

Predictive Analysis of Home Automation Systems with the Help of Machine Learning

G.Anuradha

VIT University, Vellore.

ABSTRACT

The pervasiveness of home automation has been increasing greatly in recent years due to much higher affordability and simplicity to get updates through smart phone and tablet connectivity [6]. Home automation systems are ambient intelligence systems that are designed to help people proactively, but sensibly. It is automation of the home, housework or household activity or domestic activities under the observance. The approach is based on information theory in order to convert raw data into high-level events used to represent re-cursively structured activities network services. The raw dataset to be used would be given as input for the learning algorithms [4] which is composed of data acquired through the sensors from the home under surveillance. The dataset is to be analyzed and evaluated by employing machine learning algorithms for getting the patterns from domestic activities via sensor nodes. The parameter to be considered is the timestamp of the activities under observance and hence the predicted value of timestamp [21] helps in differentiating unusual activities from the day to day activities. The user gets notified through sms regarding the bizarre activities taking place at home.

Keywords

Big data, Weka, data mining, linear regression.

1. INTRODUCTION

1.1 Background

Home is considered to be the safest place on Earth as it is supposed to provide the ultimate security excluding all the rare of accidents which are out of human control and some of the natural calamities. But, in the cases when people leave home for their some or the other work then all the valuable assets at home need to be given security. In this word, Home automation system [2] is supposed to take care of it by distinguishing unusual happenings at home. The data is captured via sensor nodes and hence data mining has to be done.

The technology community has coined the term Big Data to describe this new era of data and data management. However, there is a lot of confusion around exactly what Big Data [3] is and more importantly, how enterprises should think about Big Data within their organizations. Heterogeneity, scale,

timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata.

Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

The development of sensor networks, particularly in the last years, has extended their applicability in various domains, such as heritage preservation, environmental monitoring and human activity recognition. The size of collected data is rapidly increasing with the number and scale of deployed sensor networks and specialized methods able to deal with such scale and still satisfy application requirements are needed. Data acquisition from these sensors needs to be analyzed properly for the specific intended purposes. The close links between machine learning, statistics and data mining [20] are fairly obvious. All three areas aim at locating interesting regularities, patterns or concepts from empirical data.

Machine learning methods [11] form the core of data mining: decision tree learning or rule induction is one of the main components of several data mining algorithms.

1.2 Problem statement

In Home automation systems, all the activities under specific home are captured through sensors and hence this raw data is collected on the server via base stations present at the home. This pattern recognition of all the domestic events needs to be analyzed depending upon their timestamps in 24 hours of clock from the data acquired by sensor nodes. Different use cases are considered as a part of event types occurring at home. Timestamp prediction of future events has been done using machine learning toolkit "Weka" [21] as in time series environment.

2. DESIGN OF THE SYSTEM

2.1 Introduction

System analysis is the science dealing with analysis of complex, large scale systems and the interactions within those systems. It is the reduction of the entire system by studying the various operations performed and their relationship within the system.

The key part in the initial investigation process is gathering of information about the present available system. After studying the present system and the problems of existing systems are identified and defined. Structured analysis is the set of technique and graphical tools that allow the developer to develop a new kind of system specification that are easy to use, easily understandable to the user.

Feasibility analysis is carried out to select the best system that meets the performance requirement. The traditional smart home systems only can work

with lots of parameter from users, they only provide remote control function and timing control function. There is no intelligent control, they can't decide how and when to control these household appliances by themselves, all they can do is the user have told to them. So, we need a new system that it's user-friendly and not depend on User's settings. The new system in this paper is designed to solve these problems, users don't have to configure the system, and it will learn all the useful parameter from the users automatically. At the same time it has good portability and adaptive capacity.

2.2 System Architecture

A System Architecture is a conceptual model that defines the structure, behavior and more views of the system.

