A theoretical framework for exploratory data mining: recent insights and challenges ahead

Tijl De Bie and Eirini Spyropoulou

Intelligent Systems Lab, University of Bristol, UK
tijl.debie@gmail.com, enxes@bristol.ac.uk

Abstract. Exploratory Data Mining (EDM), the contemporary heir of Exploratory Data Analysis (EDA) pioneered by Tukey in the seventies, is the task of facilitating the extraction of interesting nuggets of information from possibly large and complexly structured data. Major conceptual challenges in EDM research are the understanding of how one can formalise a nugget of information (given the diversity of types of data of interest), and how one can formalise how interesting such a nugget of information is to a particular user (given the diversity of types of users and intended purposes). In this Nectar paper we briefly survey a number of recent contributions made by us and collaborators towards a theoretically motivated and practically usable resolution of these challenges.

1 Exploratory data mining

From the seventies of the previous century, Tukey, Friedman, and collaborators advocated complementing research into statistical tools for *confirmatory* analysis of hypotheses with the development of tools that allow the interactive and *exploratory* analysis of data [24]. The sort of techniques they proposed for this ranged from the very simple (the use of summary statistics for data, and simple visual data summarisation techniques including the box plot as well as now largely obsolete techniques such as the stem-and-leaf plot), to advanced techniques for dimensionality reduction such as projection pursuit and its variants [6, 10]. While recognising the development of confirmatory analysis techniques (such as hypothesis tests and confidence intervals, allowing us to infer population properties from a sample) as one of the greatest achievements of the twentieth century, Tukey complained that “Anything to which a confirmatory procedure was not explicitly attached was decried as ‘mere descriptive statistics’, no matter how much we learned from it.”

Since then, data has evolved in size and complexity, and the techniques developed in the past century for EDA are only rarely applicable in their basic unaltered form. Nevertheless, we argue that the problem identified by Tukey is greater than ever. Today’s data size and complexity more often than not demand an extensive exploration stage by means of capable and intuitive EDM techniques, before predictive modelling or confirmatory analysis can realistically and usefully be applied.

There are however a few important research challenges that need resolving before EDM techniques can optimally fulfil this need:

- The concept of a 'nugget of information' found in data needs to be formalised. We will refer to such a nugget of information as a *pattern*.
- To allow for automating the search for interesting patterns in data, the concept of *interestingness* of a pattern needs to be formalised mathematically. Clearly, interestingness is a subjective concept, such that the formalisation must depend on the user's perspective.
- These theoretical insights need to be turned into practical methods and eventually a toolbox for EDM 'in the wild'.

Given the nature of this Nectar paper track, in most of the remainder of this short note we will focus on our own contributions towards the resolution of these challenges. Here we only briefly mention a very incomplete list of works that have influenced our thinking or that have otherwise impacted on EDM research: In 2000 Mannila wrote a highly insightful letter in SIGKDD Explorations about frameworks for data mining [19]; Several prominent researchers advocate data compression as the key operation in the data mining process [5, 20]; Recent influential work from Mannila and others on swap randomizations has advocated the use of empirical hypothesis testing in the development of interestingness measures [8, 9, 18]; The work on tiling databases [7] has been inspirational to our earliest work on this topic. For a more comprehensive overview of data mining interestingness measures based on novelty we refer the reader to our survey paper [14]. Finally, much of our work was also inspired by applied bioinformatics research where exploratory analysis was required, and where we found that current techniques fell short [17, 16].

2 Patterns and their interestingness

Let us start by clarifying the key terminology. Let Ω be the (measurable) space to which the *data*, denoted as x, is known to belong. We will refer to Ω as the *data domain*. Then, in our work we defined the notion of a *pattern* by means of a subset Ω' of the data domain, saying that a pattern defined by $\Omega' \subseteq \Omega$ is *present* in the data x iff $x \in \Omega'$. This definition is as expressive as it is simple. Most, if not all, types of data mining patterns can be expressed in this way, including the results of frequent pattern miners, dimensionality reduction methods, clustering algorithms, methods for community detection in networks, and more.

