Supervised Rank aggregation (SRA): A novel rank aggregation approach for ensemble-based feature selection

Rahi Jain, Wei Xu

Intro

- A high dimensional data has challenges associated with:
  - model fitting
  - generalizability
  - computation complexity

- Feature selection is an important component in high dimensional data analysis

Intro

Feature selection approaches:

- base – use a **single** feature selection technique
- hybrid – uses a **sequence of multiple** feature selection techniques
- ensemble
  - **multiple models** are created from the same dataset
  - **performance** of features from these models is **pooled** and **ranked** → rank aggregation
  - **rank aggregation** based on mean, median or robust rank
  - relevant features are selected based on the cut-off of importance

Intro

The study introduces:

- **R**ank **A**ggregation approach using th **S**upervised learning (ML) → **SRA**

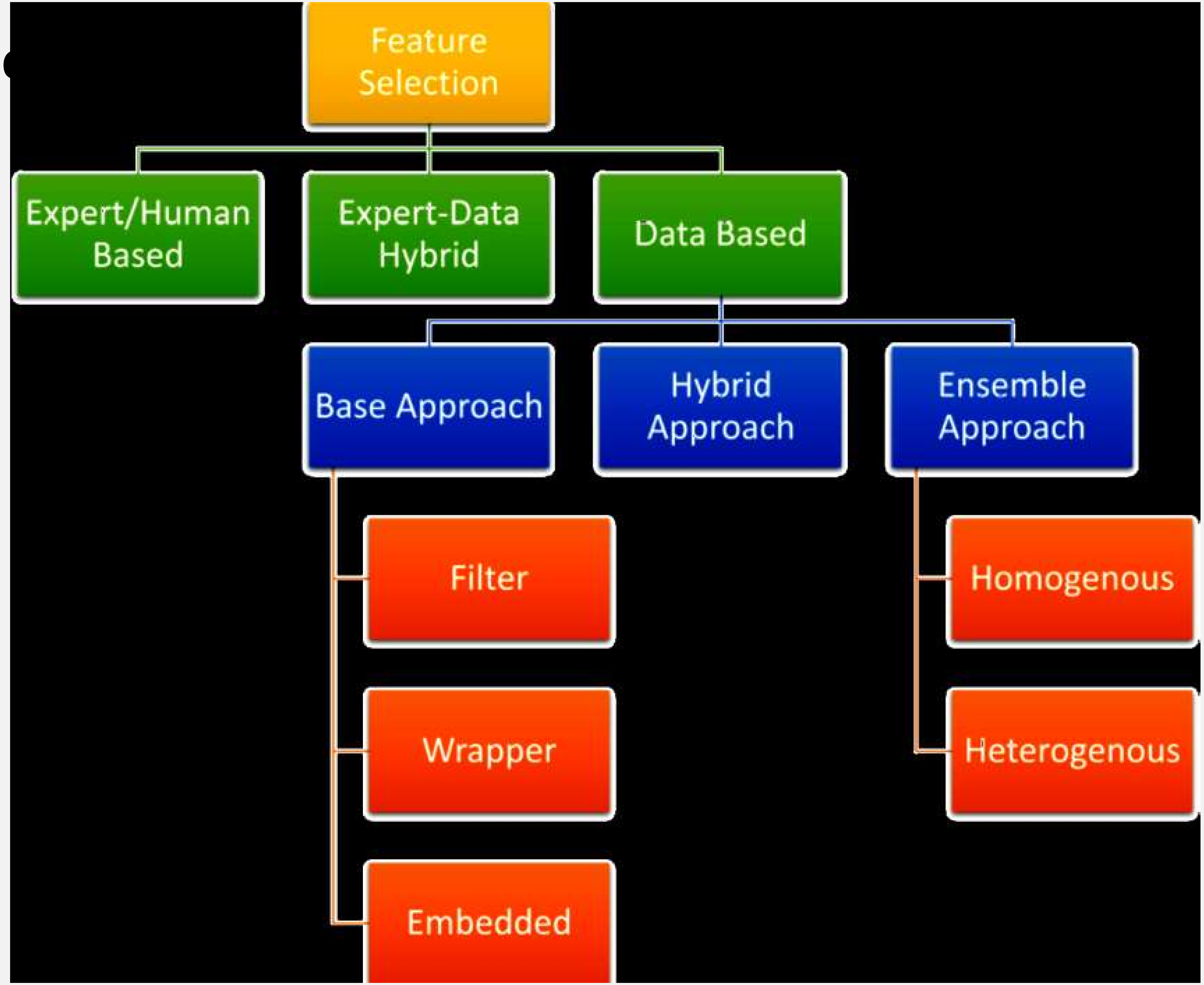
  1. building a **performance matrix** = performance of all features in all the models
  2. scoring a **performance of each model**
  3. "Supervised learning is used to find the relative rank or performance of features based on their potential to help achieve the best performance in the final data analysis."