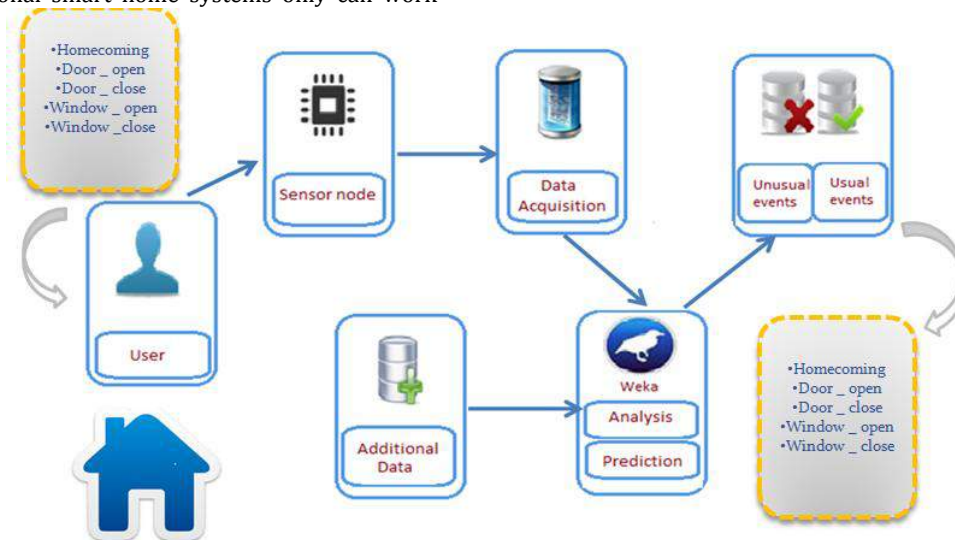


Fig 2.2 System Architecture

2.3 Methodology

2.3.1. First necessity is to have some past activities of the user from day to day life and in some cases, machine learning would not have reference past values for prediction. In that case, prediction would not be possible for some of first few days.

2.3.2. Sensors are used to capture the activities happening at home along with its corresponding timestamps.

2.3.3. This raw data contains all the events wherever sensors are fitted at home. It is the source of final predicted timestamps.

2.3.4. This data has been used as for the training purpose in WEKA toolkit [20].

2.3.5. Firstly, preprocessing of data has been done so that all the irregularities and noise gets removed. Next is forecasting, where linear regression has been used to predict the corresponding timestamps of particular event.

2.3.6. Finally, it is used for differentiating usual and unusual events.

2.4 Data Flow Diagram

Level 0:

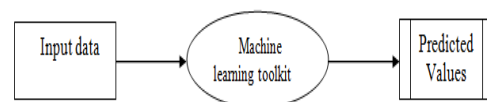


Fig 2.4.1 DFD Level-0

Level 1:

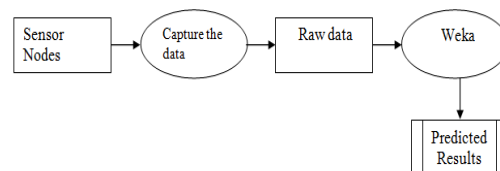


Fig 2.4.2 DFD Level-1

Level 2:

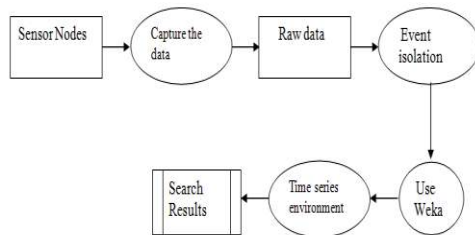


Fig 2.4.3 DFD Level-2

2.5 Activity Diagram

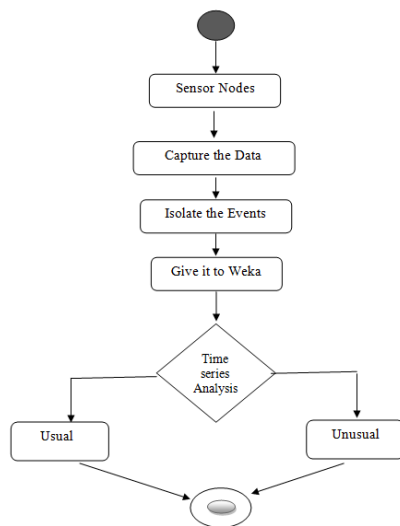


Fig 2.5 Activity Diagram

2.6 Overall Use case Diagram

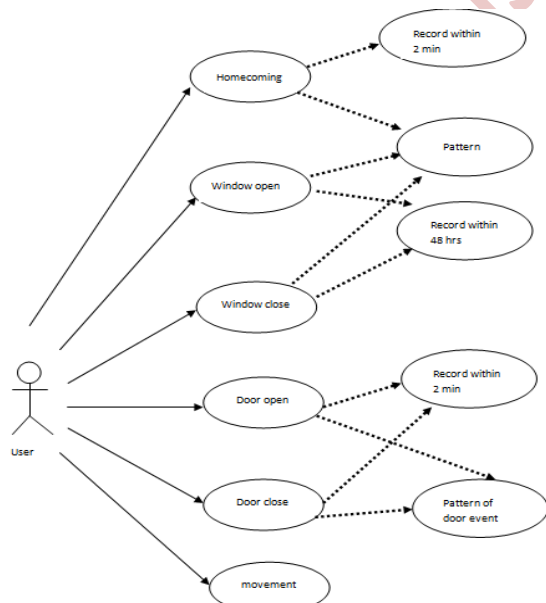


Fig.2.6 Overall use case diagram

3. APPROACHES TO BE USED

3.1 Approach 1

1. Calculate frequency of particular event (ex. homecoming) for 15 minutes interval for a particular Day of week.
2. Find the min, max frequency and calculate the probability of event falling in a particular interval.

