Data Filtering Technique for Neural Networks Forecasting

WIPHADA WETTAYAPRASIT, NASITH LAOSEN AND SALINLA CHEVAKIDAGARN
Artificial Intelligence Research Laboratory, Department of Computer Science,
Prince of Songkla University, Songkla, THAILAND

Abstract: - The weather forecast and medical prediction by neural networks would be more precise and accurate when the filtering technique was handled properly. The paper presents a method of neural networks for rainfall forecast, storm forecast, and medical prediction by using various techniques of data filtering such as moving average filtering technique, local regression filtering technique, Savitzky-Golay filtering technique, and Hamming window filtering technique. The study used weather data sets from Rio de Janeiro and Sao Paulo, Brazil, and Chonburi, Thailand. The medical data sets were from Wisconsin breast cancer database, pima-indians-diabetes, and heart disease ECG pattern from Thailand. The experimental results indicated that the local regression filtering technique gave maximum accuracy for both weather data set and medical data set.

Key-Words: - Data filtering, Neural networks, Local regression filtering

1 Introduction

In a large size of database, normally there will be incomplete data. The incomplete data may be outlier values or error data. Hence, data filtering will be needed for the preprocessing of data. Data filtering is the process of getting rid of noise such as outlier values and error data from raw data to make the data clean and be proper for further processing. The experimental results will get more accuracy if the data is clean. There are many techniques of data filtering such as moving average filtering [1], local regression filtering [2], Savitzky-Golay filtering [3,4,5], and Hamming window filtering [6].

In the process of weather forecast, there will be collecting a large amount of data. For example, data will be collected every day or every three hours. The collected data may have some outlier. And the same reason can be applied for medical prediction. Data preprocessing such as data filtering will be needed to increase the efficiency of weather forecast and medical prediction. The study of an algorithm for the weather forecast time-series data WFNN (Weather Forecast Using Neural Networks) [6] indicated that the filtered data gave higher accuracy than the non-filtered data. In the process of medical prediction using neural network [7], the breast cancer database, the pima-indians-diabetes, and the heart disease of the ECG Rhythm classification using neural network [8] can be used to predict the output data.

2 Data Filtering Technique and Neural Networks

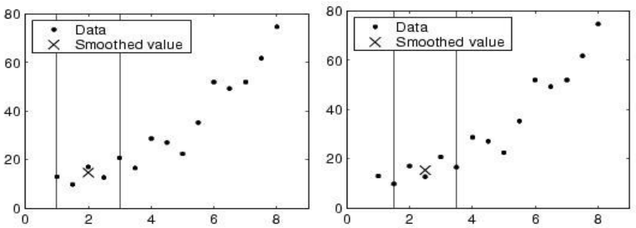
2.1 Data filtering technique

This paper will discuss on four types of sliding window data filtering techniques, which are moving average filtering, local regression filtering, Savitzky-Golay filtering, and Hamming window filtering. The details of each techniques are as follows.

2.1.1 Moving average filtering is a simple data filtering technique by calculating the average value of every point in the selected window. The calculation of new filtered data [1] is shown in equation (1)

$$(y_k)_s = \frac{\sum_{i=-n}^{i=n} y_{k+i}}{2n+1} \qquad (1)$$

where $(y_k)_s$ is the filtered data (smoothing) at point k, y_k is the data at point k (before smoothing), n is a number of points in each side (left side or right side), and $-n \leq i \leq n$. The window size equals to $2n+1$. Figure 1 shows the first and second data filtering points where n equals to 2 and window size equals to 5.



(a) 1st data filtering (b) 2nd data filtering

Fig 1. Data filtering by moving average filtering where n = 2 and the window size = 5.

2.1.2 Local regression filtering is a data filtering technique that uses the method of regression analysis. This filtering technique will specify weight for every data point in the selected window by using regression weight function as shows in equation (2)

$$w_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3 \qquad (2)$$

where w_i is regression weight of points i, x is the predictor value associated with the response value to be smooth, x_i are the nearest neighbors of x as defined by the selected window, and *d(x)* is the distance along the abscissa from x to the most distant predictor value within the selected window.

Local regression filtering can be divided into four types [2]. The a) lowess local regression uses the method of linear regression analysis. The b) loess local regression uses the method of polynomial square regression analysis. The robust function can be used to get rid of outlier values. The technique can be applied with lowess local regression which is called c) rlowess local regression and also can be applied with loess local regression which is called d) rloess local regression, respectively.

