Evaluating the Potential of SHAP-Based Feature Selection for Improving Classification Performance

Basabi Chakraborty
Iwate Prefectural University, Japan
Madanapalle Institute of Technology and Science, AP, India
And
Ashis Kumar Mandal, Bangladesh
10th March, 2025

About Me

- Received B. Tech, M. Tech and Ph. D in Radio Physics and Electronics from Calcutta University
- Worked in Indian Statistical Institute, Kolkata as Computer Engineer
- Visiting Researcher in AIC (Advanced Intelligent Communication) Systems in Sendai, Japan (1991-1993)
- Doctoral and Post doctoral research In RIEC (Research Institute of Electrical Communication) , Tohoku University, Japan (1993-1998)
- Ph. D in Information Science (1996)
- Faculty in Dept. of Software and Information Science ,Iwate Prefectural university (1998-2022)
- Visiting faculty in Dept. of Electrical and Computer Engineering, University of Western Ontario, Canada (Oct. 2006 –March 2007)
- Professor Emeritus and Distinguished Professor in Iwate Prefectural University
- Dean, School of Computing , Madanapalle Institute of Science and Technology

Contents

Introduction

Why do we need Feature Selection?

- An important process in the area of Pattern Recognition, Machine learning and Data mining.

- Dimensionality Reduction, Feature Extraction or Generation and Feature Selection/Subset Selection

Dimensionality Reduction

- Curse of dimensionality and peaking phenomenon

- The performance of a classifier depends on the interrelationship between
  - sample sizes
  - number of features
  - classifier complexity
- If table-lookup technique is adopted for classification, how many training samples are required w.r.t the number of features?

