Computers, Environment and Urban Systems 65 (2017) 28–40

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Integration of genetic algorithm and multiple kernel support vector regression for modeling urban growth

Hossein Shafizadeh-Moghadam [a,*], Amin Tayyebi [b], Mohammad Ahmadlou [c], Mahmoud Reza Delavar [c,d], Mahdi Hasanlou [c]

[a] *Department of GIS and Remote Sensing, Tarbiat Modares University, Tehran, Iran*
[b] *Geospatial Big Data Engineer, Monsanto, MO, United States*
[c] *School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran*
[d] *Center of Excellence in Geomatic Engineering in Disaster Management, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran*

ARTICLE INFO

Article history:
Received 30 November 2016
Received in revised form 10 April 2017
Accepted 10 April 2017
Available online xxxx

Keywords:
Land cover change
Support vector regression
Genetic algorithm
Kernels
Total operating characteristic

ABSTRACT

There are two main issues of concern for land change scientists to consider. First, selecting appropriate and independent land cover change (LCC) drivers is a substantial challenge because these drivers usually correlate with each other. For this reason, we used a well-known machine learning tool called genetic algorithm (GA) to select the optimum LCC drivers. In addition, using the best or most appropriate LCC model is critical since some of them are limited to a specific function, to discover non-linear patterns within land use data. In this study, a support vector regression (SVR) was implemented to model LCC as SVRs use various linear and non-linear kernels to better identify non-linear patterns within land use data. With such an approach, choosing the appropriate kernels to model LCC is critical because SVR kernels have a direct impact on the accuracy of the model. Therefore, various linear and non-linear kernels, including radial basis function (RBF), sigmoid (SIG), polynomial (PL) and linear (LN) kernels, were used across two phases: 1) in combination with GA, and 2) without GA present. The simulated maps resulting from each combination were evaluated using a recently modified version of the receiver operating characteristics (ROC) tool called the total operating characteristic (TOC) tool. The proposed approach was applied to simulate urban growth in Rasht County, which is located in the north of Iran. As a result, an SVR-GA-RBF model achieved the highest area under curve (AUC) value at 94% while the lowest AUC was achieved when using the SVR-LN model at 71%. The results show that the synergy between GA and SVR can effectively optimize the variables selection process used when developing an LCC model, and can enhance the predictive accuracy of SVR.

1. Introduction

The process of land cover change (LCC) is a function of interrelated biophysical, economic, social, cultural and political driving forces or factors (Turner, Lambin, & Reenberg, 2007). Urbanization is one of the most important dimensions of land cover systems, owing to the fact that it can have a variety of impacts on different environmental factors such as climate (Tayyebi & Jenerette, 2016), water quality (Zhou, Huang, Pontius, & Hong, 2016), the intensity of agricultural land cover (Song, Pijanowski, & Tayyebi, 2015; Tayyebi et al., 2016), and biodiversity and deforestation levels (Lambin & Geist, 2008). To develop a better understanding of the effects of urbanization on the environment, and to quantify the driving forces behind complex urban systems, the use of innovative data mining and machine learning approaches is vital (Li & Yeh, 2002; Yang & Lo, 2003). Accordingly, numerous data mining and machine learning techniques have been developed to simulate LCC over the last three decades (see Tayyebi, 2013 for more details). The availability of a wide spectrum of LCC modeling techniques has opened the opportunity for researchers to select those methods most appropriate in helping to answer their research questions (Turner et al., 2007; Tayyebi et al., 2013 and Tayyebi, Pijanowski, Linderman, & Gratton, 2014).

