Improving the Probability of Clinical Diagnosis of Coronary-Artery Disease Using Extended Kalman Filters with Radial Basis Function Network

Mashail Alsalamah(✉) and Saad Amin

Faculty of Engineering and Computing, Coventry University,
Priory Street, Coventry CV1 5FB, UK
alsalam2@coventry.ac.uk

Abstract. Kalman filters have been popular in applications to predict time-series data analysis and prediction. This paper uses a form of Extended Kalman Filter to predict the occurrence of CAD (Coronary Artery Disease) using patients data based on different relevant parameters. The work takes a novel approach by using different neural networks training algorithms Quasi-Newton and SCG with combination of activation functions to predict the existence/non-existence of CAD in a patient based on patient's data set. The prediction probability of this combination is resulted in accuracy of about 92% or above, using cross validation and thresholding to remove the limitation of time-series prediction introduced because of the Extended Kalman filter behavior.

Keywords: Coronary artery disease · Extended kalman filter · Radial basis function · Quasi-Newton and Scaled Conjugate Gradient

1 Introduction

Medical diagnosis is a vast subject and there have been several ways and methodologies, which are pure medical and depend upon the physicians. However, with advancement in computing technologies and artificial intelligence lot have been done to diagnose complex diseases using available data and patient information and symptoms. CAD is one such disease, which is a complex condition of artery blockage with high mortality figures [1] and it is not straight forward for ordinary staff to detect this disease.

A host of factors are used in the diagnosis of CAD, and these include patient's blood pressure, cholesterol, sugar levels, high BMI (overweight/obese), physical inactivity, unhealthy eating, and smoking [2]. Other factors such as age, gender, and family history of heart disease are also likely risk factors for CAD (British Heart Foundation, 2015; National Heart, Lung and Blood Institute, 2015) [3].

This research explores the application of a data fusion algorithm (Kalman filtering) in diagnosing CAD, as well as in the prediction of the need to conduct a Coronary Artery Bypass Graft (CABG) in patients identified as having CAD.

P. Perego et al. (Eds.): MobiHealth 2016, LNICST 192, pp. 269–277, 2017.
DOI: 10.1007/978-3-319-58877-3_35

Kalman filters (KF) [4] are generally used for applications such as navigation, as their ability to accurately predict next points based on the history of tracking and travel; the success of these algorithms lie in the fact that the model considers measurement noise between present and future points. The novel application of data fusion algorithm in this research precisely entails the derivation of Kalman filter equations that facilitate the recursive calculation of two terms/factors of interest stated earlier (i.e., CAD status and CABG requirement). This process uses a combination of knowledge/observations/measurements from patients, predictions from systems models and inherent noise in the observations/measurements.

The rest of the paper is divided into four sections; Sect. 2 covers the literature review to describe the work done in detection of CAD and use of Kalman Filtering in disease diagnosis; Sect. 3 will discuss the methodology and principles of this novel approach of the Extended Kalman filtering and radial basis functions. Section 4 will enlighten with the procedure of the testing and verification of the patient data set. Section 5 will present the results and analysis of the test runs and the combinations of different training and activation algorithms. Finally Sect. 6 summarizes the work and gives pointers towards the future work.

2 Literature Review

Nonetheless, over the years, Automatic Computer-Assisted Detection processes have been used in the diagnosis of CADs. With linear and logistic regression model frequently used [6–10]. Other commonly used predictive models are the Linear Discriminant Analysis, K-nearest Neighbor Classifier, Artificial Neural Network and Support Vector Machine [11–13]. These models have to some extent been shown to be of some predictive value. For example, [16] showed that the prediction accuracy of training and test sets of Linear Discriminant Analysis could be as high as 90.6% and 72.7% respectively, while [11] showed that Artificial Neural Networks could produce accuracy in test set that is as high as 81.2%. In spite these reasonable results; there are apparent limitations in most of these learning algorithms. The Linear Regression and the Linear Discriminant Analysis are both linear techniques and cannot be extended to consider non-linear modalities (variables) which are inevitable in the proper diagnosis of CAD. As a way of circumventing this issue, research in this field has also considered the use of more complex models such as the combination of Support Vector Machine with a Radial Basis Function (RBF) Kernel, Support Vector Machine optimized by particle swarm optimization or other forms of integration of two individual approaches to generate a non-linear technique [14–16]. The result of this is an improvement in the prediction accuracy of training and test sets (for example, as high as 96.9% [11]).