3.1.1 Detection

For an incoming event

1. Find out the interval it falls in.
2. Check the probability for the interval and compare with tolerance factor.
3. If probability less than tolerance factor then flag as Unusual.

3.1.2 Advantages

1. Easy to build and maintain.
2. We can start detecting from 8th day, however the quality may not be good due to less base data (for a day of week we will have single observation)

3.2 Approach 2

1. Calculate frequencies of particular event (ex. homecoming) for 15 minutes interval for a particular date.
2. Generate frequencies separately for week days and weekends.
3. Find the min, max frequency and calculate the probability of event falling in a particular interval.

3.2.1 Detection

For an incoming event (lookup appropriate weekday or weekend tables)

1. Find out the interval it falls in.
2. Check the probability for the interval and compare with tolerance factor.
3. If probability less than tolerance factor then flag as Unusual.
4. Repeat this for each summary table and find out usual/unusual based on 7, 30, all days summary.
5. The table below will give the final result.

3.2.2 Advantages

1. We can learn change in pattern quickly.
2. Week Day and Week End differentiation gives better detection based on difference routine.
3. Also takes into account the weekly, monthly and overall pattern and this should give good detection accuracy.
4. Taking weekly and monthly pattern into account should take care of any temporary changes in the pattern.
5. Checking with different age of patterns we can say usual/unusual with certain confidence (for

example of all tables flag and event as unusual then it ~100%.

6. Keeping weekly and monthly table make sure we work out of latest knowledge at the same time we also have older knowledge with us. Keeping single overall pattern table has a downside of pattern becoming stale and taking long time to understand newer pattern.

3.3 Linear Regression

Regression is the easiest technique to use, but is also probably the least powerful. This model can be as easy as one input variable and one output variable (called a Scatter diagram in Excel, or an XY Diagram in OpenOffice.org) [19]. Of course, it can get more complex than that, including dozens of input variables. In effect, regression models all fit the same general pattern. There are a number of independent variables, which, when taken together, produce a result — a dependent variable. The regression model is then used to predict the result of an unknown dependent variable, given the values of the independent variables.

In linear regression, data are modelled using linear predictor functions, and unknown model parameters are estimated from the data. Such models [20,21] are called linear models. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X . Less commonly, linear regression could refer to a model in which the median, or some other quintile of the conditional distribution of y given X is expressed as a linear function of X . Like all forms of regression analysis, *linear regression* focuses on the conditional probability distribution of y given X , rather than on the joint probability distribution of y and X , which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical application. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p -vector of regression x_i is linear. This relationship is modeled through a disturbance term or error variable ϵ_i - an unobserved random

variable that adds noise to the linear relationship between the dependent variable and regression. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n,$$

where T denoted the Transpose. Often these n equations are stacked together and written in vector form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}^T = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or reduction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y , the fitted model can be used to make a prediction of the value of y .
- Given a variable y and a number of variables X_1, \dots, X_p that may be related to y , linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j , to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y .

3.3.1 Lag creation

The Lag creation panel allows the user to control and manipulate how lagged variables are created. Lagged variables are the main mechanism by which the relationship between past and current values of a series can be captured by propositional learning algorithms. They create a "window" or "snapshot" over a time period. Essentially, the number of lagged variables created determines the size of the window. The basic configuration panel uses the Periodicity setting to set reasonable default values for the number of lagged variables (and hence the window size) created. For example, if you had monthly sales data then including lags up to 12 time steps into the past would make sense; for hourly data, you might want lags up to 24 time steps or perhaps 12.

