Predicting Size of Forest Fire Using Hybrid Model

Guruh Fajar Shidik and Khabib Mustofa

Universitas Dian Nuswantoro Indonesia,
Universitas Gadjah Mada, Indonesia
guruh.fajar@research.dinus.ac.id,
khabib@ugm.ac.id

Abstract. This paper outlines a hybrid approach in data mining to predict the size of forest fire using meteorological and forest weather index (FWI) variables such as Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), temperature, Relative Humidity (RH), wind and rain. The hybrid model is developed with clustering and classification approaches. Fuzzy C-Means (FCM) is used to cluster the historical variables. The clustered data are then used as inputs to Back-Propagation Neural Network classification. The label dataset having value greater than zero in fire area size are clustered using FCM to produce two categorical clusters,i.e.: *Light Burn*, and *Heavy Burn* for its label. On the other hand, fire area label with value zero is clustered as *No Burn Area*. A Back-Propagation Neural Network (BPNN) is trained based on these data to classify the output (burn area) in three categories, *No Burn Area, Light Burn* and *Heavy Burn*. The experiment shows promising results depicting classification size of forest fire with the accuracy of confusion matrix around 97, 50 % and Cohens Kappa 0.954. This research also compares the performance of proposed model with other classification method such as SVM, Naive Bayes, DCT Tree, and K-NN that showed BPNN have best performance.

Keywords: Forest fire Prediction, FCM, Back-Propagation Neural Network, Data Mining.

1 Introduction

Forest fire is a common natural world phenomenon. Every year millions of hectares of forests in the world are destroyed [1], between 1980 - 2007 at least 2.7 million hectares were burnt in Portugal [2]. This caused severe damages to the natural environment and resulted in loss of precious human lives. Forest fire is one of the major environmental concern that affects the preservation of forests, resulting in economical and ecological damage that causes human suffering.

Referring to Elmas [3], quick fire detection and response are effective ways in reducing the damages caused by forest fires. Various studies have been made in order to improve early fire prediction and detection systems that helps to develop response strategies during the fire. It means, one of the key successes of putting

Linawati et al. (Eds.): ICT-EurAsia 2014, LNCS 8407, pp. 316–327, 2014.

out forest fire is by providing an early warning detection. Early warning detection is related to accurate prediction of results based on determined parameters. There are three trending techniques that could be used in predicting forest fire such as the use of satellite data, infra red or smoke scanners and local sensors, for example, using the meteorological ones [2].

Safi et al [4] tried to overcome forest fire impacts by making prediction using data mining technique. A future event has always been considered a mysterious activity scientists trying to treat into scientific activities based on theories and models. Predictions in data mining can be used in identifying many real world problems such as financial forecasting and prediction of environmental applications or to test scientific understanding of the behaviour of complex systems or phenomena. The predictions are also used as a guide or basis for decision making [2].

In [5], based on the perspectives of forest fire, it is mentioned that several scientists around the world had utilized statistical approaches such as regression analysis, probabilistic analysis and artificial intelligence. Some data mining techniques have been applied in the domain of fire detection, for example by adopting meteorological data to predict forest fire [2]. Back propagation neural network and the rule generation approach [6], fuzzy c-means clustering application in the case of forest fire [5], artificial neural network to the real word problem of predicting forest fire [4], Neural Network (NN) and Support Vector Machines to predict forest fire occurrence based on weather data [7], decision tree algorithm namely C4.5 to extract a forest fire data and classifying hotspot occurrences [8].

This research aims at proposing an approach for predicting the size of forest fire occurs based on meteorological and forest weather index dataset consisting of eight variables: FFMC, DMC, DC, ISI, temperature, RH, wind and rain. The size of forest fire will be classified using Back-Propagation Neural Network into three categories,i.e: *No Burn Area, Light Burn* and *Heavy Burn.* As the label (area) in datasets of size of forest fire is numerical, before classification process, the data should be clustered into three classes. We split the dataset into two part: *the data with zero value* and *data having value greater than zero.* The process of clustering two categories dataset (light and heavy burn) is done with unsupervised method FCM in label data (area) that have value greater than zero, while for No Burn Area, it is done by selecting label (area) that have value zero. We used ten fold cross validations in separating training dataset and testing dataset with shuffled and stratified sampling. Confusion matrix and Kappa is used in evaluating the performance of the model.

