PREDICTING SMART APARTMENT-INHABITANTS DAILY ENERGY EXPENDITURE

AKEEM ABODUNRIN (San Francisco STATE UNIVERSITY, CA). DIANE COOK (PROF. AT WSU, PULLMAN WA)

Abstract

The need to know smart apartment inhabitants daily Energy Expenditure (EE) cannot be over emphasized. Knowing their EE would assist their caregivers determining their health status and deciding their health care plans. The risk of inhabitant being diabetics can also be evaluated with EE prediction. The objective of this research is to find the procedures of using data mining techniques to predict total daily energy expenditure, and this paper describes the steps taken to arrive at this goal. The steps taken include using ActiGraph to collect inhabitant hourly calorie counts, and questionnaires were used to get their corresponding activity. Hourly environmental temperature was extracted from weather forecast database, while also recording inhabitant room temperature at the appropriate time intervals. Linear Regression (LR) and M5P algorithm on WEKA were used for the EE prediction using preprocessed data from ActiGraph and questionnaires as training data sets. Performance measures and classifier error tree were used to validate the prediction data, and it clearly shows that this method is standard.

Introduction

There are many things that influence designing of EE prediction system, among them are environmental temperature, Body Mass Index (BMI), and difference in which inhabitant interact with the environment(i.e. physical activity). Physical Activity (PA) has been defined as any part of the body movement that resulted by the expansion and contraction of skeletal muscles, which yield energy expenditure (Caspersen et al. 1985). A system that is well tailored to work for inhabitants in the age range of 10-30, may not be able to predict accurate energy expenditure for those in the age range of 50-90. Warren W. Tryon (1998) agrees that physical activity that stimulates good energy expenditure is important for both physical and mental health as well as cognitive functioning. Now that some applications of this system is established, the challenges for the researchers is how to train machine, using data mining techniques to design a system that can be used to

predict inhabitant accurate energy expenditures. In this research, the first step in designing the system being addressed is data collections. Data need to be collected for temperature changes, the duration of activity, date, time, type of activity performed and calories counts. More so, the main areas of controversy in EE prediction include inappropriate linear regression equation and wrong CUT-OFF point. These controversial issues generally lead to efficiency and scalability problem. Using Weka on the datasets from ActiGraph will enable us to come out with a new linear regression equation for EE for its accurate prediction. ActiGraph will be used to collect PAL data from different inhabitants, the data would be analyzed and trained for standard EE prediction. The prediction results from different algorithm and different inhabitant data sets would be validated using Absolute error and correlation coefficient evaluation. Classifier error tree is intended to use to visualize which of the WEKA algorithm predict the most accurate energy expenditure.

Background

In this paper, for the purpose of literature review, some of the devices previously used for collecting activity data are reviewed under six criteria. These criteria include Unit of Measure, Reliability, social and environmental control, availability, validity and constrains. Among the devices that have been tested are Double-Labeled Water method (DLW), Actometer, Telemetric System (Indirect Calorimeter), Direct Calorimeter, Accelerometers (sensor based, subset of ActiGraph), Pedometers, Self Report System, and Heart-rate Monitors. We are going analyze two of these devices in order to clearly point out the need to discover new and standard device of EE data collections.

Double-labeled Water Method (DLW): This controlled method was implemented in both laboratory and field conditions, where inhabitants were allowed to move freely, and their individuals Total Energy Expenditure (TEE) was measured over a period of 10-20 days (Montoye et al. 1996). The method is implemented by administration of two isotopes elements; Hydrogen(H) and Oxygen(O_2) into the inhabitant body. These elements mix with body water and during any activity, hydrogen is eliminated from the body as water while oxygen is eliminated as water and carbon-iv-oxide. In order to know inhabitant energy

expenditure, inhabitant urine sample is taken and mass spectrometer is used to know the amount of CO2 eliminated from the body, and then a standard equation is used to determine TEE (Speakman, 1998). Mass spectrometer is measure in Dalton, and volumetric quantity of isotope elements can be calculated. After substituting the value of mass spectrometer into the equation, the numeric figure obtained depicts EE. It has capacity of producing accurate result while not also interfering with the inhabitant normal activity pattern (Schoeller and Racette, 1990). It is socially acceptable because it does not affect free movement of the inhabitants and does not have any significant medical and environmental side effects. The method is expensive to implement in large scale. Therefore, its affordability is difficulty, which makes its availability depending on researchers' preferences. Its validity for measuring EE under the laboratory condition is good (1.0-3.0%) and with precision of 4.0-7.0% (Starling 2002). Its major constrain is that it cannot differentiate between the types or patterns of activity. In addition, it is difficulty to evaluate the intensity (for example, light or hard) of activity. More so, it is costly to implement.