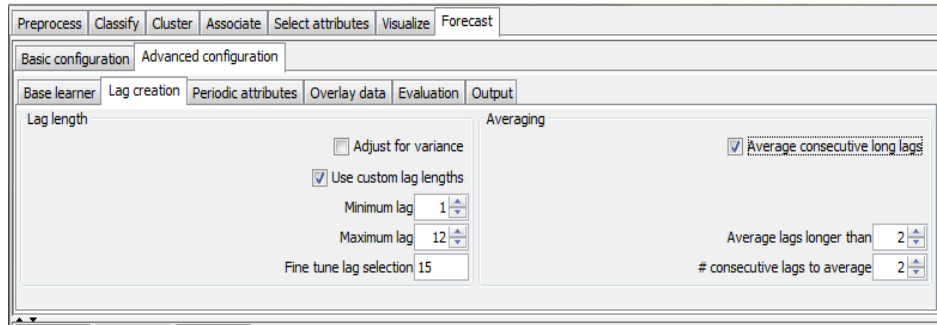


Fig.3.3.1 Lag creation under Advanced configuration

The Maximum lag text field specifies the maximum previous time step to create a lagged variable for - e.g. a value of 12 means that a lagged variable will be created that holds target values at time - 12. All time periods between the minimum and maximum lag will be turned into lagged variables. It is possible to fine tune the creation of variables within the minimum and maximum by entering a range in the Fine tune lag selection text field. In the screenshot below we have weekly data so have opted to set minimum and maximum lags to 1 and 52 respectively. Within this we have opted to only create lags 1-26 and 52.

The Periodic attributes panel allows the user to customize which date-derived periodic attributes are created. This functionality is only available if the data contains a date time stamp. If the time stamp is a date, then certain defaults (as determined by the Periodicity setting from the basic configuration panel) are automatically set. For example, if the data has a monthly time interval then *month of the year* and *quarter* are automatically included as variables in the data. The user can select the customize checkbox in the *date-derived periodic creation* area to disable, *select and* create new custom date-derived variables. When the checkbox is selected the user is presented with a set of pre-defined variables as shown in the following screenshot:

3.3.2 Periodic attributes

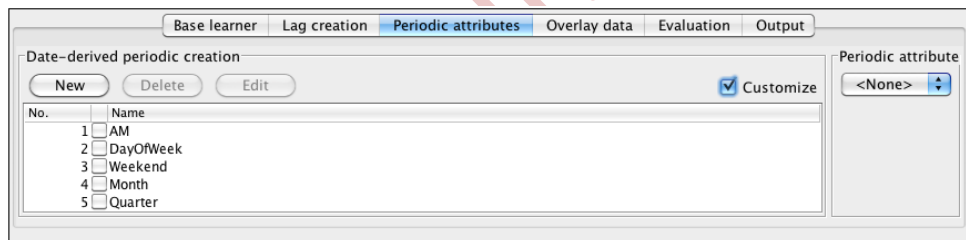


Fig.3.3.2 Periodic attributes selection

3.3.3 Overlay data

The Overlay data panel allows the user to specify fields (if any) that should be considered as "overlay" data. The default is not to use overlay data. By "overlay" data we mean input fields that are to be considered external to the data transformation and closed-loop forecasting processes. That is, data that is not to be forecasted can't be derived automatically

and will be supplied for the future time periods to be forecasted.

In the screenshot below, the Australian wine data has been loaded into the system and *Fortified* has been selected as the target to forecast. By selecting the Use overlay data checkbox, the system shows the remaining fields in the data that have not been selected as either targets or the time stamp. These fields are available for use as overlay data.

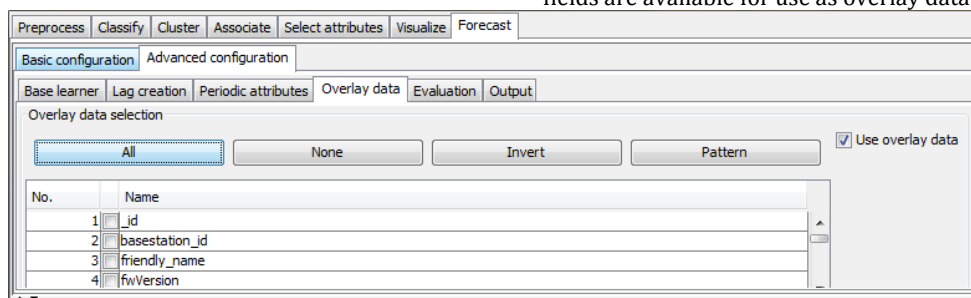


Fig.3.3.3 Overlay Data in Advanced configuration

The system will use selected overlay fields as inputs to the model. In this way it is possible for the model to take into account special historical conditions (e.g. stock market crash) and factor in conditions that will occur at known points in the future (e.g. irregular sales promotions that have occurred historically and are planned for the future). Such variables are often referred to as *intervention variables* in the time series literature.

When executing an analysis that uses overlay data the system may report that it is unable to generate a forecast beyond the end of the data. This is because we don't have values for the overlay fields for the time periods requested, so the model is unable to generate a forecast for the selected target(s). Note that it is possible to evaluate the model on the training data and/or data held-out from the end of the training data because this data does contain values for overlay fields. More information on making forecasts that involve overlay data is given in the documentation on the forecasting plugin step for Pentaho Data Integration.

