kindle direct publishing

Tushaar Gangavarapu

NLP Research Engineer
Kindle Content Experience
Amazon.com, Inc.

Data Mining – January 06, 2020

Let's Get Greedy and Genetically Ensemble the Feature Space

This work was completed at the Dept. of Information Technology, NITK Surathkal, under the guidance of Dr. Nagamma Patil

kindle | direct publishing

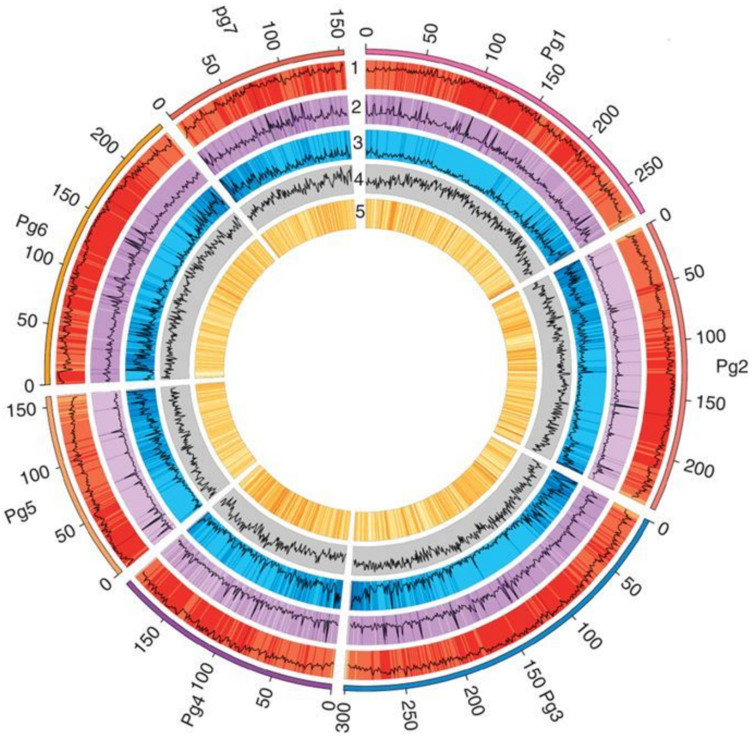
whoami

tusgan@amazon.com

Agenda

1. Feature Space and Information
2. Research Gaps in Feature Selection?
3. Greedy Hybrid Ensemble with GA
4. Intuition and Penalization
5. Conclusions and Future Directions

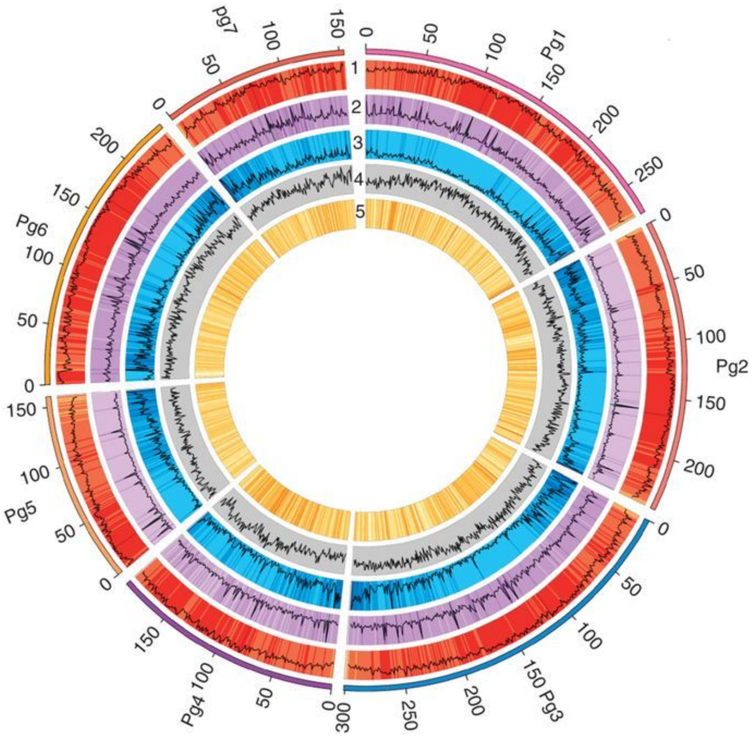
Feature Selection: Basic Idea



- Feature selection aims at facilitating how a **subset of available dimensions** can be selected
  - **Sparsity** (noise) and high-dimensionality

RK Varshney. *Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments.* Nature. 2017

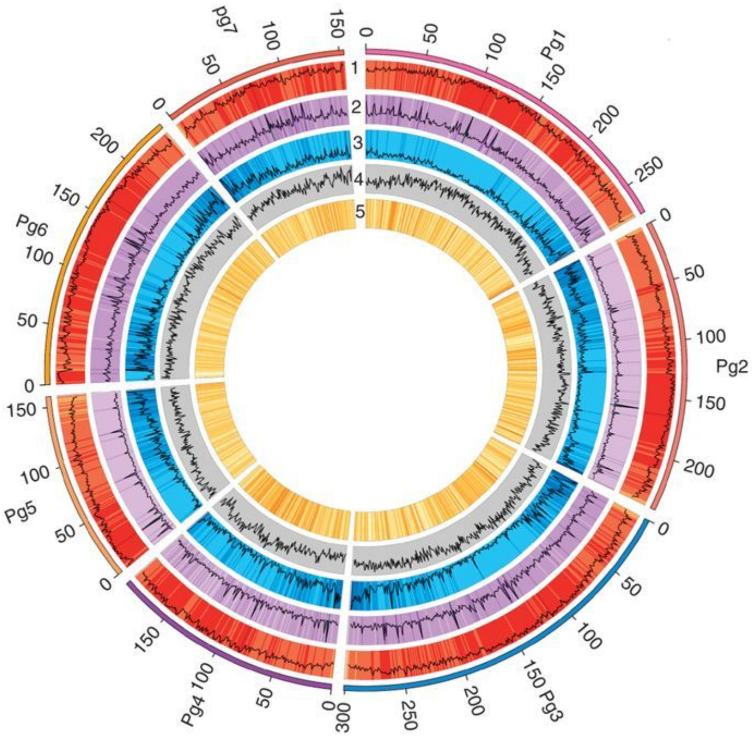
Feature Selection: Basic Idea



- Feature selection aims at facilitating how a **subset of available dimensions** can be selected
  - **Sparsity** (noise) and high-dimensionality
  - Sparse matrices often **mislead the underlying machine learners**

RK Varshney. *Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments.* Nature. 2017

