Prediction Models for a Smart Home based Health Care System

Vikramaditya R. Jakkula[1], Diane J. Cook[2], Gaurav Jain[3].
Washington State University,
School of Electrical Engineering and Computer Science,
Pullman, WA 99164, USA.

Abstract — Technology holds great potential for improvements in the field of health care used in intelligent environments, with homes becoming the centers for proactive health care. Smart health care systems at home can be used to provide such solutions. A technology-assisted smart health care system would enable elderly people to lead their independent lifestyle away from hospitals and also avoid having expensive caregivers. In this paper, we present one such solution where a prediction model in an intelligent smart home system can be used for identifying health trends over time and enable prediction of future trends which can aid in providing preventive measures.

INTRODUCTION

Today in the area of health care, a major issue is the provision of adequate and effective health for the elderly, as people aged 65 and older are the fastest growing segment of the US population; by 2030 more than 4 million Americans will be over the age 85 [1]. Studies reveal that older Americans prefer an independent lifestyle [2]. With a growing aging population that desires to maintain their independent living, the need for smart and cost effective, in-home, health care technology to replace existing home care givers becomes critical. The next generation of smart in-home systems built for health care would enable people to lead a healthier life. The major challenge for these systems would be early prediction of health variations over time, which may act as early indicators for chronic diseases.

Time series analysis is often associated with discovery of patterns (such as frequent sequences or sequences appearing with increasing or decreasing regularity) and prediction of future values (forecasting). In this paper we employ machine learning techniques using Weka, for predicting trends of health data over time, and also use Box-Jenkins seasonal arima models for forecasting health data ranges over time. The Box Jenkins seasonal arima model is frequently used for analysis and forecasting time series data [3]. We hypothesize that the Box-Jenkins arima method can be effective for analyzing health data in an intelligent environment. To validate our hypothesis, we test the method using health data collected from MavHome smart home project.

ENVIRONMENT SENSING AND DATA COLLECTION

The major goal of the MavHome project [4, 5] is to design a smart environment that acts as an intelligent agent. We define a smart environment as one with the ability to adapt the environment to the inhabitants and meet the goals of comfort and efficiency. In order to achieve these goals the house should be able to predict, reason, and adapt to its inhabitant. In MavHome, sensor network data is the primary source of data collection. The data collection system consists of an array of motion sensors which collect information through the Argus sensor network [6]. In this study, we enhance collected information with critical health parameters (for example, blood pressure and pulse) collected using digital instruments. The data collected for this study spanned a period of sixty one days, based on a single inhabitant living in the MavPad on-campus apartment (see Figure.1). Our evaluation environment is a student apartment with deployed Argus and X-10 networks. The apartment consists of a living/dining room, kitchen, bath room, and bed room. There are over 150 sensors deployed in the MavPad that include light, temperature, humidity, and reed switches.

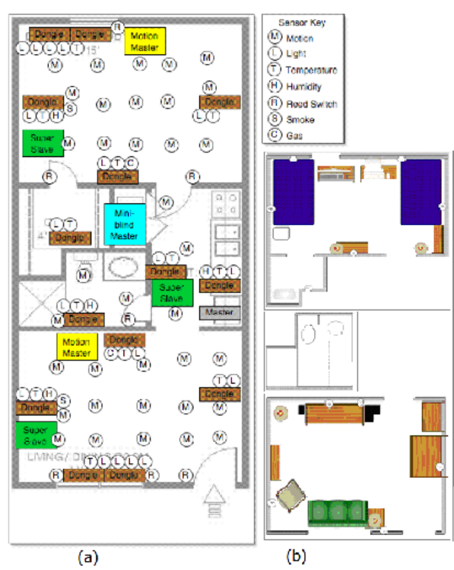

Figure 1: (a) MavHome Argus Sensor Network
(b) MavHome Apartment Environment.

