Paper

SOM an approach to Data Mining of Power Transformer Database

Non-member	Obu-Cann Kwaw	(Department of Electrical and Electronic Engineering, Tottori University)
Member	Fujimura Kikuo	(Department of Electrical and Electronic Engineering, Tottori University)
Non-member	Tokutaka Heizo	(Department of Electrical and Electronic Engineering, Tottori University)
Member	Ohkita Masaaki	(Department of Electrical and Electronic Engineering, Tottori University)
Member	Inui Masahiro	(Department of Electrical and Electronic Engineering, Tottori University)

Data mining is part of a larger area of recent research in Artificial Intelligence and Information Processing and Management otherwise known as Knowledge Discovery in Database (KDD). The main aim here is to identify new information or knowledge from database in which the dimensionality or amount of data is so large that it is beyond human comprehension. Self-Organising Map is used to analyse power transformer database from one of the electric energy providers in Japan. Furthermore, the regression aspect of SOM is also tested. Regression is achieved by searching for the Best Matching Unit (BMU) using the known vector components.

Keywords: Self-Organising Maps (SOM), Data Mining, Knowledge Discovery in Database and Best Matching Unit (BMU)

1. Introduction

The Self-Organising Map (SOM) [1] [2], developed by Professor T. Kohonen, is one of the most widely used Artificial Neural Network Algorithms. SOM is usually presented as a two-dimensional grid or map whose units (nodes) become tuned to different input data patterns. This algorithm is based on unsupervised competitive learning, training in this case is entirely data driven and the neurons or nodes on the map compete with each other. Data is useless to mankind if no meaningful information can be derived from it. The SOM is a powerful tool for data mining, knowledge discovery and visualisation of high dimensional data. In this work, SOM is applied to power transformer database from one of the electric energy providers in Japan. The aim of this work is to apply a data-mining tool based on SOM to learn more about database.

2. The SOM Algorithm

2.1 A brief introduction Based on the functions of a neuron cell of a living thing, especially the information processing ability of the human brain, Kohonen [3] developed the following equation.

$$m_i(t+1) = m_i(t) + \alpha(t)[x(t) - m_i(t)] \quad \cdots\cdots (1)$$

where $m_i(t)$ is the reference vector of the neuron (node) i at time t, and $x(t)$ is the input vector drawn from the input data set at time t. Prior to training, the reference or weight vectors of each node are initialised. Initialisation can be done using one of the three methods listed below [3].

- Random Initialisation
- Using Initial Samples
- Linear Initialisation

At time t, the cell learns this input signal, as shown in figure 1. During the time $(t+1)$, the information processing ability of the cell becomes $m_i(t+1)$. If $x(t)$ is an n-dimensional input vector, then $x(t) = [x_1, x_2, ..., x_n]$. The n-dimensional reference vector $m_i(t)$ is also expressed as $m_i(t) = [m_{i1}, m_{i2}, ..., m_{in}]$. $\alpha(t)$ is the learning coefficient factor with values between 0 and 1. Furthermore, $\alpha(t)$ reduces to 0 as learning progresses. When an n-dimensional input vector is introduced to the network, the reference vector in the network (node) that is closest to the input vector is defined as the best-matching node "winner" and its information processing ability is denoted by $m_c(t)$. The winner is selected using the following equation.

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \quad \cdots\cdots (2)$$

where m_i is the reference vector of node i and m_c is the reference vector of the winner node or unit (BMU, Best Matching unit). Prior to learning, a large reference area surrounding the winner is selected as a neighbourhood region. The reference vectors in this neighbourhood region N_c as well as the winner m_c learn the input vector x following eq.(1). The neighbourhood region reduces gradually until only the winner is trained. This forms a typical learning cycle. The next cycle begins with the introduction of the next input vector.