Actometer: Actometer is a mechanically propelled device, which is usually attached to the wrist as watch or ankle. Its mechanism is sensitive to the movement of the inhabitant weight, which is usually transferred to the hour and second hands of the modified watch through mechanical gear, and the corresponding result is read in the similar way on the normal watch. The calendar is also moved as the hours and seconds hands change (Tryon, 1985). Its unit of measurement is based on the average kinetic energy that is needed to make 4.50gram of the inhabitant weight to spin on its axis in order to effect an increment in the minute hand by 1min. It is difficult for inhabitant to have the same physical movement of either ankle or wrist for different period; therefore, the reliability of actometer is best evaluated by attaching two sensors to the same part of the body for the time interval and compares the results Morrell and Keefe (1988).

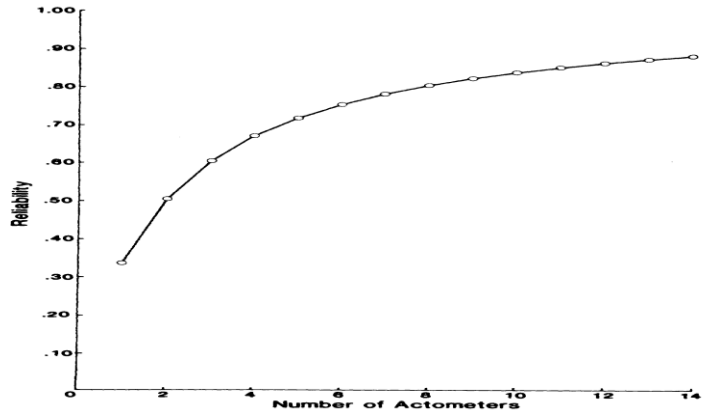


Figure 1.0: Actometer reliability chart

Actometer is less socially acceptable because inhabitant need to wear it as a wrist watch or attach to their ankle thereby influence their social acceptability in the society. Actometer is being around for long time now; it is old but readily available and not as expensive as DLW. Saris and Binkhorst (1977) validate the use of actometer by performing experimentation with use of another EE device, pedometer to determine TEE together with use of actometer and observation method at the same time. Pedometers were attached to the waist of 11 kindergarten pupils, while actometer were attached to their ankles during the school activity. The results of actometer were significantly correlate with results obtained from observational method but showed smaller, still significantly correlate with other variables with (relationship r= 0.69). The conclusion is that the actometer results means that it give reliable overall estimate of EE by the children during physical activity. However, its major constrain was discovered by Charles Johnson (1971), that no matter how the actometer is sensitive and reliable, it would not accurately showed EE in the activity of “hyperkinetic” inhabitants.

Therefore, there is a greater need for a device that is not expensive and provides accurate data sets irrespective of the hypersensitivity of the inhabitant. In addition, it is important to choose best device for data collection and algorithm for data sets preprocessing and EE prediction. ActiGraph, a decive that has capacity to monitor sensitivity in body movement and record sleeping activity is determined to use for this research. There are many algorithms worth reviewing in this paper, but we are going to analyze two of them and compare them with WEKA, our data mining technique for EE prediction.

Classification and Regression Trees (CART): *Salford* Systems Inc. invents this data mining system. It employs regression tress for prediction and decision trees for classification. It uses boosting methodology for accuracy (Han and Kamber 2006). Usually, it does not make codes available for the users.

See5 and C5.0*:* This system also has fuzzy threshold like WEKA, which is used for constructing classifier. RuleQuest invents it. See5 and C5.0 has capacity to predict a class from various attributes values. Without boosting, it shows greater error rate for experimental cases. Andreeva et al. (2003) claim that boosting reduces the error rate in the test data by 25%.

Andreeva et al. (2003) use WEKA to analyze breast cancer data from Wisconsin. In the first datasets, they used 460 randomly selected instances as training datasets. NaiveBayes Weka's algorithm indicates (98.74%) accuracy. With all these features and capabilities, Weka could provide accurate EE for the inhabitant PAL than other DM systems.

