POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXIX cycle

Advisor

Prof. Elena Baralis
Prof. Pietro Michiardi

Data Mining Algorithms for Big Data

PhD Candidate: *Fabio Pulvirenti*

EURECOM Sophia Antipolis

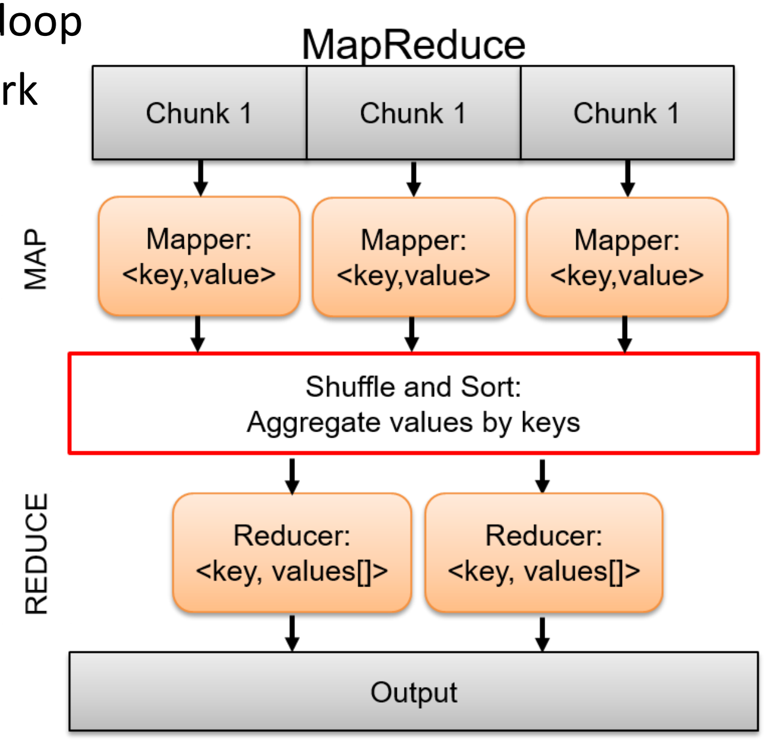
Frequent Itemset Mining for Big Data

Big Data

- Datasets so big and complex that require new architectures, techniques, algorithms, and analytics to be managed.
- Deficiency of data mining algorithms for Big Data in the state of the art
- Complexity of the problem
  - Not enough to only rearrange Data Mining algorithms
- **Research goal: design and develop advanced algorithms with scalable approaches**

