Statistical Machine Learning from VGI

Alexei Pozdnoukhov Fergal Walsh Christian Kaiser

National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland

1. Rational

The nature, volumes, and inherent uncertainty of volunteered geographic data complicate the use of data models in parametric statistical modelling. VGI stream typically is a result of human interpretations, classifications or matters-of-opinion for which there is no objective ground truth. It is hard to propose a reasonable model for data as it brings the belief that a statistician, by imagination and by looking at the low-level human-generated data, can invent a reasonably good parametric class of models for a complex mechanism behind it. While there is no doubt in usefulness of spatial statistics in understanding the aggregated geospatial data, low-level data modelling and processing requires adequate tools such as those of machine learning. Machine learning shifts focus from data models to the properties of robust non-parametric non-linear algorithms able to learn and predict from data.

This discussion paper reviews machine learning methods and tries to give a typology of available tools which can advance the use of VGI in modern geographic analysis. It overviews typical data sources and related tasks and further investigates the algorithmic requirements and implementation issues that researchers in GI science have to deal with aiming at efficient use of VGI.

2. Machine learning: generative and discriminative

Discriminative machine learning methods aim at finding a function $y=f(x)$, where input features (covariables) are vectors $x\in\Re^N$ and outputs depend on the type of problem, for example, $y\in\Re$ for regression, $y\in[1, 2, ... M]$ for M-class pattern recognition or $y\in\{-1,1\}$ for binary classification. The estimate has to be made given only empirical examples $\{(x_1, y_1), ... (x_N, y_N)\}$ of the mapping performed by an unknown mechanism which one aims to learn. In practice, a learning machine is an algorithm which from a given set of functions $F = \{f(x,\alpha), \alpha\in\Lambda\}$ selects the one which best (in some predefined sense) approximates the unknown dependency. In absence of pre-defined outputs, the problem can be stated as one of clustering (extracting similar groups) or dimensionality reduction (finding a lower-dimensional representation of data useful for further analysis or visualization). Unlabelled data, if available, can be used to improve model performance for a supervised task. This setting is known as semi-supervised learning. Generative methods of machine learning build probabilistic models of observed phenomena with tools like Gaussian Mixture Models, Hidden Markov models, Conditional Random Fields and Bayesian networks. Machine learning provides a useful toolbox in case if VGI can be encoded into feature vectors and a problem of interest is to find a useful mapping from this feature space into desired output.

2.1 Data streams and feature selection

In this discussion we focus on algorithmic learning from communication logs (e-mail, instant messenger, land lines and cell phones) and web streams of short messages available from services like Twitter, recently opened Yahoo! Firehose, etc. A bottom-up nature of VGI implicates that methods dealing with ordinal and categorical or even Boolean data will find wider application then those focused on continuous variables. It is quite rare that one meets “the temperature at (44.82, -76.34) this morning at 6:45am was -15.65°C” as a typical VGI-type measurement is rather “freezing cold as I was heading to the office this morning”. Natural Language Processing is hence an indispensable step in identification of semantic concepts related to the modelling task at hand, with a subtask of Named Entity Recognition being of particular interest for extracting geo-locations. There is a number of open source and/or freely available for academic use NLP packages such as LingPipe. A principle use of NLP is to extract location, time, movement attributes and to produce a set of vector features to encode textual content. We therefore distinguish basic data entries as Location, Timestamp, Link and Content and list typical tasks one encounters, group of methods to look at and typical data sources to consider as case studies:

Loc	Time	Link	Cont	Typical Tasks	Methods	Data
Y				Find and characterize spatial clusters	Point pattern analysis Scan Statistics	Geo-referenced internet search queries
	Y			Forecast behaviour Profile activities	Time series analysis HMM, graphical models	User activity logs
		Y		Network analysis and visualization, Community detection	Descriptive statistics, embedding and graph partitioning	Social networks
			Y	Find geographic names, localize, extract movement attributes, encode content	Natural Language Processing Named Entity Recognition	Web pages, blogs, tweets