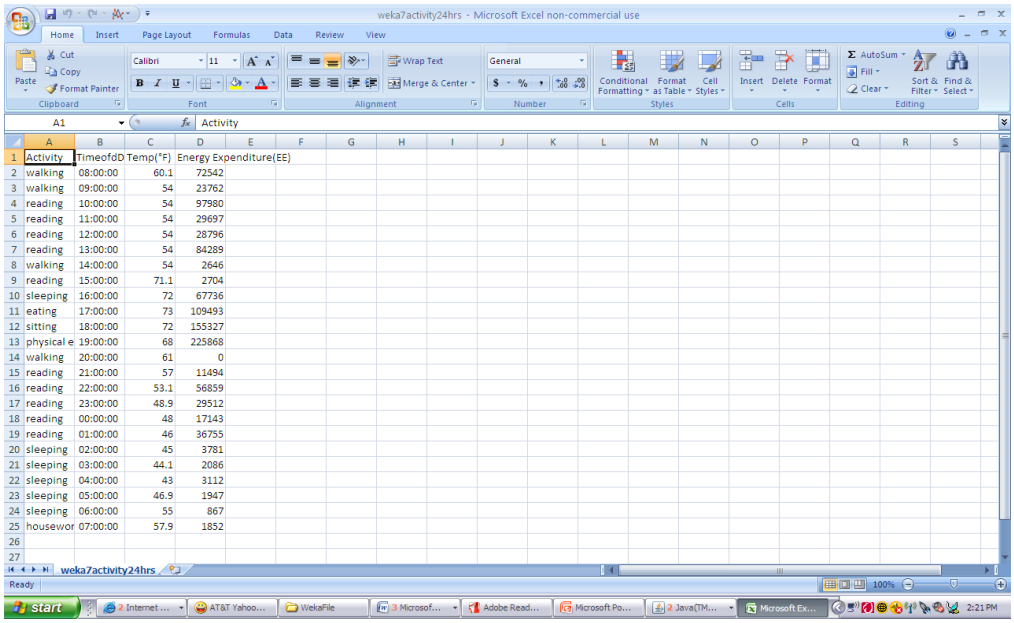
Design and Method

ActiGraph: We choose GT1M ActiGraph model to record inhabitant activity data. This device has advantage over the previous used devices because of its capability to track sleep disorders and can be used to take data for 24 hours or more continuously. It takes accurate data for both physical and sleep activities. Actigraph report the levels of activity and calories burned inform of Commas Separated Values (CSV). It categorizes the patterns of activity in light, moderate and hard mode. It shows the number of minutes in 1-hr that a mode of activity takes place. It works on the mechanism of distance travelled by the body movement and converts to mile using average stride of 18'' per step taken. The amount of EE or calories burned is presented in numeric values, which we are going to upload onto the computer for further analyze using (WEKA) data mining technique.

In this research, two male students within the age of 23-30 wore ActiGraph as wristwatch continuously for the time interval of 24-48 hours continuously. Their body mass index is different so also their race.

Figure 2.0 An ActiGraph wear as a wristwatch

However, it was established that race, sex and personality do not have any significant impact on the result of calorie burn, but BMI is an important factor. Their individual record was uploaded on to the computer using ActiWeb (software from ActiGraph Inc). We preprocess these records by adding corresponding hourly temperature and type of activity using questionnaire to fetch information and Microsoft Excel for data preparation. The processed information is saved in CSV format in order to make it compatible with WEKA data format.



	A	B	C	D
1	Activity	TimeofdD	Temp(°F)	Energy Expenditure(EE)
2	walking	08:00:00	60.1	72542
3	walking	09:00:00	54	23762
4	reading	10:00:00	54	97980
5	reading	11:00:00	54	29697
6	reading	12:00:00	54	28796
7	reading	13:00:00	54	84289
8	walking	14:00:00	54	2646
9	reading	15:00:00	71.1	2704
10	sleeping	16:00:00	72	67736
11	eating	17:00:00	73	109493
12	sitting	18:00:00	72	155327
13	physical e	19:00:00	68	225868
14	walking	20:00:00	61	0
15	reading	21:00:00	57	11494
16	reading	22:00:00	53.1	56859
17	reading	23:00:00	48.9	29512
18	reading	00:00:00	48	17143
19	reading	01:00:00	46	36755
20	sleeping	02:00:00	45	3781
21	sleeping	03:00:00	44.1	2086
22	sleeping	04:00:00	43	3112
23	sleeping	05:00:00	46.9	1947
24	sleeping	06:00:00	55	867
25	housewor	07:00:00	57.9	1852

Figure 3.0 EE Data Sets in CSV Format

Weka: This data mining systems was developed at the University of Waikato, New Zealand. Weka has greater number of machine learning tools for analysis of data and prediction. Among such tools are linear regression, ZeroR, NaïveBayes and classifiers. It has capacity to filter data and get rid of noisy through data preprocessing. It has algorithm for association mining, regression and visualization, which other reviewed Data Mining (DM) techniques lack. It has ability to split datasets and distribute to multiple hosts for analysis.

