be used to construct the machine learning models and test set will be used to test the accuracy of the models. On the other hand, the learning model generator generates a learning model for each screening item and thus multiple learning models are generated. In online mode, the recommendation server receives the user's questionnaire information and predicts whether screening is needed for an individual item. Finally, after integrating the prediction result for each item, server sends a recommendation list to the user. The proposed system has the advantage of providing a flexible and automatic prediction system by updating the data set and machine learning models accordingly when adding new rules or cases that have not been reflected so far.

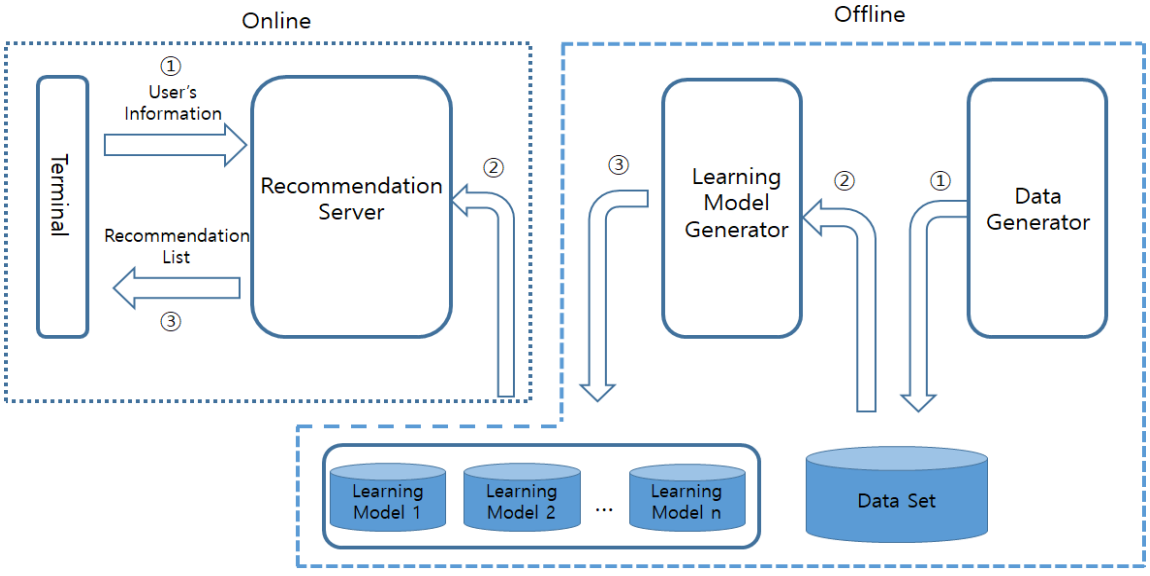

Figure 1. System Architecture for Health Screening Recommendation

2.1 Data Generator

The data generator for health screening generates a data set consisting of *features* that indicate an individual's age, gender, inspection interval, and risk factors (family history, past medical history, eating habits), and *labels* that indicate whether screening is needed, such as lung cancer screening and colon cancer screening. The guidelines issued by domestic and foreign institutions or the clinical data of hospitals were collected to generate a data set of health screening. Once the data set is generated, preprocessing step includes data cleaning which deals with the missing values and removal of unwanted characters, and feature extraction which find out which features are important for prediction. An example of a screening data set is shown in Figure 2.

In Figure 2, the data set is represented by a feature vector representing input attributes and a label vector representing output attributes. The m-dimensional feature vector is expressed as $\vec{X} = \{x_1, x_2, \ldots, x_m\}$, and the n-dimensional label vector is expressed as $\vec{Y} = \{y_1, y_2, \ldots, y_n\}$. The feature vector has health condition about the user and may have a numerical value or a category value. In addition, the label vector may have a binary classification such as whether or not a screening is necessary (0/1), or a multiclass classification classified into three or more classes.