Over the last three decades, various machine learning, data mining, statistics- and process-based LCC models have been used to understand LCC processes (Hu & Lo, 2007; Kamusoko & Gamba, 2015; Pijanowski et al., 2014; Rienow & Goetzke, 2015; Shafizadeh-Moghadam, Hagenauer, Farajzadeh, & Helbich, 2015). When using LCC models, two key issues within land change science have tended to arise (Tayyebi, 2013). First, because LCC drivers operate on a variety of spatial and temporal scales, LCC modelers always have to decide on the best way to select the appropriate and independent LCC drivers. This is due the fact that some of LCC models require the use of independent LCC drivers for modeling purposes. Second, LCC modelers also have to decide on the best LCC

[*] Corresponding author.
E-mail addresses: h.shafizadeh@modares.ac.ir (H. Shafizadeh-Moghadam), m_ahmadlou@ut.ac.ir (M. Ahmadlou), mdelavar@ut.ac.ir (M.R. Delavar), hasanlou@ut.ac.ir (M. Hasanlou).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.04.011
0198-9715/© 2017 Elsevier Ltd. All rights reserved.

model to use to model the LCC in question (Pontius et al., 2008) since users might not have a deep understanding of the existing LCC models in place and their capabilities. This limitation also arises as some LCC models are limited to one function when identifying non-linear patterns within a land cover dataset. Among the large number of LCC studies to be found within the land change science literature, these issues remain largely unresolved.

Among the LCC models available, machine learning techniques (e.g., ANN and SVM) have gained increasing attention among scholars (Kamusoko & Gamba, 2015; Pijanowski, Brown, Shellito, & Manik, 2002; Rienow & Goetzke, 2015; Shafizadeh-Moghadam, Asghari, Taleai, Helbich, & Tayyebi, 2017; Tayyebi & Pijanowski, 2014). One of the main reasons why the ANN and SVM models are so popular is that both approaches provide a variety of functions (e.g., ANN) and kernels (e.g., SVM) able to model the complexity and non-linearity of urban dynamics (Shafizadeh-Moghadam & Helbich, 2015; Tayyebi, Pijanowski, & Pekin, 2015 and Tayyebi, Pijanowski, & Tayyebi, 2011). These properties give users a choice of functions and kernels so as to model LCC. In the portfolio of available LCC models, being able to understand their performance, characteristics, strengths and limitations is a scientific requirement during the model selection phase (Pontius et al., 2008; Tayyebi et al., 2014). SVM is a well-known machine learning technique which transforms input data to a higher dimension in order to solve non-linear classification or regression problems (Cortes & Vapnik, 1995), and is independent of any prior knowledge (Rienow & Goetzke, 2015). SVM learning problems can be expressed as convex quadratic programming problems, the aim of which is to seek the global, optimal solution (Lin & Yan, 2016). SVM can, therefore, avoid the issue of local extremes, a problem that can occur when using other machine learning techniques such as ANN (Vapnik & Vapnik, 1998). SVM has been shown to be an effective tool to create LCC maps as part of an LCC modeling approach (Rienow & Goetzke, 2015; Yang, Li, & Shi, 2008). The core functionality of the SVM approach is the use of kernels which can take both linear and non-linear forms. Kernels play an important role in creating the prediction accuracy of SVM models, and the most common kernels that have been used with the SVM approach include the radial basis function (RBF), sigmoid (SIG), polynomial (PL) and linear (LN) kernels. However, we know of no study carried out that has recommended the best and most appropriate kernels to use when modeling LCC. As a result, the first objective of this research is to evaluate the influence of linear and non-linear kernels on the accuracy of the simulated urban growth maps produced using the SVM approach.

When using the LCC modeling process, several environmental and socio-economic variables can influence urban growth (Hu & Lo, 2007). As well as the large volume of satellite images available, which are often computationally expensive to process, the large number of explanatory variables involved can give rise to the issue of computational time being required (Pijanowski et al., 2014). Moreover, collinearity, which refers to the dependency among predicting factors (Dormann et al., 2013), is another issue which can arise from the use of a large number of predictors. The problem of collinearity is particularly serious when a model has to be adjusted and prepared in the light of data coming from one district or time point that used to make a prediction in another area, or data with a different or obscure collinearity structure (Pontius, Huffaker, & Denman, 2004). The selection of the most effective predictive variables is essential when modeling LCC. For example, Pal and Foody (2010) revealed that the performance of SVM varies as a function of the number of input features. In another study, Shafizadeh-Moghadam et al. (2015) showed that the design and even the types of functions used to model LCC can affect the accuracy of the model produced.

