Propagation of Uncertainty for Volunteered Geographic Information in Machine Learning

Jin Xing  
Centre for Research in Geomatics, Laval University, Quebec City, Canada  
jin.xing.1@ulaval.ca  
https://orcid.org/0000-0001-5693-3414

Renee E. Sieber  
Department of Geography, McGill University, Montreal, Canada  
renee.sieber@mcgill.ca

Abstract

Although crowdsourcing drives much of the interest in Machine Learning (ML) in Geographic Information Science (GIScience), the impact of uncertainty of Volunteered Geographic Information (VGI) on ML has been insufficiently studied. This significantly hampers the application of ML in GIScience. In this paper, we briefly delineate five common stages of employing VGI in ML processes, introduce some examples, and then describe propagation of uncertainty of VGI.

2012 ACM Subject Classification Information systems → Uncertainty

Keywords and phrases Uncertainty, Machine Learning, Volunteered Geographic Information, Uncertainty Propagation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.66

Category Short Paper

1 Background of VGI in Machine Learning

Machine Learning (ML) represents a set of methods that automatically learn from “experience” or training data with respect to given tasks. The learning can be implemented via a large body of models and algorithms, such as heuristic rules [32], decision trees [27], and cellular automata [31]. In Geographic Information Science (GIScience), ML has attracted considerable interest due to its wide applications in place recognition [34], ecology models [25], remote sensing image classification [33], transportation pattern discovery [22], and gazetteer analysis [9]. The rapid grow of ML has intensified due to the increasing ‘bigness’ of geospatial data, which describes the exalod of geographic information at unprecedented volume, velocity, and variety, as well as challenges to veracity.

Among the diverse sources of big data, Volunteered Geographic Information (VGI) is considered a main provider of input data/services [12]. For example, OpenStreetMap OSM, in which individuals have crowdsourced editable web mapping services and content, has become a powerful platform for building, training, and evaluating ML algorithms and models in GIScience [15]. VGI describes the process of obtaining geographic data or services (e.g., rating accuracy of feature labels) from large groups of users in an open call that is self-organizing via the Internet [10]. Uncertainty is innate within VGI, which means data is noisy, containing redundancies, irrelevant content, errors and biases contributed by users, who are often non-experts [26]. VGI also is disorderly, in which data may be unstructured, incorrectly ordered, mis-formatted (e.g., lacking a header), and possibly poorly geo-registered. Finally, users may be unreliable in providing consistent input and inputting within the appropriate
Table 1 Uncertainty Issues in Applying VGI for ML

<table>
<thead>
<tr>
<th>ML Process</th>
<th>Uncertainty Type</th>
<th>Examples in VGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection, Annotation, and Cleansing</td>
<td>Data Uncertainty</td>
<td>Inaccurate geolocation; spatial unevenness in data contributions; redundancies; gender, culture, and race bias in training data</td>
</tr>
<tr>
<td>Data Distribution</td>
<td>Operation Uncertainty</td>
<td>Boundary Vagueness (e.g., artificial boundaries introduced by data splitting); aggregation errors (e.g., heaping error in determining the existence of a traffic jam, binning of VGI point data)</td>
</tr>
<tr>
<td>Feature/Topic Detection</td>
<td>Representation Uncertainty</td>
<td>Interpreting location from place (from a well-defined to a poorly defined object)</td>
</tr>
<tr>
<td>Model/Algorithm Selection and Training</td>
<td>Decision Uncertainty</td>
<td>Simpler/alternate models than ML may be better like linear regression</td>
</tr>
<tr>
<td>Evaluation and Tuning</td>
<td>Service Uncertainty</td>
<td>Biased classification; Inconsistency in grading</td>
</tr>
</tbody>
</table>

Previous work in VGI’s uncertainty largely concentrates on the data quality. Researchers focused, for example, on uncertainty regarding the non-expert (e.g., skill levels and motivation), the thematic diversity of input (scattered focus relative to analysis needs), and the spatial unevenness of contributions (e.g., popularity of places relative to others) [11]. In ML, VGI is viewed primarily for its ability to provide data for ML, either as training data or general input data. It also has been employed for result evaluation and tuning of ML [18]. A worrying trend in GIScience inquiry into ML is its treatment as a big black box, where issues of data uncertainty are treated as I/O problems. We break down the black box of ML into a collection of workflow processes to briefly identify uncertainty from VGI that can occur within the ML as well as in its parameterization and refinement.

