Journal of Computational Biology and Bioinformatics Research Vol. 3(4), pp. 47-52, July 2011
Available online http://www.academicjournals.org/jcbbr
ISSN-2141-2227 ©2011 Academic Journals

Full Length Research Paper

Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods

Yong Poh Yu[1], Rosli Omar[1], Rhett D. Harrison[2], Mohan Kumar Sammathuria[3] and Abdul Rahim Nik[4]

[1]Department of Electrical Engineering, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia.
[2]Xishuangbanna Tropical Botanical Garden, Menglun, Mengla, 666303, Yunnan, China.
[3]Malaysia Meteorological Department, Jalan Sultan, 46667 Petaling Jaya, Malaysia.
[4]Forest Research Institute, 52110 Kepong, Selangor, Malaysia.

Accepted 14 June, 2011

This paper outlines two hybrid approaches to investigate the nonlinear relationship between size of a forest fire and meteorological variables (temperature, relative humidity, wind speed and rainfall). Self organizing map was used to cluster the historical meteorological variables. The clustered data were then used as inputs for two different approaches, the back-propagation neural network and the rule generation approaches. A back-propagation neural network was trained based on these inputs to classify the output (burnt area) in categorical form, namely; small, medium, large and extremely large. Several sets of rules were also generated from the data clustered by the self organizing map. Experimental results showed that both approaches gave considerable accuracy. Back-propagation neural network achieved a higher rate of accuracy than rule generation approach because the rule generation approach could not predict any criterion that goes beyond the set of rules.

Key words: Forest fire, self organizing map, back-propagation neural network, rule-based system.

INTRODUCTION

Forest fire is one type of significant disturbance to the forest ecosystem. There is increasing evidence to show that the global climate change may cause a significant effect on the forest fire (Torn and Fried, 1992; Williams et al., 2001). New evidence shows that the more forests burn the more susceptible to future burning they become (Rowell and Moore, 2000). Forest fire eventually causes destruction to the community. For example, greater than 2.7 million hectares of forest area were burnt in Portugal, from 1980 to 2005. Some fire seasons caused human deaths and losses of large territory (Cortez and Morais, 2007).

Earlier studies have shown that there is a relationship between meteorological conditions and forest fire occurrence. It is believed that fire is largely a function of meteorological variables, that is temperature, relative humidity, wind speed and precipitation (Cortez and Morais, 2007; Amiro et al., 2004). Some numerical indices incorporate these meteorological variables into their calculations. An example is the Canadian forest fire weather index (FWI). This index was adopted and used by several countries including those from developing countries (de Groot et al., 2005).

Investigation on forest fire modeling and its nonlinear relationship with meteorological conditions keep increasing. Data mining technique is one of the common approaches to determine the relationship. In the study done by Stojanova et al. (2006), they built different models based on different data mining techniques to predict the forest fire in different regions. They concluded that bagging of decision trees gave best results

*Corresponding author. E-mail: ypyu@siswa.um.edu.my Tel: 60-16-2109887.

compared to logistic regression, random forest, and decision tree. However, a study from Cortez and Morais (2007) showed that support vector machine (SVM) was the technique that gave the best performance. Thus, there is still lack of comprehensive studies on the performance and effectiveness of data mining techniques in predicting and clustering the forest fires.

In this research, two different hybrid approaches are presented. Basically, a self organizing map (SOM) was applied in the first stage to cluster the characteristics of the meteorological conditions. The clustered patterns were then used in subsequent approaches to classify the forest fire. Two approaches, namely back-propagation neural network and rule-based system, were developed.

Self organizing map was selected as the clustering method so that the system could be trained without any supervision. Self organizing map is able to cluster those samples with similar characters (inputs) into a same category (neuron) by itself, where no target or output is needed in first stage. The dimensionally reduced map was then projected to the back-propagation neural network for a supervised training. As the data were clustered at first stage, it would not be too time consuming at second stage.