Intro

Feature selection ensemble approaches:

- homogenous ensemble – **multiple datasets** created from the same data by sub-setting the samples / features / both
- heterogeneous ensemble – **single dataset** is modeled using different techniques

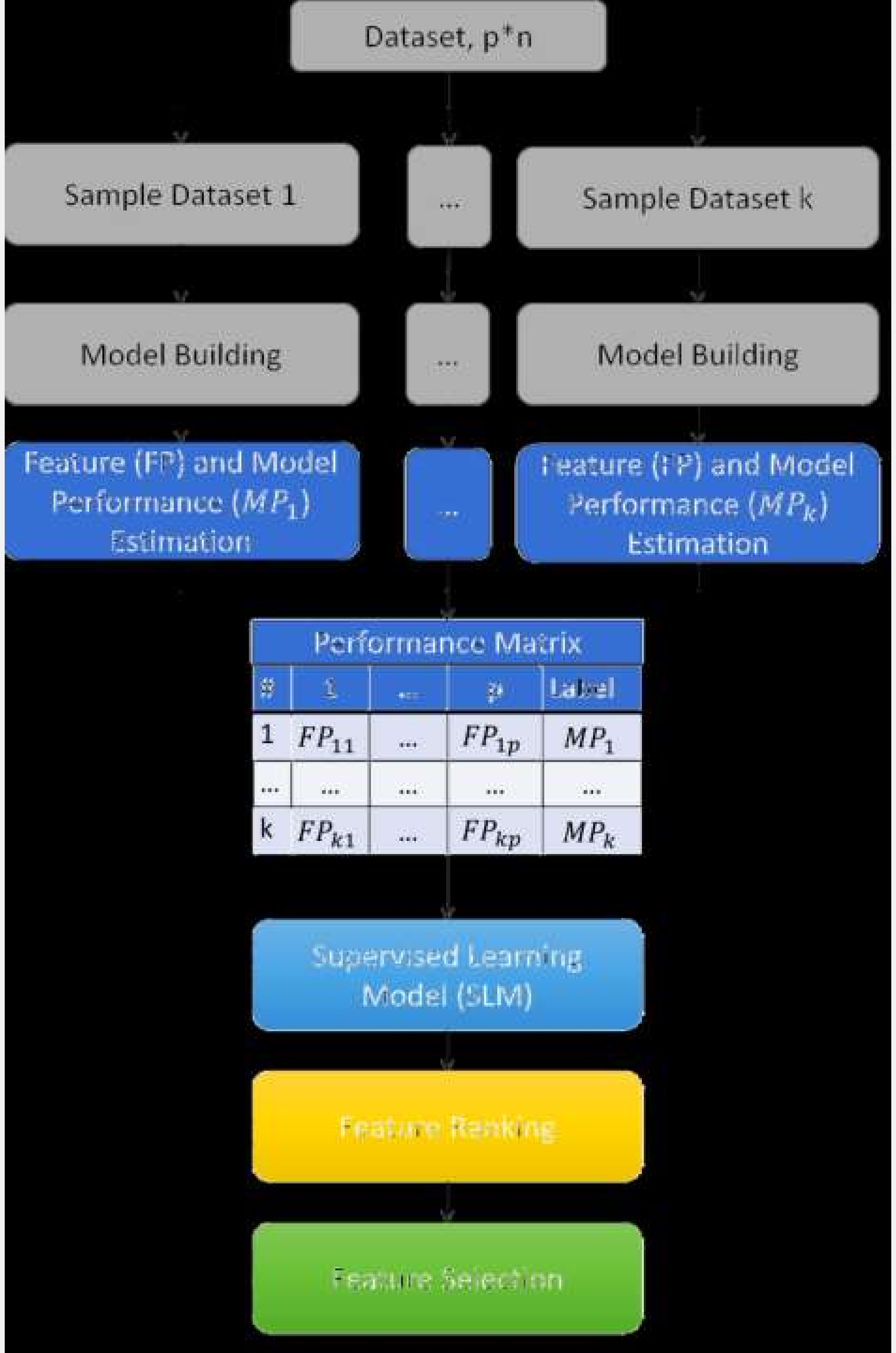
Feature sel


Feature Selection
Expert/Human Based
Expert-Data Hybrid
Data Based
Base Approach
Hybrid Approach
Ensemble Approach
Filter
Wrapper
Embedded
Homogenous
Heterogenous

Method

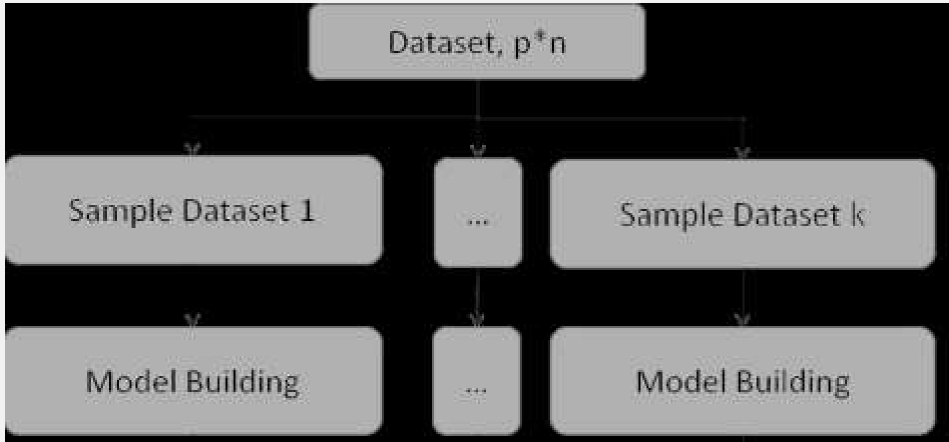
- n – sample size
- p – number of features (explanatory variables)
- homogenous ensemble – multiple bootstrap data sets
- performance matrix = feature performance and model performance from each bootstrap data set
- supervised learning algorithm (SRA) trained on the performance matrix
- final feature ranking = feature importance from SRA
- final set of feature = based on an importance cut-off

Method


Dataset, p*n
Sample Dataset 1
...
Sample Dataset k
Model Building
...
Model Building
Feature (FP) and Model Performance (MP_1) Estimation
...
Feature (FP) and Model Performance (MP_k) Estimation
Performance Matrix
1 ... p Label
1 FP_{11} ... FP_{1p} MP_1
... ...
k FP_{k1} ... FP_{kp} MP_k
Supervised Learning Model (SLM)
Feature Ranking
Feature Selection