These facts show that there are obviously more room for improvements. One important way of going about this process of improvement, which is yet to be well exploited, is by enhancing the quality of the data sets used in the prediction. This could be done by developing novel means that reduce noise and inconsistency in data. Measures of capability or predictability of algorithms such as validation error are affected by the variations in data (e.g., high level of missing data, which is inherent in the sets of data often used in CAD diagnosis/prediction as these data sets come from multiple sources – oral interviews, doctors' examinations and technical measurements with different instruments). A potential candidate for improving data quality is the Matrix Completion, which is a process that entails the addition of entries to a matrix which contains some unknown/missing values. Research has shown that how the Matrix Completion could greatly help enhance the accuracy of prediction [17, 18].

Furthermore, another step in the advancement of this realm of research could also be in the potentially novel adoption and application of hybrid Extended Kalman Filter (EKF) in the prediction/diagnosis of CAD. The Kalman Filter (KF) is a well-established estimation theory that has been in existence since the 1960 s. Though the initially designed filter provides recursive solution through a linear optimal filtering for estimating desired parameters, the extended version of the filter (i.e., Extended Kalman Filter – EKF) has the capability to handle non-linear systems/conditions [19].This learning algorithm has been used in diverse research realms and has shown excellent results in terms of prediction accuracy [21, 22]). Nonetheless, the EKF's alluring capabilities are yet to be explored in the area of Coronary Artery Disease prediction. This is most likely due to the lack of awareness about its existence, as majority of researchers in this realm of research and beyond are more familiar with the KF which can only provide a recursive solution through a linear means.

3 Methodology

The suitability of the RBF (Radial Basis Function) Neural network models for classification problems such as detection of diseases, in our case occurrence of CAD. There has been several steps involved in the determination of this classification problem, in this section a brief on the methodology steps is provided (Fig. 1).

A. Data set Classification

The study is based on Saudi Arabia population, in King Abdullah Medical City (KAMC). The obtained data set needs to be classified and pre-processed before been applied to the RBF network. The purpose of this work is to find a relationship between existence/non-existence of CAD based on the variables; namely, demographic variables like age, gender, occupation, physical variables like height, weight, smoking habit, medical history among others. There were 59 independent variables in the data. The frequency distribution of each of them were studied to ensure further modifications, if any, to fit in the model and to eliminate data entry errors. Mismatched entries were found and were considered as "no information" and tagged as 0 in required cases. The details of the analysis and classification of data based on their statistics is author's work in [19].

B. Matrix Completion

Matrix completion has been used as secondary techniques to condition the patient data. Due to several recording, interview and manual entry deficiencies the patient data has some anomalies, in certain cases missing of several or one of the major contributing fields. One of the possible solution so that the erroneous or missing data does not disturb the estimation process a simpler way is to discard the patients information,

