Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 52 (2015) 1192 – 1197

Procedia Computer Science

The 2nd International Workshop on Machine Learning and Data Mining for Sensor Networks, MLDM-SN 2015

A Fuzzy Decision Tree for Processing Satellite Images and Landsat Data

Feras Al-Obeidat[a,d,*], Ahmad T. Al-Taani[b], Nabil Belacel[c], Leo Feltrin[d], Neil Banerjee[d]

[a]*IBM, Research and Development Center, Markham-Toronto, Ontario, L3R 9Z7, Canada*
[b]*Department of Computer Science, Yarmouk University, Irbid, Jordan*
[c]*National Research Council of Canada, Moncton, Canada*
[d]*Department of Earth Sciences, Western University, London, Ontario, N6A 5B7 Canada*

Abstract

Satellite and airborne images, including Landsat, ASTER, and Hyperspectral data, are widely used in remote sensing and Geographic Information Systems (GIS) to understand natural earth related processes, climate change, and anthropogenic activity. The nature of this type of data is usually multi or hyperspectral with individual spectral bands stored in raster file structures of large size and global coverage. The elevated number of bands (on the order of 200 to 250 bands) requires data processing algorithms capable of extracting information content, removing redundancy. Conventional statistical methods have been devised to reduce dimensionality however they lack specific processing to handle data diversity. Hence, in this paper we propose a new data analytic technique to classify these complex multidimensional data cubes. Here, we use a well-known database consisting of multi-spectral values of pixels from satellite images, where the classification is associated with the central pixel in each neighborhood. The goal of our proposed approach is to predict this classification based on the given multi-spectral values. To solve this classification problem, we propose an improved decision tree (DT) algorithm based on a fuzzy approach. More particularly, we introduce a new hybrid classification algorithm that utilizes the conventional decision tree algorithm enhanced with the fuzzy approach. We propose an improved data classification algorithm that utilizes the best of a decision tree and multi-criteria classification. To investigate and evaluate the performance of our proposed method against other DT classifiers, a comparative and analytical study is conducted on well-known Landsat data.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Decision Tree; Fuzzy Classification; PROAFTN; Landsat; Remote Sensing; Satellite Images.

1. Introduction

The classification problem in data mining and machine learning paradigms is a well-known approach for decision making and prediction. A variety of research disciplines such as statistics, economics, Multiple Criteria Decision Analysis (MCDA)[1], and artificial intelligence/machine learning have addressed this generalized problem. The focus

* Feras Al-Obeidat. Tel.: +1-519-661-2111 ext 82253
E-mail address: feras@ca.ibm.com

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2015.05.157

of this study is to introduce a new classification algorithm from the two main disciplines machine learning and MCDA. This is a purely numerical methodology that will be implemented to predict the class of a pixel of the given population of multi-spectral values of Landsat Multi-Spectral Scanner image data. The Landsat data for this study was generated by the Australian Centre for Remote Sensing. The name of data is Statlog (Landsat Satellite) Data Set and is available at UCI machine learning website[2]. Some previous studies used decision tree and classification and remote sensing problem to process Landsat data including[3], [4] and [5]. Yet, no study has proposed this methodology to classify Landsat dataset.

The proposed method is a hybrid classifier that combines the two classification methods selected from a different classification perspective: the MCDA method PROAFTN (Belacel 1999)[6], and the Decision Tree (DT), which is based on a machine learning methodology[7].

Decision tree learning is a widely used method in data mining and machine learning. The strength of a decision tree can be summarized as: DT can generate understandable decision rules and run at good computational speed. On the other hand, PROAFTN has interesting characteristics, including: generating understandable rules and uses fuzzy membership degree which gives detailed information on assigning an object to a class. This advantage of PROAFTN compensates the loss of information of using strict intervals as in the case of decision trees.

However, despite the strengths of DT and PROAFTN, both methods have some limitations. For example, DT generates crisp intervals when evaluating an instance of the class. The problem of this approach is that we have strict rules for assigning objects to a class. Hence, there is no marginal area that have values between yes and no. PROAFTN has limitations too; for example to build the classification model for PROAFTN, several parameters (e.g., intervals and weights) have to be defined a priori to build the classification model. Usually, this approach requires extensive efforts to determine these intervals and might be time consuming.

To overcome the limitations of the two methods (DT and PROAFTN), in this study we combine and utilize the best of these two approaches to solve the classification problem. Hence, we will have a new hybrid classification method that utilizes the benefits of decision trees and the MCDA method PROAFTN and resolve shortcomings. Our overriding goal is to have a new efficient data mining/classification method in terms of accuracy and interpretability.