A rule-based system was selected as part of the hybrid system too. The generated sets of rules can be integrated and incorporated with other models (such as blackboard system) as well. Thus, the results from rule-based system can be used as the pre-condition and post-condition criteria in other model (McManus, 1992). For instance, in a biodiversity change blackboard model, this rule-based system can be used to predict the forest fires and subsequently, the respective biodiversity change.

DATA MINING TECHNIQUES

Self organizing map

The self organizing map (SOM) is an unsupervised learning algorithm proposed by (Kohonen, 1982). It is quite often used as a tool for clustering, classification, and data mining (Vesanto, 2000). Typically, it provides a way to reduce the topology information from high dimension to lower dimension, which is normally represented by one or two dimensional layer of neurons. Number output or target is given to the SOM for the training as it is capable of self-learning.

According to Kalteh et al. (2008), there are 3 types of procedures required to apply a SOM, namely data gathering and normalization, training and information extraction. In the data gathering and normalization step, the input variables are normalized so that all variables have equal importance in the SOM. The SOM trains itself by finding a best match unit (BMU) or winning neuron in its output map. The common criterion used to find a best match unit is Euclidean distance. Let input vector $X_i = \{x_{i1}, x_{i2}, x_{i3}\ldots x_{in}\}$ and SOM neuron $X_j = \{x_{j1}, x_{j2}, x_{j3}\ldots x_{jn}\}$, then the Euclidean distance between X_i and X_j, denoted by d_{ij}, is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (1)$$

The weights of BMU and its topological neighbouring neurons are updated in such a way as to reproduce the input pattern (Kalteh et al., 2008). The process is continued by other input vectors until convergence, this is known as incremental training algorithm. There is another type of training known as batch training algorithm. Batch training algorithm determines the BMU for each input vector. Then every BMU (and its topological neighbouring neurons) is updated based on the average of all of input vectors that fire that particular BMU. Batch training algorithm was implemented in this research.

Back-propagation neural network

Back-propagation neural network system is also a common approach in data mining (Sunar and Ozkan, 2001; Antonie et al., 2001). Back-propagation algorithm is often used to train a feed-forward multilayer perception (MLP) network. A MLP network contains two or more layers. A typical 3 layer MLP network consists of an input layer, a hidden layer and an output layer. Number of input neurons in an input layer is equal to the number of elements existing in an input vector. Hidden layer is the internal layer where the number of neurons can be chosen in trial and error manner. Output layer has the output neurons where the number of neurons is same as the number of desired output variables.

In back-propagation neural network, MLP network is trained iteratively until the difference of values (or error) between output neurons and output targets has converged. MLP forwards the input vectors or training samples from the input layer, to the hidden layer and lastly to the output layer. During the feed-forward propagation, the weight of each MLP neuron is updated based on an activation function. A common activation function is the sigmoid activation function (Gardner and Dorling, 1998).

There are many types of training algorithms that can be used to train the back-propagation neural network (Chai et al., 2008). A common training algorithm is Levernberg-Marquardt algorithm (Mas et al., 2004; Atluri et al., 1999). Levernberg-Marquardt algorithm for neural network training was developed by (Hagan and Menhaj, 1994). They concluded that Levernberg-Marquardt algorithm was much more efficient than other techniques such as conjugate gradient algorithm and variable learning rate

algorithm. They found that Levernberg-Marquardt algorithm was efficient for network that contained no more than a few hundred neurons. Thus, this algorithm is widely used now. In this research, the network that is used is also within this size.

DATA COLLECTION AND METHODOLOGY

Forest fire data

Forest fire data have been collected from the study of Cortez and Morais (2007) which are available in the UCI machine learning repository. The dataset contains forest fire occurrence, forest fire weather index (FWI) components in Montesinho Natural Park, a northeast region of Portugal. Weather observations were collected by Braganca Polytechnic Institute and integrated to the forest fire dataset. The park was divided into 81 distinct locations by placing a 9×9 grid onto the map of the park. The dataset has a total of 517 samples, from 2000 until 2007.

