Statistics and Probability Letters 136 (2018) 4–9

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Statistics in the big data era: Failures of the machine

David B. Dunson

Department of Statistical Science, Duke University, United States

ARTICLE INFO

Article history:
Available online 17 February 2018

Keywords:
Deep learning
High-dimensional data
Machine learning
Scientific inference
Selection bias
Uncertainty quantification

ABSTRACT

There is vast interest in automated methods for complex data analysis. However, there is a lack of consideration of (1) interpretability, (2) uncertainty quantification, (3) applications with limited training data, and (4) selection bias. Statistical methods can achieve (1)-(4) with a change in focus.

1. Introduction

1.1. Different cultures

The culture and ways in which the statistical community thinks of analyzing and interpreting data have been rapidly evolving in recent years, with the machine learning and signal processing communities having a fundamental impact on the rate and direction of this evolution. To set the stage for this discussion article, it is helpful to first comment on the culture and background of the machine learning and statistical communities. These comments are meant to give a "cartoon" of a complex reality, with this cartoon helpful as a starting point for discussion.

Machine learning (ML) community: tends to have its roots in engineering, computer science, and to a certain extent neuroscience — growing out of artificial intelligence (AI). The main publication outlets tend to be peer-reviewed conference proceedings, such as *Neural Information Processing Systems (NIPS)*, and the style of research is very fast paced, trendy, and driven by performance metrics in prediction and related tasks. One measure of "trendiness" is the fact that there is a strong auto-correlation in the main focus areas that are represented in the papers accepted to NIPS and other top conferences. For example, in the past several years much of the focus has been on deep neural network methods. The ML community also has a tendency towards marketing and salesmanship, posting talks and papers on social media and attempting to sell their ideas to the broader public. This feature of the research seems to reflect a desire or tendency to want to monetize the algorithms in the near term, perhaps leading to a focus on industry problems over scientific problems, where the road to monetization is often much longer and less assured. ML marketing has been quite successful in recent years, and there is abundant interest and discussion in the general public about ML/AI, along with increasing success in start-ups and industrial sector high paying jobs partly fueled by the hype.

Statistical (Stats) community: made up predominantly of researchers who received their initial degree(s) in mathematics followed by graduate training in statistics. The main publication outlets are peer-reviewed journals, most of which have a long drawn out review process, and the style of research tends to be careful, slower paced, intellectual as opposed to primarily performance driven, emphasizing theoretical support (e.g., through asymptotic properties), under-stated, and conservative.

E-mail address: dunson@duke.edu.

https://doi.org/10.1016/j.spl.2018.02.028
0167-7152/

Statisticians tend to be reluctant to market their research, and their training tends to differ dramatically from that for most ML researchers. Statisticians usually have a mathematics base including multivariate calculus, linear algebra, differential equations, and real analysis. They then take several years of probability and statistics, including coverage of asymptotic theory, statistical sampling theory, hypothesis testing, experimental design, and many other areas. ML researchers coming out of Computer Science and Engineering have much less background in many of these areas, but have a stronger background in signal processing, computing (including not just programming but also an understanding of computer engineering and hardware), optimization, and computational complexity.

1.2. Canonical application areas

Partly due to the different backgrounds and skill sets, the ML and Stats communities tend to focus on somewhat different application areas. In general in ML, the focus is almost always on automatic processing of data without any focus on uncertainty quantification or hypothesis testing. For example, a canonical area of ML is signal processing — of images, videos and audio recordings. One may want to automatically label an image with the objects in that image, identify items (e.g., an abandoned suitcase) specific to certain frames of a video, label time segments of an audio recording according to who is speaking, or convert audio into text. Also, issues of compression of signals and reconstruction with de-noising are often of interest. There is also a large focus on tech industry applications, such as recommender systems that exploit user databases and limited user information to make recommendations for new products the user may like. Other such problems include building accurate search engines, placements of ads on websites, choice of website content to maximize click through rates of ads, etc. "Big data" in ML typically means that the number of examples (i.e. sample size) is very large — e.g., in the millions or more.