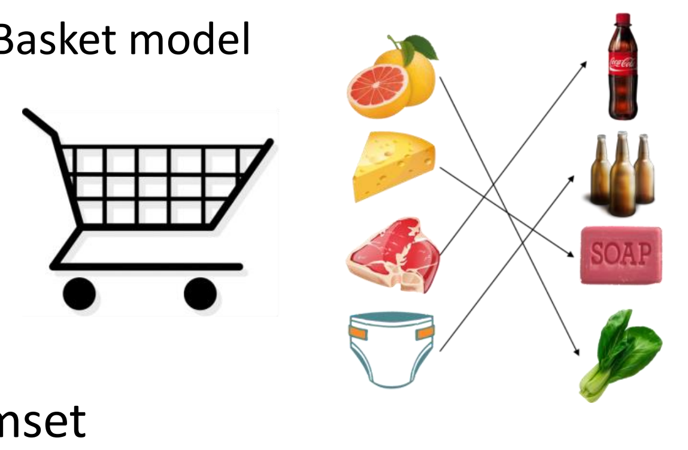
Methods & Tools

- Algorithms
- Map Reduce
  - Apache Hadoop
  - Apache Spark

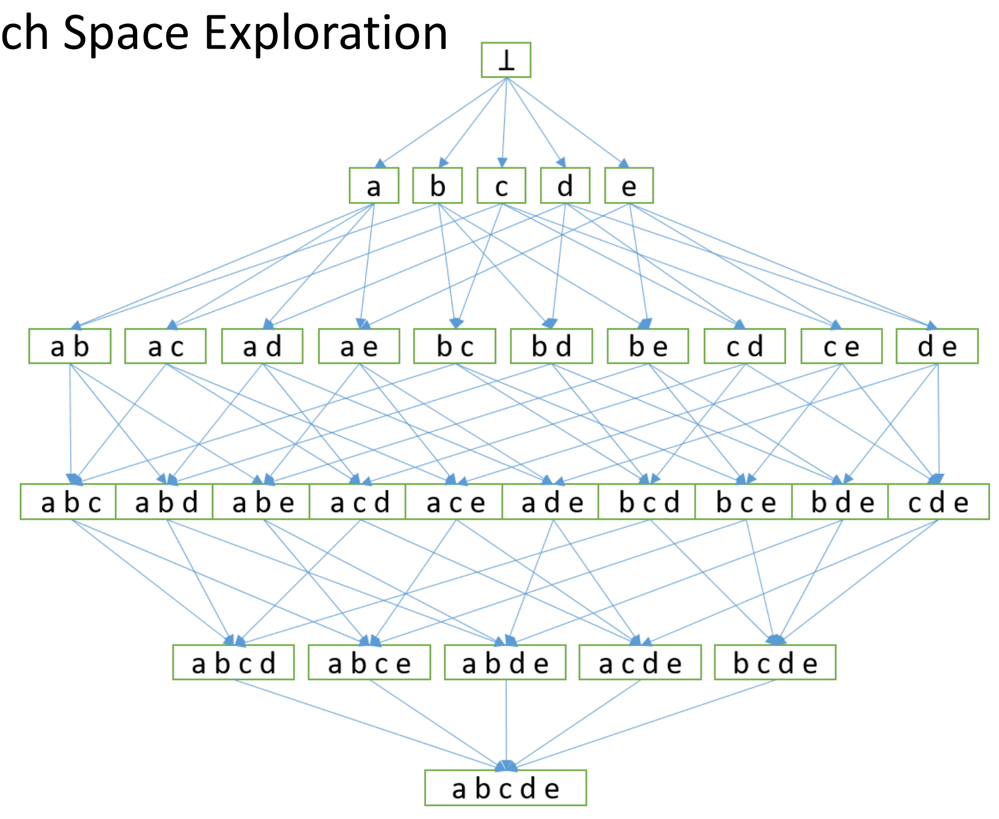


Frequent Itemset Mining

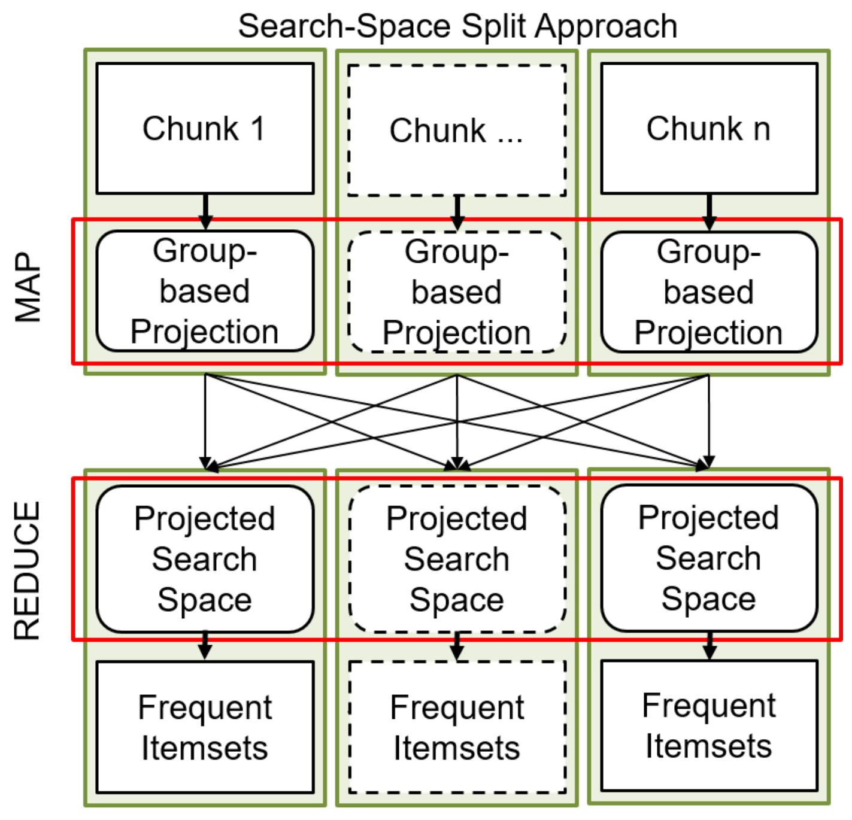
- Itemset
  - Transactional dataset
  - Itemsets
  - Market-Basket model