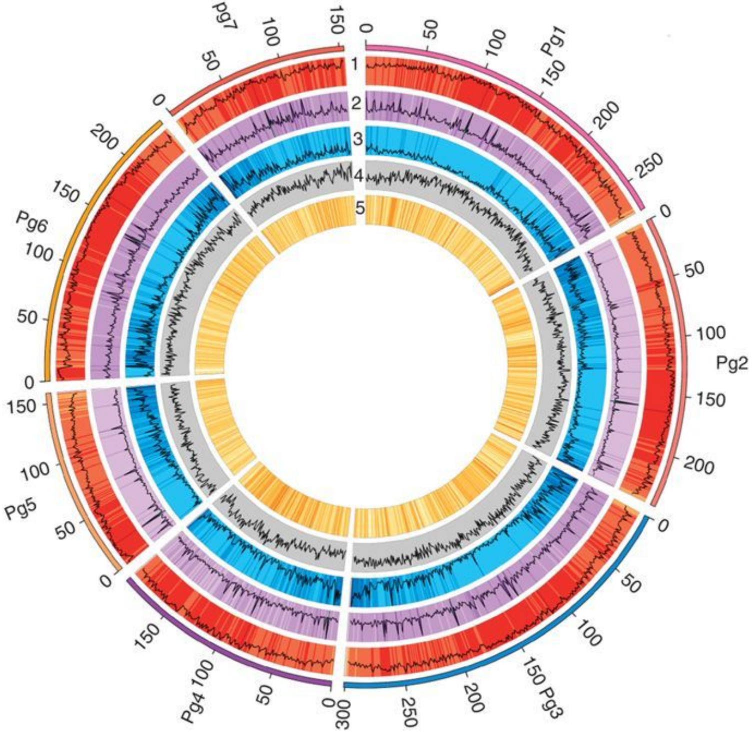
Feature Selection: Basic Idea



- Feature selection aims at facilitating how a **subset of available dimensions** can be selected
  - **Sparsity** (noise) and high-dimensionality
  - Sparse matrices often **mislead the underlying machine learners**
- **Feature extraction and engineering** *vs*. **feature selection** *vs*. **feature extraction**

RK Varshney. *Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments.* Nature. 2017

Feature Selection: Basic Idea



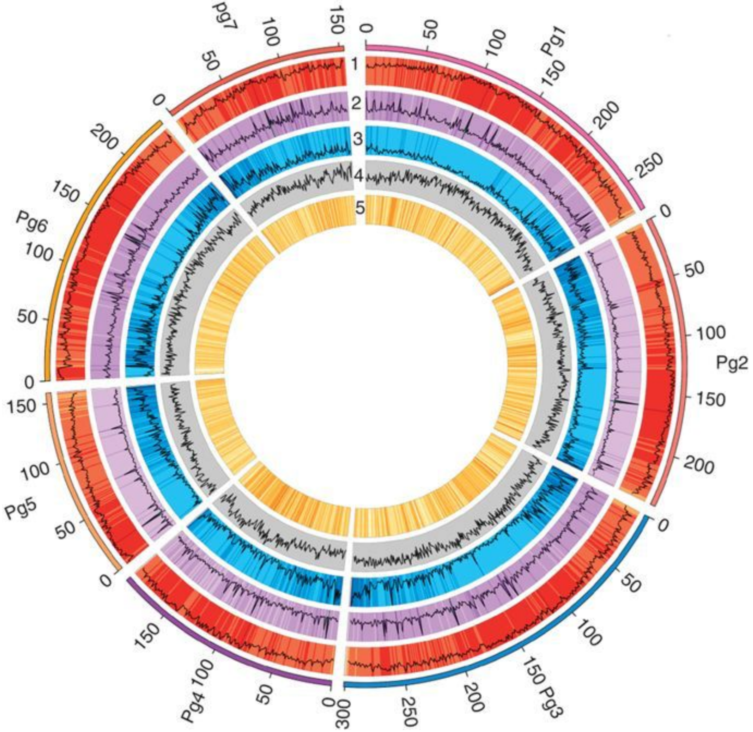
- Feature selection aims at facilitating how a **subset of available dimensions** can be selected
  - **Sparsity** (noise) and high-dimensionality
  - Sparse matrices often **mislead the underlying machine learners**

- **Feature extraction and engineering** *vs*. **feature selection** *vs*. **feature extraction**

- **Feature selection variants**: filter, wrapper, embedded, and hybrid approaches

RK Varshney. *Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments.* Nature. 2017

Feature Selection: Basic Idea

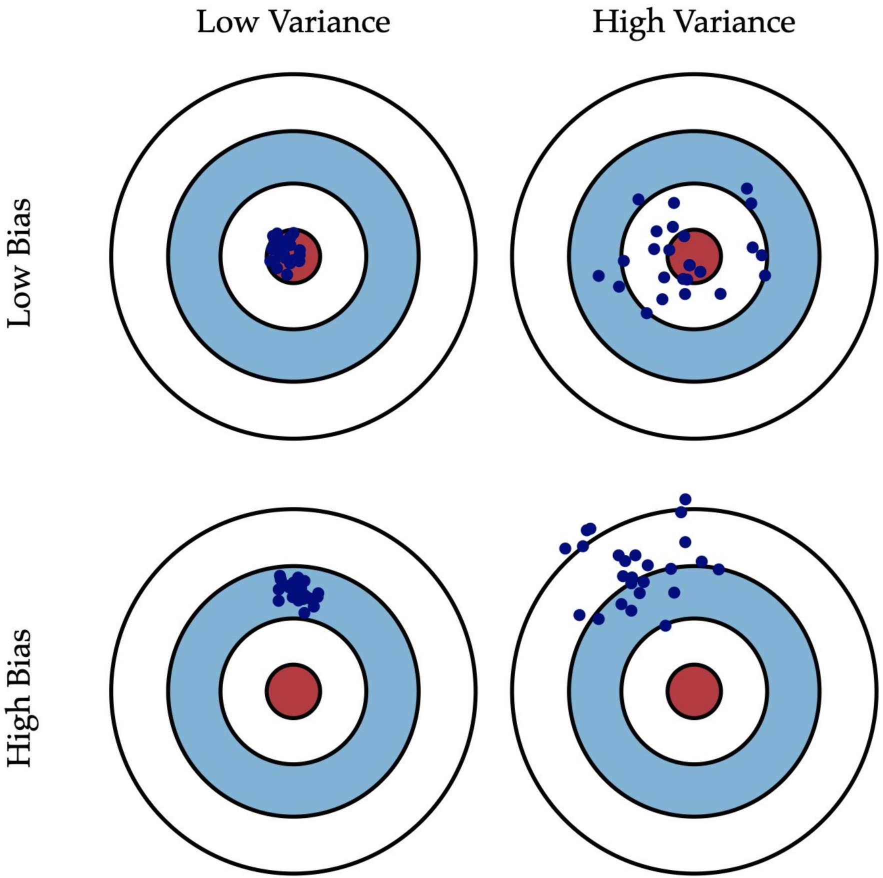


- Feature selection aims at facilitating how a **subset of available dimensions** can be selected
  - **Sparsity** (noise) and high-dimensionality
  - Sparse matrices often **mislead the underlying machine learners**
- **Feature extraction and engineering** *vs*. **feature selection** *vs*. **feature extraction**
- **Feature selection variants**: filter, wrapper, embedded, and hybrid approaches
- Bias-variance tradeoff – **bias = assumptions made by classifier**; **variance = training data variations**

RK Varshney. *Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments.* Nature. 2017

