Forest Type Classification: A Hybrid NN-GA Model Based Approach

Conference Paper · January 2016

Forest Type Classification: A Hybrid NN-GA Model Based Approach

Sankhadeep Chatterjee, Subhodeep Ghosh, Subham Dawn, Sirshendu Hore and Nilanjan Dey

Abstract Recent researches have used geographically weighted variables calculated for two tree species, *Cryptomeria japonica* (Sugi, or Japanese Cedar) and *Chamaecyparis obtusa* (Hinoki, or Japanese Cypress) to classify the two species and one mixed forest class. In machine learning context it has been found to be difficult to predict that a pixel belongs to a specific class in a heterogeneous landscape image, especially in forest images, as ground features of nearly located pixel of different classes have very similar spectral characteristics. In the present work the authors have proposed a GA trained Neural Network classifier to tackle the task. The local search based traditional weight optimization algorithms may get trapped in local optima and may be poor in training the network. NN trained with GA (NN-GA) overcomes the problem by gradually optimizing the input weight vector of the NN. The performance of NN-GA has been compared with NN, SVM and Random Forest classifiers in terms of performance measures like accuracy,

S. Chatterjee (✉)
Department of Computer Science & Engineering,
University of Calcutta, Kolkata, India
e-mail: sankha3531@gmail.com

S. Ghosh · S. Dawn
Department of Computer Science & Engineering,
Academy of Technology, Aedconagar, Hooghly 712121, India
e-mail: subhodeepghosh7@gmail.com

S. Dawn
e-mail: knightdawnxp@gmail.com

S. Hore
Department of Computer Science & Engineering,
Hooghly Engineering & Technology College, Hooghly, India
e-mail: shirshendu.hore@hetc.ac.in

N. Dey
Department of Computer Science & Engineering,
BCET, Durgapur, India
e-mail: neelanjandey@gmail.com

S.C. Satapathy et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 435,
DOI 10.1007/978-81-322-2757-1_23

precision, recall, F-Measure and Kappa Statistic. The results have been found to be satisfactory and a reasonable improvement has been made over the existing performances in the literature by using NN-GA.

Keywords Image classification · Neural network · Genetic algorithm · Random forest · SVM

1 Introduction

Image classification has drawn attention of the remote-sensing researchers due to the complexities and challenges in the context of machine learning and soft computing. Most of the remote sensing research is mainly focused on reducing the classification error due to several natural, environmental, and other effects. Several research works have been found in this regard [1–8]. The research works have proposed image classification using several methods and have tried to reduce the classification error to a greater extent. In remote sensing context the image classification can formally be defined as to classify pixels of a given image into different regions, each of which is basically a landcover type. In satellite image, each pixel represents a specific landcover area though it is highly likely that it may belong to more than one land cover type. With increasing number of instances the problem becomes more severe as the number of such pixel increases. Thus, a large amount of uncertainty may get involved during the classification task. Several Unsupervised learning methods have been proposed to handle this problem efficiently [9–12]. Unsupervised techniques like Fuzzy c-means [13], split and merge [14] and ANN [15–20] based methods [21] have been applied for the task. The problem of uncertainty during classification becomes more challenging in the Forest images as the level of variation is higher in a very small geographic area. Thus, using traditional machine learning techniques it becomes quite challenging to achieve reasonable amount of classification accuracy. The problem has been overcome by applying a SVM [22] based model with assistance of geographically weighted variables [23]. The authors have shown slight improvement over the existing methods (without using geographically weighted variables).

In the present work the authors have proposed a GA trained Neural Network model to tackle the task. The input weight vector of the NN has been gradually optimized using GA to increase the performance of NN. The application of GA in training NN has already been found to be quite satisfactory in several real life applications [24]. The proposed model has been compared with two well-known classifiers SVM and Random Forest [25, 26]. The performances of the proposed models and the other models have been measured using accuracy, precision, recall, F-Measure and Kappa Statistic.

