Final Assignment
CC62 Data Mining
Section: CC81
UPC

Willy Ugarte Rojas

1 Statement

Propose a solution to a given transactional dataset by using techniques in Pattern Mining developed in Unit 3 and Data Visualization in Unit 4.

2 Goal

Mastering the Pattern Mining and Data Visualization techniques. Starting by mining patterns from a dataset, and then visualizing them with data visualization techniques.

3 Deliverables and Important Dates

The slides and a folder containing your work (dataset, report, and slides) must be zipped with the name XXXX.zip (where XXXX is the student code) and upload it to the blackboard.

The deadline is Tuesday November 21 2018 until midnight. During the last session of week 15 each group of students will present their work during 20 minutes and respond to the questions made by a jury (composed by the teacher and the other students) during 10 minutes. The qualification will be made according to the table at the end of this document.

4 Association Rules/Pattern Mining

First, you are going to obtain results for a given dataset by using association rules and pattern mining techniques.

4.1 Algorithms

Each work must be focused in one of the following subjects:

1. **Association Rules** [Suzuki and Zytkow, 2005, Elsayed et al., 2012, Boudane et al., 2017]
2. **Conceptual Clustering** [Ouali et al., 2016, Chabert and Solnon, 2017, Aribi et al., 2018]
3. **Emerging Patterns** [Dong and Li, 1999, Novak et al., 2009, Komiyama et al., 2017]
4. **Sky Patterns** [Soulet et al., 2011, Ugarte et al., 2017]
5. **Dominance Programming/Optimal Patterns** [Négrevergne et al., 2013, Ugarte et al., 2015, Gebser et al., 2016]

Foreach, either you must implement or get an implemented version of any of the algorithms specialized for the task (e.g. Eclat for closed patterns).

4.2 Data

For the experimentations:

1. You must use at least 2 datasets:
   (a) The UCI datasets, which are classical for may mining tasks.
   (b) And any real-life dataset (e.g. SMPF datasets) that you gather from the web.
2. You must explain in detail:
   (a) your choice of data,
   (b) its provenance,
   (c) its domain field,
   (d) quality of extraction and extraction method

5 Data Visualization

You must present your results according to the visualization methods shown in the course.

- Charts
- Dashboards
- Plots

6 Report

- Your report must be written in latex under the IEEE format.
- Structure:
  1. Introduction
  2. Background (delivrable 1)
  3. Experimental Study
  4. Dataset description.
     (a) Experimental design:
        - Design a experimental protocol that verifies the functional and non-functional requirements of the pattern mining prototype.
        - Design unit tests on the components of the pattern mining prototype, based on assigned papers.
        - Design quality assurance measures for experiments.

     (b) Experimental Development:
        - Implementation of functional and non-functional tests of the pattern mining prototype according to the elaborated design.
        - Implementation of unit tests on the components of the pattern mining prototype, according to the design elaborated in the assigned papers.
        - Performs debugging on the components of the pattern mining prototype.
        - Measures the main quality attributes of the pattern mining prototype.

     (c) Analysis and interpretation of results.
        - Analyze and interpret the results generated during the functional tests of the pattern mining prototype.
        - Analyze and interpret the results of the non-functional (performance) tests performed on the pattern mining prototype.
        - Analyze and conclude based on quality measures of the pattern mining prototype.
        - Applies in an informed manner appropriate modern tools to develop the pattern mining prototype.
  5. Conclusions and perspectives

7 Schedule

This final assignment must be presented in two deliverables.

7.1 Deliverable 1

This part is to present in week 12 with an technical summary of the related papers of the asignated subject.

7.2 Deliverable 2

This part is to present in week 15 with the description of the dataset, management of association rules/pattern mining and data visualization.

References

[Aribi et al., 2018] Aribi, N., Ouali, A., Lebbah, Y., and Loudni, S. (2018). Equitable conceptual clustering using OWA operator. In Phung, D. Q., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., and Rashidi, L., editors, *Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III*, volume 10939 of *Lecture Notes in Computer Science*, pages 465–477. Springer.

[Boudane et al., 2017] Boudane, A., Jabbour, S., Sais, L., and Salhi, Y. (2017). Enumerating non-redundant association rules using satisfiability. In Kim, J., Shim, K., Cao, L., Lee, J., Lin, X., and Moon, Y., editors, *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I*, volume 10234 of *Lecture Notes in Computer Science*, pages 824–836.

[Chabert and Solnon, 2017] Chabert, M. and Solnon, C. (2017). Constraint programming for multi-criteria conceptual clustering. In Beck, J. C., editor, *Principles and Practice of Constraint Programming - 23rd International Conference, CP 2017, Melbourne, VIC, Australia, August 28 - September 1, 2017, Proceedings*, volume 10416 of *Lecture Notes in Computer Science*, pages 460–476. Springer.

[Dong and Li, 1999] Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *KDD*, pages 43–52. ACM.

[Elsayed et al., 2012] Elsayed, S. A. M., Rajasekaran, S., and Ammar, R. A. (2012). ML-DS: A novel deterministic sampling algorithm for association rules mining. In Perner, P., editor, *Advances in Data Mining. Applications and Theoretical Aspects - 12th Industrial Conference, ICDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings*, volume 7377 of *Lecture Notes in Computer Science*, pages 224–235. Springer.

[Gebser et al., 2016] Gebser, M., Guyet, T., Quiniou, R., Romero, J., and Schaub, T. (2016). Knowledge-based sequence mining with ASP. In Kambhampati, S., editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1497–1504. IJCAI/AAAI Press.

[Komiyama et al., 2017] Komiyama, J., Ishihata, M., Arimura, H., Nishibayashi, T., and Minato, S. (2017). Statistical emerging pattern mining with multiple testing correction. In *KDD*, pages 897–906. ACM.

[Négrevergne et al., 2013] Négrevergne, B., Dries, A., Guns, T., and Nijssen, S. (2013). Dominance programming for itemset mining. In Xiong, H., Karypis, G., Thuraisingham, B. M., Cook, D. J., and Wu, X., editors, *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 557–566. IEEE Computer Society.

[Novak et al., 2009] Novak, P. K., Lavrac, N., and Webb, G. I. (2009). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403.

[Ouali et al., 2016] Ouali, A., Loudni, S., Lebbah, Y., Boizumault, P., Zimmermann, A., and Loukil, L. (2016). Efficiently finding conceptual clustering models with integer linear programming. In Kambhampati, S., editor, *Proceedings of the Twenty-Fifth International Joint*

Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pages 647–654. IJCAI/AAAI Press.

[Soulet et al., 2011] Soulet, A., Raïssi, C., Plantevit, M., and Crémilleux, B. (2011). Mining dominant patterns in the sky. In *ICDM*, pages 655–664. IEEE Computer Society.

[Suzuki and Zytkow, 2005] Suzuki, E. and Zytkow, J. M. (2005). Unified algorithm for undirected discovery of exception rules. *Int. J. Intell. Syst.*, 20(7):673–691.

[Ugarte et al., 2017] Ugarte, W., Boizumault, P., Crémilleux, B., Lepailleur, A., Loudni, S., Plantevit, M., Raïssi, C., and Soulet, A. (2017). Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems. *Artif. Intell.*, 244:48–69.

[Ugarte et al., 2015] Ugarte, W., Boizumault, P., Loudni, S., and Crémilleux, B. (2015). Modeling and mining optimal patterns using dynamic CSP. In *27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, Vietri sul Mare, Italy, November 9-11, 2015*, pages 33–40. IEEE Computer Society.