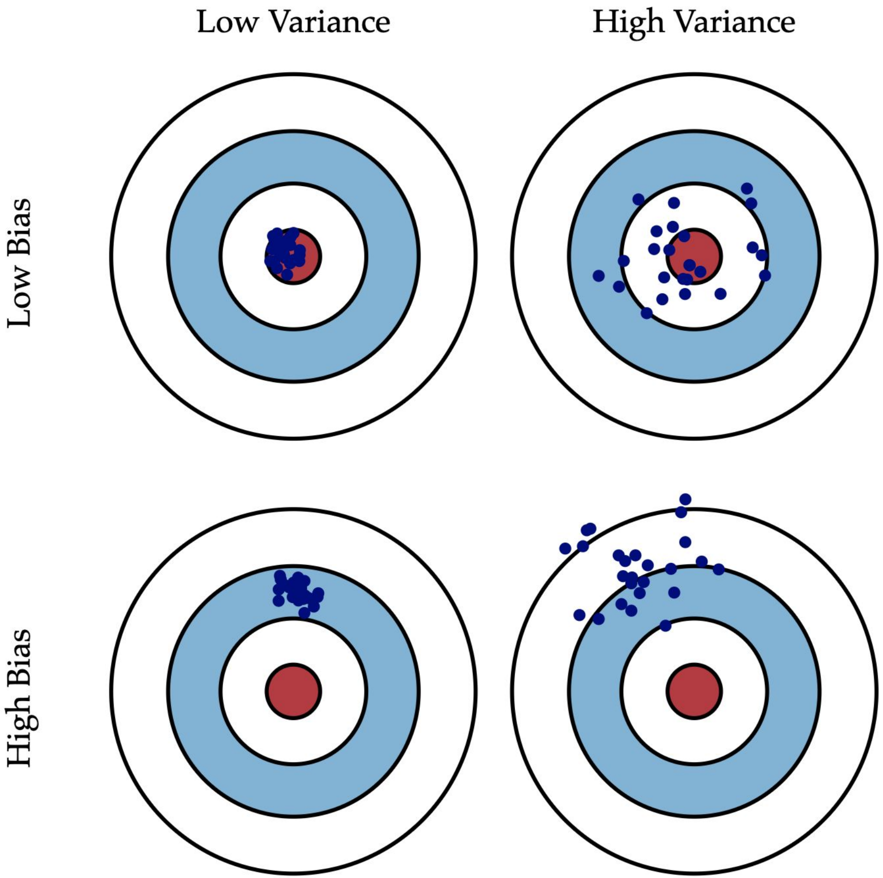
Bias-Variance Tradeoff

- **Bias**: Occurs when the classifier has limited flexibility to learn the ground truth – number of samples
  - **Low bias**: more samples



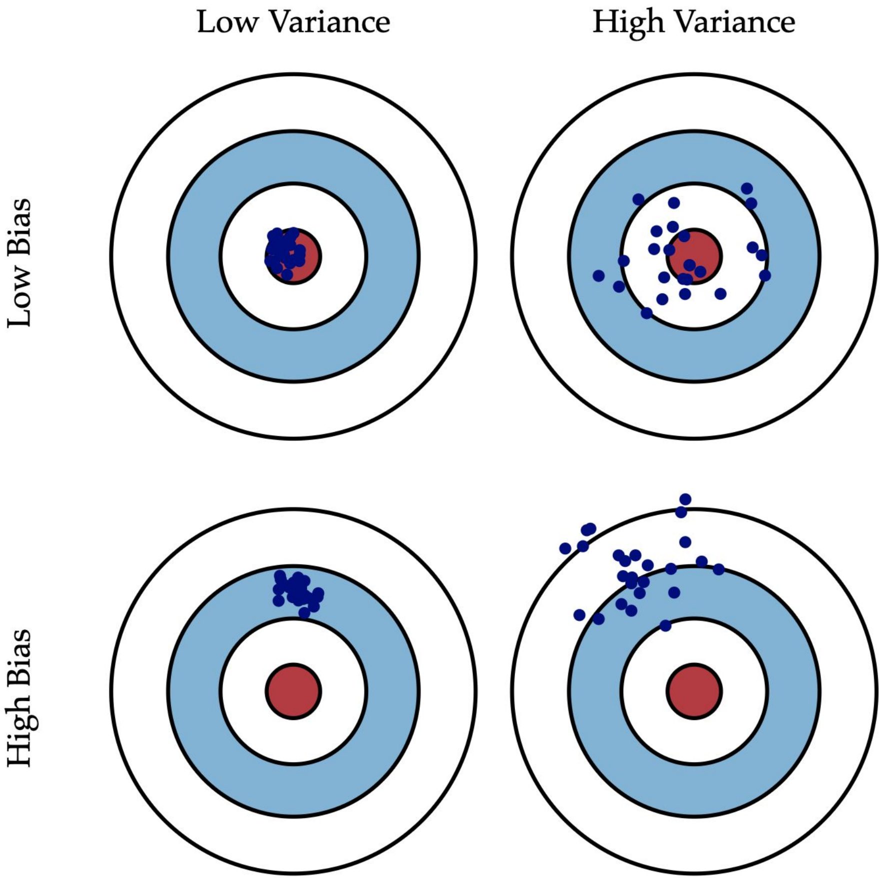
Bias-Variance Tradeoff

- **Bias**: Occurs when the classifier has limited flexibility to learn the ground truth – number of samples
  - **Low bias**: more samples
- **Variance**: Sensitivity of the classifier to the sets of training data – number of features
  - **Low variance**: less features



Bias-Variance Tradeoff

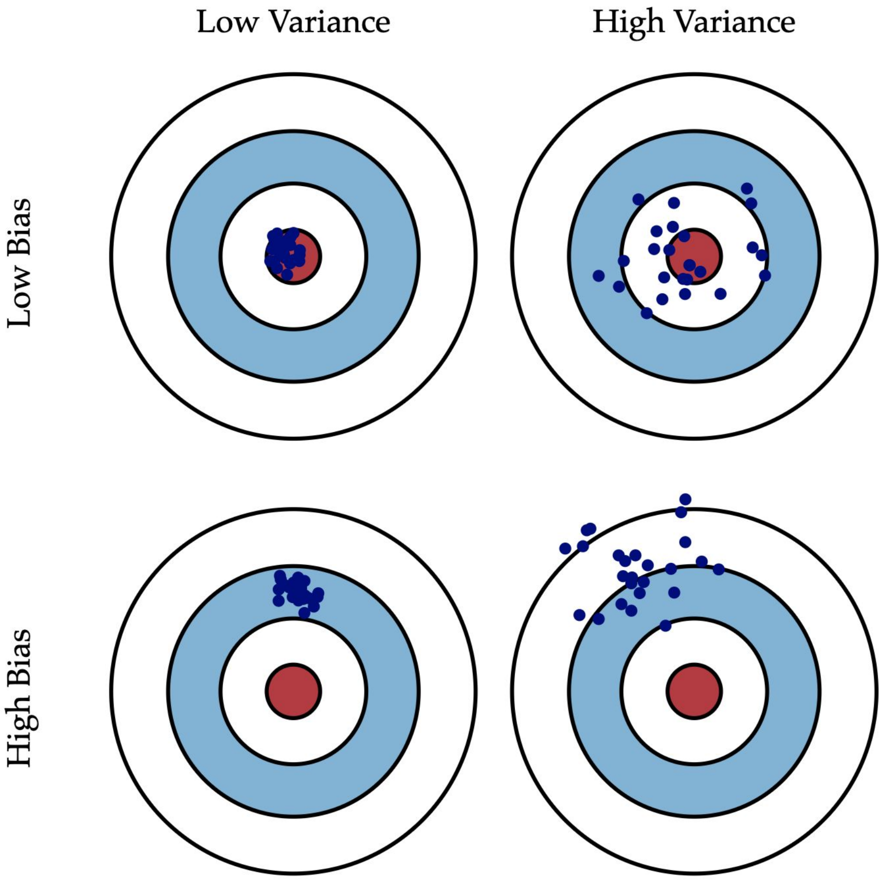
- **Bias**: Occurs when the classifier has limited flexibility to learn the ground truth – number of samples
  - **Low bias**: more samples
- **Variance**: Sensitivity of the classifier to the sets of training data – number of features
  - **Low variance**: less features
- **Total error**[Hastie 2009]: Bias^2 + variance + **irreducible error**



[Hastie 2009] T Hastie et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2009.