In our research, four meteorological variables, yield temperature, relative humidity, wind speed and rainfall, had been used to classify the size of forest fire. The data were categorized into two different sets, randomly, which were used as training dataset, and testing dataset. The training dataset contained 80% or 414 samples out of the total samples, including both of non-fire occurrence and fire occurrence samples. The remaining 103 samples were used as testing data, which were not projected to self organizing map and back-propagation neural network. Every sample was defined as a 5×1 vector, where first four elements were the meteorological variables and the last element was the burnt area of that particular sample. Thus, a 5×414 matrix was formed to represent the training samples.

Self organizing map (SOM)

Prior to the self organizing map system training, all the training data samples were normalized so that the mean of each variable was 0 and the standard deviation (SD) of each variable was 1. The training data samples were then projected to the SOM training phase using MATLAB version R2009b. Batch unsupervised weight algorithm was implemented using MATLAB. There were only four meteorological variables included in the SOM training. The burnt area variable was not used for the SOM training. As the outputs (burnt area) of testing samples may be unknown or unidentified, burnt area was excluded as part of the training inputs. Thus, every training and testing sample was defined as 4×1 vector and a 4×414 matrix was formed to be trained. Another 4×103 matrix was formed to be tested.

After the SOM was trained, training samples with similar characteristics (in terms of the meteorological variables) would be mapped to the same neurons (clusters) in the output map (Kalteh et al., 2008). In this research, each of the training samples was mapped to the neuron that had the shortest Euclidean distance with that particular sample.

Figure 1 shows the distribution of training data samples in the output map. From the figure, some neurons (clusters) had more samples, such as neuron at (2, 1) position had 45 training samples. All these 45 samples were said to have similar characteristics.

The testing data samples were then classified into the map. Euclidean distance between every SOM output neuron and the testing data sample was calculated. Similarly, each testing data sample was mapped into the neuron that had the shortest Euclidean distance.

Back-propagation neural network

The samples were classified into 16 different clusters. Every cluster was then projected to its own back-propagation neural network. Four meteorological variables of training samples in each cluster were the inputs of the neural network. The 5th element, namely burnt area, was the output of the training set in the neural network. The output (burnt area) was used in this stage for training purpose. Then, the results of testing samples (extracted from SOM map) will be projected to respective back-propagation neural network that was trained earlier.

Back-propagation neural network does not necessarily need to have testing samples with known outputs (burnt area). It is good to be used to make classification on those testing samples with unknown outputs. It is also suitable to train (and test) the SOM neurons that have no outputs, which was proposed in this research. Levenberg-Marquardt training algorithm was used for the neural network training.

The testing samples in each cluster were projected to the trained neural network to classify the burnt area of forest fire. Both of the experimental and original values of burnt area were in continuous form. In order to have better representation, the burnt area was transformed from continuous value to categorical form, namely small, medium, large, and extremely large. Empirical rule was adopted and implemented into the transformation. Empirical rule states that for a normal distribution, about 68% of the data will fall within 1 standard deviation of the mean, about 95% of the data will fall within 2 standard deviation of the mean, and about 99.7% of the data will fall within 3 standard deviation of the mean. With zero mean and unity standard deviation, the transformation was based on the following rules:

If $|normalized\ burnt\ area| < 1$, then it is small;

If $1 \le |normalized\ burnt\ area| < 2$, then it is medium;

If $2 \le |normalized\ burnt\ area| < 3$, then it is large;

If $|normalized\ \ burnt\ area| \ge 3$, then it is extremely large.

Rule-based system

Apart from the back-propagation neural network, a rule-based system was also generated from each cluster for the classification. Meteorological variables and burnt area of every training sample in each cluster were analyzed. Rule-based system was then developed based on the range of input (meteorological) variables and output variable (burnt area). IF… THEN… type of rules was implemented.

To prevent bias on the results, variables of testing samples were not used for the rule generation. The testing samples were only used to validate the effectiveness of the proposed system.

Table 1 shows some examples of the rules that were generated from SOM. Rule 1 in Table 1 can be interpreted as, “if a testing sample in SOM cluster has the normalized temperature between 0.7299 and 2.3845, normalized relative humidity between -1.3499 and -0.1316, normalized wind speed between -1.585 and -0.0406, normalized rainfall of -0.0818, then the burnt area of that particular testing sample is classified as small ”.