2.1.3 Savitzky-Golay filtering is one of data filtering techniques that has a characteristic of frequency-wave. This technique uses the simple polynomial least–square calculation, which will not filter too many details out of the data [3,4,5]. The control parameters are the size of the window and the polynomial degree. Note that the size of window has higher value than the polynomial degree. The formula for calculation on the filtered data shows in equation (3)

$$(y_k)_s = \frac{\sum_{i=-n}^{n} A_i y_{k+i}}{\sum_{i=-n}^{n} A_i} \qquad (3)$$

where $(y_k)_s$ is the filtered data (smoothing) at point k, y_k is the data at the point k (before smoothing), n is a number of positions of points in each side (left side or right side), $-n \le i \le n$, and A_i is the value of coefficient weight at position i. The window size equals to $2n+1$. This filtering technique has been applied in many fields such as decreasing noise in ultra sound image, radar image [3,5], and etc.

2.1.4 Hamming window filtering is the data filtering technique that uses signal processing technique with the method for frequency filtering of finite impulse response [6] as shows in equation (4)

$$\hat{h}(k) = h(k) \cdot w(k) \qquad (4)$$

where $\hat{h}(k)$ is a new filtered data, *h(k)* is an old data, and *w(k)* is the window function. The window function is used to improve the range of response for transition as shows in equation (5)

$$w(k) = \begin{cases} 0.54 - 0.5\cos\dfrac{2\pi k}{K-1} & ; 0 \le k \le K-1 \\ 0 & ; other \end{cases} \qquad (5)$$

where *w(k)* is the window function, k is a value of data at point k, and K is a number of data points of the window function.

2.2 Neural networks

Neural networks is one of data mining techniques to produce knowledge from database. Neural networks can be used for classification with input patterns that are linearly separable. The processing unit imitates human brain which composes of many connected neurons. The data set for supervised neural networks will be divided into two sets that are trained set and tested set. The structure of multilayer perceptron neural networks composes of three layers that are input layer, hidden layer, and output layer [9]. The neuron processing unit will calculate the summation function of input attributes times weight values of each link and pass through the activation functions. The example of activation functions are sigmoid function and tan function.

3 Data Filtering for Neural Networks Forecasting Model (DFNNF)

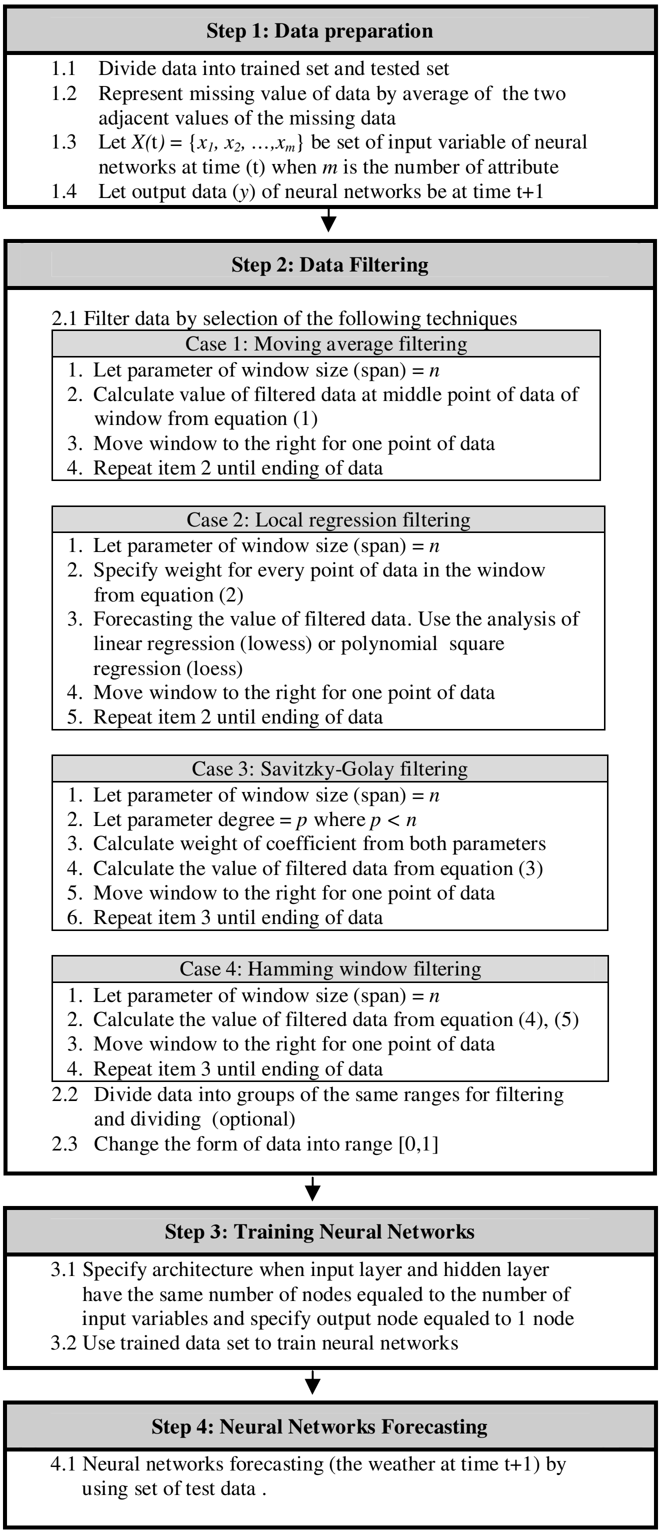
The data filtering for neural networks forecasting model has 4 steps as shows in Figure 2.