Bias-Variance Tradeoff

- **Bias**: Occurs when the classifier has limited flexibility to learn the ground truth – number of samples
  - **Low bias**: more samples
- **Variance**: Sensitivity of the classifier to the sets of training data – number of features
  - **Low variance**: less features
- **Total error**[Hastie 2009]: Bias^2 + variance + **irreducible error**
- **Ideal**: features-to-samples ratio ≪ 1
  Observed: features-to-samples ratio ≧ 1



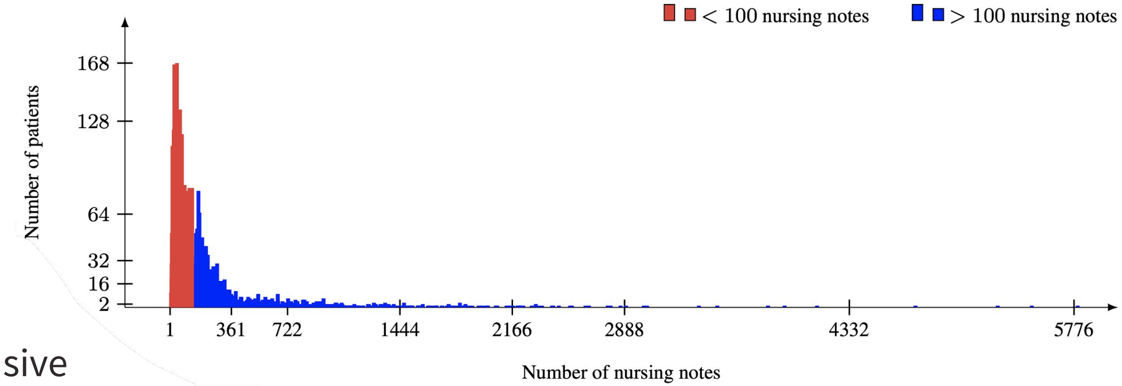
[Hastie 2009] T Hastie et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2009.

Agenda

1. Feature Space and Information
2. Research Gaps in Feature Selection?
3. Greedy Hybrid Ensemble with GA
4. Intuition and Penalization
5. Conclusions and Future Directions

Feature Space and Information Extraction

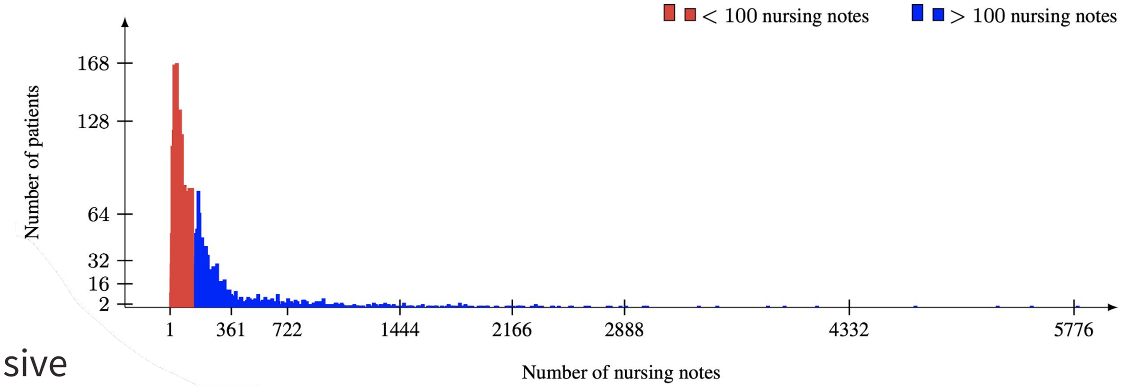
- Does **more information** lead to more informed decision making?
  - **Irrelevancy**: learnability?
  - **Redundancy**: training ?
  - **Noise**: classification errors
  - **Computational cost**: expensive



T Gangavarapu et al. *Coherence-based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification.* CoNLL. 2019.

Feature Space and Information Extraction

- Does **more information** lead to more informed decision making?
  - **Irrelevancy**: learnability?
  - **Redundancy**: training ?
  - **Noise**: classification errors
  - **Computational cost**: expensive

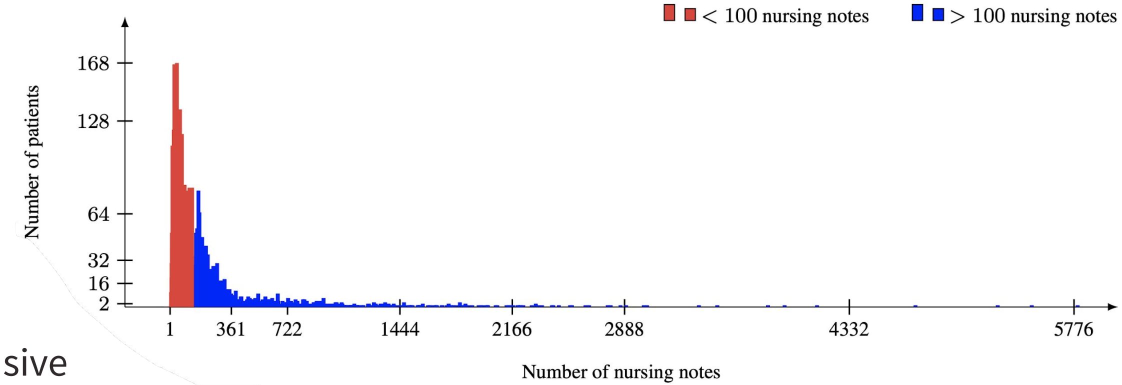


- How to **choose a feature selection approach** for the given data? – need to match with the problem structure and mine for inherent patterns in the data
  - **Intuition-based**: unreliable approach

T Gangavarapu et al. *Coherence-based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification*. CoNLL. 2019.

Feature Space and Information Extraction

- Does **more information** lead to more informed decision making?
  - **Irrelevancy**: learnability?
  - **Redundancy**: training ?
  - **Noise**: classification errors
  - **Computational cost**: expensive

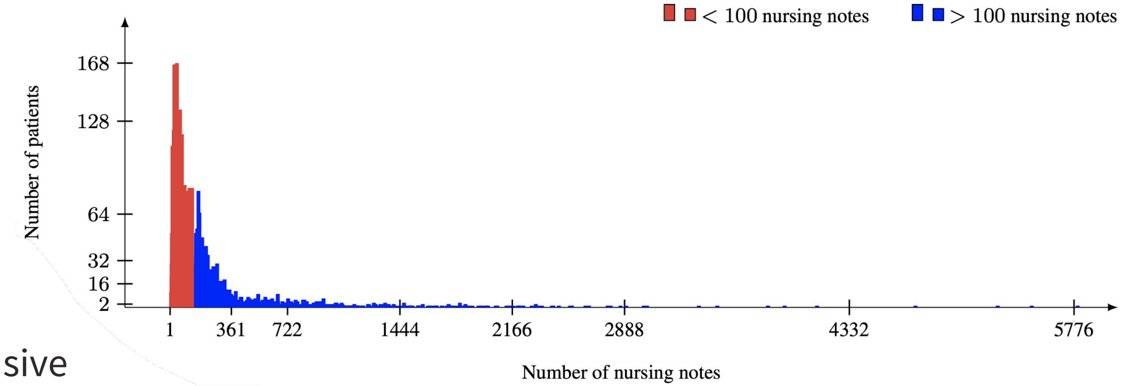


- How to **choose a feature selection approach** for the given data? – need to match with the problem structure and mine for inherent patterns in the data
  - **Intuition-based**: unreliable approach
  - **Exhaustive search**: infeasible

T Gangavarapu et al. *Coherence-based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification.* CoNLL. 2019.

