Essay 0 - Interestingness and Frequency

Najwa Laabid, University of Eastern Finland April 26, 2020

Data mining is concerned with extracting insights from data. Naturally, recurring patterns are a good place to start this exploration. A vast amount of research goes into designing methods to quantify the value of said patterns, reduce their number to a manageable size, and incorporate user's interest in both processes. Within this scope, Geng and Hamilton (2006) proposes an overview of current methods to measure interestingness, Gallo et al. (2007) presents an approach to evaluate interestingness while reducing redundancy in mined itemsets, and Webb (2011) suggests top-k mining to overcome the combinatorial search space faced by frequentist approaches. The following paragraphs evaluate the findings of each of these articles and discuss their connections.

Geng and Hamilton (2006) proposes a survey on recent methods for measuring the interestingness of mined patterns. The survey aims at summarizing the key properties of each measure as well as providing strategies for selecting the appropriate measure for a given application. Summaries and rules were the main pattern types investigated in this study. The measures used for these entities were grouped into three categories: objective, subjective and semantics-based. Despite giving a comprehensive overview of the topic, the survey reported less on subject and semantics-based measures, reflecting perhaps a gap in literature. The conclusion proposed areas for further research in the field of pattern mining, including methods to increase the user's involvement in the mining process through modeling their background knowledge and expressing their interest formally.

Gallo et al. (2007) uses interestingness to rank mined frequent itemsets. Mining based on frequency alone returns result sets of combinatorial size. Even when focusing on representative (such as *closed*, *maximal*, or *non-derivable*) patterns, the output can still be overwhelmingly large from the user's perspective. Furthermore, real-world data is noisy, which can increase redundancy within the result set. To overcome both issues (i.e., results too large for the user and redundant patterns), Gallo et al. (2007) proposes an additional processing step, in which representative patterns are ranked according to their informative value, while redundant results are identified and penalized in a statistically sound manner. This way, when the user looks at the top elements of the result set, they are guaranteed to get the most informative and most different (i.e., covering the most cases) patterns. The algorithm proposed uses an objective measure of interestingness based on hypothesis testing. One shortcoming of this method is the computationally expensive process of updating probabilities to account for redundancy. The article also lacked a comparison between their results and those of competing approaches, which may have offered a more comprehensive assessment of their algorithm.

Despite its great potential, pattern mining has had less success than anticipated. Webb (2011) blames researchers' focus on frequentist methods despite their multiple issues, including inability to capture infrequent high-value associations (the *vodka and caviar problem*), lack

of control over the number of associations discovered, and poor handling of dense data (due to the combinatorial nature of the generated rules). To overcome these limitations, Webb (2011) proposes to search for k interesting patterns under user-specified constraints, such as non-redundancy and productivity. Instead of looking for frequent patterns then assessing their worth, this approach automatically returns the k patterns with the highest value for a given interestingness measure. The paper also addresses the issue of false discoveries (defined as associations that appear relevant in a sample but may not generalize to the population) by presenting three techniques to avoid them: within-search Bonferroni correction, holdout evaluation, and randomization testing. Finally, the author emphasizes his support for expressing association mining as itemsets rather than rules, since the former model proposes more concise results, especially for an initial overview of the associations present in the data. One challenge facing the top-k mining paradigm is the choice of the interestingness measure. It may be difficult for a user to define a suitable formulation of their interest, especially in an exploratory analysis, in which the goal may be to reveal trends in the data without a foreseeable application.

The three articles look at different aspects of association mining. Overall, these readings complement the material covered in class on frequent itemset mining, by presenting advanced techniques from different paradigms. The last article raised a good point on the merits of frequentist methods, as *frequent* does not always mean *interesting*, while it can easily imply an overwhelming number of results. I still believe frequent patterns to be interesting to study in their own right. Even when they do not answer user specific queries, these patterns may hint to unsuspected trends or at least have a reason behind their multiple appearances. As for practical use, I think further research in modeling the user's interest and knowledge can push corresponding queries to their full potential. Until then, top-k and frequentist methods can evolve as two parallel paradigms, each focusing on different goals within association mining.

References

Arianna Gallo, Tijl De Bie, and Nello Cristianini. Mini: Mining informative non-redundant itemsets. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, pages 438–445, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74976-9.

Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9–es, September 2006. ISSN 0360-0300. doi: 10.1145/1132960.1132963. URL https://doi.org/10.1145/1132960.1132963.

Geoffrey I. Webb. Filtered-top-k association discovery. *WIREs Data Mining and Knowledge Discovery*, 1(3):183–192, 2011. doi: 10.1002/widm.28. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.28.