Feature Vector: $\vec{X}$ Label Vector: $\vec{Y}$

ID	Age	Gender	Pack-Years	Stop Smoking Period	Inspection Interval(Lung)	bmi	...	Lung Cancer Test	Cardiovascular Disease Test	...
1	43	1	25	10	3	32.21		0	1	
2	60	1	35	0	5	23.63		1	0	
3	39	0	0	0	1	22.94		0	1	

Figure 2. Example of Data Set for Health Screening

2.2 Learning Model Generator

In previous studies, screening items were recommended based on simple rules according to age and gender. Thus, the previous approach is static and does not produce effective results because there are not considering various considerations. In this paper, it is possible to recommend the screening items more effectively by automatically generating appropriate learning models based on the data set. If all the screening items are recommended using one learning model, there is a disadvantage that it takes a long time to perform and contains a lot of noise. Thus, we adopt multiple learning models that generate independent learning models for each screening item.

Figure 3 shows a learning model generator for health screening. The data selector selects input features and output labels for individual screening items. In general, there is no dependencies between the health screening items, and the features affecting the label are predetermined. Thus, the features and labels for a specific screening item can be selectable. For example, for the lung cancer, select the lung cancer label and its related features such as age, inspection interval, stop smoking period, and pack-years. Feature selection has the advantage of reducing learning model generation time and improving performance by removing irrelevant or redundant features.

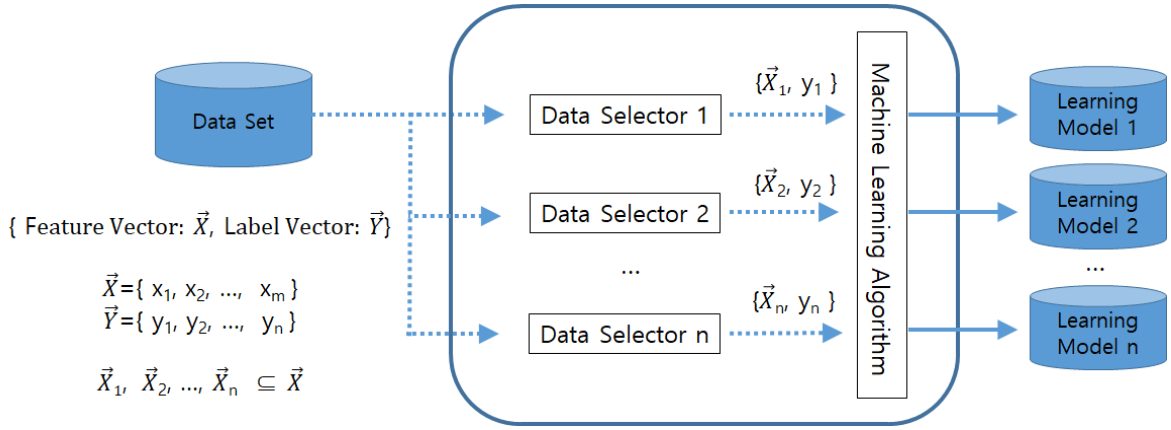

Figure 3. Learning Model Generator for Health Screening

In this paper, the machine learning model for health screening can be expressed as follows.

Given m-dimensional feature vector $\vec{X} = \{x_1, x_2, \ldots, x_m\}$, n-dimensional label vector $\vec{Y} = \{y_1, y_2, \ldots, y_n\}$, and a subset of feature vectors $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ (where, $\vec{X}_i \subseteq \vec{X}$, $1 \le i \le n$), the machine learning model f_i ($1 \le i \le n$) can be written as:

$$y_1 = f_1(\vec{X}_1), \quad y_2 = f_2(\vec{X}_2), \ldots, \quad y_n = f_n(\vec{X}_n) \tag{1}$$

In Equation (1), we generate an individual learning model f_i using $\{\vec{X}_i, y_i\}$ which is extracted by data selector. For example, a learning model for lung cancer, a learning model for colon cancer, and a learning model for stomach cancer are independently generated.

For the machine learning algorithm, supervised learning classification algorithms such as a decision tree and a random forest can be used. Decision trees are intuitive and explainable ways to classify objects but we can overfit the tree. Thus, decision trees can be used to visualize and validate models. On the other hand, random forests are an example of an ensemble learning algorithm built on decision trees. Random forests are much more robust than a single decision tree because they aggregate many decision trees to limit overfitting. Thus, random forests can yield useful results. Then, the optimal learning model is generated by changing the parameters of the learning algorithms. The final learning model is stored as a single file and used when recommending screening items. In Python, pickle can be used to serialize machine learning model to file.