Dimensionality Reduction

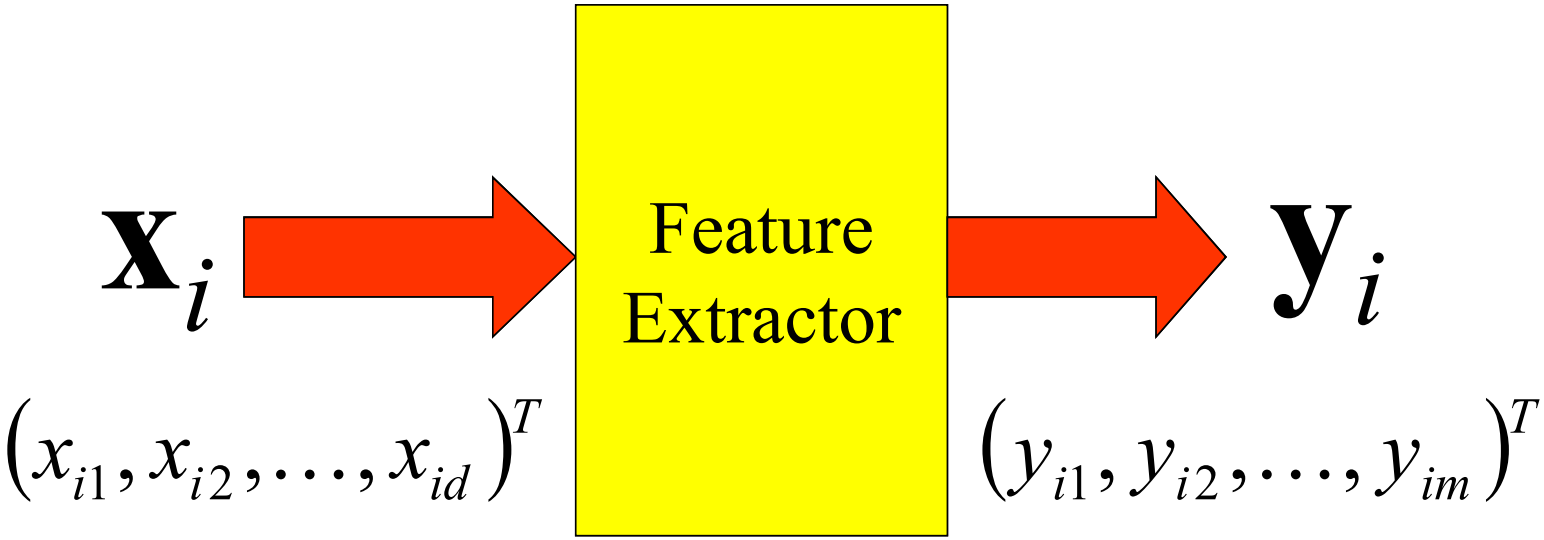
- Feature Selection / Subset selection

  Feature selection in Measurement Space

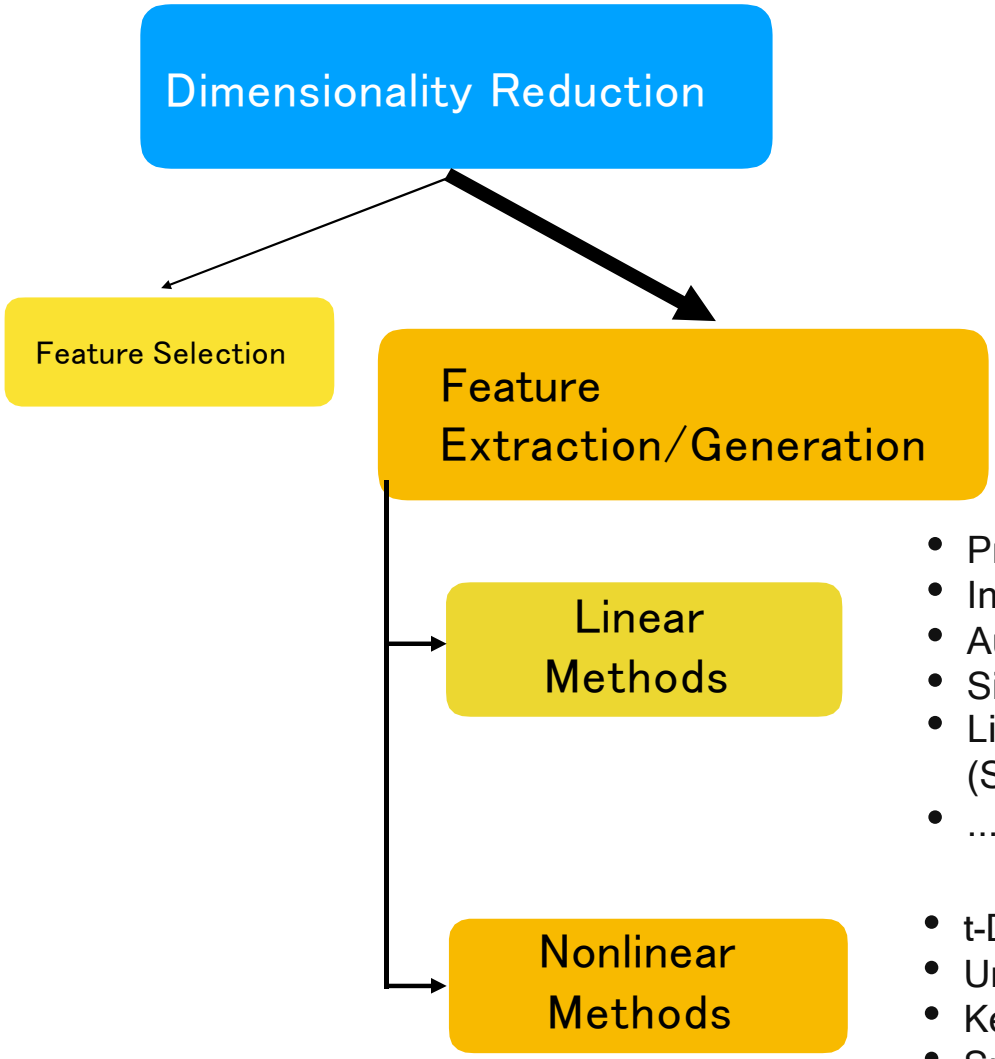
- Feature Generation/ Extraction

  Feature selection in transformed space

Feature Extractor/Generator



$m \leq d$, usually

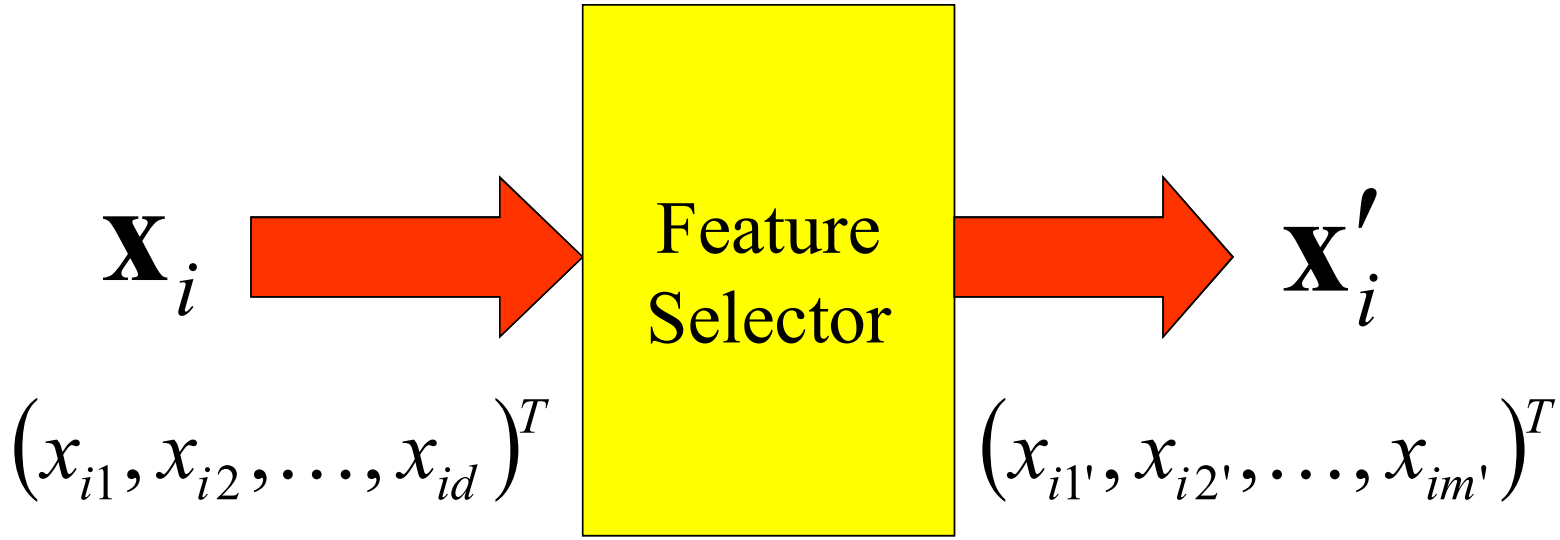


- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Autoencoders (linear act. func.)
- Singular Vector Decomposition (SVD)
- Linear Discriminant Analysis (LDA) (Supervised)
- ...

- t-Distr. Stochastic Neigh. Emb. (t-SNE)
- Uniform Manifold Approx. & Proj. (UMAP)
- Kernel PCA
- Spectral Clustering
- Autoencoders (non-linear act. func.)
- ...

