A Recommender System with Advanced Time Series Medical Data Analysis for Diabetes Patients in a Telehealth Environment

Raid Lafta[1,2], Ji Zhang[1(✉)], Xiaohui Tao[1], Jerry Chun-Wei Lin[3,4(✉)], Fulong Chen[5], Yonglong Luo[5], and Xiaoyao Zheng[5]

[1] Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, Australia
Ji.Zhang@usq.edu.au
[2] Computer Center, University of Thi-Qar, Thi-Qar, Iraq
[3] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
jerrylin@ieee.org
[4] Department of Computing, Mathematics, and Physics, Western Norway University of Applied Sciences (HVL), Bergen, Norway
[5] School of Computer and Information, Anhui Normal University, Wuhu, China

Abstract. Intelligent technologies are enjoying growing popularity in a telehealth environment for helping improve the quality of chronic patients' lives and provide better clinical decision-making to reduce the costs and workload involved in their daily healthcare. Obtaining a short-term disease risk prediction and thereby offering medical recommendations reliably and accurately are challenging in teleheath systems. In this work, a novel medical recommender system is proposed based upon time series data analysis for diabetes patients. It uses three decomposition methods, i.e., dual-tree complex wavelet transform (DTCWT), fast Fourier transformation (FFT) and dual-tree complex wavelet transform-coupled fast Fourier transform (DWCWT-FFT), with least square-support vector machine (LS-SVM) for short-term disease risk prediction for diabetes disease patients which then generates appropriate recommendations on their need to take a medical test or not on the coming day based on the analysis of their medical data. A real-life time series dataset is used for experimental evaluation. The experimental results show that the proposed system yields very good recommendation accuracy and can effectively reduce the workload for diabetes disease patients in conducting daily body tests.

Keywords: Decomposition methods · Recommender system
Diabetes disease patients · Time series prediction · Telehealth · SVM

S. Hartmann et al. (Eds.): DEXA 2018, LNCS 11030, pp. 185–192, 2018.
https://doi.org/10.1007/978-3-319-98812-2_15

1 Introduction

According to World Health Organization (WTO), chronic diseases are causing the death for 50% of people worldwide in recent years [1], and they require more and more medical attentions and resources in today's increasingly aged societies. Diabetes, one of the most common chronic diseases, is a major health problem in the world and the rates of its incidence are significantly rising [10].

Telehealth systems serve as real time and convenient platforms for healthcare practitioners and chronic diseases patients to exchange information easily in consultation, diagnosis and treatment [2], and consequently have enjoined fast developments in many countries in recent years due to fast service delivery and its low operational cost. Due to the importance of disease risk prediction on the patients' life who suffering from the chronic diseases [8] such as diabetes as well as the urgency of improving the analytic techniques used for this regard, great efforts are needed to enhance the quality of evidence-based decisions and recommendations in a telehealth environment. Diabetes disease patients often need to undertake various daily medical tests in order to monitor their overall health conditions through the telehealth system. However, in the current practice, carrying out various medical tests by diabetes disease patients every day may bring lots of inconvenience and even burden, and thus affects their overall life quality.

Generating accurate recommendations is an essential function in telehelth systems, which is often based on the prediction of patients' short-term disease risk. In literature, the assessment and prediction of various diseases have been studied by using data mining techniques and statistical tools for different healthcare and medical issues [3,4]. Although most of these studies have been achieved a reasonable level of predictive accuracy, most of them focused on the long-term medical prediction instead of short-term prediction which is studied in our work.

The major scientific contributions and features of our system are summarized as follows.

- The system utilizes three decomposition methods, including DTCWT, FFT and DWCWT-FFT;
- The statistical features extracted from these methods are then separately input into the Least Square-Support Vector Machine classifier (LS-SVM) to predict the necessity of taking body test on the next day in advance;
- We use a majority vote based ensemble technique to combine the prediction results based on the three individual decomposition methods for producing the final recommendation for diabetes patients;
- We compare our system with the existing work conducted to tackle the exactly same issue to establish the superiority of our technique.

