IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Survey on Machine Learning Algorithms for Classification and Prediction of Land Use Changes Using GIS

N. Bharanikumar[1*], P. Dhanalakshmi[2]

[1]Data Entry Operator, Examination Wing-II, Annamalai University.
[2]Dept. of Computer science and Engineering, Annamalai University, Annamalai Nagar, India

Corresponding Author:cdmbharani81@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** Large area land-cover monitoring scenarios, involving large volumes of data, are becoming more prevalent in remote sensing applications. Thus, there is a pressing need for increased automation in the change mapping process the land transformation Model (LTM), which couples geographic information systems (GIS) with artificial neural networks. The objective of this research presents the survey report based on compare the performance of three machine learning algorithms (MLAs) and prediction of land use changes in GIS. The change map generated using ARTMAP has similar accuracies to a human-interpreted map produced by the U.S. Forest Service in the southern study area (John Rogan et al 2007). ARTMAP appears to be robust and accurate for automated, large area change monitoring as it performed equally well across the diverse study areas with minimal human intervention in the classification process. GIS is used to develop the spatial, predictor drivers and perform spatial analysis on the results. The predictive ability of the model improved at larger scales when assessed using a moving scalable window metric. the individual contribution of each predictor variable was examined and shown to vary across spatial scales.At the smallest scales, quality views were the strongest predictor variable. We interpreted the multi-scale influences of land use change, illustrating the relative influences of site (e.g. quality of views, residential streets) and situation (e.g. highways and county roads) variables at different scales.

***Keywords*—**Component, Formatting, Style, Styling, Insert (key words)

I. INTRODUCTION

Changes in land use result from the complex interaction of many factors including policy, management, economics, culture, human behavior, and the environment (Dale, O'Neill, Pedlowski, & Southworth, 1993; Houghton, 1994; Medley, Okey, Barrett, Lucas, & Renwick, 1995; Richards, 1990; Vesterby & Heimlich, 1991; Wilder, 1985). An understanding of how land use changes occur is critical since these anthropogenic processes can have broad impacts on the environment, altering hydrologic cycles (Steiner & Osterman, 1988), biogeochemical dynamics (Flintrop et al., 1996), size and arrangements of natural habitats such as forests (Dale et al., 1993) and species diversity (Costanza, Kemp, & Boynton, 1993). Changes to land use can also affect local and regional economies (Bingham et al., 1995; Burchell, 1996). This paper illustrates how combining geographic information systems (GIS)and artificial neural networks (ANNs) can aid in the understanding the complex process of land use change. A GIS-based Land Transformation Model — LTM (Pijanowski, Gage, Long, & Cooper, 2000) was developed to forecast land use change over large regions. This model can be configured to use a variety of socioeconomic, political and environmental inputs. The LTM can link changes in land use to ecological process models, such as groundwater flow and solute transport (Boutt et al., 2001) and forest cover change (Brown, Duh, & Drzyzga, 2000; Brown, Pijanowski, & Duh, 2001). It can also provide local land use planners and regional resource managers with information about the potential effects of land use change on the environment. The Holy Grail of (digital) change detection is still total automation and high accuracy." (Loveland et al., 2002, p. 1098). Over the coming decades, the global effects of land-cover/use change may be as significant, or more so, than those associated with potential climate change (IPCC, 2000). In spite of this there is a lack of comprehensive information on the types and rates of land-cover/use change, and even less evidence of natural and anthropogenic causes and consequences of such change (Turner et al., 1999). As a result, several large area land-cover monitoring programs have been established over the past five years to comprehensively address this issue (Wulder et al., 2004). Monitoring programs, unlike most research-oriented studies, employ change mapping methods that require processing and interpretation of large volumes of in situ, remotely sensed and ancillary data (Cilhar 2000; Franklin&Wulder, 2002).Very large data volumes and time-

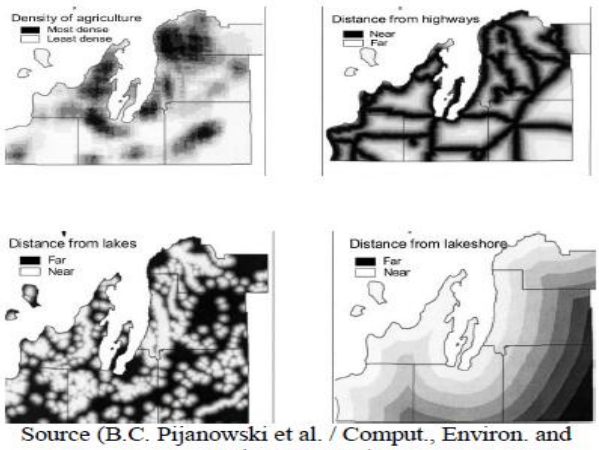
consuming data processing, integration and interpretation make automated and accurate methods of change mapping highly desirable (Aspinall, 2002; Dobson & Bright, 1994; Hansen et al., 2002; Rogan & Chen, 2004). Complex land change processes are of particular interest to researchers involved in large area monitoring (Roberts et al., 2002), where many different types of land-cover changes can occur and must be characterized (e.g., forest pest infestation, logging, wildfire, and suburbanization) (Rogan & Miller 2006). Thus, increased automation can ensure that the classification process is objective and repeatable in processing large volumes of data over complex and phenologically diverse landscapes (DeFries & Chan 2000; Gong & Xu 2003). Consequently, classification algorithm selection and performance have become particularly important, because large area change monitoring can only realistically be achieved (i.e., low cost and generalizable results) through techniques that minimize timeconsuming human interpretation and maximize automated procedures for data analysis (Woodcock et al., 2001). There is now a large body of research that demonstrates the abilities of machine learning techniques, particularly classification trees and artificial neural networks, to deal effectively with tasks involving high dimensional data (Gahegan 2003). The increased interest in MLAs can be attributed to several factors:

- Their non-parametric nature deals well with multi-modal, noisy and missing data (Hastie et al. 2001), but see Simardet al. (2000) in the case of classification trees;
- There is a significant reduction in¬ computational demands when data measurement spaces are large and complex (Foody 2003);
- They readily accommodate both categorical¬ and continuous ancillary data (Lawrence & Wright 2001);
- Users can investigate the relative¬ importance of input variables in terms of contribution to classification accuracy (Hansen et al., 1996; Foody & Arora 1997);
- They are flexible and can be adapted to¬ improve performance for particular problems (Lees & Ritman 1991) And multiple subcategories per response¬ variable can be accommodated (Gopal et al., 1999).

II. LAND USE CHANGE MODELS

Models of land use change serve as useful tools for (1) exploring the various mechanisms by which land use change occurs and the social, economic, and spatial variables that drive it (Batty & Longley, 1994; deKoning, Verburg, Veldkamp, & Fresco, 1999), projecting potential future environmental and economic impacts of land use change (Alig, 1986; Theobald, Miller, & Hobbes, 1997), and (3) evaluating the influence of alternative policies and management regimes on land use and development patterns (Bockstael, Costanza, Strand, Boyton, Bell, & Wagner, 1995). Although some approaches focus on modeling aggregate land use amounts within areal units, like counties (Alig & Healey, 1987), models that predict the spatial patterns of land use provide more information with which to evaluate the impacts of change. To address the multi-scale nature of land use change drivers, some models, like the CLUE model of de Koning et al. (1999), have used regional and global scale drivers to determine the aggregate amounts of change and geographic and landscape scale drivers to determine its pattern. This is the adopted approach for development of the Land Transformation Model. Artificial neural networks, are used to determine the location of land use change using landscape scale variables given a certain amount of change determined by regional and global scale variables. The variable values and actual instances of land use change are typically observed from historical data and used to establish functional relationships that can be used to extrapolate land use change probabilities into the future. The spatial transition models are an extension of the aspatial Markov technique and a form of stochastic cellular automata (CA; Theobald & Hobbs).

Ten predictor variables and the exclusion zones were compiled in Arc/Info Grid format (Table 1; Fig. 3 using the LTM GIS Avenue interface. The agricultural density variable represents the amount of agriculture, from the 1980 land use database,within a 1 km radius surrounding each cell. This variable describes the degree



Source (B.C. Pijanowski et al. / Comput., Environ. and Urban Systems)

Agriculture can beseen as an amenity on the landscape that attracts development. However, it is possible that agricultural land use can serve as an impediment to development, especially in this area where agricultural activities are specialized (vineyards and other fruit production) and profitable. For the variables of county roads distance, highway distance, shoreline distance, inland lake distance, and river distance, the minimum Euclidean distance to each

feature was calculated. These features serve to either improve the access of the site to larger urban areas (i.e. county roads and highways) or to increase the amenity value of a site (i.e.shoreline, inland lakes, rivers). The urban distance variable was the minimum distance of each cell to an urban cell, from the 1980 Grand Traverse County land use database.

Since access to urban services affects development patterns, it is expected that sites nearer to existing urban land uses would be more likely to develop. Distance from recreation sites, which were coded as point coverages, was also used as a predictor variable. For view quality, the height above lake level and the distance from the lakeshore were calculated for every location in the watershed. The sine of the angle of incidence of a line-of-sight from each location in the watershed to the lake was then calculated and used as a surrogate for quality of view. Larger angles represent locations that are highly elevated above the bay and are close. These locations are hypothesized to be in great demand for residential use. The exclusion zone for this execution of the model was composed of the following GIS layers: areas that were urban in 1980; locations of open water; locations of wetlands; locations of public land (e.g. local, state and federal parks) and locations of current transportation corridors.

