Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics (CONF-CIAP 2023)
DOI: 10.54254/2753-8818/5/20230332

A review of statistical analysis related to information theory and machine learning

Dongfang Lou[1]

[1] Golden Apple Jincheng NO.1 Secondary School, Sichuan, Chengdu, China, 610000

cdflsorient20@gmail.com

Abstract. With the collection and use of large amounts of data in various disciplines, it has become a necessity to analyze these data that people cannot handle by means of mathematics and information theory. A review of statistical analysis on previous work like data analysis, data mining, and machine learning are conducted in this paper. Specifically information theory in conjunction with machine learning based on basic statistical evaluation are focused on. Many mathematical instruments are used in this area, including evaluating the performance, loss functions, and activation functions. In this paper，all these methods from previous work are concluded and the future potential development in this area is discussed. Finally, this paper broadly summarizes the development history about machine learning and information theory, and discusses the pros and cons of both of them and their future development.

Keywords: information theory, machine learning, statistical analysis.

1. Introduction

As statistical analysis continues to advance, applying big data analytics to information theory and machine learning is a trend and a must. All the processing and analysis of big data is too much for humans to handle because there is too much data. there are two ways when large amounts of data are dealt, one is to reduce the amount of data by statistics or to use statistical indicators to get some characteristics of the data such as averages. The other is to use data formatting to see a large amount of data more intuitively. In this paper, the first one will be mostly focused on, using data analysis to make a review of the applications and theories behind them based on statistics.

In this article, the close connection and progressive development of statistical analysis and the above two are reviewed and discussed. Statistical methods are widely used in various areas, such as data processing, data mining, and machine learning. Within machine learning, statistical methods are implemented to calculate the loss, the difference between distributions, and the gradient to achieve backpropagation. Therefore, it would be necessary to address the applications and the relationship between statistics and machine learning.

In the first part, the development of data and information theory is described, the second part introduces machine learning based on data analytics, and the third part looks at applications of data analytics. The last part provides an overview and discussion of all the above-mentioned articles.

DOI: 10.54254/2753-8818/5/20230332

2. Literature Review

2.1. Statistics and information theory

Csisz´ar introduced a number of existing reviews on Information Theory and Statistics: Tutorials, which mainly deal with information theory related areas. Imre Csisz´ar and Paul Shields present applications of information theory in statistics based on lectures given by Csisz'ar in his early years at the University of Madrid. They choose some typical cases for a more in-depth elaboration. The text mentions the information divergence (I-divergence) or Kullback-Leibler distance or relative entropy, which plays a fundamental role in terms of statistical data [1]. Moreover, A. P. Dempster et al. proposed in 1977 an algorithm with a wide range of applicability for calculating maximum likelihood estimates based on imperfect data counted in various fields [2]. Moreover, the iterative algorithm for minimizing I-divergence under linear constraints proposed by Darroch and Ratcliff, 1986, is equivalent to a circular iteration of the explicitly executable I-projection operation [3].Imre Csiszar, 1991, proposed the possibility that there exists a vector that can contain any one of the feasible sets defined by linear constraints chosen logically consistent uniform description [4]. Shortly thereafter, Willems and T. Tjalkens introduced a program to perform "double mixing" for sequential universal data compression of binary tree sources. The encoding distribution corresponding to the tree source was weighted by using a context tree，and in 1995 a tree source containing an unknown model and unknown parameters achieved a desired encoding distribution [5]. After this, Imre Csiszár and Paul C. Shields showed that for uniformly distributed i.i.d. processes, the Bayesian estimator or the extant minimum description length algorithm (for which the BIC estimator is considered as an approximation) has significant errors.2000 [6]. Furthermore, I. Csisz´ar and F. Matu˘s refinement on the subject of I-projections, the 2003 reverse I-projections, put the focus on linear and exponential families. and introduced generalized maximum likelihood (ML) estimation based on exponential families [7].