2.2 Determination of the parameters in the Self Organising Map Program Package (SOM-PAK) SOM-PAK is constructed following the algorithm in section 2.1. The SOM learning process usually comprises of 2 steps: rough learning L1 and detailed learning L2. The total number of learning cycles can be defined in advance. The learning coefficient in eq.(1) is

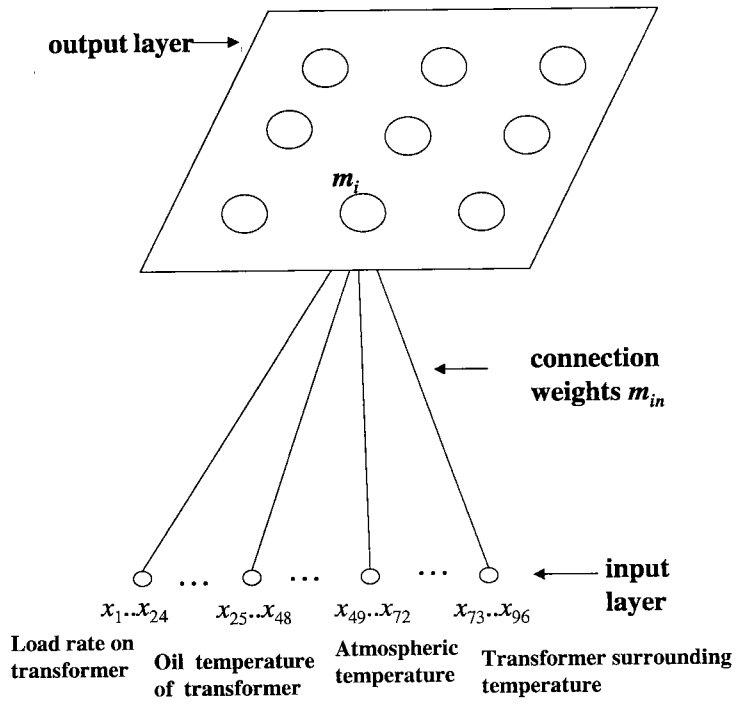


Fig. 1. Structure of the SOM adapted for data mining analysis of the power transformer database

selected as a linearly decreasing function of t by:

$$\alpha(t) = \alpha_0(1 - t/T) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (3)$$

Where α_0 is the initial value, t is the present learning cycle and T is the number of learning cycles. Other types of decay function for α can be considered. There are two types of neighbourhood functions, the bubble and the gaussian. The bubble neighbourhood function, which is the simpler of the two, is constant over the whole neighbourhood of the winner unit and zero elsewhere as shown in figure 2. However, nodes selected from a gaussian probability irrespective of the radius determine the gaussian neighbourhood region. This paper utilises the bubble neighbourhood function and the number of nodes contained in the neighbourhood decays linearly as in eq.(3).

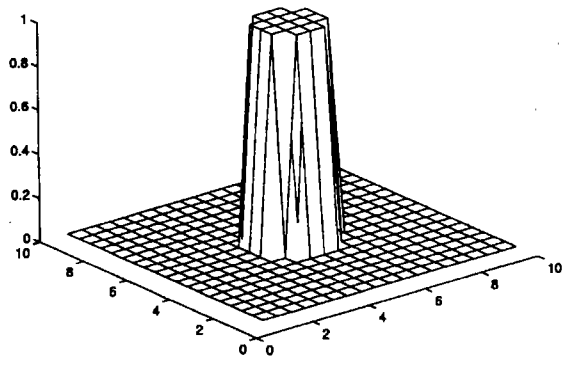


Fig. 2. The bubble neighbourhood function

3. Data Mining of Power transformer Database

The primary aim is to analyse the data to find out what information can be derived from it. In this study, the data set contains four measurements (percentage load on the transformer, transformer oil temperature, atmospheric temperature and transformer-surrounding temperature). The data was analysed to identify the day type pattern classification, season type pattern classification as well as the energy consumption pattern of the consumers. Furthermore, the regression aspect of SOM was tested. An input vector representing a particular day was deleted from the input data space. SOM was applied after which regression was achieved by searching the Best Matching Unit (BMU) using the known vector components. As an output, an approximation of the unknown components of the input vector was obtained.

3.1 **Data pre-processing** The data set for the experiment contained measurements for the percentage load on the transformer, the transformer oil temperature, atmospheric temperature and transformer-surrounding temperature. These values were recorded on the hourly bases. It is a known fact that the learning result can be affected considerably by pre-processing the data appropriately. Pre-processing the data ensures that the differences among the data are maximised. The numerical accuracy of statistical computations in connection with the SOM algorithm improves considerably when the input data is pre-processed [3]. One way in which the data can be pre-processed, is known as component scaling which is a linear transformation of each vector component.

$$x_i(new) = \frac{x_i(old) - x_l}{x_h - x_l} \quad \cdots\cdots\cdots\cdots\cdots\cdots (4)$$

