Trends Over Time In Common Statistical Methods

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 in some select cases 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.

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 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.

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.

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, the function `plot_overtime` from the `rplos` package was used. The limit for each word was set to 10,000 articles which means that 10,000 articles mentioning each of the key words were sampled and were attributed to a particular month based on their date of publication. Counts for various months were 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.

Quasipoisson models were fit for each key word with a count of articles mentioning a particular key word as the outcome and time points as the regressor, in order to identify possible differences in trends over time. More specifically, using visual inspection from the loess plots, the data were split up into a time period of increasing counts for all terms and a time period of decreasing or stable counts for all terms. For each key word, a different quasipoisson model was fit for both increasing and decreasing counts periods and the coefficients for the time points were compared both within key words and across key words.

Since the `plot_throughtime` function includes any articles that mention the key words in the counts for a particular time point, 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 using the `highplos` function which returns the API (or unique identifier) of articles that contain a particular key word. For each key word or phrase, ten of these articles were visually inspected to determine whether or not that particular statistical method was used in the analysis.

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.

Counts of Key Terms in PLOS Articles With Similar Trends over Time From 2003 to 2017
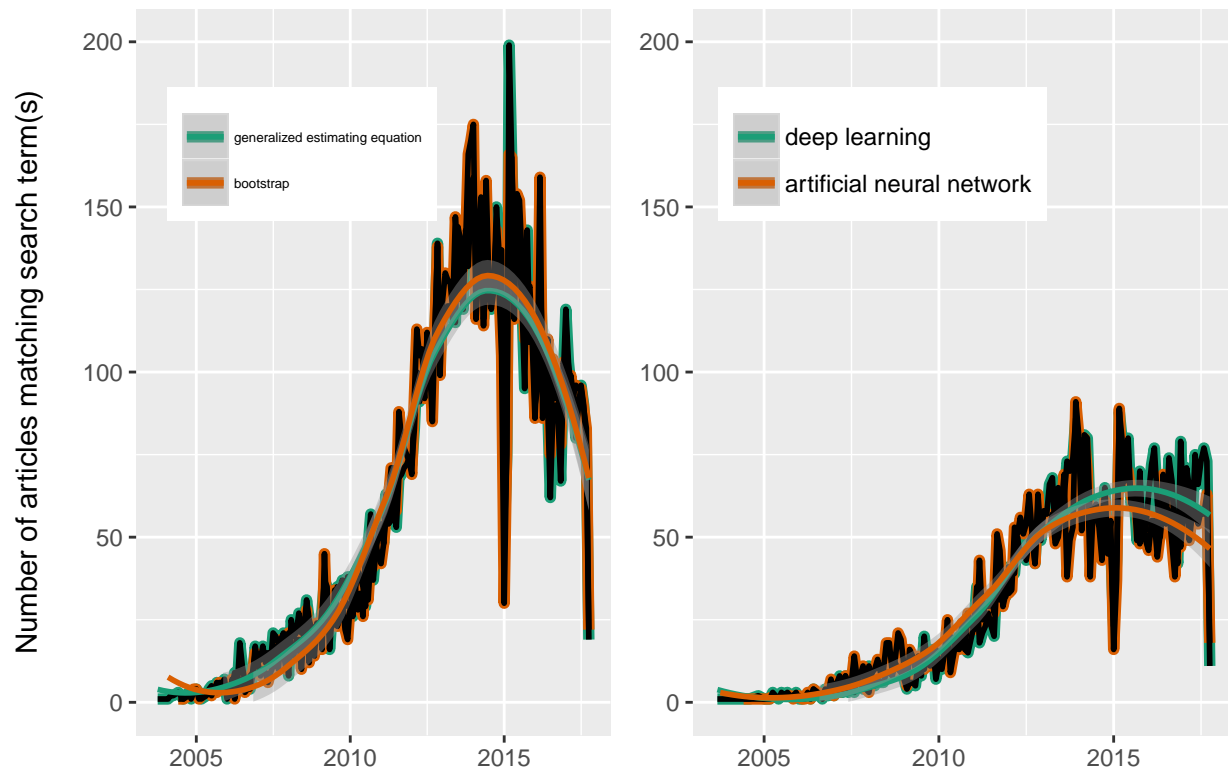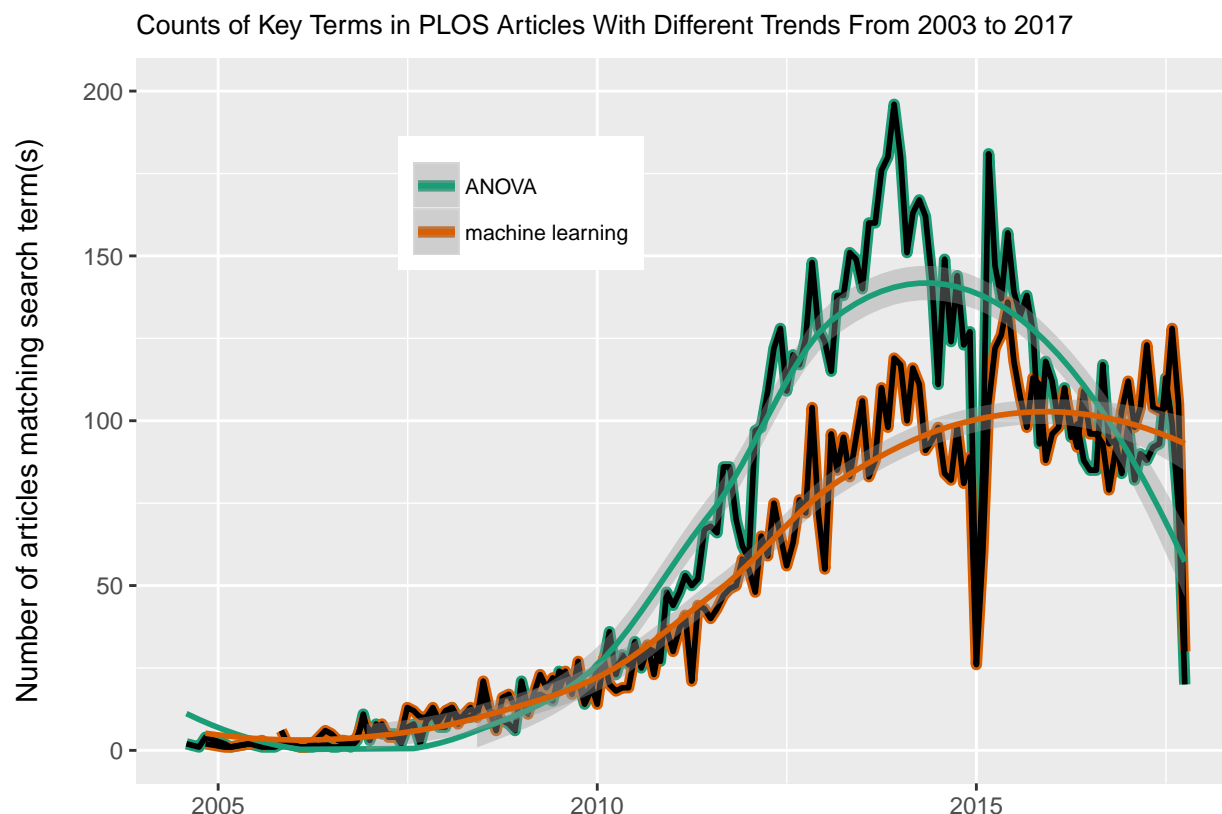

Figure 1: Key words with similar trends over time. The left graph shows the number of PLOS articles mentioning two key phrases associated with traditional data analysis techniques over time, with a loess curve fit to the data. These two phrases have similar trends over time. The right graph shows the 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.

Counts of Key Terms in PLOS Articles With Different Trends From 2003 to 2017

Figure 2: Key words with different trends over time.

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.

Finally, for each of the key phrases, the data were split into a time of ascending counts, from 2003 to the end of 2014, and a time of descending counts from the start of 2015 to October of 2017. Separate quasipoisson models were fit for each of these periods in each key phrase with time as the regressor and using the `glm` function. Of these models the coefficients related to the time point regressor that were found to be significant in both the ascending and descending models were those for the key terms ANOVA, Bayesian, linear regression, mixed effect model, generalized estimating equation, bootstrap and support vector machine. These coefficient estimates were also quite similar. Specifically, the estimates for the coefficients related to the time regressor for ascending models in the ANOVA, Bayesian, linear regression, mixed effect model, generalized estimating equation, bootstrap and support vector machines ranged from 9.6×10^{-4} to 1.83×10^{-3}, all with a p value less than 2×10^{-16}. The estimates for the coefficients related to the time regressor for descending models in the ANOVA, Bayesian, linear regression, mixed effect model, generalized estimating equation, bootstrap and support vector machines ranged from -5.2×10^{-4} to 1.827×10^{-4}, with p values ranging from 5.29×10^{-5} to 0.0477, on the cusp of what is typically considered significant.

While the models fit for the upward segments of the machine learning, deep learning and artificial neural network key words had significant coefficients related to the time point regressor, none of the coefficients for the downward segments were significant.

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.

However, it is difficult to draw any strong conclusions about this since the the coefficient associated with time regressor for the machine learning quasipoisson model from 2015 to 2017 was not significant, with a p value of 0.978. While it did show a more shallow slope than those for the key phrases associated with traditional statistical analysis methods (ANOVA, Bayesian, linear regression, mixed effect model generalized estimating equation and bootstrap), it is difficult to draw any conclusions about possible differences given the non-significant p value associated with the coefficient.

Similarly, while the loess curves from 2015 to 2017 for deep learning and neural networks, also appear to have a much shallower decrease, the time regressors for the these quasipoisson models from 2015 to 2017 were not significant, with p values of 0.978 and 0.4680 respectively.

Conclusion

Overall, while 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 quasipoisson models did not reveal significant coefficients for several modern statistical analysis methods in the period of 2015 to present day making it difficult to draw stronger conclusions about the relative popularity of these methods. 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.

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.