3.3.4 Evaluation

The Evaluation panel allows the user to select which evaluation metrics they wish to see, and configure whether to evaluate using the training data and/or a set of data held out from the end of the training data. Selecting Perform evaluation in the Basic configuration panel is equivalent to selecting Evaluate on training here. By default, the mean absolute error (MAE) and root mean square error (RMSE) of the predictions are computed. The user can select which metrics to compute in the *Metrics* area in on the left-hand side of the panel. The available metrics are:

1. Mean absolute error (MAE): $\text{sum}(\text{abs}(\text{predicted} - \text{actual})) / N$
2. Mean squared error (MSE): $\text{sum}((\text{predicted} - \text{actual})^2) / N$
3. Root mean squared error (RMSE): $\text{sqrt}(\text{sum}((\text{predicted} - \text{actual})^2) / N)$
4. Mean absolute percentage error (MAPE): $\text{sum}(\text{abs}((\text{predicted} - \text{actual}) / \text{actual})) / N$
5. Direction accuracy (DAC): $\text{count}(\text{sign}(\text{actual}_{\text{current}} - \text{actual}_{\text{previous}}) == \text{sign}(\text{pred}_{\text{current}} - \text{pred}_{\text{previous}})) / N$
6. Relative absolute error (RAE): $\text{sum}(\text{abs}(\text{predicted} - \text{actual})) / \text{sum}(\text{abs}(\text{previous}_{\text{target}} - \text{actual}))$
7. Root relative squared error (RRSE): $\text{sqrt}(\text{sum}((\text{predicted} - \text{actual})^2) / N) / \text{sqrt}(\text{sum}(\text{previous}_{\text{target}} - \text{actual})^2) / N$

The relative measures give an indication of how the well forecaster's predictions are doing compared to just using the last known target value as the prediction. They are expressed as a percentage, and lower values indicate that the forecasted values are better predictions than just using the last known target value. A score of ≥ 100 indicates that the forecaster is doing no better (or even worse) than predicting the last known target value. Note that the last known target value is relative to the step at which the forecast is being made - e.g. a 12-step-ahead prediction is compared relative to using the target value 12 time steps prior as the prediction (since this is the last "known" actual target value).

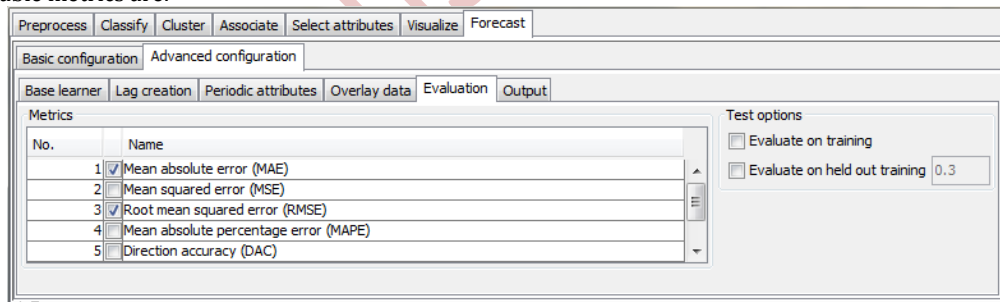


Fig.3.3.4 Evaluation under Advanced configuration

The text field to the right of the Evaluate on held out training check box allows the user to select how much of the training data to hold out from the end of the series in order to form an independent test set. The number entered here can either indicate an absolute number of rows, or can be a fraction of the training data (expressed as a number between 0 and 1).

4. MODULES

4.1 Preprocessing of Data

Data pre-processing is an important step in the data mining process.