III. MACHINE LEARNING ALGORITHMS

S-Plus classification tree (CT) The first classification tree routine tested was the one incorporated in the S-Plus statistical software package (Clark & Pregibon 1992), one of the most widely used tree algorithms in classification of land-cover (Hansen et al., 1996, Wessels et al., 2004), which employs a deviance measure to partition data set. The reduction in deviance (e.g., increase in subset homogeneity using and entropy measure) (D) is calculated as:

$$D=Ds-Dt-Du,$$

Where s represents the parent node, and t and u are splits from s. When D is maximized, the best split is identified, and the data are divided at that value. The process is repeated onthe two new nodes of the tree. The deviance for nodes is calculated from

$$D_i = -2 \sum n_{ik} \log p_{ik}$$

where n is the number of observations in class k in node i and p is the probability distribution of node i and class k. S-Plus classification trees were pruned to an optimum size based on cross-validation using ten independent subsets of the training data. This resulted in a parsimonious tree model that did not over fit the training data, thus leading to more generalizable information (Franklin, 1998). While there is generally not a single solution to pruning for all applications of classification trees, and the decision of how much to prune can affect the results, researcher did not compare different pruning trials in this work, as has been suggested by Zambon et al. (2006). We refer to S Plus classification tree as CT.

IV. EFFECT OF NOISE IN TRAINING SET

Training data errors are likely in a large area context given the disparate sources of information used and the fact that class labels can become more confused as landscape heterogeneity increases. Several authors have noted that MLAs are adversely affected by noise, which can yield very different results when included in the training phase by causing intra-class variability in the data (Simard et al., 2000; Miller & Franklin 2002). Conversely, Brodley and Friedl (1996) found classification trees to be tolerant of noisy data, and Paola and Schowengerdt (1995) found neural networks robust to training site heterogeneity. Therefore, it is instructive to examine the effect of noise in a change mapping context to provide more information on this topic, and to examine the level of robustness one may expect from MLA's.

V. CONCLUSION

The effect of training set size on algorithm performance indicates that large numbers of training and test sites are important in change mapping using MLAs. Nonetheless, reasonable accuracies were achieved at certain levels of data reduction, implying that the quality, if not the quantity of sample data was adequate. Results suggest that below a certain size, a data set is less representative of the conditions it is supposed to represent, resulting in reduced map accuracy. It might also suggest that the data sets used here have few redundant observations, or else reducing the size would have less of an effect. Careful attention should be paid to training and test set selection in a MLA context. Variations in training and test site sample size had a significant effect on the behavior of each MLA.

REFERENCES

[1] Alig, R. J. (1986). Econometric analysis of the factors influencing forest acreage trends in the southeast. Forest Science,32, 119-114.
[2] Alig, R. J., & Healy, R. G. (1987). Urban and built-up land area changes in the United States: an empirical investigation of determinants. Land Economics, 63(3), 216–226.
[3] Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. 1976). A land use and land cover classification system for use with remote sensor data. US Geological Survey, Professional Paper 964, p. 28, Reston, VA.
[4] Atkinson, P., & Tatnall, A. (1997). Neural networks in remote sensing. International Journal of RemoteSensing, 18(4), 699–709.
[5] Babaian, R., Miyashita, H., Evans, R., Eshenbach, A., & Ramimrez, E. (1997). Early detection program for prostate

cancer: results and identification of high-risk patient population. Urology, 37(3), 193–197.
[6] Batty, M., & Longley, (1994). Urban modeling in computer-graphic and geographic information system environments. Environment and Planning B, 19, 663–688.
[7] Bingham, G., Bishop, R., Brody, M., Bromley, D., Clark, E., Cooper, W., Costanza, R., Hale, T., Hayden, DeFries, R. S., & Chan, J. (2000).
[8] Multiple criteria for evaluating machine learning algorithms for land-cover classification from satellite data. Remote Sensing of Environment, 74, 503−515.
[9] Dobson, J., & Bright, E. (1994). Largearea change analysis: The Coast-watch Change analysis Project (C-CAP). Proceedings of Pecora 12, 24–26 August 1993, Sioux Falls, South Dakota (pp. 73−81).
[10] Foody, G. M., & Arora, M. K. (1997). Evaluation of some factors affecting the accuracy of classification by an artificial neural network. International Journal of Remote Sensing, 18, 799−810.
[11] Foody, G.M. (2003). Uncertainty, knowledge discovery and data mining in GIS. Progress in Physical Geography, 27(1), 113−121.
[12] Franklin, J. (1998). Predicting the distribution of shrub species in southern California from climate and terrainderived variables. Journal of Vegetation Science, 9, 733−748.
[13] G., Kellert, S., Norgaard, R., Norton, B., Payne, J., Russell, C., & Suter II, G.(1995). Issues in ecosystem valuation: improving information for decision making. Ecological Economics, 14, 73–90.
[14] Bockstael, N., Costanza, R., Strand, I., Boyton, W., Bell, K., & Wagner, L. (1995). Ecological economic modeling and valuation of ecosystems. Ecological Economics, 14, 143–159.
[15] Boutt, D. F., Hyndman, D. W., Pijanowski, B. C., & Long, D. T. (2001). Identifying potential land usederived solute sources to stream baseflow using ground water models and GIS. Groundwater, 39(1),24–34.
[16] Brown, D. G., Duh, J. D., & Drzyzga, S. (2000). Estimating error in an analysis of forest fragmentation change using North American Landscape Characterization (NALC) Data. Remote Sensing of Environment, 71, 106–117.
[17] Gopal, S., & Woodcock, C. E. (1996). Remote sensing of forest change using artificial neural networks. IEEE Transactions on Geoscience and Remote Sensing, 34, 398−404.