Feature Space and Information Extraction

- Does **more information** lead to more informed decision making?
  - **Irrelevancy**: learnability?
  - **Redundancy**: training ?
  - **Noise**: classification errors
  - **Computational cost**: expensive



- How to **choose a feature selection approach** for the given data? – need to match with the problem structure and mine for inherent patterns in the data
  - **Intuition-based**: unreliable approach
  - **Exhaustive search**: infeasible
  - **Determine heuristically**: issue of convergence

T Gangavarapu et al. *Coherence-based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification.* CoNLL. 2019.

Agenda

1. Feature Space and Information
2. Research Gaps in Feature Selection
3. Greedy Hybrid Ensemble with GA
4. Intuition and Penalization
5. Conclusions and Future Directions

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**
- **Filter-based approaches**: faster computation, but heavy dependence on correlation and classifier independence limits their accuracy

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**
- **Filter-based approaches**: faster computation, but heavy dependence on correlation and classifier independence limits their accuracy
- **Wrapper-based, embedded, and hybrid approaches**: domain adaptability and high computational cost of training, but reliable performance

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**
- **Filter-based approaches**: faster computation, but heavy dependence on correlation and classifier independence limits their accuracy
- **Wrapper-based, embedded, and hybrid approaches**: domain adaptability and high computational cost of training, but reliable performance
- **Metaheuristic search approaches**: population-based mechanism guides the search, but convergence problem and correlation-unguided search can be a bottleneck!

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**
- **Filter-based approaches**: faster computation, but heavy dependence on correlation and classifier independence limits their accuracy
- **Wrapper-based, embedded, and hybrid approaches**: domain adaptability and high computational cost of training, but reliable performance
- **Metaheuristic search approaches**: population-based mechanism guides the search, but convergence problem and correlation-unguided search can be a bottleneck!
- **Need for an ensemble**: use a set of predetermined feature selection approaches
    - **Voting-based ensemble**: simply a brute force ensemble

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**
- **Filter-based approaches**: faster computation, but heavy dependence on correlation and classifier independence limits their accuracy
- **Wrapper-based, embedded, and hybrid approaches**: domain adaptability and high computational cost of training, but reliable performance
- **Metaheuristic search approaches**: population-based mechanism guides the search, but convergence problem and correlation-unguided search can be a bottleneck!
- **Need for an ensemble**: use a set of predetermined feature selection approaches
  - **Voting-based ensemble**: simply a brute force ensemble
  - **Greedy ensemble**: penalize bad-performing selection methods and their features

Research Gaps in Feature Selection?

- **Which feature selection to use**: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? – **always an issue!**
- **Filter-based approaches**: faster computation, but heavy dependence on correlation and classifier independence limits their accuracy
- **Wrapper-based, embedded, and hybrid approaches**: domain adaptability and high computational cost of training, but reliable performance
- **Metaheuristic search approaches**: population-based mechanism guides the search, but convergence problem and correlation-unguided search can be a bottleneck!
- **Need for an ensemble**: use a set of predetermined feature selection approaches
  - **Voting-based ensemble**: simply a brute force ensemble
  - **Greedy ensemble**: penalize bad-performing selection methods and their features
- **Time and accuracy tradeoff**: use a hybrid of filter and wrapper approaches

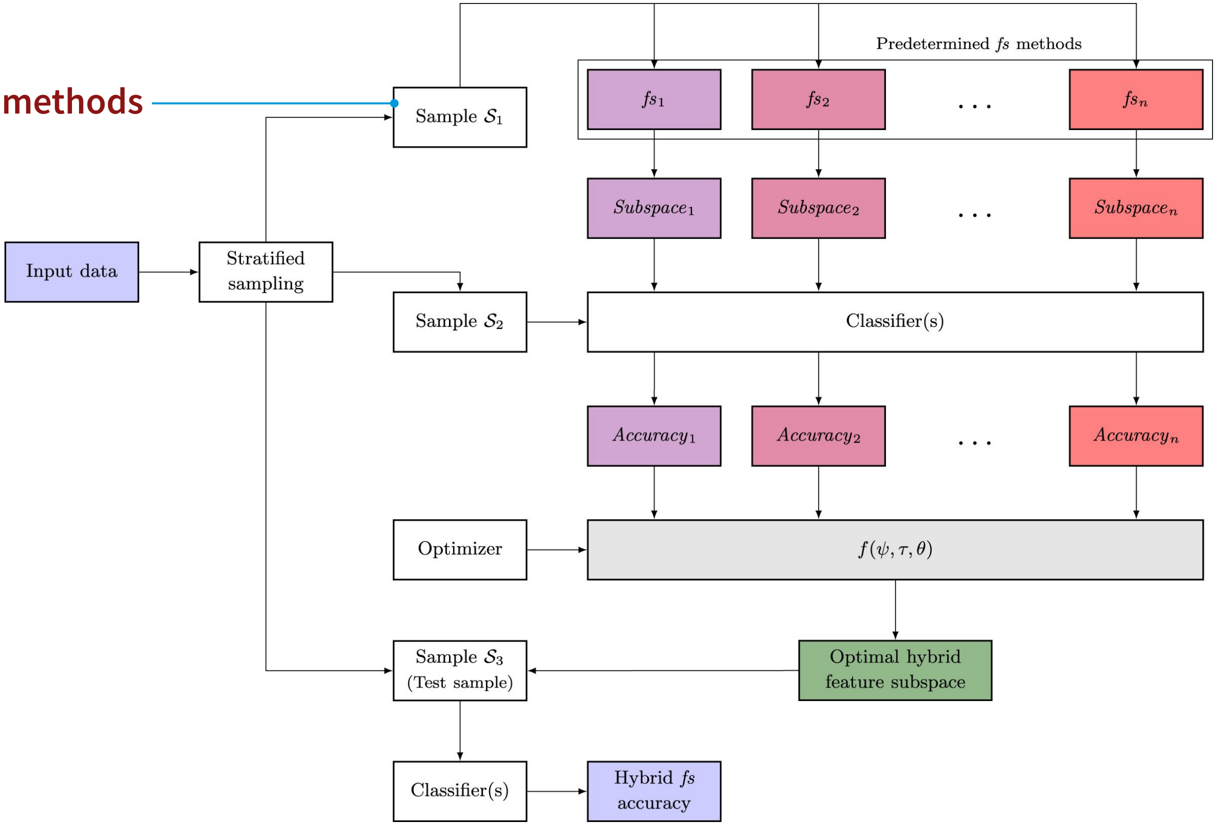
Agenda

1. Feature Space and Information
2. Research Gaps in Feature Selection?
3. Greedy Hybrid Ensemble with GA
4. Intuition and Penalization
5. Conclusions and Future Directions