possible Selections $\binom{d}{m}$

Feature Selector



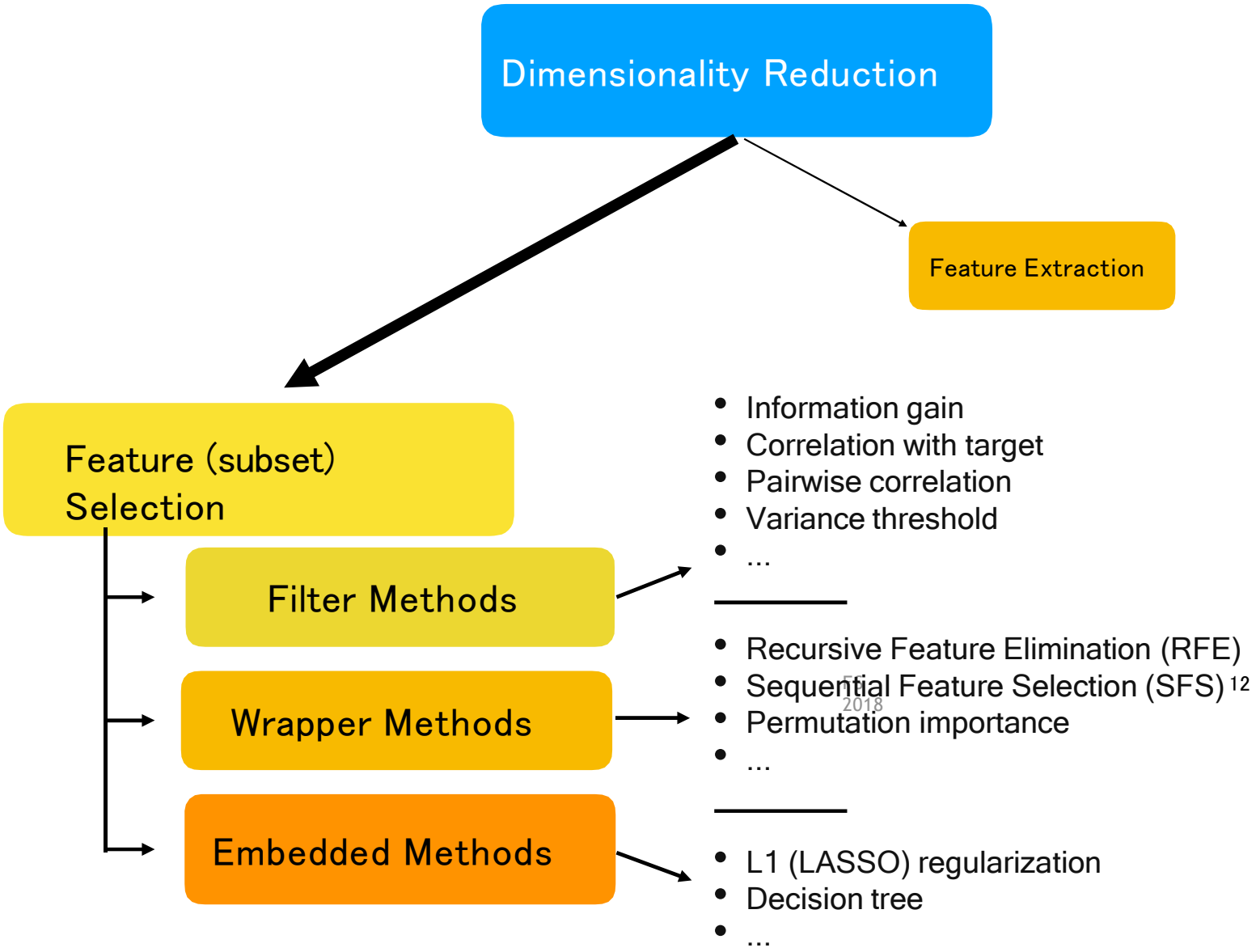
$m \leq d$, usually

Feature Subset Selection

- Given a set of *N* features, select a subset of size *m* that leads to the smallest classification error.
- Exhaustive evaluation process can guaranty the optimal subset
- No non exhaustive sequential feature selection procedure can be guaranteed to produce the optimal subset.
- For large feature set, the problem becomes NP hard

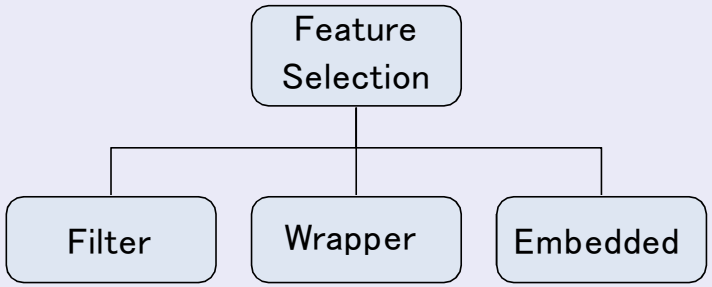
Feature Subset Selection

- Retain the discriminatory information for recognition while discard the unnecessary information
- Best two individual features do not comprise the best feature subset of two features
- Criterion Function to evaluate a feature or a set of features
- Search strategy to find the best feature subset from a number of possible candidates

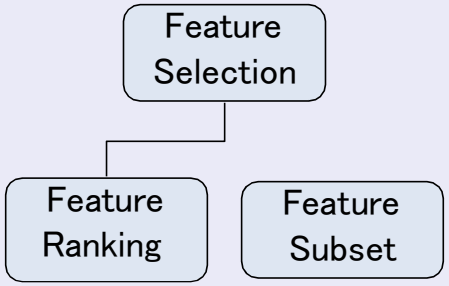

Dimensionality Reduction
Feature Extraction
Feature (subset) Selection
Filter Methods
Wrapper Methods
Embedded Methods
Information gain
Correlation with target
Pairwise correlation
Variance threshold
...
Recursive Feature Elimination (RFE)
Sequential Feature Selection (SFS)
Permutation importance
...
L1 (LASSO) regularization
Decision tree
...