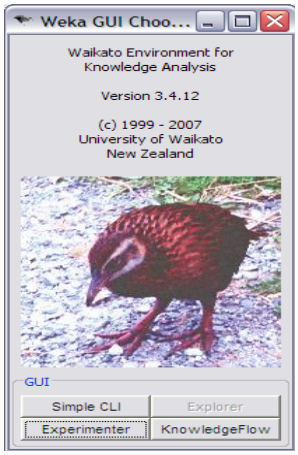


Figure 4.0 WEKA Interface

In this research, we used unsupervised preprocess algorithm on the WEKA to filter out noisy data out of the data sets for EE prediction. We accomplished this by using discretize and ReplaceMissingValues filter algorithm to standardize our attribute values. The preprocessed data is then saved as arff file that WEKA use for its data format; though it supports CSV file too. We use M5P and LinearRegression algorithms in the classification menu to use the uploaded data as training data sets to predict Inhabitant Energy Expenditure. Both algorithms produce Linear Regression Equation, Visualize tree, Numeric EE values, performance errors and visualize tree errors that worth analyze.

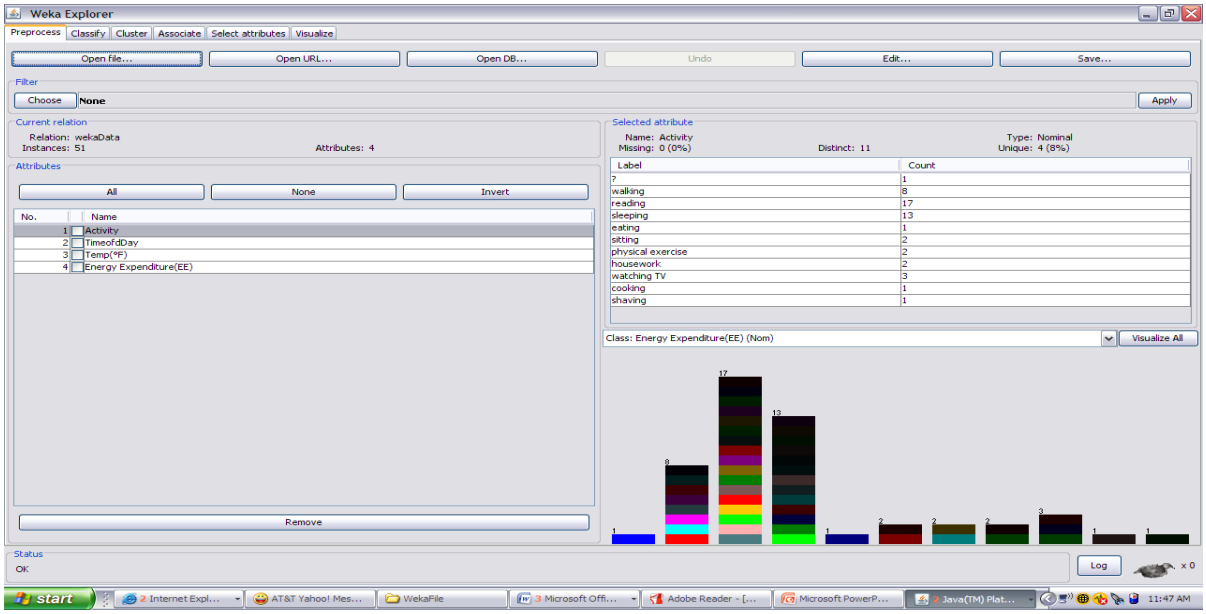


Figure 5.0 WEKA-Working Area Interface

In addition, the original data set is spilt into sub sets using spilt nodes. The goal of splitting nodes is to get sub-samples of data sets that are more original, free of noisy data than the original sample. Its methodology is such that if there are any N attributes, there will be total of N splits to consider.

Choosing best algorithm for numeric prediction, Linear regression work best if the outcome or class is numeric (in our case, Energy Expenditure), and majorities of the attributes are numeric too. The technique of this algorithm is to express class as linear combination of the attributes with specific pre-evaluated weights. It is simply represented with the following equation:

$$C = w_0 + w_1a_1 + w_2a_2 + w_3a_3 + \ldots\ldots + w_ka_k$$ *where C is the class;* $a_1, a_2, a_3, \ldots a_k$ *are the attribute values ; and* $w_0, w_1, w_2, w_3, \ldots\ldots w_k$ *are predetermined weights.* Mainly, the weights are calculated from training data sets.

The predicted value for the first instance of our EE prediction class can be written as:

$$w_0a_0^{(1)} + w_1a_1^{(1)} + w_2a_2^{(1)} + w_3a_3^{(1)} + \ldots\ldots + w_ka_k^{(1)} = \sum_{j=0}^{k} w_ja_j^{(1)}$$

We used M5P because decision tree learning is best suited to problems where instances are represented by attribute-value pairs and are described by a fixed set of attributes (e.g., TimeOfTheDay) and their values (e.g., temperature values).

