Hypothesis Testing and Statistically-sound Pattern Mining

Leonardo Pellegrina
Dept. of Information Engineering
Università di Padova
Padova, Italy
pellegri@dei.unipd.it

Matteo Riondato
Dept. of Computer Science
Amherst College
Amherst, MA, USA
mriondato@amherst.edu

Fabio Vandin
Dept. of Information Engineering
Università di Padova
Padova, Italy
fabio.vandin@unipd.it

ABSTRACT

The availability of massive datasets has highlighted the need of computationally efficient and statistically-sound methods to extracts patterns while providing rigorous guarantees on the quality of the results, in particular with respect to false discoveries. In this tutorial we survey recent methods that properly combine computational and statistical considerations to efficiently mine statistically reliable patterns from large datasets. We start by introducing the fundamental concepts in statistical hypothesis testing, including conditional and unconditional tests, which may not be familiar to everyone in the data mining community. We then explain how the computational and statistical challenges in pattern mining have been tackled in different ways. Finally, we describe the application of these methods in areas such as market basket analysis, subgraph mining, social networks analysis, and cancer genomics.

CCS CONCEPTS

• **Mathematics of computing → Contingency table analysis**;
• **Information systems → Data mining**.

KEYWORDS

Family-wise Error Rate, Hypothesis Testing, Itemset Mining

ACM Reference Format:
Leonardo Pellegrina, Matteo Riondato, and Fabio Vandin. 2019. Hypothesis Testing and Statistically-sound Pattern Mining. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332286

1 INTRODUCTION

The extraction of patterns from data has traditionally been approached from two different directions: one focusing on computational aspects, typical of data mining, which assumes that the data is the complete representation of a process/system, and one focusing on inferential aspects, typical of statistics, which considers the data as a partial and noisy collection of measurements of the underlying process/system, and which evaluates the significance of a pattern using the rigorous framework of statistical hypothesis testing. While these two points of view are deeply related, the methods developed focusing on the computational aspects have neglected the inferential aspects, and vice versa, with almost no connection till recent years. The availability of massive and rich datasets, where a gargantuan number of patterns needs to be processed and evaluated, has highlighted the need for computationally efficient methods that properly assess the statistical soundness of candidate patterns in order to avoid false discoveries. The development of such methods poses severe challenges from both the computational and the statistical side, since the multitude of candidate patterns, each corresponding to an hypothesis regarding the underlying process/system, leads to a severe multiple hypothesis testing problem. Various methods have been proposed to tackle such challenges by properly integrating computational and statistical considerations in the mining process. These methods have already been successfully applied in several areas, ranging from social networks to cancer genomics. The relevance of this area of research will only increase as analysts want to extract more and more complex patterns from larger and larger datasets.

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332286

2 TUTORIAL OUTLINE

We start with an introduction to the fundamental concepts behind statistical hypothesis testing [27, Ch. 10], and the key questions that will be answered in the rest of the tutorial. In particular, we first introduce the framework of testing a single hypothesis (defining, e.g., what a null hypothesis is) and example applications where testing hypothesis is crucial, such as in biomedical research and in the study of social networks. We then discuss fundamental tests such as Fisher's exact test [7] and the related χ^2 and Barnard's test [2]. We also briefly mention A/B testing, although the focus of the tutorial is on pattern mining where such tests are rarely used. The final part of the introduction covers issues arising from testing multiple hypotheses on the same data and how to address these issues: we outline how and why the probability of discovering false positives grows in such scenarios, and how to control for this growth by bounding different metrics, such as the Family-Wise Error Rate (FWER) [5, 13] and the False Discovery Rate (FDR) [3, 4].

In the central part, we focus on mining statistically-sound patterns. We first define the problem and highlight its computational and statistical challenges arising from the combinatorial explosion of the number of hypotheses being tested and from the sheer size of data [11, 24, 30]. We then tackle these challenges one by one. We discuss how to make the process of finding statistically significant patterns efficient from a *computational* point of view [9, 18, 20, 25]. Specifically, we discuss efficient permutation testing [9, 18], the groundbreaking LAMP method [25] which allows to apply Tarone [24]'s method to combinatorial patterns, TopKWY [20], which efficiently extracts the k most statistically significant patterns while preserving guarantees on the FWER, and SPuManTE [19], which enables significant pattern mining with unconditional tests. The *statistical* efficiency is covered next: the works presented here [15, 22, 28, 29] introduce different methods to increase the statistical power of methods to extract significant patterns while controlling the FWER, and to deal with different inferential aspects of pattern mining. This part is the core of our tutorial: in these works, statistics and data mining come together and create a positive multiplier to obtain fast and statistically-sound methods for pattern mining.

We then overview other interestingness measures and classes of patterns which, although not based on hypothesis testing, are grounded in statistics and therefore relevant to this tutorial, such as emerging [17] and discriminative patterns [12], significant association rules [10], and subgroups [1]. All these patterns are interesting on their own, and their presentation allows us to perform a comparison of different approaches. A discussion of applications of the presented methods, ranging from the mining of significant subgraphs and motifs from large graphs [23], to biomedicine [26] and computational biology [8], will be provided.