- Frequent Itemset
  - Minimum support threshold
- Association Rules
- Search Space Exploration



Partitioning Strategies

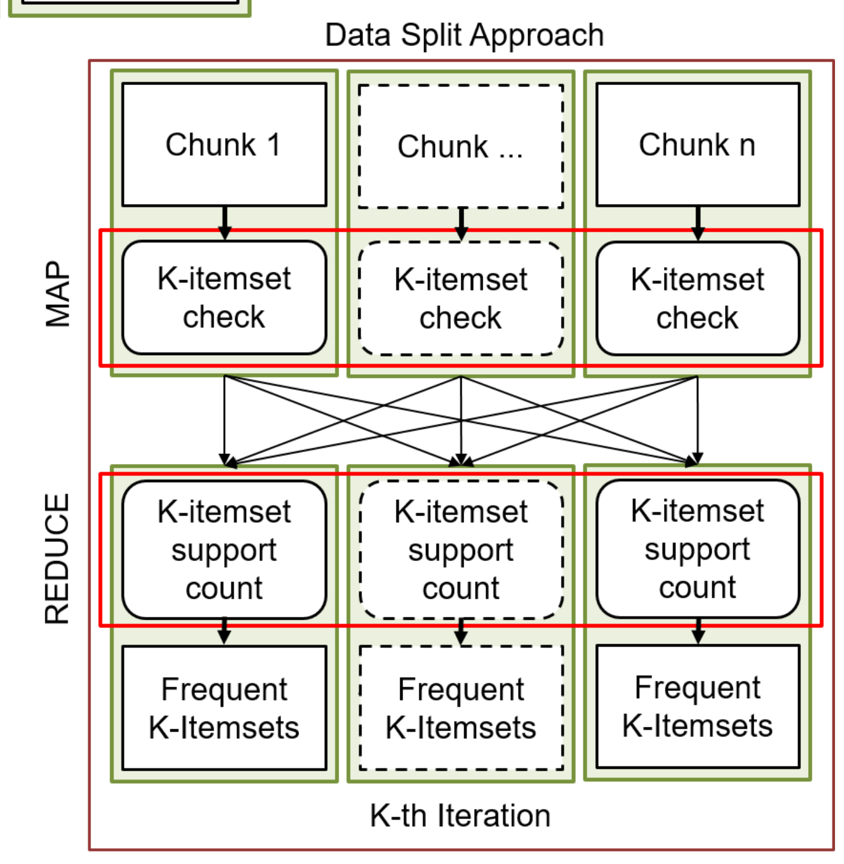


Search-Space Partitioning

- Each task explores a projected dataset which represents a subset of the search space
- Works on the complete set of input chunks, exploring just a sub-tree of the search space

Data Partitioning

- Each subtask computes the local supports of all candidate k-itemsets on one chunk of the input dataset.
- Works on the complete search space but with just one chunk of the input data.