EXPERIMENTATION EVALUATION

The goal of this experiment is to evaluate a series-based forecasting model which would be a part of smart healthcare systems in smart homes. The expectations from this experiment are better predictive accuracy with a limited data from the smart healthcare system using time series data. These predictions would enable the care-givers and patients to prepare for possible critical situations before they actually occur. The basic idea behind self-projecting time series forecasting models is to find a mathematical formula that will approximately generate the historical patterns in a time series. A time series is a historical record of some activity, with measurements taken at equally-spaced intervals with a consistency in the activity and the method of measurement. Box-Jenkins forecasting models are based on statistical concepts and principles and are able to model a wide spectrum of time series behaviors [7]. We use Box-Jenkins methodology in analysis and forecasting rather than developing our own because the Box-Jenkins methodology is widely regarded to be the most efficient forecasting technique, and is used extensively, especially for univariate time series [10]. Compared to other techniques, Box-Jenkins forecasting provides some of the most accurate short-term forecasts. However, a limitation is that it requires a large amount of data which is a current problem leading to lower accuracy.

EXPERIMENTATION AND RESULTS

The experiment consists of two parts. The first part uses a support vector machine algorithm found in Weka [8] to predict whether health trends are increasing, decreasing or constant. The second experiment utilizes automated forecasting tools, such as Phicast [9], which enables us to use the Box-Jenkins method for forecasting the range of health data values.

The first experiment employs the SVM using Weka to predict trends in the health data values over time. We are looking to predict an increasing, decreasing or constant trend over time. For this experiment we use a training set consisting of 40 days of health parameter data which was collected from the inhabitant in the MavHome apartment. For the second experiment, we included observations from one additional day as it was made available. Thus for the second experiment we did use 41 days of health parameter data. In our prediction experiment, we use observed trend accuracy as the performance measure. This is defined as the number of correctly classified observed instances divided by the total number of observed instances. We also look at the error rates, which is different from trend accuracy and is defined as the incorrectly classified instances divided by the total classified instances in the observed output test set data. Most of the experiment is run using the Weka environment and in all of the experiments reported here percentage split was used as the evaluation technique. This consists of dividing the data into two subgroups. The first subgroup, called the training set, is used for building the model for the classifiers. The second subgroup, called the test set, is used for calculating the accuracy of the constructed model.

The subgroup or the test set consisted of 20 data points for experiment one and 21 data points for experiment two. The total dataset consisted of 41 data points. We also did perform 3-fold cross validation on the datasets for experiment one only and the observed results are given below in Table 4. We have chosen 3 fold cross validation because of the size of the dataset used for the experimentation.

Table 1: Predicted systolic trend observations.

Inst No.	Actual	Predicted	Error
1	2:decrease	2:decrease	No
2	1:increase	1:increase	No
3	1:increase	1:increase	No
4	2:decrease	2:decrease	No
5	2:decrease	2:decrease	No
6	2:decrease	2:decrease	No
7	1:increase	1:increase	No
8	1:increase	1:increase	No
9	1:increase	1:increase	No
10	2:decrease	2:decrease	No
11	2:decrease	2:decrease	No
12	1:increase	1:increase	No
13	2:decrease	1:increase	Yes
14	1:increase	1:increase	No
15	2:decrease	2:decrease	No
16	2:decrease	2:decrease	No
17	2:decrease	2:decrease	No
18	2:decrease	2:decrease	No
19	1:increase	2:decrease	Yes
20	2:decrease	2:decrease	No

Table.1 display's the predicted systolic data trends and compares them with actual trends of the data collected from the inhabitant. We observe the accuracy to be fairly high with an error rate of 10%. Table 2 displays predicted trends in diastolic data values and compares them with actual trends of the data collected from the inhabitant. We observe that the accuracy was less and error rate was 67%, because this method implements a sequential minimal optimization algorithm for training a support vector classifier and converts nominal values to binary for prediction. In Table 3 we compare predicted trends to actual trends in pulse rate values on the data collected from the inhabitant. These predictions have a high accuracy and an error rate of 5%.

Table 2: Predicted diastolic trend observations.