$i = 1, 2, \ldots, n$ where n is the number of components in the input vector x, $x_i(new)$ is the normalised value of component i in the input vector x, $x_i(old)$ is the original value of component i in the input vector x, x_l is the value of the minimum component in the input vector x and x_h is the value of the maximum component in the input vector x. The input data is normalised so that the minimum and maximum values in the input vector x, are set to zero and one respectively. The component values of the training data can also be scaled by ensuring that the mean is zero and the variance is one. Component scaling is applied to ensure that no component will have excessive influence or control on the training results by virtue of its higher absolute value. This procedure is reversible, after training the results can be reversed to report the original unscaled data. Other options of pre-processing the component values of the input vector such as histogram equalisation and filtering can also be applied. Figure 3 depicts the hourly temperature of the transformer oil of a typical day in the data set.

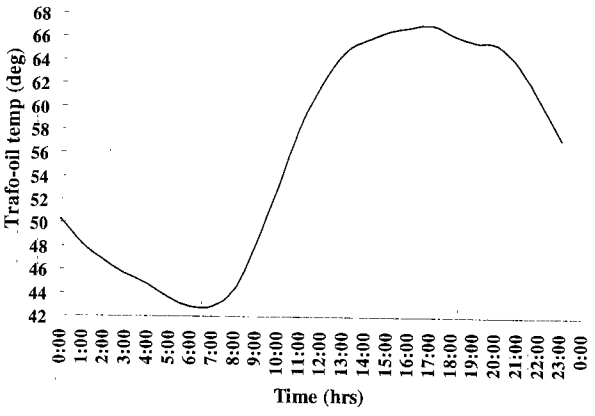


Fig. 3. Hourly transformer oil temperature curve

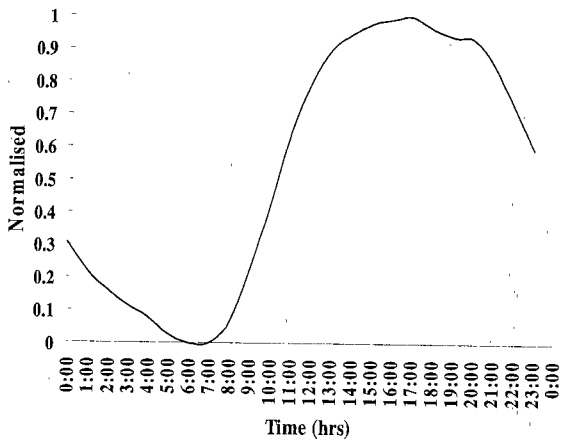


Fig. 4. Normalised hourly transformer oil temperature curve

Figure 4 depicts the normalised hourly temperature of the transformer oil of figure 3. The other input patterns, atmospheric temperature and transformer-surrounding temperature were also normalised using the same formula.

3.2 **Day Type Pattern Classification** In this experiment, SOM is utilised for the classification of hourly transformer oil temperature patterns. Consumption pattern classification is the first step to prediction of power demand, otherwise known as load forecast [4]. Consumption pattern can be said to be a function of the social habits of the consumers. In load forecasting, it is imperative that the social habits of the consumer be studied and considered in order to come up with a good approximation of the load. SOM is used as a pattern recognition tool to identify the different load patterns in the data. An input pattern comprises of the 24 hourly transformer oil temperature readings for a day. Since the learning process is to identify new day types and merge together similar day types, the learning rate factor was initialised to a value of 0.008 and the learning cycles to 50. Figure 5 shows the feature map for March 1996 after 50 iterations of training. Table 1 is a summary of the results obtained in figure 5.

