Dr. Enrico Glaab
Biomedical Data Science Group
Luxembourg Centre for Systems Biomedicine
Université du Luxembourg
7, Avenue des Hautes Fourneaux
L-4362 Esch-sur-Alzette
Luxembourg
Tel: +352-466644-6186 (Office)
Email: enrico.glaab@uni.lu

Luxembourg, 2020-11-24

Review of the Doctoral Thesis

High-performance exploration and querying of selected multi-dimensional spaces by Miroslav Kratochvíl

Overview

The thesis presents new computational methods for querying, visualizing and exploring multi-dimensional datasets, with a focus on facilitating and speeding up biological applications related to the querying of chemical compound data, and visualizing and interactively exploring flow and mass cytometry data.

Chapter 1 discusses the chemical search in large small-molecule compound databases, providing on overview of the problem background, introducing index structures, the Sachem database cartridge and federated search applications. It presents current challenges in heterogeneous chemical search and possible solution strategies.

Chapter 2 discusses new efficient techniques for the visualization of high-dimensional data, with a focus on biological applications in interactive processing and analysis of flow and mass cytometry data. It explains the developed EmbedSOM approach as well as variant approaches, and how they can be used for different practical purposes, such as computationally assisted cytometry data analysis. Additionally, it covers the indexation of large cytometry datasets and the relevance of dedicated indexation approaches for multimedia browsing and retrieval applications.

Main strengths

- The doctoral project has led to significant new scientific contributions, in particular in the fields of chemical compound search and the exploration and visualization of high-dimensional cytometry data, with further applications of the developed algorithms extending to many other fields, e.g. multimedia retrieval.

- The thesis is well written and clearly structured. Formal definitions are provided for the key concepts, and sufficiently detailed pseudo-code is presented for the main algorithms. Figures and illustrations are adequate and informative, and clear information is provided on which contributions the doctoral candidate has made to the publications.

- The publication record is very strong, with 11 peer-reviewed articles, 5 of which representing the main contributions of the thesis, and 2 further manuscripts not yet published (page xi). The publications also include first-author manuscripts in very well-regarded journals in the field, such as the Journal of Cheminformatics.

Main weaknesses

- The evaluation of the developed algorithms focuses strongly on runtime performance comparisons and some example applications; however, in particular for the low-dimensional representation approach EmbedSOM it would have been useful to see more comprehensive quantitative benchmark comparisons, applying other related approaches (FlowSOM, t-SNE, UMAP, IsoMap, LLE, etc.) and EmbedSOM to multiple simulated and real-world datasets that represent different common data analysis scenarios, and examining the benefits and shortcomings of different methods systematically using multiple performance metrics (e.g. preservation of data point distances according to different distance measures, utility of the transformed data for clustering applications, etc.), and comparing the achieved performances using a statistical test.

- The thesis does not contain a dedicated and detailed outlook session that would explain the reader what the key future challenges in the field are, what the more detailed next follow-up steps could be, and how specifically the developed tools will be further maintained, or could potentially be complemented by new developments. Some possible further extensions are mentioned briefly on the Conclusions page and in other parts of the text, but a more detailed and comprehensive discussion of possible follow-up research is not presented (while the doctoral candidate may plan to move to a different field after the PhD, a detailed theoretical discussion of how the research could be followed up would still be a relevant component of the thesis).

- Since the applications of the developed algorithms mainly focus on biological data analysis, it would have been useful to show more specific examples of how using the algorithms on real-world datasets can lead to new biological findings. In particular, for the EmbedSOM approach, it would be interesting to see whether applying the method to further cytometry datasets or other types of omics data might enable the discovery of new biologically relevant clusters (e.g. via subsequent application of density-based clustering approaches) which would not be detected as easily by other dimension reduction approaches (or any other examples of how applying EmbedSOM can lead to new specific biological insights, or help to address a specific biological question, would be helpful for the reader as a practical illustration of the benefits the method can provide).

Comments

Page 6: In definition 3 of “structure similarity”, the author requires that the corresponding function is symmetric and satisfies the identity of indiscernibles – here, it would be relevant to also provide background information on why certain typical properties of generic metric functions, such as the triangle inequality, are not always fulfilled by structure similarity functions.

Page 7: Fingerprint-based screening approaches have arguably lost much of their influence in pharmaceutical applications of virtual screening due to the publication of tree-based representations and associated fast screening approaches, such as the “feature tree” methodology (see Lessel et al., J Chem Inf Model, 2009; Ehrlich and Rarey, Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011), which captures both topological and chemical properties when quantifying compound similarity. While the problem of finding a maximal common subgraph (MCS) isomorphism is indeed NP-hard, MCS problems become less complex for certain graph types, and in particular the MCS between two trees can be computed in polynomial time using dynamic programming. It would therefore be relevant and informative to discuss if and to which extent approaches presented in the thesis could be extended or complemented by related approaches that can also operate on the newer, more commonly used tree-based compound representations.

Pages 13 and 14: When the text refers to “a practical query that asks for derivable compounds with likely biological activity” or suggests at the bottom of page 14 that “the availability of protein interaction data in ChEMBL and UniProt may provide sufficient material for quantifying relations between structure and biological activity of the small molecules”, it would be important to point out that structural (and also chemical) compound similarity is not a reliable indicator of shared biological activity. Even a small change in just a single hydrogen atom can turn an active drug compound into an inactive compound, or a safe substance into a toxic substance. These limitations should be pointed out early in the thesis.