2.2. Machine learning based on statistics

The main areas related to information theory include the following reviews. Qing He et al. conducted a general overview of current research on machine learning algorithms based on big data. In addition, parallelization as a mainstream approach for processing big data, some parallel algorithms are presented in the article and then the unresolved machine learning research based on big data problem is presented [8]. Christian Bizer et al. presented a conclusion that engineering of big data is not only urgently needed, but the purposeful application of engineering to the processing of big data is more important. and describe and analyze this issue from four main different perspectives [9]. More importantly, Simon discusses the differences between machine learning and human learning. He says that in machine learning, once an adapted program is debugged, it can be processed and run in a computer as a way to do the same amount of work that would take humans tens of times longer to complete [10]. Furthermore, A. Sagheer et al. proposed a fast feature extraction method based on self-organizing maps (SOM), known as FSOM. FSOM overcomes the traditional slowness. Not only that, they investigated the superiority of the new method [11]. Next, in Quevedo's paper, an algorithm is proposed to rank the input variables as a way to compare their usefulness in the learning task. This algorithm combines simple and classical techniques, on top of the base, which allows the construction of a faster algorithm that improves the computational speed [12]. After that, Hua et al. checked some approaches of basic feature-selection in settings involving thousands of characteristics. The data used to support the conclusion are from both synthetic data baed on model and real data. distribution models were defined and involved the difference between numbers of markers (useful features) and non-markers (useless features). Different kinds of relations among the features was also in discussion. Under this framework, the performances of these algorithms for different classifiers and distribution models were evaluated.[13]. In the end,

DOI: 10.54254/2753-8818/5/20230332

Papadimitriou and Sun described previous discover in applying Map-Reduce, through the process from amateur data to mature data, on a significant mining mission. Co-clustering are specifically focused on, as it has been used in many domains such as text mining, filtering collaboratively and bio-informatics. The authors resented the distributed co-clustering framework, which provides practical methods for distributed data in a pre-process way [14].

2.3. Applications

There are many related applications using information theory and machine learning. To begin with, R.Iniesta and D. Stahl introduced the concept of Big Data in different fields, especially in the psychiatric research area. In the psychiatric study, the benefits of using statistical learning to processing is discussed, from different aspects of practical clinical and academic [15]. What's more, Jiao and Du reviewed high frequency of used evaluation methods and performance metrics for predictors of Bioinformatics，as machine learning technology has been applied in many existing predictors of bioinformatics[16]. On the other hand, Abodayeh et al. provided an idea of a combination of statistical testing hypothesis and machine learning in order to improve fault detection performance in photovoltaic (PV) systems [17]. After that, Ramalho et al. proposed that mechanistic models and statistical models given by measurable data should be used to predict indoor air quality (IAQ), which is defined by the density of pollutions of indoor air. Practical statistical modelling methods were revisited and discussed their outstanding points and drawbacks [18]. Wei Yu and Griffith discussed the uncertainties of quantifying in energy demanding fields. They established methods to build distributions of energy usage based on statistical modelling analysis [19]. Ultimately, Attneave and Fred summarized existing informational methods used in psychological research and illustrates the methods of calculating some of the measures including quantitative expressions of uncertainty and redundancy from qualitative examples, informational methods for analyzing sequences of events, and so on [20].

In summary, the application areas of information theory and machine learning include psychiatric research, biological prediction, circuit fault detection, mathematical modeling, psychological research, etc.

3. Discussion

3.1. Information Theory and Statistics

Information theory, as a part of statistics, is more applied in practice than the purely mathematical part of it.Coding and Cryptography, for example, both use the mathematical logic embedded in information theory. In the early days of messaging, the problem of using only one or two bits to represent a particular meaning was discovered, because there was no way to avoid interference during transmission, transfer, and reception, so characters containing only one or two bits were prone to errors and therefore could not be correctly identified. And with the trend of processing tens and hundreds of bits in a single character, the cost of transportation is rising. Furthermore, information theory, as the basis of modern communication, provides tools for traditional descriptive statistics: the Kullback-Leibler divergence, the Jensen-Shannon divergence, and the Wasserstein metric. These tools are more or less based on entropy calculations. Entropy, as a quantity that describes "uncertainty", is widely used in various fields such as artificial intelligence (AI) to generate avatars. When AI needs to approximate the probability of the pixel distribution of the generated "face" to the natural case, a tool is needed to determine the probability distribution. The Kullback-Leibler distance is the first and slightly crude tool that has been used, with the disadvantage of asymmetry. And this drawback was also refined by the Jensen-Shannon divergence. Similarly, the second tool has the inevitable drawback of being mathematically unsolvable in the face of discrete random variables when the data are brought in. The last one, the Wasserstein metric, can circumvent all these problems, however, as an equation that needs to find all possible solutions, it requires too much arithmetic power to be universally applicable.

DOI: 10.54254/2753-8818/5/20230332

3.2. Machine Learning and Statistics

Machine learning has been called glorified statistics, as most of the machine learning tasks need more or less statistical methods to process the data processing For example, the decision tree-based algorithms, including using the Gini factor or the entropy to calculate the information gain or the information gain ratio are part of the statistics. This then evolved to other variants of decision tree-based methods, including the bagging tree, AdaBoost, gradient boosting, XGB, random forest, etc. All the other methods are some kind of ensemble of the original model, which, statistically, gives better results than the original single model.