High Dimensional Frequent Pattern mining

High dimensional datasets

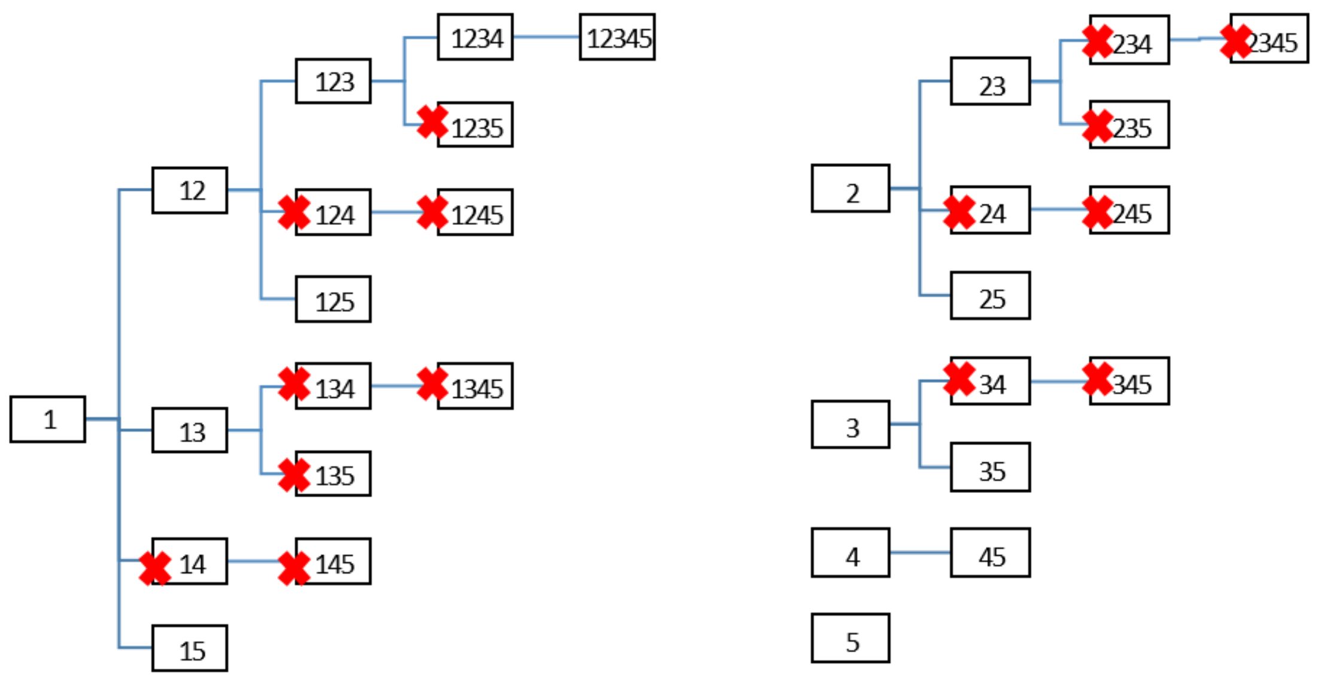
- State of the art analysis
  - Many approaches dealing with large number of transactions
  - No support for problems characterized by a large number of attributes (hundreds of thousands)
  - E.g., Bioinformatics, Smart Cities, …