RESULTS AND DISCUSSION

Self organizing map outputs

Table 2 summarizes the distribution of the training

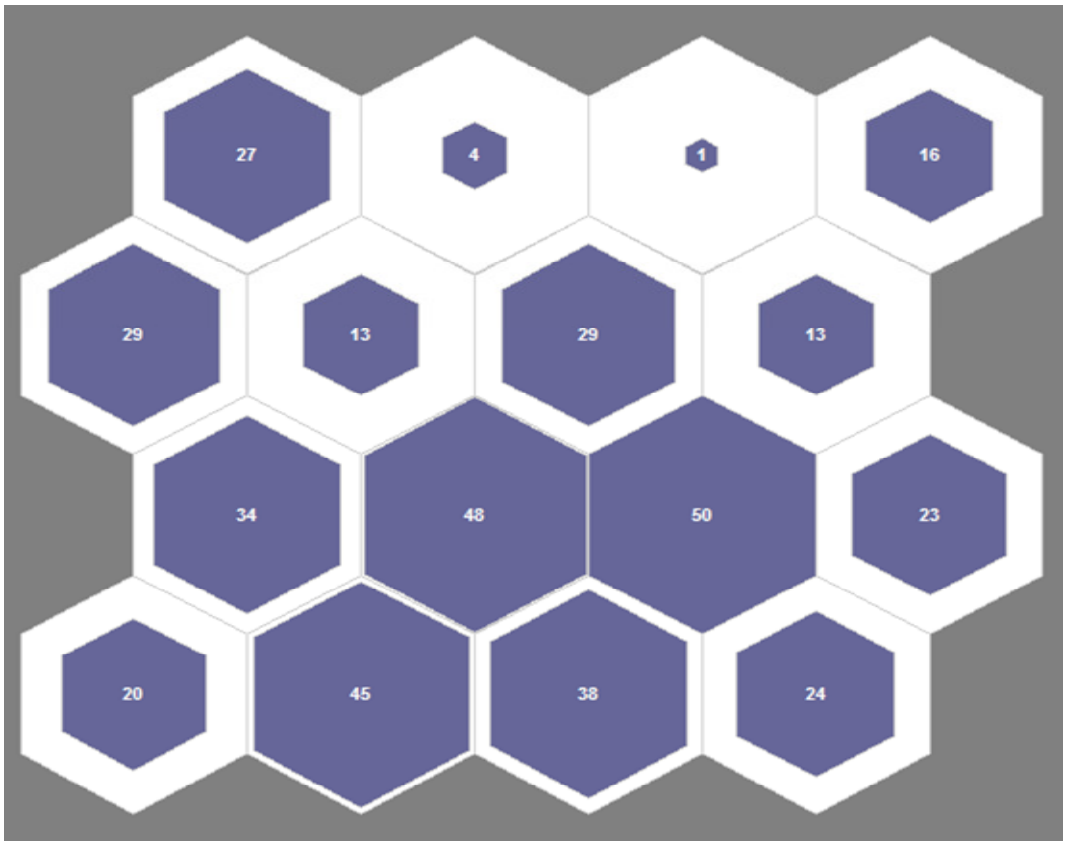

Figure 1. SOM output map after training. The SOM output neuron is represented by a hexagon. A number is shown in each SOM neuron to indicate the number of training samples that is mapped into that SOM neuron. The more number of training samples in the neuron, the more area is shaded.

samples and testing samples in the output map. The ratio of number of testing samples over number of training samples for each cluster was between 0 to 47%.

Comparison between outputs of back-propagation neural network and rule-based system approaches

Table 3 summarizes the accuracy of the burnt area classification for the testing samples, based on back-propagation neural network and rule-based system approaches. Overall, it provides agreement to the hypothesis that SOM is a good clustering method where the samples with similar characteristics are mapped into the same cluster. For this research, a single SOM could not predict the burnt area since the burnt area was not projected into the SOM training.

