A Critical View on Automatic Significance-Filtering in Pattern Mining

Florian Lemmerich and **Frank Puppe**
lastname@INFORMATIK.UNI-WUERZBURG.DE

Artificial Intelligence and Applied Computer Science Group,
University of Wuerzburg,
Am Hubland, 97074 Wuerzburg, Germany

Editors: Wilhelmiina Hämäläinen, Francois Petitjean, and Geoff Webb

Abstract

Statistically sound validation of results plays an important role in modern data mining. In this context, it has been advocated to disregard patterns that cannot be automatically confirmed as statistically valid by the available data. In this short position paper, we argue against a mandatory automatic significance filtering of results.
Keywords: Data Mining, Pattern Mining, Statistics, Filtering, Position paper

1. Introduction

Knowledge discovery in databases is deeply rooted in classical statistics. In one of its most popular definitions, it has been characterized as the "non-trivial process of identifying *valid*, novel, potentially useful and ultimately understandable patterns" (Fayyad et al., 1996). Here, *valid* explicitly describes the desire for statistically sound patterns. This avoids the reporting of patterns, which are caused only by random fluctuations in the data.

In this spirit, the selection of interesting patterns is in many approaches strongly influenced by statistical considerations. Interestingness measures (also called quality functions or evaluations functions) for pattern selection are very often inspired by statistical hypothesis tests, e.g., the χ^2-test (Morishita and Sese, 2000), the binomial measure (Klösgen, 2002), or the Kolmogorov-Smirnov-Test for numerical data (Jorge et al., 2006). An even stricter way to secure statistic validity is, to apply an arbitrary interestingness measure, but to filter all discovered patterns, which do not pass a statistical significance test (Webb, 2011). In order to facilitate the identification of statistically sound patterns additional techniques are applied to handle the *multiple comparisons problem* (Holm, 1979): If many different hypotheses are investigated (as it is usually done in data mining), then some candidates will pass standard significance tests with unadapted significance values by pure chance, see for example (Webb, 2006). Approaches to achieve statistically sound result patterns despite this issue include (i) utilizing a holdout set, (ii) applying an improved Bonferroni-like adaption of the significance threshold (Webb, 2006, 2007), (iii) employing randomization techniques (Mannila, 2008; Duivesteijn and Knobbe, 2011), or (iv) applying bootstrap-based

approaches to control false discoveries (Lallich et al., 2006). These filterings are typically applied automatically without showing the unfiltered results to the end user.

Focusing on the (statistically) significant patterns has been practiced with large success in different domains. This, however, may lead in particular inexperienced data miners to automatically test candidate patterns on statistical significance and reject patterns that do not pass a significance test altogether, regardless of the application scenario. In this short position paper, we argue *against* automatic rejection of statistically insignificant patterns.

One goal of this paper is to encourage critical assessment regarding automatic statistical filtering of data mining results. In particular, we claim that automatic significance filtering of results should be considered carefully in the application context and should not be regarded as mandatory in general. By pointing out some deficiencies of current statistical filtering approaches, we also imply potential directions for future research. In addition, we also provide an (incomplete) list of factors that should be considered for the decision regarding automatic rejection of not-significant results.

Statistical significance testing in general has been criticised for several reasons, some of them fundamental, see for example (Carver, 1978; Cohen, 1994; Ziliak and McCloskey, 2008). In this paper, we do not follow these lines of reasoning, but focus solely on arguments that are specific to *automatic* significance testing of *data mining* results, especially of (supervised) pattern mining.

The rest of the paper is structured as follows: Section 2 presents our arguments against automatic filtering of data mining results based on statistical significance testing. It is structured in four main propositions. Section 3 provides a list of factors, which should be considered if automatic significance filtering is in debate. The paper concludes in Section 4 with a short summary.

2. Propositions

Proposition 1 *Arguably, Data Mining is intended for* ***hypothesis generation****, not hypothesis testing. Data Mining is an* ***exploratory*** *task.*

The ultimate goal of each data analysis is to understand the characteristics and relations of the variables in the dataset. In traditional statistics, exploratory data analysis techniques (Tukey, 1977) have been developed as a complement method to confirmatory statistics. Exploratory data analysis makes use of descriptive statistics and visualization techniques in order to get an overview of the dataset and to generate promising hypotheses. These can then be confirmed or rejected by traditional tests on statistical significance separately. It has been argued, that the task of data mining primarily excels in the exploratory part of the data analysis. For example, the function of subgroup discovery (an important method of data mining) has been explained as "a convenient hypothesis generator for further analysis, not as a statistical oracle that can be blindly trusted" (Wrobel, 2001).

Although a statistical validation of result patterns is desired in many target applications, this is not the case in all scenarios. For example, it might be already advantageous to get a hint to the cause for a certain problem for further investigation. This investigation can be performed by additional data analysis methods, but also by completely different application-inherent methods.

Proposition 2 *If significance testing is performed in an independent process step, better results can be achieved.*

Amongst others, there is a simple and effective standard way to perform exploratory data mining as well as automatic sound statistical validation with the available data: the *holdout* approach splits the data in an exploratory part and a holdout part, which is used for statistical validation (Webb, 2006). However, it can be favorable to perform the holdout evaluation not automatically as part of the data mining process, but in a separated process step that is preceded by an inspection of the patterns by human experts. This yields several advantages:

First, by manual inspection experts can reduce the number of hypotheses, which are validated using the holdout set. E.g., they can remove patterns, which are (semantically) redundant to each other, or they can select only those patterns, which are potentially useful for the target application. Since less patterns are validated with the holdout set after the reduction, the applied correction factor for the multiple comparison procedure (e.g., a Bonferroni or Holm-Bonferroni correction) is less strict, thus allowing more patterns to pass the significance test.

Second, the choice of the significance test can be optimized. The selection of the best-fitting statistical test for a certain task is a non-trivial task. The choice depends strongly on the distributional characteristics of the test data. These characteristics may vary in different parts of the data. For example, a numeric target attribute (as right-hand-side of a rule) might be gaussian distributed for some patterns in the data, but not for the overall dataset. However, in current pattern mining systems, which employ significance filtering, only one fixed significance test is applied for all patterns. To achieve applicability for a wide range of datasets, usually a nonparametric significance test is used. These are often less powerful than their more specialized counterparts. This potentially causes a pattern to be rejected as insignificant, although the pattern could be confirmed by a more powerful statistical test chosen by a human after a manual inspection of the pattern.

Third, if significance testing is performed in a separate phase (apart from the automatic mining algorithm), some application scenarios also allow for simple, target-oriented acquisition of additional validation data. In particular for patterns that occur only seldom in the data, the overall holdout set must be huge in order to confirm this kind of patterns. However, if additional data can be acquired specifically for this pattern, a few dozens of additional cases might be sufficient.

Of course, manual investigation of data mining results requires increased efforts in comparison to automatic validation of the patterns directly in the data mining algorithm. It might, however, also result in more, better, and statistically valid results. Therefore, in many application scenarios it might be favorable to perform data mining first (as hypothesis generation) and test hypotheses later in a separate step. This might specifically be true, if an iterative and interactive process model is applied, as it has been recommended for data mining.

Proposition 3 *Adaptations of statistical significance tests for the multiple comparisons problem are not suited for interactive and iterative mining. The progression of significance values during exploration can be counter-intuitive.*

It has been widely acknowledged that successful knowledge discovery requires an interactive and iterative approach. However, current adaptation methods for sound statistical data mining are heavily focused on a single run of an automatic discovery algorithm. To the authors' knowledge, sound corrections of significance values for the multiple comparison problem are not integrated in any current interactive pattern mining system.

In theory, each hypothesis, which is displayed (or considered for display by an algorithm) influences the correction factor for the multiple comparison procedure. Therefore, the more patterns are interactively explored, the less powerful the statistical significance tests get. This not only is inconvenient to implement, but also discourages the central goal of interactive data exploration: the user should freely navigate in the data in order to understand relations and data characteristics. If it has to be carefully considered, which patterns are displayed during the interactive mining, in order to not lose power for statistical tests, free exploration is discouraged and an (for domain experts) already difficult task is further complicated. Effects of applying the correction for the multiple comparison problems can also be very counter-intuitive. For example, the statistics of a pattern can show a statistically significant deviation at first. However, after more patterns are interactively explored, the same deviation gets insignificant, since a bigger correction factor has to be applied.

The problem gets even more serious, if one considers the iterative process of data mining tasks. As proposed for example by the CRISP-DM process model (Shearer, 2000), it is not uncommon to perform pre-processing steps (such as discretization of numeric attributes), then perform a core data mining algorithm and later adapt the pre-processing steps (e.g., using a different discretization method) with respect to the acquired results for a second iteration. To achieve a statistically sound procedure, patterns explored in the first iteration are also to be considered for the correction factor in subsequent iterations. This severely limits the power of the applied significance tests.

A simple solution again could be to consider the interactive exploration only as a hypothesis generation step and perform statistical testing later on separated holdout data. A convenient, integrated solution for statistically sound, interactive and iterative pattern mining has to the authors' solution not yet been proposed.

Proposition 4 *Significance does not reflect interestingness. Significance tests are (too) strongly dependent on the amount of available data.*

The effects of rejecting insignificant patterns strongly depend on the size of the used datasets. If the dataset is overall relatively small, then rarely appearing patterns will almost always get filtered out, even if they describe very strong correlations in the data. On the other hand, patterns with a large coverage might pass the test despite showing only weak correlations. If there is only limited data available, then data mining results get dominated by high coverage patterns, if automatic rejection of insignificant patterns is performed.

However, in many application scenarios strong influence factors are especially of interest, even if they are less generally applicable. As an example, consider a study in a medical domain: Here, finding a candidate for a small group of patients (e.g., those having certain comorbidities) that can almost always be succesfully treated with a certain drug provides far more practical advantages than validating that the treatment success differs marginally between two large groups, e.g., between genders. This preference is also reflected by a trend in applications sciences, which increasingly utilize additional methods that go beyond null

hypothesis significance testing, e.g., by measuring the *effect size* (Nakagawa and Cuthill, 2007; Kelley and Preacher, 2012). Furthermore, such weak influences described by high coverage patterns are probably more likely to be already known previously.

3. Decision factors for automatic filtering

In a specific target application a variety of factors should influence the decision on automatic significance tests for pattern filtering. These include:

- Does the target application per se require statistically valid results, such as in scientific applications or empirical medical research? In these cases, statistical filtering is of course inherently motivated.
- How many potentially interesting patterns are found (without significance testing)? If there are many patterns, additional significance filtering can help to focus on the most relevant ones. However, also larger correction factors have to be applied in this case.
- How much manpower is available for the manual inspection of the results? More manpower allows to inspect also insignificant results, discouraging automatic filtering. Since more hypotheses can be rejected manually than automatically, later significance testing of the remaining candidates on a separate test set will be more powerful.
- How knowledgeable and experienced is the target audience regarding the interpretation of statistical data? Care should be exercised when presenting non-significant results to inexperienced users.
- How closely are domain experts involved in the mining process? Statistical filtering is more difficult to apply if several iterations of data mining results are investigated by domain experts and mining parameters are adjusted according to their feedback. Consider saving a part of the data for hypothesis confirmation after the actual mining process in this case.
- What is the level of interactivity used in the mining process? Automatic filtering is currently optimized for purely automatic tasks.
- Can additional data be acquired for specific hypotheses, and at which cost? If additional data is available, then interesting, but not confirmed hypotheses are more relevant.
- Are non-statistical methods available that can confirm a hypothesis with reasonable effort? If hypotheses can be approved by other means, strict statistical testing might not be necessary.

Of course, these factors may point in opposite directions, so a trade-off is necessary most of the time. The actual decision on automatic rejection of not-significant patterns has to be made according to the actual problem at hand.

4. Conclusions

In this position paper, we argued against strict automatic rejection of data mining results, which do not pass a statistical significance test. Furthermore, we presented a list of factors that should influence the decision on automatic significance testing.

So, do these arguments imply that pattern mining should by no means employ statistical significance testing? Not at all. In fact, not considering the statistical significance of the result in particular with respect to the implications of the multiple comparisons problem might be one of the most critical mistakes one can make in the assessment of data mining results. However, automatically rejecting all result patterns, which do not pass a certain significance may come not too far behind. An experienced data analyst should always carefully consider the task at hand and the application context, e.g., the usage of the result patterns, the ability to acquire additional data for specific patterns and the capacities for the manual inspection of results.

References

Ronald P Carver. The case against statistical significance testing. *Harvard Educational Review*, 48(3):378–399, 1978.

Jacob Cohen. The earth is round ($p < .05$). *American psychologist*, 49(12):997–1003, 1994.

Wouter Duivesteijn and Arno J. Knobbe. Exploiting False Discoveries – Statistical Validation of Patterns and Quality Measures in Subgroup Discovery. In *Proceedings of the 11th International Conference on Data Mining (ICDM)*, pages 151–160, 2011. ISBN 978-1-4577-2075-8. doi: 10.1109/ICDM.2011.65.

Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI magazine*, 17(3):37–54, 1996.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Alipio M. Jorge, Paulo J. Azevedo, and Fernando Pereira. Distribution Rules with Numeric Attributes of Interest. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 247–258, 2006.

Ken Kelley and Kristopher J Preacher. On effect size. *Psychological methods*, 17(2):137–152, 2012.

Willi Klösgen. Data Mining Tasks and Methods: Subgroup Discovery: Deviation Analysis. In Willi Klösgen and Jan M. Zytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, pages 354–361. Oxford University Press, Inc., 2002.

Stéphane Lallich, Olivier Teytaud, and Elie Prudhomme. Statistical inference and data mining: false discoveries control. In *Compstat 2006 – Proceedings in Computational Statistics*, pages 325–336. Springer, 2006.

Heikki Mannila. Randomization techniques for data mining methods. In Paolo Atzeni, Albertas Caplinskas, and Hannu Jaakkola, editors, *Advances in Databases and Information Systems*, volume 5207 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85712-9. doi: 10.1007/978-3-540-85713-6_1.

Shinichi Morishita and Jun Sese. Traversing Itemset Lattices with Statistical Metric Pruning. In *Proceedings of the 19th ACM Symposium on Principles of Database Systems (PODS)*, pages 226–236, 2000.

Shinichi Nakagawa and Innes C Cuthill. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605, 2007.

Colin Shearer. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5:13–22, 2000.

John Wilder Tukey. *Exploratory Data Analysis*. Addison Wesley, 1977.

Geoffrey I. Webb. Discovering Significant Rules. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 434–443, 2006.

Geoffrey I. Webb. Discovering Significant Patterns. *Machine Learning*, 68(1):1–33, 2007.

Geoffrey I. Webb. Filtered-top-k Association Discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):183–192, 2011.

Stefan Wrobel. Inductive Logic Programming for Knowledge Discovery in Databases. In Sašo Džeroski and Nada Lavrač, editors, *Relational Data Mining*, pages 74–101. Springer, 2001.

Stephen Thomas Ziliak and Deirdre N McCloskey. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2008.