ID	Att_1	Att_2	Att_3	…
1	Black	No	1	…
2	Red	Yes	4	…
…	…	…	…	…

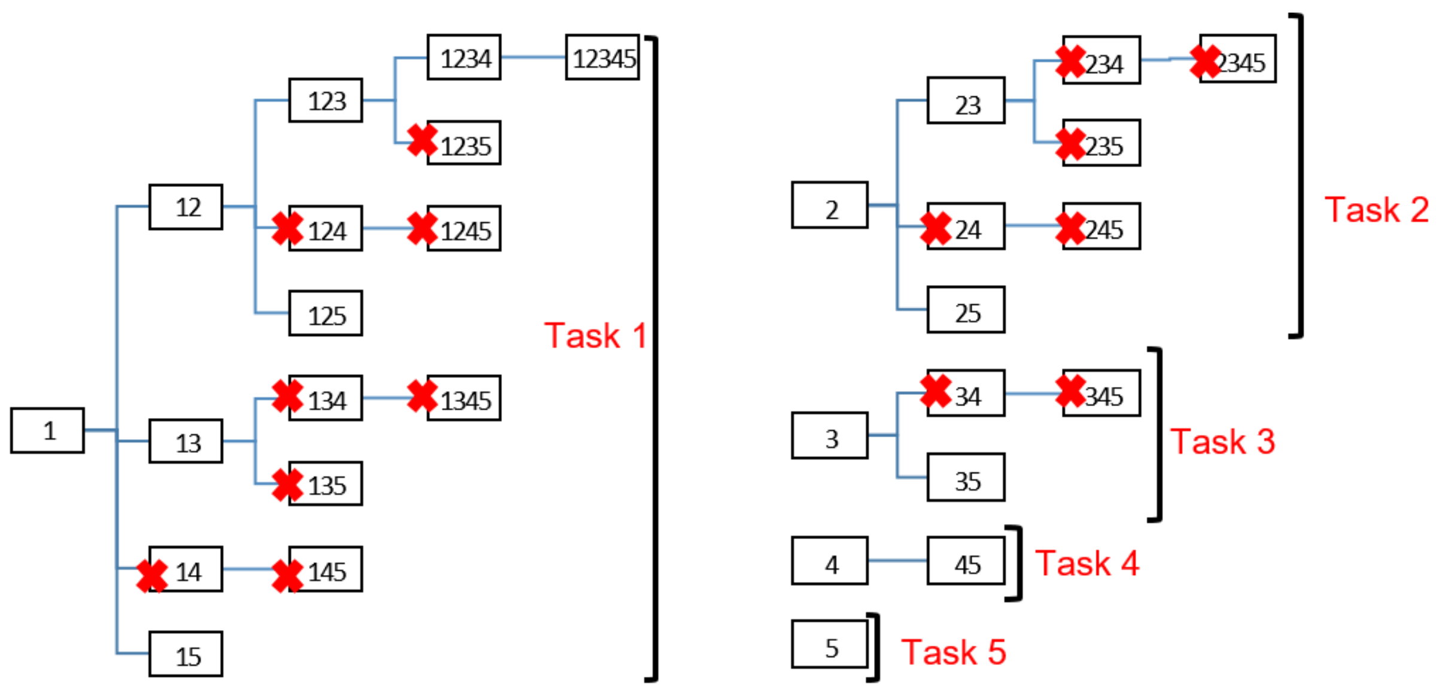
PaMPa-HD

- High-Dimensional frequent pattern miner
  - Row enumeration tree
- Fast and Scalable
  - Hadoop MapReduce
  - Swaps to the disk in case of memory issues and starts a new iteration
- Outperforming the state-of-the-art frequent pattern miners with high dimensional datasets