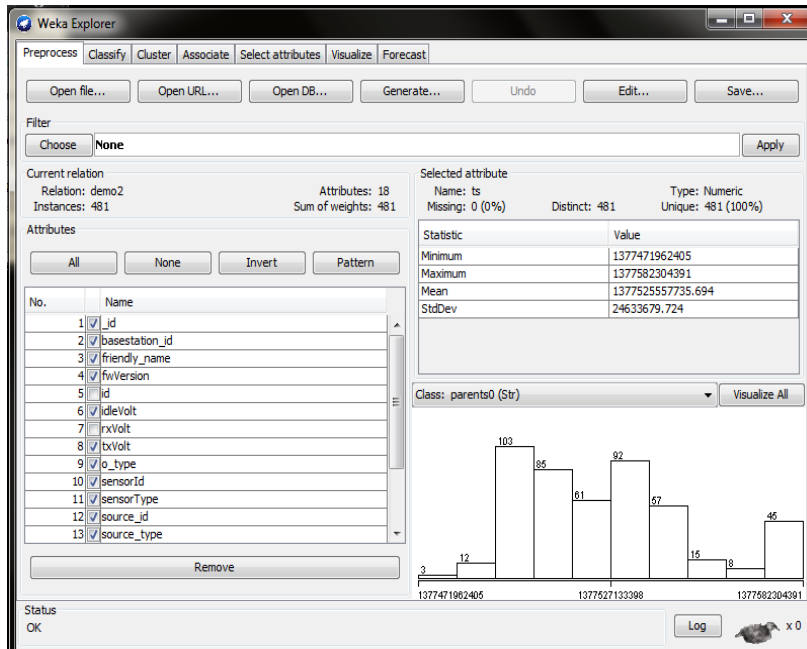


Fig.4.1 Data preprocessing

The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc.

Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

4.2 Select attributes

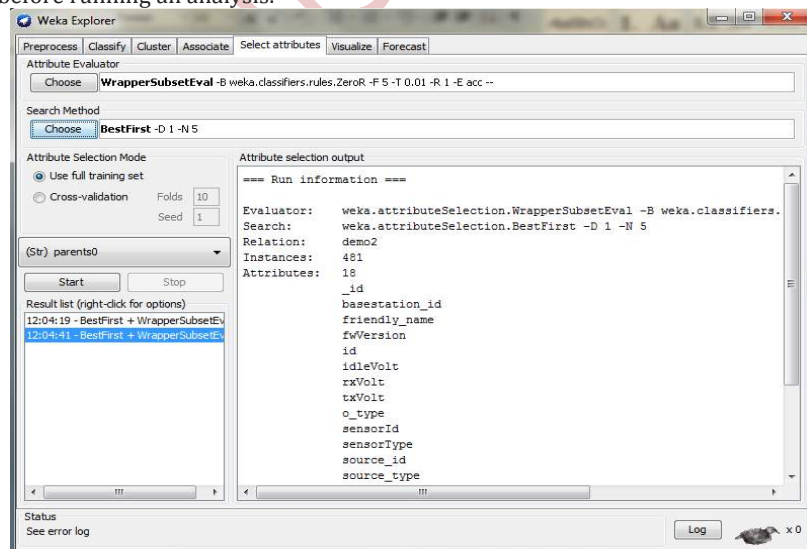


Fig.4.2 Selection of attributes

This panel allows you to configure and apply any combination of weka attribute evaluator and search method to select the most pertinent attributes in the dataset. This attribute helps in evaluation of the attributes.

4.3 Time Series Analysis and forecasting

The Weka Forecasting plugin is a transformation step for PDI4.x that is similar to the Weka Scoring Plugin.

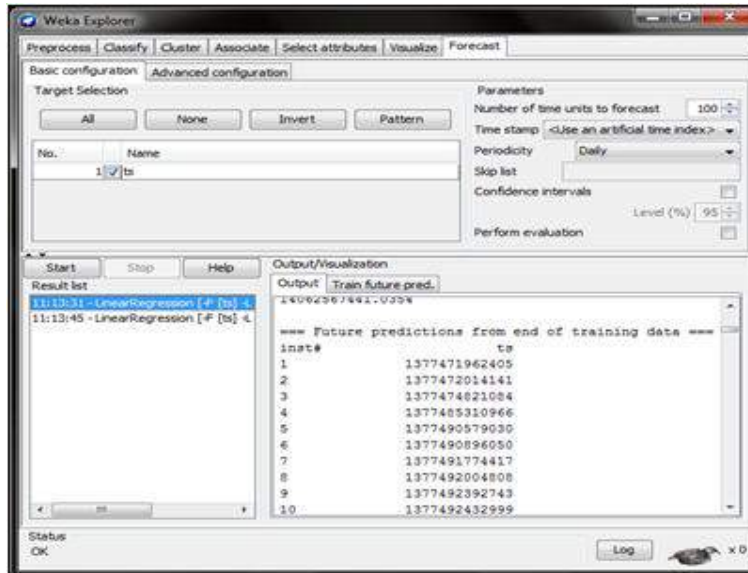


Fig.4.3 Time series forecasting with original timestamp

It can import time series forecasting model created in Weka's time series analysis and forecasting environment and use it to generate a forecast for future time steps beyond the end of incoming historical data.