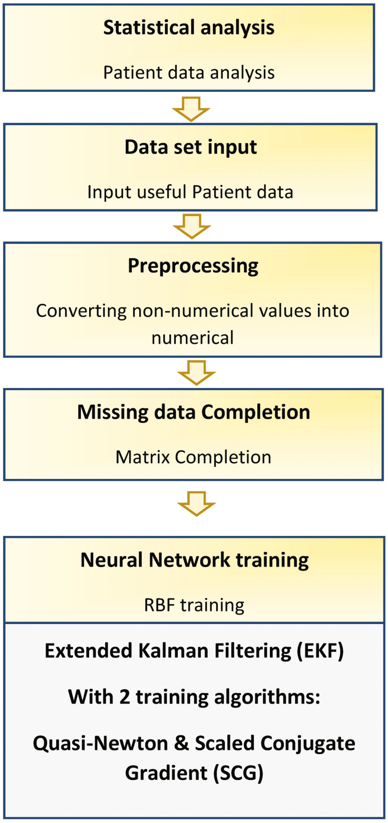

Fig. 1. Methodology overview

however due to scarcity of the patient information from the source (check acknowledgement section), there was a requirement not to discard any such data. Therefore Matrix completion techniques were used to improve the prediction efficiency. Although this has not provided a high percentage of accuracy, though helped to use patient information with less number of missing fields and higher confidence in the MC results have been included in the study. Exact Matrix Completion via Convex Optimization is used in this work to improve the data quality and availability.

C. Kalman Matrix

The Kalman Filter gain is a time-varying gain matrix. It is given by the algorithm presented below. In the expressions below the following matrices are used and shown in Fig. 2.

- Auto-covariance matrix (for lag zero) of the estimation error of the corrected estimate:
- Auto-covariance matrix (for lag zero) of the estimation error of the predicted estimate:

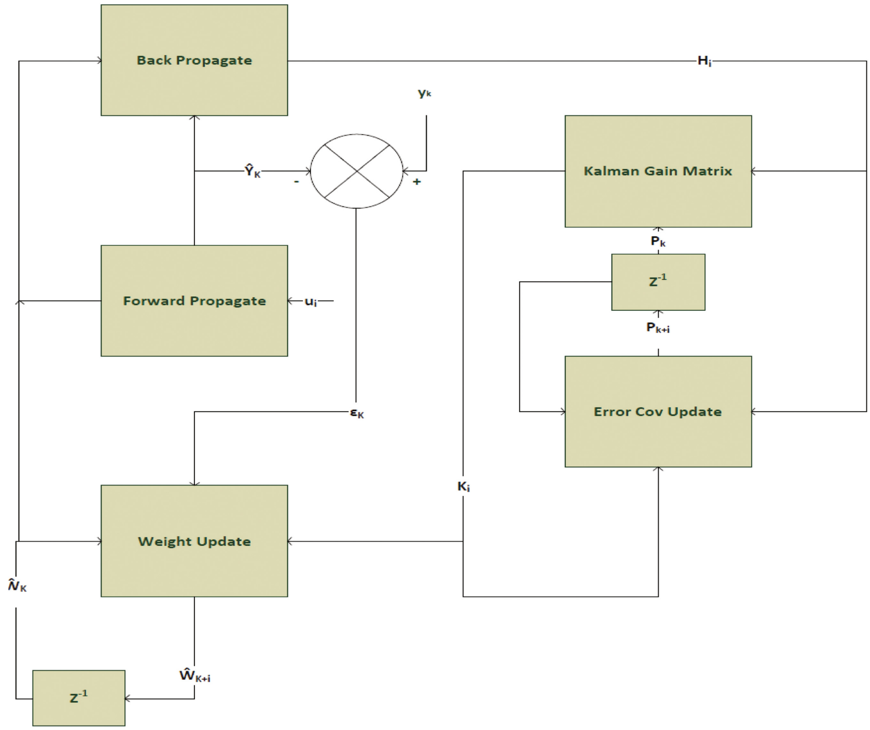

Fig. 2. Calculation of Kalman filter gain

- The transition matrix A of a linearized model of the original nonlinear model calculated with the most recent state estimate, which is assumed to be the corrected estimate xc(k):
- Given the continuous-time nonlinear process model. Linearize it at the operating point to obtain
- Then calculate A = A discrete as the discretized version of A continuous. Forward method of discretization in manual calculations can be used; however in this case Matlab is used to discretize this function.

4 Procedure

The main theme of this classification solution is to optimize the network and then validate the results. This is mainly divided into three steps i.e. train test and validate.

If enough data is available then we can divide the data for training and testing based on the inputs and the outputs. This way several models trained on the training set will be available to be applied on the test set. The best result on the test set based on these several trained models is then considered as the optimal simulation.