In statistics, there has been and continues to be more focus on application areas in which scientific inferences are a primary goal and the over-arching emphasis is not necessarily on building a black box for prediction or optimization of some utility function (e.g. based on revenue). For example, many statisticians work closely in collaborative projects with scientists, with a focus being on using data collected from new experiments or observational data sources to improve scientific understanding. My own ongoing primary collaborative areas include genomics, neurosciences, and ecology. In each of these areas, it has become common to collect high dimensional, complex and intricately structured data. Often the dimensionality of the data vastly exceeds the available sample size, and the fundamental challenge of the statistical analysis is obtaining new insights from these huge data, while maintaining reproducibility/replicability and reliability of the results. It is crucial to not over-state the results and appropriately characterize the (often immense) uncertainty to avoid flooding the scientific literature with false findings. Experimental replicability is broadly recognized as a key challenge for science today, but ML does not typically think of replicability in the scientific sense. In ML, replicability tends to mean that the code has been made available so that the results can be reproduced for the same examples, while for statisticians replicability means that another scientist can repeat the experiment, obtain new data, and reach the same scientific conclusions (meaning the statistical analysis results are similar).

1.3. Dangers of ML in non-ML applications

ML algorithms have clearly had dramatic success in a number of ML-style application areas, leading to understandable interest in the field and in the broader community in leveraging upon this success. Particularly notable is the performance of so-called deep learning methods, such as convolutional neural networks. These approaches define a highly complex and over-parameterized model that is built up in multiple layers, leading to the "deep" terminology. Key characteristics of deep neural nets include: (i) huge numbers of parameters; (ii) substantial flexibility; and (iii) a multiscale specification. There have been dramatic developments in recent years in algorithms for fitting of deep neural nets to very large data sets, leading to state-of-the-art and arguably transformative performance in certain machine learning tasks. Exploiting modern computational resources and TensorFlow, it has become straightforward to apply deep learning to a wide variety of data sets and settings. Deep learning is most suited to applications in which (a) data have a clear spatial and/or temporal structure (as in image/video/audio processing); (b) huge labeled data sets are available (e.g., human labels of images via Mechanical Turk); (c) the interest is in a black box for prediction and there is no interest in statistical inferences or uncertainty quantification.

In many ML application areas (a)–(c) hold, stimulating the excitement and hype related to *deep learning*. However, in the "statistics-style" application areas that I have spent my career working on, it instead tends to be the case that the data are complex without such a clear spatial or temporal structure as one obtains in signal processing. For example, I may have a very high-dimensional vector of biomarkers collected for each patient in a medical study, with these biomarkers including expression levels of different genes, single nucleotide polymorphisms (SNPs), demographic characteristics, etc. In addition, I never have the luxury of having millions of labeled observations; instead I would be lucky to have 1000 subjects and often I have more like 10 or 100. Finally, scientists are essentially never satisfied with a black box for prediction; they focus their studies on improving mechanistic understanding even if better prediction (e.g., of a medical outcome) is a key part of the goal. Hence, in sharp contrast to much of the hype, I find deep learning of no use whatsoever in the vast majority of application areas encountered in not only my own collaborative research but also that of most statisticians.

Of course, deep learning is only one particularly popular class of ML algorithms, and there are many other ML algorithms that are more useful in the types of applications that I tend to encounter. For example, there has been a huge emphasis in ML

and in the associated statistical literature on developing methods for dimensionality reduction through feature selection, learning of lower-dimensional structure in high-dimensional data, and low rank approximations among others. There has also been abundant interest in penalization methods that enable fitting of statistical models to high-dimensional data having insufficient sample size and/or labels. I can use many of these methods in the applications that I work on, *but* there is a huge danger in doing so due to the lack of uncertainty quantification and reproducibility of the results. If my collaborators were able to collect a new data set under similar conditions and I were to apply the same methods, I may obtain substantially different results. This lack of reproducibility may be an intrinsic characteristic of high-dimensional low to moderate sample size data, but it is particularly problematic when there is nothing in the statistical results we provide to our scientific collaborators to warn them of the problems. If uncertainty in inferences were appropriate characterized, the scientists may well decide that their inferences need to be made substantially less ambitious — e.g., through focusing on "coarser" scale hypotheses, a concept I will touch on later in this article.

