Classification Rule Learning for Automated Building Generalization

Khelifa Djerriri
Centre des Techniques Spatiales
Arzew, Algeria
djerriri.k@gmail.com

Bouhadjar Meguenni
Centre des Techniques Spatiales
Arzew, Algeria
bmeguenni@cts.asal.dz

Abstract

Deriving small-scale geographic data based on large-scale one is a frequent problem in mapping large data sets. Various map generalization techniques, such as simplification, displacement, elimination, and aggregation, must therefore be applied. In this study, we focused on the elimination and aggregation of polygons in building layer, for which each building in a large scale was classified as eliminated, retained, aggregated. Attribute selection and rule-based classification algorithms were then used for selecting and classifying the building objects. Vector maps of 1:5,000 scale and 1:25,000 scale were used to perform experiments. Obtained results showed the effectiveness of the proposed approach.

Keywords: Map generalization, Rule based classification, Attribute selection.

1 Introduction

Map providers produce various maps at different scales and themes. A challenging problem for them is deriving small-scale maps from larger-scale ones. The current used processes are time-consuming, costly, and complicated. The process of transforming large-scale maps into small-scale maps is called map generalization, which is intended to improve the readability of maps and maintain their essential information [Brassel and Weibel 1988, McMaster and Shea 1992, K.S Ruas et Plazanet 1996].

Typical map generalization operators include simplification, displacement, elimination, and aggregation. Elimination is needed to remove small buildings, such as sheds or isolated buildings. Displacement is needed to separate buildings that would be too close to each other in the desired map scale, or to move buildings further from roads. Aggregation groups buildings into larger units of built-up blocks if the buildings are not to be separately shown [Lee et al. 2017].

Many studies have been conducted on automated map generalization in the GIS/cartography field [Lee et al. 2017, Wang et al. 2017]. The above operators were developed based on these studies, which determined how map features should be represented at a small scale. Among the various data themes, the building features have attracted much research attention in the field of map generalization because of their man-made shape and complex spatial distribution.

2 Learning Classification Rules

In this study, generalization process is considered as classification problem, where each building polygon in a smaller scale, the following output classes were assigned: “eliminated,” “retained,” and “aggregated”.

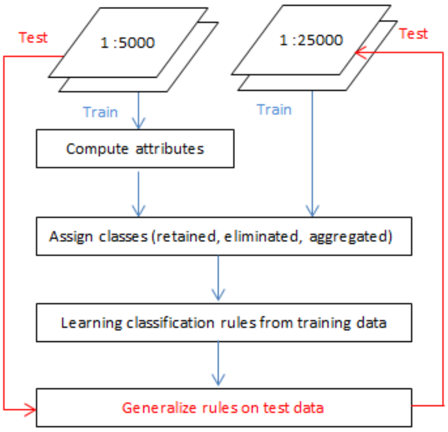
Interpretability is an added value sought in classifiers that are built within supervised machine-learning. Given a set of training data, a common task is to extract a model that predicts the class label of an unseen example. The generalization or discriminating capacity of a classifier measures the number these predictions of unseen examples that are correct.

In machine learning and data mining [Urbanowicz and Moore 2009, Fürnkranz et al. 2012], decision tree and rule induction algorithms possess the desired ability to build understandable models. They share the goal of finding regularities in data that can be expressed in the form of an IF-THEN rule.

3 Proposed methodology

A four-staged methodology is proposed in this study with the goal of extracting useful rules for map generalization .We mainly focused on three operators, retaining, elimination and aggregation. The flowchart of the methodology is shown in Figure 1. The source data for testing, the building layer of a digital map with a scale of 1:5000 and 1:25000.