2.3 Recommendation Server

Figure 4 shows the structure of the recommendation server for health screening that can be performed in real time. First, the user sends questionnaire information about health condition. An example is {“ID”: 1, “Age”: 43, “Gender”: “M”, “Height”: 169, “Weight”: 92, “Daily smoking”: 20, “Smoking period”: 25, …}. Second, the feature generator derives the features from the questionnaire information provided by the user. For example, BMI is calculated based on height/weight, and smoking pack years is calculated by multiplying the number of packs of cigarettes smoked per day by the number of smoking years. Third, the feature selector extracts the features for each screening item. For example, in relation to the lung cancer, features such as the inspection interval, non-smoking period, age, and smoking pack years are extracted. Fourth, the screening predictor predicts whether a screening is needed by using features and the learning model as inputs. Finally, the prediction aggregator collects the results of the individual screening prediction and sends the recommendation results $\vec{Y}' = \{y'_1, y'_2, \ldots, y'_n\}$. Examples of recommendation lists can be expressed as JSON

format {"ID": 1, "lung cancer": 0, "liver cancer": 0, "colorectal cancer": 0, "cardiovascular": 1, … }, where 0 indicates that no test is required, and 1 indicates that a test is required. The test result value may have binary data or multiple classes classified into three or more classes.

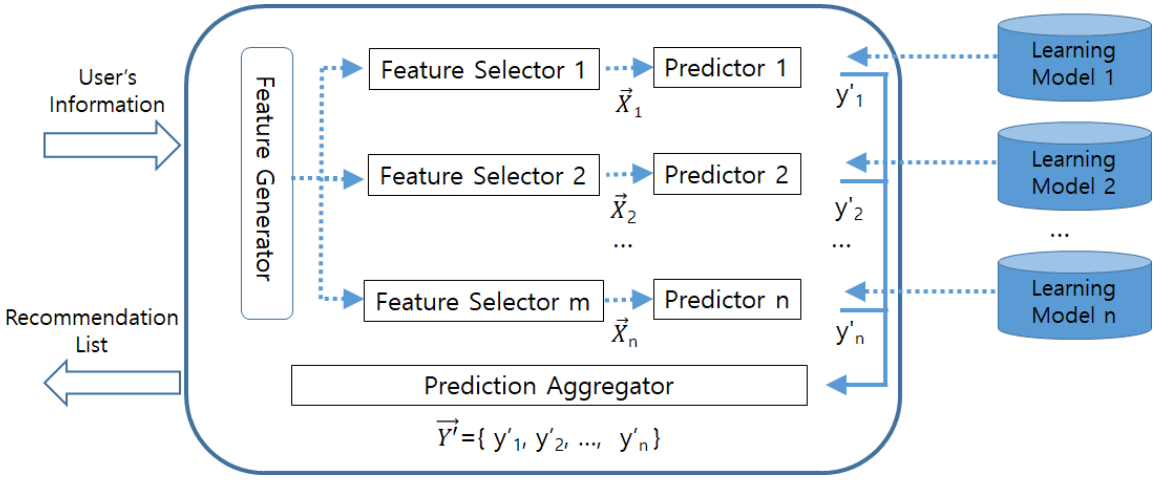

Figure 4. Recommendation Server for Health Screening

3. Experiment

To evaluate the performance of the recommendation system, we measured 1) the prediction accuracy of health screening items and 2) the response time of the recommendation system. First, we can obtain confusion matrix from scikit-learn [10], which visualizes the accuracy of a classifier by comparing the actual and predicted labels. In binary classification, the count of true negatives(TN) is C_{00}, false negatives(FN) is C_{10}, true positives(TP) is C_{11} and false positives(FP) is C_{01}. We can compute the prediction accuracy from the confusion matrix: accuracy = (TP+TN)/(TP+TN+FP+FN). Table 1 shows confusion matrix and prediction accuracy for 7 most common cancers and cardiovascular disease. In this experiment, the data set contains 23 feature variables and 10,000 observations. The data set is split in a way that 75% of the observations fall under the training set and 25% of the observations are used for testing the model. Next, we built the decision tree classifier, fit the model to the training data and get the confusion matrix. For lung cancer in Table 1, the accuracy is equal to (2,387+113)/ (2,387+0+0+113) = 1 where there were 2,387 true negatives and 113 true positives. From Table 1, the accuracy of a classifier comes out as 100% or very close. Note that we have a 4-class classification problem for cervical cancer to classify as no-screening, Pap test, Human papilloma virus, and both (Pap test + Human papilloma virus).

