Maximum Relevancy, Minimum Redundancy

Michael Matějů

AI Squad

July 30, 2023

Outline

Motivation

- Mirek Pavelka point this algorithm to me
- Recent popularity of the algorithm due to UBER's paper (2019).
- The original paper from 2005.

Motivation

- Mirek Pavelka point this algorithm to me
- Recent popularity of the algorithm due to UBER's paper (2019).
- The original paper from 2005.

Motivation

- Mirek Pavelka point this algorithm to me
- Recent popularity of the algorithm due to UBER's paper (2019).
- The original paper from 2005.

Introduction

Feature selection methods allow you to choose the best features in your dataset, with respect to some criteria. Feature selection methods can be divided into three families:

- Filter Methods: Feature selection is made as part of the pre-processing, that is, before we train a model. We filter out features that perform poorly based on some criteria (e.g. correlation).
- Wrapper Methods: Iteratively choose a subset of features, train your model, and choose the best combination. Very time consuming.
- Embedded Methods: Embedded methods take advantage of the feature importance estimations which are embedded in the algorithm. E.g. Random Forest.

Introduction

Feature selection methods allow you to choose the best features in your dataset, with respect to some criteria. Feature selection methods can be divided into three families:

- Filter Methods: Feature selection is made as part of the pre-processing, that is, before we train a model. We filter out features that perform poorly based on some criteria (e.g. correlation).
- Wrapper Methods: Iteratively choose a subset of features, train your model, and choose the best combination. Very time consuming.
- Embedded Methods: Embedded methods take advantage of the feature importance estimations which are embedded in the algorithm. E.g. Random Forest.

Introduction

Feature selection methods allow you to choose the best features in your dataset, with respect to some criteria. Feature selection methods can be divided into three families:

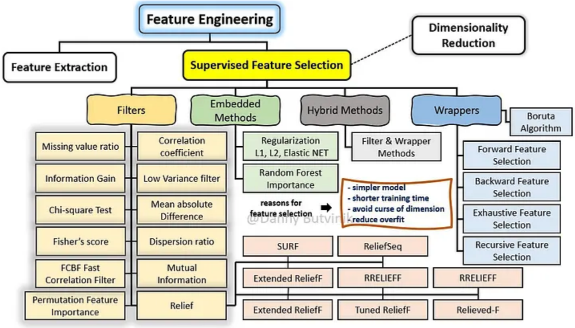
- Filter Methods: Feature selection is made as part of the pre-processing, that is, before we train a model. We filter out features that perform poorly based on some criteria (e.g. correlation).
- Wrapper Methods: Iteratively choose a subset of features, train your model, and choose the best combination. Very time consuming.
- Embedded Methods: Embedded methods take advantage of the feature importance estimations which are embedded in the algorithm. E.g. Random Forest.

Introduction

Feature selection methods allow you to choose the best features in your dataset, with respect to some criteria. Feature selection methods can be divided into three families:

- Filter Methods: Feature selection is made as part of the pre-processing, that is, before we train a model. We filter out features that perform poorly based on some criteria (e.g. correlation).
- Wrapper Methods: Iteratively choose a subset of features, train your model, and choose the best combination. Very time consuming.
- Embedded Methods: Embedded methods take advantage of the feature importance estimations which are embedded in the algorithm. E.g. Random Forest.

Cheat Sheet


Feature Engineering
Dimensionality Reduction
Feature Extraction
Supervised Feature Selection
Filters
Embedded Methods
Hybrid Methods
Wrappers
Boruta Algorithm
Missing value ratio
Correlation coefficient
Regularization L1, L2, Elastic NET
Filter & Wrapper Methods
Forward Feature Selection
Information Gain
Low Variance filter
Random Forest Importance
Backward Feature Selection
- simpler model
- shorter training time
- avoid curse of dimension
- reduce overfit
reasons for feature selection
Chi-square Test
Mean absolute Difference
Exhaustive Feature Selection
Fisher's score
Dispersion ratio
SURF
ReliefSeq
Recursive Feature Selection
FCBF Fast Correlation Filter
Mutual Information
Extended ReliefF
RRELIEFF
RRELIEFF
Permutation Feature Importance
Relief
Extended ReliefF
Tuned ReliefF
Relieved-F

MRMR

MRMR works iteratively. At each iteration, it identifies the best feature (according to a rule) and adds it to the basket of selected features. Once a feature goes into the bucket, it cannot ever come out.

