Group based Unsupervised Feature Selection

The 24th Pacific-Asia Conference on
Knowledge Discovery and Data Mining (PAKDD 2020)

Authors:

Kushani Perera **(University of Melbourne)**

Jeffrey Chan **(RMIT University)**

Shanika Karunasekera **(University of Melbourne)**

Outline

- Introduction
- Motivation
- Our Approach
- Experimental Results
- Conclusion & Future Work

Introduction

- High dimensional datasets are used in:
  - Document clustering, gene selection, image matching
- Feature selection:
  - Improves prediction accuracy
  - Reduces computation costs
  - Produce simple learning models
- Most of the real world data is unlabelled
- Unsupervised feature selection is challenging

Motivation

- External sources of correlations within feature groups to improve the usefulness of unsupervised feature selection

- Feature grouping methods:
  - Pixels in images - Spatial locality
  - Words in text data - Semantics
  - Genes in genomic data - Gene Ontology terms

- Only few unsupervised feature selection methods use feature group information
  - k-medoid (KM)
  - Hierarchical Unsupervised Feature Selection (HUFS)

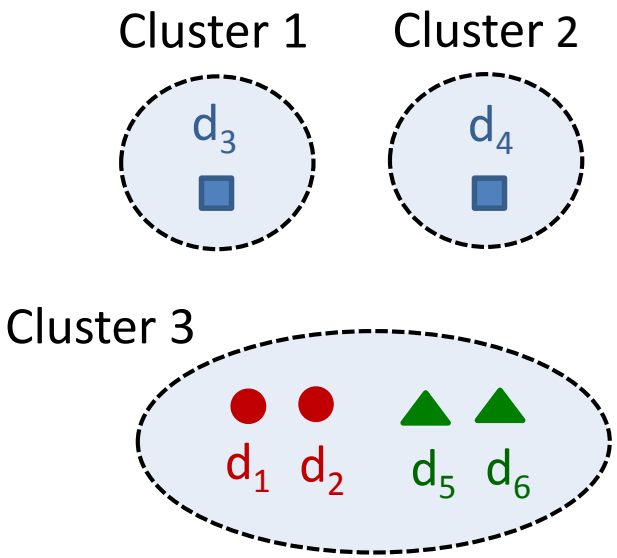
Instance-Feature based Feature Selection

	Bank	Patient	Cell	Google	Document Type
Document 1 (d_1)	**13**	0	0	0	Business
Document 2 (d_2)	**10**	0	0	0	
Document 3 (d_3)	0	**20**	0	0	Health
Document 4 (d_4)	0	0	**16**	0	
Document 5 (d_5)	0	0	0	**13**	Technology
Document 6 (d_6)	**1**	0	0	0	

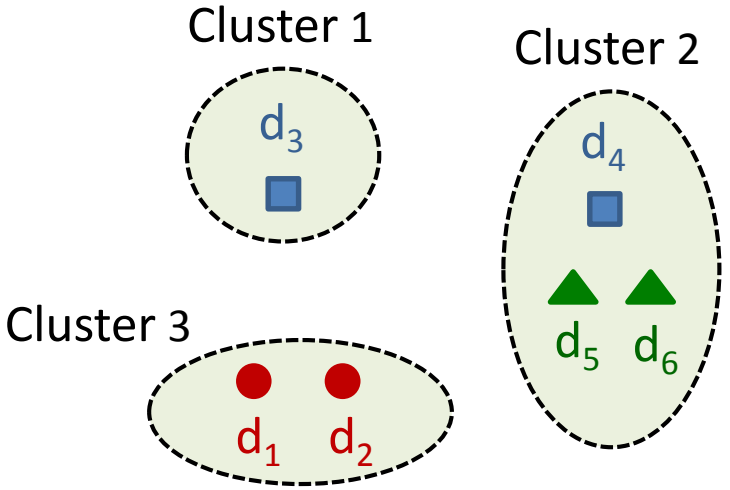
l_Bank = 0.39, l_Patient = 0.1.06, l_Cell = 1.06, l_Google= 1.1

