An analysis of scientific papers published in PLOS journals suggests modern machine learning techniques are overtaking traditional statistical tecnhiques

Sophie Berube

9/28/2017

Abstract

This study uses articles from all nine PLOS journals to examine trends over the past decade in data analysis techniques used by the scientific community. Specifically, modern data analysis techniques such as machine learning, deep learning and neural networks that have become increasingly popular among applied scientists, and in order to meet the increasing demands of complex and large data sets scientists have begun to use these methods in lieu of traditional statistical methods like linear regression, generalized estimating equations and bootstraps. By measuring the number of articles mentioning key phrases associated with these modern and traditional statistical methods over time, this analysis shows a very preliminary trend that might suggest decreasing use of traditional statistical methods since 2015. The analysis also shows, for some key words, that the modern statistical methods have overtaken traditional methods in terms of raw counts which might signal a broader shift in the scientific community towards machine learning, deep learning and neural networks for data analysis. Moreover, analysis of Poisson regression models shows, in some cases, a significant interaction between the subject of the paper and the time of publication, suggesting that the trends in the number of papers mentioning a particular key word might differ significantly over time according to the subject of the paper.

Introduction

Over the past decade, machine learning algorithms have become increasingly accessible to the scientific community as a means of data analysis. Through the use of various coding languages such as R and Python, scientists have successfully implemented machine learning algorithms in a variety of applied fields like neuroscience, genetics, proteomics and cell and molecular biology. Under the umbrella of machine learning advances, various other tools have been developed by statisticians, mathematicians and computer scientists and are beginning to make a significant impact in the larger scientific community. Deep learning has begun to emerge as one of these modern methods of analysis, meeting the scientific community's ever growing needs for tools that can handle increasingly complex and large data sets.[1] Similarly, artificial neural networks seek to harness the properties of the nervous system to analyze a variety of data and rely on similar machine learning related concepts to attack a range of complex and large problems.[2]

The following analysis uses a sample of the articles in the various journals of the Public Library of Science to assess whether modern statistical methods like machine learning, deep learning and neural networks are surpassing traditional statistical methods like ANOVA, linear regression, bootstraps and Bayesian methods in popularity across applied scientific fields and whether these time trends interact significantly with the scientific field of the paper.

The Public Library of Science or PLOS began publishing articles in 2003 in PLOS Biology and has since added eight journals to the repertoire in a variety of fields including medicine, computational biology, genetics, clinical trials, pathogens and tropical diseases. PLOS ONE, the largest of the PLOS journals, publishes more articles annually than any other current scientific journal.[3] An open access, pay to publish model ensures that articles are peer reviewed to verify that scientific standards are met but submissions are not turned down based on perceived lack of importance or impact to a particular field's advancement the way they would be in many prestigious mainstream journals like Nature or Cell. This makes PLOS journals a fairly adequate

representation of average scientific research over the past decade, and thus an appropriate database to assess statistical methodologies used by the scientific community over the past decade.

Another important component of the PLOS structure is the breadth of fields the articles cover. There are multiple different kinds of journals published in the PLOS family, however PLOS ONE has the highest annual number of published articles by a sizable margin, thus comparisons across journals might prove difficult. Conveniently, each article is associated with key phrases that accurately classify the field each article is associated with.

The R package `rplos` [4] which accesses articles from all nine currently published PLOS journals since their respective inceptions, was used in this analysis.

Methods

First, a list of ten key words representing both modern and traditional data analysis techniques was assembled by looking through a variety of PLOS articles and based on fundamental techniques discussed in most basic statistics textbooks. The five phrases meant to represent traditional statistical analysis methods were ANOVA, Bayesian, linear regression, mixed effect model, generalized estimating equation and bootstrap. Similarly phrases words meant to represent modern statistical methods are machine learning, deep learning, support vector machine, artificial neural network.

In order to visualize their distribution over all published articles in PLOS over time a sample of ten thousand articles across all PLOS journals that contain the key word was taken, then using the publication dates of these ten thousand articles, the counts were assembled and plotted from the period 2003 to October of 2017 and a loess curve was fit to these plots. Terms were plotted pairwise and each possible pairing of key phrases was observed with the corresponding loess curves. See section 1 of the supplemental code.

