Incremental Learning for Alzheimer's Disease on Medical Cloud Service Environment

Chen-Shie HO* and Yu-Mei CHANG
58, Sihchuan Rd. Sec. 2, Pan-Chiao District, New Taipei City 22061, Taiwan
*Corresponding author

Keywords: Machine learning, Data mining, Alzheimer's disease.

Abstract. According to previous survey data, there estimated about 150,000 people suffering from Alzheimer's disease in Taiwan, in which over 65 years old population’s prevalence rate is 4 to 5% and the prevalence rate doubled every increasing in the age of 5. For population of 85 years of age and older, half to one-third of them have the chance of suffering from Alzheimer's disease. If the disease can be accurately predicted before it occurred, or if it can be correctly classified after the occurrence of disease then it will be of great help on the treatment or prevention of this disease. In this paper, the cloud care environment for incremental learning will be considered. The experimental results show that in the case of dynamic learning, the ensemble model with multiple classifiers can achieve more than 80% accuracy, and the results of this study will be applied to hospitals for further modeling for other interested target disease.

Introduction

The advance on information technology has brought revolutionary changes in people's living environment. Cloud-based Internet computing associate with Internet of things (IoT) and various terminal devices will play the important role in the future. The benefits of cloud-based healthcare service systems include providing more convenient access and experience for healthcare facilities in more immediate and cost-effective manner. In response to the future wise medical needs, the hospital will need cloud-based technical support, such as: (1) The need for storage and reservation of ultra-volume of radiology and ultrasound data for patients. (2) The need for increasing data exchange between the hospitals because of the sharing of patients’ medical records. (3) The need to fit the requirement for implementation of long distance medicine and care centers for the hospitals. The purpose of the cloud industry for the government is to establish a ‘personal electronic medical record exchange cloud services’ and ‘intelligent medical system cloud services’ as the future of medical identities, build decentralized intelligent medical network, reduce unnecessary duplication of investment waste, and ultimately improve the quality of medical services. It is hoped that defensive surveillance before diseases, effective control and support of the system after the disease can be achieved and zero medical can be realized through interventions of information and communication technology. In Taiwan, the ‘Medical Cloud’, ‘Care Cloud’ and ‘Wellness Cloud’ with the National Electronic Health Record (EMR) are integrated to ‘Health Cloud’ by Ministry of Health and Welfare [1]. It’s assumed that everyone needs the ‘Care Cloud’ when one becomes old. An associated tele-care program is provided using cloud IT technology, and currently can be seen on long-term caregivers. Home care and community care have not yet identified a sustainable service model but only for institutional care. The ‘Health Cloud’ provides the aim of fitness, health and combines tourism, exercise, health food, and consumer electronics products. The ‘Medical Cloud’ will help to establish the 4P medicine as Personalization, Participation, Prediction and Prevention. If the issues about patient privacy information on the public cloud security can be well-handled and carefully established, the cloud-based medical service system will change the face of traditional service model with unlimited potential.

On the other hand, machine learning techniques provide principled approaches for developing sophisticated, automatic, and effective algorithms for analysis of high-dimensional and multimodal
medical data. As a branch field of machine learning, data mining is part of the so-called Knowledge Discovery in Databases (KDD) [2], which uses a number of statistical analysis and modelling methods to find useful patterns and relationships from the data [3]-[4]. In general, data mining could include the following four functions: (1) Classification: In accordance with the analysis of the properties of different categories to be defined, classify the new instances to a group (class). (2) Estimation: According to the existing continuity of the value of the relevant attribute information, an unknown value of a property is obtained. Techniques used include statistical methods of correlation analysis and regression analysis. (3) Prediction: According to the object properties of the past observations, the future value of the property can be estimated. The techniques used include regression analysis, time series analysis and neural network-like methods. (4) Clustering: The heterogeneous samples are divided into several homogeneous groups (clusters). Homogeneous grouping is equivalent to segmentation in marketing terms, but it is assumed that segments are not defined in advance, and the segments are naturally generated in the data. The techniques used include k-means and agglomeration. Numerous studies have shown that data mining techniques can be used effectively in medical diagnosis [6]-[8]. The classification techniques in data mining will be used in this paper to achieve the aim of computer-aided disease diagnosis.

As the application of cloud computing in various area increases, many medical organizations have started to provide cloud-based interactive services for patients or normal users. The cloud system allows users to access and update patients’ medical information, which is of great help to medical workers for verifying patients’ identification and giving proper treatments to patients. The information then can be wirelessly transmitted between medical personnel through the cloud system. Due to the variety of medical web services, the traffic volume passing through the cloud grows dramatically. The data streams came from device everywhere arrives at the cloud in random manner and hence the real time data analysis becomes the challenge for service provider. In such situation, traditional batch learning for data streams become improper because of model bias for fragment data and infeasible to continuously updated data. Incremental learning is a method of machine learning algorithms, which input data is been reading consecutively and is used to extend the existing model knowledge. It represents a dynamic technique of supervised learning system construction that be applied when training data becomes available gradually over time or its size is out of a main computer system memory limits. The algorithms that apply incremental learning are known as incremental algorithms. The main process of incremental learning is constituted of an existing machine learning model and a new input batch. In addition, the aim is the model be adapted to new data without forget the existing knowledge. However after the model adaption, only some statistics of the data are stored and the original input data is refused. In this paper, the cloud-based clinical decision support system is constructed and the dynamic incremental learning model is established. In order to increase the effectiveness and accuracy of the ensemble model more effectively, two types of queues are prepared, and the full use of these resources will make the operation of the model more effective.