In each iteration i a score is computed for each feature f according to formula:

$$score_i(f) = \frac{relevance(f|target)}{redundancy(f|\text{features selected until }(i-1))},$$

where we have to specify the relevancy and redundancy functions.

MRMR

MRMR works iteratively. At each iteration, it identifies the best feature (according to a rule) and adds it to the basket of selected features. Once a feature goes into the bucket, it cannot ever come out.

In each iteration i a score is computed for each feature f according to formula:

$$score_i(f) = \frac{relevance(f|target)}{redundancy(f|\text{features selected until } (i-1))} \quad ,$$

where we have to specify the relevancy and redundancy functions.

Feature importance

There are couple of variants to calculate the feature importance from the original UBER paper. For discrete features:

- Mutual Information Difference (MID):

$$f^{MID}(X_i) = I(Y, X_i) - \frac{1}{|S|} \sum_{X_s \in S} I(X_s, X_i)]$$

- Mutual Information Quotient (MIQ):

$$f^{MID}(X_i) = I(Y, X_i) / \left(\frac{1}{|S|} \sum_{X_s \in S} I(X_s, X_i)] \right)$$

Feature importance

There are couple of variants to calculate the feature importance from the original UBER paper. For discrete features:

- Mutual Information Difference (MID):

$$f^{MID}(X_i) = I(Y, X_i) - \frac{1}{|S|} \sum_{X_s \in S} I(X_s, X_i)]$$

- Mutual Information Quotient (MIQ):

$$f^{MID}(X_i) = I(Y, X_i) / \left(\frac{1}{|S|} \sum_{X_s \in S} I(X_s, X_i)] \right)$$

Feature importance

For continuous features:

- F-test Correlation Difference (FCD):

$$f^{MID}(X_i) = F(Y, X_i) - \frac{1}{|S|} \sum_{X_s \in S} \rho(X_s, X_i)]$$

- F-test Correlation Quotient (FCQ):

$$f^{MID}(X_i) = F(Y, X_i) / \left(\frac{1}{|S|} \sum_{X_s \in S} \rho(X_s, X_i)] \right)$$

where ρ is Pearson correlation.

Feature importance

For continuous features:

- F-test Correlation Difference (FCD):

$$f^{MID}(X_i) = F(Y, X_i) - \frac{1}{|S|} \sum_{X_s \in S} \rho(X_s, X_i)]$$

- F-test Correlation Quotient (FCQ):

$$f^{MID}(X_i) = F(Y, X_i) / \left(\frac{1}{|S|} \sum_{X_s \in S} \rho(X_s, X_i)] \right)$$

where ρ is Pearson correlation.

Feature importance

Another extensions:

- RDC algorithm for redundancy function (The randomized dependence coefficient)
- Model-based Feature importance for relevance:
  - Random Forest Correlation Quotient
  - Random Forest RDC Quotient
  - Pure Random Forest

From benchmarks, RF variances are good if down-stream model is Random Forest. FCQ is pretty close and works on more models.

Feature importance

Another extensions:

- RDC algorithm for redundancy function (The randomized dependence coefficient)
- Model-based Feature importance for relevance:
  - Random Forest Correlation Quotient
  - Random Forest RDC Quotient
  - Pure Random Forest

From benchmarks, RF variances are good if down-stream model is Random Forest. FCQ is pretty close and works on more models.

Feature importance

Another extensions:

- RDC algorithm for redundancy function (The randomized dependence coefficient)
- Model-based Feature importance for relevance:
  - Random Forest Correlation Quotient
  - Random Forest RDC Quotient
  - Pure Random Forest

From benchmarks, RF variances are good if down-stream model is Random Forest. FCQ is pretty close and works on more models.

Example

F Statistic

		Target Income
Not Selected	Mother's IQ	42.30
	Father's IQ	40.51
	Age (integer)	35.37
	Height	12.86
	Partner's height	8.08

Pearson Correlation

		Selected: Age	Selected: IQ		Average
Not Selected	Mother's IQ	0.01	0.67	→	0.34
	Father's IQ	0.00	0.64	→	0.32
	Age (integer)	1.00	0.06	→	0.53
	Height	0.01	0.00	→	0.01
	Partner's height	0.01	0.02	→	0.02

MRMR Score

Not Selected	Mother's IQ	42.30/0.34 =	124.41
	Father's IQ	40.51/0.32 =	126.59
	Age (integer)	35.37/0.53 =	66.76
	Height	12.86/0.01 =	2572.00
	Partner's height	8.08/0.02 =	538.67

Illustration of what happens at iteration 3. [Figure by Author]

The End

Thank you for your attention and patience.