In order to asses whether or not there is an interaction between time and field a Poisson model was fit in the following way for each of the ten key phrases.

$$log(\mu) = \beta_0 + \sum_i \beta_i * year_i + \sum_j \beta_j * field_j + \sum_{i,j} \beta_{ij} * year_i * field_j$$

Where our outcome y_{ij} is the number of articles mentioning one of the ten key words (ANOVA, Bayesian, linear regression, generalized estimating equation, bootstrap, support vector machine, deep learning, machine learning and artificial neural network), that were published in year i and pertain to scientific field j.

The years in the regression range from 2004 to 2017, the year 2003 was not considered in the model, since PLOS had just been set up and very few articles were published that year. The fields were chosen by sampling ten thousand articles and extracting the eight most common fields associated with articles. They are: Biology and life sciences, Physical sciences, Medicine and health sciences, Research and analysis methods, Earth sciences, Computer and information sciences, Engineering and technology and Social sciences. The counts y_{ij} were obtained by sampling ten thousand articles for each key phrase across all PLOS journals and counting the number of articles published in each year for each of the eight fields identified. See section 2 of the supplemental code.

In the above processes, all articles that mentioned a key word were eligible for sampling. Since we are trying to assess which statistical methods scientists are using, this method assumes that if an article mentions a key word associated with a particular statistical technique, then that technique was actually used in the study. In order to test the validity of this assumption, a false discovery rate was estimated to give an idea of how many articles might mention a key word without it being used as a statistical analysis method in the paper. This rate was estimated by randomly sampling ten articles for each key word or phrase. These articles were visually inspected to determine whether or not that particular statistical method was used in the analysis. See section 3 of the supplemental code.

Results

The estimate of the false discovery rate was found to be 16 out of 100 or 16%. Importantly, the number of falsely discovered articles was not consistent across all key phrases. Specifically, the phrases generalized estimating equation, deep learning and artificial neural network had a higher number of these incorrect identifications, specifically of ten scanned articles the key phrase generalized estimating equations was incorrectly identified as a method of analysis in seven of ten articles, deep learning was incorrectly identified as a method in two of ten articles as was the phrase artificial neural networks. All other phrases were incorrectly identified in none, or one article.

The pairwise plots with loess curves reveal very similar curves, in other words low distance between the two curves for any time point, for the following pairs of terms. ANOVA and: linear regression, mixed effect model, generalized estimating equation and bootstrap. Bayesian and: linear regression,mixed effect model, generalized estimating equation,support vector machine and bootstrap. Linear regression and: mixed effect model, generalized estimating equation and bootstrap. Mixed effect model and: generalized estimating equation, support vector machine and bootstrap. Generalized estimating equation and: support vector machine and bootstrap. Machine learning and: support vector machine and bootstrap. Deep learning and artificial neural network.