Other taxonomies tend to focus on classifying ML methods (e.g., supervised, unsupervised, and reinforcement learning) and application areas (e.g., computer vision, natural language processing, and speech recognition) [16]. The importance of uncertainty and its propagation have not been highlighted. We view the interaction between VGI and ML as five stages throughout the processing of VGI: data collection and cleaning, data distribution, feature/topic detection, model/algorithm selection and training, and evaluation and tuning.

2 A General Framework for Integrating Geospatial Crowdsourcing and ML

Our framework (Table 1) follows the standard ML workflow (data collection and cleaning, splitting of training from testing data, model training, evaluation, parameter tuning) [28] and adds components from big data handling [21] and ML computation [4] for de-/re-composition. Since the five stages may occur iteratively (e.g., the evaluation result could be fed back to the training process to improve accuracy), uncertainty also can propagate if we fail to attend to the origin of the uncertainty.
2.1 Data Collection and Cleaning

The primary utility of VGI in ML is for training and, more generally, input data. Training refers to data used by ML to calculate its parameters/weights so that input data generates expected outputs. Geospatial content is available across a wide range of VGI. It can be raster (landscape photographs) and vector (social checkins, binned aggregations of points); structured (Twitter metadata) and unstructured (Twitter text), explicit (x,y’s, placenames in hashtags) and implicit (colloquial names for neighborhood), absolute (latitude/longitude) and relative (concepts of home), passive (geo-fencing) and active (Amazon Mechanical Turk-AMT). It can be static or dynamic (harvesting of Flickr geotags at point in time or movement data), compensated or voluntary (AMT or VGI) [19]. Considerable research has been conducts to assess uncertainty with various VGI (cf., [14]).

Like other crowdsourced content, VGI data contains considerable error, vagueness, and ambiguity, and is vulnerable to malicious contributions (e.g., via GPS spoofing). As suggested above, this is the richest area of current research so this section is admittedly brief. Most research on the negative impact of ML focuses on the issue of algorithmic bias due to input data [26]. Location often serves as a proxy for race so one needs to debias on the basis of primary variable as well as data which functions as its surrogate [1]. Often dibiasing requires human intervention (cf., gendered word2vec example in [2]) so this stage also can utilize crowdsourcing. Geographic unevenness in data contributions can further distort ML output, for example the low OSM participation in Africa or the differential accuracy of OSM in urban areas versus rural regions [29]. Privacy protections, like the EU’s General Data Protection Regulation, will increase distortions in VGI as whole swaths of data are removed or masked [6]. Lastly, much of VGI is streamed, which requires new sampling techniques (e.g., reservoir sampling) to normalize temporal spikes or redundancies.

2.2 Data Distribution

The attraction of VGI to ML is both in its source (geosocial media) and its potential as big data. The latter likely requires de-/re-composition to distribute the computing. Data distribution may suffer from disorder in VGI because geographic data has its own internal topology and geometry that can be destroyed by arbitrary decomposition or splitting. For example, rectangular decomposition can distort the boundary of geographic objects and increase output uncertainty [5]. Most VGI is point-based and may need to be binned. A more sophisticated feature type, a polygon like a hexagon, does not easily alleviate the problem and any aggregation is subject to modifiable areal unit problems [24] that can alter ML output.

ML can be employed to reduce uncertainty in data distribution. Felzenszwalb et al. [7] employed latent support vector machine to decompose the original raster data into multiple object-based rectangles to lower boundary distortions. Temporal disorder in VGI, such as burstiness of reporting of natural disasters, could be addressed by decomposition with parallel processing.