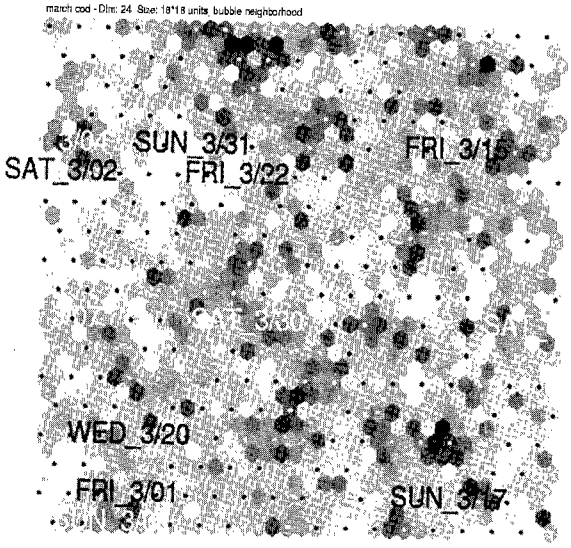


Fig. 5. SOM for transformer oil temperature in March 1996

This procedure was repeated for the other months in the year and the results compared. 20th March 1996 (Wednesday) was the only holiday in March. From the results of figure 5 and Table 1, 20th March (Wednesday) was located at node 256 together with 24th March (Sunday), which was a weekend. The load pattern for 16th March (Saturday) was missclassified as a weekday as a result of incorrect readings during data collection. Furthermore, the weekend patterns were also classified seperately as Saturday patterns and Sunday patterns. Comparing these with results obtained when the other months were analysed, it was observed that new output nodes, each corresponding to a group of input patterns with a particular feature emerged after the training process. It was also observed that not all the days appeared on the map, this is because days with similar transformer oil temperature patterns are mapped to the same output node as the same day type as shown in Ta-

ble. 1†. The day types identified from the feature maps are summarised in Table 2. By the use of the SOM, the load patterns can be classified into day types patterns.

Table 1. Summary of the SOM for March 1996

Output node	Date(day)
73	4(M), 5(T), 6(W), 12(T), 13(W), 14(TH), 16(S), 19(T), 21(TH), 25(M)
73	26(T), 27(W), 29(F)
78	31(SU)
87	15(F), 18(M)
91	2(S)
97	22(F)
181	7(TH), 8(F), 11(M)
188	30(S)
198	9(S), 23(S)
256	20(W), 24(SU)
286	1(F), 28(TH)
303	17(SU), 10(SU)
309	3(SU)

Table 2. Summary of the day type classification for March 1996

Day type	Output Node
Weekdays	73, 87, 181, 286
Saturdays	91, 188, 189
Sundays and Holiday	78, 303, 309, 256

3.3 Season Type Pattern Classification In this section, SOM is applied in the classification of the input patterns into the various seasons of the year namely winter, spring, summer and autumn. The data set for the SOM comprises of 4 input measurements (percentage load on the transformer, the transformer oil temperature, atmospheric temperature and transformer-surrounding temperature) for the months of January, March, July and September 1996. Each of the four months (January, March, July and September) is represented by a day. The assumption here is that the input patterns for each season will have a similar pattern and therefore can be represented by a day's pattern. These are hourly measurements hence each input vector is made up of 96 components, which have been normalised as in section 3.1. For this experiment since the input patterns are to be separated into the various seasons, detailed training is required. It is therefore necessary that the learning cycles as well as the learning rate factor be higher than that used for the day type pattern classification. The learning rate factor was initialised to a higher value of 0.03 and the learning cycles to 5000. Figure 6 illustrates the seasonal feature map obtained for the year 1996. The four input patterns, which represent the various seasons of the year, were classified into four separate regions on the map. These four regions are separated by boundaries coloured in grey, which determines the similarity or dissimilarity among the input patterns used. The lighter the grey the higher the degree of similarity.