4.4 Visualization

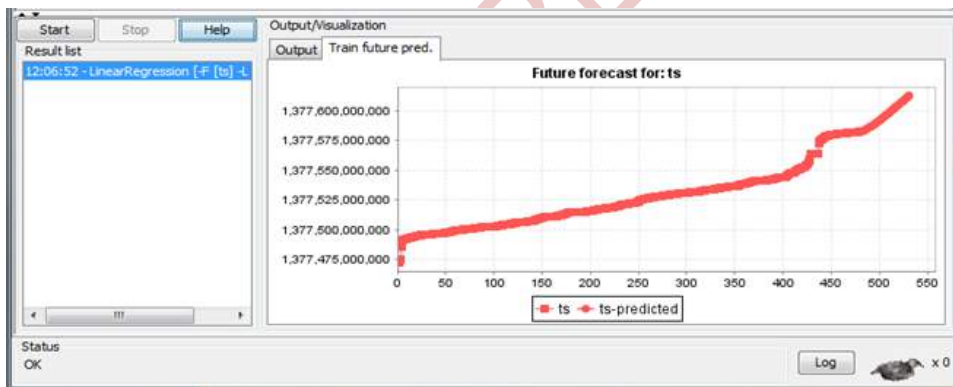


Fig.4.4 Trained future prediction showing timestamp and no of events.

The above diagram shows original and predicted data in plotted graph form. Y-axis denotes time in microseconds and X axis denotes no of events.

4.5 Histogram

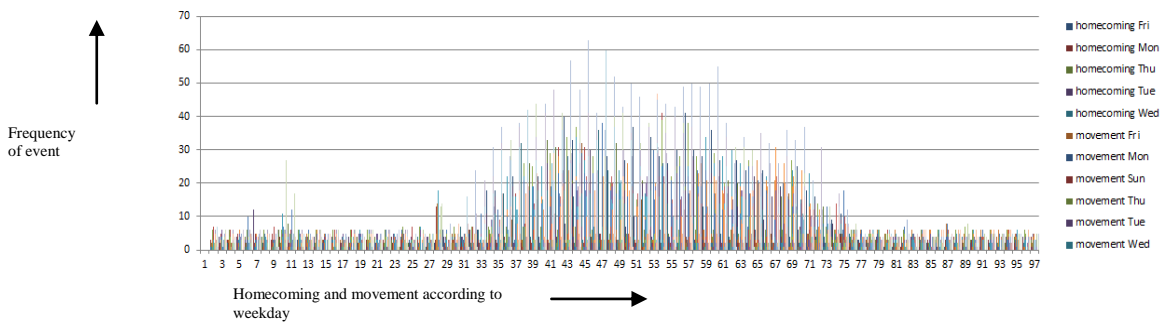


Fig.4.5.1 Histogram showing movement and homecoming events depending upon weekdays

happenings at the home having Home automation systems.

The above histogram shows the overall month's data of events homecoming and movement

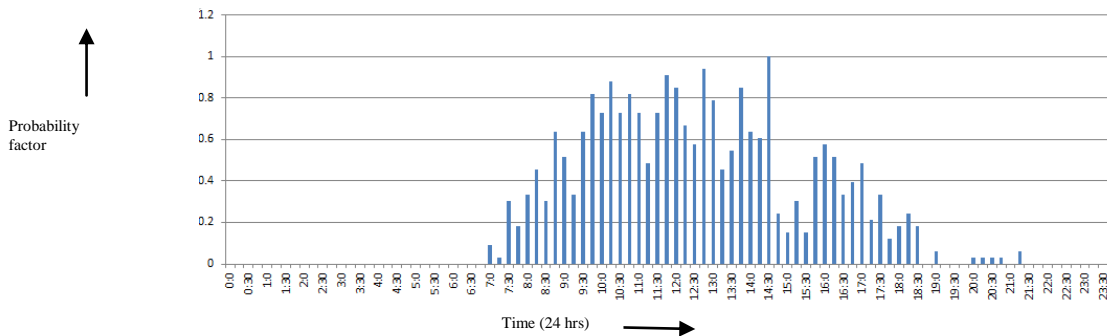


Fig.4.5.2 Homecoming event divided into 24 hrs

This histogram shows the data of event homecoming happening at the single day with respect to 24 hrs of timestamp which has been divided into 15 minutes across the whole 24 hrs.

5. CONCLUSION

The proposed framework is to support the intelligent home network service along with the user-patterns, using the time series analysis in the data mining technique. Machine learning tool Weka has played remarkably important role to forecast the values of timestamp regarding the events captured by sensor nodes successfully. The predicted values have been used to classify as usual and unusual events happening at home. Hence, the activities at home can be easily kept under the surveillance.