In the final third part, we focus on more advanced material. Specifically, we show how to remove the assumptions on the data generating process [6], which have classically been used to make the problem more tractable. We also discuss how to weight hypotheses in a data-dependent way, with the goal of increasing the statistical power [14]. The materials covered here are recent developments that should interest the attending researchers, as will the potential future directions that complete the tutorial.

The outline of the tutorial is the following.

1. **Introduction and Theoretical Foundations**
   - 1.1 Testing a single hypothesis: setting, basic concepts, and applications [27, Ch. 10]
   - 1.2 Fundamental tests: Fisher's test [7], χ^2 test [27, Section 10.3], Barnard's test [2], A/B testing [16]
   - 1.3 The challenge of testing multiple hypotheses: Family-Wise Error Rate [5] and False Discovery Rate [3]
   - 1.4 Taming the challenge: the Bonferroni-Holm procedure [13] and the Benjamini-Yekutieli correction [4]
2. **Mining Statistically-Sound Patterns**
   - 2.1 Computational and statistical challenges in pattern mining [11, 24, 30]
   - 2.2 Computational aspects: LAMP [25], permutation testing [9, 18], TopKWY [20], SPuManTE [19]
   - 2.3 Statistical aspects: hold-out approach and layered critical values [28, 29], a threshold for significant pattern mining [15], true frequent itemsets [22]
   - 2.4 Other statistical measures: emerging patterns [17], discriminative patterns [12], significant association rules [10], significant subgroups [1]
   - 2.5 Applications: subgraph mining [23], cancer genomics [26], computational biology [8], and survival analysis [21]
3. **Recent developments and advanced topics**
   - 3.1 Removing assumptions on the data generating process [6]
   - 3.2 Data-dependent hypothesis weighting [14]
   - 3.3 Conclusions, future directions, and discussion

REFERENCES

[1] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[2] George Alfred Barnard. A new test for 2×2 tables. *Nature*, 156:177, 1945.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. of the Royal Statistical Society. Series B*, pages 289–300, 1995.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[5] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[6] Leena Choi, Jeffrey D. Blume, and William D. Dupont. Elucidating the foundations of statistical inference with 2×2 tables. *PloS ONE*, 10(4):e0121263, 2015.

[7] Ronald A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. of the Royal Statistical Society*, 85(1):87–94, 1922.

[8] Yoshinori Fukasawa, Junko Tsuji, Szu-Chin Fu, Kentaro Tomii, Paul Horton, and Kenichiro Imai. Mitofates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Molecular & Cellular Proteomics*, 14(4):1113–1126, 2015.

[9] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. *ACM Trans. on Knowledge Discovery from Data (TKDD)*, 1(3):14, 2007.

[10] Wilhelmiina Hämäläinen and Matti Nykänen. Efficient discovery of statistically significant association rules. In *2008 Eighth IEEE International Conference on Data Mining*, pages 203–212. IEEE, 2008.

[11] Wilhelmiina Hämäläinen and Geoffrey I Webb. A tutorial on statistically sound pattern discovery. *Data Mining and Knowledge Discovery*, 33(2):325–377, 2019.

[12] Zengyou He, Simeng Zhang, and Jun Wu. Significance-based discriminative sequential pattern mining. *Expert Systems with Applications*, 2018.

[13] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[14] Nikolaos Ignatiadis, Bernd Klaus, et al. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577, 2016.

[15] Adam Kirsch, Michael Mitzenmacher, Andrea Pietracaprina, Geppino Pucci, Eli Upfal, and Fabio Vandin. An efficient rigorous approach for identifying statistically significant frequent itemsets. *J. of the ACM*, 59(3):12, 2012.

[16] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the Web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[17] Junpei Komiyama, Masakazu Ishihata, Hiroki Arimura, Takashi Nishibayashi, and Shin-ichi Minato. Statistical emerging pattern mining with multiple testing correction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 897–906. 2017.

[18] Felipe Llinares-López, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten Borgwardt. Fast and memory-efficient significant pattern mining via permutation testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 725–734. 2015.

[19] Leonardo Pellegrina, Matteo Riondato, and Fabio Vandin. SPuManTE: Significant Pattern Mining with Unconditional Testing. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*, 2019.

[20] Leonardo Pellegrina and Fabio Vandin. Efficient mining of the most significant patterns with permutation testing. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, 2018.

[21] Raissa T Relator, Aika Terada, and Jun Sese. Identifying statistically significant combinatorial markers for survival analysis. *BMC Med. Genomics*, 11(2):31, 2018.

[22] Matteo Riondato and Fabio Vandin. Finding the true frequent itemsets. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 497–505. SIAM, 2014.

[23] Mahito Sugiyama, Felipe Llinares-López, Niklas Kasenburg, and Karsten M. Borgwardt. Significant subgraph mining with multiple testing correction. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 37–45. SIAM, 2015.

[24] Robert E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.

[25] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32):12996–13001, 2013.

[26] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18 (3):507–522, 2011.

[27] Larry Wasserman. *All of Statistics: A concise course in statistical inference.* Springer, 2013.

[28] Geoffrey I. Webb. Discovering significant patterns. *Machine learning*, 68(1):1–33, 2007.

[29] Geoffrey I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2-3):307–323, 2008.

[30] Peter H. Westfall and Stanley S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* Wiley-Interscience, 1993.