InstNo.	Actual	Predicted	Error
1	2:decrease	1:increase	Yes
2	1:increase	1:increase	No
3	1:increase	1:increase	No
4	2:decrease	1:increase	Yes
5	2:decrease	1:increase	Yes
6	3:constant	1:increase	Yes
7	1:increase	1:increase	No
8	1:increase	1:increase	No
9	1:increase	1:increase	No
10	2:decrease	1:increase	Yes
11	2:decrease	1:increase	Yes
12	3:constant	1:increase	Yes
13	1:increase	2:decrease	Yes
14	1:increase	1:increase	No
15	2:decrease	1:increase	Yes
16	1:increase	1:increase	No
17	2:decrease	1:increase	Yes
18	2:decrease	1:increase	Yes
19	2:decrease	1:increase	Yes
20	1:increase	1:increase	No

Table 3: Predicted pulse trend observations.

Instance No	Actual	Predicted	Error
1	2:decrease	2:decrease	No
2	1:increase	1:increase	No
3	1:increase	1:increase	No
4	2:decrease	2:decrease	No
5	2:decrease	2:decrease	No
6	2:decrease	2:decrease	No
7	1:increase	1:increase	No
8	1:increase	1:increase	No
9	1:increase	1:increase	No
10	2:decrease	2:decrease	No
11	2:decrease	2:decrease	No
12	1:increase	1:increase	No
13	1:increase	2:decrease	Yes
14	1:increase	1:increase	No
15	2:decrease	2:decrease	No
16	2:decrease	2:decrease	No
17	2:decrease	2:decrease	No
18	2:decrease	2:decrease	No
19	1:increase	1:increase	No
20	2:decrease	2:decrease	No

We also have performed cross validation evaluation on the experiment testing data and the obtained outputs are displayed in Table 4.

Table 4: Observation of cross validation on datasets.

	Systolic	Diastolic	Pulse
Correctly Classified Instances	19	15	19
Incorrectly Classified Instances	1	5	1
Mean absolute error	0.2333	0.3111	0.2333

The second experiment evaluates the ability of ARIMA methods to perform effective prediction or forecasting of health data value ranges.. This involves the use of an automated forecasting tool, which predicts the range of values for the time series. The experiment compares predicted values with actual test data. The performance is then measured as the ratio of test cases predicted accurately. The figures plotted for comparing the predicted range with the actual data values are illustrated in fugures 2,3 and 4. In figure 2 we compare the collected systolic data values to the predicted range values over the time series data. We observe that this had 62% accuracy in the prediction. Note the observations summarized in Table 5.

Table 5: Comparing the Actual Value to the Predicted Range of the Systolic Data values collected.

Instance #	Actual Recorded Systolic	Lower Predicted Value	Upper Predicted Value	Error [Yes/No]
1	134	126.4	142.1	No
2	129	127.93	144.52	No
3	132	128.13	144.44	No
4	134	128.32	144.56	No
5	126	128.1	144.7	Yes
6	123	130.11	144.8	Yes
7	119	129.32	145.34	Yes
8	122	128.2	145.37	Yes
9	135	129.33	145.6	No
10	139	127.76	145.45	No
11	133	126.63	145.44	No
12	132	129.11	145.78	No
13	134	128.21	146.34	No
14	124	129.34	146.37	Yes
15	153	129.57	146.45	Yes
16	140	129.76	146.76	No
17	137	130.95	146.95	No
18	135	131.2	147.15	No
19	102	131.34	147.34	Yes
20	145	131.53	147.5	No
21	123	131.73	147.73	Yes

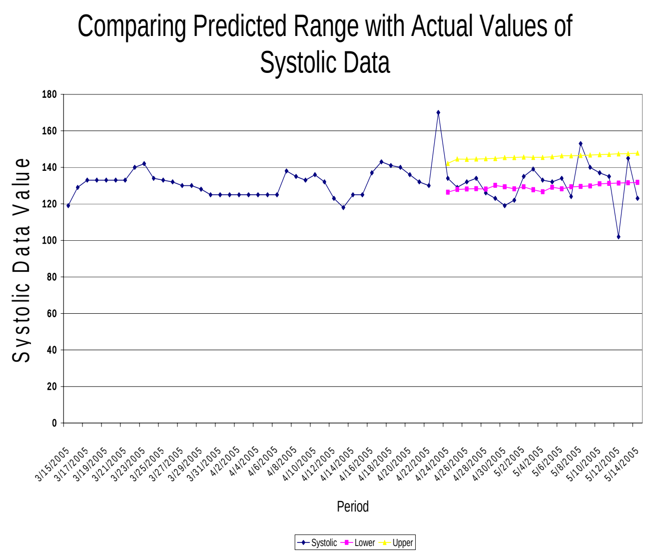

Figure.2: Comparing Predicted Range with Actual Values of Systolic Data.

