Statistically-sound Knowledge Discovery from Data

Matteo Riondato*

Abstract

Knowledge Discovery from Data (KDD) has mostly focused on understanding the available data. *Statistically-sound* KDD shifts the goal to understanding the partially unknown, random Data Generating Process (DGP) process that generates the data. This shift is necessary to ensure that the results from data analysis constitute new knowledge about the DGP, as required by the practice of scientific research and by many industrial applications, to avoid costly *false discoveries*.

In statistically-sound KDD, results obtained from the data are considered as *hypotheses*, and they must undergo *statistical testing*, before being deemed significant, i.e., informative about the DGP.

The challenges include *(1)* how to subject the hypotheses to *severe testing* to make it hard for them to be deemed significant; *(2)* considering the simultaneous testing of *multiple hypotheses* as the default setting, not as an afterthought; *(3)* offering flexible statistical guarantees at different stages of the discovery process; and *(4)* achieving scalability along multiple axes, from the size of the data to the number and complexity of hypotheses to be tested.

Success for Statistically-sound KDD as a field will be achieved with *(1)* the introduction of a rich collection of null models that are representative of the KDD tasks, and of the existing knowledge of the DGP by field experts; *(2)* the development of scalable algorithms for testing results for many KDD tasks on different data types; and *(3)* the availability of benchmark dataset generators that allow to thoroughly evaluate these algorithms.

1 Introduction

The Data Mining and Knowledge Discovery from Data (KDD) research community has created ingenious algorithms for many tasks (e.g., pattern mining, anomaly detection, graph analysis), on all kinds of data (from transactional to sequence datasets, to graphs, to time series), both static and time-evolving. These methods are used by practitioners and companies for all kinds of data analysis, from logistics, to cybersecurity, to customer analysis. KDD algorithms also found their way to research labs in all sciences, such as microbiology and genomics, to study combinations of gene mutations, or protein interactions [10, 21–23]. Nevertheless, most KDD methods lack the ability to give *strong statistical guarantees on their results*. A *rigorous statistical assessment* of the results, performed through *statistical hypothesis testing* [15] is a *necessary* step in the modern process of scientific discovery. This need goes beyond scientific labs: KDD practitioners in industry need their results to be *trustworthy and actionable*, in order for decision-making based on them to be effective. For example, when using subgraph-based methods for detecting cyberattacks in computer networks, it is necessary not only that as many as possible true attacks are detected, but also that as few false flags as possible are raised. *False discoveries* may arise because the available data gives only a partial, noisy, representation of the random Data Generation Process (DGP).

Our blue sky idea, which we call *statistically-sound KDD*, transforms the field of data mining by shifting the focus of the KDD process from extracting information about the dataset, to *obtaining new understanding* (i.e., knowledge) of the DGP. The need for this shift has long been recognized by the KDD community [27], but progress has so far been limited.

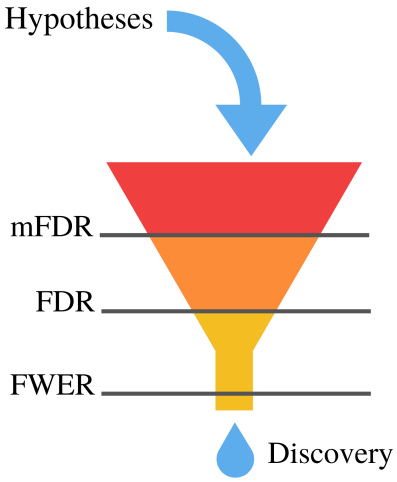

Figure 1: The discovery funnel, with the stages of hypothesis testing and the different measures for false discovery control at each stage.

2 Challenges

Algorithms for statistically-sound KDD consider the results (e.g., patterns, anomalies, clusters, graph/edge/vertex properties) as *hypotheses*, and perform *statistical tests* on them, marking as *(statistically) significant* those for which there is sufficient

*Amherst College. mriondato@amherst.edu. This work was supported in part by NSF award 2006765.

evidence that they give *new knowledge about the DGP*, and discarding the others as due to the randomness of the process itself or not offering new information. To make *statistically-sound KDD* possible, we must address the following challenges.