Greedy Filter–Wrapper Hybrid Ensemble

Feature selection using the chosen methods

Feature space: #features(dataset)



Greedy Filter–Wrapper Hybrid Ensemble

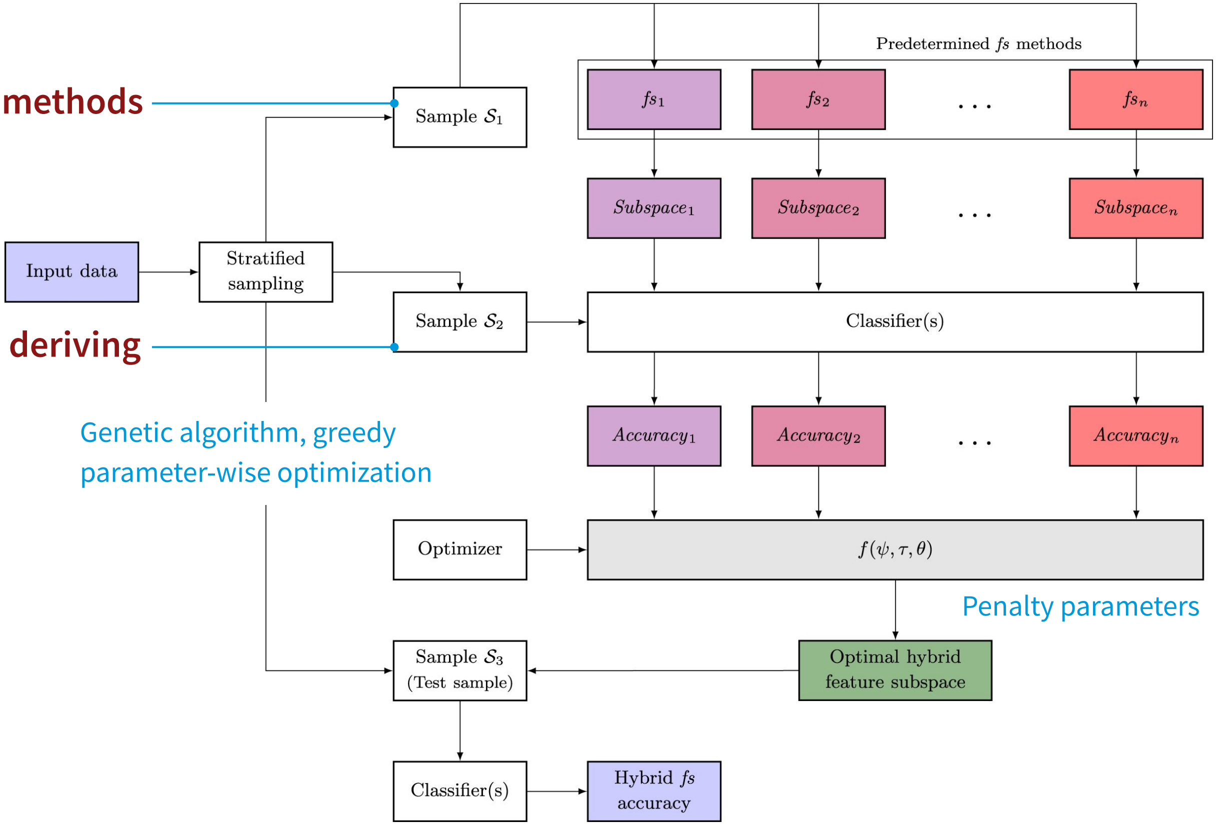
Feature selection using the chosen methods

Feature space: #features(dataset)

Evaluation of selected features and deriving the hybrid feature subspace

Feature space: #features(S_1)

Genetic algorithm, greedy parameter-wise optimization



Greedy Filter–Wrapper Hybrid Ensemble

Feature selection using the chosen methods

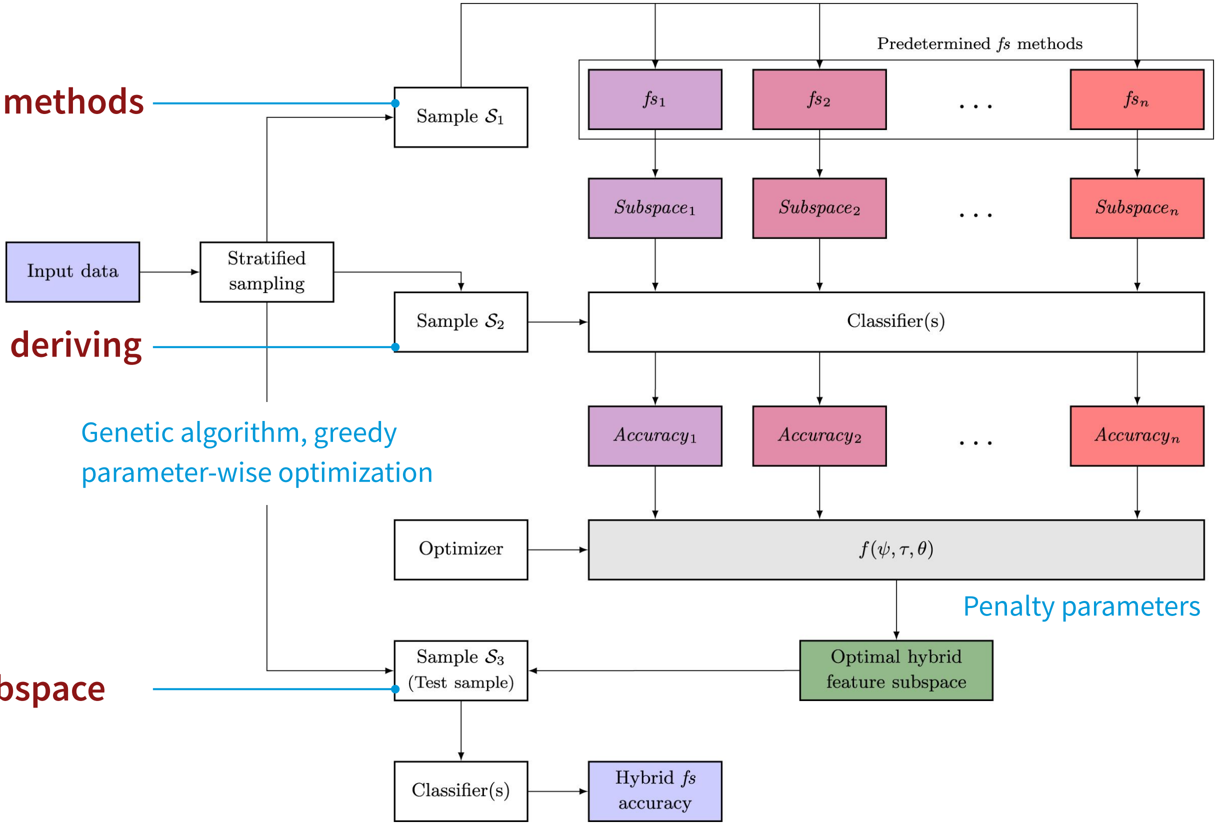
Feature space: #features(dataset)