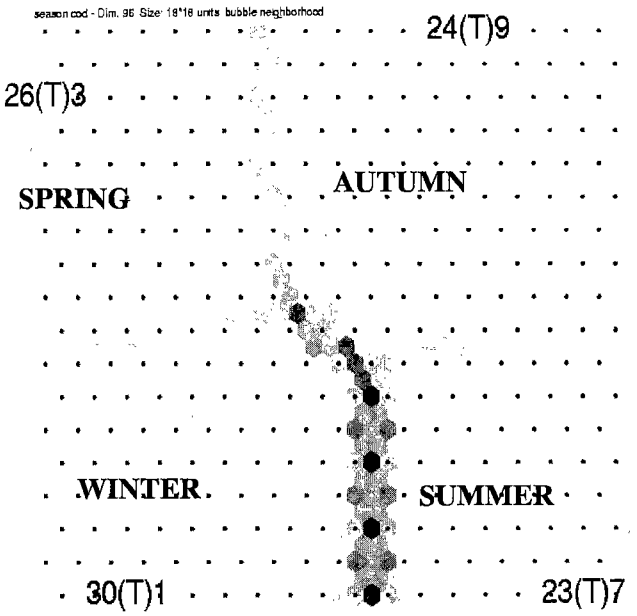


Fig. 6. Seasonal SOM for 1996

From figure 6 it can be observed that winter is located on the bottom left portion of the map, summer on the bottom right portion of the map, autumn on the top right and spring on the top left. Between winter and summer, the darkest grey separation was identified on the map. This is because during winter the weather is very cold, but heating is mostly done with paraffin other than electric energy. In summer a lot of electric energy goes into cooling thereby resulting in a drastic increase in energy consumption during summer. The next in the degree of grey is the difference between winter and autumn. The difference between winter and spring is last in the degree of grey. Spring is a bit cold so the electric energy consumption pattern is closer to that of winter. During this time, consumers are still using paraffin as the fuel for heating. Comparing the electric energy consumption patterns for spring and summer, it can be concluded that the pattern for spring is closer to autumn than that of summer. This is because in autumn, the heating fuel is also paraffin. So by just looking at the SOM of figure 6, a lot can be said about the electric energy consumption patterns of the inhabitants of the area being served by that transformer. Furthermore, if one is confronted with an unlabeled data set, by use of this map, the season within which this data was recorded can be identified.

3.4 Regression based SOM In this section, the Self-Organising Maps is used as a tool for regression. An input vector representing a particular day is deleted from the input data space. SOM is applied after which regression is achieved by searching for the Best Matching Unit (BMU) using the known vector components [5]. Since SOM is being used to predict the omitted part

†The date in Table 1 has been shortened for lack of space. WED-3/20 (20th March, Wednesday) is represented as 20(W)