The remaining of this paper will be organised as follows: chapter two talks about related works, chapter three talks about fundamentals of the approach, chapter four describe the research method used to predict the size of forest fire. The rest of the papers are discussion on the results and, the last chapter is conclusion and future work of this research.

2 Related Works

Satoh et al [9] developed a system for predicting the dangers of a forest fire. A simulation of dangers related to forest fire was developed not only using the previous weather condition, but also coupled with data on population density and some other factors.

Cortez and Morais [2] used five different data mining techniques to predict the burnt area of forest fire using Support Vector Machines (SVM) and Random Forests. With four distinct features likes spatial, temporal, Fire Weather Index components and weather variables (such as temperature, relative humidity, rain and wind), it was found that the best configuration was reached using Support Vector Machine, which is capable of predicting the frequent burnt areas due to small fire.

A study to increase the Fuzzy C-means model intelligently using a flexible termination criteria for the clustering of forest fire was conducted by Illadis et al[5]. This approach enables the algorithm to be more flexible and human-like in an intelligent way. It also avoids possible infinite loops and unnecessary iterations.

Decision tree C4.5 algorithm is implemented to predict the location of the incident hotspots in Rokan Hilir district, Riau province, Indonesia [8]. The dataset consists of hotspot locations, human activity factors, and land cover types. The human activity factors include city center locations, road network and river network.

Safi et al [4] applied artificial neural networks to the real world problem of predicting forest fire, using back propagation learning algorithm. Yu et al [6] conducted a research investigating the nonlinear relationship between the size of a forest fire and meteorological variables (temperature, relative humidity, wind speed and rainfall) using two hybrid approaches. At first phase Self Organizing Map is used to cluster the data. Than, in second phase the clustered data were used as inputs for two different approaches, the back-propagation neural network and the rule generation approaches.

Sakr et al [10] applied a description and analysis of forest fire prediction methods based on Support Vector Machines to predict the fire hazard level of a day, where the algorithm depended on previous weather conditions. Moreover, in [7], Sakr et al try to reduced a set of weather parameters utilizing relative humidity and cumulative precipitation to estimate the risk of the output, to predict the occurrence of forest fire by comparing two artificial intelligence-based methods: Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

Based on the above existing researches, this paper proposes an approach to predict the size of forest fire using hybrid model, between Fuzzy C-Means (FCM) clustering technique and Back-Propagation Neural Network (BPNN) classification technique in processing meteorological and forest weather index data as input.

3 Fundamentals

3.1 Forest Fire

Forest fire as a kind of common natural disaster possibly makes a great danger to people living in the burnt forest as well as to wildlife. Such disaster may be caused by lightning, human negligence or arson that can burn thousands of square kilometers. According to Brown and Davis [11], there are three types of forest fire namely: *ground fire, surface fire* and *crown fire*.

3.2 Data Mining

Data mining can be seen as a process of discovering patterns in large volume of data having meaningful information [11]. The process must be automatic or (more usually) semi-automatic. Among several existing methods commonly applied in data mining, in this research, clustering and classification are chosen to be implemented. Clustering technique is used to cluster the size of forest burning size area having value greater than zero into four clusters, while classification technique is to determine which type of burning size that will probably occur based on meteorological data.

Fuzzy C-Means (FCM). is one of popular fuzzy clustering techniques that used for finding similarities in data and putting similar data into several groups, has been proposed by Dunn [12] in 1973 and then later modified by Bezdek [13] in 1981. It is an approach where the data points have their membership values with cluster center, to be updated iteratively. The detail explanation of FCM algorithm, could be seen at [14].

Back-Propagation Neural Network (BPNN). is classification technique highly dependent on the network structure and training process that has better learning rate [15]. The number of input layer nodes, hidden layer and output layer in BPNN will determine the structure of the network.

Back-propagation learning process requires a pair of input vectors and the target vectors. The output vector of each input vector will be compared with the target vector. This measurement is necessary in order to minimize the difference between the output vector and the target vector.

