Applied Soft Computing 35 (2015) 136–150

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Automatic classification of high resolution land cover using a new data weighting procedure: The combination of *k*-means clustering algorithm and central tendency measures (KMC–CTM)

Süleyman Savaş Durduran *

Necmettin Erbakan University, Faculty of Engineering and Architecture, Department of Geomatics Engineering, 42075 Meram, Konya, Turkey

ARTICLE INFO

Article history:
Received 2 June 2015
Received in revised form 19 June 2015
Accepted 20 June 2015
Available online 29 June 2015

Keywords:
Data pre-processing
Data weighting
The combination of *k*-means clustering algorithm and central tendency measures (KMC–CTM)
Urban land cover
Classification

ABSTRACT

Information on a well-scale urban land cover is important for a number of urban planning practices involving tree shade mapping, green space analysis, urban hydrologic modeling and urban land use mapping. In this study, an urban land cover dataset received from the database of UCI (University of California at Irvine) machine learning was used as the urban land cover data. This dataset is the urban area located in Deerfield Beach, FL, USA. Separately, this dataset is a high definition atmospheric image consisting of 9 different urban land covers. The characteristics of a multi-scale spectral, magnitude and formal tectology were used to sort out and classify these different images. The dataset comprises a total of 147 features and land covers of 9 different areas involving trees, grass, soil, concrete, asphalt, buildings, cars, pools and shadows. A new data weighting method was recommended to classify these 9 different patterns automatically. This recommended data weighting method is based on the combination of the measures of central tendency composed of mean value, harmonic value, mode and median along with the *k*-means clustering method. In the data weighting method, the data sets belonging to each class within the dataset are first calculated by using *k*-means clustering method, after which the measures of central tendency belonging to each class are calculated, as well. The measure of central tendency belonging to each class is divided by the set central value belonging to the class in question, as the result of which the data weight coefficient of that class has already been calculated. This calculation process is performed separately for 9 different land covers, and afterwards, these data weighting coefficients found are multiplied by the dataset, and thus, the dataset has been weighted. In the second stage, on the other hand, 3 different classification algorithms containing *k*-NN (*k*-nearest neighbor), extreme learning machine (ELM) and support vector machine (SVM) were used to classify 9 different urban land covers after the data weighting method. In determining the educational and test data sets, the 10-fold cross validation was used. When classification through raw data was performed along with *k*-NN (for $k = 1$), ELM and SVM classification algorithms, the overall classification accuracy obtained was 77.15%, 84.70% and 84.79%, respectively.

When classification through data weighting method (the combination of *k*-means clustering and mode measure) along with *k*-NN (for $k = 1$), ELM and SVM classification algorithms was made, the overall classification accuracy obtained proved to be 98.58%, 98.62% and 98.77%, respectively. The obtained results suggest that the urban land cover in an atmospheric image via the recommended data weighting method was classified as 9 different areas with a high classification success rate.

1. Introduction

In this study, a new method automatically classifying the land cover of an urban area located in Deerfield Beach, FL, USA [1,2] which comprises trees, grass, soil, concrete, asphalt, buildings, cars, pools and shadows was recommended in 9 different ways. First of all, the subject as to what a land cover is and why it is important was explained; later on, the topic of remote sensing which is the data acquisition stage of an urban land cover was described. The use of GIS (geographic information system) in planning an urban land cover was described, as well. The use of urban land cover dataset in the literature that was acquired from the database of UCI

* Tel.: +90 332 3237926/4026; fax: +90 332 3237926.
E-mail address: ssdurduran@konya.edu.tr

http://dx.doi.org/10.1016/j.asoc.2015.06.025

(University of California at Irvine) machine learning [3] was explained along with the conducted studies mentioned in the literature. Information on a well-scaled urban land cover is of importance for a number of urban planning practices that comprise tree shade mapping, green space analysis, urban hydrologic modeling and urban land use mapping. A land cover is defined as the characteristics and the physical combination of land/areal elements at the surface of the earth [1–3].

The distribution of a land cover is of great significance for scientists and authorities in order to do the mapping of the patterns of a land cover from global and regional to local sizes and to display the changing world better, since it has a major impact on climate and environment. The climate research committee of the National Research Council [4] emphasized the fact that the distribution of a land cover would have a prominent effect on the radiation balance of the Earth due to the fact that any change in the land cover might have an impact on the vaporization, transpiration and the heat flow of the ground surface. For instance, tree eaves/fringes absorb the solar radiation resulting in the decrease in the land surface temperature over the ground. On the other hand, the increase in the stable/resistant urban patterns within the urban area (asphalt, concrete and paving stones) prevents the underground waters likely to cause a potential risk of flood from leaking out. For this reason, displaying the land cover accurately is inevitable for the decision makers involved in public policy planning and the management of world's resources. As one of the areas of the study, remote sensing has attracted the attention of researchers for the past twenty years. In this field of research, various computer-assisted techniques were used for the data analysis comprising artificial neural networks, decision trees, genetic programming, statistical machine learning and other analysis methods [5,6]. An image through remote perception can be used for a number of practices that involve land use imaging, research missions, predicting environmental damages, city planning, radiation imaging, growth regulation, land evaluation and herbal production review. The classification of this image is an indispensable part of such practices in particular [5,6].