The only observed problem with WEKA Classification trees is that it can be unstable sometimes and small variations in the data (such as that made by randomization) can cause very different looking trees being generated. The future researchers can look at more suitable algorithms to solve this problem (may be using user defined stepwise algorithm through experimenter)

Results

Using Classify panel, we were able to train our data sets and test learning scheme to perform classification and regression. Originally, for the first subject with higher BMI, there are four attributes, Activity(nominal), TiimeofdDay(numeric), Temp(temperature in ^{o}F, numeric) and Energy Expenditure(the class value expected to predict. There are twenty instances altogether. M5P decision tree algorithm predicts EE with 97% accuracy at 0.9777 correlations co-efficient. The predicted values are the node of the tree in the figure below.

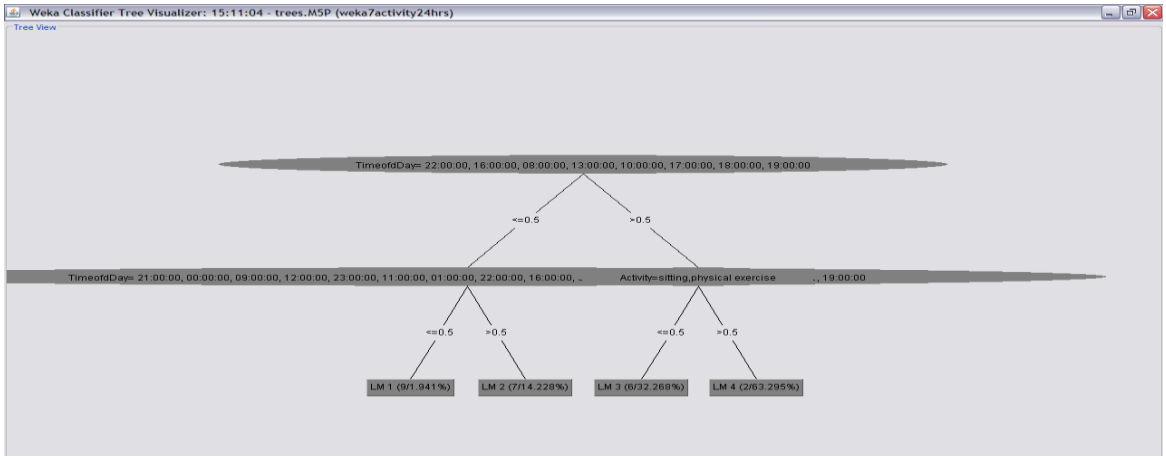


Figure 6: M5P tree

In this research, contrary to general opinion, Linear Regression function produces accurate prediction than the decision tree. The linear regression equation produced is 99% accurate with correlation coefficient of 0.9999. Its basic output is shown in the figure below:

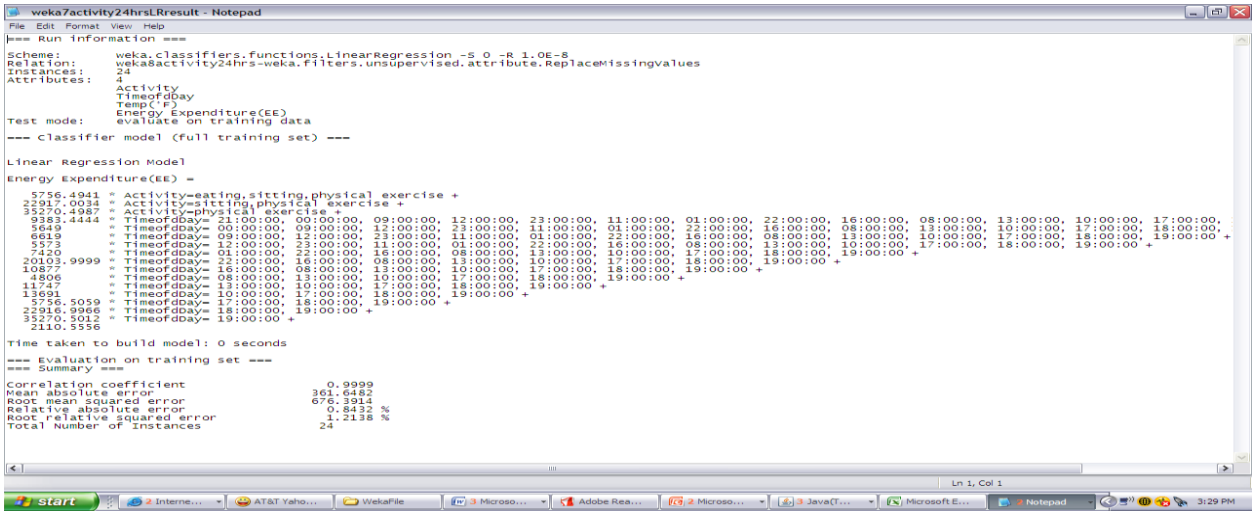


Figure 7.0 LinearRegression Function Output

We use M5 rules for the second subject data set because of few numbers of instances. In this case, there are four attributes and nine instances. M5 rules predict EE for this data set with 0.9745 correlation co efficient and higher root relative square error of 22.5%. The accuracy of this sample is not good as the first subject, though, he has lower BMI. This inaccuracy may be due to few numbers of instances and too much noisy data in the training data.

Evaluation

In this research, we based our evaluation of the predicted results on the percentage accuracy and the performance error of the instance using classification algorithms for the prediction. The table below depicts disparities in the performances of two different algorithms used for the first subject EE prediction.

	M5P (full training Set)	Linear Regression(full training set)
Correlation coefficient	0.9777	0.9999
Mean absolute error	7009.6725	361.6482
Root mean squared error	11711.4451	676.3914
Relative absolute error	16.3434 %	0.8432 %
Root relative squared error	21.0169 %	1.2138 %

Table 1.0: Performance measures for Energy Expenditure (EE) prediction models

It is observed that Linear Regression method gives better prediction value than M5P. Because, its Correlation Coefficient is greater than that of M5P, and it has the smaller values for each error measure. This signifies the success rate at which instances of the training data contributes to the EE prediction. For our own data sets,

Linear Regression is better. In addition, compare with previous numeric predictions for other systems, it is ranked 98%.

Moreover, using Visualize Classifiers errors, the figures below represent each data point with crosses, which indicate absolute error for those instances. The fewer numbers of crosses in figure 8.0a compare with figure 8.0b shows that LinearRegression is better-quality in this research.