Evaluation of selected features and deriving the hybrid feature subspace

Feature space: #features(S_1)

Evaluation of the hybrid feature subspace

Feature space: hybrid



Agenda

1. Feature Space and Information
2. Research Gaps in Feature Selection?
3. Greedy Hybrid Ensemble with GA
4. Intuition and Penalization
5. Conclusions and Future Directions

Greedy Hybrid Ensemble: Scoring Scheme

❖ Scoring of features (`featScore`) and selection techniques (`accScore`)

$$\texttt{featScore} = \begin{cases} \dfrac{|FS| - \rho_f + 1}{|FS|} & f \in \text{ranked } FS \\ 1/\,|FS| & f \in \text{unranked } FS \\ -1/\,|FS| & f \notin FS \end{cases}$$

General trends in parameter optimization: lower value of ψ, higher value of τ, and fine tuning of θ

Greedy Hybrid Ensemble: Scoring Scheme

- Scoring of features (`featScore`) and selection techniques (`accScore`)

$$
\texttt{featScore} = \begin{cases} \dfrac{|FS| - \rho_f + 1}{|FS|} & f \in \text{ranked } FS \\ 1/\,|FS| & f \in \text{unranked } FS \\ -1/\,|FS| & f \notin FS \end{cases}
$$

$$
\texttt{accScore} = \frac{|M| - \rho_m + 1}{|M|}
$$

index(m) + 1

General trends in parameter optimization: lower value of ψ, higher value of τ, and fine tuning of θ

Greedy Hybrid Ensemble: Scoring Scheme

- **Scoring of features** (`featScore`) and **selection techniques** (`accScore`)

$$\texttt{featScore} = \begin{cases} \dfrac{|FS| - \rho_f + 1}{|FS|} & f \in \text{ranked } FS \\ 1/\,|FS| & f \in \text{unranked } FS \\ -1/\,|FS| & f \notin FS \end{cases}$$

$$\texttt{accScore} = \frac{|M| - \rho_m + 1}{|M|}$$

index(m) + 1

- **Penalty parameters for greedy ensembling** of base feature subspaces
  - **Accuracy penalty (ψ)**: reduces the impact of accuracy scores = `accScore`/ψ
  - **Feature penalty (τ)**: increases the negative impact of the feature scores = `featScore`×τ

General trends in parameter optimization: lower value of ψ, higher value of τ, and fine tuning of θ

Greedy Hybrid Ensemble: Scoring Scheme

- **Scoring of features** (`featScore`) and **selection techniques** (`accScore`)

$$\texttt{featScore} = \begin{cases} \dfrac{|FS| - \rho_f + 1}{|FS|} & f \in \text{ranked } FS \\ 1/\,|FS| & f \in \text{unranked } FS \\ -1/\,|FS| & f \notin FS \end{cases}$$

index(m) + 1

$$\texttt{accScore} = \frac{|M| - \rho_m + 1}{|M|}$$

- **Penalty parameters for greedy ensembling** of base feature subspaces
  - **Accuracy penalty (ψ)**: reduces the impact of accuracy scores = `accScore`/ψ
  - **Feature penalty (τ)**: increases the negative impact of the feature scores = `featScore` $\times\,\tau$

- **Overall feature scoring** and hybrid feature selection (θ)

$$\texttt{overallScore} = \sum_{m}^{M} \texttt{featScore}(f) \times \texttt{accScore}(m) \longrightarrow \text{Threshold-based feature selection}$$

General trends in parameter optimization: lower value of ψ, higher value of τ, and fine tuning of θ

Greedy Hybrid Ensemble: Scoring Scheme

- **Scoring of features** (`featScore`) and **selection techniques** (`accScore`)

$$\texttt{featScore} = \begin{cases} \dfrac{|FS| - \rho_f + 1}{|FS|} & f \in \text{ranked } FS \\ 1/\,|FS| & f \in \text{unranked } FS \\ -1/\,|FS| & f \notin FS \end{cases}$$

$$\texttt{accScore} = \frac{|M| - \rho_m + 1}{|M|}$$

index(m) + 1

- **Penalty parameters for greedy ensembling** of base feature subspaces
  - **Accuracy penalty (ψ)**: reduces the impact of accuracy scores = `accScore`/ψ
  - **Feature penalty (τ)**: increases the negative impact of the feature scores = `featScore` × τ
- **Overall feature scoring** and hybrid feature selection (θ)

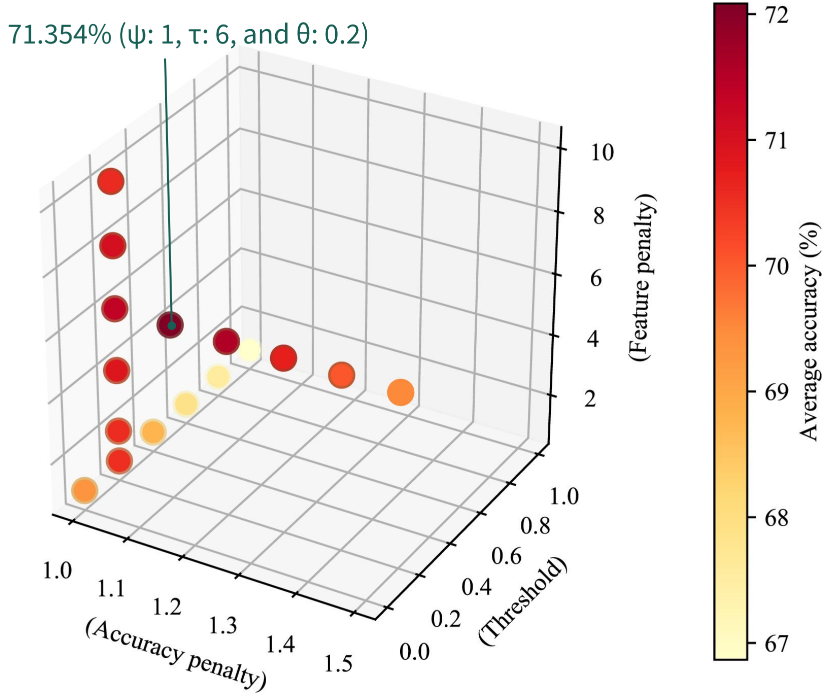
$$\texttt{overallScore} = \sum_{m}^{M} \texttt{featScore}(f) \times \texttt{accScore}(m) \longrightarrow$$

Threshold-based feature selection

- **Optimization of penalty parameters** (ψ, τ, and θ): genetic algorithm, greedy optimization, …

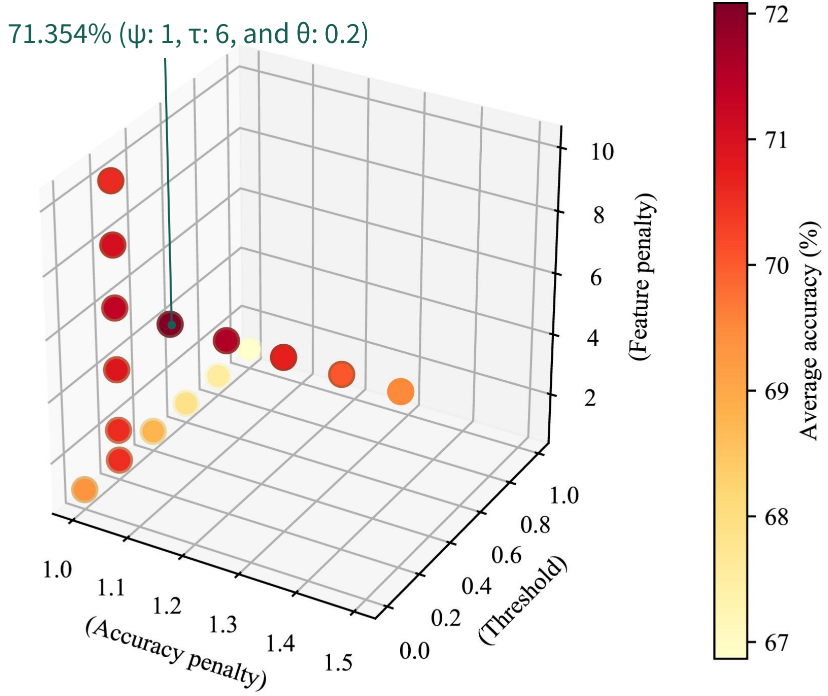
General trends in parameter optimization: lower value of ψ, higher value of τ, and fine tuning of θ

Optimization of Penalty Parameters

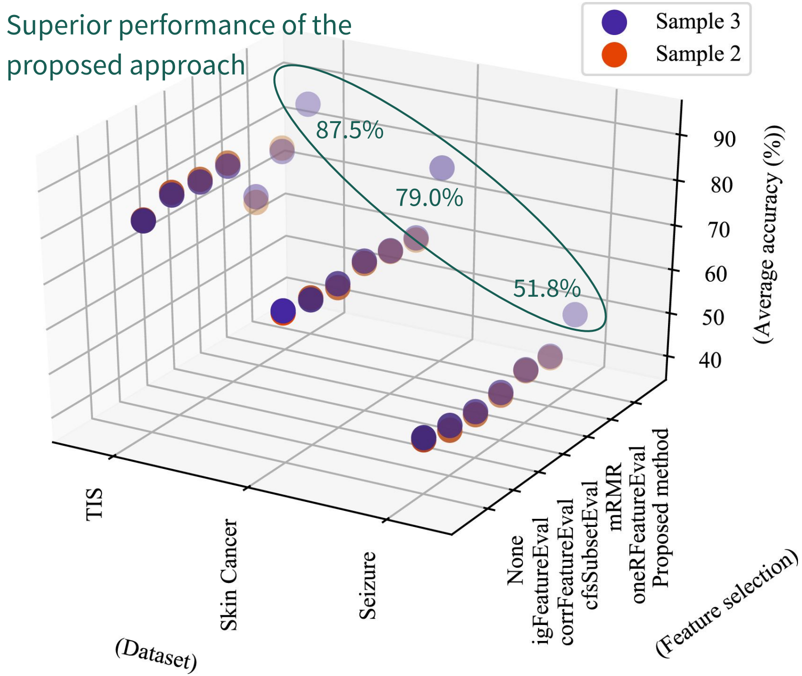


The effect of ψ, τ, and θ on Skin Cancer dataset
(greedy parameter-wise optimization)

Optimization of Penalty Parameters



The effect of ψ, τ, and θ on Skin Cancer dataset
(greedy parameter-wise optimization)



The effect of ψ, τ, and θ on Skin Cancer dataset
(genetic algorithm (N = 50, $p_c = 0.6$, $p_m = 0.1$))

Agenda

1. Feature Space and Information
2. Research Gaps in Feature Selection?
3. Greedy Hybrid Ensemble with GA
4. Intuition and Penalization
5. Conclusions and Future Directions

Conclusions and Future Directions

- Proposed a penalty based **greedy filter–wrapper hybrid ensemble approach** to facilitate optimal feature selection

Conclusions and Future Directions

- Proposed a penalty based **greedy filter–wrapper hybrid ensemble approach** to facilitate optimal feature selection
- Ensemble **greedily selects the features** from the subspaces obtained from the predetermined base selection methods

Conclusions and Future Directions

- Proposed a penalty based **greedy filter–wrapper hybrid ensemble approach** to facilitate optimal feature selection
- Ensemble **greedily selects the features** from the subspaces obtained from the predetermined base selection methods
- Specific **performance dependent penalty parameters** were used to penalize the base feature subspaces essential to achieve the optimal ensembling of those subspaces

Conclusions and Future Directions

- Proposed a penalty based **greedy filter–wrapper hybrid ensemble approach** to facilitate optimal feature selection
- Ensemble **greedily selects the features** from the subspaces obtained from the predetermined base selection methods
- Specific **performance dependent penalty parameters** were used to penalize the base feature subspaces essential to achieve the optimal ensembling of those subspaces
- At any point in time, only a stratified sample and not the entire dataset is not used for computation; the **computational complexity is significantly reduced**

Conclusions and Future Directions

- Proposed a penalty based **greedy filter–wrapper hybrid ensemble approach** to facilitate optimal feature selection
- Ensemble **greedily selects the features** from the subspaces obtained from the predetermined base selection methods
- Specific **performance dependent penalty parameters** were used to penalize the base feature subspaces essential to achieve the optimal ensembling of those subspaces
- At any point in time, only a stratified sample and not the entire dataset is not used for computation; the **computational complexity is significantly reduced**
- We leverage **effective heuristic search strategies** including the greedy parameter-wise optimization and the GA to obtain optimal values of the penalty parameters

Conclusions and Future Directions

- Proposed a penalty based **greedy filter–wrapper hybrid ensemble approach** to facilitate optimal feature selection
- Ensemble **greedily selects the features** from the subspaces obtained from the predetermined base selection methods
- Specific **performance dependent penalty parameters** were used to penalize the base feature subspaces essential to achieve the optimal ensembling of those subspaces
- At any point in time, only a stratified sample and not the entire dataset is not used for computation; the **computational complexity is significantly reduced**
- We leverage **effective heuristic search strategies** including the greedy parameter-wise optimization and the GA to obtain optimal values of the penalty parameters
- The proposed method introduces additional (penalty) parameters which **require prior training** to obtain the optimal setting in advance

Further Reading

[1] Gangavarapu, Tushaar, and Nagamma Patil. *A novel filter-wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets.* Applied Soft Computing (2019): 105538.

[2] Tu, Qiang, Xuechen Chen, and Xingcheng Liu. *Multi-strategy ensemble grey wolf optimizer and its application to feature selection.* Applied Soft Computing 76 (2019): 16-30. Accessible: sciencedirect/science/article/pii/S1568494618306793.

[3] Min, Fan, Qinghua Hu, and William Zhu. *Feature selection with test cost constraint.* International Journal of Approximate Reasoning 55.1 (2014): 167-179.

[4] Dong, Hongbin, et al. *A novel hybrid genetic algorithm with granular information for feature selection and optimization.* Applied Soft Computing 65 (2018): 33-46. Accessible: sciencedirect/science/article/pii/S1568494618300048.

[5] Masood, Mustafa K., Yeng Chai Soh, and Chaoyang Jiang. *Occupancy estimation from environmental parameters using wrapper and hybrid feature selection.* Applied Soft Computing 60 (2017): 482-494.

[6] Chandrashekar et al. *A survey on feature selection methods.* Computers & Electrical Engineering 40.1 (2014).

> Nanos gigantum humeris insidentes

Thank you ~