Step 1. Data preparation: Data have been divided into trained set and tested set and missing data will be replaced by average of the two adjacent values of the missing data.

Step 2. Data Filtering: Select one of the four filtering techniques to filter the raw data. For the filtering and dividing technique (optional), both data of trained set and tested set will be calculated to find range of each input attribute. We can calculate the distance (range) of each group by *r = (max-min) / q* where *r* is the range in each group, *q* is the number of group, *max* is the maximum value of data, and *min* is the minimum value of data. Next step is changing the data into range [0,1] before training the neural network. The calculation can be performed by *new data = (original data – min range) / (max range – min range).*

Step 3. Training neural networks: The multilayer perceptron neural networks structure used backpropagation algorithm and sigmoid activation function. User will specify the number of input nodes, the number of hidden nodes, and the number of output node. The number of input nodes is equal to the number of input attributes and the number of output nodes is one. In this step we will use the trained data set.

Step 4. Neural networks forecasting: This step is weather forecast for time-series data (at time t+1) or medical prediction by using data set from tested set.



Step 1: Data preparation

1.1 Divide data into trained set and tested set
1.2 Represent missing value of data by average of the two adjacent values of the missing data
1.3 Let $X(t) = \{x_1, x_2, \ldots, x_m\}$ be set of input variable of neural networks at time (t) when m is the number of attribute
1.4 Let output data (y) of neural networks be at time t+1

Step 2: Data Filtering

2.1 Filter data by selection of the following techniques

Case 1: Moving average filtering
1. Let parameter of window size (span) = n
2. Calculate value of filtered data at middle point of data of window from equation (1)
3. Move window to the right for one point of data
4. Repeat item 2 until ending of data

Case 2: Local regression filtering
1. Let parameter of window size (span) = n
2. Specify weight for every point of data in the window from equation (2)
3. Forecasting the value of filtered data. Use the analysis of linear regression (lowess) or polynomial square regression (loess)
4. Move window to the right for one point of data
5. Repeat item 2 until ending of data

Case 3: Savitzky-Golay filtering
1. Let parameter of window size (span) = n
2. Let parameter degree = p where $p < n$
3. Calculate weight of coefficient from both parameters
4. Calculate the value of filtered data from equation (3)
5. Move window to the right for one point of data
6. Repeat item 3 until ending of data

Case 4: Hamming window filtering
1. Let parameter of window size (span) = n
2. Calculate the value of filtered data from equation (4), (5)
3. Move window to the right for one point of data
4. Repeat item 3 until ending of data

2.2 Divide data into groups of the same ranges for filtering and dividing (optional)
2.3 Change the form of data into range [0,1]

Step 3: Training Neural Networks

3.1 Specify architecture when input layer and hidden layer have the same number of nodes equaled to the number of input variables and specify output node equaled to 1 node
3.2 Use trained data set to train neural networks

Step 4: Neural Networks Forecasting

4.1 Neural networks forecasting (the weather at time t+1) by using set of test data .

Fig 2. Data filtering for neural networks forecasting (DFNNF) model.