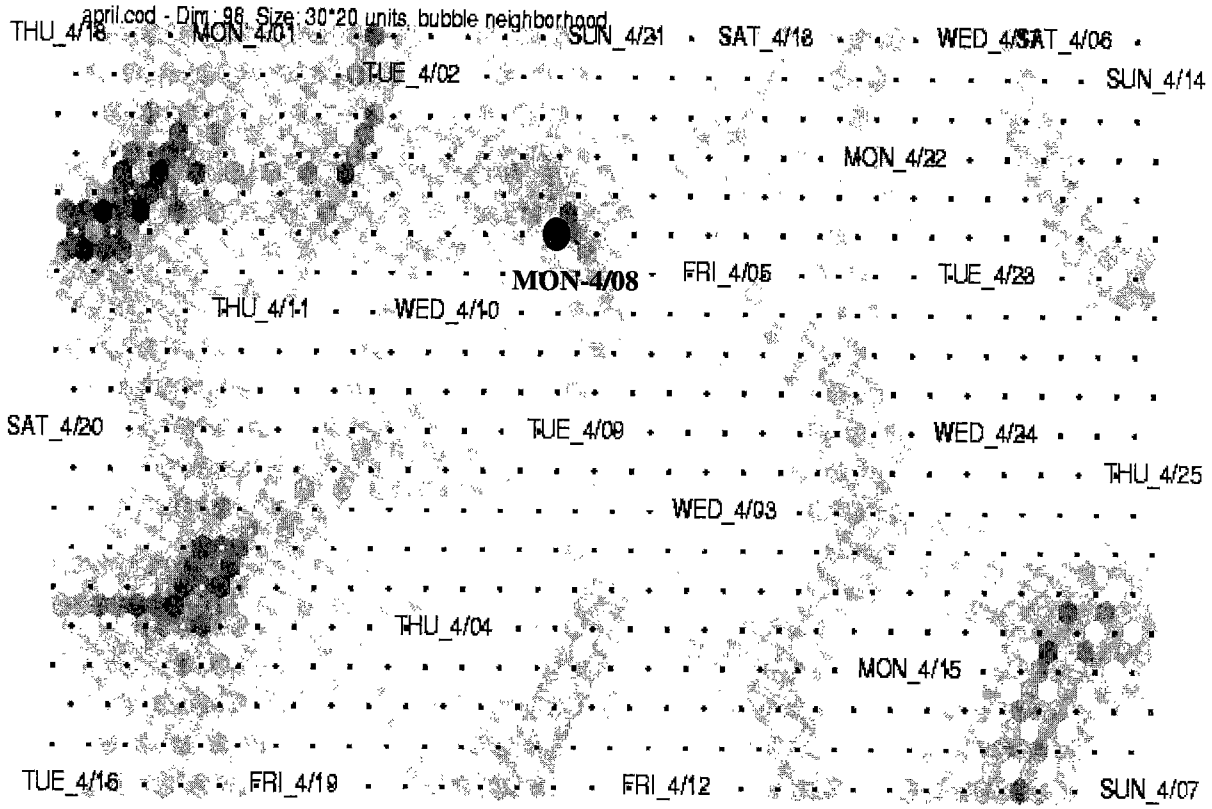


Fig. 7. SOM for April 1996. The input vector for 8th April (monday) was omitted. The large filled circle was identified as the BMU for 8th April (monday). The nomenclature of labeling for figure 7 is as follows, 8th April (Monday) is represented by MON-4/08.

of the data, an even more detailed training is required than that for season type pattern classification. In view of this, learning cycle of 20,000 was chosen; the learning rate factor was maintained at 0.03. As an output, an approximation of the unknown components of the input vector was obtained. The data set for the SOM comprised of 4 input measurements (percentage load on the transformer, the transformer oil temperature, atmospheric temperature and transformer-surrounding temperature). These are hourly measurements hence each input vector is made up of 96 components, which have been normalised as in section 3.1. The map was developed on a 600 (30 X 20) unit grid. Figure 7 shows the feature map for the month of April. In this data set, the data vector for 8th April (Monday) was omitted. Using eq.(5), all the 600 nodes on the grid were compared to the original data for 8th April (Monday) and the node with the minimum value was selected as the Best Matching Unit.

$$Err = \sum_{j=1}^{n}(x_j - m_{ij})^2 \quad \cdots\cdots\cdots\cdots\cdots\cdots (5)$$

where x_j and m_{ij} are the j-th component value of the n-th dimensional input data and i-th unit (node) respectively. Figure 8 compares the parts (percentage load on the transformer, the transformer oil temperature, atmospheric temperature and transformer-surrounding temperature) of the input data vector for 8th April (Monday), which was omitted to that which was identified as the BMU after the training process. It was observed that though 8th April (Monday) was excluded from the input data, the SOM was able to learn the other input vectors very well so as to predict with a good precision, a data vector not included in the input data space. Under heavy load, the temperature of the oil for cooling and insulating the transformer windings rises. This causes deterioration of the dielectric strength of the oil, which is used to insulate and cool the transformer thereby reducing the life span of the transformer. Prediction of the oil temperature rise is therefore necessary so that the necessary countermeasures such as load transfers and load shedding can be implemented to forestall physical damage to the transformer. Figure 9 is a comparison of the SOM to conventional method currently been used by the service provider for predicting the transformer oil temperature on the 8th April (Monday). The two methods were compared with the recorded data for 8th April (Monday) using the mean squared error (MSE) in eq.(6).

$$MSE = \frac{1}{N}\sum_{j=0}^{N-1}(r_j - p_j)^2 \quad \cdots\cdots\cdots\cdots\cdots\cdots (6)$$

where r_j is the recorded data, p_j is the predicted data and N is the number of components per data vector. The MSE for the SOM prediction was 0.31 whereas that for the conventional method was 8.38.