The paper is organized as follows. Section 2 briefly overview the Decision Tree and PROAFTN methodology. Section 3 explains the proposed methodology. In Section 4 experimentation and computational results are reported. Finally, conclusions and future work are presented in Section 5.

2. Decision Tree and PROAFTN

2.1. Decision Tree

The most well-known algorithm in machine learning literature used for building DT is the C4.5, which is an extension of Quinlan's earlier ID3 algorithm[8] ,[9]. The classification model in DT is constructed from the training samples based on recursive computation, *e.g.,* entropy, information gain or gain ratio used in C4.5 for each attribute in the available/remaining data. An entropy measure of a set of objects is calculated as follows:

$$H(N) = -\sum_{c\in C} p(c)\, log_2\, p(c) \tag{1}$$

where N represents the dataset; $p(c)$ is the proportion of instances in the dataset of class c, and C represents the set of classes.

The information gain can be calculated as:

$$IG(A) = H(N) - \sum_{t\in T} p(t)H(t) \tag{2}$$

Where T represents the splitted subsets created from N and $H(t)$ is the entropy of the subset t. DT usually stops the learning process when all leaves are purely classified, where in some cases the pure nodes could be reached on subsets of attributes; or when no more attributes or instances are remaining for further partitioning.

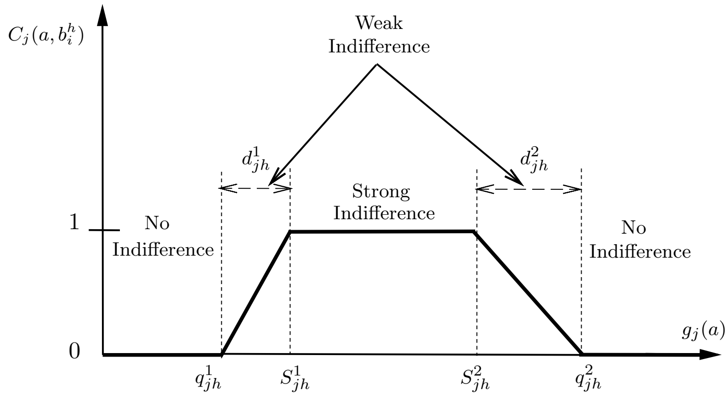

Fig. 1. Graphical representation of the computation of $C_j(a, b_i^h)$

2.2. PROAFTN

PROAFTN is fuzzy method that belongs to the class of supervised learning to solve classification problems and has been applied to the resolution of many real-world practical problems [10, 11]. The following subsections describe the required parameters, the classification methodology, and the procedure used by PROAFTN.

2.2.1. Initialization

From a set of N objects known as a training set, consider a as an object which requires to be classified; assume this object a is described by a set of m attributes $\{g_1, g_2, ..., g_m\}$ and C classes $\{c^1, c^2, ..., c^C\}$. Given an object a described by the score of m attributes, the different steps of the procedure are as follows:

For each class c^h, we determine a set of L_h prototypes. For each prototype b_i^h and each attribute g_j, an interval $[S_j^1(b_i^h), S_j^2(b_i^h)]$ is defined where $S_j^2(b_i^h) \geqslant S_j^1(b_i^h)$.

The fuzzy approach is introduced through thresholds $d_j^1(b_i^h)$ and $d_j^2(b_i^h)$ to define the pessimistic interval $[S_j^1(b_i^h), S_j^2(b_i^h)]$ and the optimistic interval $[S_j^1(b_i^h) - d_j^1(b_i^h), S_j^2(b_i^h) + d_j^2(b_i^h)]$. The ($S^1$ to S^2) will contain the most true-like values.

2.2.2. Computing the Fuzzy Indifference Relation

After obtaining intervals, the following stage is to calculate the fuzzy membership degree between an object a and the prototype b_i^h, which mathematically presented as $I(a, b_i^h)$. The calculation of $I(a, b_i^h)$ is identified by:

$$I(a, b_i^h) = \sum_{j=1}^{m} w_j^h C_j(a, b_i^h) \tag{3}$$

where w_j^h is the weight that measures the importance of a relevant attribute g_j of a specific class c^h:

$$w_j \in [0, 1], \sum_{j=1}^{m} w_j^h = 1$$

$C_j(a, b_i^h)$ is the degree that measures the closeness of the object to the prototype as graphically presented in Fig. 1.

