universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Framework for Visual Cluster Analysis and Consensus Clustering“

verfasst von / submitted by

Christian Permann, BSc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree

MSc.

Wien, 2020 / Vienna, 2020 Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet	A 066 921
Studienrichtung lt. Studienblatt:/ degree programme as it appears on the student record sheet	Masterstudium Scientific Computing
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Inform.Univ. Dr. Claudia Plant

Danksagungen

Danke!

Contents

List of Figures

List of Tables

Part I

Introduction and Preliminaries

Chapter 1

Introduction

The generation of clusterings is quite a difficult task when little knowledge on the data is available. There exist a large number of clustering algorithms that all aim to find a fitting partitioning of the data, but they all come with different assumptions on the data. Taking k-Means as an example, the data is assumed to be Gaussian distributed which might not actually be the case for many data-sets, making it unsuitable for clustering such data.

As a solution for this, visual frameworks have been proposed that aim to help finding the right clustering by evaluating the results of different algorithms and parameter settings via quality metrics and ranking the clustering solutions. One of these frameworks is Clustervision [5] whereby different algorithms are run, their solutions are visualized and ranked according to the average of five different quality metrics. The user is then able to look at the candidates through different visualizations and choose a solution as final result. A problem with this approach is though, that each metric is biased towards a specific resulting cluster structure, possibly ranking it higher than others even if it does not fit the data properly. Even the authors of Clustervision themselves mention that "the effectiveness of these metrics in gauging the quality of the clustering is also difficult to determine due to the lack of ground truth". Also, as many solutions were calculated and only one of them is chosen in the final decision, a lot of calculated knowledge is ignored.

To overcome this problem, I propose to not use quality metrics for the choice of the best clustering, but to rely on robustness indicators for this selection. As robustness is difficult to evaluate by itself many different algorithms with different parameters should be used, analyzing where multiple methods or parameter sets find a similar result. Clusterings that are similar to many other results can be seen as robust and finding a consensus of those can result in a better overall clustering. To find such similar clusterings or groups of similar clusterings meta-clustering can be used. To then compute final candidates, each group of similar results can be merged to a consensus result, capturing the essence of the group.

Based on this idea I created a tool which will be further described in Chapter 5. Additionally, the tool aims to include the expert user into the process of finding the result. Most research on consensus clustering, as also seen in the survey of Vega-Pons and Ruiz-Shulcloper [7], only briefly mention how generating or pre-selecting base clusterings can impact the result of consensus clustering. For this, my tool allows the user to visually explore the clusterings created by the simple clustering algorithms, facilitating the choice of which ones to merge for computing a final result.

The rest of this thesis is organized as follows:

Part II

Theory

Chapter 2

Related Work

2.1 Clustering Algorithms

A lot of research as been put toward the topic of clustering, resulting in a large number of different methods. [9]

2.2 Visual Clustering Frameworks

Clustervision [5]

VISTA [2] ClusterMap and user result evaluation

iVIBRATE [3] extension VISTA?

simple visualizations ELKI implementation [6]

Frameworks with specific data-sets in mind:

Propose: "An Evolutionary and Visual Framework for Clustering of DNA Microarray Data" [1], tool for clustering DNA data with visualization

Even Memes [4]

2.3 Consensus Clustering

consensus cluster plus [8]

Chapter 3

Methods

3.1 Clustering

3.1.1 OPTICS

3.2 Consensus Clustering

3.2.1 DICLENS

3.3 Other Methods

3.3.1 Hungarians Method

Part III

Implementation

Chapter 4

Implementation Remarks

4.1 Programming Language and Tools

For this implementation my language of choice was Java. With Java there are lots of available implementations for different clustering methods available, including the Frameworks ELKI [CITATION] and WEKA [CITATION]. Additionally, it is easy to create custom visualizations using the low level draw() method of JComponents in Swing. To improve the performance in some parts of the tool Java 1.8 was chosen as lowest required version, as it allows to use of parallel streams to easily implement parallelism.

