Bc. Kristián Barna
Bc. Marián Pukančík
Bc. Martin Čermák
Bc. Matúš Galba

MINING SEQUENCE DATA

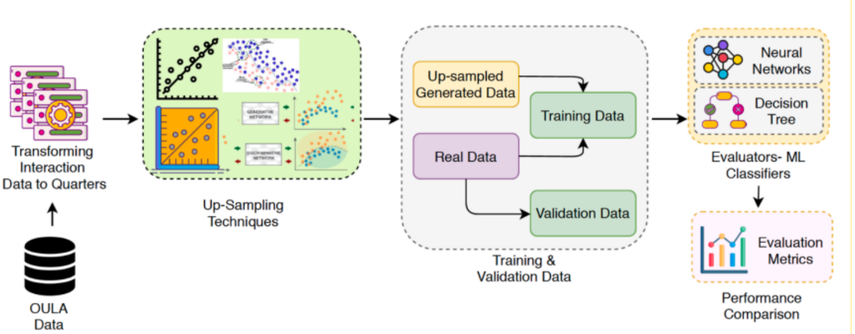
BALANCING SEQUENTIAL DATA TO PREDICT STUDENTS AT-RISK USING ADVERSARIAL NETWORKS

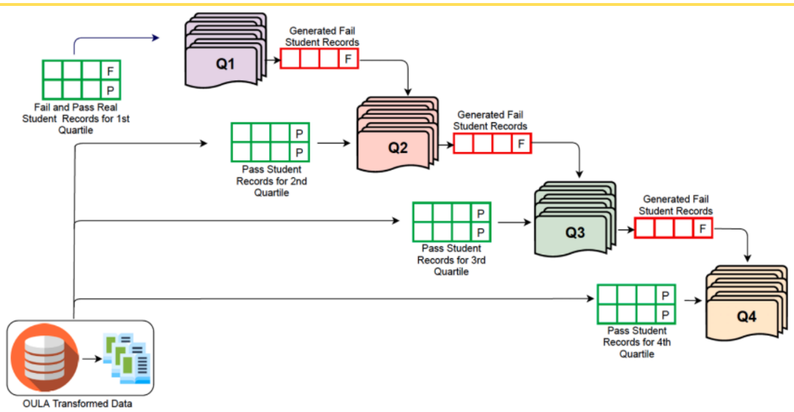
Motivation: Due to pandemic situation in late 2020 many online courses appeared and the data about students and their interactions with online platforms incremented massively. This provide researches with multiple dimensions to explore the area of students behavior.

The issue in machine learning is imbalance in data classes. In students pass/fail datasets tend to be much more "passed" students instances" then the "failed" ones.

Data: This study utilizes OULA (Open University Learning Analytics) dataset to eliminate its class imbalance between "pass" and "fail" students. The data are transformed into "quarterly" sequential format and each quarter is appended with next quarter. Then up-sampling approaches are used to eliminate mentioned class imbalances.

The objectives of this study can be described in three steps:
1. The raw data is transformed into a temporal sequential format.
2. Modified Generative Adversarial Networks are implemented for up-sampling instances of minority class - in sequential quarterly setting.
3. Comparison and evaluation of proposed approaches.





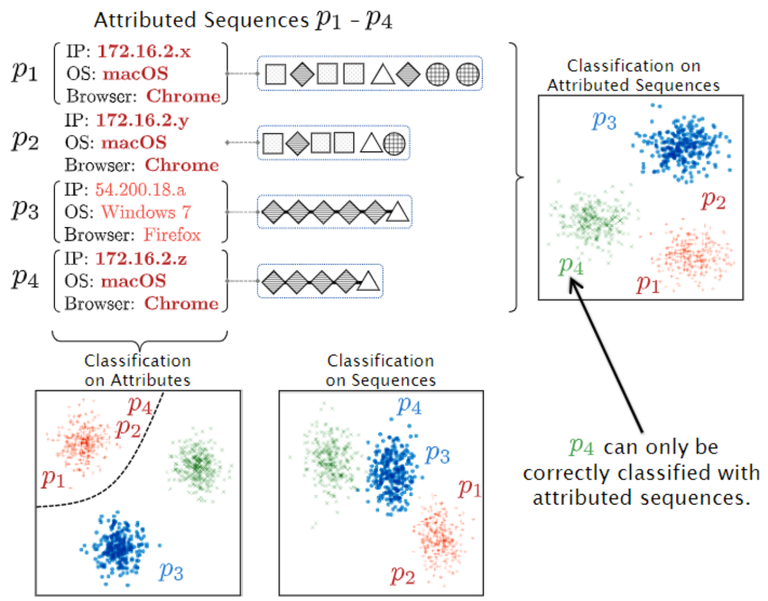
Hajra Waheed, Muhammad Anas... Balancing sequential data to predict students at-risk using adversarial networks, Computers & Electrical Engineering, Volume 93, 2021, 107274, ISSN 0045-7906

ATTENTION MODEL FOR ATTRIBUTED SEQUENCE CLASSIFICATION

Motivation: Classification over sequential data has seen a lot of applications from information retrieval, anomaly detection to genomic analysis. However, recent innovations in sequence classification learn from not only the sequences but also the associated attributes, called attributed sequences. This allows to find new classes that wouldn't be visible when using only one or another.