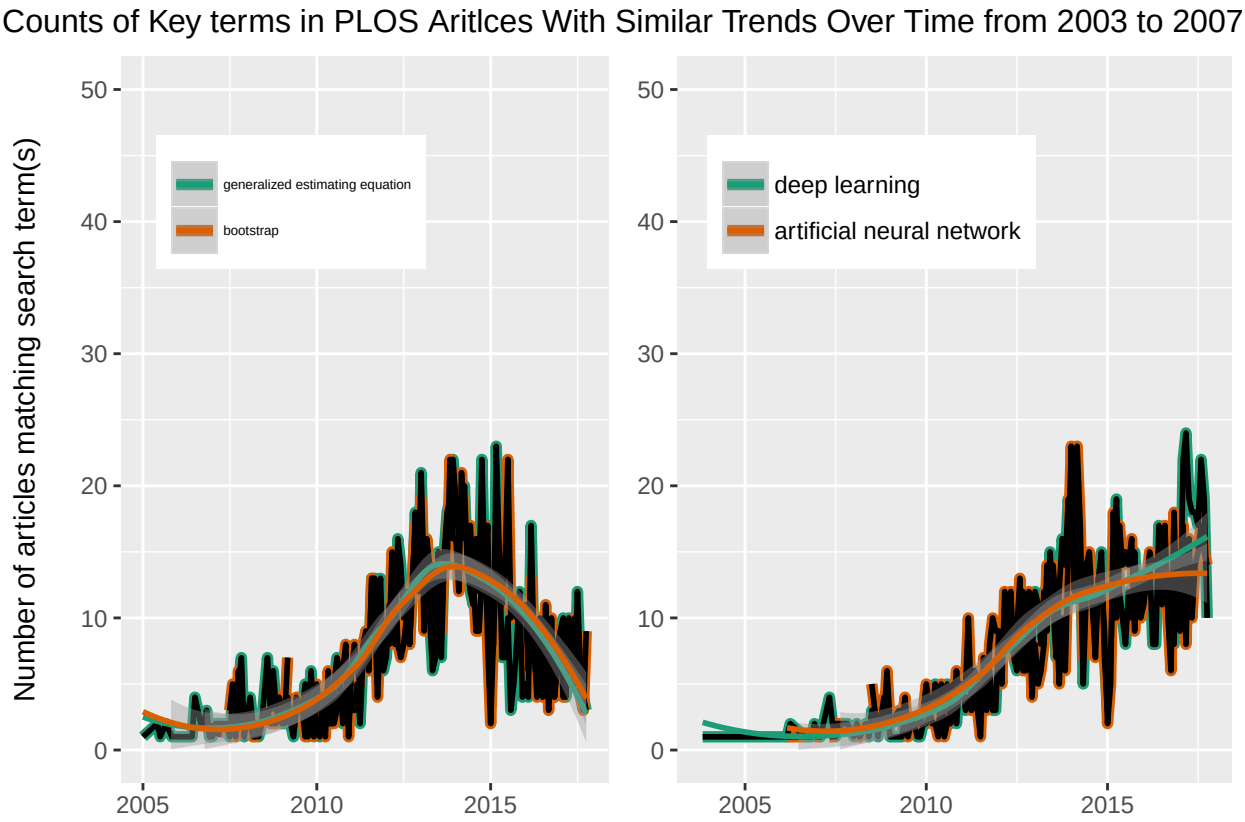

Figure 1: **Older statistical techniques have comparable trends to each other and new statistical tecnhinques have comprable trends to each other.**The left graph shows the number of PLOS articles mentioning two key phrases associated with traditional data analysis techniques over time, with a loess curve fit to the data. These two phrases have similar trends over time. The right graph shows the number of PLOS articles mentioning two key phrases associated with modern data analysis techniques over time with a loess curve fit to the data. These two phrases also have similar trends over time.

The pairwise plots with loess curves reveal different curves, in other words high distance between the two curves for several time points,for the following pairs of terms. ANOVA and: machine learning, support vector

machine, deep learning and artificial neural network. Bayesian and: machine learning, deep learning and artificial neural network. Linear regression and: machine learning, support vector machine, deep learning and artificial neural network. Mixed effect model and: machine learning, deep learning and artificial neural network. Generalized estimating equation and: machine learning, deep learning and artificial neural network. Machine learning and: deep learning, bootstrap and artificial neural network. Support vector machine and deep learning. Support vector machine and artificial neural network. Deep learning and bootstrap. Bootstrap and artificial neural network.

However these different curves have interesting patterns within them. Across all terms, the loess curves are increasing from years 2003 to 2015 and decreasing after 2015. In the pairwise comparison plot between ANOVA and machine learning, while ANOVA has a steep rate of decrease after 2015 machine learning has much more shallow rate of decrease after 2015, and while the counts for articles displaying machine learning are consistently below those for articles displaying ANOVA, sometime during late 2015 and into early 2017, the loess lines for the two counts cross and more articles display machine learning than ANOVA. The similar phenomenons can be observed for ANOVA and support vector machines, AONOVA and deep learning, ANOVA and artificial neural networks, Bayesian and machine learning, Bayesian and deep learning, Bayesian and artificial neural network, linear regression and machine learning, linear regression and deep learning, linear regression and artificial neural network, mixed effect model and machine learning, mixed effect model and deep learning, mixed effect model and artificial neural network, generalized estimating equation and machine learning, generalized estimating equation and artificial neural network,generalized estimating equation and artificial neural network, machine learning and bootstrap, deep learning and bootstrap and bootstrap and artificial neural network.