4.2 Interfaces and Extendability

4.3 Clustering Algorithms

4.4 Meta-Clustering

4.5 Consensus Clustering

Chapter 5

The Tool

5.1 I/O and Data Visualization

bla

5.2 Selection of Algorithms and Parameters

bla

5.3 Meta-View and Consensus Clustering

bla

Part IV

Testing

Chapter 6

Experiments

bla

Part V

Concluding Thoughts

Chapter 7

Future Work

7.1 Improvements for Tool

bla

7.2 Research Consensus Clustering

bla

7.3 Visual Frameworks

bla

Chapter 8

Conclusion

8.1 Lessons learned

bla

8.2 Reflection of Work

bla

Appendices

Appendix A

Abstract

A.1 English abstract

Finding a good clustering solution for an unexplored data-set is a non-trivial task. Due to the large number of clustering algorithms which usually have lots of parameters, clustering results may differ strongly from each other and the underlying ground truth. With only little knowledge on the data the evaluation of which result best represents the underlying cluster structure is difficult. To find a fitting selection for the result, there exist visual frameworks that aim to simplify this choice by ranking the results according to quality measures. As those measures also have the downside of being biased towards specific structures (weather or not they fit the data) they are problematic for selecting a final result. For this reason, I propose to purely use indicators of robustness for the creation of a clustering result. This is done by meta-clustering results from different clustering algorithms and results and calculating consensus clusterings from each group of simmilar results. Additionally this process is supported through visualizations, giving the expert user the possibility to use his knowledge to further improve on the final result.

A.2 Deutsche Zusammenfassung

Eine gute Clustering Lösung für wenig erforschte Daten zu finden ist eine komplexe Aufgabe. Wegen der großen Anzahl an Clustering Algorithmen, welche meist auch viele verschiedene Parameter benötigen, können sich die Ergebnisse stark untereinander, aber auch von dem richtigen Ergebnis, unterscheiden. Mit nur wenig Wissen über die Daten ist auch die Evaluierung welches Ergebnis am nähersten zu der der unterliegenden Wahrheit, beziehungsweise am besten der Struktur der Daten entspricht eine schwere Aufgabe. Um eine solche Auswahl besser treffen zu können wurden visuelle Frameworks erschaffen, die mittels Qualitäts-Metriken die verschiedenen Ergebnisse bewerten und gereiht anzeigen. Da diese Metriken aber auch das Problem haben gewisse Strukturen in Ergebnissen zu bevorzugen zeigen sie sich wiederum bei der Entscheidung über das endgültige Ergebnis als problematisch. Aus diesem Grund schlage ich vor die Eigenschaft wie robust ein Ergebnis ist für die finale Entscheidung heranzuziehen. Um dies zu tun werden die Clusterings auf Meta-Ebene nochmals geclustert, wobei ähnliche Ergebnisse in einer Gruppe mittels Consensus Clustering zu einer Lösung zusammengeführt werden. Dieser Prozess wird weiters durch Visualisierungen unterstützt, so dass ein Experte mit Hilfe seines Wissens die Lösung möglicher Weise noch weiter verbessern kann.

Bibliography

[1] Castellanos-Garzón, J., and Díaz, F. An evolutionary and visual framework for clustering of dna microarray data. *Journal of Integrative Bioinformatics 10* (12 2013).

[2] Chen, K., and Liu, L. A visual framework invites human into clustering process. pp. 97 – 106.

[3] Chen, K., and Liu, L. Ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Trans. Inf. Syst. 24*, 2 (Apr. 2006), 245–294.

[4] Dang, A., Moh'd, A., Gruzd, A., Milios, E., and Minghim, R. *An Offline–Online Visual Framework for Clustering Memes in Social Media*. Springer International Publishing, Cham, 2017, pp. 1–29.

[5] Kwon, B. C., Eysenbach, B., Verma, J., Ng, K., deFilippi, C., Stewart, W. F., and Perer, A. Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics 24* (2018), 142–151.

[6] Schubert, E., Koos, A., Emrich, T., Züfle, A., Schmid, K. A., and Zimek, A. A framework for clustering uncertain data. *Proc. VLDB Endow. 8*, 12 (Aug. 2015), 1976–1979.

[7] Vega-Pons, S., and Ruiz-Shulcloper, J. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence 25* (May 2011), 337–372.

[8] Wilkerson, M. D., and Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics 26*, 12 (04 2010), 1572–1573.

[9] Xu, D., and Tian, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science 2* (08 2015).