4 Experimental Results

There were 6-data sets for the experiment, which the rainfall data set was from Chonburi province in Thailand [10], the storm data sets were from Rio de Janeiro and Sao Paulo in Brazil [11], the medical data set, pima-indians-diabetes from UCI data set [12], heart disease ECG (Electro Cardio Graphy) pattern in Thailand, and Wisconsin breast cancer data set [12].

4.1 The rainfall data set from Chonburi province in Thailand:

The data were received from the Meteorological Department of Thailand [10] and composed of 17,520 records (6 years). The data set was collected every three hours at 01:00 AM, 04:00 AM, 07:00 AM, 10:00 AM, 1:00 PM, 04:00 PM, 07:00 PM, and 10:00 PM on each day. There were 7-input variables that were x_1: quantity of cloud (0 to 10 parts of the sky), x_2: temperature of dew point (Celsius), x_3: weather pressure (hextopascal), x_4: relative humidity (percent), x_5: temperature (Celsius), x_6: wind speed (knot), x_7: direction of wind, and 1-output variable that was y: quantity of rain (millimeter).

Step 1: Divide data into trained set (5 years) of 14,600 records and tested set (1 year) of 2,920 records.

Step 2: Filter the data of both trained set and tested set by using the four-filtering techniques. Figure 3 shows the example of filtering for the data of dew points by various techniques. The window size was specified equal to 5 where the x-axis was the day and y-axis was the temperature of dew point. The experiment result indicated that the filtered data received was different for each technique. Figure 4 shows comparison of time using in data filtering of each technique. The technique used longer time was local regression filtering with rloess and rlowess. The local regression filtering with loess and lowess used moderate time. The filtering techniques that used less time were Hamming window filtering, moving average filtering, and Savitzky-Golay filtering.

Step 3: The neural networks architecture was 7:7:1 with 7-input nodes (input attributes), 7-hidden nodes, and 1-output node. Then filtered data of neural networks were trained by trained set.

Step 4: The neural networks were tested by tested set. Table 1 shows accuracy (at time t+1) from the data filtering and the data filtering and dividing comparing with different size of window. The experiment found that loess local regression data filtering technique gave highest accuracy at 95.6% with the window size was equal to 3. For the experiment of data filtering and dividing, loess and rloess local regression gave highest accuracy at 94.5% with the window size was equal to 3. Figure 5 shows comparison of accuracy of various techniques between data filtering and data filtering and dividing with window size is equal to 3. Note that the data filtering gave higher accuracy than the data filtering and dividing techniques.

