Topic ID: 296-IT-EN-01

Data Clustering and Segmentation

Supervisor: doc. Ing. Petr Šilhavý, Ph.D.

Consultant: doc. Ing. Radek Šilhavý, Ph.D.

Department: Department of Computer and Communication Systems

Programme: Information Technologies

Abstract:

The proposed dissertation project focuses on the design and implementation of a novel adaptive clustering algorithm to support system classification in the early phases of software development planning. The research is motivated by the practical need to categorize systems and projects under incomplete and evolving information, where system features, documentation quality, and available characteristics may change over time. The proposed algorithm is intended to adapt dynamically to changing feature sets, handle missing or partially observed characteristics, and remain robust under heterogeneous project descriptions.

The research aims to investigate techniques for analyzing similarity in systems' descriptions and to evaluate whether similarity-driven approaches improve the accuracy and stability of downstream estimation tasks (e.g., effort or size estimation) when compared to conventional baselines. The work will therefore cover the analysis and design of data segmentation and clustering strategies, including the evaluation of kernel functions and penalization mechanisms for weighting or discounting specific observations within datasets (e.g., sliding windows and weighted-window schemes). A key outcome of the topic is the development of mathematical models and an implemented prototype for feature subset selection from system characteristics, enabling adaptive clustering under evolving data and supporting reproducible empirical evaluation.

Literature:

[1] P. Mohagheghi, B. Anda, and R. Conradi, "Effort estimation of use cases for incremental large-scale software development," pp. 303-311, 2005, doi: 10.1109/icse.2005.1553573.

[2] G. Claeskens and N. L. Hjort, Model selection and model averaging (Model Selection andModel Averaging). 2008, pp. 1-312.

[3] D. F. Hendry and J. A. Doornik, "Empirical model discovery and theory evaluation:automatic selection methods in econometrics," Empirical Model Discovery and Theory Evaluation, 2014.

[4] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, Bayesian data analysis, third edition (Bayesian Data Analysis, Third Edition). 2013, pp. 1-646.

R. Silhavy, P. Silhavy, and Z. Prokopova, "Evaluating subset selection methods for use casepoints estimation," Inform Software Tech, vol. 97, pp. 1-9, 2018/05/01/ 2018, doi:https://doi.org/10.1016/j.infsof.2017.12.009.