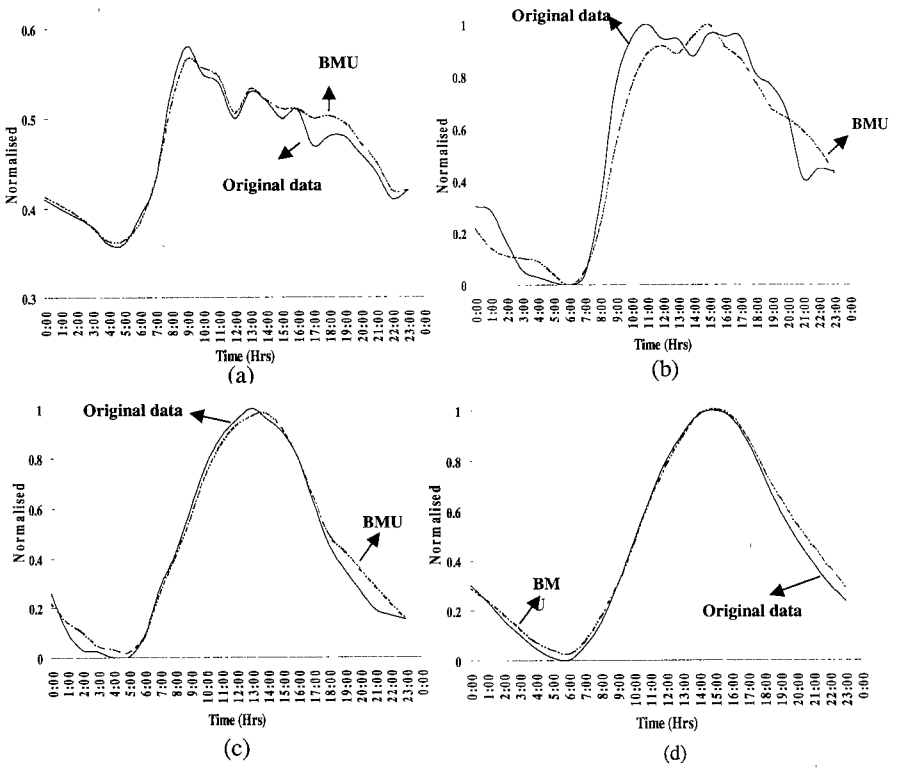


Fig. 8. Comparison of the parts of the input data vector for 8th April (Monday), which was omitted and the BMU after the training process. (a) Percentage loading of the transformer, (b) transformer oil temperature, (c) atmospheric temperature and (d) transformer surrounding temperature. The solid and broken lines refer to the original data and the Best Matching Unit (BMU) respectively.

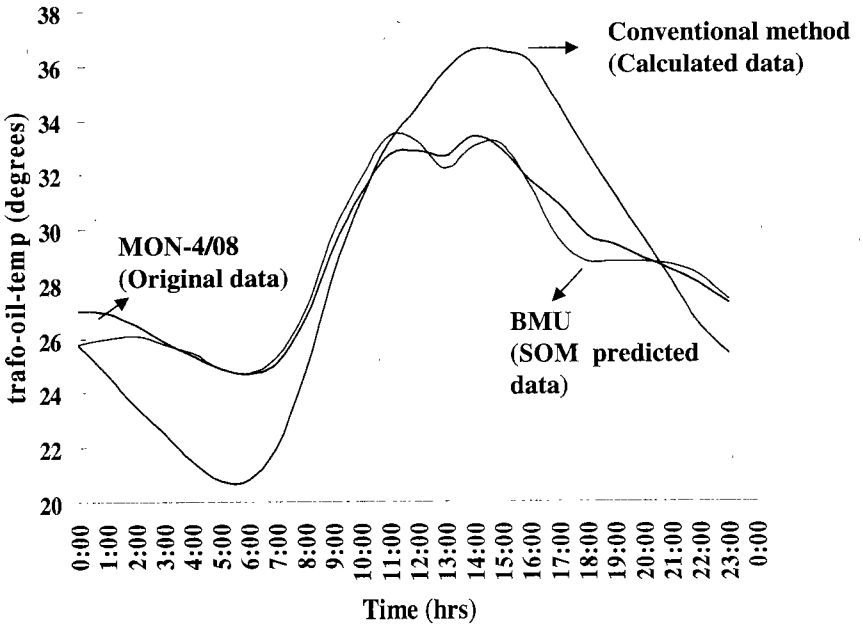


Fig. 9. Predicted values for the transformer oil temperature of 8th April (Monday) by SOM and the conventional method using calculation.

The SOM resulted in a better approximation of the data for 8th April (Monday) as compared to the conventional method for data prediction. With the conventional method, prediction of the oil temperature of the transformer according to the operational guideline of oil filled transformers [6] is based on various constants such as time constant necessary for the calculation of the corresponding optimal cooling conditions of the installation environment of the transformer. This is a difficult task and it also limits the prediction accuracy

of the method. The hourly temperature changes are calculated using the following equations:

$$\theta_{OK_1} = \theta_{ON}(\frac{(K_1)^2R+1}{R+1})^m \quad \cdots\cdots\cdots\cdots\cdots (7)$$

where θ_{ON} is the highest rise in temperature at normal load, K_1 is the ratio of the load P_1 to the normal load P_N, R is the ratio of loss at normal load to loss at no load and m is a constant determined by the mode of cooling. θ_{ON} and R can be obtained from data from factory test conducted on the transformer at the time of manufacture. Furthermore, after more than a unit time has elapsed after a change of transformer load from K_1P_N to K_2P_N, the rise in oil temperature is as shown in eq.(8).

$$\theta_{O(K_1\sim K_2)} = (\theta_{OK_2}-\theta_{OK_1})(1-\exp(\frac{-t}{\tau}))+\theta_{OK_1} \quad (8)$$

where τ is the time constant of the oil temperature change.

