Journal of Scientometric Res.2019;8(2s):s39-s43
http://www.jscires.org

Relevance of Innovations in Machine Learning to Scientometrics

Gowri Srinivasa

Center for Pattern Recognition and Department of Computer Science and Engineering, PES University, EC Campus, Bengaluru, Karnataka, INDIA..

ABSTRACT

Machine learning envisages building models that either classify, predict, cluster or determine the relative relevance of features to a problem and the associations between them. This paper briefy describes how these tasks are relevant to Scientometrics. Through this brief survey of selected tasks, it is observed that most solution approaches in Scientometric literature are built on the strong foundation of understanding and debating in uencing factors and the process of feature engineering, requiring the descriptors to be intuitive and methods used for classication, prediction, etc., to be amenable to interpretation. Recent trends in machine learning, particularly, deep learning methods, however, pose an interesting question: can we build models that automatically determine what features are important and thereby bypass the step of feature engineering? This paper discusses how such techniques could also be harnessed in Scientometrics.

Keywords: Feature engineering, Machine learning, Deep learning, Scientometrics.

Correspondence
Gowri Srinivasa
Professor, Department of Computer Science and Engineering, PES University, EC Campus, Hosur Road, Bengaluru 560100, Karnataka, INDIA.
Email: gsrinivasa@pes.edu

Received: 16-02-2019
Revised: 30-04-2019
Accepted: 11-06-2019
DOI: 10.5530/jscires.8.2.23

INTRODUCTION

Most computational tasks in Scientometrics can be understood broadly to involve the design or application of features to gain an insight to the (relative) impact of innovation or research of institutions, scientists and avenues of knowledge dissemination (such as journals, conference proceedings, etc.). Eliciting relevant numerical descriptors for quantifying such impact seems to require a deep understanding of the influencing factors. For instance, the Hirsch number or *h-index*, a relatively popular measure for the citation index of an author, is computed as

$$\max_k \min(c(k), k),$$

where $c(k)$ is the number of citations of the kth publication, listed in the descending order of citations.[1] This measure does not favor a large number of poor quality publications or a very small number of highly cited articles. Similarly, the *THE-QS* based on academic prestige through features that quantify the quality of teaching, research, citations, international outlook, etc. As with any quantitative measure, ever since these quantifiers were proposed, there have been several debates on the advantages, limitations and comparisons with alternatives or variations of these measures.[2-5] However, our interest in these measures is to note that the selection of features for such problems requires some knowledge of the domain world University rankings ranks institutions

Given the definition of each measure, automating their computation on relevant data is fairly straightforward. Indeed, it would appear that most tasks in Scientometrics involve computation of scores from data, once the measures are defined. However, when we are required to perform computations that build on the insight gained from past data for prediction (such as: predict the ranking of an institution at the end of the year, based on the measures computed a few months into the year) or tasks such as mining underlying patterns (for instance, what should an institution X focus on to improve its ranking in the coming years? what strategy should a publication employ to ensure articles in a particular area are read and cited?) or other descriptive tasks such as grouping institutions, individuals or journals with a similar subset of parameters or filtering publications in a certain research area, etc., techniques such as classification, regression, clustering, association rule-mining, etc., borrowed from machine learning prove to be useful.

The subsequent section presents a brief survey of machine learning techniques used in Scientometrics, followed by a summary of the recent innovations in machine learning (particularly, the power of deep learning networks) and a discussion of their relevance to Scientometrics.

MACHINE LEARNING IN SCIENTOMETRICS

Machine Learning techniques can be broadly discussed under the heads of supervised and unsupervised learning. Supervised learning comprises tasks such as classification and prediction for which models are designed for training data with target annotations. Unsupervised learning or clustering deals with finding groups of similar data points. This latter task does not presuppose any annotation of the training data. The goal of supervised learning is to achieve a model that maps an input to its target output for the training set, replicating the logic behind the annotation process on test data and hence produce *expected* results. In the case of unsupervised learning, the goal is to group similar data points or partition the feature space to natural groupings. Thus, besides the advantage of not requiring the data to be annotated (a tedious and time consuming effort), there is scope to discover novel underlying patterns and gain an insight to the data through unsupervised learning.[6]