The simplicity of this definition further allows us to reason about the interestingness of a pattern in terms of how it affects a user's beliefs about the data. To achieve this, we have opted to represent the beliefs of a user by means of a probability measure P defined over the data domain Ω, to which we refer as the *background distribution*. The interestingness of a pattern is then related to how the background distribution is affected by revealing a pattern to a user, i.e. the degree to which revealing a pattern enhances the user's belief attached to the actual value of the data under investigation.

To do this, several issues need to be studied, such as how to come up with a sensible background distribution without putting too large a burden on the user, how the revealing of a pattern affects the background distribution, how a change in background distribution should be translated into interestingness, and the cost (e.g. in terms of mental energy or processing capacity) presented to a user when processing the revealed pattern.

In answer to these questions, in [2, 1, 4] we presented formal arguments demonstrating that a robust approach to quantifying interestingness is based on three elements: (1) inferring the background distribution as the one of maximum entropy subject to constraints that formalise the user's prior beliefs about the data; (2) the quantification of the *information content* of the pattern, as minus the logarithm of the probability $P(x \in \Omega')$ under this background distribution that the data belongs to the restricted domain Ω' defined by the pattern; and (3) trading off this information content with the *length of the description* required to communicate the pattern to the user.

Most commonly the purpose of the data miner is to obtain as good an understanding of the data (overall information content of the set of patterns revealed) within specific bounded resource constraints (overall description length of all the patterns revealed). Initially in [2] and later more formally in [1], we argued that this amounts to solving a weighted budgeted set coverage problem. While this problem does not allow for an efficient optimal solution, it can be approximated provably well in a greedy way, iteratively selecting the next best pattern. Hereby, the next best pattern is defined as the one that maximizes the ratio of its information content (given the current background distribution) to its description length. Thus, matching this common usage setting, we proposed to formalize the interestingness of a pattern as the ratio of its information content and its description length, called its *information ratio* (or compression ratio). It represents how densely information is compressed in the description of the pattern.

3 Data and users in the real world

Initially we demonstrated our theoretical results on the particular data mining problem of frequent itemset mining [2] for a relatively simple type of prior beliefs (namely the row and column sums), and for a simple type of pattern (namely a tile [7]). In our later work we extended it in the following directions:

- Using more complex types of pattern (in casu noisy tiles) [11] as well as allowing more complex types of prior beliefs to be taken into account on simple binary databases, such as tile densities and itemset frequencies [12].
- Expanding these ideas toward real-valued data, for local pattern types [15] as well as global clustering pattern types [3, 13].
- The development of a new expressive pattern syntax for multi-relational data with binary and n-ary relationships, the formalisation of subjective interestingness for a certain type of prior information, and the development of efficient algorithms to mine these patterns [21–23].

4 An encompassing toolbox for exploratory data mining?

We believe there is significant value to be gained by further expanding these theoretical insights as well as the practical instantiations of the framework. We hope and anticipate that this may ultimately result in a modular and expandable toolbox for EDM that can be applied to data as it presents itself in real-life, and that is effectively usable by experts and lay users alike.

Most real-world structured data is multi-relational data in some way, including simple binary and attribute-value tables, traditional (relational) databases, (annotated) graphs, as well as RDF data and the semantic web. We therefore believe that a general EDM toolbox could most easily be built upon our recent work on multi-relational data mining. In this work we developed a new pattern syntax for multi-relational data with categorical attribute values, an associated interestingness measure along the lines of the advocated framework (demonstrated for a simple but important type of prior beliefs), as well as efficient mining algorithms [21–23].

Of course, in order to mature into a fully fledged EDM toolbox, this starting point requires a number of advances. Some of these, however, we have already partially developed for simpler types of data. For example, the resulting toolbox will need to be able to deal with real-valued data, which requires the definition of a new multi-relational pattern syntax and the adaptation of the prior belief types for real-valued data developed in [15, 3, 13] to the multi-relational case. Another required step will be the incorporation of more complex types of prior information also for categorical data, along the lines of our previous work on single-relational data [12].

