Towards Proximity Graph Auto-Configuration: an Approach Based on Meta-learning

Rafael S. Oyamada, Larissa C. Shimomura, Sylvio Barbon Junior, and

Daniel S. Kaster.

Universidade Estadual de Londrina

CAPES

CNPq

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

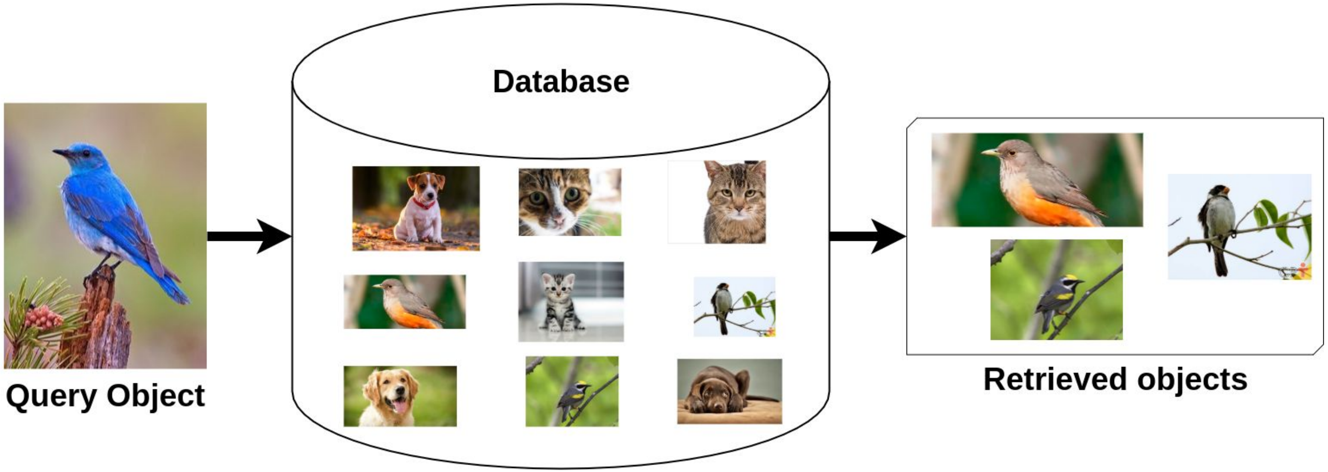
Summary

- Introduction and Concepts
  - Similarity Searches
  - Proximity Graphs
  - Meta-learning
- Contribution
- Experimental results
- Conclusion

Universidade Estadual de Londrina

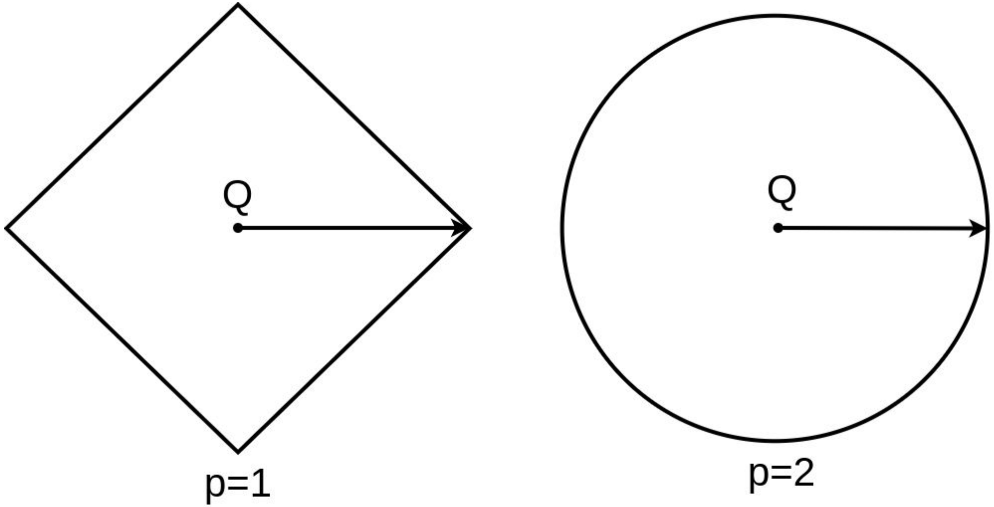
Busca por similaridade

Retrieving complex data (image, video, audio, etc) through its similarities.