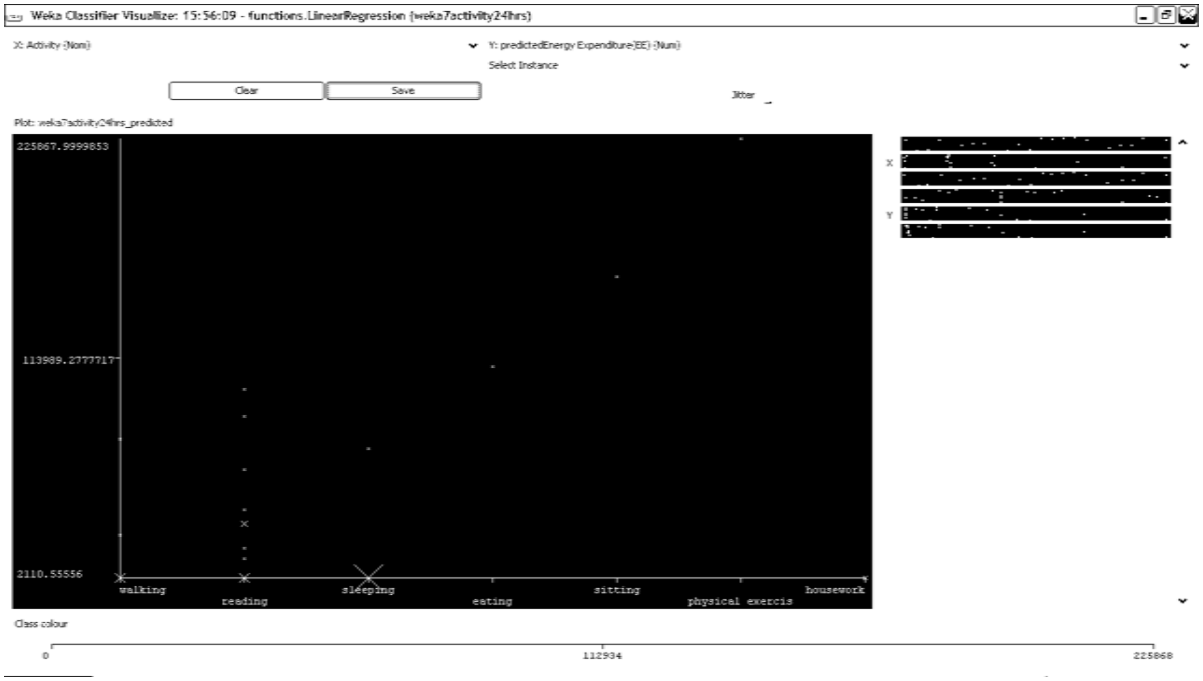


Figure 8.0a LinearRegression Visual Error Graph

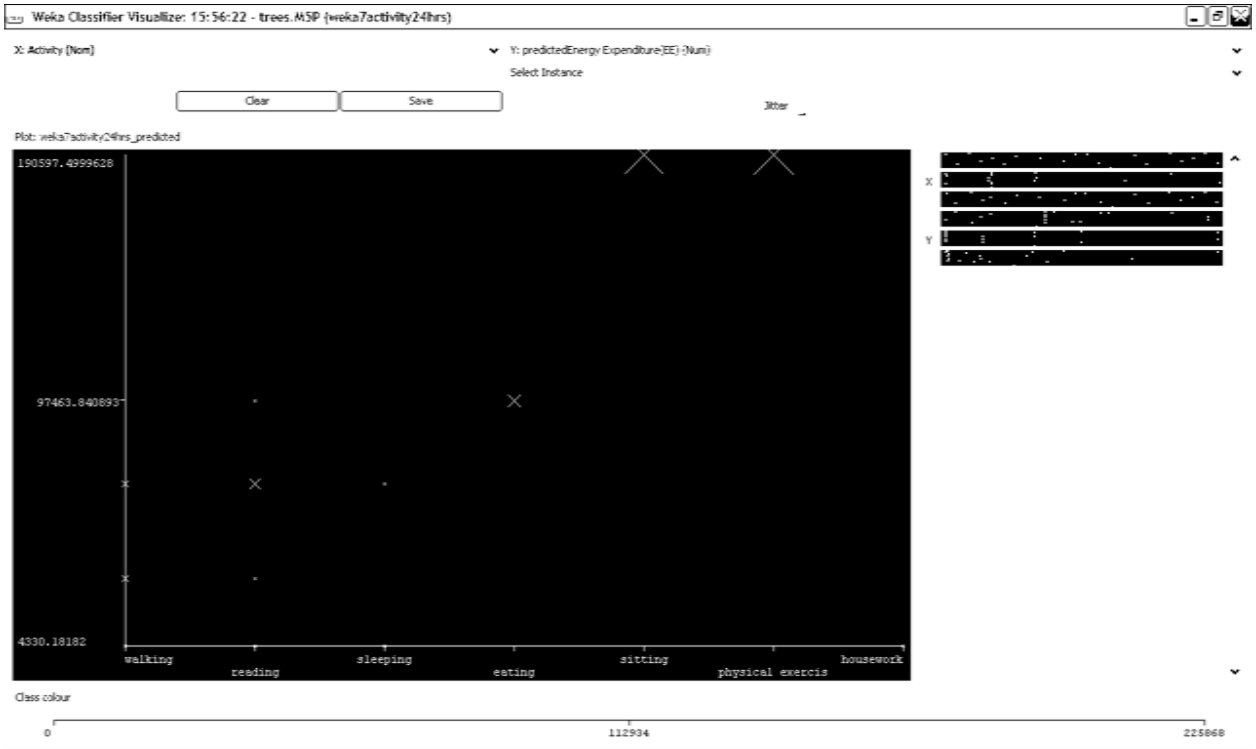


Figure 9.0b M5P (decision Tree) Visualize Error

The figure below shows the outlier of training data sets in this research. Predicted Energy Expenditure values are plotted against time of the day. The graph shows uniformity in the EE between 8.00am and 2.00pm. These values increased drastically between 3.00pm and 10.00pm, and reduce to the minimum level between 10.00pm and 6.00am, which depicts the period that subject is sleeping. It can be conclude that inhabitant energy expedition largely depends on the time of the day.