6. REFERENCES

- [1] Larsen, K.R. ; Hovorka, D. ; West, J. ; Birt, J. ; Pfaff, J.R. ; Chambers, T.W. ; Sampedro, Z.R. ; Zager, N. ; Vanstone, B., " Theory Identity: A Machine-Learning Approach", System Sciences (HICSS), 2014 47th Hawaii International Conference on ,pp: 4639 - 4648,2014.
- [2] Cottone, P. ; Gaglio, S. ; Re, G.L., "User activity recognition for energy saving in smarthomes", Sustainable Internet and ICT for Sustainability (SustainIT),pp:1-9,2013.
- [3] Katal, A. ; Wazid, M. ; Goudar, R.H., "Big data: Issues, challenges, tools and Good practices", Contemporary Computing (IC3), Sixth International Conference on,pp:404-409,2013.
- [4] Sharma,S.; Agrawal,J.; Agarwal,S, "Machine learning techniques for data mining: A survey", Computational Intelligence and computing research (ICCIC),IEEE International conference on,pp:1-6,2013.
- [5] Balar, A. ; Malviya, N. ; Prasad, S. ; Gangurde, A. "Forecasting consumer behavior with innovative value proposition for organizations using big data analytics", Computational Intelligence and Computing Research (ICCIC), IEEE International Conference on ,pp:1-4,2013.
- [6] Antunes, M. ; Gomes, D. ; Aguiar, R., " Towards behaviour inference in smart environments",Future Internet Communications (CFIC), 2013 Conference on,pp: 1 - 8,2013.
- [7] Yi-Cheng Chen ; Yu-Lun Ko ; Wen-Chih Peng, "An Intelligent System for Mining usage patterns from appliance data in smart home environment", Technologies and Applications of Artificial Intelligence (TAAI) Conference,pp:319-322,2012.
- [8] Jihua Ye ; Qi Xie ; Yaohong Xiahou ; Chunlan Wang, "The Research of an adaptive smart home system",Computer Science & Education (ICCSE), 7th International Conference on ,pp:882-887,2012.
- [9] Lei Jiang ; Suhui Luo ; Jiaming Li, "An Approach of Household Power Appliance Monitoring Based on Machine Learning",Intelligent Computation Technology and Automation(ICICTA), Fifth International Conference on,pp:577-580,2012.
- [10] Begoli, E.Horey, J., "Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture(WICSA) and European Conference on Software Architecture (ECSA),Joint Working IEEE/IFIP Confence on,PP:215-218,2012.
- [11] WuYuntian , "Based on Machine Learning of Data Mining to Further Explore", Computer Science and Information Processing (CSIP), International Conference on, pp: 1235 - 1238 ,2012.
- [12] Wu Yuntian, " Based on Machine Learning of Data Mining to Further Explore", Computer Science and Information Processing (CSIP), 2012 International Conference on ,pp: 1235 - 1238,2012.
- [13] Mirkin, B., " Data analysis, mathematical statistics, machinelearning, data mining: Similarities and differences",Advanced Computer Science and Information System (ICACSIS), 2011 International Conference on ,pp: 1 - 8,2011.

- [14] Das, B. ; Chao Chen ; Dasgupta , N. ; Cook , D. J. ; Seelye, A. M., "Automated Prompting in a Smart Home Environment", Data Mining Workshops (ICDMW), IEEE International Conference on ,pp:1045-1052,2010.
- [15] Das, B. ; Chao Chen ; Dasgupta, N. ; Cook, D.J. ; Seelye, A.M., "Automated Prompting in a Smart Home Environment", Data Mining Workshops (ICDMW), 2010 IEEE International Conference on ,pp:1045-1052,2010.
- [16] Hak Soo Kim ; Jin Hyun Son," User-pattern analysis framework to predict future service in intelligent home network", Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on ,Volume: 2,pp: 818-822,2009.
- [17] El Mahrsi, M.K. ; Vignes, S. ; Hebrail, G. ; Picard, M.-L.," A data stream model for home device description", Research Challenges in Information Science, Third International Conference on ,pp:395-402,2009.
- [18] Abouel Nasr, E.S. ; Al-Mubaid, H.,"Mining process control data using machine learning", Computers & Industrial Engineering, 2009. CIE 2009. International-13th Conference on ,pp: 1434 - 1439,2009
- [19] Bin Wang ; Yunhui Yin ; Kaijun Dong," A Machine Learning Arithmetic for Home Auto control System", Intelligent Control and Automation, WCICA, The Sixth World Congress on ,Volume: 1,pp:4055-4059,2006.
- [20] Mannila,H., " Data mining : machine learning , statistics , and databases" , Scientific and Statistical Database Systems, Proceedings., Eighth International Conference on ,pp: 2 - 9,1996.
- [21] Holmes ,G.;Donkin, A.; Witten,Ian H.,"WEKA:a machine learning workbench",Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on,pp: 357 - 361,1994.

IJSHRE