D. Cross Validation

The discussed above can introduce a bias towards a particular data set. Therefore, in this case the data set can be portioned and swapped as testing and training sets to negate this bias and also to calculate an average based on these different partitions.

The data set in this system has been divided to three sections training, testing and validation.

5 Observations and Results

The experiment uses the hospital data set and makes three different combinations, by swapping the training, testing and validation sets. In this experiment the two different training algorithms will be used with the different combinations of the data sets to see the performance of the training algorithms

E. Without Matrix Completion

The results of the four different combinations are given in the Fig. 3 and the Table 1 shows the details, including training errors, testing errors, validation error and the best iteration in each training algorithm for each of the three sets.

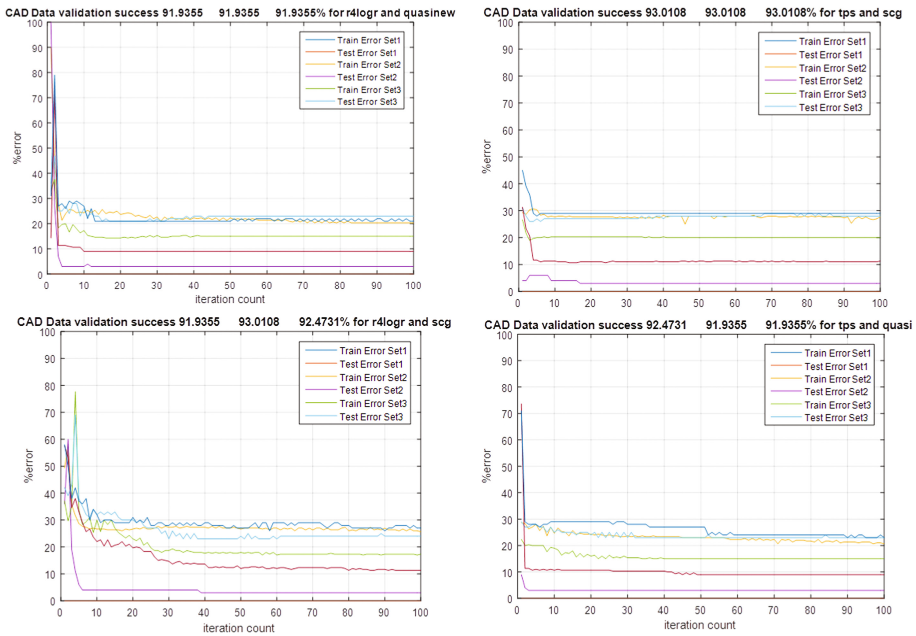

Fig. 3. Three combinations of data sets and 4 different training and activation functions without matrix completion

Table 1. Summary of the best result obtained from the observations shown for Fig. 1

Combinations of training algorithms & activation functions		Best results of Data sets														
		Set No 1					Set No 2					Set No 3				
Training Algo	Act Fcn	Train Error %	Test Error %	Valid Error %	Accuracy %	Best Iter	Train Error %	Test Error %	Valid Error %	Accrcy %	Best Iter	Train Error %	Test Error %	Valid Error %	Accuracy %	Best iter
Quasi New	R4R	21.33	3.00	8.06	91.9%	4	14.67	21	8.06	91.9%	15	9.00	21.00	8.06	91.9%	13
Quasi New	Tps	26.33	3.00	7.53	92.47%	3	15.00	23.00	8.06	91.9%	22	9.00	23.00	8.06	91.9%	90
SCG	R4R	27.33	3.00	6.99	93.01%	30	27.33	3.00	8.06	91.4%	39	12.00	26.00	7.53	92.47%	58
SCG	Tps	27.67	3.00	6.99	93.01%	17	19.00	26.00	6.99	93.01%	3	11.67	28.00	6.99	93.01%	5

*Training Algo: Training Algorithm
*Act Fcn: Activation Function
*Valid Error: Validation error
*Best iter: Best Iteration

F. With Matrix Completion

The results of the four different combinations are given in the Fig. 4 and the Table 2 shows the details, including training errors, testing errors, validation error and the best iteration in each training algorithm for each of the three sets after Matrix completion.