Y	Y			Find clusters, describe evolution, identify typical routines	Point pattern analysis, clustering Space-Time Scan Statistics	Wearable sensors (GPS, cell phones, laptops)
	Y	Y		Dynamic network analysis, community evolution	Dynamic graph partitioning, Generative models, CRF	Cell phone, blog communities
		Y	Y	Visualization Community detection	Embedding, Multi-Dimensional Scaling	Blogs, Tweets, Web pages
Y		Y		Spatial clustering Spatial interaction modelling	Dynamic graph partitioning,	Cell phone, blogs, Twitter
Y			Y	Topic spread detection Geo-targeted marketing	Clustering, Classification, location-allocation models	Web search queries, twitter, blogs
	Y		Y	Interests evolution	NLP, HMM, Graphical models	Blogs, search queries
Y	Y	Y		Spatial patterns in social networks	Spatial clustering and graph part.	Cell phone, Twitter
	Y	Y	Y	General ICT research, WWW	General ICT research	WWW
Y		Y	Y	Spatial communities with common interests	Clustering, graph partitioning	Blogs, Twitter
Y	Y		Y	Clusters and evolution of interest focus	Clustering, Space-Time ScanStat	Twitter
Y	Y	Y	Y	Any Information Retrieval tasks	All of the above	Twitter

3. Software frameworks and implementations

As a software concept, current Geographic Information Systems rapidly approach a saturation point: it is a matter of time as we will digitize our static environments and speed up raster processing to keep regular updates. Industries have demonstrated their advances on this way with success of such products as Google Earth, MS Virtual Earth and recently launched arcgis.com. Wide-scale web crawling is also a domain rapidly turning into commercial services such as spinn3r.com which offers API to their pre-processed and indexed data of internet blogs.

Having said that, it is very noticeable that state-of-the-art in academic research in VGI analysis and usage mainly concerns basic data visualization. The mainstream activity is probably a simple point pattern analysis of VGI data which have not gone far beyond famous cholera investigation by Dr. J. Snow in London in 1856. Very few examples in academic domain exist where technical implementation issues have been resolved and a full operational chain of more sophisticated methods involving crawling, NLP, filtering and outlier detection, clustering, classification and user-oriented visualization are implemented and can be used to pursue further developments. It is clear that academic research has to identify fundamental problems, recognize bottlenecks and concentrate on providing better methods for this chain, considering future GIS as knowledge extraction and decision-support platforms.

3.1. Map-Reduce for Machine Learning models

VGI will increase in volume dramatically as more and more users gain access to internet and join spatially and temporally traceable services. Computational approach to handle these streams will focus on using multi-core and cloud computing paradigms. To take advantage of the potential speed up and gain industrial interest to methods being developed in GI community, one has to keep in mind if parallel programming is applicable to implement proposed algorithms. One can expect higher interest to algorithms that fit the statistical query model. Algorithms that calculate sufficient statistics or gradients fit this model if these calculations can be batched and expressed as a sum over data points to allow them to be easily parallelized on multi-core computers. Hopefully, a Google map-reduce paradigm is applicable to a variety of learning and spatial statistics algorithms including locally weighted linear regression, k-means, logistic regression, naive Bayes classifiers, a form of linear support vector machines, ICA and PCA, discriminant analysis, expectation-maximization methods, and certain back-propagation neural networks.

For example, in locally weighted linear regression it is required to find a solution of the normal equations $A\mu = b$, where $A = \sum w_i (x_i x_i^T)$ and $b = \sum w_i (x_i y_i)$. If weights w_i are chosen to be decaying with geographical distance from regression points this is known as Geographically Weighted Regression for which a comprehensive statistical inference framework is elaborated. For the required summation form, one divides the computation among different mappers of two sets to compute A and b. Two reducers respectively sum up the partial values for A and b, and the algorithm finally computes the solution via distributed numerical linear algebra. Focus on geographical localities appears to be a further advantage.

4. Conclusions

As industries open access to real-time VGI data, build services to process vast amounts of texts from raw web logs and provide APIs to stimulate research, GI community has to be aware of current data handling and processing paradigms to provide adequate algorithms and make VGI data “speak for themselves”. It is timely to turn our attention to transform GIS from data handling to knowledge extraction tool. In this talk we will share our experience of working with VGI data, processing communication logs from spatially aggregated instant messenger service and geo-referenced nation-wide cell phone records of more than million users and present software solutions we exploit and develop for this research.