SENSITIVITY ANALYSIS OF SUPPORT VECTOR MACHINE LAND USE CHANGE MODELLING METHOD

Mileva Samardžić-Petrović¹, Branislav Bajat¹, Miloš Kovačević¹, Suzana Dragićević²

¹ Faculty of Civil Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
² Spatial Analysis and Modeling Laboratory, Department of Geography, Simon Fraser University, 8888 University Dr., Burnaby, B.C., V5A1S6, Canada

This study was supported by an award from the Ministry of Education, Science and Technological Development of the Republic of Serbia (Project No. III 47014) to the first three authors, and by a Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant awarded to the fourth author.
Support Vector Machine

General binary classification case:

• Binary training set \((x_i, y_i), x_i \in \mathbb{R}^n, y_i = \{-1, 1\}, i = 1, ..., m\)

• Hyper-plane \(h\) that best separates the samples of binary classes \(h: wx + b = 0\)

• The best separation: \(\max M = 2 \|w\|^{-1}\)

**Radial Basis Function (RBF):**

\[
\begin{align*}
  k(x, y) &= \exp\left(-\gamma \|x - y\|^2\right) \\
  f(x) &= \text{sgn} \sum_{i=1}^{m} \alpha_i y_i k(x_i \cdot x) + b^*
\end{align*}
\]

**Soft margin:**

\[
\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i} \varepsilon_i, \text{ w.r.t: } 1 - \varepsilon_i - y_i (w \cdot x + b) \leq 0, -\varepsilon_i \leq 0, i = 1, ..., m
\]

**Optimal combination of:**

- \(C\) - the penalty for misclassified training points
- \(\gamma\) - control the radius of influence of each support vector
Finding a subset of most informative attributes

- Improving accuracy of the model,
- Reducing model complexity,
- Reducing the time required to train (learn).

Used methods for attribute selection

- Info Gain - IG,
- Gain Ratio - GR,
- Correlation-based Feature Subset - CFS
Study area:

- Covered approximately 17.5 x 8 km
- Population of about 161,000 - Census 2011
Data representation:

- Three orthophoto images for the years 2003, 2007 and 2011
- Maps of actual LU classes for the years 2003 and 2010
- Population census data for the years 2003 and 2011

Study area - rectangular cell grid of 10m spatial resolution.

Each grid cell - n-dimensional real vector $x^t (x^t = <x_1^t, x_2^t, \ldots, x_n^t>)$

Coordinate $x_i^t$ - the value of the i-th attribute associated with the cell $x$, at a particular time $t$. 

ArcGIS, ESRI 2011
Training and test datasets:

Model building - based on data from 2003 and 2007

- Model building:
  - Training dataset
  - $x^{t-1} \rightarrow y^t$
  - $y^t$ actual land use class $t$
  - $x^{2003} \rightarrow y^{2007}$

Model validation based on data 2007 and 2011

- Model validation:
  - Test dataset
  - $x^t \rightarrow y^t$
  - $y^{t+1}$ actual land use class at $t+1$
  - $x^{2007} \rightarrow y^{2011}$
  - Validation
  - $y_{t+1}^p$ predicted land use class at $t+1$
## Attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>ed. City centre Euclidean distance of grid cell to Belgrade centre</td>
</tr>
<tr>
<td>$x_2$</td>
<td>ed. Centre municipality Euclidean distance of grid cell to municipality centre</td>
</tr>
<tr>
<td>$x_3$</td>
<td>ed. River Euclidean distance of grid cell to the closest rivers (Danube and Sava)</td>
</tr>
<tr>
<td>$x_4$</td>
<td>ed. Green Euclidean distance of grid cell to the closest big green areas (100×100 m)</td>
</tr>
<tr>
<td>$x_5$</td>
<td>ed. Railway Euclidean distance of grid cell to the closest railway lines at time $t-1$</td>
</tr>
<tr>
<td>$x_6$</td>
<td>ed. Highway Euclidean distance of grid cell to the closest highway at time $t-1$</td>
</tr>
<tr>
<td>$x_7$</td>
<td>ed. Main road Euclidean distance of grid cell to the closest main road at time $t-1$</td>
</tr>
<tr>
<td>$x_8$</td>
<td>ed. str. I category Euclidean distance of grid cell to the closest street I category at time $t-1$</td>
</tr>
<tr>
<td>$x_9$</td>
<td>ed. str. II category Euclidean distance of grid cell to the closest street II category at time $t-1$</td>
</tr>
<tr>
<td>$x_{10}$</td>
<td>PCI Population Change Index between two census 2003 and 2011</td>
</tr>
<tr>
<td>$x_{11}$</td>
<td>Class $(t-1)$ Land use class at time $t-1$</td>
</tr>
<tr>
<td>$y$</td>
<td>Class $(t)$ Land use class at time $t$</td>
</tr>
</tbody>
</table>
Data Sampling

Changes in land use from 2003-2011 – 3%

- The training and test datasets contained all changed and an equal number of unchanged cells that were uniformly distributed over the area, thereby preserving original distribution over the classes.

This balanced dataset would produce a more informative and realistic dataset for
  - model creation - with less bias towards the majority class,
  - model evaluation.
Attribute Selection

Weka, software
## Attribute Selection

<table>
<thead>
<tr>
<th>$S_{CFS}$</th>
<th>$S_{IG}$</th>
<th>$S_{GR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class ($t-1$)</td>
<td>Class ($t-1$)</td>
<td>Class ($t-1$)</td>
</tr>
<tr>
<td>PCI</td>
<td>ed. Highway</td>
<td>PCI</td>
</tr>
<tr>
<td>ed. Highway</td>
<td>ed. City centre</td>
<td>ed. Highway</td>
</tr>
<tr>
<td>ed. Main road</td>
<td>ed. str. I category</td>
<td>ed. City centre</td>
</tr>
<tr>
<td>ed. Green</td>
<td>ed. Green</td>
<td>ed. Main road</td>
</tr>
</tbody>
</table>
Sensitivity of model on SVM parameters and selected attributes

Weka, software
Sensitivity of model on SVM parameters and selected attributes

Ground true

Predicted

Real land use class

SVM model outputs of land use class

Land use classes:
Agricultural
Wetlands
Traffic areas
Infrastructure
Residential
Commercial
Industry
Special use
Green areas

Land use change form 2007 to 2011

GIScience 2014, Vienna - September, 23-26
Conclusion and future works:

**Conclusion:**
- Using selected attributes by the CFS and GR methods resulted in a:
  - simple SVM model with better performance
  - less possibility to be overfitted with higher values of parameters

**Future works:**
- Explore sensitivity analysis using datasets covering different study areas and with different land use change dynamics
- The $k$ number of selected attributes was not investigated as the optimal number of attributes selected by IG and GR
THANK YOU FOR YOUR ATTENTION