2 Proposed Recommender System

2.1 An Overview of Our System

Figure 1 illustrates the overall architecture of our recommender system used for diabetes patients in the telehealth environment. First, the time series medical

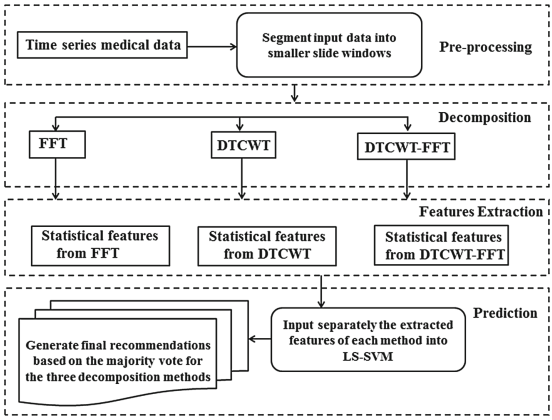

Fig. 1. The architecture of our recommendation system

data of a given patient is pre-processed, which is performed off-line, by segmenting them into smaller overlapped sliding windows based on the size of the sliding window used in the data analysis. Then, the three decomposition methods – DTCWT, FFT and DTCWT-FFT – are separately applied to decompose the segmented time series data of patients. The LS-SVM is used with each decomposition method to test its ability to classify the patient's condition. The final recommendation is then taken based on the ensemble mechanism using the majority vote approach for the three decomposition methods in order to produce a binary accurate recommendation concerning whether the patient needs to take a medical test on the coming day or not.

2.2 Dual Tree Complex Wavelet Transformation

The drawbacks of DWT are ameliorated by using the Dual Tree Complex Wavelet Transformation (DTCWT) which offers a better time-frequency representation of signals [5]. It is an improved version of wavelet transformation that is designed to tackle some limitations in the discrete wavelet transform.

In our system, DTCWT is adopted to decompose the input time series data into sub-bands of *delta, theta, alpha, beta* and *gamma*. Each DTCWT coefficient has two parts real and imaginary. As a result, ten sub-bands in total obtained after four-level decomposition (five sub-bands for each part).

From each frequency sub-band, six different statistical features can be extracted. The extracted features are mean of coefficients of the absolute values, average power of the coefficients, standard deviation of the coefficients, ratio of the absolute mean values of coefficients of adjacent sub-bands, kurtosis of the coefficients and skewness of the coefficients respectively.

2.3 Fast Fourier Transformation

The Fast Fourier Transformation (FFT) is one of the most efficient techniques used to compute the Discrete Fourier Transformation (DFT) and its inverse. In many studies, it is used as a windowing technique like the wavelet transformation [6,9].

For each sliding window, five frequency bands (i.e., *alpha*, *beta*, *gamma*, *delta*, and *theta*) are obtained using the fast Fourier transformation.

From each frequency band, eight different statistical features can be extracted. The extracted features are denoted by X_{Min}, X_{Max}, X_{SD}, X_{Med}, X_{Mean}, X_{RG} (for Range), X_{FQ} (for the First Quartile) and X_{SQ} (for the Second Quartile), respectively. The best performing features are dataset dependent. Some series data are symmetrically distributed while others may have a more skewed distribution.

In our work, the extracted features from each frequency band are grouped into one vector and used as the input to the LS-SVM to predict the patient's condition.

2.4 Hybrid Method of Dual Tree Complex Wavelet Transform with Fast Fourier Transform (DTCWT-FFT)

In this method, we apply 1D dual-tree complex wavelet transform to the input time series data of patients for four-level DTCWT decomposition, and then applies the fast Fourier transform to each DTCWT sub-bands and takes the magnitude of these coefficients. In this way, the levels of the Fourier spectrum vectors are used as a features set and the LS-SVM is used to classify the input time series data into one of two classes: test required or no test required.

The proposed method for generating short-term medical recommendation can be summarized as follows:

1. Get the input time series data of patients $x(n)$ where $n \in [1, N]$.
2. Apply four-levels DTCWT decomposition to the input time series data. Let the output time series be y{1}, y{2}, y{3}, y{4} and z{4} for levels 1, 2, 3, and 4 respectively.
3. Apply forward FFT to y{1}, y{2}, y{3}, y{4} and z{4} and then take the logarithm of the Fourier spectrum. Let the generated features be F{1}, F{2}, F{3}, F{4}, and F{5} respectively.
4. All the generated features vectors enter to the LS-SVM classifier to classify the input time series data of patient.

3 Experimental Results

3.1 Diabetes Dataset

The diabetes dataset obtained from the Repository of Machine Learning Databases by Washington University [7]. The collected data contain measurements taken multiple times per day from 70 patients. Blood glucose measurements, symptoms and insulin treatments were recorded with timestamps for each patient, over the course of several weeks to months.

Each record in the diabetes dataset consists of four fields about the date of measurement, time of measurement, the code of measurement and the value of measurement.

For the purpose of evaluating our system, the dataset is divided into two parts: the training set and the testing set. The three transfer methods are trained using the training set and then validated using the testing set as the ground truth result. In our study, 75% of the dataset was partitioned as the training data while the remaining 25% was used as testing data.