Feature Subset Selection

Taxonomy of Feature Selection: evaluation

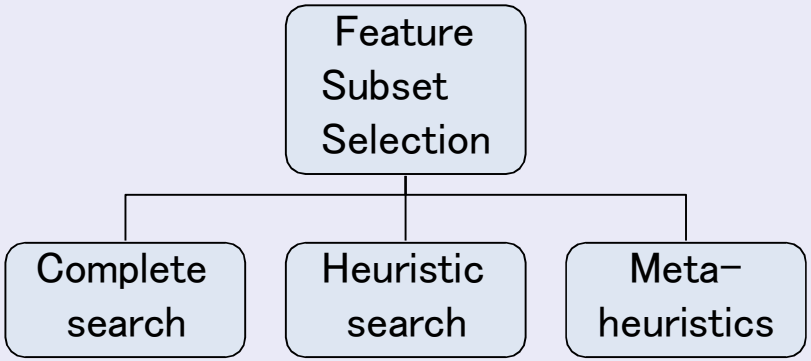


Taxonomy of Feature Selection: selection type



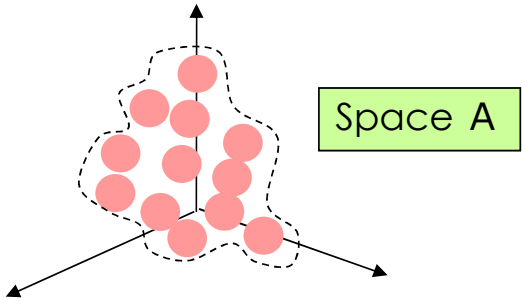
- Feature Subset Selection
  - Finding best subset is NP-hard problem
  - Combinatorial optimization problem
  - For n feature, total Possible number of subset is $2^n - 1$

Search Strategy for Feature subset

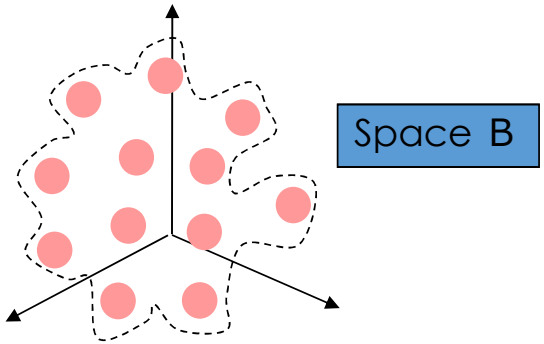


Goodness of Feature Space

- Small intra-class variation

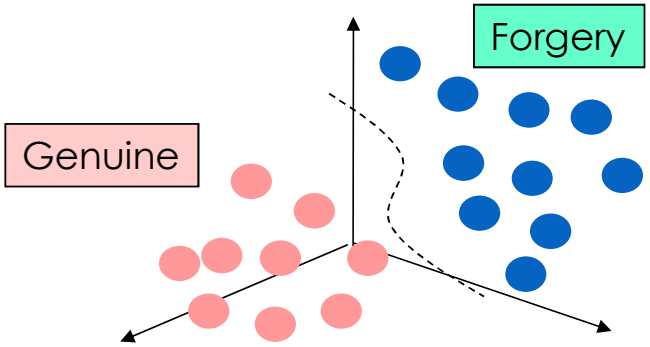


Stability

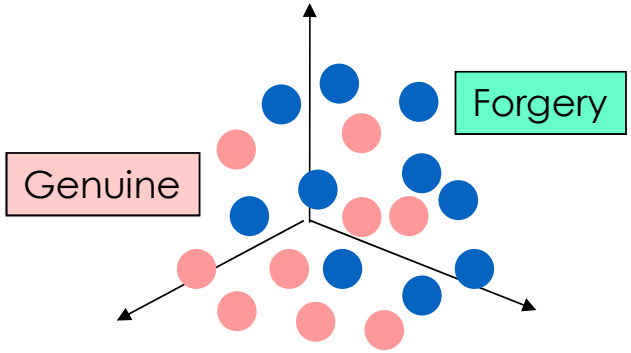


Instability

- Large inter-class variation



Easy Verification



Difficult Verification

How to measure quality of a feature subset

- Classification accuracy
- Relevancy
- Redundancy (Correlation)
- Feature – Feature Interaction
- Number of features (cardinality)
- Interpretability
- Stability
- Computational time

Feature Evaluation Measures

- Statistical and Mathematical Measures
- Distance based, dependency/relevance based
- Information content, classifier error
- Consistency based measure
- Soft computing based Measures
- Fuzzy set and Rough set
- Correlation and mutual information

Shapley Value: Background and Definition

- An old idea from Cooperative Game Theory (L. S. Shapley 1953, Won 2012 Nobel Memorial Prize in economics), originally not related to AI/ML.
- It is now the basis of a popular technique for Explainable Machine Learning.
- Shapley Value is a technique for allocating credit to players in a coalitional game (Transferable Utility TU game).

