# Capstone Project Plan

Tahsin Mostafa

Supervision: Dr. Osmar R. Zaiane

### Motivation and objectives

Association rule mining [1] is the task of extracting co-occurrence relationships between items in a dataset. The typical algorithms for discovering these associations require very difficult-to-set parameters, such as support and confidence. In some applications where association rules are used, these parameters may have little meaning, making them even more difficult to set or interpret.

A less-known algorithm for mining association rules is the King-fisher [2]. It does not have the aforementioned parameters but relies on statistical significance. One issue with the King-fisher algorithm is that it assumes that each row in the dataset is represented as a vector of 0s and 1s with columns referring to attribute-value pairs, where 0 means an absence of the pair, and 1 means a presence. This can be problematic in terms of run-time and memory requirements of the algorithm when there are large numbers of attribute-value pairs in the dataset; as a result, the algorithm is unusable for large datasets, particularly with large dimensionality.

The objective is to reimplement the algorithm in a way that it can work with a more efficient representation of the dataset- where each row is represented by the attribute values themselves. This should greatly reduce the run-time and memory requirements and make it feasible for the algorithm to be used on large datasets. What makes King-fisher efficient is the Branch-and-Bound optimization strategy used. The project also requires the investigation of whether this optimization strategy can still be adapted to the new representation. Further, the project entails carrying out several experiments to determine the correctness of the algorithm and make sure it is scalable for large datasets.

The developed algorithm has the potential of improving several works related to developing associative classifiers [3], [4], explainability of black-box machine learning models [5], among others. Therefore, we will also measure this impact, particularly on the associative classifier.

### Experiments

1) Experiment for correctness: Is the reimplemented kingfisher algorithm able to generate the same frequent item sets as the original kingfisher algorithm for a given dataset.

2) Experiment for scalability: Is the reimplemented kingfisher algorithm able to generate frequent item sets faster than the original algorithm for large datasets; particularly with number of attribute-value pairs greater than 30. This is because the original kingfisher is unable to generate frequent itemsets for these large datasets.

3) Experiment for effectiveness on downstream tasks: Build an associative classifier [3], [4] based on the new kingfisher algorithm and compare performance with classifiers that were based on the original kingfisher algorithm.

## Expected results

We expect that the reimplemented algorithm will:

1) output correct frequent itemsets

2) output results faster than the original algorithm for large datasets

3) improve the performance of associative classifiers.

## Deliverables

1. Report on difference between binary and itemset representation of dataset in terms of efficiency of algorithms to find frequent item sets.

2. Implementation of the Kingfisher algorithm using the itemset representation in C++.

3. Experiments determining the correctness, scalability and effectiveness on downstream tasks of the reimplemented algorithm.

4. Final Report

## Timeline of deliverables

1. Deliverable 1: Jan 11 - Jan 17 (1 week)

2. Deliverable 2: Jan 18 - Feb 18 (4 weeks)

3. Deliverable 3: Feb 19 - March 19 (4 weeks)

4. Deliverable 4: March 20 - April 10 (3 weeks)

## Main Readings/Reference

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

[2] Hämäläinen, W. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. Knowl Inf Syst 32, 383–414 (2012). https://doi.org/10.1007/s10115-011-0432-2

[3] Li J, Zaiane OR. Exploiting statistically significant dependent rules for associative classification. Intelligent Data Analysis. 2017;21(5):1155-1172. doi:10.3233/IDA-163141

[4] Li, J., & Zaiane, O.R. (2015). Associative Classification with Statistically Significant Positive and Negative Rules. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*.

[5] Mohammad Motallebi, Md Tanvir Alam Anik, and Osmar R. Zaïane. 2023. Explaining Decisions of Black-Box Models Using BARBE. In Database and Expert Systems Applications: 34th International Conference, DEXA 2023, Penang, Malaysia, August 28–30, 2023, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 82–97. https://doi.org/10.1007/978-3-031-39821-6_6