Assignment 2: Patterns

The **deadline** is May 31st at 14:00 Saarbrücken standard-time. You are free to hand in earlier. You will have to choose **one** topic from the list below, read the articles, and hand in a **report** that critically discusses this material and answers the assignment questions. Reports should summarise the key aspects, but more importantly, should include original and critical thought that show you have acquired a meta level understanding of the topic – plain summaries will *not* suffice. All sources you use should be appropriately referenced, any text you quote should be clearly identified as such. The expected length of a report is between 3 to 5 pages, but there is no limit.

For the topic of your assignment, choose one of the following:

1. *Elementary, my dear Watson*

   In data mining the goal is to find interesting structure of your data, things that somehow stand out. There are many ways to define 'standing out'. Significance testing based on background knowledge is one of them, but can, again, be done in different ways. There are two main schools. Gionis et al. [1] propose to measure significance using (swap) randomization, whereas De Bie argues to use Maximum Entropy (MaxEnt) modelling [2]. (Variants exist for all sorts of data types.) What are the key differences between the two approaches? What are the differences in background knowledge that can be incorporated? What are the differences in how the background knowledge is treated? And, what are hence the effects on using either method to infer whether a pattern is surprising or not? Which do you think is better, more practical, etc? Why?

   Within the MaxEnt school, there exist two sub-schools. In addition to [2], read [3]. What are the key differences between the models? What are the differences in type of knowledge they can incorporate? Can we use both models to test significance of the same types of structures/patterns? Are the two approaches unitable? Does it make any sense to have the type of background information of [2] incorporated into [3], and how about vice versa? If it does, sketch how you think this would work.

2. *Squeezing it*

   Mining *sets* of patterns that together describe the data well has effectively solved the pattern explosion. The question remains, how to score, and how to mine such sets? This are difficult questions. The first determines what the ideal solution looks like, whereas the second determines what we can possibly find. Both involve choices that have far-reaching consequences that are not always easy to oversee.

   To summarise event sequences, Tatti and Vreeken [4], for example, proposed the SQS algorithm. They use MDL, the Minimum Description Length principle, to define their score, and propose algorithms to both score the data, as well as to discover good pattern sets directly from data. A few years later, Fowkes and Sutton [5] take a related but slightly different probabilistic approach that does not directly punish gaps, and does allow patterns to interleave. Recently, Bhattacharyya and Vreeken [6] presented SQUISH, which can discover interleaving and nested patterns, as well as considers a richer class of patterns than both SQS and ISM.

   Your assignment, if you choose to accept is, is to read and analyse these papers critically, and connect the dots. Basic questions that can help you on your way, include the following. Are there any hidden, or obvious, biases and assumptions in the scores, in the cover, or search algorithms that may influence the results? If so, how? What are the advantages of the probabilistic over the MDL based score? How different are they really? What are the implications of not punishing gaps, in theory and in practice? What about the comparison between the different methods, are these convincing, fair, or are they comparing apples and oranges? Does ISM fare well on discovering the types of patterns they are after? How about interleaving? Why are the results as they are? (And, are the experiments presented fairly?)

   (Bonus) Squish much faster discovers a model that is at least as good as SQS, yet convergence takes a while. How could we speed this up, in a principled way? Also, the SQS-Search procedure requires many passes over the data, how could we reduce this and still (likely) obtain good models? Further, is it possible to include some of the ideas

of ISM back into SQS or Squish? How?

3. *So Significant, So Interesting*

   Once upon a time, frequency was thought to be a good measure for the interestingness of a pattern – you want to know what has been sold often, after all. After realising that the resulting set of patterns was more often than not enormous in size, as well as extremely redundant, research attention shifted to find better measure of interestingness. A natural approach then seemed to only report those patterns for which the frequency is *significant* with regard to some background model. Unsurprisingly, this turned out to be much harder than expected.

   Read both Brin et al. [7] and Webb [8]. Both give examples of how to identify patterns that are somehow deviating from an expectation, both consider a lot more information than simply the expected frequency under the marginals, yet both approaches are otherwise quite different. Analyse what the core ideas are of both approaches, give a succinct summary, and give a detailed discussion on how they differ, what in your view the strong and weak points of both approaches are. Is either of this measure the ultimate measure of interestingness for itemsets, or are there further improvements possible? Discuss a few (potential) downsides of the self-sufficiency approach, and give general (or specific) ideas on how we could address these.

   (Yes, Brin is Sergey Brin from Google.) (Although in [8] Webb does not give an algorithm for mining self sufficient itemsets, a few years back we showed how we can mine these efficiently [9].)