2 Proposed NN-GA Model

GA was proposed by Holland (1975) to implement the theory of natural selection to solve optimization problems. The GA starts solving a problem by using a set of initial solutions. And it continuously applies crossover and mutations on the solutions to produce better offspring. The survival of any offspring depends on the fitness which is decided by the problem definition of the problem being solved. GA has lesser chance of getting trapped into local optima. Thus, it can be better choice than the traditional methods. The method of applying GA can be summarized as follows;

1. **Generation of initial population** 'N' numbers of chromosomes are randomly generated. Each chromosome is actually an array of random real weight values, biologically genes; they vary in between '0' to '1'.
2. **Calculating fitness values** A fitness function has to be defined, using it the fitness of each individual solution (chromosome) has to be evaluated. RMSE of NN training is used as the fitness function.
3. **Selection** The smaller the RMSE, higher is the chance of getting selected for the next generation. $RMSE_i$ denotes the fitness function value of ith solution. The selection procedure works as follows:

   3.1. $RMSE_i$ is calculated for each solution in population.
   3.2. All $RMSE_i$ are aggregated or averaged together to find $RMSE_{ave}$
   3.3. A random value ($RMSE_r$) is selected from predefined closed interval $[0, RMSE_{ave}]$
   3.4. For all solutions $RMSE_r - RMSE_i$ is calculated and if the result of the subtraction is less than or equal to '0' the ith individual is selected.
   3.5. The process goes on until the number of solutions selected for next generation (mating pool) is equal to the number of solutions in the population initially.

4. **Cross-over** The selected chromosomes take part in cross-over where the after selecting cross-over points on the chromosome the genes at the right of that point for both the chromosomes taking part get exchanged. And it creates two new individual.
5. **Mutation** Genes of same chromosome take part in this phase. Genes from randomly selected position are swapped to create new individual solution.
6. **Termination condition** Finally the termination condition is checked. In the present work number of generation has been selected as terminating condition. When the user given number of generation is reached the best possible individual is selected as the optimized weight vector, otherwise it starts from step 2 again.

Figure 2 depicts the flowchart of NN trained with GA. The GA block in the flowchart is separately depicted in Fig. 1.

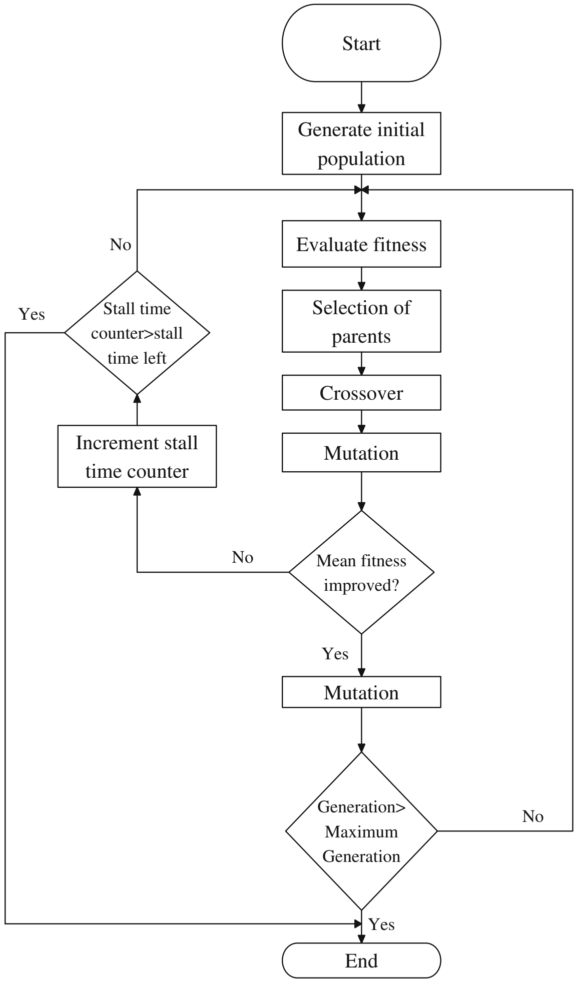

Fig. 1 The genetic algorithm which has been followed to optimize the input weight vector

3 Dataset Description

The dataset [23] used in the current work includes information of forested area in Ibaraki Prefecture, Japan (36° 57_N, 140° 38_E), approximately 13 km × 12 km. The landscape consists mainly of *Chamaecyparis obtusa* (Hinoki, or Japanese Cypress) planted forest ('h' class), *Cryptomeria japonica* (Sugi, or Japanese Cedar) planted forest ('s' class) and mixed natural forest, along with other land cover types (agriculture, roads, buildings, etc.) [23] which have been mapped based on the spectral characteristics at visible-to-near infrared wavelengths of the satellite images taken by ASTER satellite imagery. There are all total 27 attributes which are; spectral information in the green, red, and near infrared wavelengths for three dates (Sept. 26, 2010; March 19, 2011; May 08, 2011 (Total 9 attributes). Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 's' class (Total 9 attributes) and Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 'h' class (Total 9 attributes).