In other machine learning-based methods, statistical methods are also widely applied. For example, the neural network, considered as the multiple layer perceptron-based algorithms including the basic MLP and the other algorithms like a convolutional neural network, all have statistical methods used to improve the performance. Some use it to calculate the loss function, and statistically when people try to understand why CNN can perform better on image-based programs, or transformer-based algorithms are trying to understand languages, statistical measures provide a great insight into how these methods can be well explained. This could help people have a better understanding of AI, especially model developers.

On the other hand, the statistical measures can also be used to compare the difference between the output and the input distribution, which is widely used in generation tasks with the GAN algorithm. This is to say, that when two objects are statistically similar, they can be the same or belong to the same group. This makes the ‘fake’ output cannot be distinguished from the ‘real’ input data.

4. Conclusion

In this paper, the main development history of informatics and statistics-based machine learning is reviewed and discussed, and also certain reflections on these two aspects are discussed and analyzed and future trends are discussed. The paper could not include more detailed literature due to the limit of space, and the hard stab focuses on the two aspects mentioned above. What’s more, data can be better explored and analyzed in more areas in the future. There could be many potential cross-discipline research areas covering information theory-based applications, such as generating information theory-based features in machine learning, and data mining, and applying more advanced calculation methods to describe the trend and pattern of a certain dataset.

References

[1] Csiszár, I., & Shields, P. C. (2004). Information theory and statistics: A tutorial. Foundations and Trends® in Communications and Information Theory, 1(4), 417-528.

[2] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

[3] Csiszar, I. (1989). A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. The Annals of Statistics, 1409-1413.

[4] Csiszar, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. The annals of statistics, 19(4), 2032-2066.

[5] Willems, F. M., Shtarkov, Y. M., & Tjalkens, T. J. (1995). The context-tree weighting method: Basic properties. IEEE transactions on information theory, 41(3), 653-664.

[6] Csiszár, I., & Shields, P. C. (2000). The consistency of the BIC Markov order estimator. The Annals of Statistics, 28(6), 1601-1619.

[7] Csiszár, I., & Matus, F. (2003). Information projections revisited. IEEE Transactions on Information Theory, 49(6), 1474-1490.

[8] HE Qing,LI Ning,LUO Wen-Juan,SHI Zhong-Zhi （2014）. A Survey of Machine Learning Algorithms for Big Data. Pattern recognition and artificial intelligence, 27(4) : 327-336.

[9] Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2012). The meaningful use of big data: four perspectives--four challenges. ACM Sigmod Record, 40(4), 56-60.

DOI: 10.54254/2753-8818/5/20230332

[10] Simon, H. A. (1983). Why should machines learn?. In Machine learning). Morgan Kaufmann. 25-37.
[11] Sagheer, A., Tsuruta, N., Taniguchi, R. I., Arita, D., & Maeda, S. (2006, August). Fast feature extraction approach for multi-dimension feature space problems. In 18th International Conference on Pattern Recognition (ICPR'06) . IEEE. 3, 417-420.
[12] Quevedo, J. R., Bahamonde, A., & Luaces, O. (2007). A simple and efficient method for variable ranking according to their usefulness for learning. Computational statistics & data analysis, 52(1), 578-595.
[13] Hua, J., Tembe, W. D., & Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition, 42(3), 409-424.
[14] Papadimitriou, S., & Sun, J. (2008, December). Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining. In 2008 Eighth IEEE International Conference on Data Mining. 512-521.
[15] Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. Psychological medicine, 46(12), 2455-2465.
[16] Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quantitative Biology, 4(4), 320-330.
[17] Fazai, R., Abodayeh, K., Mansouri, M., Trabelsi, M., Nounou, H., Nounou, M., & Georghiou, G. E. (2019). Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems. Solar Energy, 190, 405-413.
[18] Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J. C., & Mandin, C. (2019). Machine learning and statistical models for predicting indoor air quality. Indoor Air, 29(5), 704-726.
[19] Yu, W., An, D., Griffith, D., Yang, Q., & Xu, G. (2015). Towards statistical modeling and machine learning based energy usage forecasting in smart grid. ACM SIGAPP Applied Computing Review, 15(1), 6-16.
[20] Attneave, F. (1959). Applications of information theory to psychology: A summary of basic concepts, methods, and results.