1. Severe testing Statistical assessment of hypotheses must be *severe* [17], i.e., it must be hard for results to be deemed significant. Severe testing requires *representative null models*. A *null model* is a collection of *possible* datasets that the unknown process *may* generate, and a probability distribution over this collection. The null model captures, in some sense, what is assumed or already known about the data generating process, and the results are assessed against it to understand in what way they cannot be explained by the existing knowledge or assumptions. The choice of the null model by the user must be *deliberate and informed*, as the meaning of "statistically significant" depends on the null model. For example, the results deemed significant under one null model cannot in general be compared to those deemed significant under a different null model. Nevertheless "all models are wrong, but some are useful" (George E. P. Box), and some null models may be more appropriate for testing the significance of the results of a KDD task than others, because they *more closely represent the settings of the task*. Defining such models is therefore imperative. A second requirement of severe testing is that the quantities used to perform the tests (e.g., test statistics, empirical p-values) are *conservative*, to avoid wrongly marking results as significant. Statistically-sound KDD methods must satisfy these requisites and perform severe testing, to ensure a trustworthy assessment of the results.

2. Testing multiple hypotheses The results of most KDD tasks are composed by a large number of quantities (e.g., the collection of interesting patterns, or a score for each vertex in a graph). More importantly, the practice of science today *requires* testing *multiple hypotheses*: a scientist does not formulate a single promising hypothesis, and then tests it with a well-crafted experiment on "perfect" data. Rather, scientists consider a *family $\mathcal{H}$ of hypotheses*, a member of which *may* explain the phenomenon under study. For example, no molecular biologist would ever test the single hypothesis that one specific combination of gene mutations is much more often present in individuals with some disease than in healthy individuals. They would instead ask whether *any* combination of gene mutations is significantly more frequent among individuals with the disease than healthy individuals, thus testing one hypothesis per combination of mutations. The process of scientific research is then akin to a multi-stage distillation process, or to a *funnel with intermediate filters* (Fig. 1): the entire *family $\mathcal{H}$ of hypotheses* is "poured" into the funnel, and the intermediate filters, which represent different *stages of hypothesis testing* (discussed below), prevent unpromising hypotheses, i.e., those deemed to be non-significant on the available data, from proceeding further. Any hypothesis that "drips" out of the funnel is considered a *discovery*. The hypotheses arriving at each filtering stage are *tested simultaneously on the same data*, highlighting the need for a *multiple-hypothesis first* approach to testing.

3. Offering flexible statistical guarantees Multiple stages of hypothesis testing are *necessary*: passing a single (severe) test is not sufficient to declare a discovery: it just gives partial evidence that the hypothesis is worth further investigation. There are two *competing goals* in designing the hypothesis testing stages: *(1) minimizing false discoveries*, i.e., false hypotheses that appear to be significant on the available data *due to the randomness in the data generating process and possibly in the testing procedures*; and *(2) maximizing statistical power*, i.e., the probability that a true hypothesis is deemed significant. A procedure can avoid any false discovery (resp. guarantee all true hypotheses are deemed significant, thus achieving maximum statistical power) by simply not marking any hypothesis as significant (resp. marking all hypotheses as significant), but this procedure would incur in zero statistical power (resp. would maximize the number of false discoveries). Thus, it is necessary to balance these two goals. The trade-off point may differ depending on "how deep into the funnel" the testing is performed: at the early stages, it is convenient to tilt towards statistical power, while at the last stage, minimizing the probability of false discoveries becomes imperative, as it is the last chance. The statistical literature offers different metrics to quantify the *guarantees for false discovery control*, e.g., the Family-Wise Error Rate (FWER) [6], the False Discovery Rate (FDR) [5], and the marginalized FDR (mFDR) [11]. Statistically-sound KDD methods must offer *flexible guarantees* by controlling these measures, in order to be applicable at every stage of the discovery process.