- Famously derived from a set of *fairness axioms*

Cooperative game notation

- Set of *players* $D = \{1, \dots, d\}$
- A *game* is given by specifying a value for every coalition $S \subseteq D$
- Mathematically represented by a *characteristic function*: (defines the payoff for each coalition)

$$v: 2^D \mapsto \mathbb{R}$$

- Grand coalition value $v(D)$, null coalition $v(\emptyset)$, arbitrary coalition $v(S)$

Example

- Let Set of *players* D = 1, 2, 3
- Let the payoffs $v(\emptyset) = 0$; $v(\{1\}) = 7$; $v(\{2\}) = 11$; $v(\{3\}) = 14$; $v(\{1,2\}) = 18$;

$v(\{1,3\}) = 21$; $v(\{2,3\}) = 23$; $v(\{1,2,3\}) = 25$;

A ***solution concept*** defines an allocation principle through which rewards/credits can be distributed among players. ***Solution vector*** is a specific allocation.

Set of ***permutations of the player set*** (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2) ,(3,2,1) . In the permutation (3,2,1), ***predecessor*** set of the 1st player is (3,2).

Marginal Contribution

- $v(\emptyset) = 0$; $v(\{1\}) = 7$; $v(\{2\}) = 11$; $v(\{3\}) = 14$; $v(\{1,2\}) = 18$;

$v(\{1,3\}) = 21$; $v(\{2,3\}) = 23$; $v(\{1,2,3\}) = 25$;

Permutation	Player 1	Player 2	Player 3
(1,2.3)	7	11	7
(1,3,2)	7	4	14
(2,1.3)	7	11	7
(2,3,1)	2	11	12
(3,1,2)	7	4	14
(3,2,1)	2	9	14
Shapley Value	32/6	50/6	68/6

Shapley Value

- The Shapley value (SV) assigns a vector of credits to each game (in $\mathbb{R}^d$, one credit per player)
- The Shapley value of a player is the average marginal contribution of the player to the value of the predecessor set over every possible permutation of the player set.
- Given an ordering π, each player contributes when added to the preceding ones
- SV is the average contribution across all orderings

Fairness axioms: Properties of Shapley value

Consider a game v and credit allocations $\phi(v) = [\phi_1(v), \ldots, \phi_d(v)]$. We want to satisfy the following properties:

- ***(Efficiency)*** The credits sum to the grand coalition's value, or $\sum_{i \in D} \phi_i(v) = v(D) - v(\emptyset)$
- ***(Symmetry)*** If two players (i, j) are interchangeable, or $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq D$, then $\phi_i(v) = \phi_j(v)$
- ***(Null player)*** If a player contributes no value, or $v(S \cup \{i\}) = v(S)$ for all $S \subseteq D$, then $\phi_i(v) = 0$
- ***(Linearity)*** The credits for linear combinations of games behave linearly, or $\phi(c_1 v_1 + c_2 v_2) = c_1 \phi(v_1) + c_2 \phi(v_2)$, where $c_1, c_2 \in \mathbb{R}$

Lloyd Shapley, "A value for n-person games" (1953)

Shapley Value

- The Shapley value (SV) is the only function ϕ: G ↦ $\mathbb{R}^d$ to satisfy these properties
- Given by the following equation:

$$\phi_i = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|! \ (|D| - |S| - 1)!}{|D|!} [v(S \cup \{i\}) - v(S)]$$

Approximations of Shapley Values

- SV computation requires an exponential number of characteristic function evaluation related to every possible permutations leading to exponential time complexity
- Monte Carlo Permutation Sampling->
- Every iteration a random sample from the set of permutations is selected.
- The marginal contribution of the players are scaled down by the number of players in the permutation and added to the approximated SV from the previous iteration

Stratified Sampling for Variance Reduction

- In further improvement of Monte Carlo estimation, **stratified sampling** was proposed by dividing the permutations into homogeneous non overlapping sub populations.
- Development of different definitions of **strata** like when the player is in a specific position or to minimize the uncertainty of the estimate etc..
- **Other techniques:**
- Instead of random samples, ergodic but not independent samples
- Bayesian Mote Carlo approach

Shapley Value in Machine Learning

- Consider features as players
- Consider model behavior as profit
  - e.g., the prediction, the loss, etc.
- Then, use Shapley Values to quantify each feature's impact/importance for the outcome.