3.2 Performance Evaluation Measurements

To evaluate the performance of the proposed system, we have proposed three performance metrics for this work, namely *accuracy*, *workload saving* and *risk*. Accuracy refers to the percentage of correctly recommended days against the total number of days for which recommendations are provided. Workload saving refers to the percentage of the total number of days when recommendations are provided for skipping the medical test against the total number of days in the training set. Risk refers to the percentage of the days with risky recommendation against the total number of days in the training set.

3.3 Recommendation Effectiveness of Our System

Recommendation Effectiveness of Our System Using FFT. Based on our previous work [8,9], it was found that the obtained prediction results of our system were not good enough when the features were not appropriately selected from a time series data and vice versa. Thus, the statistical features of FFT were tested separately to evaluate the prediction accuracy of the proposed system.

Furthermore, the patient discrimination ability of the eight statistical features $\{X_{Min}, X_{Max}, X_{SD}, X_{Med}, X_{Mean}, X_{RG}, X_{FQ}, X_{SQ}\}$ is performed using t-test. The p-values of the eight statistical features of five waveforms for two different classes including taking a test or not needed using t-test are presented in Table 1. It can be seen that the last four waves (*theta, alpha, beta and gamma*) with the statistical features of rang, mean, median, standard deviation, max, and min provide a significantly difference ($P < 0.003$). Thus, the six features of $\{X_{Min}, X_{Max}, X_{SD}, X_{Med}, X_{Mean}, X_{RG}\}$ of four waves are extracted from FFT sub-bands were used to evaluate the performance of our system for predicting the patient's condition one day in advance. The vector of features is then entered into LS-SVM classifier to decide whether a given patient needs to take a medical measurement on day in advance or not.

Based on the results in Table 3, it can be seen that when using the six statistical features with the four waves, our system can achieve an accuracy over 90%, a workload saving over 63% while the risk is lower than 5 %, indicating that our recommendation system is highly accurate and able to significantly reduce the workload for chronic diabetes disease patients to take up their daily medical tests with a low health risk.

Recommendation Effectiveness of Our System Using DTCWT. In this experiment, the extracted statistical features from the high-frequency sub-bands (i.e., y1, y2, and y3) were also tested separately to evaluate the prediction accuracy of the proposed system.

To achieve the best possible performance of our system, the statistical features extracted from the high-frequency sub-bands (i.e., y1, y2, and y3) are quantified using t-test. Table 2 shows the P-values of the six features extracted from five waveforms of need or not to take a medical test. On the basis of these results, we can be clearly observed that the five statistical features of average power, ratio of mean values, standard deviation, skewness and kurtosis extracted from the sub-bands (*alpha, beta and gamma*) are provide a highly deference ($P < 0.001$). Hence, the five statistical features with the three sub-bands DTCWT are selected to present the time series data of patient.

Our findings showed that combining all the five statistical features for the three high-frequency sub-bands yielded a high prediction accuracy with an average accuracy of 91%, workload saving 63% and risk 4%. The obtained results showed that the statistical features were able to reveal the characteristics of time series data of patients, and to identify patient's condition for short-time disease risk prediction.

Table 1. p-values of statistical features for all waveforms

Waveforms	p-values of X_{FQ}	p-values of X_{SQ}	p-values of X_{SD}	p-values of X_{MAX}	p-values of X_{RG}	p-values of X_{Min}	p-values of X_{Med}	p-values of X_{Mean}
Delta	0.0425	0.0215	6.5214E−09	3.1245E−04	0.0695	0.0425	0.0525	4.2586E−11
Theta	0.0612	0.0325	2.3125E−06	6.2548E−07	0	3.5271E−05	4.2154E−06	0
Alpha	0.4325	0.0345	0	0	5.6243E−05	0	0	0
Beta	0.2154	0.6258	0	5.3641E−09	0	8.2546E−10	0	2.2547E−11
Gamma	0.9457	0.5054	8.9554E−07	0	0	0	0	0

Table 2. p-values of statistical features extracted form the five waveforms of DTCWT

Waveforms	p-values of mean	p-values of average power	p-values of ratio of mean values	p-values of standard deviation	p-values of skewness	p-values of kurtosis
Delta	0.0914	0.0754	0.0541	8.7894E−06	0.0465	0.0254
Theta	0.0456	0.1245	0.3254	0.0312	0.2584	0.4512
Alpha	0.1524	0	0	0	0	5.2364E−10
Beta	0.6324	0	5.8254E−09	0	0	0
Gamma	0.5214	7.5417E−05	0	6.2548E−07	7.5588E−09	0

Recommendation Effectiveness of Our System Using a Hybrid Transform Method (DTCWT-FFT). In order to improve this method, the present study applies the a dual-tree complex wavelet to the input time series data of patients for four-level DTCWT decomposition, and then applies the fast Fourier

transform to each DTCWT sub-bands and takes the magnitude of these coefficients. The extracted features are entered to the LS-SVM classifier to decide whether the patient who is suffering from chronic diabetes disease needs to take a medical body test one day in advance or not.