4. Scaling along multiple axes Data mining methods should be scalable in terms of the dataset size, but algorithms for statistically-sound KDD must also scale well along the axes of the *number* and *complexity* of the hypotheses to be tested. As an example, the Human Protein Reference Database [20] protein-protein interaction network has $\approx$ 19000 proteins and $\approx$ 37000 interactions between them, and scientists are interested in understanding the significance of relatively small connected subgraphs in this network, representing pathways in cancer cells. There are more than 10^{13} sub-

graphs of size 8, each corresponding to an hypothesis. It is imperative that statistically-sound KDD methods can extract such a large number of patterns and test the corresponding hypotheses as fast as possible. But "scalability" must be considered not only in the computational sense, but also w.r.t. the *statistical properties* of false discovery control and power. With reference to Fig. 1, hypotheses that arrive at each filtering stage are *tested simultaneously on the same data*. But most procedures for multiple hypothesis testing are not designed according to this paradigm, rather they assume that *each test is performed in isolation*. This assumption creates *computational and statistical drawbacks*: computations are unnecessarily replicated, *limiting the scalability* and throughput of the filtering stage, while considering each test as an separate task *prevents from leveraging the structure (broadly defined) of the set of hypotheses being tested*, resulting in *lower statistical power*, i.e., fewer discoveries. Algorithms for statistically-sound KDD must exploit this structure and scale well along both computational and statistical axes.

5. Offering practical methods on different tasks on rich data Long gone is the era when KDD methods only had to deal with binary tabular data, static graphs, and similar "simple" data. Today, practitioners want to *extract complex information from rich, evolving datasets*. Statistically-sound KDD methods must be able to assess results obtained on such datasets (e.g., utility transactional datasets, attributed graphs, multi-valued time series), while taking into account their *dynamic nature*, as in temporal networks. Additionally, many of the aforementioned challenges are, although "theoretically" solved given the many methods available (e.g., to control the FWER or the FDR), *not solved in practice*. To find application among practitioners, methods for statistically-sound KDD must be *designed around the practitioners' needs*, and their implementations must be *thoroughly evaluated* to ensure that their performance is satisfactory.

3 Achieving Success

Some initial work towards statistically-sound KDD has been done [12, 19], but it mostly failed to address the challenges: it tested hypotheses w.r.t. simple null models, while using approximate test statistics with no correction; when controlling for multiple hypotheses, it mostly focused on the FWER, thus not offering the desired flexible guarantees; it was rarely scalable, as it, in the best cases, relied on Markov-Chain Monte-Carlo (MCMC) methods with slow mixing time; and it was limited to simple data and tasks, such as binary transactional datasets for itemset mining, or static graphs. To achieve Statistically-sound KDD, we propose to we tackle the challenges as follows.

- Define *realistic null models for different KDD tasks* (1^{st} challenge), informed by the needs of practitioners, from industry to scientific research labs. To start, we suggest starting from well-established tasks such as various forms of *pattern mining* [1–3, 24], evaluation of vertex/edge properties such as centrality measures [18] and graph structural properties such as clustering coefficients and core decomposition. Another promising task is the statistically-sound identification of anomalies, e.g., in network traffic which may correspond to security breaches or attacks.

- Derive *simultaneous confidence intervals for p-values*, to be used in the testing procedures to ensure that control of false discoveries is at the user-specified level at all times, as required by severe testing (1^{st} challenge), not just asymptotically. As a starting point, uniform convergence results based on *variance-aware Rademacher Averages* [4, 7, 14] can likely be used to derive tight confidence intervals with little impact on statistical power.

- Design *methods to directly test multiple hypotheses* (2^{nd} challenge), by embracing the *resampling-based approach to hypothesis testing* [25], which, by approximating the distribution of the p-values, enables the *flexible guarantees* that we seek (3^{rd} challenge), by controlling, as needed, the FWER and/or the (m)FDR [8, 9, 26]. Resampling-based methods leverage the *structure of the hypothesis family*, resulting in the desired high statistical power, and scaling well with the family's size and complexity (4^{th} challenge).

- Develop *efficient sampling procedures to quickly generate datasets* from the null model, as required by the resampling-based approach. Such algorithms should include both *fast-mixing MCMC methods* [16] and *exact-sampling approaches* [13], that scale well along multiple axes (4^{th} challenge).

- Subject the methods to a *thorough empirical evaluation* (5^{th} challenge) by assessing their scalability along both computational and statistical axes (4^{th} challenge). To help thorough evaluation of statistically-sound KDD methods beyond this project, we suggest the *development of artificial dataset generators* which allow the KDD researchers to plant true hypotheses in the generated data, so as to evaluate the tightness in the control of false discoveries and in the statistical power.