2.3 Feature/Topic Detection

ML is designed in large part to recognize patterns, generate rules, approximate functions, and classify data sets. An important use of VGI in ML can be for feature or topic detection (e.g., forest, alternate route to avoid traffic jam). We lack explicit control over the feature representation in VGI. Users may not provide feature identification as planned or neural networks may fail to extract useful features from noisy VGI. For example, uncertainty in
placename makes it difficult to infer locations; “downtown nearby” could be interpreted as multiple locations [8]. Although iterative feature/object detection in ML can reduce uncertainty, there is no easy way to clean data to better disambiguate place to a location and location to a place. This resembles the challenge of NLP regarding semantic modeling to disambiguate slang (e.g., “bad”, “hot”, “sick”) in ML. Aggregation (pattern detection) is a likely outcome of ML that is based on VGI and therefore is subject to Sorites paradox and modifiable areal unit problems here as well (e.g., how many cars constitute a jam; how many trees constitute a forest).

The temptation for users new to ML is to treat it as a blackbox, an algorithm amongst many in a software library. Treating ML as a black box means that ML cannot necessarily accommodate the geography of VGI. For example, max pooling, which is a widely used method to pass features from one layer of neural network to another, is considered problematic in convolutional neural network by Sabour et al. [30] because max pooling lacks topology. In another example, a word embedding algorithm may produce very different vectors to represent “pub” and “bar” due to the surrounding content, which may then require multiple detection iterations.

2.4 Model/Algorithm Selection and Training

Which ML model or algorithm achieves the highest accuracy with a given input dataset and features? What is the best way to calculate the weights or parameters of the ML model/algorithm? Should we rely on a single ML model/algorithm or combine several ones together? These questions are difficult in ML and there are no clear answers. VGI can potentially assist this selection process with existing knowledge about model/algorithm selection and training strategies (think a wiki of appropriate ML) [23]. However, knowledge contributed via VGI may be unreliable because of a “follow the crowd” mentality with little investigation into alternate approaches [17]. Deep neural network is increasingly popular in ML research but a linear regression may be more appropriate, considering the quality of the data at hand and the ease of an ML implementation.

2.5 Evaluation and Tuning

Performance of ML algorithms needs to be evaluated with datasets different from the training process. VGI plays a pivotal role in collecting evaluation datasets and crowdsourcing can play a role in the evaluation process. To avoid overfitting (i.e., model is too closely fitted to the training data), ML scientists usually employ cross-validation, which can reduce the influence of uncertainty from VGI training data. Evaluation can be conducted with crowdsourcing services, such as the translation validation within the Google Translate Community [20] or Captcha [3]. Here, issues similar to data collection re-emerge, with potential biases introduced by the evaluators, who may be drawn from a particular gender, race, class, or skill level. These issues resemble the social approach to assessing spatial data accuracy in [13], in which the focus shifts from the uncertainty of the contribution to that of the contributor. One may wish to implement ranking or rating systems to improve confidence in the validators.

3 Propagation of Uncertainty in ML and Conclusion

In this paper, we propose a general framework to explore VGI uncertainty in ML. This includes the concrete importance of VGI for training data as well as the use of crowdsourcing for model/algorithm selection and performance evaluation in ML.
Uncertainty also can propagate across the ML workflow. Uncertainty in data collection can make data distribution more difficult because we do not know the appropriate aggregation size or scale. Without adequate cleaning, noisy data can generate messy features or false positives that will invalidate the chosen ML models and algorithms. Crowdsourcers bring their own bias to the evaluation of ML, which can influence the training of ML for parameter tuning. Disagreements during the cross validations may generate inconsistency in iterations of ML and force us to re-run the process. Where possible, it is critical to identify uncertainty at each stage to minimize the propagation of uncertainty. However, the cost (e.g., human intervention) of reducing the uncertainty in the early stages of ML (e.g., data collection and cleaning) is generally less than later stages (e.g., evaluation and tuning), so it is useful for us to consider at which stages it is appropriate to insert geographic crowdsourcing and crowdsourcers.

References


Amelia McNamara and Aran Lunzer. Exploring the effects of spatial aggregation, 2016.