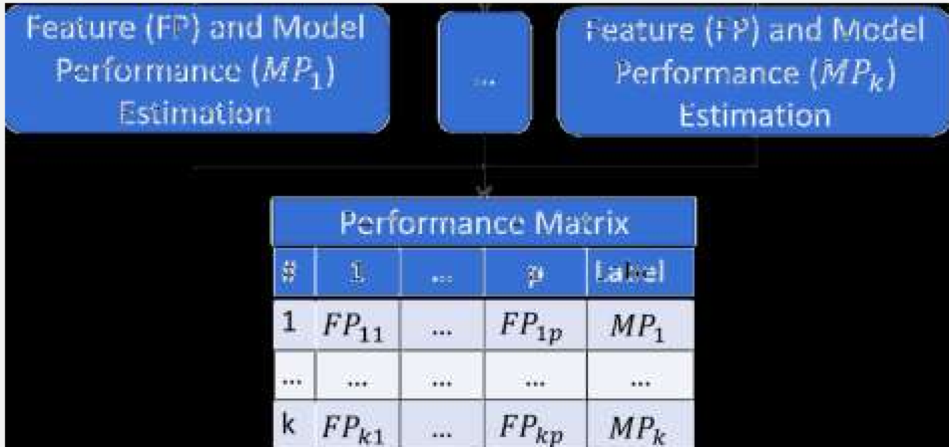
Method – sample preparation and modelling

- **k** sample data sets of size **n** – sampling with replacement from the original data set
- each data set has **q** features – sampling randomly from original **p** features
- Ridge regression – a model for each data set with **q** features



Method – performance matrix

- **k** sample data sets x **p+1** → **p** features + **1** model fit
- Performance matrix **k** x **q+1**
- MP = model performance → RMSE^{-1} (root mean square error)
- FP = feature performance → effect estimate (???)



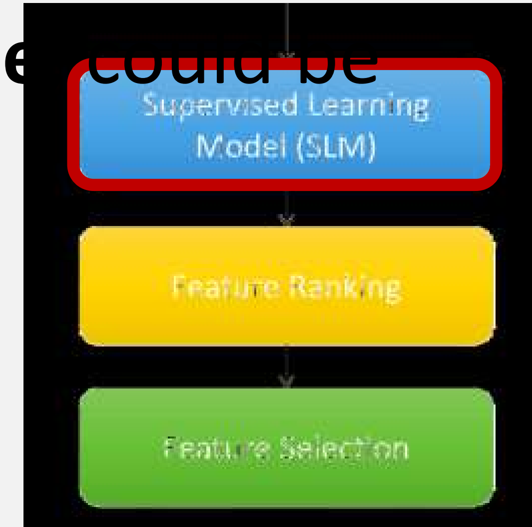
Method – sra

- supervised learning model
- created from the performance matrix

Performance Matrix

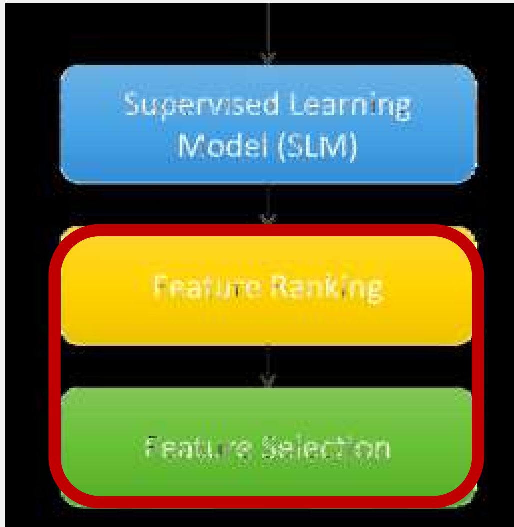
#	1	...	p	label
1	FP_{11}	...	FP_{1p}	MP_1
...	...	...	...	...
k	FP_{k1}	...	FP_{kp}	MP_k

- $\boldsymbol{MP} = g\left(\sum_{i=1}^{p} FP_i\right)$
- $g\left(\sum_{i=1}^{p} FP_i\right)$ → “determined by ML technique”
- “Currently, only ML techniques like **penalized regression** and **decision trees** which could **provide feature importance** could be used.”
- Why only these two ^ ???