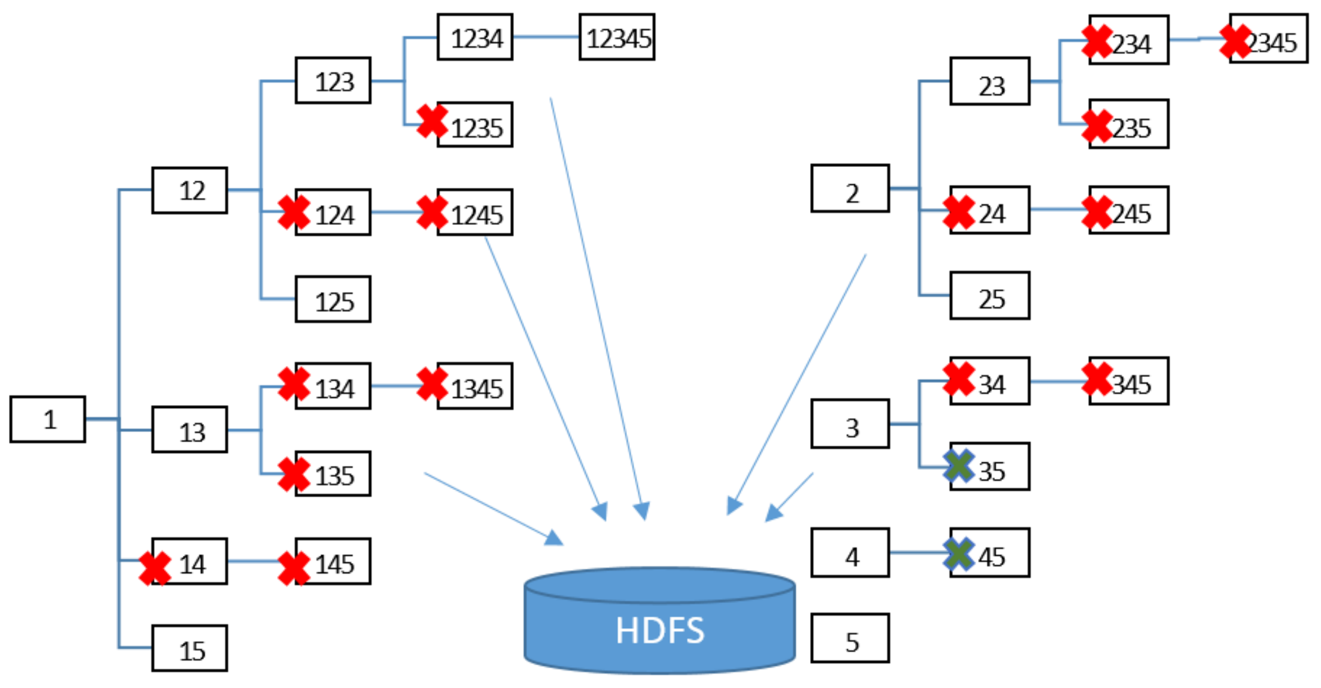
1. Row enumeration tree and pruning



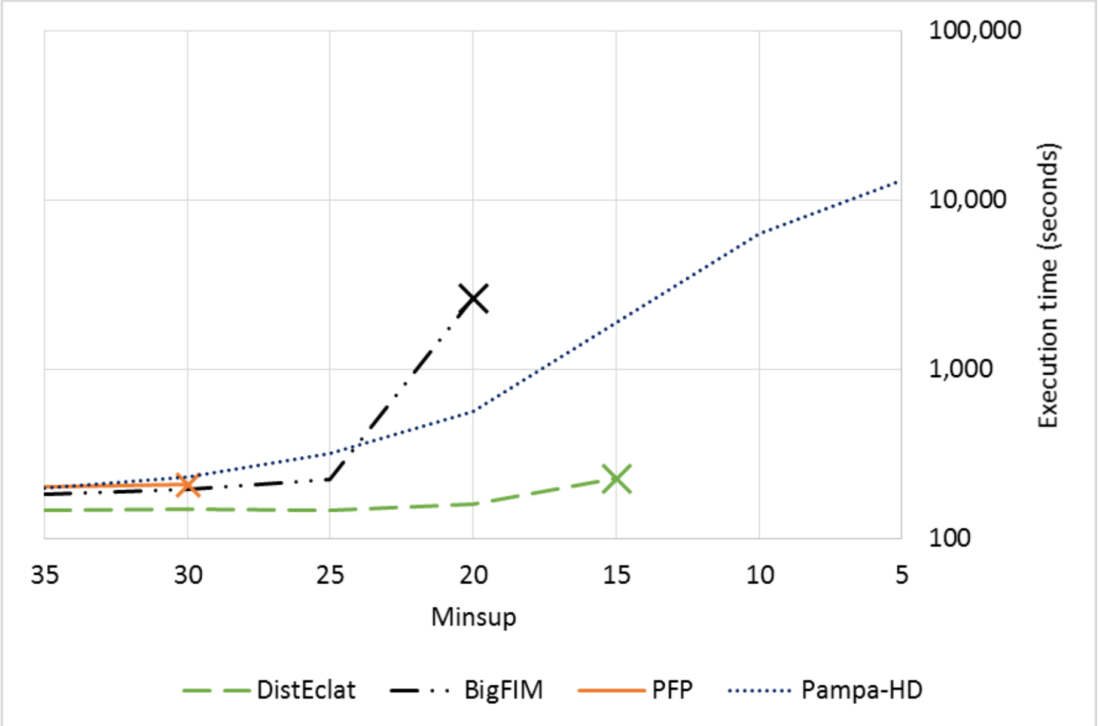
2. Parallelization



3. Synchronization for additional pruning



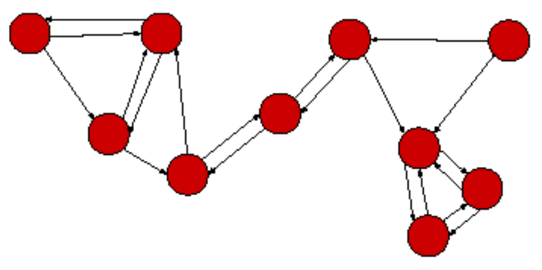
Performance



Related Activities

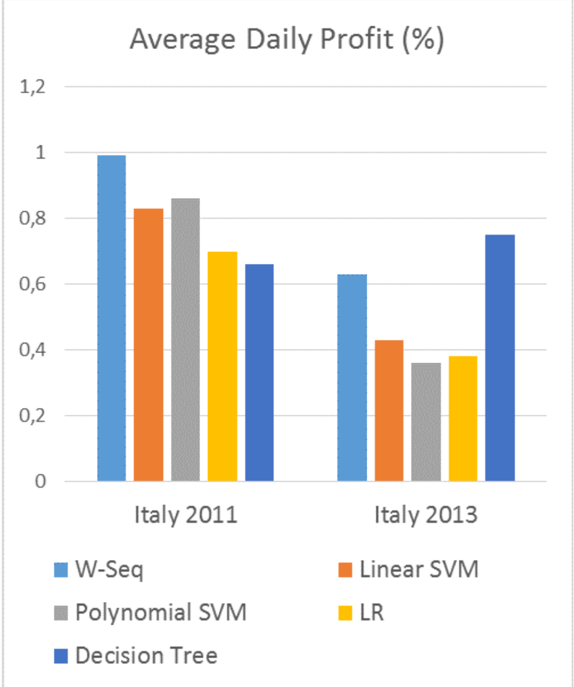
Streaming Distributed K-NN Graph

- Approximated algorithm
- Trade-off Accuracy vs Performance
- Streaming nature



Predictive Modeling for stock intraday trading

- Analysis of the historical values of the stock prices
- Weighted sequence mining and regression techniques
- Mutual influence between multiple stocks



Conclusion

- The target of my PhD is to thoroughly analyze the distributed and scalable data mining environment and making a step forward to fill in the discovered gap
- We focused on high-dimensional distributed frequent itemset mining
- Distributed algorithms and frameworks have been the travel companions of this 3-years journay
- Opportunity to deepen also other branches of data mining such as time-series analysis, clustering, classification and K-nn approaches.

References

1. Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Fabio Pulvirenti, and Pietro Michiardi. PaMPa-HD: a Parallel MapReduce-based frequent Pattern miner for High-Dimensional data. HDM 2015 @ IEEE ICDM 2015.
2. Daniele Apiletti, Paolo Garza, Fabio Pulvirenti. A review of scalable approaches for frequent itemset mining. BIGDAP 2015 @ ADBIS 2015.
3. Thibault Debatty, Fabio Pulvirenti, Pietro Michiardi, Wim Mees. Fast distributed k-nn graph update. BigGraphs 2016 @ IEEE BigData 2016.
4. Elena Baralis, Luca Cagliero, Tania Cerquitelli, Paolo Garza, Fabio Pulvirenti. Discovering profitable stocks for intraday trading. Submitted.