Thus, a subsequent approach is needed to predict the burnt area, yields the back-propagation neural network or rule-based system.

One of the back-propagation neural network issues is the non-representativeness of training samples (Chang et al., 1993). If the training samples are not representative of the testing samples, the network may not be classified very well due to the limitation of the training samples (the training samples are too few). For instance, a testing sample with the normalized burnt area of 12.1952 (extremely large burnt area) was classified as a small burnt area by the neural network system.

The result extracted from back-propagation is slightly better than the result extracted from rule-based system. The proposed rule-based system has a weakness where if a testing sample has a criterion that goes beyond the set of rules, then the system cannot recognize and classify it accurately. As a back-propagation neural network can be trained and used to classify all kinds of patterns, it has a better position to achieve higher accuracy. Future studies may include the SOM training with 5 variables (4 meteorological variables and burnt area) and its hybrid approaches of back-propagation neural network or rule-based system.

Comparison with existing methods

The datasets were used by other researchers using

Table 1. Examples on rule generation (from SOM cluster 2).

Rules	1	2
Normalized Temperature TP	0.7299 ≤ TP ≤ 2.3845	0.9587 ≤ TP ≤ 1.5748
Normalized Relative Humidity RH	-1.3499 ≤ RH ≤ -0.1316	-1.0528 ≤ RH ≤ -0.5001
Normalized Wind Speed WS	-1.585 ≤ WS ≤ -0.0406	-0.5532 ≤ WS ≤ -0.5532
Normalized Rainfall RF	-0.0818 ≤ RF ≤ -0.0818	-0.0818 ≤ RF ≤ -0.0818
Burnt area BA	SMALL	MEDIUM

Table 2. Distribution of dataset in SOM output map.

Position of SOM output neuron	Number of training samples	Number of testing samples	Number of testing samples/number of training samples (%)
1	20	6	30
2	45	12	27
3	38	11	29
4	24	9	38
5	34	16	47
6	48	10	21
7	50	7	14
8	23	8	35
9	29	2	7
10	13	2	15
11	29	6	21
12	13	2	15
13	27	5	19
14	4	0	0
15	1	0	0
16	16	7	44
Total	414	103	20

different approaches, such as the work done (Cortez and Moraisv 2007; Ku Ruhana and Khor, 2009). Cortez and Moraisv (2007) concluded that support vector machine (SVM) was the best among the approaches they used. They found that it was better to use weather conditions (meteorological variables) rather than FWI variables. Also, the spatial and temporal variables (irrelevant variables) will not improve the performances of SVM. Reason was not stated to support or discuss the results. Ku Ruhana and Khor (2009) used sliding window technique to extract the patterns. The inputs were also based on meteorological variables. The rules were then generated from the patterns extracted from the sliding window technique. They concluded that the proposed method is able to produce a high-accuracy result. Irrelevancy issue was not discussed in the paper.

Issues on irrelevant variables may occur in the data mining model. If irrelevant variables are chosen, accuracy and performance may be greatly influenced. In this paper, SOM approach is suggested to be one of the alternatives to reduce the impact. As SOM algorithm is able to reduce the data dimensions, irrelevancy issue can be reduced to minimum. Thus, future studies may include some other conditions that potentially contribute to the forest fires, such as topology factors, types of forest and location, etc.

Conclusion

The nonlinear relationship between size of forest fire and meteorological variables (temperature, relative humidity, wind speed and rainfall) was investigated using self organizing map together with its hybrid approaches, namely back-propagation neural network or rule-based system. The rules or outputs generated from these approaches can be used to classify the size of forest fire. This study was wholly based on the qualitative analysis on the 4 meteorological variables to predict the size of forest fire. It can be much more challenging in a real time forest fire management and analysis. More data mining techniques are needed to analyze the relationships. As

Table 3. Accuracy of burnt area prediction.