Acknowledgements Most importantly we are grateful to Kleanthis-Nikolaos Kontonasios, as well as to Jilles Vreeken, Mario Boley, and Matthijs van Leeuwen, all of whom have made important contributions to the development of the vision advocated in this Nectar paper. This work was partially funded by EPSRC grant EP/G056447/1.

References

1. T. De Bie. An information-theoretic framework for data mining. In *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD11)*, 2011.
2. T. De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, 23(3):407–446, 2011.
3. T. De Bie. Subjectively interesting alternative clusters. In *Proceedings of the 2nd MultiClust workshop: Discovering, Summarizing and Using Multiple Clusterings*, 2011.
4. T. De Bie, K.-N. Kontonasios, and E. Spyropoulou. A framework for mining interesting pattern sets. *SIGKDD Explorations*, 12(2), December 2010.

5. C. Faloutsos and V. Megalooikonomou. On data mining, compression, and kolmogorov complexity. *Data Mining and Knowledge Discovery*, 15:3–20, 2007.
6. J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis. *Computers, IEEE Transactions on*, 100(9):881–890, 1974.
7. F. Geerts, B. Goethals, and T. Mielikäinen. Tiling databases. In *Discovery Science*, volume 3245, pages 278–289, 2004.
8. A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data*, 1(3):14, 2007.
9. S. Hanhijarvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila. Tell me something I don't know: Randomization strategies for iterative data mining. In *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD09)*, pages 379–388, 2009.
10. P. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.
11. K.-N. Kontonasios and T. De Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, 2010.
12. K.-N. Kontonasios and T. De Bie. Formalizing complex prior information to quantify subjective interestingness of frequent pattern sets. In *Proc. of the 11th International Symposium on Intelligent Data Analysis (IDA)*, 2012.
13. K.-N. Kontonasios and T. De Bie. Subjectively interesting alternative clusterings. *Machine Learning*, 2013.
14. K.-N. Kontonasios, E. Spyropoulou, and T. De Bie. Knowledge discovery interestingness measures based on unexpectedness. *WIREs Data Mining and Knowledge Discovery*, 2(5):386–399, 2012.
15. K.-N. Kontonasios, J. Vreeken, and T. De Bie. Maximum entropy modelling for assessing results on real-valued data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2011.
16. K. Lemmens, T. De Bie, T. Dhollander, S. D. Keersmaecker, I. Thijs, G. Schoofs, A. De Weerdt, B. De Moor, J. Vanderleyden, J. Collado-Vides, K. Engelen, and K. Marchal. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in escherichia coli. *Genome Biology*, 10(R27), 2009.
17. K. Lemmens, T. Dhollander, T. De Bie, P. Monsieurs, K. Engelen, J. Winderickx, B. De Moor, and K. Marchal. Inferring transcriptional module networks from ChIP-chip-, motif- and microarray data. *Genome Biology*, 7(R37), 2006.
18. J. Lijffijt, P. Papapetrou, and K. Puolamki. A statistical significance testing approach to mining the most informative set of patterns. *Data Mining and Knowledge Discovery*, December 2012.
19. H. Mannila. Theoretical frameworks for data mining. *SIGKDD Explorations*, 2000.
20. A. Siebes, J. Vreeken, and M. van Leeuwen. Item sets that compress. In *SIAM Conference on Data Mining*, 2006.
21. E. Spyropoulou and T. De Bie. Interesting multi-relational patterns. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2011.
22. E. Spyropoulou, T. De Bie, and M. Boley. Mining interesting patterns in multi-relational data. *Data Min. Knowl. Discov.*, 2013.
23. E. Spyropoulou, T. De Bie, and M. Boley. Mining interesting patterns in multi-relational data with n-ary relationships. In *Proceedings of the International Conference on Discovery Science (DS)*, 2013.
24. J. Tukey. Exploratory data analysis. *Reading, MA*, 231, 1977.