$$I(a, b_i^h) = \sum_{j=1}^{m} w_j^h C_j(a, b_i^h) \tag{4}$$

2.2.3. Evaluation of the Membership Degree

The membership degree between the object a and the class c^h is calculated based on the indifference degree between a and its nearest neighbor in B^h:

$$d(a, c^h) = \max\{I(a, b^h_1), I(a, b^h_2), ..., I(a, b^h_{L_h})\} \tag{5}$$

2.2.4. Assignment of an Object to the Class

The last step is to assign the object a to the right class c^h:

$$a \in c^h \Leftrightarrow d(a, c^h) = \max\{d(a, c^i)/i \in \{1, ..., C\}\} \tag{6}$$

3. Classification Methodology

This section presents the methodology used to construct the classification model using the combined decision tree and fuzzy approach adopted by PROAFTN. As discussed earlier, to apply PROAFTN we need to obtain several parameters $\{S^1_j(b^h_i), S^2_j(b^h_i), d^1_j(b^h_i), d^2_j(b^h_i)\}$ for each attribute in each class. In this study, the DT is performed as a prior step to obtain these parameters through running the supervised discretization techniques introduced by Fayyad and Irani [12]. The discretization step is based on the calculation of entropy and information as described in equations: 1 and 2. To determine the values for $d^1_j(b^h_i)$ and $d^2_j(b^h_i)$, an adjustment is applied on $S^1_j(b^h_i)$ and $S^2_j(b^h_i)$ to allow more flexibility in assigning patterns to the closest classes: $d^1_j(b^h_i) = S^1_j(b^h_i) - \beta S^1_j(b^h_i)$ and $d^2_j(b^h_i) = S^2_j(b^h_i) - \beta S^2_j(b^h_i)$ where $\beta \in [0, 1]$.

The induction approach is given in Algorithm 1. The branches are selected recursively using the decision tree algorithm to compose the prototypes from N based on the calculation of entropy and information gain. During learning, once the sub-attribute value of the best information gain is found, it's added to the prototype b_i. The learning proceeds until all attributes and sub-attributes are examined. Finally, from the generated decision tree the PROAFTN prototypes are constructed for each class. Finally, these prototypes are submitted to PROAFTN to classify new instances.

Algorithm 1 Building the Classification Model

```
1: i : prototype's index; h : class index;
2: g : attribute's index; j : sub-attribute's index
3: Calculate the entropy of every attribute g using the data set N
4: Split the set N into subsets using the attribute for which entropy is minimum (information gain is maximum)
5: Start with the node that has best information gain and make a decision tree node containing that attribute
6: Recursively proceed on subsets g_j using remaining attributes
7: if The the value of best information gain is found then
8:     Choose intervals for prototype b^h_i for class c^h
9: else
10:    Assign w^h_j = 0 for this value
11:    Go next sub-attribute
12: end if
13: Proceed with learning by recursively visiting step 5 for the remaining dataset N'
```

4. Experimental Work

4.1. Dataset Description

The datasets used in our experimental work are available on the public domain of the University of California at Irvine (UCI) Machine Learning Repository database [2]. The dataset consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification is associated with the central pixel in each neighborhood.

Table 1. Comparisons of accuracy versus time complexity for different approaches.

Approach	Accuracy (%)	Time (seconds)
Proposed	88.29	10
ID3	82.00	11
C4.5	85.71	9

Table 2. Decision Tree $C4.5$ Performance (approx. to three digits).

Categories	Count	TPRate	FPRate	Precision	Recall	FMeasure	AUC
1. red soil	1533	0.952	0.022	0.941	0.952	0.947	0.967
2. cotton crop	703	0.946	0.008	0.947	0.946	0.947	0.969
3. grey soil	1358	0.884	0.037	0.884	0.884	0.884	0.925
4. damp grey soil	626	0.554	0.054	0.555	0.554	0.555	0.752
5. soil with vegetation stubble	707	0.792	0.022	0.835	0.792	0.813	0.882
6. very damp grey soil	1508	0.851	0.058	0.84	0.851	0.845	0.9

The aim is to predict this classification based on the given multi-spectral values. In the sample database, the class of a pixel is coded as a number. This data is widely used by most researchers to evaluate and benchmark their work[2].

The dataset consists of 6435 instances; each instance is described with 36 attributes (4 spectral bands x 9 pixels in neighborhood) categorized over six groups which identify the types: red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, or very damp grey soil. All attributes are numeric and their values range from 0 to 255.

4.2. Experiments and Results

The proposed learning methodology was implemented in Java and run in a Linux machine. We applied the algorithm to the discussed Landsat data. We conducted a comparative study with C4.5 and ID3 algorithms, which were implemented in Weka[13]. We used the default settings using the stratified 10-fold cross-validation.