Town planners are in need of solutions that appeal to day-to-day business requirements while promoting the ability to predict and respond to the chronic urban problems and market fluctuation for the future. The success of town planners in fighting with chronic urban problems are greatly determined by the ability to use efficient and effective tools and to apply the practice of planning decision support systems. Today, town planners utilize GIS (geographic information system) throughout the world in numerous practices. GIS tools may provide data visualization, modeling, and necessary planning platform for analysis and cooperation [7,23].

In the literature, there are two studies regarding the "urban land cover dataset" classification acquired from the database of UCI (University of California at Irvine) machine learning [3]. In a study conducted by Brian A. Johnson [1], a two-stage-classification procedure was proposed to classify the urban land cover dataset. First, a hierarchy of a seven image segmentation was formed for an urban image by means of different scales, and the preliminary classification values were executed for each of the segmentations by using a classification algorithm. The success rate obtained proved to be 82.10%. In the second study conducted, on the other hand, Johnson et al. proposed a multi-scale approach to classify the land cover in a high-definition urban image. They obtained 78.11% overall accuracy and 0.727 kappa coefficient values without any Super-object information. After the Super-object information had been assigned into the images, 84.42% overall accuracy and 0.804 kappa coefficient values were obtained.

In this article, a new data weighting method to classify nine different land covers in the urban land cover dataset prepared by Johnson et al. was proposed. This data weighting method consists of three stages. In the first stage, set centers are found by using the k-means clustering algorithm for each class within the dataset. Afterwards, in the second stage, the central tendency measurements comprising the statistical values like mean value, harmonic value, mode and median for each class within the dataset are calculated. In the final stage, on the other hand, the data weighting coefficient belonging to each class is calculated by dividing the central tendency measurement belonging to that class by the set central value of that class. This calculation is performed separately for 9 different land covers, and then these data weighting coefficients found are multiplied by the dataset, and thus, the dataset has been weighted. Following the data weighting process, three different classification algorithms were used to classify 9 different land covers. The kappa coefficient values obtained prior to the application of the data weighting process (the combination of k-means clustering and mode measure) were as follows: 0.30 for k-NN (for $k=1$) classifier algorithm, 0.45 for ELM classifier, and 0.45 for SVM classifier. The kappa coefficient values obtained in the wake of applying the data weighting process (the combination of k-means clustering and mode measure) were as follows: 0.90 for k-NN (for $k=1$) classifier algorithm, 0.90 for ELM classifier, and 0.91 for SVM classifier. It is seen that promising results have been obtained in the classification of nine different land covers with the combination of the proposed data weighting method and classification algorithms.

Table 1
The number of data points of nine different land covers in urban land cover dataset [3].

The names of the land cover in the dataset	The number of data points
Trees/shrubs	106
Grass	116
Buildings	112
Concrete	122
Asphalt	59
Vehicles	36
Pools	29
Soil	34
Shadow	61
Total	675

2. Material and method

2.1. Urban land cover dataset

In this study, urban land cover dataset acquired from UCI (University of California at Irvine) machine learning database was used as the urban land cover data [3]. This dataset is an urban area situated in Deerfield Beach, FL, USA [1,2]. Separately, this dataset is a high-resolution atmospheric image consisting of 9 different urban land covers. The characteristics of a multi-scale spectral, magnitude and formal tectology were used to sort out and classify these different images. The dataset comprises a total of 147 features and land covers of 9 different areas involving trees, grass, soil, concrete, asphalt, buildings, cars, pools and shadows. There are a total of 675 data within the dataset. The data distribution according to nine different land covers is given in Table 1. The color infrared image of the study area is given in Fig. 1.

There are a total of 147 features within the urban land cover dataset, and these features comprise the spectral, magnitude, formal and textural properties of the image in question. While forming the features/properties, the spectral, magnitude, formal and textural properties of the image consist of 21 features.

Afterwards, these features were repeated for each coarse scale (scale_20, scale_40, scale_60, scale_80, scale_100, scale_120, and scale_140), and 147 features were obtained. Table 2 shows what scale information the features used within the dataset belong to.