From my perspective statistics is increasingly contaminated by ML-style analyses applied to stats-style data sets and problems; this is leading to a crisis in progress and understanding in scientific fields, as well as in policy making. In the subsequent sections of this short discussion paper, I will attempt to make some of the key issues clear through a couple of case studies (Section 2), a discussion of the role of UQ in scientific inferences (Section 3), comments on the crucial role of sampling and selection bias (Section 4), and a brief discussion (Section 5). I have excluded references due to space constraints.

2. Case studies

2.1. Neuroscience and brain networks

In my (biased) view, one of the most exciting areas of scientific advancement in recent years relates to the understanding of the brain. I have several collaborations that seek to better understand brain networks, and how these networks relate to variation among individuals in behavior and cognitive traits. This problem can be attacked from different angles — (i) mouse electrophysiology studies: an array of electrodes are inserted into different regions of the mouse brain & brain activity is recorded through a wireless device, while also recording behavior through a video; (ii) human brain connectomes: using diffusion tensor and structural MRI, one estimates the locations of white matter fiber tracts (acting as highways for neural activity and communications) in the brain for each individual in a study; for these same individuals, many different traits are measured.

These two application areas have a common feature with many "modern" scientific studies. In particular, the number of study subjects (n) is vastly smaller than the dimension of the data being measured (p). In addition, the data are complex and geometrically structured, and the choice of p is somewhat arbitrary. In particular, the measurement technology collects data at such a fine resolution that one would never use observations on the finest measurement scale directly in the statistical analysis, as this would lead to intractable computation and statistical problems. In mouse electrophysiology studies conducted at Duke, each experiment on each mouse collects data for millions of time points or more, but there are typically only $\sim$10 mice available. In brain connectome studies, we may have access to up to $\sim$1000 individuals, but it is intractable to statistically analyze the data on the individual voxel level due to (a) the impossibility of aligning different individuals' brains at that resolution level; and (b) the absurdly massive number of pairs of voxels.

Hence, in this and many other application areas, one can choose the *resolution* at which to analyze the data, and hence the value of p. An interesting question is how would the typical "modern" statistician go about analyzing these data? The primary scientific interest is in inferring how brain networks relate to outcome variables. A usual approach would be to apply pre-processing to reduce the rich geometrically structured data into a simpler form amenable to automatic analysis using off-the-shelf machine learning algorithms. In particular, we would like to reduce the data for individual i to a response variable y_i and a vector of predictors $x_i = (x_{i1}, \ldots, x_{ip})$. The response variable may correspond to a particular trait of interest (e.g., IQ) for individual i, while x_{ij} may consist of a binary indicator of any structural connections between the jth pair of brain regions, for $j = 1, \ldots, p = R(R-1)/2$, with R the number of regions the brain is segmented into. Then, one can use random forests or some other flexible regression/classification algorithm to predict y_i from x_i. Such a predictive black box is interesting (if accurate) but one also wants to obtain interpretability. Hence, it is important to also include variable selection — for example, Lasso or one of its many variants could be used to identify the pairs of brain regions whose connections relate to the response. Unfortunately, depending critically on the choice of R (and hence p), such an approach will often be quite unreliable — producing many errors in practice, and leading to a tendency to badly over-interpret the results. In addition, the analysis output does not warn the user of the lack of reliability and the substantial uncertainty in the results.

2.2. Fair decisions and predictive algorithms

There are numerous decisions that are made by various authorities based on their own judgment and experience; these decisions can have an enormous impact on society as a whole and on individuals. Some examples include whether and how to regulate car emissions, patrol locations and decisions of who to stop and search in policing, sentencing and bail decisions in the criminal justice system, hiring decisions, salary levels, selection of grants to fund, and tenure decisions in academics. Of course whenever there is an individual or a small group of individuals in charge of making such decisions, there is substantial room for the decisions to not be entirely "fair" and objectively based on the data at hand but instead

driven in part by implicit or explicit biases. Such biases may lead to under-regulation of pollution, policing that targets certain minority communities, more severe sentencing for individuals within those communities, and hiring/salary/tenure decisions driven in part by demographic factors.