4 Conclusion

Shifting from KDD to statistically-sound KDD is absolutely needed as the abundance of data, which is now a fact in many scientific areas and in many KDD use cases, is accompanied by an abundance of questions that are asked on these data, and when the answers to these questions are used for decisions that may impact a large portion of the population, e.g., to create policies or develop drugs. False discoveries in such settings are just not acceptable. Like others before us [27], we call the research community to action: there are fascinating computational and statistical challenges to solve, and the opportunity to have a large impact, from enabling a faster and higher-throughput scientific discovery pipeline, to ensuring better use of data by private companies and governments.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB '94, pages 487–49. 1994.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In ICDE'95, pages 3–14. 1995.

[3] M. Atzmueller. Subgroup discovery. *Wiley Interdisc. Rev.: Data Mining and Knowl. Disc.*, 5(1): 35–49, 2015.

[4] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Nov):463–482, 2002.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B (Methodol.)*, 57(1):289–300, 1995.

[6] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. Regio Istit. Sup. Sci. Econ. Commerc. Firenze*, 8:3–62, 1936.

[7] C. Cousins and M. Riondato. Sharp uniform convergence bounds through empirical centralization. In NeurIPS'22, pages 15123–15132. 2020.

[8] S. Denkowska. Controlling the effect of multiple testing in big data. *Math. Econ.*, 10(17):5–16, 2014.

[9] S. Dudoit, H. N. Gilbert, and M. J. van der Laan. Resampling-based empirical bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study. *Biometr. J.*, 50(5):716–744, 2008.

[10] E. Ferkingstad, L. Holden, and G. K. Sandve. Monte Carlo null models for genomic data. *Stat. Sci.*, 30(1):59–71, 2015.

[11] D. P. Foster and R. A. Stine. α-investing: A procedure for sequential control of expected false discoveries. *J. Roy. Stat. Soc. Ser. B (Methodol.)*, 70(2):429–444, 2008.

[12] W. Hämäläinen and G. I. Webb. A tutorial on statistically sound pattern discovery. *Data Mining Knowl. Disc.*, 33(2):325–377, 2019.

[13] S. Jenkins, S. Walzer-Goldfeld, and M. Riondato. SPEck: mining statistically-significant sequential patterns efficiently with exact sampling. *Data Mining Knowl. Disc.*, 36(4):1575—1599, 2022.

[14] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Th.*, 47 (5):1902–1914, July 2001.

[15] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 4 edition, 2022.

[16] D. A. Levin and Y. Peres. *Markov chains and mixing times*. Am. Math. Soc., 2nd edition, 2017.

[17] D. G. Mayo. *Statistical inference as severe testing*. Cambridge Univ. Press, 2018.

[18] M. E. J. Newman. *Networks – An Introduction*. Oxford Univ. Press, 2010.

[19] L. Pellegrina, M. Riondato, and F. Vandin. Hypothesis testing and statistically-sound pattern mining. In KDD '19, pages 3215–3216. 2019.

[20] S. Peri et al.. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, 13 (10):2363–2371, Oct 2003.

[21] R. T. Relator, A. Terada, and J. Sese. Identifying statistically significant combinatorial markers for survival analysis. *BMC Med. Genom.*, 11(2):31, 2018.

[22] J. Sese, A. Terada, Y. Saito, and K. Tsuda. Statistically significant subgraphs for genome-wide association study. In *Stat. Sound Data Mining*, pages 29–36, 2014.

[23] A. Terada, K. Tsuda, and J. Sese. Fast Westfall-Young permutation procedure for combinatorial regulation discovery. In BIBM'13, pages 153–158. 2013.

[24] T. Truong-Chi and P. Fournier-Viger. A survey of high utility sequential pattern mining. In *High-Utility Pattern Mining*, pages 97–129. 2019.

[25] P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

[26] D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Infer.*, 82(1-2):171–196, 1999.

[27] A. Zimmermann. The data problem in data mining. *SIGKDD Explor.*, 16(2):38–45, 2014.