To achieve a better performance, therefore, it is important to pay careful attention to the optimum selection of the model's features. For this purpose, several statistical techniques such as principle component analysis (PCA; Dormann et al., 2013) and factor analysis, as well as evolutionary techniques such as the genetics algorithm (GA) have been used in previous studies (Shan, Alkheder, & Wang, 2008). PCA can be applied on continuous data; however, LCC explanatory factors tend to consist of a combination of continuous and categorical data groupings (Dormann et al., 2013). GA uses stochastic search methods inspired by natural evolution principles (Davis, 1991; Engelbrecht, 2007) to select the most effective LCC drivers. For example, Tang, Wang, and Yao (2007) coupled GA with Markov chain models to carry out a feasibility study of the potential for remote sensing to predict future landscape change. They found GA to be useful in helping to demonstrate spatial information in a spatio-temporal model. On the basis of SVR and GA, Nieto, Fernández, de Cos Juez, Lasheras, and Muñiz (2013) suggested a hybrid approach known as a GA–SVR to predict the presence of cyanotoxins in the Trasona reservoir in northern Spain, but did not investigate the role of various kernels in the performance of the model, nor did they compare their suggested model against SVR without the GA component added. The second objective of this study is to combine the GA approach with various SVM kernels, the aim being to reduce the high dimensionality levels of the input variables and provide the optimum selection of predictive variables.

Carrying out an accuracy assessment of the suitability and accuracy of the simulated maps produced by a model is of great importance (Pontius & Schneider, 2001), and a variety of calibration metrics have been used in land change science to this end (Tayyebi et al., 2014). Among them, receiver operating characteristic (ROC) – one of the most common accuracy matrices – has received a lot of attention in land change science circles (e.g., Hu & Lo, 2007; Rienow & Goetzke, 2015). ROC evaluates the predictive ability of a suitability map for binary classification, using a reference map for each given threshold, varying from 0 to 1. However, ROC has recently been criticized by scientists (e.g., Golicher, Ford, Cayuela, & Newton, 2012; Pontius & Si, 2014); for example, for failing in cases where some types of error are more important than others (Pontius & Parmentier, 2014). ROC also fails to reveal the size of each entry in the contingency table for each threshold. As an alternative, Pontius and Si (2014) recently introduced the total operating characteristic (TOC) approach to rectify the limitations of ROC. The area under the curve for TOC is the same as for ROC. In this study we compared the performance of both ROC and TOC in the evaluation of SVR kernels. To confirm the efficiency of the proposed framework, SVR with varying kernels was implemented with and without the GA, and the results were compared.

2. Study area and dataset

2.1. Study area

In this study, we used our models on the city of Rasht, the capital of Gilan Province in Iran and the largest and most populous city on the Caspian Sea coast. The city is located at 37° 53′ N and 49° 58′ E, and has an area of approximately 180 km^2 (Fig. 1). In its 2011 census, the county's population was 920,000. The city has six districts, these being Khomam, Khoshke Bijar, Kuchesfahan, Lashte Nesha, Sangar and Central. The conjunction of the Caspian Sea coast with the plains and mountainous regions set behind has made urban Rasht one of the major tourist centers in Iran, attracting thousands of tourists annually.

In recent decades the city has experienced increasing population growth and urban expansion. Similar to other large Iranian cities and provincial capitals, industrialization is a key feature in this region. In Gilan Province as a whole, the people, services, industry and investment are concentrated in Rasht, and the city's economic growth has influenced most of the province's peripheral cities. Due to urban expansion in recent decades, many peripheral villages have been absorbed into Rasht's urban zones.

2.2. Dataset

The information required for this study was derived from Landsat satellite images, plus we extracted road networks from topographic