There has been some thought that machine learning algorithms can replace or augment decisions made by a judge or other authority to improve the fairness of the decisions.

Unfortunately, off-the-shelf ML algorithms applied to existing data sources will inherit issues present in the data upon which they are trained. Hence, if the data are obtained through a biased measurement process, then the ML algorithm-based predictions will inherit those biases. For example, suppose that African American men commit no more crimes than Caucasian women but that police (a) assign their patrols predominately within African American communities; and (b) are significantly more likely to stop an African American man for questioning and/or search. Then, effectively, the police are over-sampling African American men, and unless this sampling bias is accounted for, any ML algorithm will predict that an African American man is much more likely to commit a crime. An "objective" ML-driven policing and sentencing strategy may then decide it is appropriate to target African American men, potentially even increasing the bias through a feedback loop. In general, issues of selection bias are hugely important and are almost always ignored in the ML literature. Selection bias is particularly problematic in large observational data sets, as the sampling process is often complex and unknown, and hence difficult to adjust for in the analysis.

3. Uncertainty quantification in scientific inferences

One of the key disadvantages of most ML methods, which also include approaches developed by statisticians and published in the statistical literature, is the inability to quantify uncertainty. It has become standard practice in high-dimensional data settings to focus on producing a point estimate – e.g., via solving an optimization problem, which incorporates a penalty to effectively reduce the dimensionality of the problem. There is an immense literature defining different types of penalties and efficient algorithms for optimization in an amazing variety of cases. Statistical articles proposing such approaches in leading journals tend to include asymptotic theory justifying the methodology. As opposed to traditional asymptotics, which let $n \to \infty$ while fixing the dimensionality p of the parameter of interest θ, modern asymptotics attempt to mimic the high-dimensional nature of the problem by letting $p \to \infty$ with n, potentially even at a faster rate. Under some serious restrictions (e.g., the truth is sparse, the design matrix is nearly orthogonal, the non-zero signals are large enough, etc.), it is often possible to provide positive asymptotic support in terms of (a) ability to find the true low dimensional structure (e.g., zero coefficients); and (b) accurately estimate the parameters.

Such a seemingly strong asymptotic justification can obscure the fact that the methodology being proposed (a) just produces a point estimate with no measure of uncertainty; and (b) is justified in finite samples only through some limited simulation study assessing error rate relative to other point estimation methods. Consider the human brain connectomics case study. If we apply such point estimation methods to obtain a single sparse point estimate without any notion of uncertainty, we and our scientist collaborators are almost fully in the dark in terms of how confident we should be in the results. Obtaining perfect results when the sample size goes to infinity under overly idealized conditions gives us very little reassurance in the small sample sizes we are faced with. In fact, we know given the huge statistical challenges that we likely have many false positives and negatives in our results. If we try simple heuristics, such as holding some observations out and repeating the analysis, we often obtain significantly different results. Unfortunately, the culture within the statistical community is overly focused on producing strong *positive* results even if this requires making unrealistic assumptions. It would be substantially more impactful to have theory that really attempts to describe positive or *negative* behavior depending on realistic science-based assumptions. If the problem is simply too ill-posed given the data at hand and the focus of inferences, then I for one would really like to know that I am attempting an impossible task — suggesting we must be less ambitious.