4. Conclusion

The Self-Organising Map (SOM) is a very versatile and flexible tool for database exploration. It has excellent visualisation characteristics, which gives the user an opportunity to visualise the inherent features of the data set. It also has very effective clustering capabilities. In its application to the power transformer database, the SOM provides an in-depth knowledge about the consumption pattern of the consumers in a particular day or season. The information gathered from the database by the use of SOM, provides the energy provider with a guide as to the consumption pattern of its consumers in any particular day or season. SOM was also found to be very effective in regression and this could be a very effective tool for predicting unknown as well as incomplete data vectors in a database. This could be very vital information for planning engineers in their load forecasting and other prediction and planning activities.

(Manuscript received September 6, 2000, revised March 22, 2001)

References

(1) T. Kohonen: "Self-Organizing formation of topologically correct feature maps", Biological Cybernetics, 43(1):59-69,1982.

(2) T. Kohonen: "The Self-Organizing Map", Proceedings of IEEE, 78:1464-1480,1990.

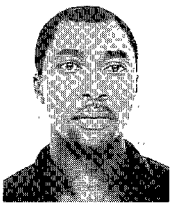
(3) T. Kohonen: "Self-Organizing Maps", Springer-Verlag, 1997.

(4) A. Murray: "Applications of Neural Network", Kluwer Academic Publishers, pp 157-189.

(5) O. Simula, J. Vesanto, E. Alhoniemi and J. Hollmen: "Analysis of complex systems using the Self-Organising Map", Proceedings of the International Conference on Neural Information Processing and Intelligent Information Systems, Vol. 2,1313-1317, 1997.

(6) Technical report of the Institute of Electrical Engineers of Japan, Part 1,No.143: "The operation guideline of the oil filled transformers (in Japanese) ", (1978).

Obu-Cann Kwaw (Non-member) was born in Accra, Ghana, on December 10, 1964.



He received the B.Sc in Electrical Engineering in June 1991 from the University of Science and Technology, Kumasi, Ghana.
In March 2000 he received the M.Sc degree from the Department of Electrical and Electronic Engineering, Tottori University, Japan. He is currently a Ph.D student at the Department of Electrical and Electronic Engineering, Tottori University, Japan. Research Interests are in Neural networks and its applications. Currently working on Self Organising Maps and its application to power systems.

Fujimura Kikuo (Member) was born on September 9th, 1962.

He received B.E. and M.E. degree from Tottori University.
He received Ph.D in Information Science from Kyushu Institute of Technology in March 2000.
He is currently working as a research associate in the Department of Electrical and Electronic Engineering of Tottori University. Research interests are in applications of neural networks and Self-Organising Maps. He is a member of IEICE and JNNS.

Tokutaka Heizo (Non-member) was born on September 28th, 1937.

He received the B.Sc in Science in 1960 from the Department of Physics, Osaka University.
In 1970 he received the Ph.D in Physics from the Department of Physics, University of York (UK).
He is currently a Professor at the Department of Electrical and Electronic Engineering, Tottori University, Japan. Research Interests are in Self Organising Maps and its applications to Data Mining Problems and Control of Industrial Processes.

Ohkita Masaaki (Member) was born in Osaka, Japan, on January 18, 1943.

He received the B.E. degree in Electrical Engineering in 1966 from Osaka Technical College, Osaka, Japan.
He received the M.E. and Ph.D degrees in 1968 and 1988 respectively from the University of Osaka, Osaka Prefecture, Japan.
He joined Tottori University, Japan in 1968 and since 1993, he has been a Professor in the department. Research Interests are in application of fuzzy theory, autonomous mobile robots, and analysis of power systems.

Inui Masahiro (Member) was born in Osaka, Japan, on July 28, 1951. He received the B.E. degree in 1974 from Tottori University, Tottori, Japan.

From 1974-1998 he worked with DAIHEN corporation, Engineering Department.
In 1997 he enrolled as a Ph.D student at the Department of Electrical and Electronic Engineering, Tottori University.
Research Interests are in fuzzy theory and its applications to power systems.