Instance-Feature based Feature Selection

Cluster results for
{Bank, Patient, Cell}



Cluster results for
{Bank, Patient, Google}



Business Health Technical

Our Approach

- Propose a framework for unsupervised feature selection methods to exploit feature group information from external sources of knowledge
- Use this framework to incorporate feature group information in to Laplace Score objective
- Propose two optimisation methods:
  - Quadratic optimisation
  - Greedy optimisation

Modelling Feature Group Information

G =

	Bank	Patient	Cell	Google
Bank	0	0	0	0
Patient	0	0	**1**	0
Cell	0	**1**	0	0
Google	0	0	0	0

U =

	Bank	Patient	Cell	Google
Bank	**1**	0	0	0
Patient	0	**1**	0	0
Cell	0	0	**1**	0
Google	0	0	0	0

Objective: $\min ||UGU||_{1,1}$

Modelling Feature Group Information

S = {Bank, Patient, Cell}

UGU =

	Bank	Patient	Cell	Google
Bank	0	0	0	0
Patient	0	0	**1**	0
Cell	0	**1**	0	0
Google	0	0	0	0

$||UGU||_{1,1} = 2$

S = {Bank, Patient, Google}

UGU =

	Bank	Patient	Cell	Google
Bank	0	0	0	0
Patient	0	0	0	0
Cell	0	0	0	0
Google	0	0	0	0

$||UGU||_{1,1} = 0$

Reformulating Feature Selection Objective

Q =

	Bank	Patient	Cell	Google
Bank	0.39	0	0	0
Patient	0	1.06	0	0
Cell	0	0	1.06	0
Google	0	0	0	1.1

UQU =

	Bank	Patient	Cell	Google
Bank	0.39	0	0	0
Patient	0	1.06	0	0
Cell	0	0	1.06	0
Google	0	0	0	0

Objective: $\min ||UQU||_{1,1}$

Group Laplace Score (GLS)

Feature Selection Objective:

$$\min ||UQU||_{1,1} + \lambda . ||UGU||_{1,1} \text{ subject to } ||U||_{1,1} = k$$