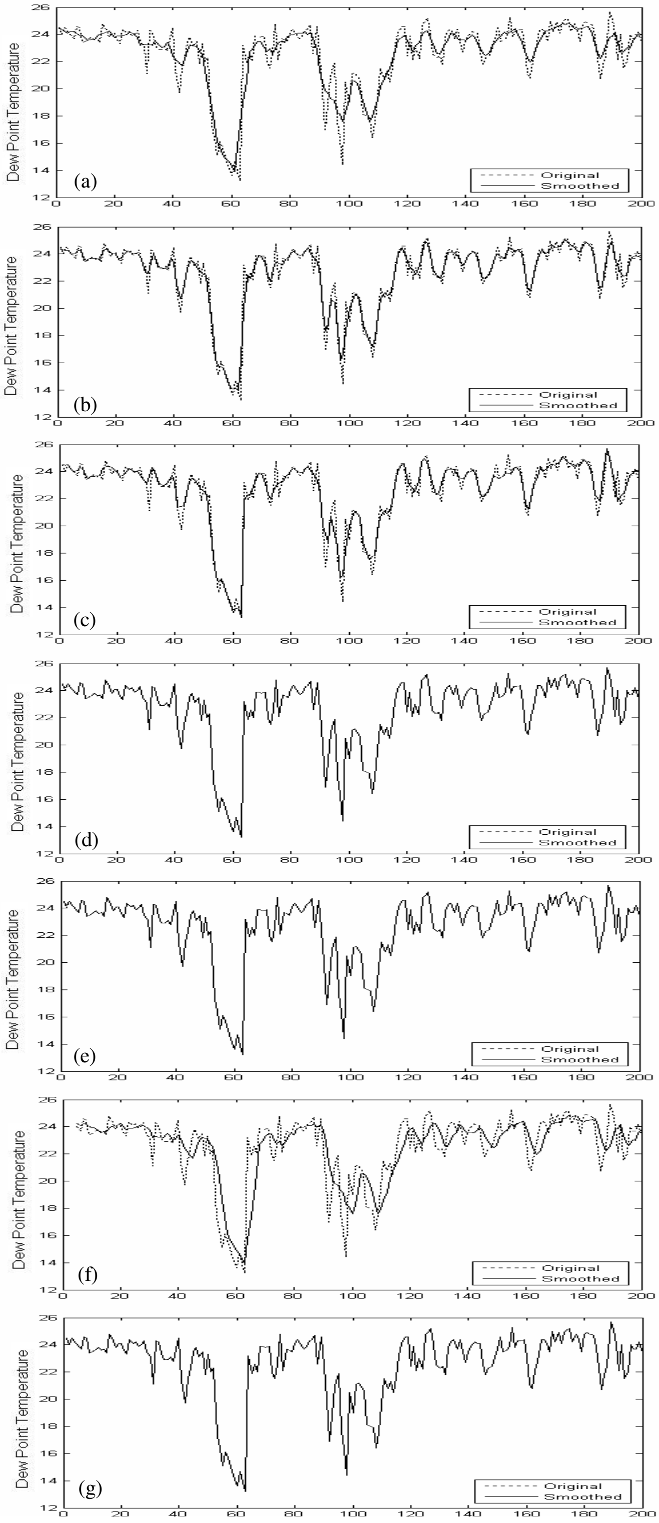


Fig 3. Varoius data filtering techniques where window size = 5 for Chonburi dew point temperature: (a) Moving average (b) lowess (c) rlowess (d) loess (e) rloess (f) Hamming window and (g) Savitzky-Golay polynomial degree = 4

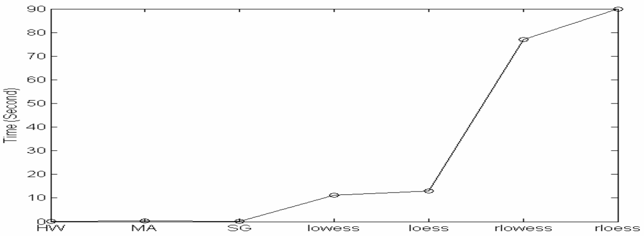


Fig 4. Shows data filtering comparison of time among each technique for Chonburi rainfall forecast.

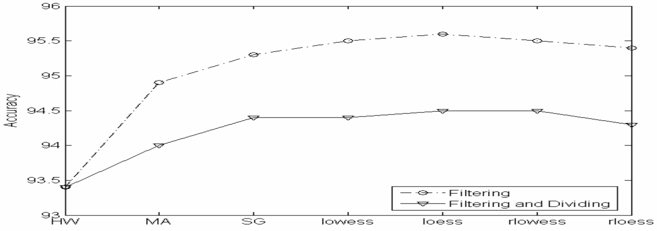


Fig 5. Comparison of accuracy window size = 3 for Chonburi rainfall forecast.

Table 1. The data filtering and the data filtering and dividing accuracy for Chonburi rainfall forecast.

Technique	Size of window					
	Filtering			Filtering and Dividing		
	3	5	7	3	5	7
Moving average	94.9	94.7	94.0	94.0	94.1	93.4
lowess	95.5	95.0	95.0	94.4	94.3	93.9
rlowess	95.5	94.5	94.6	**94.5**	94.0	93.9
loess	**95.6**	95.5	95.1	**94.5**	94.1	94.3
rloess	95.4	95.5	94.7	94.3	93.5	93.8
Savitzky-Golay	95.3	95.5	95.0	94.4	94.2	94.3
Hamming window	93.4	93.3	93.1	93.4	93.4	93.4

4.2 The storm data set from Rio de Janeiro:

There ware 10-input variables [11] that were x_1: highest temperature (Fahrenheit), x_2: lowest temperature (Fahrenheit) x_3: highest temperature of dew point (Fahrenheit), x_4: lowest temperature of dew point (Fahrenheit), x_5: highest relative humidity (percent), x_6: lowest relative humidity (percent), x_7: highest weather pressure (inch), x_8: lowest weather pressure (inch), x_9: wind speed (mile/hour), x_{10}: quantity of cloud (0 to 10 parts of sky), and 1-output variable y: storm which is represented by 1 (storm) or by 0 (no storm).