In Figure 3, we compare the collected diastolic data values to the predicted range values over the time series data and observe that it had 62% accuracy for predicted values. The observations are summarized in Table 6.

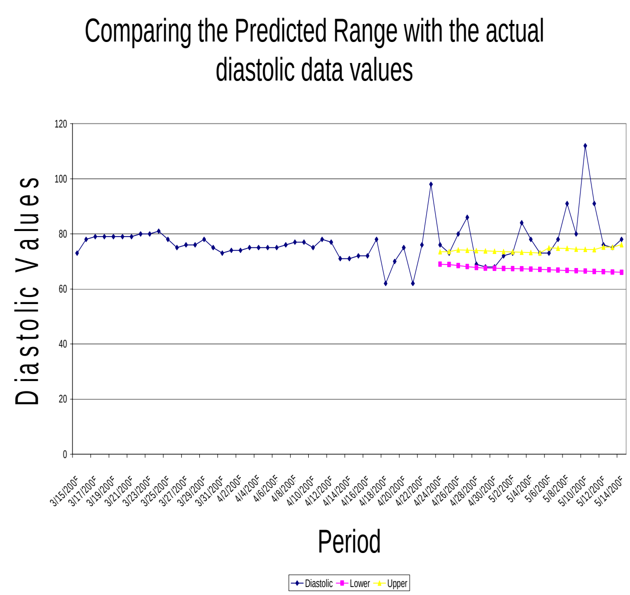

Figure.3: Comparing Predicted Range with Actual Values of Diastolic Data.

Table 6: Observations of comparing the Actual Value to the Predicted Range of the Diastolic Data values collected from the inhabitant.

Instance #	Actual Recorded Diastolic	Lower Predicted Value	Upper Predicted Value	Error [Yes/No]
1	76	69	73.5	Yes
2	73	68.88	73.56	No
3	80	68.45	74.15	Yes
4	72	68.13	74.04	No
5	69	67.78	73.92	No
6	68	67.56	73.8	No
7	68	67.5	73.68	No
8	72	67.43	73.56	No
9	73	67.38	73.44	No
10	84	67.32	73.32	Yes
11	72	67.21	73.21	No
12	73	67.09	73.09	No
13	73	66.97	74.93	No
14	78	66.85	74.78	Yes
15	91	66.73	74.73	Yes
16	80	66.61	74.45	Yes
17	112	66.49	74.36	Yes
18	71	66.37	74.27	No
19	76	66.26	75.22	No
20	75	66.14	75.18	No
21	78	66.02	76.08	Yes

In Figure 4 we compare collected pulse data values to the predicted range. We observe that this had 76% predictive accuracy. The observations are summarized Table 7 and a comparison is shown in Figure 4.

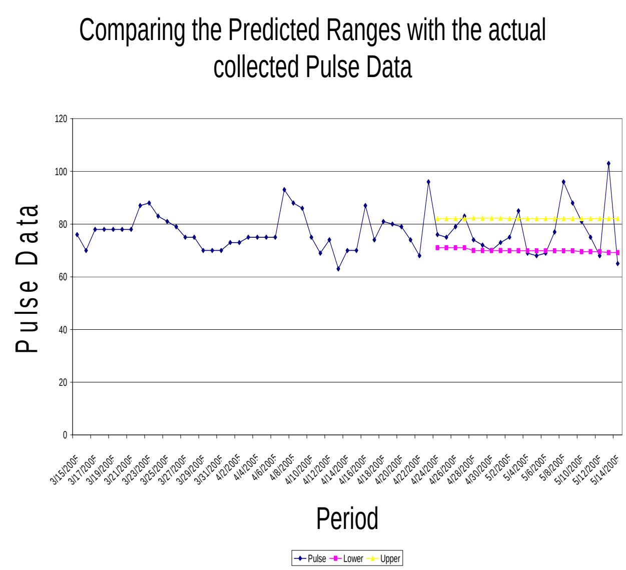

Figure.4: Comparing Predicted Range with Actual Values of Pulse Data.

