Data Mining, Data Fusion and Analysis of Massive Distributed Data Sets

Amy Braverman
Jet Propulsion Laboratory, California Institute of Technology

Modern data used in the study of the Earth's climate and environment are no longer monolithic, static entities. There was a time when they were: central data repositories held large collections of heterogenous data. Users typically searched to find information using tools available through a webpage, and downloaded what they found. Assembling the right data set for a unique scientific study could take weeks or months because of large data volume, different formats and other incompatible characteristics of the data. Today, many scientific communities have moved in the direction of decentralization. There are many data providers, and web services that allow remotely callable processes are commonplace. This creates both new challenges and new opportunities for the practice of data analysis and for the development of statistical methods used therein. In particular, there are three general areas where this new landscape demands some new thinking: data mining, data fusion, and analysis of distributed data (as distinct from distributed computing).

Data mining has been around for a long time. It's defined in various ways, and one convenient definition is the automated, exploratory data analysis on massive data sets. Data mining is different in different domains and application areas, but where very large remote sensing data sets are concerned the first order problem is simply to understand the content of those data sets. Techniques for this fall under the general heading of unsupervised learning; other data mining activities, such as searching for specific patterns and supervised learning already presuppose some knowledge about the data. Many unsupervised learning methods can be described as clustering algorithms, that attempt to partition the data into classes of similar items. The underlying motivation is to approximate the empirical distribution of the data so that this empirical estimate can studied as a denoised proxy for the original data. At JPL we have implemented this idea for two Earth observing missions including the Multi-angle Imaging SpectroRadiometer (MISR). MISR observes surface characteristics using multiple view angles and wavelengths and has been used to monitor the size and extent of inland water bodies. We "compress" very large remote sensing data sets by stratifying them in some sensible way (e.g., in space and time) and providing distribution estimates for all strata. The collection of stratum distribution estimates is a compact description of the contents of the data. By examining how these distributions change across strata, we may also gain insight into the explanatory effects of the stratification variables. The nominal objective of this approach is to make a smaller version of the data that is faithful to the original but more easily moved, accessed, and understood. The question for the modern era is how to do this when different components of the data live in different places. Can the method be performed on pieces and the compressed pieces combined? How does this depend on the computational capabilities of the nodes housing the data and the need for real-time interaction? Can the fidelity of different parts of the compressed data to their original counterpart be changed dynamically in response to users' needs?

Data fusion, like data mining, has a long history and a myriad of definitions arising from various applications. The most prominent application has historically been to image processing for military and intelligence purposes. Multiple images of the same scene can be combined to present a more complete picture than is possible from the individual images in isolation. Data mining can be applied to the fused image to identify features of interest (e.g., tanks, trucks, missile launchers). In science and decision support for policy making and disaster response, the fusion problem has mainly been addressed through the activities of individual scientists, and through tools like geographic information systems and image registration technologies, respectively. Popular methods include averaging to coarse common resolution, "nearest neighbor match-up" among pixels, and linear or bilinear interpolation to common locations. However, if the fused data are to be used for scientific inference or policy making, then taking account of uncertainties and quantifying them in the output is essential. For that reason, we take the position that data fusion is a statistical inference problem: estimate the true quantity (or quantities) of interest at a specified set of locations from all available data (not just observations precisely at the locations of interest), which may have different spatial and temporal resolutions and different statistical characteristics (e.g., biases, measurement errors). Every estimate must be accompanied by a standard error. At JPL we have implemented a methodology called Spatio-Temporal Data Fusion (STDF) for fusing data from two satellites to estimate CO2 concentrations in the lower atmosphere. The same technology could be applied to the fusion of multiple data sources to estimate snow and precipitation from satellite data, or to fuse satellite data with in-situ or other ground station information. The methodology lends itself to some parallelization: parts of the algorithm can be performed on input data sets independently and at the locations where they reside. To fully adapt the methodology to perform fusion of distributed data sets would require careful consideration of how data can be moved over the network in order to compute quantities requiring joint information.

Finally, the analysis of distributed data is a new research area. The previous discussions of data mining and data fusion hint at some of the fundamental issues, but those issues exist even in the context of much simpler forms of data analysis. Consider, for example, the problem of computing the covariance between two variables where one variable's data lives on one computer and the other's data lives on another computer. The covariance is $Cov(x, y) = N^{-1}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$. $\bar{x}$ and $\bar{y}$ can each be computed on the machine housing their data, but $\sum_{i=1}^{N} x_i y_i$ requires that the x and y data values be brought together somehow. It may be most efficient to move x to the y computer, or visa versa, or to do this in pieces; move the data to common location in batches and perform the calculation in pieces, assembling the final cross-product sum at the end. Which of these options makes the most sense will likely depend on dynamic characteristics of the network and users' changing needs. If the system architecture is given, then the problem is to optimize the analysis for that architecture. On the other hand, if the system is yet to be designed, then what design is most efficient given a set of data analysis objectives? To what degree might users be willing to accept approximations to their calculations in exchange for greater speed and efficiency? Can that trade-off be quantified? Is it generally possible to define a canonical set of data analysis functions that can be assembled through work-flows to carry out more complex analyses, or would any such set be too restrictive?

These are some of the challenges I see in the analysis of massive remote sensing data sets for climate and environmental science. I imagine a future data analysis system in which a scientist working inside their favorite language or package issues remote web service calls to dynamically assemble and analyze a virtual data set brought together in real time. The same infrastructure with a web page front end could provide decision makers with a set of customized tools for visualizing and assessing uncertainties in virtual data sets they themselves define. Perhaps some of these ideas can be adapted to be useful for aquatic ecosystems sustainability and management.