EDAMI Project semester 2020z

Contact information:

Grzegorz Protaziuk, PhD, email: g.protaziuk@ii.pw.edu.pl

Consultations: to schedule a zoom meeting during project consultation hours (Thursday, 14-16) , please email me in advance.

The aim of the project

Implementation of a data mining algorithm and testing its properties (scaling, influence of parameters values on results)

Tools

All programming tools (IDE, languages, libraries) legally available for this type of project are allowed, however in the case of an "exotic" (not popular) programming language, it should be agreed with me.

Datasets

In experiments publicly available datasets should be used, e.g.:

https://archive.ics.uci.edu/ml/datasets.php,
http://fimi.uantwerpen.be/data/,
https://github.com/deric/clustering-benchmark,
http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

Important Dates

02.12.2020–deadline for choosing project topic (by WUT mail, please include [EDAMI project] in the title of the message). Please, include in the message three preferred topics in the order of preference. The projects will be assigned according to order of receiving mails and the preference. For projects realized by teams please address the message to all members of the team.

18.12.2020 – deadline for presentation of the key elements of the proposed solution during the meeting. Some writing materials (e.g. ppt presentation) should be prepared for that meeting.

22.01.2021 – the final date for submitting the final version of documentation. The final project report should be accompanied by the source and executable files, used input data as well as resulting output data. Each student (group) has to present to me how his/her program works at the latest by January 22th, during my consultation hours . The report should be sent at least two working days before the presentation.

Report - requirements

The final project report should contain:

- a description of the task (made assumptions);
- a description of the form of input and output data;
- a description of all important design (e.g. class diagram(s)) and implementation issues;
- an objective and a way of carrying out experiments
- a presentation of the performed experiments including a description of input data;
- obtained results (qualitative and/or quantitative, also performance time) – conclusions;
- all other issues you find worth presenting;
- description how to use/install a software.

Topics

1. Implementation of an algorithm for discovering association rules (Apriori, Eclat, …) allowing hierarchy application. (1-2 students).
2. Implementation of an algorithm for discovering association rules (Apriori, Eclat, …) with items occurrence constraints. (1-2 students).

3. Implementation of an algorithm for mining negative association rules (1-2 students)
   Literature
   Maria-Luiza AntonieOsmar R. Zaïane “Mining Positive and Negative Association Rules: An Approach for Confined Rules " PKDD 2004
4. Implementation of an algorithm for mining generalized association rules (1-2 students)
   Literature:
   Srikant, Ramakrishnan, and Rakesh Agrawal. "Mining generalized association rules." (1995): 407-419.
5. Implementation of an agglomerative hierarchical clustering (1 student).
6. Implementation of Neighborhood-Based Clustering algorithm (1 student).
   S. Zhou, Y. Zhao, J. Guan, and J. Huang, “A Neighborhood-Based Clustering Algorithm,” in Advances in Knowledge Discovery and Data Mining
7. Implementation of K-medoids (PAM) algorithm with application of parallel execution (1 student).
8. Implementation K-medoids (PAM) algorithm for clustering objects with nominal and numerical attributes. (1 student).
9. Implementation K-means algorithm for clustering objects with nominal and numerical attributes. (1 student).
10. Implementation of CN2 algorithm for rules induction. (1-2 students)
    Literature: Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. Machine Learning, 3, 261–283
11. Implementation of an algorithm (GSP (generalized), SPADE, PrefixSpan , ERMiner) for frequent pattern discovery (1-2 students)
    Literature
- R. Srikant, R. Agrawal , “Mining sequential patterns: Generalizations and performance improvements,” In Proceedings of International Conference on Extending Database Technology, pp. 3–17, 1996,
- Zaki, M. J. ,“SPADE: An efficient algorithm for mining frequent sequences”, Machine learning, vol.42.no.1-2,pp.31-60,2001,
- Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M., “Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach”, IEEE Trans. Knowledge and Data Engineering, vol. 16, no. 10, pp. 1-17, 2001,
- Founier, P., Zida, S., Guenieche, T. and Tseng V., “ERMiner: Sequential Rule Mining using Equivalence Classes”, Advanced in intelligent data Analysis, 13th InternationalSymposium, pp . 108-119, 2014.

12. Student’s task agreed with me.