There is a small and growing literature seeking to address the lack of uncertainty quantification in high-dimensional inferences; for example, focused on penalized optimization methods, such as Lasso. There is also a Bayesian literature, which attempts to approximate the full posterior distribution quantifying uncertainty instead of simply producing a point estimate. However, the frequentist literature on uncertainty quantification in high-dimensional settings is still quite young and limited in scope, while current Bayesian methods have key unresolved issues — (i) it is difficult to scale up sampling methods, such as MCMC, to very high dimensions, while fast approximations to posterior distributions (e.g., variational methods) can badly under-estimate uncertainty; (ii) even scalable sampling methods may have considerable errors in approximating posterior summaries quantifying uncertainties of interest; and (iii) it is not at all clear how well the exact posterior under a Bayesian method for high-dimensional data actually does quantifying uncertainty — in order to get good performance (empirically and in terms of asymptotic guarantees), it is often necessary to employ strong priors (e.g., favoring very sparse values); such priors may lead to an overly concentrated posterior distribution that only provides reasonable UQ under unverifiable and strong assumptions about the true data generating model. Ideally, our priors would be chosen to *accurately* reflect the actual knowledge and science available in an application area — usual high-dimensional variable selection and/or shrinkage priors do not make much sense in this regard. We should avoid putting too much information in the prior even if this means the resulting posterior is too vague to distinguish between competing hypotheses of interest. It is a fact of life in high-dimensional settings that available data often will not be definitive — such "negative" results should be embraced and not hidden.

Given these issues, I would say that in the brain network applications and in many other areas (indeed most scientific areas), we currently lack the necessary tools to provide useful and reliable results to our scientist collaborators. We certainly do not want to only provide a predictive black box, along with an estimate of the important variables. Instead, we crucially need tools to tell us how reliable our variable selection decisions are given the sample size, dimensionality, and correlation structure of the data at hand. We need negative results that will tell us to be less ambitious about the types of inferences we are attempting — perhaps we simply cannot examine brain networks at too fine of a resolution given statistical limitations, and hence R should be chosen to be less than some bound. How do we choose this bound in a principled manner? We need accurate and interpretable measures of uncertainty in our results. We need tools to include more knowledge and structure into the analysis to improve performance. Often the data are not a matrix in their "native" form and by including more geometric constraints, and limiting ad hoc pre-processing, we may improve efficiency and insights. Often scientists know a lot about constraints that should be imposed other than simple sparsity, low rank and other black box assumptions. It is likely not possible in most scientific fields to make progress using generic machine learning methods that are agnostic to how the data are collected and to the background knowledge in the field. Instead we need carefully thought out and targeted statistical approaches developed for scientific applications.

4. Issues with sampling, selection bias and measurement error

There has been huge interest and hype around the potential of mining large data sets using ML methods to address many different types of problems. One big issue with such attempts is the selection and measurement process under which such large data sets are collected. Typically, in conducting statistical inferences, the focus is on estimating a particular parameter of interest (say θ), which represents a characteristic of some population P of interest. For example, θ may represent the proportion of obese individuals in the US population or some particular sub-group of interest, such as males aged 13–18. If we had a simple random sample $X = (X_1, \ldots, X_n)$ from P, with $X_i = 1$ indicating the ith individual in the sample is obese and $X_i = 0$ otherwise, then we could simply estimate θ using the sample average. However, "big" data sets are essentially never simple random samples, but are instead collected under some complex and unknown measurement process. For example, suppose instead of a simple random sample, you conduct a web survey in which individuals choosing to answer the survey indicate whether or not they are obese. Then, the average of the resulting sample may be very far from the true parameter of interest even if the sample size is enormous due to selection bias and measurement errors. Selection bias comes in because the individuals responding to the survey represent some population Q that may differ very substantially from the population of interest P in all sorts of factors including obesity. Perhaps Q has significantly greater proportions of high SES Caucasian males over 40. Measurement error comes in because individuals may not know their obesity status and/or may misreport their status. Without having some notion of how Q differs from P or the magnitude of measurement error, statistical estimators of θ based on the available sample may be completely flawed to the point of being useless. Having bigger data does not really solve this issue — it just *decreases the statistical uncertainty in estimating the wrong quantity* (in this case, an attribute of Q instead of P).