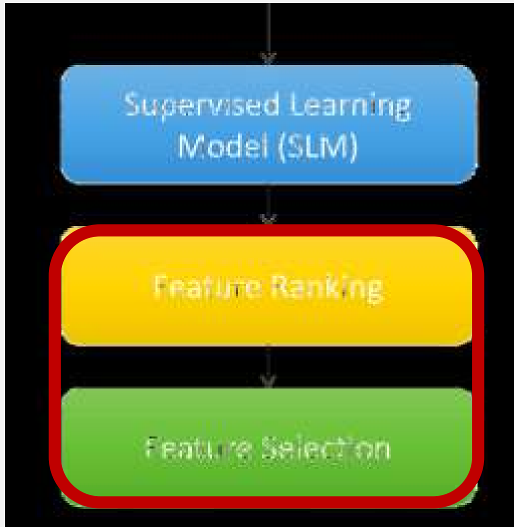
Method – feature selection

- importance for each feature estmated by $\boldsymbol{MP} = g\left(\sum_{i=1}^{p} FP_i\right)$ is used to **select target features**
- features with more importance = **target features** = most relevant in achieving high model performance
- goal → estimate threshold for **target features** along their ranking
  - predefined threshold
  - rule-based threshold estimation
  - unsupervised learning based threshold estimation



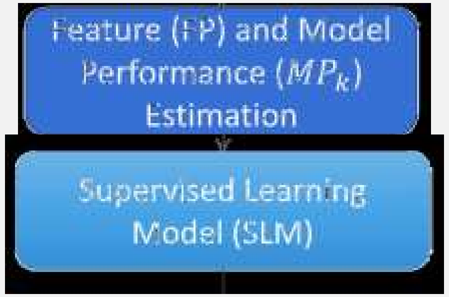
Method – threshold estimation

- unsupervised learning based threshold estimation
- 1D K-means on importance of FP_i obtained from $\boldsymbol{MP} = g\left(\sum_{i=1}^{p} FP_i\right)$
- FP_i clustered into two groups
  - — cluster with a higher mean = important features
  - — cluster with a lower mean = unimportant features



Simulation

- a linear model $\boldsymbol{y} = \boldsymbol{b}_0 + \sum_{i=1}^{p} \boldsymbol{b}_i \boldsymbol{x}_i + \boldsymbol{e}$
- Simulated covariance between $\boldsymbol{x}$
- Models for particular feature set – Ridge regression
- Models for $\boldsymbol{MP} = g\left(\sum_{i=1}^{p} FP_i\right)$
  - SRA-Lasso
  - SRA-Ridge
  - SRA-RF
  - Each model «using optimized hyperparameter values»



Simulation - scenarios

Scenario	β (Non-Zero coefficients)	p	Sample Size		σ	k
			Train (n)	Test		
A	$\{\beta_i \mid i = \{1, \ldots, 10\}\} = \{0.9, \ldots, 0.9\}$	75	100	500	0.25	300
B	$\{\beta_i \mid i = \{1, \ldots, 10\}\} = \{0.5, \ldots, 0.5\}$	100	100	500	0.25	100
C	$\{\beta_i \mid i = \{1, \ldots, 15\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, 0.4\}$	175	275	500	0.25	100
D	$\{\beta_i \mid i = \{1, \ldots, 15\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, 0.4\}$	75	275	500	0.25	100
E	$\{\beta_i \mid i = \{1, \ldots, 15\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, 0.4\}$	75	225	500	0.25	200
F	$\{\beta_i \mid i = \{1, \ldots, 20\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, -0.8\}$	125	225	500	0.25	200

Results – selection of target features (features with nonzero effect)