Greedy method: Selects one feature (f) at a time such that

$$f = \min\_ x \in S' \quad l_x + \lambda . \frac{w_i}{\alpha_i}$$

λ = User defined parameter

L_x = Laplace score of feature x

G_i = Feature group of feature x

w_i = No. of features selected from G_i / No. of all selected features

α_i = Weight of G_i

S' = Unselected feature subset

Experiment Datasets

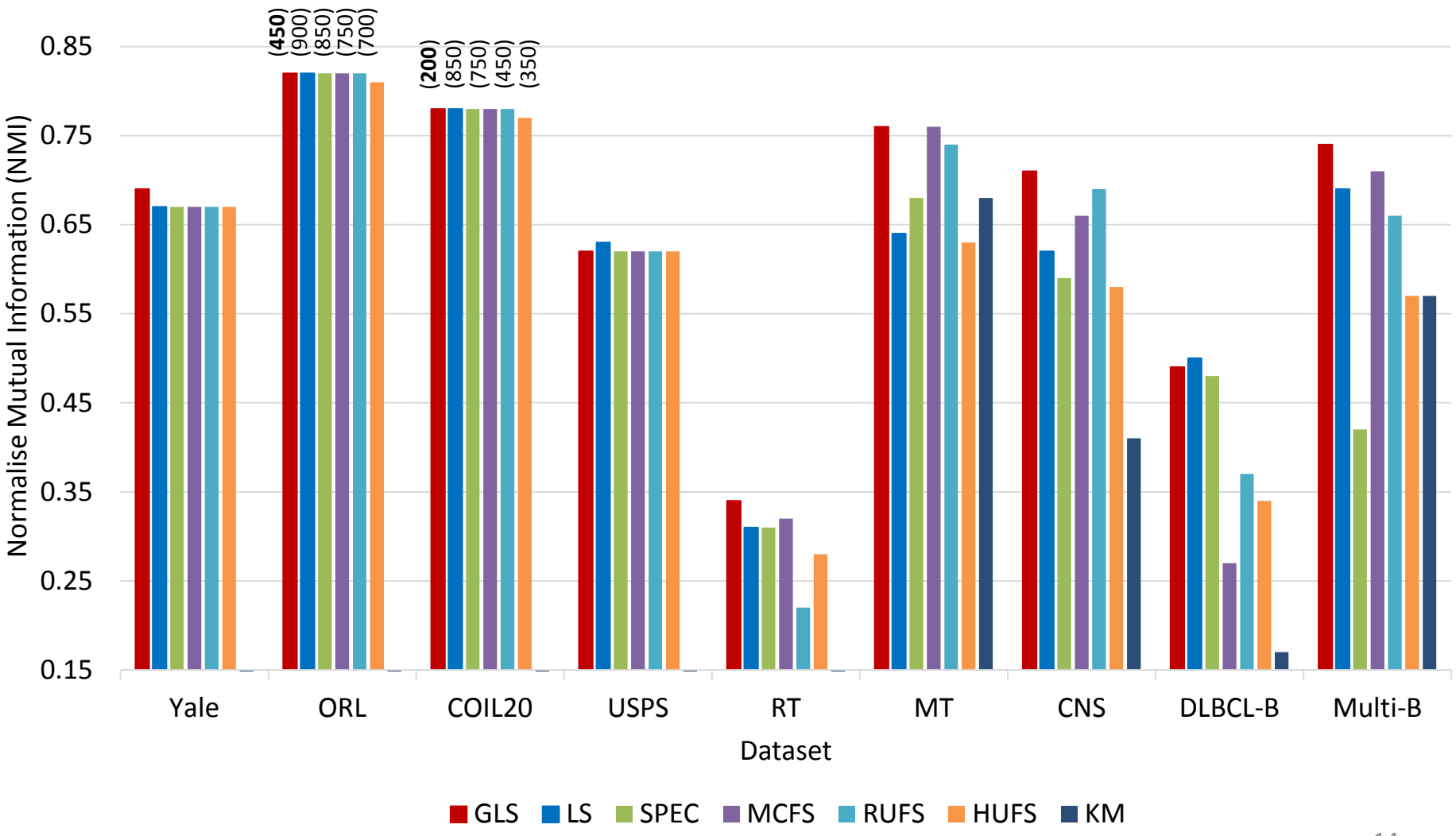
Dataset	Type	# m	# n	# c	Feature Grouping
Yale	Image	1,024	165	15	Spatial locality
ORL	Image	1,024	400	40	Spatial locality
COIL20	Image	1,024	1,440	20	Spatial locality
USPS	Image	256	9,298	10	Spatial locality
Reuters (RT)	Text	3,068	294	6	Semantics (WordNet)
Multi-tissue (MT)	Genomic	1,000	103	4	Gene ontology
CNS	Genomic	989	42	5	Gene ontology
DLBCL-B	Genomic	661	180	3	Gene ontology
Multi-B	Genomic	5,565	32	4	Gene ontology

m: No. of features, **n**: No. of instances, **c**: No. of classes

Baselines

- Laplace Score (LS)
- Spectral Feature Selection (SPEC)
- Multi Cluster Feature Selection (MCFS)
- Robust Unsupervised Feature Selection (RUFS)
- Hierarchical Unsupervised Feature Selection (HUFS)
- k-Medoid (KM)