Step 1: The data set was collected every day. The data set composed of 1,794 records. Divide data into trained set of 1,457 records and tested set of 337 records.

Step 2: Filtering the data by using the four-filtering techniques. The time used for data filtering of each technique indicated that the technique used longer time was local regression filtering with rloess and rlowess. The loess and lowess used moderate time. The filtering technique that used less time was Hamming window filtering, moving average filtering, and Savitzky-Golay filtering.

Step 3: The neural networks architecture was 10:10:1.

Step 4: The experiment for data filtering technique in Figure 6, which was rloess local regression gave highest accuracy at 92.6% with the window size was equal to 5. For the experiment of

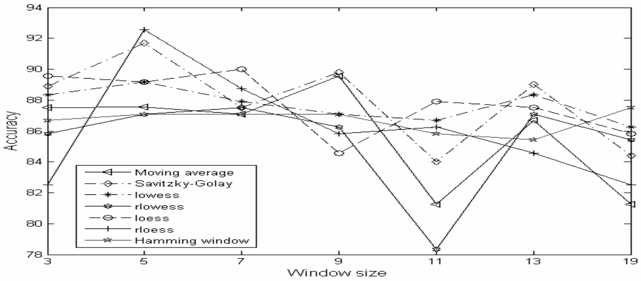


Fig 6. The data filtering accuracy of Rio de Janeiro for storm forecast.

data filtering and dividing, rloess local regression gave highest accuracy at 91.7% with the window size was equal to 5.

4.3 The storm data set from Sao Paulo:

There were 10-input variables [11] that were x_1: highest temperature (Fahrenheit), x_2: lowest temperature (Fahrenheit), x_3: highest temperature of dew point (Fahrenheit), x_4: lowest temperature of dew point (Fahrenheit), x_5: highest relative humidity (percent), x_6: lowest relative humidity, x_7: highest weather pressure (inch), x_8: lowest weather pressure (inch), x_9: wind speed (miles per hour), x_{10}: quantity of clouds (0 to 10 parts of the sky) and 1-output variable that was y: data of storm that would be represented by 1 (storm) or by 0 (no storm).

Step 1: The data set was collected every day. The data set composed of 1,420 records. Divide data into trained set of 1,083 records and tested set of 337 records.

Step 2: Filter the data by using the four-filtering techniques. The experimental result indicated that rloess local regression filtering technique used more time on data filtering. The technique used longer time was local regression filtering with rloess, rlowess. The technique used moderate time was loess and lowess. The techniques that used less time were Savitzky-Golay, moving average, and Hamming window.

Step 3: The neural networks architecture was 10:10:1.

Step 4: Figure 7 shows the experimental result of data filtering. The experiment for the data filtering technique, which was rloess local regression gave highest accuracy at 88.7% with the window size was equal to 5. For the experiment of data filtering and dividing, which was loess local regression gave highest accuracy at 86.4% with the window size was equal to 5.

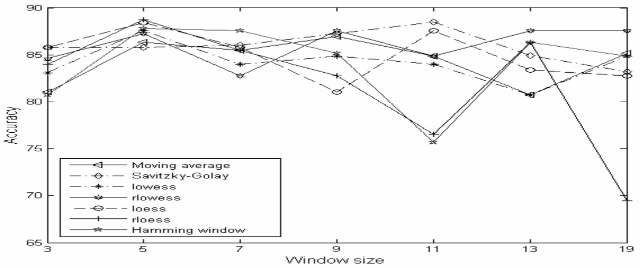


Fig 7. The data filtering accuracy of Sao Paulo for storm forecast.

4.4 The pima-indians-diabetes data set:

There were 8-input variables [12] that were x_1: number of pregnant times, x_2: plasma glucose concentration with 2 hours in an oral glucose tolerance test, x_3: diastolic blood pressure (mm Hg), x_4: triceps of skin fold thickness (mm), x_5: 2-hour serum insulin (mu U/ml), x_6: body mass index (weight in kg/(height in m^2)), x_7: diabetes pedigree function, x_8: age (years), and 1-output variable that was y: data of diabetes (yes or no). The data set composed of 768 records. Divide data into trained set of 600 records and tested set of 168 records. Figure 8 shows the experimental result of data filtering. The experiment for the data filtering technique, which was rloess local regression gave highest accuracy at 79.8% with the window size was equal to 3. For the experiment of data filtering and dividing, which was rloess local regression gave highest accuracy at 78.6% with the window size was equal to 5.