Method Description

Clinical Decision Support System (CDSS)

The framework of CDSS considered in this study is depicted in Figure 1. A CDSS is a health information technology system that is designed to provide physicians and other health professionals with clinical decision support (CDS), that is, assistance with clinical decision-making tasks. Here the function of it is to help the doctors to make the differential diagnosis. Initially, the existing evidence database is fed into the system, and it will dominate the diagnosis since new medical decision pattern is not enough to modify the decision. As the diagnostic result data stream increase and fed into the cloud from distributed medical stations, the dataset will be used to produce the learning model. The learning result will used to update the knowledge database and provide the analytic model to the doctor to make the diagnosis decision. The confirmed result of disease will be fed into the validation

385
set associated with the existing dataset. The fetching strategy in which frequency to retrieve data can be decided by the CDSS system depending on the data arrival rate.

![Diagram of CDSS used in this study.](image)

**System Model**

The proposed system model is depicted in Figure 2. The critical components are described as follows:

(a) Dynamic database: Used to store the existing model. The model is represented by the rule form, and can be fetched by the selected linear or non-linear candidate classifier easily.

(b) Model selector: In this module, appropriate classifier combination is generated to classify the new arrived data.

(c) The filter: Used to decide whether to feed the new diagnostic result into the dynamic database or not. Note the result will also be fed into the method selector module to change the weight of each method in the combination.

The ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. The first step is to create multiple classification/regression models using some training dataset. Each base model can be created using different splits of the same training dataset and same algorithm, or using the same dataset with different algorithms or any other method. Here the logistic regression, K-nearest neighbor and neural network methods are adopted. The model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes. The source code of the system is implemented by Python with Scikit-learning tool. The incremental data is generated by random sampling the medical data acquired from the cooperating hospital.

**Experiment Study**

**Dataset Description**

The data set consist of 374 patients’ records collected from Far Eastern Memorial Hospital (FEMH) located at Pan-Chhiao, New Taipei City, Taiwan. The section of neurology in FEMH is one of the training centers for neurologist in Taiwan. It’s clinical service includes treatment of all kinds of neurological diseases. The cases are afforded by medical history data center under the admission of institutional review board (IRB). Among the patient records, only 41 patients are confirmed to be Alzheimer's disease (AD) patients while the remaining 333 are not. To train the classification model as fair as possible, the training data fed into the system are generated by fetching these personal medical data with ratio of about 2:1. The 10-fold crossover validation scheme is adopted to validate and compare the prediction performance.
**Evaluation Criteria**

The performance of the prediction model can be validated by some parameters including instant accuracy (or error rate), ratio accuracy and window size variation. The instant accuracy is calculated in terms of sensitivity and specificity at some time $t$. The ratio accuracy indicates the average accuracy for all classifier when majority/minority dataset changes. The window size variation of the model represents the size of required buffer window to hold needed best knowledge so far using combined old and new data. The confusion matrix showed in Table 1 can be used to derive the accuracy. It displays the count of correct/incorrect predictions, and compares the actual values in the test set with the predicted values in the train set.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Once the number of predicted instance is counted into each entry of the table, these numbers will be used to derive the parameters as follows:

- **Sensitivity** = $\frac{TP}{TP+FN}$
- **Specificity** = $\frac{TN}{FP+TN}$
- **Accuracy** = $\frac{(TP+TN)}{(TP+FP+TN+FN)}$

**Performance Study**

The results of instant accuracy with accumulated samples are plotted in Figure 2. It is calculated when an updated classifier model is constructed after new data arrives. This procedure doesn’t consider the imbalanced data set problem here. The neural network model is with Gaussian radial-based activation functions. In initial period, the accuracy is dominated by the domain expert knowledge to classify the class of the test data. When the volume of arriving data increases, the constructed model will play more significant role to classify the data using updated classification rules. It shows the ensemble model has excellent average performance even in the imbalance of dataset situations.

![Figure 2. Result of predicted accuracy with accumulated samples.](image)

The windows size variation on model construction is plotted in Figure 3. The maximum size for saving historical data is assigned to 100. The new data will fed into the window buffer if the classification result is wrong compared with the confirmed disease result. Two queues are prepared to load the majority and minority dataset. The results show that the window size is related to the data arrival and distribution pattern, that is, the order of input data will have influence on the buffer size. In general, this variance makes the real-time model construction is impossible if the condition that one update after one data arrives is allowed.
Conclusion

This study implements an incremental learning classifier for Alzheimer's disease against the target dataset, and the proposed ensemble model is established associated with techniques including dynamic knowledge bases, method selectors, filters, and window buffers. The results of the performance studies show that the classifier performs well even for the imbalanced clinical dataset and which implies it suitably be implemented in the cloud decision-making environment. Further studies include the different arrival rates on the overall performance of the classifier, dynamic optimization for classifier, real-time processing model and processing for imbalanced data.

Acknowledgement

This study was financially supported by Far Eastern Memorial Hospital and Oriental Institute of Technology.

References