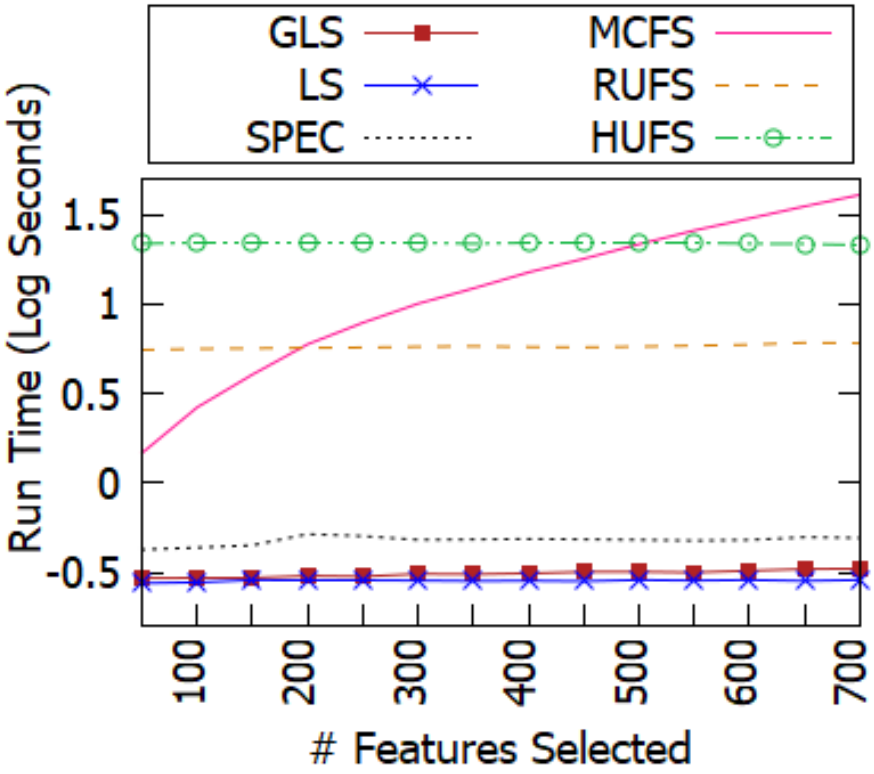
Maximum Clustering Performance



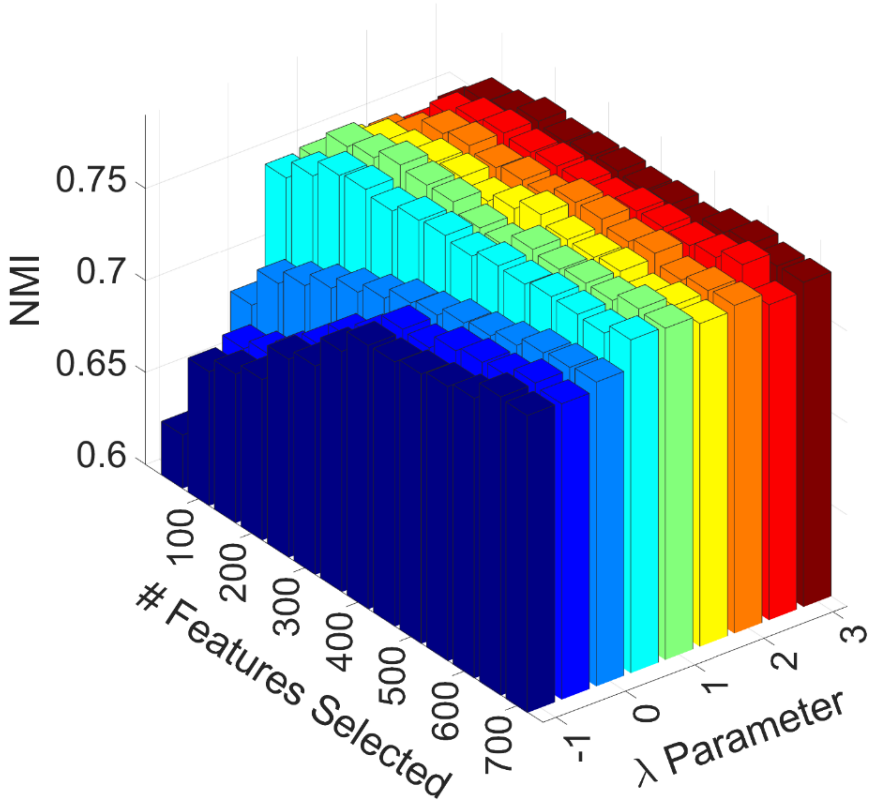
Average Accuracy Algorithm Rankings

	Yale	ORL	COIL20	USPS	RT	MT	CNS	DLBCL-B	Multi-B
GLS	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LS	3	5	5	2	2	5	5	4	3
SPEC	2	6	4	6	3	6	6	2	7
MCFS	4	3	3	3	4	2	3	3	2
RUFS	6	2	2	5	6	4	2	7	4
HUFS	5	4	6	4	5	3	4	5	6
KM	-	-	-	-	-	7	7	6	5

Runtime and Parameter Sensitivity Results



Run time variation (COIL20)



Accuracy variation with λ (COIL20)

Conclusion and Future Work

Conclusion

- A framework which facilitates unsupervised feature selection methods to exploit feature group information as an external source of information

- A concrete implementation using Laplace Score method

Future Work

- Experimenting the proposed framework for other unsupervised feature selection methods

Thank you!