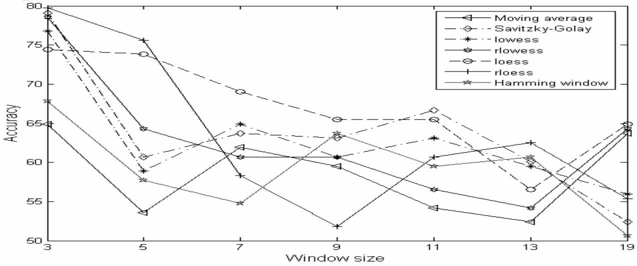


Fig 8. The data filtering accuracy for pima-indians-diabetes data set.

4.5 The heart disease ECG pattern:

The data were collected from 4 hospitals in Thailand which were Prince of Songkla hospital, Chulalongkorn hospital, Rajavithi hospital, and Phyathai2 hospital. There were four types of Sino-Atrial node (SA node) of malfunction (heart disease). There were 5-input variables that were x_1: P wave, x_2: P-R interval, x_3: QRS complex, x_4: rhythm, x_5: rate and 1-output variable that was y: type of heart disease (yes or no). The data set composed of 870 records. Divide data into trained set of 600 records and tested set of 170 records. Figure 9 shows the experimental result of data filtering type 4. The experiment for the data filtering technique, which were rloess, lowess, and loess local regression gave highest accuracy at 93.5% with the window size was equal to 3. For the experiment of data filtering and dividing loess and lowess local regression gave highest accuracy at 90.0% with the window size was equal to 3.

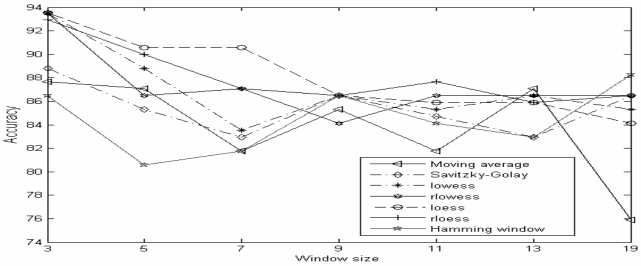


Fig 9. The data filtering accuracy for heart disease type 4 data set.

4.6 Wisconsin breast cancer data set:

There were 9-input variables [12] that were x_1: clump thickness, x_2: uniformity of cell size, x_3: uniformity of cell shape, x_4: marginal adhesion, x_5: single epithelial cell size, x_6: bare nuclei,

x_7: bland chromatin, x_8: normal nucleoli, x_9: mitoses and 1-output variable that was y: breast cancer (yes or no). The data set composed of 699 records. Divide data into trained set of 500 records and tested set of 199 records. Figure 10 shows the experimental result of WBCD for data filtering technique. The experiment for the data filtering technique rloess local regression gave highest accuracy at 99.0% with the window size was equal to 3. For the experiment of data filtering and dividing rloess local regression gave highest accuracy at 98.5% with the window size was equal to 5.

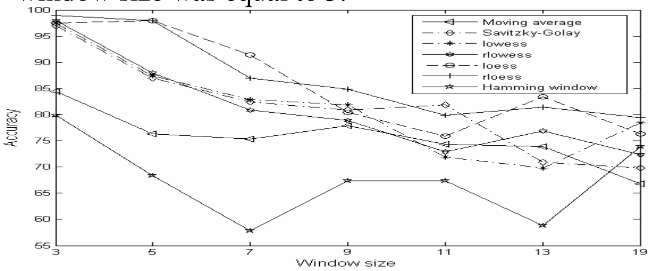


Fig 10. The data filtering accuracy for WBCD.

Table 2 shows the comparison from six-data sets among the filtering data, the non-filtering data, and the filtering and dividing data at window size equals to 3 of rloess local regression. The experimental result shows that the filtering data gave higher accuracy than the non-filtering data, and the filtering and dividing techniques.