Figure 1: proposed methodology flowchart


1 :5000
1 :25000
Test
Test
Train
Train
Compute attributes
Assign classes (retained, eliminated, aggregated)
Learning classification rules from training data
Generalize rules on test data

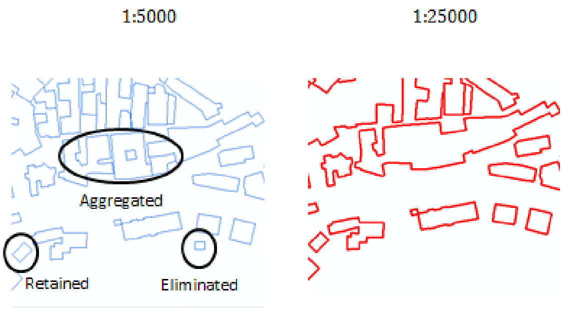
3.1 Building polygons characterization

We had to generate training data to apply the machine learning technique to this process. Training data are required for applying classification algorithms, and they must contain input features and output classes. We extracted the input features from attributes of 1:5000 vector layer and to generate the output class by comparing the areas of 1:5000 and 1:25,000 data. Various geometric and topological attributes were used as input features.

3.2 Polygons labeling

For the representation of each building in a smaller scale, the following attributes were used as output classes: “0-eliminated,” “1-retained,” and “2-aggregated”. As suggested in [Lee et al. 2017].] the labeling was performed using the overlapping areas between 1:5000 and 1:25,000 buildings. The criterion for identification of two building objects in both datasets as representing the same building object in reality was the overlapping area, which was over 80%. Figure 2 shows the three cases taken into account during the buildings labelling process.

Figure 2: the three cases taken into account



3.3 Feature selection

The desired generalization rules should be as simple as possible. Thus a feature selection step was necessary to eliminate the less informative attributes in the classification process.

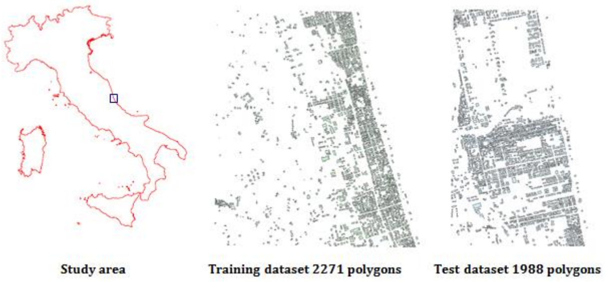
3.4 Classification rules learning

The final stage is the generation of the classification rules and writes them in an understandable way such as decision trees or IF-Then rules, then apply these rules on unseen dataset 1:5000 to test their generalization ability.

4 Used dataset

The 1:5000 and 1:25000 building layers used in this study were downloaded from the OpenGeodata of Abruzzo region in Italy “Il portale dei dati aperti della Regione Abruzzo” (opendata.regione.abruzzo.it).The primary goal in this study was to derive the 1:25000 scale. We selected 1:5000 and 1:25000 data as source data because these datasets showed more prominent differences than the other available datasets. In total, 2271 buildings were used for the training data and 1988 for the test data (Figure 3). The building shapes had various forms.

Figure 3: study area and used training and test datasets



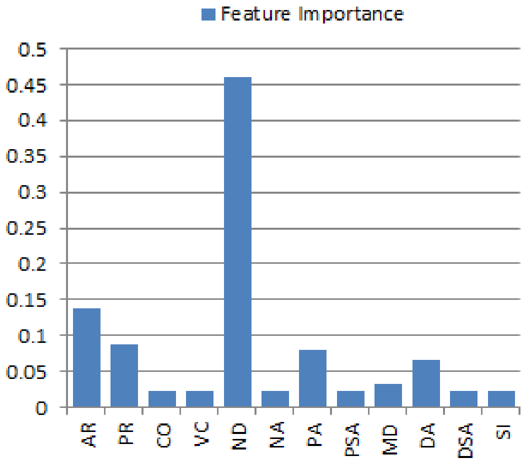
5 Results and discussion

Random Forests, an assembling classification tree, that provides feature importance index, we iteratively eliminate features with less important index until the mean decrease accuracy is stable [Breiman 2001, Guan et al. 2012]

A set of topological and polygon shape indices are computed for each 1:5000 polygon. This includes: distance to closest polygon (ND), area (AR), perimeter (PR), compactness (CO), perimeter / area (PA), perimeter / square root of the area (PSA), maximum distance (MD), maximum distance / area (DA), maximum distance / square root of the area (DSA) and shape index (SI).

The feature importance values obtained using the random forest classifier, are presented in Figure 4.

Figure 4: feature importance for each attribute



According to figure 4 Area and Distance to neighbor polygon are the most importance attributes. Thus, these two features are considered in the classification process.

In our study, as mentioned earlier, we applied the following classification rule learning methods: C4.5 decision tree (DT), Repeated Incremental Pruning to Produce Error Reduction rules (RIPPER), PARtial decision Trees (PART) and the Unordered Fuzzy Rule Induction (FURIA). The different rule learning experiments are conducted using the well-known datamining Weka software [Quinlan 1993, Cohen1995, Frank and Witten 1998, Hall et al. 2009, Hühn 2009].

The rules obtained by the C4.5 DT algorithm are the following:

```
NEAR_DIST <= 1.482725
|  Area5k <= 652.9: aggregated (420.0/57.0)
|  Area5k > 652.9
|  |  NEAR_DIST <= 0.13345
|  |  |  Area5k <= 772: eliminated (5.0/1.0)
|  |  |  Area5k > 772: retained (8.0/1.0)
|  |  NEAR_DIST > 0.13345: aggregated (6.0/1.0)
NEAR_DIST > 1.482725
|  Area5k <= 49.5: eliminated (184.0/3.0)
|  Area5k > 49.5: retained (1648.0/41.0)
```

Where numbers between brackets represents: first is the number of polygons covered by the rule and the second is the number of misclassified ones.

The rules obtained by the RIPPER algorithm:

(Area5k <= 49.5) and (NEAR_DIST >= 1.528121) => Class=eliminated (184.0/3.0)
(NEAR_DIST <= 1.479654) and (Area5k <= 190) => Class=aggregated (265.0/29.0)
(NEAR_DIST <= 1.483) and (NEAR_DIST >= 0.22422) => Class=aggregated (122.0/19.0)
(NEAR_DIST <= 0.22269) and (Area5k <= 508.2) => Class=aggregated (38.0/9.0)
=> Class=retained (1662.0/46.0)

The rules obtained by the PART algorithm

NEAR_DIST > 1.482725 AND
Area5k > 49.5 AND
NEAR_DIST <= 19.910824: retained (1517.0/29.0)
NEAR_DIST <= 1.482725 AND
Area5k <= 652.9: aggregated (420.0/57.0)
Area5k <= 60.6 AND
Area5k <= 48.6: eliminated (181.0/2.0)
NEAR_DIST > 19.911294 AND
Area5k > 60.6: retained (126.0/7.0)
NEAR_DIST > 0: aggregated (15.0/6.0)
Area5k > 772: retained (8.0/1.0)
: eliminated (4.0/1.0)

The rules obtained by the FURIA algorithm

(NEAR_DIST in [-inf, -inf, 1.482725, 1.509036]) and (NEAR_DIST in [0.118258, 0.13345, inf, inf]) => Class=aggregated (CF = 0.88)
(NEAR_DIST in [-inf, -inf, 0, 1.170677]) and (Area5k in [-inf, -inf, 217.4, 218.6]) => Class=aggregated (CF = 0.9)
(NEAR_DIST in [-inf, -inf, 0, 1.222604]) and (Area5k in [-inf, -inf, 508.2, 556.7]) => Class=aggregated (CF = 0.86)
(NEAR_DIST in [1.538829, 1.545353, inf, inf]) and (Area5k in [56.2, 57, inf, inf]) => Class=retained (CF = 0.98)
(Area5k in [49, 50.6, inf, inf]) and (NEAR_DIST in [1.482725, 1.509036, inf, inf]) and (NEAR_DIST in [-inf, -inf, 18.213694, 19.911294]) => Class=retained (CF = 0.98)
(Area5k in [1162.5, 1249.7, inf, inf]) and (NEAR_DIST in [-inf, -inf, 0, 0.842308]) => Class=retained (CF = 0.93)
(Area5k in [-inf, -inf, 49.5, 50.6]) and (NEAR_DIST in [1.435002, 1.528121, inf, inf]) => Class=eliminated (CF = 0.97)

The confusion matrix from the PART algorithm outputs is presented in Table 1.