The observations made from Table 5 reveal that the accuracy was 62%, Table 6 shows an accuracy of 62% and Table 7 shows an accuracy of 76%. The accuracy was low as this method needs more data for accurate predictions. Data is collected continuously, so as more data is available we expect to see an increase in the predictive accuracy. The data does include a few outliers which also affect the accuracy of the results. The first model has higher accuracy, thus making it more effective than the second model which predicts the health data ranges. These models can be more effective when used together.

Table 7: Observations of comparing the Actual Value to the Predicted Range of the Pulse Data values collected from the inhabitant.

Instance #	Actual Recorded Pulse	Lower Predicted Value	Upper Predicted Value	Error [Yes/No]
1	76	71	82	No
2	75	71	82	No
3	79	71	82	No
4	82	71	82	No
5	74	69	82	No
6	72	69	82	No
7	70	69	82	No
8	73	69	82	No
9	75	69	82	No
10	85	69	82	Yes
11	69	69	82	No
12	71	69	82	No
13	69	69	82	No
14	77	69	82	No
15	96	69	82	Yes
16	88	69	82	Yes
17	81	69	82	No
18	75	69	82	No
19	70	69	82	No
20	103	69	82	Yes
21	65	69	82	Yes

CONCLUSION AND FUTURE WORK

Generally, predicting time series data is difficult. In this case, however, we show that there was good overall prediction performance. We have used the Box-Jenkins method with relatively small amounts of data, keeping in mind the use of this tool over time, as the amount of data continuously increases. As this method performs its best with large data [10], the performance will steadily improve. We observe that the prediction models act as useful components to the health care system in smart homes. Future work would include improving the prediction, collecting more data over time which would help improve prediction and detecting and avoiding outliers to increase the prediction accuracy. Providing accurate tele-health care systems in smart homes, would be a breakthrough for the healthcare industry.

ACKNOWLEDGEMENT

This work is supported by NSF grant IIS-0121297.

REFERENCES

[1]. Anderson, R. N. Method for Constructing Complete Annual U.S. Life Tables. National Center for Health Statistics. Vital and Health Stat, Series 2(129) (1999)

[2]. T. K. Hareven. Historical Perspectives on Aging and Family Relations. 2001. Handbook of Aging and the Social Sciences. 5th Edition. 141-159.

[3].Box, G.E.P. and G.M. Jenkins. Time series analysis: Forecasting and control, San Francisco: Holden-Day.(1970)

[4]. G. Michael Youngblood, Lawrence B. Holder, and Diane J. Cook. Managing Adaptive Versatile Environments.Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom).(2005)

[5]. D. J. Cook, M. Youngblood, E. Heierman, K. Gopalratnam, S. Rao, A. Litvin, F. Khawaja. MavHome: An Agent-Based Smart Home. Proceedings of the IEEE International Conference on Pervasive Computing and Communications. 521-524.(2003)

[6].G. Michael Youngblood. http://mavhome.uta.edu /argus. (2006)

[7].Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. Time Series Analysis: Forecasting and Control. Third Edition. Prentice Hall. (1994)

[8].Ian H. Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann, San Francisco.(2005)

[9]. Rob J. Hyndman, Anne B. Koehler, Ralph D. Snyder and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods http://www.elsevier.com/locate/ijforecast. (2002).

[10] O. D. Anderson, Box-Jenkins in government, a development in official forecasting. Statist. News 32,14-20 (1977).

[11] S. Wheel Wright and S. Makridakis, Forecasting Models for management, Wiley, New York (1975).