Distance functions

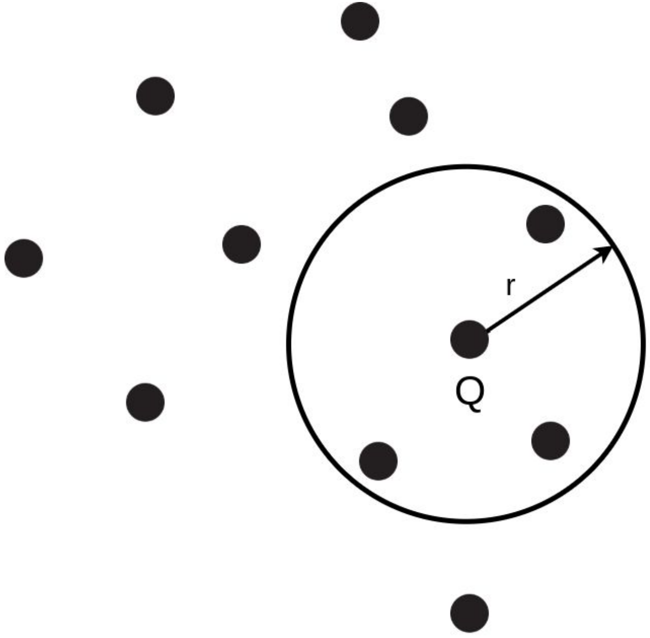
- Distance functions to measure the similarity between a pair of feature vectors.
- Lp norms: Manhattan (L1), Euclidean (L2)



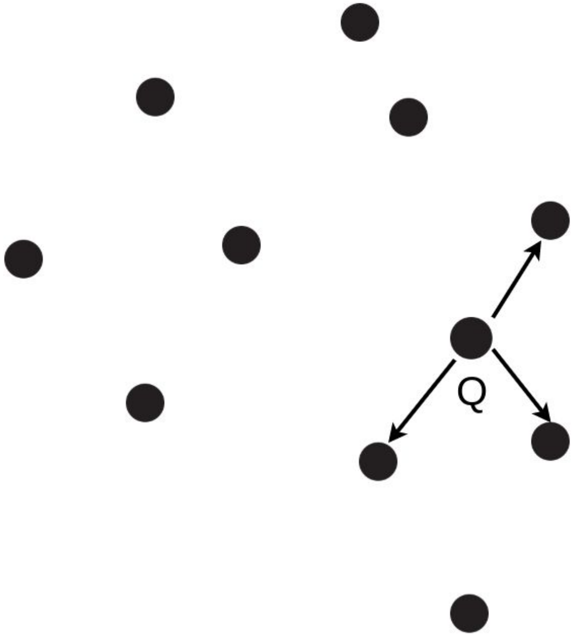
Universidade Estadual de Londrina

Similarity Queries

Range query



k-NN query (k=3)



Universidade Estadual de Londrina

Index structures for similarity searching

- Tree-based methods;
- Hash-based methods;
- Permutation-based methods;
- **Graph-based methods.**

Proximity Graphs

- A proximity graph is a graph $G=(V, E)$, in which each pair of vertices $(u, v) \in V$ is connected by an edge $e=(u, v)$ iff u and v satisfy a given property P;

Proximity Graphs

- Popular approaches are based on *k-NN* graphs or navigable small-world graphs (*NSW*);
- **Sensible to construction and search parameters.**