Data: The first part of the dataset consists of log files that contain user sessions logged in an information system of a firm in the form of attributed sequences.
The attributes include office name, system configuration and the sequence consist of click activities invoked by the user. Second part of the dataset comes from the online game Wikispeedia which requires the players to click through from a given start page to an end page in fewest clicks. The sequence consist of the pages visited in this path, ant the attributes include time spent per click, category of start path etc.

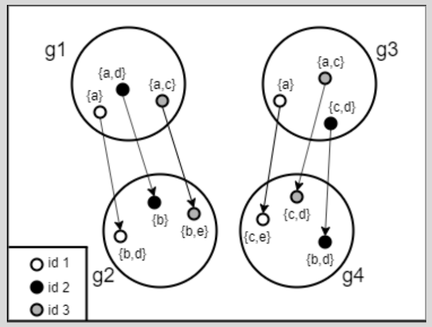
Approach: AMAS is the first framework that employs neural moment as a way to classify attribute sequences. It consists of fully connected neural network which transforms the attributes into attribute vector. It is concatenated with a LSTM which learns from the sequence part of the data. The goal is to minimize the cross-entropy between predicted and true labels.



Zhongfang Zhuang, Xiangnan Kong, Elke Rundensteiner: AMAS: Attention Model for Attributed Sequence Classification, 2019 SIAM International Conference on Data Mining (SDM), p. 109 - 117

LOCATION-BASED PARALLEL SEQUENTIAL PATTERN MINING

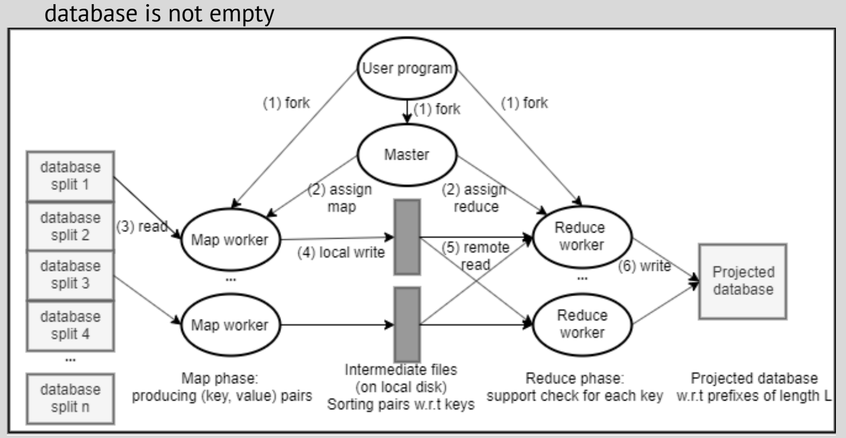
Motivation: An algorithm to mine frequent patterns from location-based sequential data stored in big databases.
Contribution: An approach for mining more specific type of sequential data. Mining location-based sequences could bring benefit to many industries by providing necessary services to customers in more efficient ways and thus, generating more income.
A location-based sequence: has following form: <l_1l_2...l_t>, where each 'l_i' is a location-based itemset of a form: l_i = [a, g], where 'g' belongs to region set and 'a' is an itemset.



ID	SEQUENCE
1	<[a,g1][bd, g2][a,g3][ce,g4]>
2	<[ad,g1][b, g2][cd,g3][bd,g4]>
3	<[ac,g1][be, g2][ac,g3][cd,g4]>

Approach: Two algorithms are presented: Naïve Location-based PrefixSpan (NLPS) and MapReduce Location-Based PrefixSpan (MRLPS). NLPS algorithm extends existing PrefixSpan algorithm to consider location data. MRLPS utilises multiple machines using the MapReduce framework.

- user program splits the (projected) database into N pieces and then creates several copies of itself , one master, the rest are called workers
- worker reads input split and calls Mapper() function which produces pairs (key, value), where "key" is a prefix of predefined length L and "value" is postfix
- these pairs are saved and sorted w.r.t keys
- other worker calls the Reducer() function, checks the support for each "key" and writes "values" into projected database
- the process is repeated for prefixes of length L+1 until the projected database is not empty



Byoungwook Kim, Gangman Y: Location-Based Parallel Sequential Pattern Mining Algorithm, IEEE Access (Volume 7), pp. 128651 - 128658 (2019)