Table 2. The comparison of accuracy from six-data sets (window size = 3) of rloess.

Data set	Filtering	Non-Filtering	Filtering and dividing
Pima Indian Diabetes	**79.8**	76.2	78.6
Sao Paulo	**88.7**	86.9	85.2
Rio de Janeiro	**92.6**	89.6	91.7
Heart disease type1	**96.5**	95.3	81.2
Heart disease type2	**98.8**	98.8	97.7
Heart disease type3	**95.9**	95.9	81.8
Heart disease type4	**92.9**	90.0	86.5
Chuburi Province	**95.4**	95.4	94.3
WBCD	**99.0**	99.0	95.5

5 Discussion and Conclusion

Each smoothing method has its characteristic as follows. Hamming window is the frequency filtering of finite impulse response in time domain. Moving average filtering takes the average of neighboring data points. Savitzky-Golay filtering is a generalized moving average where the method derives the filter coefficients by performing an unweighted liner least square using a polynomial of the specified degree. When lowess filtering (a first degree polynomial) and loess filtering (a second degree polynomial) are local regression techniques, then use locally weighted linear square fitting.

This paper presents various data filtering techniques for weather data and medical data. The experimental results concluded as followings. 1) The size of window for various filtering techniques would effect on the accuracy of data. The higher value of window size was, the less accuracy of the forecasting would be., 2) Data filtering of each technique used different time. Techniques that used least time were moving average filtering, Savitzky-Golay filtering, and Hamming window filtering. Techniques that used moderate time were lowess and loess local regression filterings. Techniques that used most time was rlowess and rloess local regressions., 3) The data filtering technique gave higher accuracy than the filtering and dividing technique., and 4) Local regression filtering technique gave highest accuracy of forecasting. Hence, the selecting of data filtering techniques needs to consider on the size of the window, time used, and accuracy of forecasting.

6 Acknowledgement

This study has been supported by research fund in the year 2007 from the Faculty of Science at Prince of Songkla University, Thailand.

References:

[1] C. Vaiphasa, “Consideration of Smoothing Techniques for Hyperspectral Remote Sensing”, ISPRS, Journal of Photogrammetry & Remote Sensing, 2006, pp.91-99.

[2] Mathwork Matlab[online].available:http://www.mathwork .com/ access/helpdesk/help/toolbox/curvefit

[3] C. Chinrungrueng and A. Suvichakorn , “Fast Edge-Preserving Noise Reduction for Ultrasound Images”, IEEE Transactions on Nuclear Science, 2001, pp. 849-854.

[4] T. Tarumi, W.S. Gary, J. C. Roger, and T. K. Robert, “Infinite Impulse Response Filters for Direct Analysis of Interferogram Data from Airborne Passive Fourier Transform Infrared Spectrometry”, Vibrational Spectroscopy, vol. 37, issue1, 2005, pp. 39-52.

[5] C. Chinrungrueng, “Combining Savitzky-Golay Filters and Median Filters for Reducing Speckle Noise in SAR Images”, IEEE International Conference on Man and Cybernetics, vol. 1, 2003, pp. 690-696.

[6] W. Wettayaprasit and P. Nanakorn, “Feature Extraction and Interval Filtering Technique for Time-series Forecasting Using Neural Networks”, in Proc. 2006 IEEE International Conferences on Cybernetics and Intelligent Systems (CIS), 2006, pp. 635-640.

[7] W. Wettayaprasit and C. Lursinsap, “Neural Rule Extraction Based on Activation Projection with Certainty Factor”, IEEE International Conference on Neural Networks, 2002, pp. 1730-1735.

[8] G. E. Oien, N. A. Bertelsen, T. Eftestol and J. H. Husoy, “ECG Rhythm Classification Using Artificial Neural Networks”, IEEE Digital Signal Processing Workshop Proceedings, 1996, pp. 514–517.

[9] J. M. Zurada, “Introduction to Artificial Neural System”: West publishing, 1992.

[10] Meteorological Department of Thailand. [online]. available: http:// www.tmd.go.th

[11] M. Jeffrey, Internet Weather Service “Weather underground” [online]. available: http://www.wundergrou nd.com

[12] J. Mertz and P. M. Murphy. (2005, December) University of California at Irvine (UCI) repository of machine learning databases. [Online]. available: ftp://ftp.ics.uci. edu/pub/machine-learning-databases