Parameters of major impact

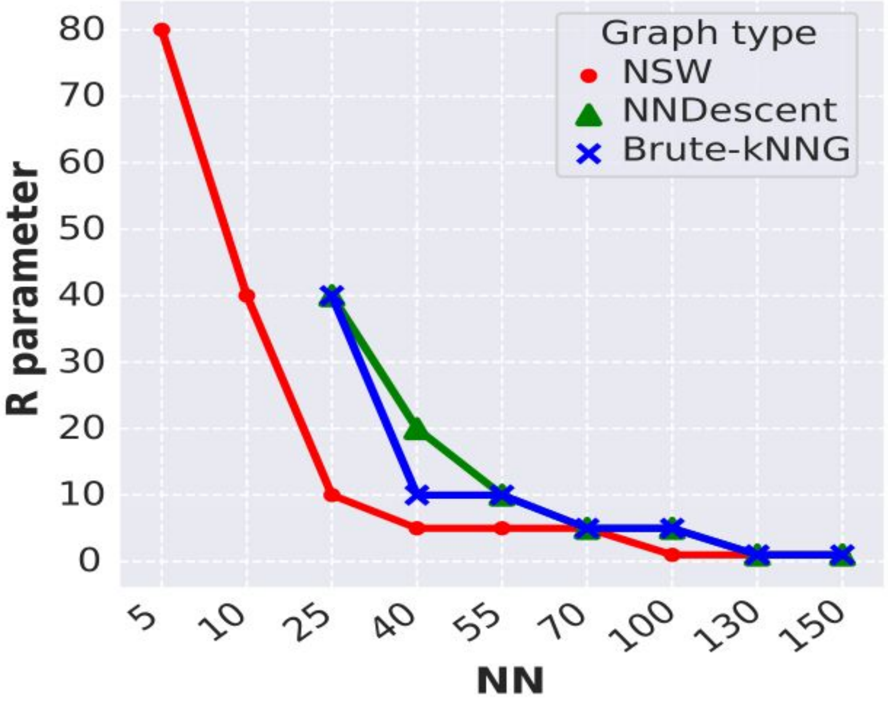
- Construction: number of nearest neighbors (NN)
- Query: number of restarts (R)
    - Regarding the *GNNS* algorithm

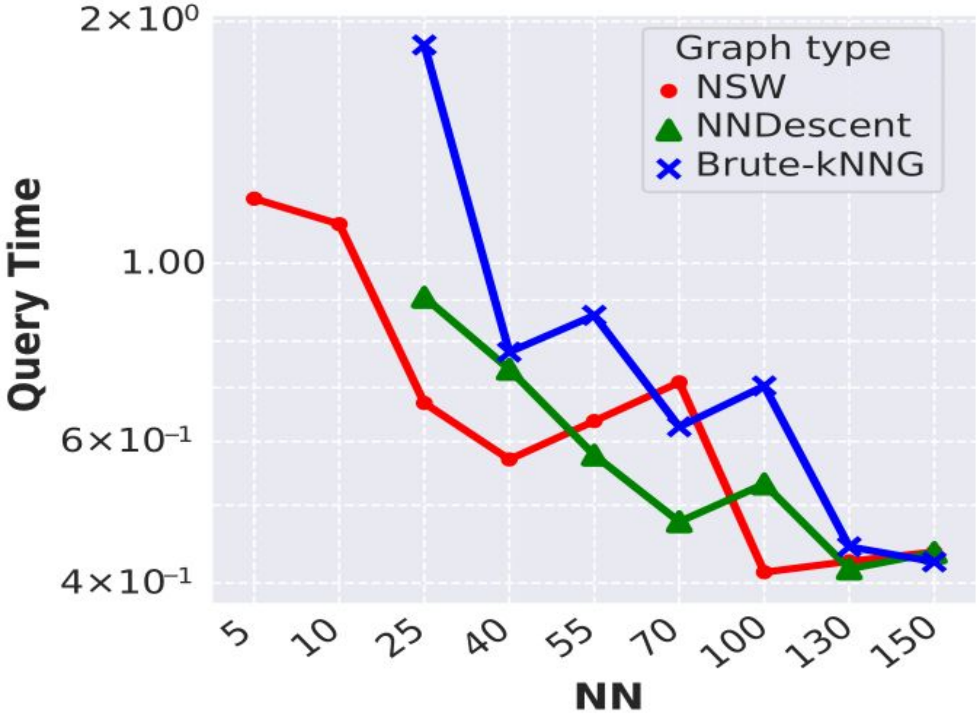
Usually chosen through grid search steps

Example: impact of parameters

- Choosing the best graph type and its configuration for a given dataset for achieving a minimum recall rate (0.95)
- Considering different optimization criteria
  - Memory usage, or
  - Query time