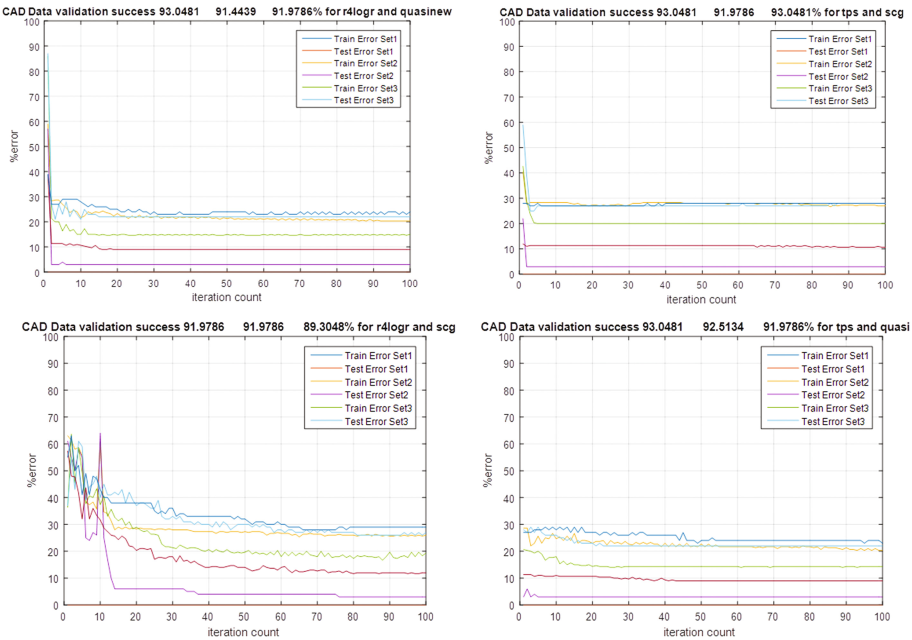

Fig. 4. Three combinations of data sets and 4 different training and activation functions with matrix completion

Table 2. Summary of the Best result obtained from the observations shown for Fig. 2

Combinations of training algorithms & activation functions		Best results of Data sets														
		Set No 1					Set No 2					Set No 3				
Training Algo	Act Fcn	Train Error %	Test Error %	Valid Error%	Accuracy %	Iter no	Train Error %	Test Error %	Valid Error%	Accuracy %	Iter no	Train Error %	Test Error %	Valid Error %	Accuracy %	Best iter
Quasi New	R4R	28.33	3.00	6.95	93.0%	2	20.00	21.00	8.56	91.4%	3	9.00	23.00	8.02	91.9%	29
Quasi New	Tps	28.67	3.00	10.70	93.0%	1	14.67	22.00	7.49	92.5%	21	9.00	23.00	8.02	91.9%	49
SCG	R4R	26.00	3.00	8.02	91.9%	76	11.33	27.00	8.02	91.9%	81	12.33	28.00	10.70	89.3%	66
SCG	Tps	27.67	3.00	6.95	93.0%	2	23.67	25.00	8.02	91.9%	3	11.33	27.00	6.95	93.0%	3

*Training Algo: Training Algorithm
*Act Fcn: Activation Function
*Valid Error: Validation error
*Best iter: Best Iteration

6 Conclusion

This work deals with the prediction of Coronary artery disease by applying the patient information as a set of data feed to the RBF neural network. The training of these neural networks is based on the prediction strength of Extended Kalman filters and combination of training algorithms. The work still need to be investigated and that suggests that it's possible that ensemble of the classifiers trained on different part of the dataset might have a greater performance.

Acknowledgment. We are very thankful to King Abdullah Medical City in Saudi Arabia for providing the patient data to be used in this study.