Shapley value in Feature Subset Selection

- **Contribution Selection Algorithm (CSA)** – developed by Cohen in 2005, 2007, an iterative algorithm with forward or backward selection. Wrapper like technique with Shapley value based feature evaluation.
- At each step it ranks each feature according to its contribution/ shapley value and either adds feature with highest value or delete feature with lowest value.
- The steps are repeated until the contribution values (sum) of the resulting feature subset exceeds a threshold value.
- The simulation results with benchmark data sets showed that the performance of CSA with backward elimination is the best.

Shapley value in Feature Subset Selection

- Shapley value based feature ranking as a filter method is also used in several works (Sun et al , 2012).

- F. Afgah, A.Raji , 2014, 2015 → Several papers on medical data used shapley value to compute the importance of feature or feature subset in filter or wrapper methods.

- The shapley value based feature subset selection algorithms showed higher performance regarding classification accuracy compared to existing algorithms by simulation experiments.

Evaluation of ML/DL models

- Performance in terms of average accuracy is important
- Cost of wrong decision is low
- Highly one dimensional metric
- Problems with highly complex real world systems
- Risk sensitive systems – Medical diagnosis, Financial prediction etc
- Safely critical systems : Vehicle control, autonomous decision systems etc.

Problems with real world system

- Complex real-world systems
  - Risk-sensitive systems
    - E.g. Medical diagnosis, Financial modeling/prediction
  - Safety-critical systems
    - E.g. Cockpit decision support

Cost of a bad decision can be very high

- Accuracy is not the only objective
- Need for a multi-dimensional perspective

In general, it seems like there are few fundamental problems –

- We don’t trust the models
- We don’t know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model makes similar mistakes as humans ?
- How to change model when things go wrong ?

Interpretability is one way we try to deal with these problems

Explainable Machine Learning

- Interpretability : Comprehending what the model is doing (How ?)
- Explainability: Summarizing the reasons for model behavior, causes of decision (Why?)
- .Explainable models are interpretable by default, but the reverse is not always true
- Building Interpretable model is the first step.

Shapley value for Interpretable feature subset selection

- Shapley value has been used as **locally interpretable model agnostic tool** for explainable Machine Learning like **LIME** or **DeepLift.**
- Shaply Value based Error Apportioning (SVEA) ->developed in 2020 by Tripathi et al. for **interpretable feature subset selection.** This algorithm is used for feature subset selection before training of the classification model. Shapley values are used here as feature contribution to the training error.

SHAP (Linear Regression Approximation)

- SHAP = *SHapley Additive exPlanations*

- Popularized use of Shapley values in ML
  - Also used in earlier work by Lipovetsky & Conklin (2001), Strumbelj et al. (2009), Datta et al. (2016)

- SHAP uses Shapley values to explain individual predictions

Lundberg & Lee, "A unified approach to interpreting model predictions" (2017)

SHAP

- SHAP has been proposed as a unified measure of feature importance. Shapley values of a conditional expectation function of the original model.
- It assigns each data point of each feature an importance value for a particular prediction.
- The definition of SHAP is designed to closely aligned with the Shaply regression
- The exact computation of SHAP values is challenging
- If we assume feature independence when approximating conditional expectation, SHAP can be estimated directly using Shaply value sampling method.

Variants of SHAP (Approximations)

- Model Agnostic
- Kernel SHAP -> Linear LIME + Shapley Values

Linear LIME uses a linear explanation model.

Model Specific

Linear SHAP, Max SHAP, Deep SHAP (DeepLift + Shapley Values)

Tree SHAP: Tree SHAP is an algorithm to compute exact SHAP values for Decision Trees based models.

Contd..

- SAGE (Shapley Additive Global Importance) : For assessing the role of each feature globally. A model agnostic method that quantifies feature importance while accounting for feature interactions (Lundberg, 2020)
- Shap-select, a feature selection framework integrating SHAP values with statistical significance testing to provide an efficient heuristic for feature selection. The framework runs a regression of the target on the SHAP values of the original features, on the validation set, and filters features based on their coefficients and statistical significance. (Kraev 2024)

Assessing SHAP as a Feature Selection Tool

- SHAP approximates Shapley values through weighted linear regression or through different assumptions about feature dependence for ensemble tree models.
- Local explanations are accurate than global one and summing up local contributions can give a better global picture.
- Simulation experiments (Wilson et al 2020) with limited benchmark data sets showed better performance compared to many common feature selection algorithms.