As presented in Table 1, the overall accuracy generated by the proposed method is 88.29%. In contrast, the decision trees C4.5 and ID3 have achieved a lower classification accuracy of 85.72% and 82.00%, respectively. Also it is worth noting that the time for building the PROAFTN classification model was reasonable compared with the classifiers. One can clearly notice that when using the hybrid approach - decision tree and PROAFTN - the overall accuracy has improved compared with using a standalone ID3 and C4.5. Hence, the proposed approach has significantly improved in terms of efficiency and classification accuracy of 88.29%.

For further analytical study, other performance measures were used. These measures are: recall, precision, and F_1 measure. Where $TPRate$ refers to true positive rare and $FPRate$ refers to false positive rate. We also compared the area under the Receiver Operating Characteristic (ROC) curve (AUC) and the time to construct the classification model. The detailed performance of the proposed approach, C4.5 and ID3 for each class is summarized in Tables 2, 3, and 4.

5. Conclusion

The objective of this paper was to propose a new data classification algorithm based on a fuzzy approach to process Landsat satellite images. The proposed method uses a hybrid approach through the combination of a decision tree and the MCDA classifier PROAFTN.

Decision tree classifiers such C4.5. ID3 are efficient classifiers in terms of speed and accuracy. However, in some cases especially when dealing with many numeric data as presented in this study, decision trees models could not generate good accuracy. To improve the accuracy, we used the fuzzy PROAFTN to circumvent the problem of crisp intervals, generated by decision tree. This led to better efficiency results and interpretable models.

Table 3. Decision Tree *ID*3 Performance (approx. to three digits).

Categories	Count	TPRate	FPRate	Precision	Recall	FMeasure	*AUC*
1. red soil	1533	0.913	0.023	0.939	0.913	0.926	0.940
2. cotton crop	703	0.804	0.021	0.850	0.804	0.826	0.888
3. grey soil	1358	0.848	0.041	0.871	0.848	0.859	0.900
4. damp grey soil	626	0.543	0.068	0.501	0.543	0.521	0.744
5. soil with vegetation stubble	707	0.778	0.046	0.716	0.778	0.746	0.873
6. very damp grey soil	1508	0.843	0.060	0.841	0.843	0.842	0.893

Table 4. Proposed Method Performance (approx. to three digits).

Categories	Count	TPRate	FPRate	Precision	Recall	FMeasure	*AUC*
1. red soil	1533	0.962	0.022	0.942	0.962	0.952	0.974
2. cotton crop	703	0.953	0.007	0.952	0.953	0.952	0.973
3. grey soil	1358	0.921	0.019	0.938	0.921	0.929	0.949
4. damp grey soil	626	0.65	0.042	0.645	0.65	0.648	0.805
5. soil with vegetation stubble	707	0.82	0.02	0.853	0.82	0.836	0.898
6. very damp grey soil	1508	0.862	0.05	0.855	0.862	0.858	0.908

In conclusion, the new proposed method has shown to be a promising classification tool and merits further investigation. However, further improvements could be investigated, which include: (i) refining attributes and the best set of features from the dataset; (ii) utilize other popular satellite remote sensing imagery such as hyperspectral (AVIRIS) data; and (iii) extend the comparative study to include different classification methods from the machine learning paradigm (e.g, K-Means clustering). The latter is an interesting area of investigation that could consider the study of how classification performance can vary as a function of the input data as some methodologies of classification may actually perform differently depending on the statistics of the image data.

References

1. B. Roy. Multicriteria methodology for decision aiding. *Kluwer Academic*, 1996.
2. UCI: https://archive.ics.uci.edu/ml/index.html.
3. Sebastian Rahtz. A survey of Tex and graphics. Technical Report CSTR 89-7, Department of Electronics and Computer Science, niversity of Southampton, UK, 1989.
4. Rick Lawrence, Andrew Bunn, Scott Powell, and Michael Zambon. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90(3):331 – 336, 2004.
5. Christopher J. Moran and Elisabeth N. Bui. Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science*, 16(6):533–549, 2002.
6. N. Belacel. Multicriteria assignment method PROAFTN: methodology and medical application. *European Journal of Operational Research*, 125(1):175–183, 2000.
7. T. Mitchell. *Machine Learning*. McGraw Hill, 1997.
8. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo CA, 1993.
9. J. R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.
10. N. Belacel and M. Boulassel. Multicriteria fuzzy assignment method: A useful tool to assist medical diagnosis. *Artificial Intelligence in Medicine*, 21(1-3):201–207, 2001.
11. N. Belacel, Q. Wang, and R. Richard. Web-integration of PROAFTN methodology for acute leukemia diagnosis. *Telemedicine Journal and e-Health*, 11(6):652–659, 2005.
12. U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *XIII International Joint Conference on Artificial Intelligence (IJCAI93)*, pages 1022–1029, 1993.
13. H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems, 2005.