Table 1: Confusion matrix.

aggregated	retained	eliminated	<-- classified as
372	24	2	aggregated
26	1614	1	retained
37	13	182	eliminated

The number of generated rules, overall classification accuracy, kappa coefficient and time taken to build each model are provided in Table 2.

Table 2: obtained classification statistics for each algorithm.

Algorithm	Nb. Rules	Acc. (%)	Kappa	Time (s)
DT	6	95.42	0.8939	0.02
RIPPER	5	95.33	0.8915	0.12
PART	7	95.46	0.8952	0.10
FURIA	7	95.24	0.8899	0.72

Statistics in Table 2 are computed based on obtained classification confusion matrices. The overall accuracies of each algorithm are: DT, 95.42%; RIPPER, 95.33%; PART, 95.46%; and FURIA, 95.24%. All the four algorithm show high accuracies (above 95%) and Kappa coefficients (above 0.88) .DT showed high accuracy and significantly lower time taken to build classification model than the other three algorithms, but the obtained model is a decision tree instead of IF-THEN rules. PART algorithm showed the highest accuracy and kappa coefficient and the generated rules are of type IF-THEN.

6 Conclusion

In this study, we applied differents classification rules learning (DT, PART, RIPPER, FURIA) to retained, eliminate and aggregate building polygons when updating scale from 1:5000 to 1:25000. Various topological and geometric properties of the building layer were used. The accuracy of each algorithm was also evaluated.

The proposed technique can be used to other scale (e.g., 1:1000 and 1:50,000 scale). It is simple to apply this method to the new datasets, as long as the building is labeled by the proposed method or other methods. Results can be obtained by applying the above-noted machine learning algorithms, the output class resulting from labeling, and the attribute as the input feature, which will affect the output class.

This study had some limitations and more building generalization cases should be addressed such as displacement and simplification.

Moreover, extracting useful rules from black-boxed classifiers such as SVM and Artificial neural networks can be considered in future works.

References

Brassel, K.E.; Weibel, R. A review and conceptual framework of automated map generalization. Int. J. Geogr. Inf. Syst. 1988, 2, 229–244.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cohen, W. W. (1995). Fast effective rule induction. In Machine Learning Proceedings 1995 (pp. 115-123).

Guan, H., Yu, J., Li, J., & Luo, L. (2012). Random forests-based feature selection for land-use classification using lidar data and orthoimagery. International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences, 39, B7.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Hühn, J., & Hüllermeier, E. (2009). FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, 19(3), 293-319.

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). Foundations of rule learning. Springer Science & Business Media.

Il portale dei dati aperti della Regione Abruzzo http://opendata.regione.abruzzo.it/ [Accessed November 2017].

Lee, J., Jang, H., Yang, J., & Yu, K. (2017). Machine Learning Classification of Buildings for Map Generalization. ISPRS International Journal of Geo-Information, 6(10), 309.

McMaster, R.B.; Shea, K.S. Generalization in digital cartography. In Spatial Data Handling; Association of American Geographers: Washington, DC, USA, 1992; pp. 6.1–6.18.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Ruas, A., & Plazanet, C. (1996, August). Strategies for automated generalization. In Proceedings of 7th International Symposium on Spatial Data Handling (Vol. 1, No. 6).

Urbanowicz, R. J., & Moore, J. H. (2009). Learning classifier systems: a complete introduction, review, and roadmap. Journal of Artificial Evolution and Applications, 2009, 1.

Zhang, C. and Zhang, S., 2002. Association rule mining: models and algorithms. Springer-Verlag.

Wang, L., Guo, Q., Liu, Y., Sun, Y., & Wei, Z. (2017). Contextual Building Selection Based on a Genetic Algorithm in Map Generalization. ISPRS International Journal of Geo-Information, 6(9), 271.