RA technique		Scenarios					
		A	*B*	*C*	*D*	*E*	*F*
		Target Features (%) [μ(95%CI)]					
Existing	CVRA	100 (100-100)	100 (100-100)	46 (45-47)	46 (45-47)	47 (47-47)	51 (48-53)
	MARA	100 (100-100)	100 (100-100)	87 (81-93)	97 (95-99)	85 (78-93)	66 (56-76)
	MeRA	100 (100-100)	100 (100-100)	47 (47-47)	47 (47-47)	47 (47-47)	53 (51-55)
	MedRA	100 (100-100)	100 (100-100)	47 (47-47)	47 (46-49)	47 (47-47)	53 (50-56)
	MIRA	62 (48-76)	94 (89-99)	95 (91-98)	77 (72-82)	85 (82-89)	88 (86-89)
	RRA	99 (97-100)	99 (97-101)	47 (47-47)	48 (46-50)	47 (46-48)	52 (50-54)
	SDRA	78 (67-89)	71 (67-75)	34 (29-39)	40 (35-45)	35 (27-42)	39 (32-46)
	tRA	100 (100-100)	100 (100-100)	46 (45-47)	45 (44-47)	47 (47-47)	51 (48-53)
	WRA	100 (100-100)	100 (100-100)	49 (45-53)	58 (58-58)	58 (58-58)	37 (34-40)
SRA	Lasso	92 (87-97)	37 (18-56)	41 (37-45)	58 (51-65)	63 (57-69)	46 (38-54)
	RF	98 (95-100)	53 (43-63)	63 (54-73)	67 (57-76)	65 (56-75)	61 (53-69)
	Ridge	100 (100-100)	99 (97-100)	95 (89-100)	95 (92-99)	100 (100-100)	92 (88-96)

Results – F1

RA technique		Scenarios					
RA technique		*A*	*B*	*C*	*D*	*E*	*F*
RA technique		F1 Score[μ(95%CI)]					
Existing	CVRA	1.00 (1.00-1.00)	0.93 (0.89-0.97)	0.63 (0.62-0.64)	0.63 (0.62-0.64)	0.64 (0.64-0.64)	0.67 (0.65-0.69)
	MARA	0.58 (0.55-0.61)	0.70 (0.68-0.72)	0.60 (0.55-0.64)	0.83 (0.8-0.86)	0.65 (0.61-0.69)	0.44 (0.39-0.49)
	MeRA	0.83 (0.8-0.86)	0.81 (0.80-0.82)	0.64 (0.64-0.64)	0.64 (0.64-0.64)	0.64 (0.64-0.64)	0.69 (0.67-0.71)
	ModRA	0.85 (0.82-0.87)	0.81 (0.80-0.82)	0.64 (0.64-0.64)	0.64 (0.63-0.65)	0.64 (0.64-0.64)	0.69 (0.67-0.71)
	MIRA	0.41 (0.33-0.49)	0.60 (0.53-0.67)	0.79 (0.73-0.85)	0.75 (0.72-0.78)	0.70 (0.67-0.73)	0.76 (0.74-0.79)
	RRA	0.90 (0.87-0.93)	0.82 (0.80-0.83)	0.64 (0.64-0.64)	0.65 (0.63-0.66)	0.64 (0.63-0.65)	0.68 (0.67-0.7)
	SDRA	0.30 (0.27-0.32)	0.22 (0.20-0.24)	0.07 (0.06-0.07)	0.16 (0.15-0.18)	0.16 (0.14-0.17)	0.12 (0.10-0.14)
	tRA	1.00 (1.00-1.00)	0.92 (0.88-0.96)	0.63 (0.62-0.64)	0.62 (0.61-0.64)	0.64 (0.64-0.64)	0.67 (0.65-0.69)
	WRA	0.40 (0.38-0.42)	0.33 (0.32-0.34)	0.14 (0.13-0.16)	0.29 (0.27-0.3)	0.30 (0.28-0.32)	0.18 (0.17-0.2)
[illegible]	Lasso	0.96 (0.93-0.98)	0.33 (0.16-0.50)	0.58 (0.53-0.62)	0.73 (0.67-0.79)	0.77 (0.72-0.81)	0.62 (0.35-0.7)
	RF	0.75 (0.69-0.81)	0.29 (0.24-0.33)	0.64 (0.57-0.71)	0.73 (0.66-0.8)	0.77 (0.71-0.83)	0.70 (0.64-0.76)
	Ridge	1.00 (1.00-1.00)	0.99 (0.98-1.00)	0.97 (0.94-1.00)	0.98 (0.96-0.99)	1.00 (1.00-1.00)	0.95 (0.93-0.98)

Results – predictive performance

RA technique		Scenarios					
		A	*B*	*C*	*D*	*E*	*F*
RA technique		Predictive Performance (1/RMSE) [μ(95%CI)]					
Existing	CVRA	3.50 (3.29-3.71)	3.82 (2.90-4.75)	0.81 (0.79-0.84)	0.84 (0.81-0.86)	0.83 (0.80-0.85)	0.75 (0.72-0.77)
	MARA	2.67 (2.48-2.90)	3.65 (2.77-4.54)	1.73 (1.43-2.03)	2.42 (1.97-2.86)	1.43 (1.01-1.85)	0.58 (0.51-0.65)
	MeRA	2.94 (2.56-3.31)	3.67 (2.83-4.51)	0.82 (0.80-0.84)	0.84 (0.82-0.87)	0.83 (0.80-0.85)	0.76 (0.73-0.79)
	MedRA	2.96 (2.55-3.36)	3.67 (2.83-4.51)	0.82 (0.80-0.84)	0.85 (0.82-0.89)	0.83 (0.80-0.85)	0.76 (0.73-0.79)
	MIRA	0.80 (0.27-1.34)	2.58 (2.00-3.17)	2.45 (1.97-2.93)	1.45 (1.29-1.61)	1.74 (1.57-1.91)	1.31 (1.25-1.37)
	RRA	2.92 (2.42-3.42)	3.54 (2.61-4.46)	0.82 (0.80-0.84)	0.87 (0.83-0.91)	0.84 (0.81-0.87)	0.75 (0.73-0.77)
	SDRA	1.10 (0.53-1.68)	1.03 (0.96-1.10)	0.68 (0.66-0.7)	0.77 (0.74-0.8)	0.71 (0.66-0.76)	0.53 (0.47-0.58)
	tRA	3.50 (3.29-3.71)	3.79 (2.91-4.67)	0.81 (0.79-0.83)	0.84 (0.81-0.86)	0.83 (0.80-0.85)	0.75 (0.72-0.77)
	WRA	2.18 (1.96-2.39)	2.62 (2.39-2.85)	0.45 (0.44-0.45)	0.47 (0.45-0.48)	0.46 (0.45-0.46)	0.37 (0.36-0.38)
SRA	Lasso	2.17 (1.32-3.02)	0.77 (0.51-1.03)	0.79 (0.76-0.82)	1.00 (0.85-1.16)	1.09 (0.96-1.21)	0.72 (0.66-0.79)
	RF	2.62 (2.07-3.17)	0.88 (0.77-0.99)	0.70 (0.62-0.78)	0.87 (0.46-1.27)	0.73 (0.55-0.9)	0.55 (0.50-0.59)
	Ridge	3.50 (3.29-3.71)	3.83 (2.91-4.75)	2.58 (2.07-3.08)	2.65 (2.02-3.28)	2.98 (2.51-3.44)	1.87 (1.46-2.28)

Key points

- Supervised Rank Aggregation methods are better than rule-based rank aggregation methods for ensemble-based feature selection
- ???
- SRA Ridge could give much better discrimination between true and noise features as well as predictive performance than rule-based rank aggregation methods
- SRA could be useful in detecting the genomic features like methylation sites which could have biological relevance