4 Experimental Methodology

The experiment is conducted on the dataset [23] obtained from UCI Machine Learning Repository. The experiments are performed by using Support Vector Machine (LibSVM) [27], Random Forest and real coded NN, NN-GA classifiers. For NN scaled conjugate gradient algorithm [28] has been used as the learning algorithm. The algorithm is well known and benchmarked against traditional back-propagation and other algorithms. The basic flow of experiment opted in the present work is as follows

1. **Preprocessing** The following preprocessing is done on the dataset before the classification
   (a) **Data Cleaning**—The data might contain missing values or noise. It is important to remove noise and fill up empty entries by suitable data by means of statistical analysis.
   (b) **Data Normalization**—the needs to be normalized before classification task is carried on to reduce distance between attribute values. It is generally achieved by keeping the value range in between −1 to +1.
2. After preprocessing the datasets are divided into two parts. One of which is used as training dataset and the other as testing dataset. In the present work two third (70 %) of the data is used as training data and rest (30 %) as testing data.
3. In the training phase the training dataset is supplied to different algorithms respectively to build the required classification model.
4. In the testing phase the classification models obtained from the training phase is employed to test the accuracy of the model.

To measure the performance and to compare the performances we use several statistical performance measures like accuracy, precision, recall, Kappa statistic [29], True positive rate (TP rate), and F-measure. The performance measuring parameters are calculated from the confusion matrix [30] which is a tabular representation that provides visualization of the performance of a classification algorithm. The objective function of Genetic algorithm (RMSE) is defined as follows; RMSE [31] of a classifier prediction with respect to the computed variable v_{c_k} is determined as thesquare root of the mean-squared error and is given by:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}\left(v_{d_k} - v_{c_k}\right)^2}{n}}$$

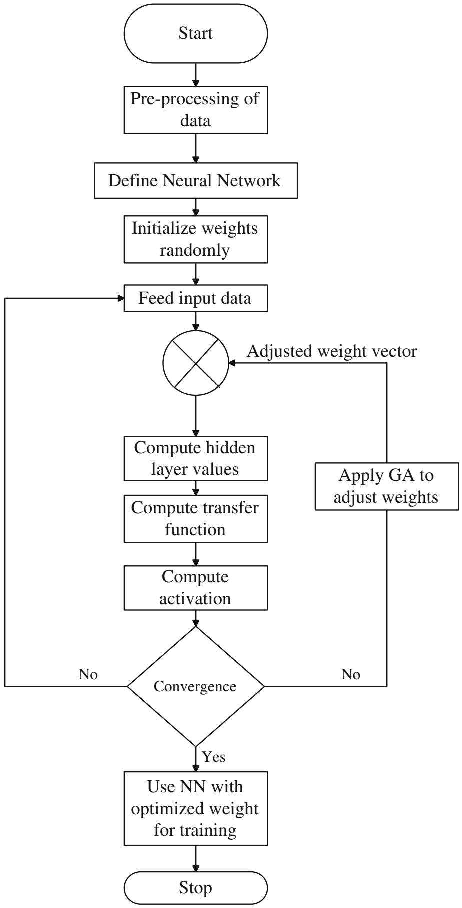

Fig. 2 Flowchart of NN training using genetic algorithm depicted in Fig. 3

where, v_{d_k} denotes the originally observed value of kth object and v_{c_k} denotes the predicted value by the classifier. The genetic algorithm based optimization of input weight vector has been implemented by following the Fig. 2. The different parameters used as inputs are as follows (Table 1).

5 Results and Discussion

The experiments have been carried out on the dataset described in Sect. 3. The experimental methodology has been described in Sect. 4. Table 2 reports the experimental results for Neural Network, Support Vector Machine (LibSVM), Random Forest and NN-GA classifiers. The experimental results suggest that the performance of Neural Network (trained with scaled conjugate gradient descent

Table 1 Genetic algorithm setup for input weight vector optimization

Maximum number of generation	1000
Population size	500
Crossover probability	0.2
Mutation	Gaussian
Crossover	Single point crossover
Selection	Roulette
Stall Time Limit	75 s

Table 2 Performance measures of different algorithms

	NN	SVM	Random forest	NN-GA
Accuracy	85.35	85.99	82.17	95.54
Precision	86.31	84.67	86.37	94.66
Recall	82.32	82.78	78.83	95.76
F-Measure	84.27	83.71	82.59	95.21
Kappa statistic	0.793	0.804	0.746	0.938

algorithm) is moderate with an accuracy of 85.35 %, precision 86.31 %, recall 82.32 %, F-Measure 84.27 % and Kappa Statistic 0.793 while the SVM (LibSVM) has performed almost same as the NN but with negligible. The experimental results suggest that Random Forest (which is considered to be one of the best tree based classifiers) may not be suitable for the classification of image classification of Forest or other areas which involves sufficient amount of uncertainty during classification. Though, the precision (86.37 %) of the classifier is better than NN and SVM. In the column four of Table 2 the experimental result of the proposed model has been shown. The objective of the GA was to reduce the RMSE (as described in Sect. 4). The model has performed significantly well than all the other classifiers in this study with an accuracy of 95.54 %, precision 94.66 %, recall 95.67 %, F-Measure 95.21 % and Kappa Statistic 0.938. Figure 3 depicts the different performance measures for the classifiers under consideration. The comparative analysis has revealed that the NN-GA is superior not only in terms of accuracy but also in terms

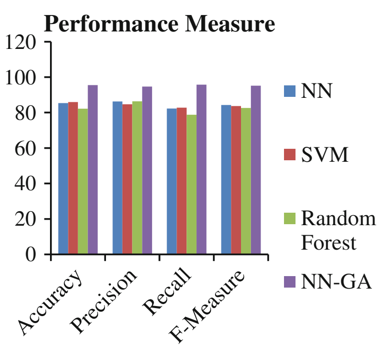

Fig. 3 Comparison of performance measure of NN, SVM, Random Forest and NN-GA classifiers

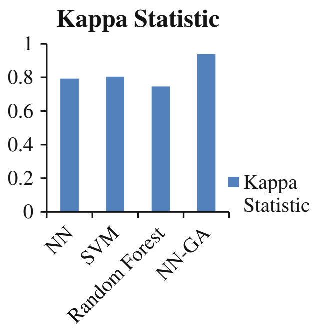

Fig. 4 Comparison of kappa statistic of NN, SVM, random forest and NN-GA classifiers

of precision, recall and F-Measure. Figure 4 depicts the same for Kappa Statistic. The plot establishes the superiority of NN-GA once again.

6 Conclusion

The present work has proposed a Genetic algorithm trained Neural Network model to classify two forest type along with one mixed class of forest of Japanese Cedar Japanese Cypress. The experimental results have suggested the superiority of NN-GA over the NN (trained with scaled conjugate gradient algorithm) for classification of pixels in Forest images. The study has also established that the NN-GA is a better classifier than support vector machines for this task. A satisfactory amount of improvement has been found over the existing work in literature. Besides the previous works have compared the performances of algorithms mainly based on accuracy which is not a good metric for performance measure as, accuracy of an algorithm varies greatly if number of instances varies in different classes. The present work have analyzed the performances of algorithms in terms of several performance measuring parameters like accuracy, precision, recall, F-Measure and Kappa Statistic to provide a vivid picture of the exact performance of the algorithms and to have a fair comparison of algorithms.

References

1. Gong, P. and Howarth, P.J., 1992, Frequency-based contextual classification and gray-level vector reduction for land-use identification. Photogrammetric Engineering and Remote Sensing, 58, pp. 423–437.
2. Kontoes, C., Wilkinson, G.G., Burrill, A., Goffredo, S. and Megier, J., 1993, An experimental system for the integration of GIS data in knowledge-based image analysis for remote sensing of agriculture. International Journal of Geographical Information Systems, 7, pp. 247–262.

3. Foody, G.M., 1996, Approaches for the production and evaluation of fuzzy land cover classification from remotely-sensed data. International Journal of Remote Sensing, 17, pp. 1317–1340.
4. San Miguel-Ayanz, J. and Biging, G.S., 1997, Comparison of single-stage and multi-stage classification approaches for cover type mapping with TM and SPOT data. Remote Sensing of Environment, 59, pp. 92–104.
5. Stuckens, J., Coppin, P.R. and Bauer, M.E., 2000, Integrating contextual information with per-pixel classification for improved land cover classification. Remote Sensing of Environment, 71, pp. 282–296.
6. Franklin, S.E., Peddle, D.R., Dechka, J.A. and Stenhouse, G.B., 2002, Evidential reasoning with Landsat TM, DEM and GIS data for land cover classification in support of grizzly bear habitat mapping. International Journal of Remote Sensing, 23, pp. 4633–4652.
7. Pal, M. and Mather, P.M., 2003, An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86, pp. 554–565.
8. Gallego, F.J., 2004, Remote sensing and land cover area estimation. International Journal of Remote Sensing, 25, pp. 3019–3047.
9. Bandyopadhyay, S., Maulik, U.: Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification, Pattern Recognition, 35(2), 2002, 1197–1208.
10. Bandyopadhyay, S., Maulik, U., Mukhopadhyay, A.: Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery, IEEE Transactions on Geoscience and Remote Sensing, 45(5), 2007, 1506–1511.
11. Bandyopadhyay, S., Pal, S. K.: Pixel Classification Using Variable String Genetic Algorithms with Chromosome Differentiation, IEEE Transactions on Geoscience and Remote Sensing, 39 (2), 2001, 303– 308.
12. Maulik, U., Bandyopadhyay, S.: Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification, IEEE Transactions on Geoscience and Remote Sensing, 41 (5), 2003, 1075– 1081.
13. Cannon, R. L., Dave, R., Bezdek, J. C., Trivedi, M.: Segmentation of a Thematic Mapper Image using Fuzzy c-means Clustering Algorithm, IEEE Transactions on Geoscience and Remote Sensing, 24, 1986, 400– 408.
14. Laprade, R. H.: Split-and-merge Segmentation of Aerial Photographs, Computer Vision Graphics and Image Processing, 48, 1988, 77–86.
15. Hecht-Nielsen, R., Neurocomputing. Addison-Wesley, Reading, MA, 1990.
16. Schalkoff, R.J., Artificial Neural Networks. McGraw-Hill, New York, 1997.
17. Jain, A.K., Mao, J., Mohiuddin, K.M., Artificial neural networks: a tutorial. Comput. IEEE March, pp. 31–44, 1996.
18. K.J. Hunt, R. Haas, R. Murray-Smith, "Extending the functional equivalence of radial basis function networks and fuzzy inference systems", IEEE Trans Neural Networks Vol.7, No.3, pp.776–781, 1996.
19. D.S. Broomhead, L. David, "Radial basis functions, multivariable functional interpolation and adaptive networks", In Technical report Royal Signals and Radar Establishment, No. 4148, p. 1–34,1998.
20. J. Han, M. Kamber, "Data mining: concepts and Techniques", 2nd ed. Morgan and Kaufmann, pp. 285–378, 2005.
21. Baraldi, A., Parmiggiani, F.: A Neural Network for Unsupervised Categorization of Multivalued Input Pattern: An Application to Satellite Image Clustering, IEEE Transactions on Geoscience and Remote Sensing, 33, 1995, 305–316.
22. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273–297.
23. Johnson, B., Tateishi, R., Xie, Z. (2012). Using geographically weighted variables for image classification. Remote Sensing Letters, 3(6), 491–499.

24. S. Chatterjee, R. Chakraborty, S. Hore: “A Quality Prediction method of Weight Lifting Activity”. Proceedings of the Michael Faraday IET International Summit - 2015 (MFIIS-2015), An IET International Conference On September 12 – 13, 2015 in Kolkata, India, (in press).
25. V. Y., Kulkarni, and P. K. Sinha, “Random forest classifiers: a survey and future research directions”, Int. J. Adv. Comput, Vol.36, No. 1, pp. 1144–1153, 2013.
26. L. Brieman, “Random Forests”, Machine Learning, Vol. 45, pp. 5–32, 2001.
27. Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.
28. Møller, Martin Fodslette. “A scaled conjugate gradient algorithm for fast supervised learning.” Neural networks 6, no. 4 (1993): 525–533.
29. Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguist, Cambridge, MA, USA: MIT Press, 22(2):249–54, 1996.
30. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote SensEnviron, 62(1):77–89, 1997.
31. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: empirical comparisons. Int. J Forecast, 8:69–80, 1992.