The proposed system using a hybrid method yields an improved performance compared with the previous two methods. Our system using a hybrid method able to improve the accuracy performance from 90% to 93% while there is no significant difference in the value of workload saving where the workload saving rate is a very close to the two previous methods. The recommendation risk of our system is also lower than the two competitive approaches.

Recommendation Effectiveness Based on the Majority Vote of the Three Decomposition Methods. First, we apply the three methods to process the time series medical data to facilitate the subsequent data analytic. Then, the statistical features extracted from each decomposition method are separately entering into the least square-support vector machine classifier to predict the necessity of taking body test. The final recommendation of a given medical measurement is considered based on applying the majority vote of the three decomposition methods to decide whether the patient who is suffering from chronic diabetes disease needs to take a medical body test one day in advance or not.

Based on Table 3, the proposed method based on majority vote technique had the ability to classify the patients' condition with a high accuracy over 96%, while the risk is lower than 1.5%.

Table 3. The averaged performance of the three decomposition methods and the proposed method

Method used	Accuracy (%)	Saving (%)	Risk (%)
FFT	90.90	63.40	04.75
DTCWT	91.00	63.20	04.30
DTCWT-FFT	93.60	64.00	03.20
Proposed method	**96.00**	**64.50**	**01.50**

4 Conclusions and Future Research Directions

In this work, we propose a recommendation system supported by three decomposition methods including dual-tree complex wavelet transform, fast Fourier transform and dual-tree complex wavelet-coupled fast Fourier transform– to provide the patients suffering from chronic diabetes disease with appropriate

recommendations in a telehealth environment. This study applies three decomposition methods which effectively analyze the medical time series data and input separately the extracted statistical features from each method to the LS-SVM to generate the accurate, reliable recommendations for chronic diabetes disease patients. The final recommendation is taken according to the majority vote of the three decomposition methods.

In future, we will apply other ensemble techniques, such as Adaboost and boosting, to generate recommendations and conducting a comparative study on those different ensemble models.

Acknowledgment. This research was partially supported by Guangxi Key Laboratory of Trusted Software (No. kx201615), Shenzhen Technical Project (JCYJ20170307151733005 and KQJSCX20170726103424709), the general research project of National Science Foundation of China (No. 61572036, No. 61672039, No. 61772034) and Anhui Provincial Natural Science Foundation (1808085MF172).

References

1. Kuh, D., Shlomo, Y.B.: A Life Course Approach to Chronic Disease Epidemiology. Oxford University Press, Oxford (2004)
2. Dewar, A.R., Bull, T.P., Malvey, D.M., Szalma, J.L.: Developing a measure of engagement with telehealth systems: the mHealth technology engagement index. J. Telemed. Telecare **23**, 248–255 (2017)
3. Mohktar, M.S., et al.: Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. Artif. Intell. Med. **63**(1), 51–59 (2015)
4. Krishnaiah, V., Narsimha, D.G., Chandra, D.N.S.: Diagnosis of lung cancer prediction system using data mining classification techniques. Int. J. Comput. Sci. Inf. Technol. **4**(1), 39–45 (2013)
5. Das, A.B., Bhuiyan, M.I.H., Alam, S.S.: Classification of EEG signals using normal inverse Gaussian parameters in the dual-tree complex wavelet transform domain for seizure detection. Signal Image Video Process. **10**(2), 259–266 (2016)
6. Deo, R.C., Wen, X., Qi, F.: A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy **168**, 568–593 (2016)
7. AIM-94 data set provided by Michael, K., MD. Ph.D. Washington University, St. Louis, MO, US. https://archive.ics.uci.edu/ml/datasets/diabetes
8. Lafta, R., et al.: An intelligent recommender system based on predictive analysis in telehealthcare environment. Web Intell. **14**(4), 325–336 (2016). IOS press
9. Lafta, R., et al.: A fast fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS (LNAI), vol. 10234, pp. 654–670. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57454-7_51
10. Temurtas, H., Yumusak, N., Temurtas, F.: A comparative study on diabetes disease diagnosis using neural networks. Expert Syst. Appl. **36**(4), 8610–8615 (2009)