SHAP for Feature Selection

- SHAP uses local explainability to build surrogate models to black box learning models. It changes/perturbs the input to test change in model prediction. If it does not change much for a feature, that feature for that particular data point is not important.
- Need to set SHAP characteristic function properly to use it as a feature selector. Absolute SHAP value should represent the importance of features.

Feature Subset Selection Using Shapley Values General Algorithm

1. **Define the Coalition Game**:
   - Each feature represents a "player" in the game.
   - The "value" of a coalition (subset of features) is the performance (e.g., accuracy, R^2, or any relevant metric) of the model trained with those features.
2. **Compute Marginal Contributions**:
   - For each subset of features, calculate the difference in model performance when adding a feature to the subset.
   - Average the marginal contributions of each feature across all subsets to compute Shapley values.
3. **Rank Features**:
   - Rank features based on their Shapley values. Features with higher values contribute more significantly to the model's performance.
4. **Select Subsets**:
   - Use the ranked features to select subsets for your final model based on a threshold (e.g., cumulative Shapley value, top-k features, etc.).

Simulation Experiments

- We conducted preliminary simulation experiments with 10 bench mark data sets from UCI ML repository.
- We used popular 3 rank based feature selection algorithms using different feature evaluation criteria, Chi-square, Correlation-based and Mutual Information based Feature Selection with linear SHAP versions.
- We took different percentage of top ranked feature subsets for evaluation using linear classifier.

Comparison of Classification Accuracy (50% of the features selected)

Dataset	Linear Shap	Chi-Square	Mutual Information	Correlation
Breast-W	0.97	0.97	0.97	0.97
Clean	0.81	0.78	0.77	0.78
Hepatitis	0.76	0.77	0.77	0.77
Parkinsons	0.76	0.76	0.78	0.80
Promoters	0.76	0.78	0.79	0.78
Qsar-biodeg	0.82	0.77	0.80	0.80
Sonar	0.75	0.75	0.76	0.74
Spect	0.69	0.70	0.70	0.70
Spectf	0.78	0.78	0.77	0.78
Wisconsin	0.96	0.93	0.93	0.93

Simulation Results

- As a feature selection tool, liner SHAP exhibits similar accuracy for 50% to 75% reduction of feature set compared to full set.
- Linear SHAP demonstrated better performance in 3 out of 10 data sets, similar in 6 and slightly poor in one data set
- On the average, it is found that Linear SHAP's performance is not statistically different from any other three approaches.

Merits of Shapley value for feature selection

1. **Fairness and Interpretability**:
   - Shapley values are grounded in game theory, ensuring a fair distribution of contributions among features.
   - They provide interpretable results by showing how each feature or subset impacts model performance.
2. **Handles Feature Interactions**:
   - Shapley values account for interactions among features. This is particularly beneficial in datasets with correlated or dependent features.
3. **Model-Agnostic:**
   - They work with any model, making them flexible for use across various machine learning algorithms.
4. **Feature Redundancy**:
   - By evaluating all subsets, Shapley values can highlight redundant
5. **Comprehensive Evaluation**:
   - By considering all subsets of features, Shapley values provide a holistic assessment, avoiding biases that simpler methods might introduce.

Limitations and Challenges

- **Computational complexity**

The primary drawback of Shapley values is their computational cost. Approximations, like Monte Carlo sampling, reduce this cost but can introduce variability in the results.

Sensitivity to Model and Data -> changes is either can lead to different feature ranking, might not generalize well.

Local vs Global -> impacts performance for selecting global subset of features

Sensitivity to noisy data:

- **Causal Relationship:** Shapley value of a variable do not reflect their causal relationship with the target of interest.

Limitations and Challenges

- As Shapley value is computationally challenging, varieties of approximations are used for implementations whose theoretical basis and assumptions are different, leading to different results to the same problem.
- Need for proper setting of the feature selection problem.
- Combine with other techniques to improve performance.

Conclusions

- In summary, Shapley value is an effective tool for feature subset selection, particularly when feature interactions matter. However, computational costs and noise sensitivity require careful management. For practical applications, approximate methods and hybrid approaches often strike the best balance between performance and feasibility.
- There is scope for research and various experiments to develop better algorithms.