Page 20 onwards & page 57 onwards: For a new low-dimensional data representation approach, such as EmbedSOM, one would typically expect to see a quantitative benchmark comparison to other methodologies in terms of different performance metrics, e.g. assessing how well local and global features are preserved in terms of different distance or dissimilarity measures, or comparing clustering results obtained on the transformed data using multiple cluster validity indices. Even if embeddings on subsets of the data can be recomputed quickly, the capacity of a dimension reduction approach to find adequate structure-preserving data representations without additional recomputation cycles or further interactive involvement of the user is practically relevant, and results generated in a fully automated fashion could be compared on multiple benchmark datasets for the newly proposed method and existing algorithms.
Possibly, as in other areas of data science where the “no free lunch theorem” applies, no algorithm will be superior on all problem instances – but to decide for which types of data the proposed methodology or other approaches are more suitable would require a more comprehensive evaluation on different types of input data, e.g. using simulated datasets with known properties or multiple representative real-world datasets for different practical settings.

Pages 24-25: The text suggests that low-dimensional visualization approaches such as t-SNE, UMAP, and IsoMap require a kNN-graph construction, which has a runtime complexity of O(n log n) in the best case. However, in practice, approximate nearest neighbor search has been described to be sufficient for most of these algorithms, and efficient t-SNE implementations have been applied successfully to datasets with millions of samples without a requirement for specialized hardware (see publications by van der Maaten et al. listed on https://lvdmaaten.github.io/tsne). Thus, considering also that there are many different implementations available for t-SNE, UMAP etc. with different runtime requirements than those for the implementations used in the manuscript on page 66 in Figure 2, for the comparative evaluation of methods, rather than focusing mainly on runtime performance, it would be more informative to assess in further detail the utility of the generated outcome visualizations for practical clustering and data interpretation applications, e.g. showing an example where EmbedSOM provides some biological insights that could not be obtained as easily from other methods.

Page 24 & page 60: The thesis mentions that instead of using a SOM, even a random selection of landmarks can produce good results. Given that the section starting at page 65 suggests that alternative land mark generation methods may even improve visualization, it is not clear whether there is still a sufficient benefit of using SOM at all for landmark generation, and more practical examples should be given to illustrate under which circumstances which method for landmark generation is recommended, and whether using SOM shows a benefit in a particular example. In Figure 3 on page 67, the results derived from the t-SNE landmarks appear to be most human-

interpretable and the densely clustered regions seem to match well with the tissue of origin, as compared to both SOM and GQTSOM landmarks, but a single dataset may only provide rough indications. Here, it would be useful to see more examples for other datasets (e.g. using different types of simulated datasets with diverse distributions and noise levels), in order determine the strengths and weaknesses of different approaches and to be able to make conclusions for which settings which methodology is most adequate.

Page 62: The manuscript on the updated EmbedSOM version suggests to use certain default parameters, indicating that these parameters worked well in a majority of test cases. However, it is not fully clear how these parameter choices were evaluated and how significant their influence on the results is. To provide more information on how the parameters influence the outcomes, it would be useful to show detailed comparisons for a few scenarios. Similarly, also for different distance metrics, a comparison of a few different selections would be useful, in order to observe their specific influence on the results.

General comment: Vague terms such as “tremendous” (page 3 and page 14), “relatively huge” (page ix) and “relatively large” (page 26) should be omitted or replaced by a more specific description.

Formatting

- Page IX: Replace “gives extended overview” by “gives an extended overview”
- Page 4: Replace “appliation” by “application”
- Page 8: Replace “despite the upper compound forms a proper substructure” by “unless the upper compound forms a proper substructure”
- Page 13: Replace “In result, “ by “As a result, “
- Page 13: Replace “relational databases as such” by “relational databases”
- Page 14: Replace “different identifier for the same molecule” by “different identifiers for the same molecule”
- Page 15: Much of the text in Figure 2 is too small to read, increasing the size and adding the legend below the figure would help to increase readability.
- Page 16: Replace “Despite the performance” by “Although the performance”
- Page 17: Replace “causing to substructure search to fail” by “causing the substructure search to fail”
- Page 19: Replace “utilization of the automated techniques that could aid fast and precise dissection the measured samples” by “utilization of automated techniques that could aid fast and precise dissection of the measured samples”
- Page 20: Replace “the data can be fitted to original” by “the data can be fitted to the original”
- Page 21: Replace “features describe by the map” by “features described by the map”
- Page 25: Replace “internal structure these clusters” by “internal structure of these clusters”; and replace “summarized Fig. 10” by “summarized in Fig. 10”

- Page 33: Replace “may work sufficiently for” by “may work sufficiently well for”
- Page 37: Replace “that enabled development” by “that enabled the development”
- Page 49: It is not fully clear what is meant with “an interoperable wrap” (is the software only available as interoperable web-service or also as interoperable software library, and in which format?), this should be specified more clearly
- Page 63: Replace “find approximate embedding” by “find an approximate embedding”
- Page 64: Replace “occupy positionson” by “occupy positions on”

Summary

The doctoral candidate has written a well-structured and well-researched thesis, which is complemented by a strong publication record. New contributions made for the efficient and interactive exploration of high-dimensional data are timely and have significant practical relevance for the fields of cheminformatics, cytometry data analysis, multimedia retrieval, as well as other areas of data science.

It would have further strengthened the thesis, if apart from the useful runtime performance comparisons and example applications, more detailed quantitative comparisons of the proposed dimension reduction approaches and other alternative methodologies applied to multiple simulated and real-world benchmark datasets had been presented. As a strong point, the thesis provides detailed theoretical discussions of the benefits and limitations of the developed algorithms, as well as detailed definitions, pseudo-code algorithm descriptions and adequate illustrations.

Overall, the published articles which represent the author’s main contribution and the thesis itself prove the candidate’s ability for both systematic and creative scientific work. Therefore, I recommend to accept this thesis for the graduation and to award the PhD degree to Miroslav Kratochvíl.

Luxembourg, 24th November, 2020

Enrico Glaab