The issue of selection bias and measurement error in estimating a simple quantity such as the proportion of obese individuals is relatively obvious, but such issues are common broadly in much more complex settings. Consider, for example, the recent strong interest in using medical records data to improve health care practice and medical knowledge. The "old school" way to conduct a medical study comes in one of several flavors. The gold standard is the randomized clinical trial — in this case, there are a variety of treatments that can be given to a patient having some condition and patients are randomized to treatment groups. This randomization significantly reduces issues with selection bias and unmeasured confounding making treatment efficacy comparisons relatively straightforward. However, randomized clinical trials are very expensive and can only be conducted in specialized settings in which there is a new treatment of interest, but not yet evidence that the treatment is better than a previous treatment. Hence, it is more common to conduct observational epidemiology studies, with the most common design being the case-control study. Such studies collect data on a sample of cases (e.g., individuals having some disease or condition of interest) along with a set of controls that are chosen to be similar to the cases except they do not have the disease. The typical focus of inference is the exposure odds ratio obtained from a logistic regression model, adjusting for covariates. Epidemiologists tend to think carefully about the covariates to adjust for and try to limit the impact of unmeasured confounders. When the covariates have a very different distribution for cases and controls, propensity score matching or adjustment methods are often recommended.

In general, medical records data are automatically collected by a medical system, and it is typically very difficult to retrospectively ascertain based on the records the selection mechanism by which patients end up in the database. For example, I have been recently involved in projects focusing on analyzing medical records data collected at Duke including detailed monitoring information (e.g., on blood pressure) while the patient is in the operating room, along with information on the type of procedure and some limited additional information on the patient and their stay in the hospital. Suppose that the θ of interest is the increase in risk of morbidity following surgery for patients having a high A1C value relative to patients with a low A1C value, and that the population of interest P is all patients having non-cardiac elective surgeries. Estimating θ based on the available Duke medical records database, we encounter multiple challenging selection bias issues. These include that Duke is in general a referral hospital, and hence may not obtain an "average" selection of patients even from among our region of North Carolina. There may be less routine surgeries and a mix of patients having more extreme health conditions at Duke than other health centers. In addition, physicians are already applying some process in selecting

patients eligible for surgery. Patient with indicators of metabolic syndrome (high A1C is one indicator) may be recommended to delay surgery until symptoms of the syndrome are reduced. It is very difficult to properly think through and statistically adjust for all of these issues retrospectively using a database that has been collected over an extended period of time using a varying, complex and unknown selection process. The resulting inferences may be highly unreliable and biased.

My goal in presenting these issues is not to rule out the use of medical records and other big data sets (e.g., from the web) to conduct scientific inferences and inform policy. However, it is very important to keep in mind the enormous impact that selection bias and measurement errors can have on statistical inferences. Big data that are subject to substantial selection bias and measurement errors, without information in the data about the magnitude, sources and types of errors, should not be used to inform important decisions without substantial care and skepticism. Currently many ML researchers are charging ahead aggressively without a full knowledge of such issues. The fair prediction algorithms case study provides another example of the dangers of such a practice. It is crucial to at least be aware of such issues and attempt to the extent possible to adjust for them, and include as a crucial component of uncertainty quantification the impact of selection and measurement error.

5. Discussion

In this short discussion article, I have attempted to provide a brief overview of what I see as the role of statistics in the era of big data — the theme of this special journal issue. I view myself as a statistician with an active interest and research agenda focused on developing and applying machine learning methods. My own research tends to be fundamentally application-driven, and I want to develop practically useful methods that can lead to new scientific insights and that can ideally inform policy. I work closely with scientists in a wide variety of research areas ranging from neuroscience to genomics to epidemiology to ecology. In scientific applications collecting high-dimensional and complex data, there is a fundamental danger to applying current ML-style statistical methods. These include the lack of uncertainty quantification, the inability to provide a warning that we are being too ambitious and should attempt "coarser scale" inferences, and the lack of accounting for selection bias and the sampling frame under which the data were developed. "Modern" statistical theory and methods essentially take a ML mindset to attacking high-dimensional data problems, and hence also do not currently provide much in the way of useful solutions to these pressing problems. I am hoping that this article and the corresponding discussions in this special issue stimulate much more of a focus on developing statistically well grounded methodology for reliably and reproducibly conducting scientific inferences and making policies on the basis of "big data." Such developments will likely require a close collaboration between the Stats and ML-communities and mindsets. The emerging field of data science provides a key opportunity to forge a new approach for analyzing and interpreting large and complex data merging multiple fields.

Acknowledgments

The author thanks Anirban Bhattacharya, Daniele Durante, James Johndrow and Kristian Lum for helpful comments on a draft.