Table 1. Prediction accuracy for cancers and cardiovascular disease

disease	Confusion Matrix	Prediction Accuracy
lung cancer	[2,387 0 0 113]	1
liver cancer	[2,216 0 0 284]	1
colorectal cancer	[1,512 0 0 988]	1

stomach cancer	[1,265 1 0 1,234]	0.9996
breast cancer	[2,012 0 0 488]	1
thyroid cancer	[2,144 0 0 356]	1
cervical cancer	[1,584 0 0 0 0 333 0 0 0 0 274 0 309]	1
cardiovascular	[2,149 0 0 351]	1

Second, we measured the amount of time between a client request and server response. A client request can contain multiple user's information. For this experiment, we built a prototype API using Python and the Flask web framework [11] and run 10 experiments on local machine (MacBook Pro 2.6GHz Quad-core Intel Core i7). Table 2 shows the minimum(min), average(avg) and maximum(max) values of response time for each request. From the table, we can see that a request including 1,000 users can be processed within 1 second.

Table 2. Response Time of Recommendation Algorithm

Users	Response Time(sec)		
	min	avg	max
100	0.820	0.837	0.844
1,000	0.906	0.914	0.930
10,000	1.694	1.711	1.726

4. Conclusions

In this paper, we proposed a recommendation system for health screening based on machine learning models. When offline, a synthetic data set is generated according to institutional guidelines and clinical results, and a machine learning model is generated for each screening item. When online, the recommendation server provides a recommendation list of screened items in real time using the customer's health status and machine learning model. There is an advantage of maximizing the effectiveness and reducing the cost by recommending personalized screening items based on a health condition. As a result of experiments on the 7 most common cancers and cardiovascular disease, it was found that the prediction accuracy was high and the real-time response was possible. The method proposed in this paper can generate a flexible and automatic prediction model in response to changes in the new screening environment. In the future, we will study how to recommend hospitals that best match the proposed screening items.

Acknowledgement

This research was financially supported by Hansung University.

References

[1] Hyun Su Kim, et al., "National Screening Program for Transitional Ages in Korea: A New Screening for

Strengthening Primary Prevention and Follow-up Care," *Journal of Korean Medical Science*, 2012.
DOI: https://doi.org/10.3346/jkms.2012.27.S.S70

[2] The American Cancer Society medical and editorial content team, *https://www.cancer.org/healthy/find-cancer-early/cancer-screening-guidelines/american-cancer-society-guidelines-for-the-early-detection-of-cancer.html*.

[3] Consumer Reports, *https://www.consumerreports.org/men-s-health/mens-health-checklist-for-every-age*.

[4] Chi-Seo Jeong, et al., "Adaptive Recommendation System for Tourism by Personality Type," *International Journal of Internet, Broadcasting and Communication(JIIBC)*, Vol.12, No.1, pp. 55-60, 2020.
DOI: https://doi.org/10.7236/IJIBC.2020.12.1.55

[5] Yoonjung Kim, et al., "Disease risk prediction system using correlated health indexes," *International Journal of Advanced Smart Convergence(IJASC)*, Vol.7, No.4, pp. 1-9, 2018.
DOI: https://doi.org/10.7236/IJASC.2018.7.4.1

[6] National Cancer Information Center, 7 most common cancer guideline, *https://www.cancer.go.kr/lay1/bbs/S1T261C263/B/35/list.do.*

[7] O. Simeone, "A Very Brief Introduction to Machine Learning With Applications to Communication Systems," *IEEE Trans. Cognitive Communications and Networking*, 4(4), pp. 648–664, 2018.
DOI: https://doi.org/10.1109/TCCN.2018.2881442

[8] Jehad Ali, et al., "Random Forests and Decision Trees," *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 5, No 3, pp. 272-278, September 2012.

[9] M. Awad, R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Apress, pp.39-66, 2015.

[10] Scikit-learn: machine learning in Python, *https://scikit-learn.org.*

[11] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, O'Reilly Media, 2018.