Who Evaluates the Evaluator? Reconsidering Validation of Classification Processes under Big Data

Samuel Stehle

Pennsylvania State University

Keywords: classification, evaluation, interestingness, spatio-temporal applications

Despite well-calibrated methods for discovering and analyzing patterns in big collections of data, much of the evaluation of these processes is an underappreciated but necessary step toward furthering the field methodologically. This short position paper encourages careful consideration of the veracity of classification algorithms and other processes critical to making sense of big data, and seeks to direct these efforts away from current approaches of 'validation' toward a strategy of 'evaluation.' It contends that data-first methodologies and the resulting use of exploratory procedures for big data necessitates that we consider evaluation an iterative process and one that does not rely on a single optimal solution, as is implied by the use 'validation' methods.

There is truth in the contentions that big data is not a new concept to geographers and spatial methods. Spatial data is high-dimensional, and data volume has always threatened the limit of software capabilities for storage and analysis. However, the discipline of geography, and particularly GIScience, has undergone many methodological changes attributable to big data, from cloud computing (Shekhar, Gunturi, Evans, & Yang, 2012) to greater emphasis on interdisciplinary collaboration (Kitchin, 2014). Here, I consider a suite of methods that are increasingly important to spatial applications of big data concerning automatic classification and clustering. Such methods are specifically important for exploratory data analysis given their unsupervised nature. Unsupervised classification algorithms are one subset of methods important for data analysis, though their untrained nature makes for a perfect example of how an underconsidered evaluation strategy can be problematic and counter productive.

Following from Färber et al, (2010) more attention should be paid to the process of evaluating classification algorithms. There are two reasons for using classification on a collection of big data. First, as mentioned before, is for exploratory purposes. Although one never begins data analysis without a theoretical idea of what insight is contained in the information, an exploratory analysis seeks to determine some of the themes taking place. The second seeks to compare new insight with expected insight established through previous exploratory analysis or theory. Even with the data-driven emphasis of big data, theory remains a significant driver of expectation and successful analysis. It is this expectation that is often used to evaluate classification processes, creating a gold-standard solution to compare against.

Färber et al (2010) dispel the veneration of the gold standard method against which results are evaluated. As long as "the whole point in performing unsupervised methods in data mining is to find previously unknown knowledge," (Färber et al., 2010, p. 3) then comparing against a known result is not helpful. Classification methods themselves have been tested and become common tools in areas from text modeling (Blei, Ng, & Jordan, 2003) to image analysis (Pal, 2005), so as we apply them to spatial data, the goal is increased insight, rather than additional proof in the function of a method. The evaluation of such methods should be optimized to be critical of the insights they provide, rather than the ability of the method to provide a fully expected result.

Additionally, classification methods may not be designed to produce a single optimal solution (Färber et al., 2010). Multiple parameterizations, random components, and partial cluster assignments make evaluation a complex and imperfect task. Although some measures, such as maximum likelihood

may be used to compare sets of outputs, quantitative measures of cluster validity seek to ensure that clusters diverge from one another at a maximum potential. Many methods allow for classified data to exist proportionally in more than one cluster, representing a legitimate convergence of clusters, rather than a measurable divergence. With near infinite possible parameter combinations for many methods, it is clear that some evaluation process is necessary to ensure correct inputs and justify insightful results.

I provide two potential measures of the effectiveness of classification. First is a series of evaluation measures meant to consider the 'interestingness' of patterns derived from data analysis. Ranging from the quantitative to the subjective, measures such as *unexpectedness, actionability,* (Silberschatz & Tuzhilin, 1996), *conciseness, generality, novelty,* and *diversity* (Guo, 2003) are offered. *Conciseness* and *generality* can be implemented as rules in an iterative analysis, while *actionability* and novelty require subjective input. Interesting/insightful results need not satisfy all of these measures, but explicit consideration of some in both their objective and subjective interpretations would benefit the classification process by providing substantial interrogation of classes and assignments.

Second, and finally, I suggest the use of spatio-temporal context which grounds our geographic insight in the real world. Because an unexpected or actionable pattern may exist in one place or time, but not exhibit those interesting characteristics in another, spatio-temporal context and comparison is crucial for evaluation. Big data encourages this exploration, as spatial and temporal data slices should be assumed to contain potentially wide-ranging results with varying adherence to established theories.

Insight comes from a combination of both realistic and unexpected results. Increasingly in big data applications, we seek new explorations and insight that is impossible to glean from targeted samples and established standards for comparison. This position paper more than anything hopes to reevaluate how we evaluate the insights under big data processes, specifically reducing the inclination to 'validate' them against some standard solution. Evaluation is an iterative process of quantified and subjective consideration, requiring contributions from spatio-temporal and big data applications.

Acknowledgements

Thank you to the panelists and audience members of the Spatial Data Mining and Big Data Analytics panel at the Annual Meeting of the Association of American Geographers 2016 for encouraging and contributing to conversation on the topic of evaluation and big data.

References:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Bibliometrics, 3*, 993-1022.

Färber, I., Günnermann, S., Kriegel, H.-P., Kröger, P., Müller, E., Schubert, E., . . . Zimek, A. (2010). *On Using Class-Labels in Evaluation of Clusterings.* Paper presented at the Procedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington D.C.

Guo, D. (2003). Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization, 2*(4), 232-246.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1). doi:10.1177/2053951714528481

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217-222. doi:10.1080/01431160412331269698

Shekhar, S., Gunturi, V., Evans, M. R., & Yang, K. (2012). *Spatial big-data challenges intersecting mobility and cloud computing.* Paper presented at the MobiDE Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, New York, NY.

Silberschatz, A., & Tuzhilin, A. (1996). What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 970-974.