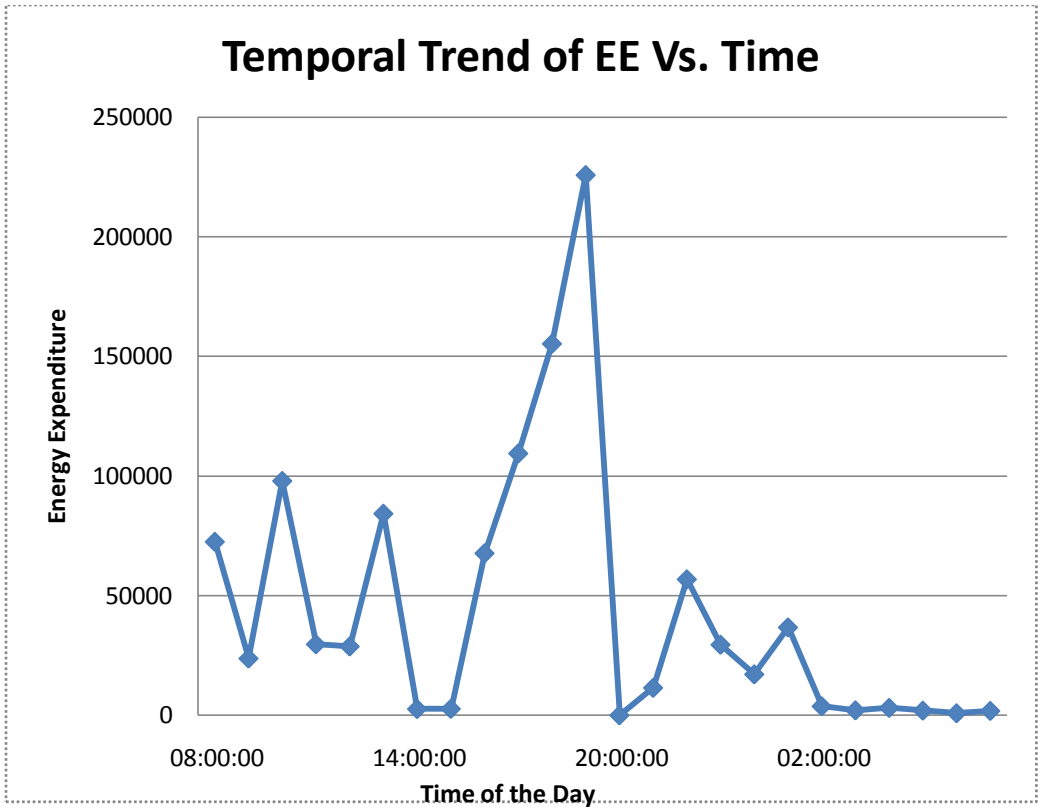


Figure 9.0: Training Data Sets Temporal trend

Conclusion

Accurate Energy Expenditure (numeric) prediction mainly depends on the training data sets. If it consists of noisy data, and it is not filter through preprocessing, the probability of getting accurate prediction result is minimal. In this research, the choice of ActiGraph to collect inhabitant calorie counts assist in minimizing outliers of data. The use of questionnaire enhances normalization of training data sets and using WEKA allows us to filter the noisy out of it. More importantly, ActiGraph allowed us to monitor the sleeping activity and took data for a continuous period. WEKA is a powerful data-mining tool that provides different algorithms for prediction. It allows us to upload our data in CSV format, which other data mining techniques do not support. WEKA enhances error performance validation in order to evaluate our results. In this research, we were able to solve efficiency and scalability problem that always associate with data mining. The major problem noticed in this research was too many missing data in the second subject data sets, though

WEKA allowed us to filter it, but it shows us up in the error percentage. In the long run, we were able to predict inhabitant daily Energy Expenditure (EE) without error rate due to variance in age, and individual physical composition

We recommend that the future researchers in Energy Expenditure prediction focus on how inhabitant daily calories-intake and their BMI affects the physical activity level, and consequently their corresponding EE predicted values.

Acknowledgement

The authors acknowledge the effort of colleagues in the WSU REU programme and everybody in the Artificial Intelligent Laboratory. Without the effort of WSU-EECS departmental staffs and professors, this research work might have been just ordinary mirage, we say thank you.

References

Ainslie PN, Reilly T. Westerterp KR (2003) Estimating Energy Expenditure. A review of techniques with particular reference to Double labeled water. Sport Med 33.

Charles F Johnson (1971): Hyperactivity and the Machine: Actometer, Black Publishing(1971).

Ian H. Witten, Eibe Frank(2005) Data Mining : practical machine learning tools and techniques. Morgan Kaufmann Publishers.

Han J and Kamber M. (2006) Data Mining: Concept ad Techniques. Morgan Kaufmann Publishers (2006)

Valanou E.M; Bamia C; Trichopoulou A(2006) Methodology of Physical-Activity and energy-expenditure assessment: a review. J Public Health 14 : 58-65.

Wang J, Han J (2004) BIDE: efficient mining of frequent closed sequences. In: Proceeding of the 2004 international conference on data engineering (ICDE), Boston, MA, pp 80-90.

http://www.togaware.com/datamining/survivor/Usage5.html

http://books.google.com/books?id=IYc2muhCbmEC&pg=PT216&dq=weka+tutorial&sig=9Q-bCTugldc4ujSTywDCdaJpc4A#PPT200,M1

http://www.theactigraph.com/?gclid=CNOf9eWo5pQCFSBciAodLhWnRQ

http://www.cs.waikato.ac.nz/ml/weka/

http://en.wikipedia.org/wiki/Weka_(machine_learning)