R (left) and Query Time (right) varying NN





Universidade Estadual de Londrina

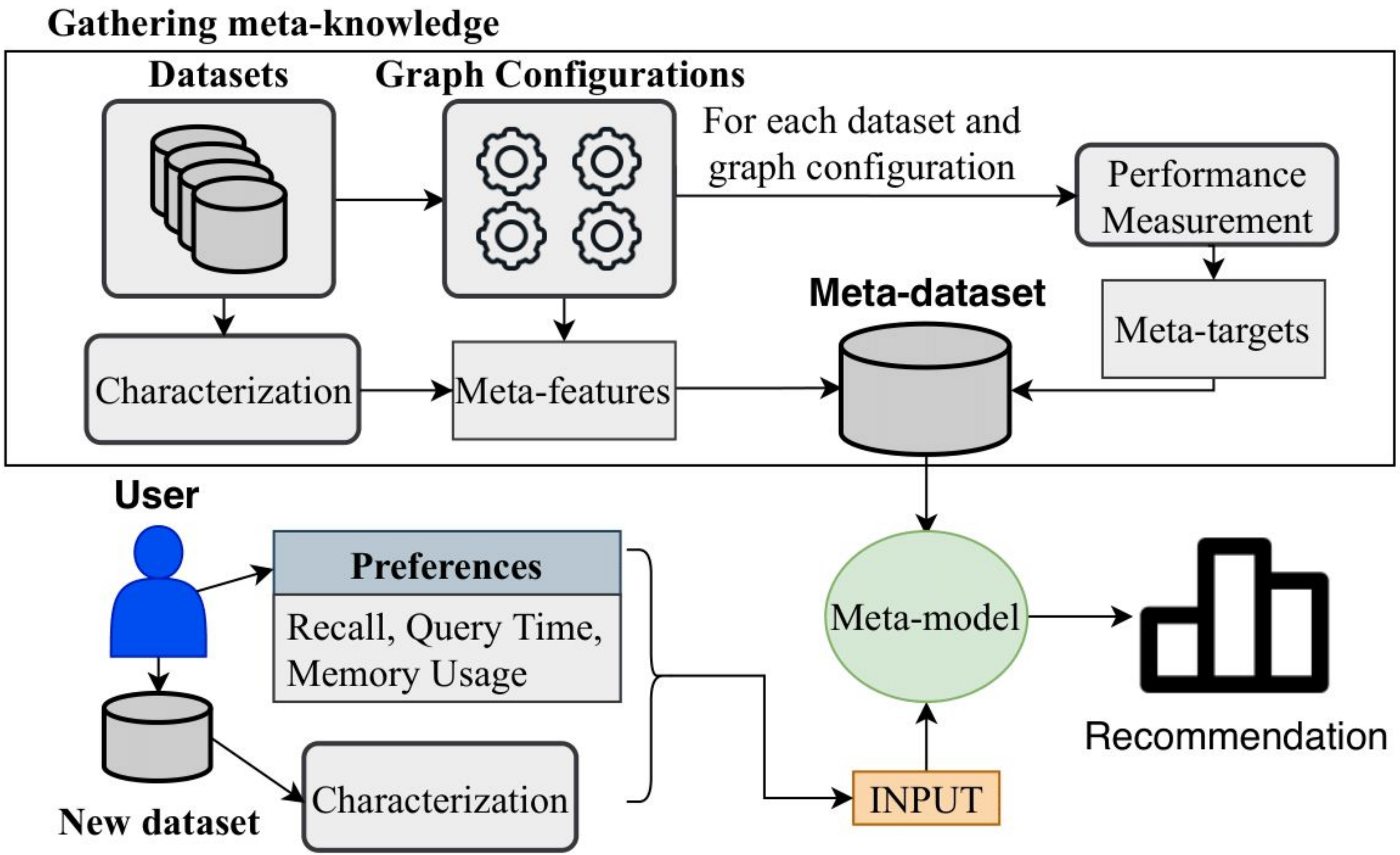
Contribution

An intelligent system, based on meta-learning techniques, capable of recommending a suitable proximity graph, together with its settings for a given dataset.

Meta-learning

- “Learning accross experiences”;
  - Gathering knowledge from several problems to learn how to provide suitable solutions in future.
- Algorithm selection, parameter recommendation, performance prediction, and etc;
  - Popular in machine learning community.

Proposal



Gathering meta-knowledge
Datasets
Graph Configurations
For each dataset and
graph configuration
Performance
Measurement
Meta-dataset
Meta-targets
Characterization
Meta-features
User
Preferences
Recall, Query Time,
Memory Usage
Meta-model
Recommendation
New dataset
Characterization
INPUT