4. *Simple Subgroups* (Hard)

   Loosely speaking, in subgroup discovery we are after discovering subpopulations of our data that are a) selectable with a simply interpretable query, i.e. a pattern, and b) that exhibit a different distribution over the target variable than we see for the global distribution. Of course, there are many ways to define what makes a good subgroup—and over time many such definitions have been proposed. Commonly used scores include weighted relative accuracy for discrete targets, and mean-shift for continuous-valued targets.

   Only surprisingly recently, however, we realized that 'standing out' by itself is not necessarily a virtue. If we want to use subgroups to better predict the value of the target attribute, or want to use them to explain the target attribute using simple terms, only standing out does not suffice.

   Read the following two papers, one by Song et al. [10], and one by Boley et al. [11]. Focus your discussion on what is common between these two (technically very different) approaches, and how they (try to) solve what problem. Discuss critically on whether they achieve this goal, be it in general, or in the specific use case they consider. Last, go into detail what is different between the two approaches, beyond the obvious.

   (Bonus) Read the recent paper by Kalofolias et al. [12], who propose a variant of subgroup discovery in which we additionally consider a *control* variable. Ignoring the technical solution they propose, but rather focusing on the general problem where this control variable could be either continuous-valued or discrete valued, could we use the ideas of Song et al. [10] and Boley et al. [11] to build a stronger method? Or, do simple notions of (not) standing out suffice here?

Return the assignment by email to tada-staff@mpi-inf.mpg.de by **31 May, 1400 hours**. The subject of the email must start with [TADA]. The assignment must be returned as a PDF and it must contain your name, matriculation number, and e-mail address together with the exact topic of the assignment.

Grading will take into account both Hardness of questions, as well as whether you answer the Bonus questions.

References

You will need a username and password to access the papers outside the MPI network. Contact the lecturer if you don't know the username or password.

[1] Gionis, A., Mannila, H., Mielikäinen, T. & Tsaparas, P. Assessing Data Mining Results Via Swap Randomization (../papers/a2/gionis06swaprnd.pdf). In Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, pages 167-176, ACM, 2006.

[2] De Bie, T. Maximum entropy models and subjective interestingness: an application to tiles in binary databases (../papers/a2/tijl11maxent.pdf). Data Mining and Knowledge Discovery, 23(3):407-446, Springer, 2011.

[3] Mampaey, M., Tatti, N. & Vreeken, J. Tell Me What I Need To Know: Succinctly Summarizing Data with Itemsets (../papers/a2/mampaey11mtv.pdf). In Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, pages 573-581, ACM, 2011.

[4] Tatti, N. & Vreeken, J. The Long and the Short of It: Summarizing Event Sequences with Serial Episodes (../papers/a2/tatti12sqs.pdf). In Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China, ACM, 2012.

[5] Fowkes, J. & Sutton, C. A Subsequence Interleaving Model for Sequential Pattern Mining (../papers/a2/fowkes16ism.pdf). In Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA, 2016.

[6] Bhattacharyya, A. & Vreeken, J. Squish: Efficiently Summarising Event Sequences with Rich Interleaving Patterns (../papers/a2/bhattacharyya17squish.pdf). In Proceedings of the SIAM International Conference on Data Mining (SDM), Houston, TX, SIAM, 2017.

[7] Brin, S., Motwani, R. & Silverstein, C. Beyond Market Baskets: Generalizing Association Rules to Correlations (../papers/a2/brin97beyond.pdf). In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Tucson, AZ, pages 265-276, ACM, 1997.

[8] Webb, G.I. Self-sufficient itemsets: An approach to screening potentially interesting associations between items (../papers/a2/webb10self.pdf). ACM Transactions on Knowledge Discovery from Data, 4(1):1-20, 2010.

[9] Webb, G. & Vreeken, J. Efficient Discovery of the Most Interesting Associations (../papers/a2/webb14selfsufs.pdf). ACM Transactions on Knowledge Discovery from Data, 8(3):1-31, ACM, 2014.

[10] Song, H., Kull, M., Flach, P. & Kalogridis, G. Subgroup Disocvery with proper scoring rules (../papers/a2/song16proper.pdf). , 2016.

[11] Boley, M., Goldsmith, B.R., Ghiringhelli, L. & Vreeken, J. Identifying Consistent Statements about Numerical Data with Dispersion-Corrected Subgroup Discovery (../papers/a2/boley17dispersion.pdf). Data Mining and Knowledge Discovery, 31(5):1391-1418, Springer, 2017.

[12] Kalofolias, J., Boley, M. & Vreeken, J. Efficiently Discovering Locally Exceptional yet Globally Representative Subgroups (../papers/a2/kalofolias17rawr.pdf). In Proceedings of the 17th IEEE International Conference on Data Mining (ICDM), New Orleans, LA, IEEE, 2017.

(http://mmci.uni-saarland.de)