In BPNN it begins with the initialization of weights and thresholds at random. The weights are updated in each iteration to minimize the Mean Square Error (MSE) between the output vector and the target vector, where the detail information of BPNN was explain in [16].

3.3 Preprocessing

There are several steps in data mining preprocessing [17], such as data cleansing, data integration, data reduction and data transformation. In this research, data

transformation is used to normalize the data. Data normalization is useful for classification involving neural networks or distance measurements such as nearest neighbour classification and clustering. Beside that, it can affect to speed up the learning rate of BPNN for classification. In this research Min-max normalization is applied to perform a linear transformation on the original data [11], where the formula could be seen at (1). The data of eight variables or attribute used in this research will be transform in new range with min value is 0 and max value is 1.

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(New\_max_A - New\_min_A) + New\_min_A \qquad (1)$$

Where min_A is existing minimum value and $maxA$ is existing maximum values of an attribute A. vi is existing data value in attribute A that will be mapped to current data value v_i' in the new range [0 , 1] $[New\_min_A,\ New\_max_A]$.

4 Research Method

Fig.1 depicts the overall process in this research, describing the position the proposed hybrid model in predicting size of forest fire.

4.1 Data Collection

Data on forest fire are collected from the study by Cortez and Morais, available in the UCI machine learning repository [2] . The dataset contains 12 variable with their respective labels, forest fire weather index (FWI) components in Montesano Natural Park, a northeast region of Portugal. Weather observations are collected by Braganza 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 year 2000 until 2007. This research only select 8 variables to be considered: FFMC, DMC, DC, ISI, Temperature, RH, Wind and Rain.

4.2 Splitting the Dataset

In this steps, the dataset is split into two categories. The process of splitting data is conducted by selecting label dataset(Area). The label (Area) that have value zero, it means have not any total burn area size will be separate from label (Area) that have value more than zero as showed in Fig. 2. After that, all data with zero value will be categorized as data "No Burn Area". Otherwise, the label (Area) data with value more than zero will be cluster by FCM to categorized as data Light Burn or Heavy Burn.

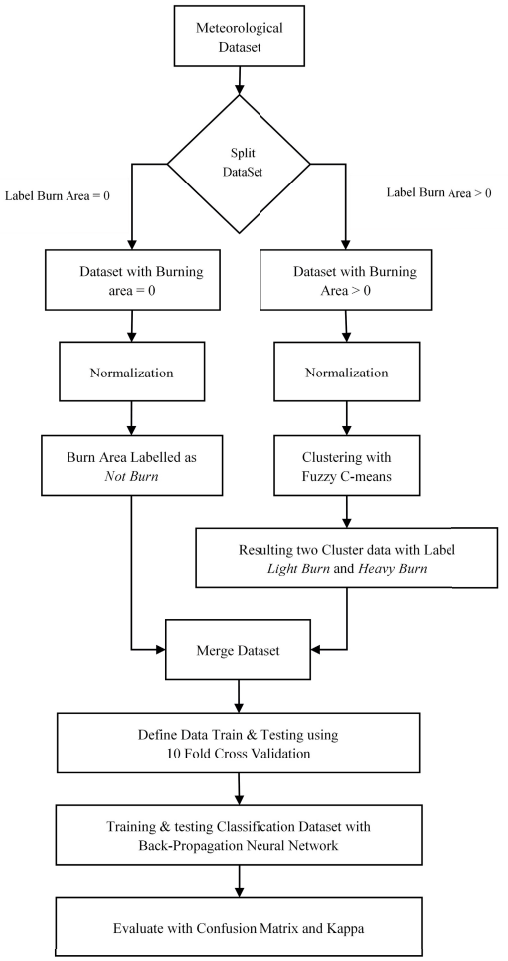
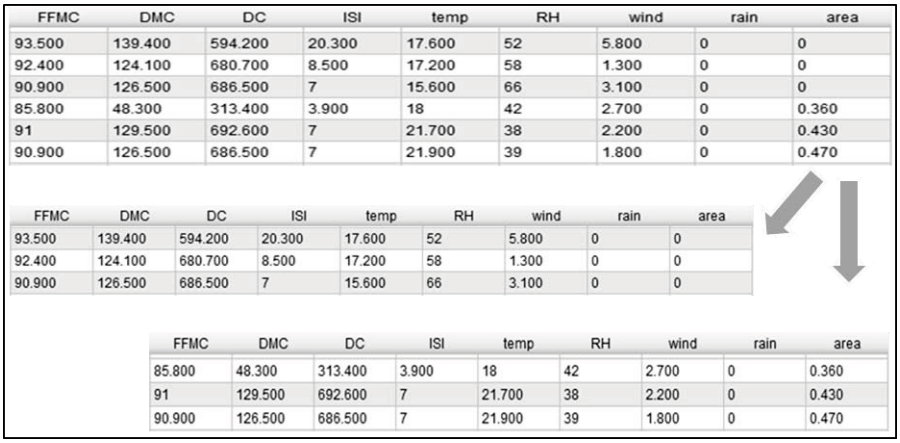


Fig. 1. Research Method Outline

4.3 Normalization

After splitting dataset into two categories between dataset that have value zero or more than zero in attribute label, we continue with the process of normalization. Normalization process in this research uses equation (1) with min max normalization. The normalization process, only transform 8 variables that will be used in clustering and classification process such as FFMC, DMC, DC, ISI, temperature, RH, wind, and rain. This process results in minimum and maximum values between $[0, 1]$ in dataset. The sample of process before and after normalization could be seen in Fig.3 and Fig.4.



FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
93.500	139.400	594.200	20.300	17.600	52	5.800	0	0
92.400	124.100	680.700	8.500	17.200	58	1.300	0	0
90.900	126.500	686.500	7	15.600	66	3.100	0	0
85.800	48.300	313.400	3.900	18	42	2.700	0	0.360
91	129.500	692.600	7	21.700	38	2.200	0	0.430
90.900	126.500	686.500	7	21.900	39	1.800	0	0.470

FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
93.500	139.400	594.200	20.300	17.600	52	5.800	0	0
92.400	124.100	680.700	8.500	17.200	58	1.300	0	0
90.900	126.500	686.500	7	15.600	66	3.100	0	0

FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
85.800	48.300	313.400	3.900	18	42	2.700	0	0.360
91	129.500	692.600	7	21.700	38	2.200	0	0.430
90.900	126.500	686.500	7	21.900	39	1.800	0	0.470

Fig. 2. Sample of Process Split Dataset

FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
93.100	157.300	666.700	13.500	21.700	40	0.400	0	2.470
93.100	157.300	666.700	13.500	26.800	25	3.100	0	0.680
93.100	157.300	666.700	13.500	24	36	3.100	0	0.240
93.100	157.300	666.700	13.500	22.100	37	3.600	0	0.210
91.900	109.200	565.500	8	21.400	38	2.700	0	1.520
91.600	138.100	621.700	6.300	18.900	41	3.100	0	10.340

Fig. 3. Sample Dataset before Normalization

FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0.960	0.538	0.771	0.595	0.627	0.259	0	0	2.470
0.960	0.538	0.771	0.595	0.791	0.074	0.300	0	0.680
0.960	0.538	0.771	0.595	0.701	0.210	0.300	0	0.240
0.960	0.538	0.771	0.595	0.640	0.222	0.356	0	0.210
0.945	0.373	0.651	0.352	0.617	0.235	0.256	0	1.520
0.941	0.472	0.717	0.278	0.537	0.272	0.300	0	10.340

Fig. 4. Sample Dataset after Normalization

4.4 Fuzzy C-Means Clustering Dataset

The process of categorizing dataset into two categorise of size of fire is done in this phase. Fuzzy C-Means here will cluster the data based on eight Meteorological variables. Since FCM is unsupervised method, it will automatically categorise the dataset into two categorise by default: cluster_0 (as Light Burn) and custer_1 (as Heavy Burn).

We observed several distance similarity measurements algorithm in FCM such as Correlation Similarity, Cosine Similarity, Dice Similarity, Inner Product Similarity, Jaccard Similarity, Overlap Similarity, Kernel Euclidian Distance,

Manhattan Distance, Chebychev Distance, Euclidean Distance, Canberra Distance, Dynamic Time Warping Distance to achieve the best performance of classification BPNN .

4.5 Merge Dataset

After clustering process has been done, the data that has been categorize as No Burn Area will be merged with the data that has been cluster by FCM. Therefore, after this process we will have the dataset that contain label No Burn Area, Light Burn, and Heavy Burn. As you can see in Fig.5.

FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0.965	0.479	0.692	0.362	0.475	0.435	0.645	0	no_burn_area
0.933	0.047	0.021	0.219	0.475	0.141	0.645	0	no_burn_area
0.862	0.113	0.077	0.320	0.476	0.148	0.500	0	light_burn
0.862	0.113	0.077	0.320	0.476	0.148	0.500	0	light_burn
0.722	1	1	0.146	0.476	0.642	0.500	0	heavy_burn
0.942	0.258	0.838	0.139	0.479	0.282	0.355	0	no_burn_area

Fig. 5. Sample of Merge Dataset

4.6 Back-Propagation Neural Network Architecture

The architecture of Back-Propagation Neural Network in this research uses only one hidden layer, where the learning rate has been fixed at $\beta = 0.3$ and the maximum number of iteration is $\alpha = 500$. The detail steps of BPNN could be seen at [16]. Fig.6 is showed the architecture of BPNN.

5 Result Evaluation and Discussion

5.1 Performance Measurement

After classification process, to assess the performance results of our proposed hybrid method for predicting forest burning size, we used confusion matrix [11] to measures accuracy of classifier can be calculated by equation (2) and Cohen's Kappa statistic measurement [18] to assess inter-rater reliability when observing categorical variables can be calculate by equation (3).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + TN + FP + FN} \tag{2}$$

$$Kappa = \frac{Observed\ Agreement - Expected\ Agreement}{1 - Expected\ Agreement} \tag{3}$$

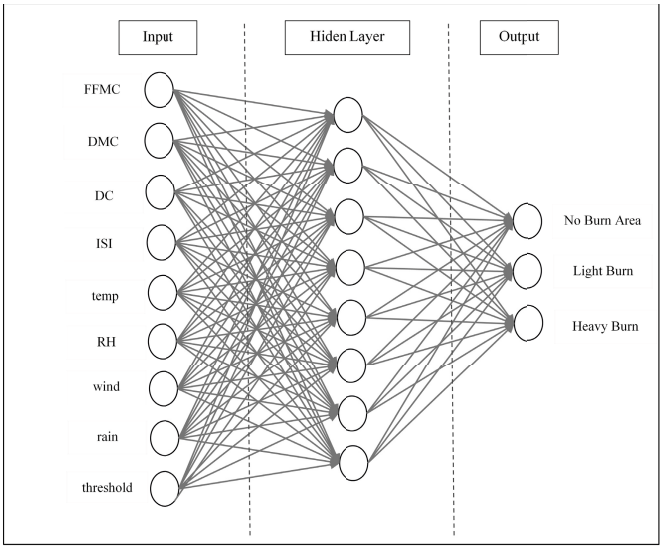


Fig. 6. Back-Propagation Neural Network Architecture

5.2 Experiment Result

This research used RapidMinner tools to conduct the experiment. All the process from normalization phase, clustering, classification until Validation and Evaluation were conducted in RapidMiner. Based on the results displayed in Table 1 and Table 2, it is shown performance of Back-Propagation neural network could achieve best results with accuracy around 97.50% and index of Cohens Kappa 0.961. The best performance of proposed model are gathered with combination

Table 1. Performance of Hybrid Model in Stratified Sampling

Type of Distance Similarity in FCM	Accuracy	Kappa
Correlation Similarity	95.74%	0.933
Cosine Similarity	**97.10%**	**0.954**
Dice Similarity	96.91%	0.945
Inner Product Similarity	96.71%	0.936
Jaccard Similarity	96.91%	0.945
Overlap Similarity	91.30%	0.861
Kernel Euclidian Distance	96.14%	0.938
Manhatan Distance	96.13%	0.935
Chebychev Distance	95.74%	0.930
Euclidean Distance	96.33%	0.941
Canberra Distance	93.42%	0.897
Dynamic Time Warping Distance	86.26%	0.780

Table 2. Performance of Hybrid Model in Shufled Sampling

Type of Distance Similarity in FCM	Accuracy	Kappa
Correlation Similarity	95.75%	0.932
Cosine Similarity	**97.50%**	**0.961**
Dice Similarity	97.30%	0.952
Inner Product Similarity	96.34%	0.927
Jaccard Similarity	97.30%	0.952
Overlap Similarity	89.56%	0.833
Kernel Euclidian Distance	96.92%	0.951
Manhatan Distance	96.90%	0.947
Chebychev Distance	96.91%	0.949
Euclidean Distance	96.92%	0.951
Canberra Distance	92.65%	0.884
Dynamic Time Warping Distance	84.56%	0.754

of BPNN with FCM that used Cosine Similarity. Besides that, to show the performance of BPNN, we also compare the performance of another classification method such as SVM, KNN, DCT, and Naive Bayes, that include with same clustering technique FCM in categorizing the dataset. The results comparison could be seen at Fig 7 and Fig.8.

The proposed Hybrid model for predicting the size of forest fire indicates a promising result. Compared with other methods such as SVM, K-NN and DCT Tree, the proposed method is still showing better performance, more over Naive Bayes and Random Forest have lowest performance classification with accuracy less than 74% and Cohens Kappa 0.54.

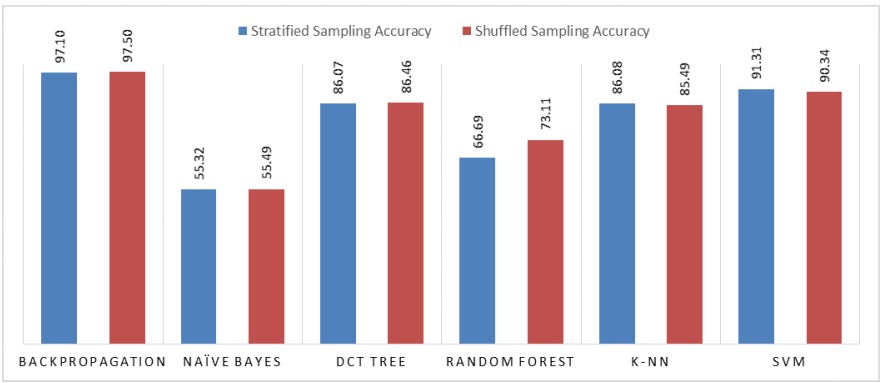


Fig. 7. Results of Accuracy Confusion Matrix Performance

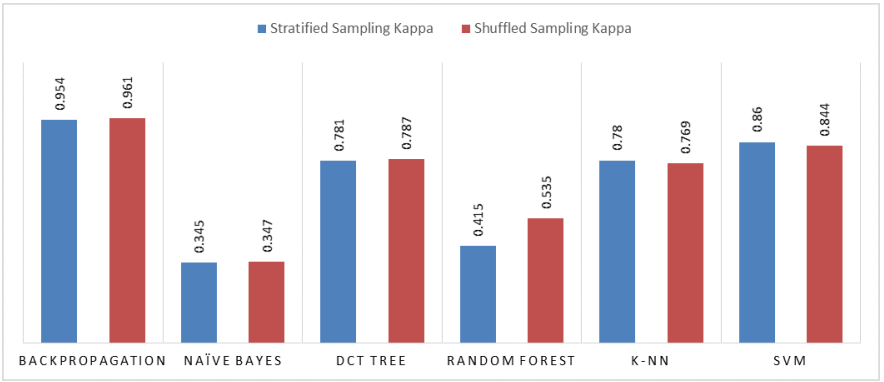


Fig. 8. Results of Cohen's Kappa Performance

The overall approach of experiment in this study is different to the existing work done by Cortez and Morais [2] that also used same dataset. However, in their study used twelve variables which our approach used eight variables. Besides that, they only evaluate pure prediction methods such as Neural Network, SVM, Naive Bayes, Multiple Regression and Decision Trees without combining cluster methods that provide burn area prediction in numerical results without categorizing the type of result forest burning size.

6 Conclusion

This research has proposed an alternative hybrid model capable of predicting the size of forest fire by combining Fuzzy C-Means and Back-Propagation Neural Network method. The model which incorporates meteorological and forest weather index variables (FFMC, DMC, DC, ISI, temperature, RH, wind and rain) has been shown to be successfully classify the level of burning into three categories: *No Burn Area, Light Burn* and *Heavy Burn*. The evaluation of the proposed model has showed promising results with accuracy of confusion matrix around 97.50% and Kappa 0.961. It is also found that cosine similarity method in FCM shows better performance than other similarity distance measuring algorithms under simulation. For the future work, the model will be implemented as web services and integrated with meteorological sensor to build early warning of forest fire prediction system.

References

1. Alonso-Betanzos, A., Fontenla-Romero, O., Guijarro-Berdinas, B., Hernndez-Pereira, E., Paz-Andrade, M.I., Jimenez, E., Legido, J.L., Carballas, T.: An intelligent system for forest fire risk prediction and fire fighting management in galicia. Expert Syst. Appl. 25(4), 545–554 (2003)

2. Cortez, P., Morais, A.: A data mining approach to predict forest fires using meteorological data. In: Neves, J., Santos, M.F., Machado, J. (eds.) EPIA 2007, pp. 512–523 (2007)
3. Elmas, C., Sonmez, Y.: A data fusion framework with novel hybrid algorithm for multi-agent decision support system for forest fire. Expert Syst. Appl. 38(8), 9225–9236 (2011)
4. Safi, Y., Bouroumi, A.: A neural network approach for predicting forest fires. In: 2011 International Conference on Multimedia Computing and Systems (ICMCS), pp. 1–5 (2011)
5. Iliadis, L., Vangeloudh, M., Spartalis, S.: An intelligent system employing an enhanced fuzzy c-means clustering model: Application in the case of forest fires. Computers and Electronics in Agriculture 70(2), 276–284 (2010); Special issue on Information and Communication Technologies in Bio and Earth Sciences
6. Yu, Y.P., Omar, R., Harrison, R.D., Sammathuria, M.K., Nik, A.R.: Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods. Journal of Computational Biology and Bioinformatics Research 3, 47–52 (2011)
7. Sakr, G.E., Elhajj, I.H., Mitri, G.: Efficient forest fire occurrence prediction for developing countries using two weather parameters. Engineering Applications of Artificial Intelligence 24(5), 888–894 (2011)
8. Sitanggang, I., Ismail, M.: Hotspot occurrences classification using decision tree method: Case study in the rokan hilir, riau province, indonesia. In: 2010 8th International Conference on ICT and Knowledge Engineering, pp. 46–50 (2010)
9. Satoh, K., Weiguo, S., Yang, K.T.: A study of forest fire danger prediction system in japan. In: Proceedings of the 15th International Workshop on Database and Expert Systems Applications, pp. 598–602 (2004)
10. Sakr, G., Elhajj, I., Mitri, G., Wejinya, U.: Artificial intelligence for forest fire prediction. In: 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp. 1311–1316 (2010)
11. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2005)
12. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters (1973)
13. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers (1981)
14. Chattopadhyay, S., Pratihar, D.K., Sarkar, S.C.D.: A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms. Computing and Informatics 30(4), 701–720 (2011)
15. Singh, D., Dutta, M., Singh, S.H.: Neural network based handwritten hindi character recognition system. In: Shyamasundar, R.K. (ed.) Bangalore Compute Conf., p. 15. ACM (2009)
16. Eleyan, A., Demirel, H.: PCA and LDA based Neural Networks for Human Face Recognition, Number June, Viena, Austria (2007)
17. Han, J., Kamber, M.: Data mining: concepts and techniques. Kaufmann, San Francisco (2005)
18. Byrt, T., Bishop, J., Carlin, J.B.: Bias, prevalence and kappa. Journal of Clinical Epidemiology 46(5), 423–429 (1993)