Experiments

Universidade
Estadual de Londrina

Datasets

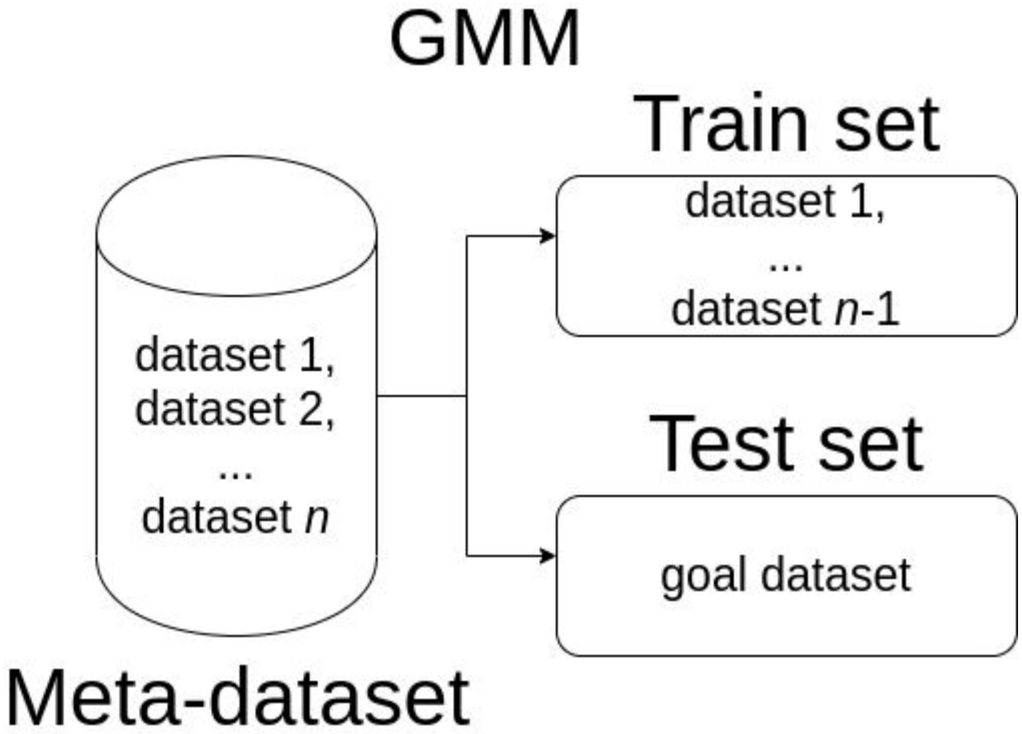
	Title	Size	Dimensions
Real	Color Moments	68,040	9
	Texture	68,040	16
	Color Histogram	68,040	32
	MNIST	70,000	784
	ANN-SIFT1M	1,000,000	128
	Properties	**Values**	
Synthetic	Size	$\{10^4, 10^5, 10^6\}$	
	Dimensionality	{8, 32, 128}	
	Gaussian distribution	{1, 5, 10}	
	Number of clusters	{1, 10, 100}	

Experimental setup

- C++ NMSLib for performance measurements
  - Brute Force *k-NNG*, *NNDescent*, and *NSW*
- k-NN queries using the Euclidean distance
- One meta-model for each performance measurement (recall and query time)
- Random Forests for meta-model induction
  - Scikit-learn default parameters

Tuning strategies: generic (no tuning)