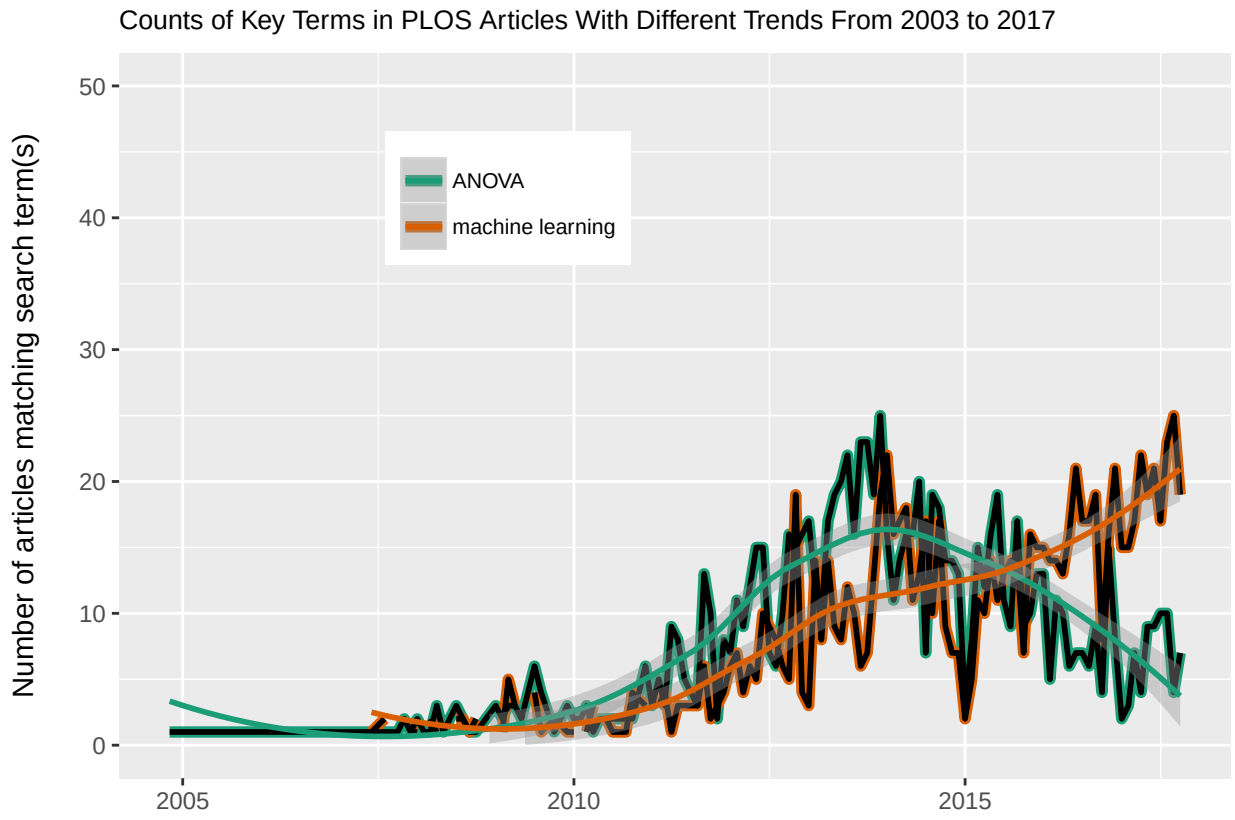

Figure 2: **Older statistical methods have different trends over time than modern statistical methods.**This graph shows the number of PLOS articles mentioning two key phrases, one associated with traditional data analysis techniques over time (ANOVA) and one associated with modern data analysis techniques over time (machine learning). Loess curves are fit to both data sets, while ANOVA shows a sharp decrease from 2015 onward machine learning is almost flat for this time period.

Finally, when looking at the pairwise graphs, some curves, while far apart, seem to have a similar shape after 2015. These include machine learning and artificial neural network and machine learning and deep learning.

For the significant estimates of the Poisson regression models see section 2 of supplemental code. Significant interaction coefficients are present in every model except for the one fit for ANOVA as the key word.