Classification

Machine learning for classification can be abstracted to a system that preprocesses the raw input (involves tasks such as cleaning the data to eliminate erroneous data, identify missing data, outliers, etc., and take an appropriate action – such as interpolating some missing data or eliminating anomalies, to render the data amenable for further tasks in the process pipeline), extracts features (and possibly, even selects a subset of relevant features) and creates a model based on the features. The process of creating a model or 'training the classifier' is iterative and refinement is based on evaluating the model through cross-validation (presenting some of the annotated data (not used for training) to 'test' the system). Suitable adjustments to the model may be made to ensure the errors are minimal, there is no systemic 'bias', that the model does not 'overfit' the training data and to assess the significance of the average performance (or describe how repeatable is the performance on one set of validation data) through multiple folds of validation.[7]

An example of a classification task in Scientometrics could be automatically classifying the category of a citation to be able to retrieve the most relevant/ useful references when required. Garzone and Mercer start with the observation a citation has multiple purposes such as paying homage to predecessors, acknowledging the use of some equipment or technique, questioning, agreeing with some result, etc. They present a rule-based classifier to label 35 categories of citations.[8] Another example is the related task of determining the polarity of a citation (positive or negative). This has been accomplished using linguistic features that describe the context of the reference.[9] It is noteworthy that the solution approach to these tasks hinge on the choice of meaningful features that capture the essence of the relevant content as well as the careful design of rules, which happen to be intuitive and easy to interpret, to achieve a meaningful outcome.

Prediction

Prediction of a target variable is the task of building various types of models, typically a weighted function of the current and/or 'past' data and other parameters that influence the outcome of target variable, to forecast the value it might take on at a specified time in the future.[10] A popular approach to predict values is through designing regression models whose output is a real number *vis-a-vis* a category or a number signifying a class label as with classification. For instance, the number of citations an article could be expected to have based on the number of authors, institutions, citations of individual authors, etc., is a predictive task.[11] A recent study has proposed the use of relevant features with neural networks to predict articles that would be highly cited.[12] Another interesting study has shown that the most cited articles (in Medical Research) can be predicted based on the number of tweets within the first three days of the article being published.[13] Quantile regression has been used to model the probability distribution of the future citation count of articles.[14] Through this it has been shown that potential long term impact of various articles can be predicted. Another study has shown that the keywords of the abstract (modeled using a bipartite graph) can be used to determine the number of citations in the future, with articles having higher citations conforming to the mainstream.[15] There have been a number of such studies that have compared the merits of various predictive models, most notably variants of regression.[16,17]

a. Feature Analysis

As we noted earlier, each of the foregoing methods requires an understanding of the influencing factors that are most relevant to the problem. A straight-forward supervised approach to classification is the k-nearest neighbor method (abbreviated as KNN). This method matches a set of features of a sample that needs to be annoated with others in the training set. It then selects a label that is common to the k most similar samples.[18]

While prediction tasks are typically accomplished through studying correlations between features, there are more sophisticated techniques used and interesting questions that can be asked when we automate the process of finding appropriate factors. For instance, what makes an article influential?

Multivariate analysis has been performed to elicit this information.[19] Likewise, associations between features can be studied to come up with recommendations. As an example, research collaborations have been suggested based on predicting the link between research centers doing similar work using random forest classifiers.[20] Gini Index has been used in this study to determine the relative importance of features in determining the recommendation. Another multivariate analysis technique to understand the relative importance of features and a method traditionally used to arrive at a subset of weighted features that serve as strong predictors is the principal component analysis (abbreviated as PCA).[21,22] Association rules are implications or bijections that describe the relationship between two features. It seems the one of the most natural methods to arrive at relationship between explanatory variables. Even though there is much to be explored with mining of association rules, there have been a few examples of how insightful this can be. For instance, co-occurance of keywords with authors has been used to mine for frequent patterns resulting in association rules for authors and keywords or Journals and keywords.[23]

Clustering

Clustering is a process of grouping data based on the similarity between the features (most methods seek to minimize inter-class similarity and maximize intra-class similarity). This is a popular approach in Scientometrics for two reasons: aggregating data helps summarize the results for data points that are similar and there is no need for a large dataset of annotated data. Unlike the case of supervised learning where class labels or category boundaries may require some justification, the patterns that emerge through clustering can be used to gain some insight. The aspects that need some attention when using clustering are: the choice of similarity (or dissimilarity) measures – how well does it capture the inherent relationship between features? and the other is the choice of clustering algorithm. There are a number of approaches that can be used for clustering. The most popular approaches are agglomerative hierarchical clustering and DBSCAN. The former results in a dendrogram that augurs for a neat visual representation. The latter, DBSCAN, is a density based clustering method that takes into account the lack of homogeneity in the spread of data. An example of aggregation in Scientometrics is the clustering of journals and category labels at various levels.[24] For the DBSCAN, an example could be of arriving at a paper recommendation based on the proximity of citations.[25]

It is often seen that methods are rarely used in isolation, but in combination. For instance, an interesting problem is that of tracking changes in trends. In particular, changes in patent citation networks (i.e., clusters) have been studied over time to describe growth (an increase in the number of citations), contraction (a reduction in the number of citations), merging and splitting of citation networks, and the birth and death of a network from an existing one. Subjective and objective measures have been combined for the task with the hope the method identifies, for instance, the advent of new technological areas before the US Patents Office recognizes them.[26] Another multivariate model analyzes citation networks of articles to infer that for a higher h-index, it is advisable to publish with a large number of co-authors, particularly those who have been highly cited.[27] For this, the authors consider a network of co-authors that is centered around an individual author ('ego-centric networks').

Since the design of a new heuristic or modeling approach in machine learning is not the objective in Scientometrics *vis-a-vis* the choice of features and interpretation of the outcome, few papers in the area have detailed explanations of the mathematical underpinnings of the methods used or algorithmic details such as parameter turning. A broad overview of Machine Learning methods and how they apply to Scientometrics can be culled from.[28]

INNOVATIONS IN MACHINE LEARNING

Most of the effort in Machine Learning was focused on finding representations of data that are descriptive (for clustering or predictive analysis) or discriminative (for classification), understanding their interrelationships (correlation, multivariate analysis, association mining) and arriving at meaningful subsets (principal component analysis), etc. These tasks presupposed an understanding of the domain and an ability to preprocess the data followed by the design meanignful features. An enormous innovation in machine learning has been to outsource the task of feature engineering to machines. Central to this innovation is the question: can machines determine, on their own, representations of the data that matter for a task? It turns out that this is possible remarkably well through, what is quickly evolving to be a tool of choice across fields, deep learning.[29] This has had a particularly high impact for problems in which the dimensionality of the original data is huge, for which there is a very large volume of data points and the complexities are prohibitive to manually comb through the data to annotate training images and engineer meaningful features, such as classification of over a million images belonging to over a 1000 categories.

Deep learning has at its core an artificial neural network – the same idea used in the foregoing section to explain the principle of classification – a model that maps the input to a target label. The only difference is that there is a nonlinear function that computes the weighted sum of input features. While a single layer neural network returns a nonlinear map of some weighted combination of the features, it was explained that a network with multiple such layers partitions the feature space

through arbitrary unions of finite intersections, representing different regions corresponding to the categories.[30] Building on that principle, when multiple nodes in each layer are stacked upon multiple such layers, hence the name 'deep neural network', it manages to extract features that are relevant. And, through multiple epochs of training, adjusts the weights assigned to these features to arrive at a meaningful decision.[31] In fact, it has been shown that such a deep neural networks can outperform traditional feature selection and dimensionality reduction approaches such as PCA.[32]

Since the need for explicit feature engineering is obviated, different combinations of the number of layers in a deep learning network, the nonlinear functions used, the loss function based on which the weights are optimized, choice of learning rate and regularization procedures to overcome overfitting and mushrooming of off-the-shelf pre-trained models and computational tools to code deep learning architectures there has been an implosion of scientific papers on the theoretical aspects of deep learning and even more on the application of deep learning to solve problems in various fields. Some of the techniques and their progression have been summarized in various surveys.[33,34] The applications of deep learning in computer vision, natural language processing and time series analysis have also been summarized in recent surveys.[35-37]

Relevance of Deep Learning to Scientometrics

What role would deep learning have to play in Scientometrics that relies on the design of features that can be understood and discussed? Since a lot of the predecessor work on machine learning hinges on analysis of content, this can be done with even more content using deep learning. For instance, when similarity groupings between journal articles were done, proximity measures had to be defined. Keywords of articles do not always match similar papers accurately and extending the matches to keywords extracted from the content could be colored by the length of the paper, context, etc. These are circumvented through use of language embedding models with deep learning. A word embedding, such as Word2Vec for example, coverts every word to a d-dimensional vector (typically 100–300 dimensions have been found empirically to be useful), rendering words used in similar contexts to be more similar than words that are literally closer.[38] Thus, words such as king and prince would have vector representations with a smaller distance between them than word pairs such as king and kind or prince and price.

Language models have been used to good effect with semantic ranking of papers in PubMed.[39] Similarly, content can be studied for proximity between citation contexts using such language embedding models to arrive at more meaningful reference retrieval systems. Language models can also be used for sentiment analysis of the reference context having a positive or negative connotation.[40] Deep Learning can be used with unsupervised learning to group similar content (document clustering).[41] Further heuristics, such as citations or frequently used keywords, etc., can be extracted from these that can be interpreted. Node representations through deep learning architectures can be used to discover network communities within large domains of scientific publishing.[42]

Treading with Caution

For low-resource data, overfitting is a problem with large network architectures. Transfer learning has found to be useful.[43] It remains to be seen if models built for tasks in other domains can be retrained with less effort for similar tasks in Scientometrics to achieve meaningful outcomes.

If neural networks were treated as a black-box for not being able to interpret the weights and partitioning of the feature space, deep learning networks have proved to be a 'blacker' box, in that even the features are not easily amenable to interpretation. It has also been shown through applications that deep learning is prone to errors in adversarial settings. This has limited the use of deep learning in fields such as healthcare, where it is imperitive for a computational model to be robust and 'transparent'. There has been some effort to address these limitations in the recent times.[44]

The spotlight in the recent times has also turned towards understanding metaheuristics for deep learning. What loss functions work better for an application? What assumptions on the data/ loss function expedite convergence? How should the learning rate be selected to avoid local minima? Is it possible for fewer epochs of training or smaller training sets to be used to achieve high performance measures achieved with vast amounts of high dimensional data and training over several epochs? For instance the Saha-Bora Activation Function (SBAF) has been used to explain the rise in ranking of the journal Astronomy and Computing over its predecessors.[45] Perhaps, a close look at the theoretical underpinnings of the methods can lend a deeper insight to the features that matter and pave way for harnessing the power of deep learning more effectively in Scientometrics in the future.

CONCLUSION

The field of Scientometry has benefited from computational advancements in Machine Learning in the past. Some instances include the analysis of social media postings to forecast the citations a Journal article might receive, analyzing the sentiment of a citation to determine if it has been used to strengthen an argument or rebut it and the context of a citation to retrieve relevant references. We have also noted the complexity of designing heuristics to rank institutions or journals or quantify the scientific impact of an author as these are beset with some bias inherent to how the measure is defined. The explicit

choice of influencing factors has led to debates about the relative importance of features and paved way for new measures to evolve. The advances in Machine Learning in recent times, particularly deep learning that obviates the need for explicit feature engineering, has proved to be most useful in other domains such as computer vision and linguistics to solve a plethora of problems considered computationally intractable earlier. Given that, by design, deep learning takes away the transparency of features, it remains to be seen how the community will take to adopting these methods for Scientometrics. While the limitations are obvious, it can be argued to eliminate human biases. As suggested by empirical evidence, computational methods that require little intervention can be used to explain perplexing trends in the data. However, the choice of these methods would require a deep insight to the workings and foundations of the methods. It seems like a possibility that the right use of deep learning methods may even lend some new insights in Scientometrics.

REFERENCES

1. Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences. 2005;102(46):16569-72.
2. Costas R, Bordons M. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of informetrics. 2007; 1(3):193-203.
3. Waltman L, Van Eck NJ. The inconsistency of the hindex. Journal of the American Society for Information Science and Technology. 2012;63(2):406-15.
4. Jeremic V, Bulajic M, Martic M, Radojicic Z. A fresh approach to evaluating the academic ranking of world universities. Scientometrics. 2011;87(3):587-96.
5. Radojicic Z, Jeremic V. Quantity or quality: what matters more in ranking higher education institutions?. Current Science. 2012;158-62.
6. Duda H, Hart PE, Stork DG. Pattern Classification. John Wiley and Sons. 2012.
7. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife and cross-validation. The American Statistician. 1983;37(1),36-48.
8. Garzone M, Mercer RE. Towards an automated citation classifier. In Conference of the Canadian Society for Computational Studies of Intelligence. Springer, Berlin, Heidelberg. 2000; 337-346.
9. Athar A, Teufel S. Context-enhanced citation sentiment detection. In Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies. Association for Computational Linguistics. 2012;597-601.
10. Harrell JF, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems and suggested solutions. Cancer Treatment Reports. 1985;69(10),1071-77.
11. Fu LD, Aliferis C. Models for predicting and explaining citation count of biomedical articles. Proceedings of the American Medical Informatics Association Annual Symposium. 2008: 222.
12. Wang F, Fan Y, Zeng A, Di Z. Can we predict ESI highly cited publications?. Scientometrics. 2019;1-17.
13. Eysenbach G. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet research. 2011;13(4).
14. Stegehuis C, Litvak N, Waltman L. Predicting the long-term citation impact of recent publications. Journal of informetrics. 2015;9(3):642-57.
15. Klimek P, Jovanovic AS, Egloff R, Schneider R. Successful fish go with the flow: citation impact prediction based on centrality measures for termdocument networks. Scientometrics. 2016;107(3):1265-82.
16. Thelwall M, Wilson P. Regression for citation data: An evaluation of different methods. Journal of Informetrics. 2014;8(4):963-71.
17. Ajiferuke I, Famoye F. Modelling count response variables in informetric studies: Comparison among count, linear and lognormal regression models. Journal of Informetrics. 2015;9(3),499-513.
18. Wang M, Yu G, An S, Yu D. Discovery of factors influencing citation impact based on a soft fuzzy rough set model. Scientometrics. 2012;93(3):635-44.
19. Haslam N, Ban L, Kaufmann L, Loughnan S, Peters K, Whelan J, *et al.* What makes an article influential? Predicting impact in social and personality psychology. Scientometrics. 2008;76(1):169-85.
20. Guns R, Rousseau R. Recommending research collaborations using link prediction and random forest classifiers. Scientometrics. 2014;101(2):1461-73.
21. Julian K, Rigby J. Telling the whole story: finding structures in bibliometric information using PCA.
22. Salvador MR, Lpez-Martnez RE. Cognitive structure of research: scientometric mapping in sintered materials. Research Evaluation. 2000;9(3):189-200.
23. Li F, Li C, Tian Y. Applying association rule analysis in bibliometric analysis-a case study in data mining. In Proceedings. The 2009 International Symposium on Computer Science and Computational Technology. Academy Publisher. 2009; 431-4.
24. Zhang L, Liu X, Janssens F, Liang L, Glnzel W. Subject clustering analysis based on ISI category classification. Journal of Informetrics. 2010;4(2):185-93.
25. Habib R, Afzal MT. Paper recommendation using citation proximity in bibliographic coupling. Turkish Journal of Electrical Engineering and Computer Sciences. 2017;25(4):2708-18.
26. Rdi P, Makovi K, Somogyvri Z, Strandburg K, Tobochnik J, Volf P, *et al.* Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics, 2013;95(1):225-42.
27. Mccarty C, Jawitz JW, Hopkins A, Goldman A. Predicting author h-index using characteristics of the co-author network. Scientometrics. 2013;96(2):467-83.
28. Ibez Martn A. Machine Learning in Scientometrics. Doctoral dissertation (ETSI Informatica). 2015.
29. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 2012;1097-105.
30. Chen YQ, Thomas DW, Nixon MS. Generating-shrinking algorithm for learning arbitrary classification. Neural Networks. 1994;7(9),1477-89.
31. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
32. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504-7.
33. Deng L, Yu D. Deep learning: methods and applications. Foundations and Trends in Signal Processing. 2014;7(34):197-387.
34. Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015;61:85-117.
35. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience. 2018.
36. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. ieee Computational Intelligence Magazine. 2018; 13(3):55-75.
37. Gamboa JCB. Deep learning for time-series analysis. 2017; arXiv preprint arXiv:1701.01887.
38. Goldberg Y, Levy O, word2vec Explained: deriving Mikolov *et al.*'s negative-sampling word-embedding method. arXiv preprint arXiv. 2014; 1402.3722.
39. Gargiulo F, Silvestri S, Fontanella M, Ciampi M, De Pietro G. A deep learning approach for scientific paper semantic ranking. In International Conference on Intelligent Interactive Multimedia Systems and Services. Springer, Cham. 2018; 471-81.
40. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies 2011;1: 142-50.
41. Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. In International Conference on Machine Learning. 2016;478-87.
42. Xie Y, Gong M, Wang S, Yu B. Community discovery in networks with deep sparse filtering. Pattern Recognition. 2018;81:50-9.
43. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2010;22(10):1345-59.
44. Papernot N, McDaniel P. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. 2018; arXiv preprint arXiv:1803.04765.
45. Saha S, Sarkar P, Mathur A, Basak S. Model Visualization in understanding rapid growth of a journal in an emerging area. arXiv:1803.04644. 2018.