Generic meta-model

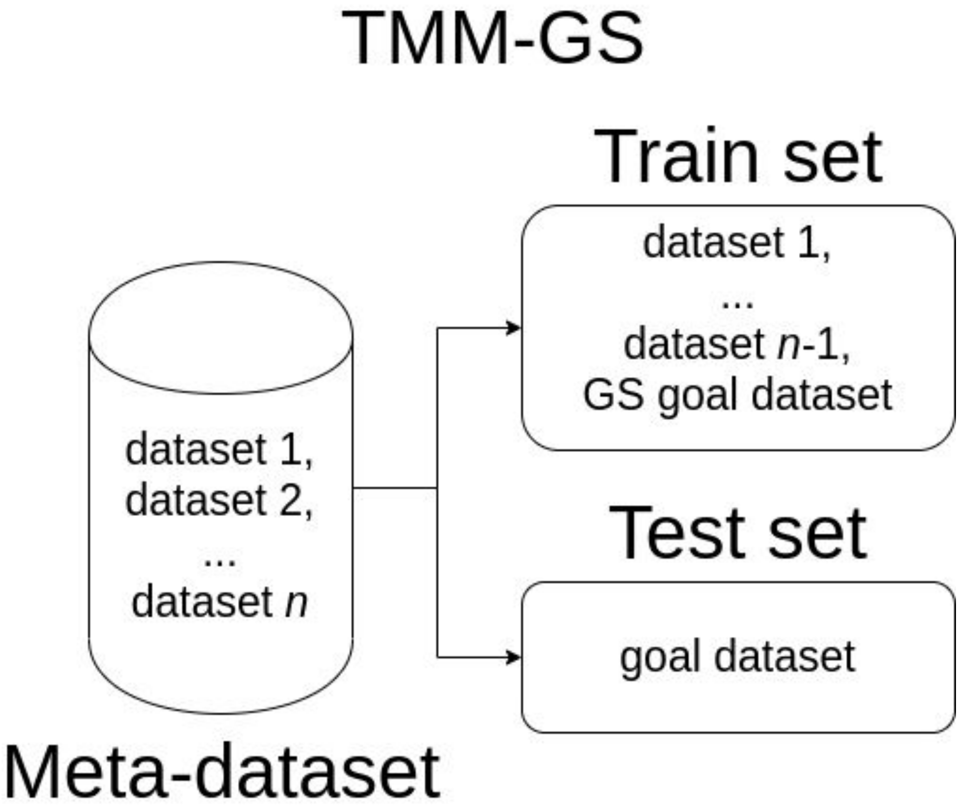


Universidade Estadual de Londrina

Tuning strategies: add grid search

Tuned meta-model:

Grid Search

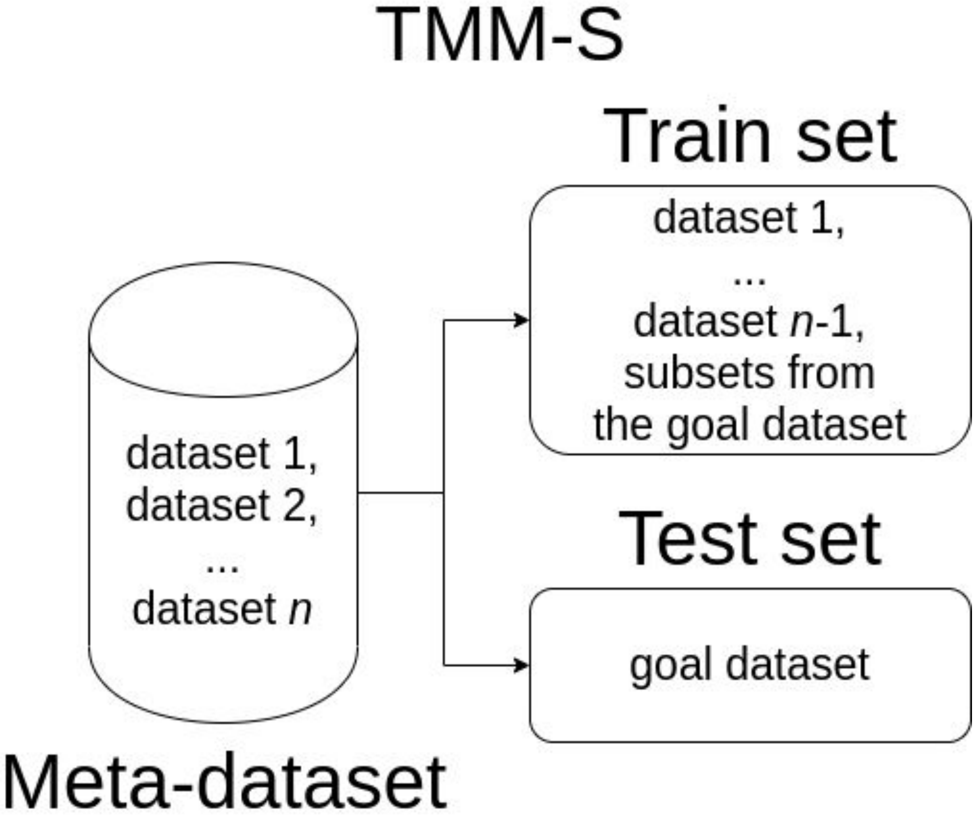


Universidade Estadual de Londrina

Tuning strategies: add grid search on subsets

Tuned meta-model:

Subsets



Universidade Estadual de Londrina

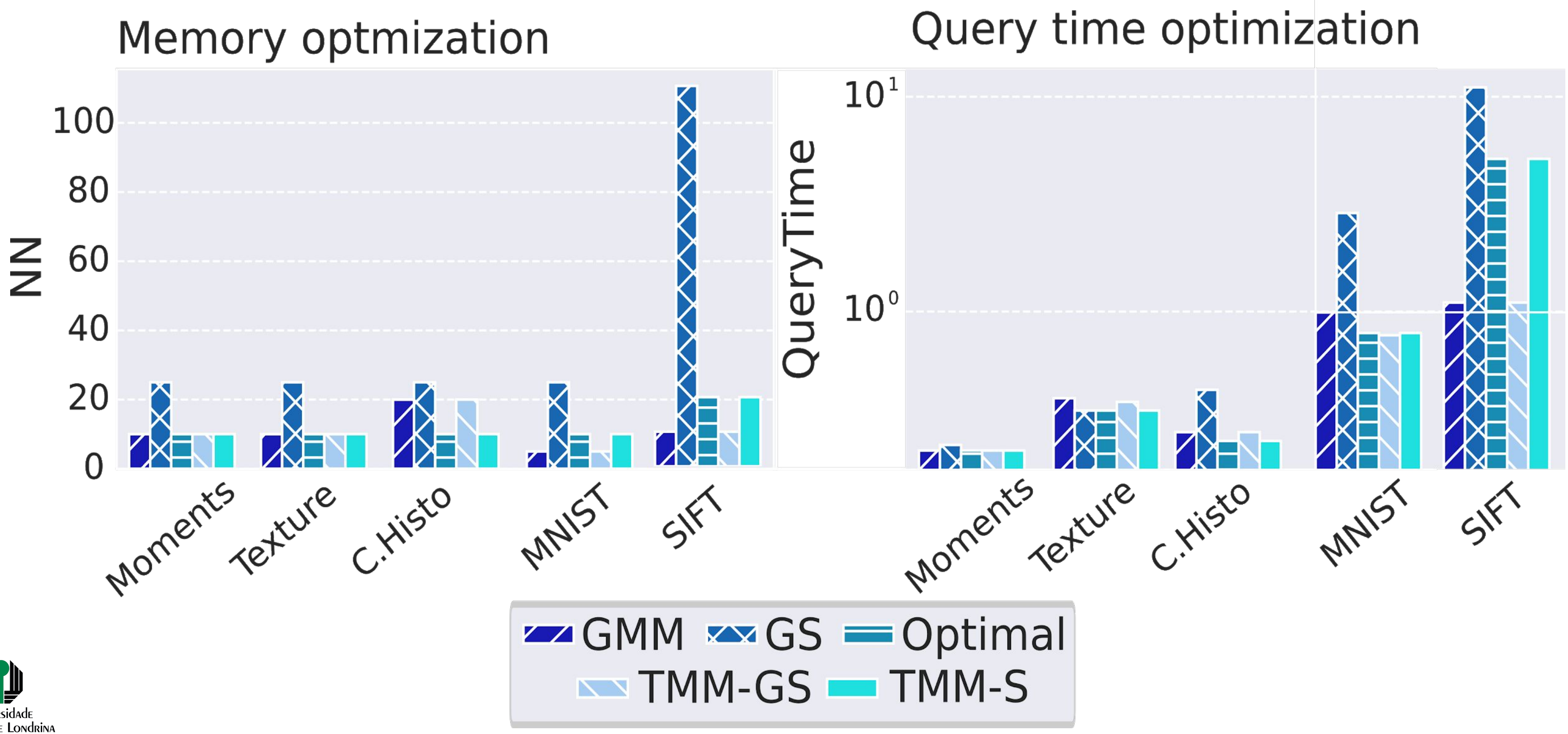
Accuracy evaluation: r-squared and RMSE

Goal Dataset	GMM				TMM-GS				TMM-S			
	Recall		QueryTime		Recall		QueryTime		Recall		QueryTime	
	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE	r^2	RMSE
Histogram	0.350	0.135	0.980	0.249	0.605	0.130	0.961	0.338	0.996	0.012	0.998	0.068
MNIST	0.765	0.111	0.694	1.097	0.617	0.173	0.920	0.559	0.997	0.014	0.998	0.068
Moments	0.955	0.034	0.989	0.179	0.973	0.031	0.979	0.241	0.991	0.019	0.998	0.065
SIFT	0.807	0.132	0.932	0.524	0.568	0.247	0.803	0.932	0.983	0.049	0.984	0.260
Texture	0.978	0.024	0.962	0.344	0.990	0.022	0.951	0.378	0.996	0.012	0.998	0.058