Discussion

Looking at terms whose graphs are similar in pairwise comparisons, traditional statistical analysis methods (ANOVA, Bayesian, linear regression, mixed effect model generalized estimating equation and bootstrap) appear to have similar trends. Notably, support vector machines appear to have similar count trends over time to several of the more traditional methods. Support vector machines fall under the umbrella of machine learning, though the modern representation of support vector machines that is most applicable to data from applied science disciplines was published by Cortes and Vapnik in 1995 [5] which has allowed time for computer scientists, mathematicians and statisticians to make code for this technique accessible to a variety of disciplines. This might explain the similarity in trend to more traditional statistical methodologies. Moreover certain similarities in trends between terms can be explained by their intertwined development process. For instance, deep learning and artificial neural networks seem to share similar trends over time, which could be attributed to the fact that the development of deep learning was an integral part of advances in artificial neural networks[6].

Looking at terms whose graphs are different in pairwise comparisons, machine learning provides an interesting contrast to each of the traditional approaches. Especially in the post 2015 period, where the number of machine learning articles seems to stay almost steady while the number of articles mentioning each of the traditional approaches declines steadily. In some cases, including ANOVA and linear regression, machine learning counts which are lower than ANOVA and linear regression counts from 2003 to 2016, surpass ANOVA and linear regression counts around 2016. This might indicate a general shift away from linear regression and other simpler traditional statistical methods and towards machine learning techniques that are often better suited to handle the increasingly complex and large data sets in applied scientific fields.

The Poisson models reveal some significant interaction terms shown in Table 1, suggesting that in some cases there is an interaction between time and field, in other words there are different effects across disciplines for a particular year for the counts, however these don't follow any clear patterns and seem to be evenly dispersed between old and new terms and across time points. This could have to do with the non linearity of the relationship between counts and time as is evidenced in Figures 1 and 2. In order to establish a more clear relationship and in an effort to answer the question of whether these time trends interact significantly with the scientific field of the paper, a more complex model would likely be required to handle this non-linearity.

Some limitations of the study should be considered. First, the key phrases meant to represent the older and newer statistical techniques were chosen by carrying out a quick survey of current literature, there are several other terms that might also be representative of older and newer statistical techniques which could produce different results than the ones obtained here. Second, the total number of articles published by PLOS is not constant across years, in order to be able to draw stronger conclusions about both time trends and possible interactions between time and field, the counts should be normalized to account for variation in total number of PLOS published articles. Finally, the random samples of PLOS journals used to fit the Poisson regression model and to create the plots is being used to make inferences about trends in the broader scientific community which includes a wide variety of journals outside PLOS, a more extensive sampling of peer reviewed journals would give a more complete picture of trends in statistical techniques being used by the broader scientific community.

Conclusion

Overall, the pairwise graph comparisons with associated loess curves did suggest that modern statistical analysis techniques like machine learning, deep learning and artificial neural networks might be overtaking

traditional statistical methods. The Poisson regression models suggest that in some cases there could be a significant interaction between time and the field the article pertains to, however there are no clear trends across the key phrases, nor are there evident patterns across particular time periods or fields. Further studies that include more articles from a wider source of journals might provide further insight into these trends and the use of more advanced prediction models might allow for stronger conclusions about whether or not techniques like machine learning, deep learning and neural networks are in fact overtaking methods like linear regression, generalized estimating equations and bootstraps and whether or not these time trends interact significantly with the field associated with the content of the articles.

References

1. Schmidhuber,J (2015). Deep Learning. Scholarpedia, 10(11):32832
2. Wang SC. (2003) Artificial Neural Network. In: Interdisciplinary Computing in Java Programming. The Springer International Series in Engineering and Computer Science, vol 743. Springer, Boston, MA 3.Morrison, Heather (5 January 2011). "PLoS ONE: now the world's largest journal?". Poetic Economics Blog. Retrieved 16 January 2011
3. Scott Chamberlain, Carl Boettiger and Karthik Ram (2014). rplos: Interface to PLoS Journals search API.. R package version 0.4.0. https://github.com/ropensci/rplos
4. Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.
5. Schmidhuber,J (2016). Deep Learning in neural networks: An overview. Neural Networks(61):85-117.