References

1. WHO: WHO World Health Organization, January 2015. http://www.who.int/mediacentre/factsheets/fs317/en/. Accessed 6 Mar 2016
2. NIH: What is Coronary Heart Disease, National Institute of Health, 3 October 2015. http://www.nhlbi.nih.gov/health/health-topics/topics/cad. Accessed 6 Mar 2016
3. British Heart Foundation: Risk factors (2015). https://www.bhf.org.uk/heart-health/risk-factors. Accessed 6 Mar 2016
4. Wan, E.A., Van Der Merwe, R.: The unscented Kalman filter for nonlinear estimation. In: The IEEE Conference on Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000, AS-SPCC (2000)
5. Julier, S.J., Uhlmann, J.K.: New extension of the Kalman filter to nonlinear systems. In: AeroSense 1997 International Society for Optics and Photonics (1997)
6. Genders, T.S.S., Steyerberg, E.W., Hunink, M.G.M., Nieman, K.: Prediction model to estimate presence of coronary artery disease: Retrospective pooled analysis of existing cohorts. Br. Med. J. **344**, 1–13 (2012)
7. Yamada, H., Do, D., Morise, A., Atwood, J.E., Froeliche, V.: Review of studies using multivariable analysis of clinical and exercise test data to predict angiographic coronary artery disease. Prog. Cardiovasc. Dis. **39**, 457–481 (1997)

8. Pryor, D.B., Harrell, F.E., Lee, K.L., Califf, R.M., Rosati, R.A.: Estimating the likelihood of significant coronary artery disease. Am. J. Med. **75**, 771–780 (1983)
9. Diamond, G.A., Forrester, J.S.: Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. N. Engl. J. Med. **300**, 1350–1358 (1979)
10. Kalman, R.: A new approach to linear filtering and prediction problems. J. Basic Eng. **83**, 35–45 (1960). Transaction of the ASME
11. Heydari, S.T., Ayatollahi, S.M.T., Zare, N.: Comparison of artificial neural networks with logistic regression for detection of obesity. J. Med. Syst. **36**(4), 2449–2454 (2012)
12. Hongzong, S., Tao, W., Xiaojun, Y., Huanxiang, L., Zhide, H., Mancang, L., BoTao, F.: Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease. West Indian Med. J. **56**(5), 451–457 (2007)
13. Hedeshi, N.G., Abadeh, M.S.: Coronary artery disease detection using a fuzzy-boosting PSO approach. Comput. Intell. Neurosci 6 (2014)
14. Karabulut, E.M., İbrikçi, T.: Effective diagnosis of coronary artery disease using the rotation forest ensemble method. J. Med. Syst. **36**, 3011–3018 (2012)
15. Oh, S.: Matrix Completion: Fundamental Limits and Efficient Algorithms. Stanford University, California (2010)
16. Wu, T.T., Lange, K.: Matrix completion discriminant analysis. Comput. Stat. Data Anal. **92**, 115–125 (2015)
17. Arif, M., Malagore, I.A., Afsar, F.A.: Detection and localization of myocardial infarction using K-nearest neighbor classifier. J. Med. Syst. **36**, 279–289 (2010)
18. Comak, E.: A biomedical decision support system using LS-SVM classifier with an efficient and new parameter regularization procedure for diagnosis of heart valve diseases. J. Med. Syst. **36**, 549–556 (2010)
19. Salamah, M.: The Statistic analysis Study of Coronary-Artery Disease Data Based on King Abdullah Medical City in Saudi Arabia (KAMC-CAD). Coventry (2016)
20. Haykin, S.: Kalman Filtering and Neural Networks. Wiley, New York (2001)
21. Sepasi, S., Ghorbani, R., Liaw, B.Y.: Improved extended kalman filter for state of charge estimation of battery pack. J. Power Sources **255**, 368–376 (2014)
22. Sun, X., Jin, L., Xiong, M.: Extended Kalman filter for estimation of parameters in nonlinear state-space models of biochemical networks. PLoS ONE **3**(11), 1–13 (2008)