Position of SOM output neuron	Number of testing samples	Number of successful prediction	
		Back-propagation network	Rule-based system
1	6	5	4
2	12	9	10
3	11	10	8
4	9	9	7
5	16	15	14
6	10	9	9
7	7	5	6
8	8	8	6
9	2	2	2
10	2	2	1
11	6	5	4
12	2	2	2
13	5	5	4
14	0	0	0
15	0	0	0
16	7	7	7
Total	103	93	84

SOM is able to reduce the data dimension, irrelevancy issue can be reduced. Future studies may include some other conditions that potentially contribute to the forest fires, such as topology factors, types of forest and location, etc.

REFERENCES

Amiro BD, Logan KA, Wotton BM, Flannigan MD, Todd JB, Stocks BJ, Mattell DL (2004). Fire weather index system components of large fires in the Canadian boreal forest. Int. J. Wildland Fire, 13: 391–400.

Antonie M, Zaiane OR, Coman A (2001). Application of Data Mining Techniques for Medical Image Classification. In: Second International Workshop on Multimedia Data Mining (MDM/KDD), pp. 94-101.

Atluri V, Hung CC, Coleman TL (1999). An artificial neural network for classifying and predicting soil moisture and temperature using Levenberg-Marquardt algorithm. In: Proceedings IEEE Southeastcon '9, pp.10-13.

Chai SS, Veenendaal B, West G, Walker JP (2008). Back propagation Neural Network for Soil Moisture Retrieval Using NAFE'05 Data: A Comparison of Training Algorithms. In: Proceedings of the International Society for Photogrammetry and Remote Sensing (ISPRS) XXIth Congress, pp. 1345-1350.

Chang W, Bosworth B, Carter GC (1993). Empirical results of using back-propagation neural networks to separate single echoes from multiple echoes. Neural Networks, IEEE Trans., 4(6): 993-995.

Cortez P, Morais A (2007). A data mining approach to predict forest fires using meteorological data. In: Proceedings of the 13th Portugese Conference on Artificial Intelligence, pp. 512-523.

de Groot WJ, Field RD, Brady MA, Roswintiarti O, Mohamad M (2005). Development of the Indonesian and Malaysian fire danger rating systems. Mitigation Adaptation Strat. Global Change, 12(1): 165-180.

Gardner MW, Dorling SR (1998) .Artificial neural networks (the multi-layer perceptron) – a review of applications in the atmospheric sciences. Atmos. Environ., 32: 2627–2636.

Hagan MT, Menhaj MB (1994). Training feedforward networks with the Marquardt algorithm. Neural Networks, IEEE Trans., 5(6): 989-993.

Kalteh AM, Hjorth P, Berndtsson R (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environ. Model. Softwar, 23(7): 835-845

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

Ku Ruhana KM, Khor JY (2009). Pattern Extraction and Rule Generation of Forest Fire Using Window Sliding Technique. Comput. Inf. Sci., 2(3): 113-121

Mas JF, Puig H, Palacio JL, Sosa-Lopez A (2004). Modelling deforestation using GIS and artificial neural networks. Environ. Model. Software, 19(5): 461-471.

McManus JW (1992). Design and Analysis Techniques for Concurrent Blackboard Systems. PhD Thesis. The College of William and Mary in Virginia. pp. 35-41

Rowell A, Moore DPF (2000). Global review of forest fires. International Union for Conservation of Nature and Natural Resources. p. 3

Stojanova D, Panov P, Kobler A, Dzeroski S, Taskova K (2006). Learning to Predict Forest Fires with Different Data Mining Techniques. In: Conference on Data Mining and Data Warehouses, pp. 255-258.

Sunar F, Ozkan C (2001). Forest fire analysis with remote sensing data. Int. J. Remote Sens., 22: 2265-2277.

Torn MS, Fried JS (1992). Predicting the impact of global warming on wildfire. Clim. Change, 21: 257-274.

Vesanto J (2000). Using SOM in data-mining. Dissertation, Helsinki University of Technology. pp. 26-38

Williams AAJ, Karoly DJ, Tapper N (2001). The Sensitivity of Australian Fire Danger to Climate Change. Clim. Change, 49: 171-191.