Recommendations

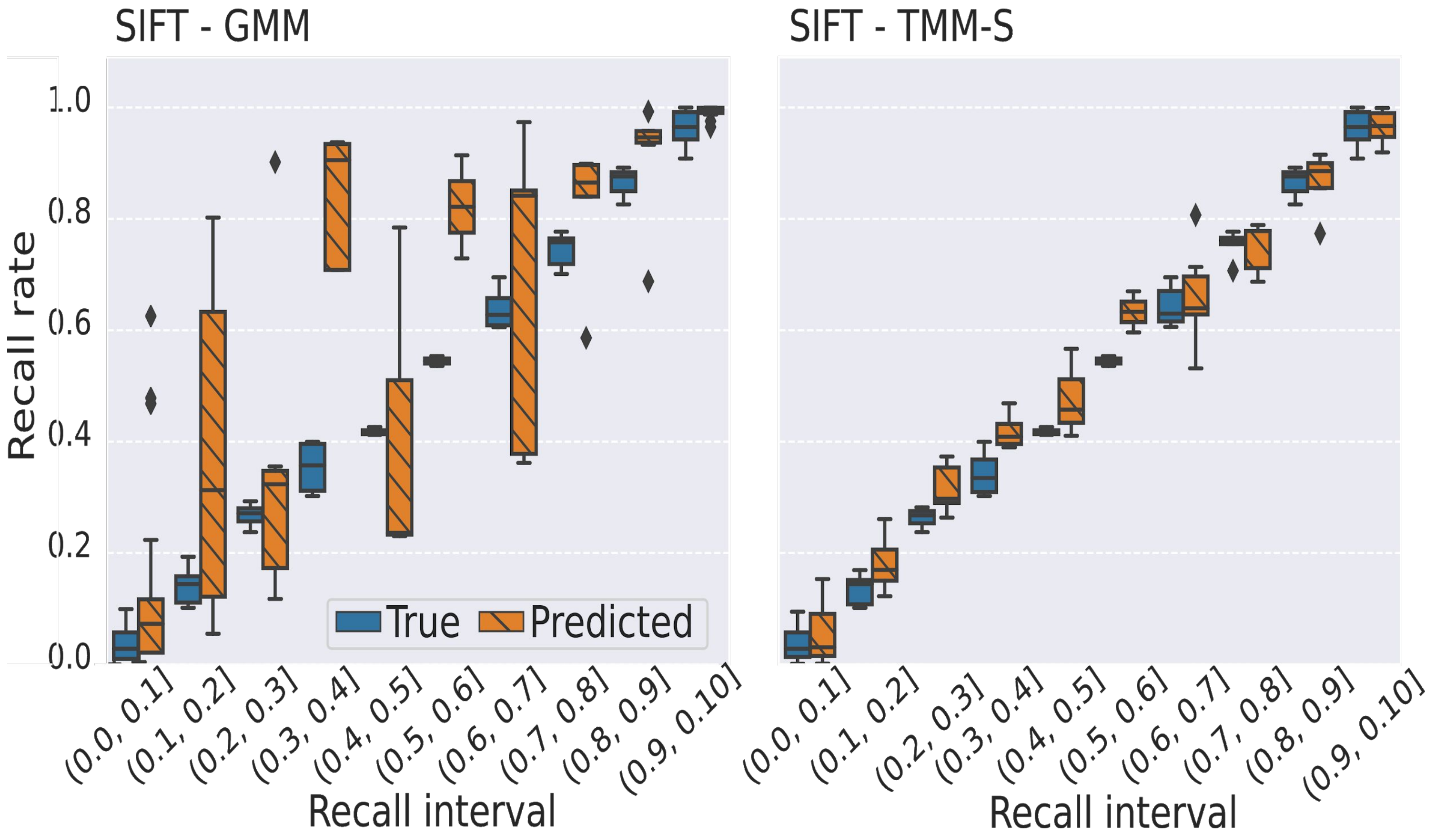
- Optimal: best graph configuration achieved from all results
- Grid search: best graph configuration achieved from a reduced parameter space
    - $NN = \{1, 25, 70, 150\}$
    - $R = \{1, 10, 40, 120\}$

Recommendation according to different criteria



Universidade Estadual de Londrina

Predictions per interval



Universidade
Estadual de Londrina

Conclusion and future works

- Overall, our approaches overcome the grid search method
- The TMM-S is able to reach optimal results in most cases
- Explore more dataset descriptors
- Increase the meta-dataset with more image datasets

Universidade Estadual de Londrina

Thank you!

Contact: rseidi.oyamada@uel.br

Universidade Estadual de Londrina

CAPES

CNPq

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY