12º Workshop em Engenharia e Tecnologia Espaciais
6, 7, 13 e 14 de novembro de 2021

# Seleção de atributos em dados de telemetria de satélite

**Ivan Márcio Barbosa [1], Maurício Gonçalves Vieira Ferreira[1], Milton de Freitas Chagas Júnior[1]**

[1]Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

Aluno de Doutorado do curso de Engenharia e Gerenciamento de Sistemas Espaciais - CSE.

[1]Coordenação de Rastreio, Controle e Recepção de Satélites - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

[1]Serviço de Relações Institucionais - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

ivan.barbosa@inpe.br

***Resumo.*** *Esse trabalho de pesquisa utiliza os métodos filter, wrapper e embedded para seleção de atributos em um conjunto de dados de telemetria de um satélite de coleta de dados do INPE. Ele objetiva a seleção dos melhores atributos considerando como variável dependente (y) a telemetria TM130. A metodologia utilizada foi análise bibliográfica sobre ciência de dados, estatística, matemática, aprendizado de máquina etc., a análise exploratória desses dados e a utilização dos algoritmos SelectKBest, SelectFromModel, Recursive Feature Elimination, Sequential Feature Selection e LassoCV. Conclui-se que o método wrapper é o método mais custoso do ponto de vista computacional, que o método embedded selecionou 23 atributos dentre os 135 atributos possíveis com um tempo de execução muito bom e que 07 telemetrias (TM114, TM117, TM118, TM128, TM129, TM131 e TM132) foram selecionadas pelos três métodos de seleção de atributos com diferentes tempos de execução.*

**Palavras-chave:** Seleção de atributos; Aprendizado de máquina, Telemetria, Satélite.

## 1. Introdução

A cada dia são criados e instalados dispositivos, sensores e equipamentos que geram uma vasta quantidade de dados. Esses dados podem ser categóricos (nominal ou ordinal) ou numéricos (discreto ou contínuo).

Os dados possuem fontes distintas (satélites, *web pages*, *tweets*, planilhas Excel, textos, imagens etc.), formatos distintos (binário, texto, csv etc.), tipo de dados distintos (*string*, *float*, inteiro, objeto, data, *NaN* etc.) e tamanhos que podem variar de alguns *bytes* a *gigabytes*.

Os dados de um sistema espacial se enquadram em três categorias básicas: *house keeping*, *atitude* e carga útil. Os dados (temperatura, pressão, corrente, tensão, estado das redundâncias etc.) de *house keeping*, às vezes conhecidos como dados de parâmetros de engenharia, precisam ser monitorados para verificar a saúde e o status operacional dos equipamentos a bordo da espaçonave. Os dados de atitude são de uma variedade de sensores

de sol, da Terra e de estrelas, giroscópios e acelerômetros (FILLERY; STANTON, 2003). Os dados de carga útil são recebidos pela estação terrena de recepção, armazenados, processados e disseminados para a comunidade científica e tecnológica.

A quantidade de dados de telemetria (*house keeping*) gerados diariamente é muito grande e tende a crescer proporcionalmente ao tempo de vida útil da missão espacial e do número de instrumentos a bordo dela. A análise exploratória dos dados e a análise temporal realizada manualmente tem alto impacto financeiro nas atividades de operação de solo, é mais propensa a erros, reduz a significativamente a possibilidade de encontrar novos *insights*, padrões ainda não conhecidos, associações em grandes volumes de dados, possui baixa eficiência e baixa eficácia.

Como o volume e a variedade dos dados crescem exponencialmente é necessária a utilização da Ciência de Dados que, segundo (BOSCHETTI; MASSARON, 2016), é um domínio de conhecimento relativamente novo que requer a integração bem-sucedida de álgebra linear, modelagem estatística, visualização, linguagem de programação, análise de gráficos, aprendizado de máquina, inteligência de negócios, armazenamento e a recuperação de dados.

O termo *machine learning* (do Inglês, aprendizado de máquina) foi cunhado pelo americano Arthur Lee Samuel (1959) que o definiu como "a capacidade de aprender sem ser explicitamente programada".

O aprendizado de máquina é um subcampo da ciência da computação que visa obter *insights* a partir dos dados históricos. Algoritmos, modelos e ferramentas de aprendizado de máquina podem ser aplicados em diferentes áreas do conhecimento como, por exemplo, na engenharia, na educação, na administração, nos negócios, na área médica, em vendas, em sistemas espaciais etc.

Considerando que os dados de telemetria do satélite de coleta de dados do INPE possuem alta dimensionalidade com milhões de amostras e 136 variáveis e que isso pode produzir resultados redundantes, provocar *overfitting* no modelo de aprendizado de máquina, exigir excessivo tempo de processamento e complexa análise dos dados etc., faz se necessário a redução dos dados.

Segundo (CUESTA; KUMAR, 2016), a dimensionalidade de um modelo é o número de atributos independentes no conjunto de dados. Para reduzir a complexidade do modelo, precisamos reduzir a dimensionalidade sem sacrificar a precisão. Quando trabalhamos com dados multidimensionais complexos, precisamos selecionar os recursos que podem melhorar a precisão da técnica que estamos utilizando. Às vezes, não sabemos se as variáveis são independentes ou se compartilham algum tipo de relacionamento. Precisamos de critérios para encontrar as melhores características e reduzir o número de variáveis.

A redução de dimensionalidade é a operação de eliminar alguns atributos ou variáveis do conjunto de dados de entrada e criar um conjunto restrito de recursos que contém todas as informações necessárias para prever a variável de destino de uma maneira mais eficaz e confiável. A redução do número de variáveis geralmente também reduz a variabilidade do resultado e a complexidade do processo de aprendizado (bem como o tempo necessário) (BOSCHETTI; MASSARON, 2016).

Há diferentes técnicas para redução da dimensionalidade dos dados. Dentre essas técnicas podemos citar a *feature selection* (do Inglês, seleção de atributos) e *feature extraction* (do Inglês, extração de atributos).

De acordo com (CUESTA; KUMAR, 2016), a seleção de atributos possibilita a seleção de um subconjunto de variáveis para obter melhores tempos durante a fase de treinamento ou para melhorar a precisão do modelo. Na análise de dados, encontrar as melhores variáveis para o nosso problema geralmente é guiado pela intuição e não sabemos o valor real de uma variável até testá-la. No entanto, podemos utilizar métricas como correlação e informações mútuas, o que pode nos ajudar, fornecendo distância entre as variáveis. O coeficiente de correlação é uma medida de quão forte é a relação entre duas variáveis e informação mútua refere-se a uma medida de quanto uma variável informa sobre outra.

Segundo (CUESTA; KUMAR, 2016), a extração de atributos é uma forma especial de redução de dimensionalidade realizada por uma transformação de um espaço de alta dimensão (conjunto de dados multivariado), para obter um espaço de menor dimensão (as que são mais informativas). A extração de variável é amplamente utilizada no processamento de imagens, visão computacional e mineração de dados.

A seleção de atributos objetiva:

- ✓ A simplificação do modelo;
- ✓ Evitar a maldição da dimensionalidade;
- ✓ A redução da variância e, consequentemente, do *overfitting*;
- ✓ A redução do tempo de execução;
- ✓ Melhor performance do modelo de aprendizado de máquina.

Há basicamente três tipos de métodos que podem ser utilizados na seleção de atributos de um conjunto de dados. Os 3 métodos são:

1. Métodos filtro

Os métodos de seleção de atributos através de filtros aplicam uma medida estatística para atribuir uma pontuação a cada atributo. Os atributos são classificados pela pontuação e selecionados para serem mantidos ou removidos do conjunto de dados. Os métodos são frequentemente univariados e consideram o atributo de forma independente ou em relação à variável dependente. Alguns exemplos de métodos filtro incluem o teste de Qui-quadrado, ganho de informação e coeficiente de correlação (BROWNLEE, 2020a).

Alguns algoritmos de seleção de atributos que utilizam o método filtro são: *Univariate feature selection* e *SelectFromModel*.

A *Univariate feature selection* (do Inglês, seleção univariada de atributos) seleciona os melhores atributos com base em testes estatísticos univariados. Pode ser implementado através dos métodos de transformação *SelectKBest* (remove todos, exceto os atributos de maior pontuação), *SelectPercentile* remove todos os atributos, exceto a porcentagem de pontuação mais alta especificada pelo usuário), e *GenericUnivariateSelect* (permite realizar a seleção univariada de atributos com uma estratégia configurável) (SCIKIT-LEARN DEVELOPERS, 2020).

*SelectFromModel* é um meta-transformador que pode ser utilizado junto com qualquer estimador que atribua importância a cada atributo por meio de um atributo específico (como *coef_*, *feature_importances_*) ou por meio de um *importance_getter* que pode ser chamado após o *fit*. Os recursos são considerados sem importância e removidos se a importância correspondente dos valores do atributo estiver abaixo do parâmetro de limite fornecido. (SCIKIT-LEARN DEVELOPERS, 2020).

2.Métodos *Wrapper*

Os métodos *Wrapper* (do Inglês, invólucro, embalagem) consideram a seleção de um conjunto de atributos como um problema de busca, onde diferentes combinações são preparadas, avaliadas e comparadas com outras combinações. Um modelo preditivo é utilizado para avaliar uma combinação de atributos e fornecer uma pontuação com base na precisão do modelo (BROWNLEE, 2020a).

Alguns algoritmos de seleção de atributos que utilizam o método *wrapper* são: (*Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS) Forward Selection, Backward Elimination, Stepwise selection*.

O *Recursive Feature Elimination* (do Inglês, eliminação recursiva de atributos) é um algoritmo de seleção de atributo do tipo wrapper. Isso significa que um algoritmo de aprendizado de máquina diferente é fornecido e utilizado no núcleo do método, é empacotado pelo *RFE* e utilizado para selecionar atributos. Isso contrasta com as seleções de atributos com base no método filtros que pontua cada atributo e seleciona os atributos com a maior (ou menor) pontuação (BROWNLEE, 2020b).

O *RFE* descreve uma abordagem incremental para a seleção de atributos onde atributos são gradualmente eliminados de um modelo mais inclusivo, criando um modelo menos inclusivo até o número mínimo de atributos a serem incluídos (NIELSEN, 2020).

O *Sequential Feature Selection* (do Inglês, seleção de atributo sequencial) pode ser executado para frente ou para trás.

O *Forward-SFS* é um algoritmo guloso que encontra iterativamente o melhor novo atributo para adicionar ao conjunto de atributos já selecionados. Inicia-se com nenhum atributo e procura aquele atributo que com maior pontuação durante a validação cruzada. Uma vez que o primeiro atributo é selecionado, repete-se o procedimento adicionando um novo atributo ao conjunto de atributos já selecionados (SCIKIT-LEARN DEVELOPERS, 2020).

O *Backward-SFS* inicia com todos os atributos e remove avidamente os atributos do conjunto de dados. O parâmetro *direction* controla se o SFS inicia para frente ou para trás (SCIKIT-LEARN DEVELOPERS, 2020).

3. Método *Embedded*

Os métodos *embedded* (do Inglês, incorporado) aprendem quais atributos contribuem melhor para a precisão do modelo enquanto o modelo está sendo criado. O tipo mais comum de método *embedded* de seleção de atributos são os métodos de regularização.

Os métodos de regularização também chamados de métodos de penalização, introduzem restrições adicionais na otimização de um algoritmo preditivo (como um algoritmo de

regressão) que influencia o modelo em direção à complexidade inferior (menos coeficientes) (BROWNLEE, 2020a).

Alguns algoritmos de seleção de atributos que utilizam o método *embedded* são: *Lasso*, *Elastic Net* e *Ridge Regression*.

### 3.1 *Lasso* e *Ridge*

O L1 (também chamado de *Lasso*) reduz alguns coeficientes a zero, tornando seus coeficientes esparsos (MASSARON; MUELLER, 2015).

O L2 (também chamado de *Ridge*) reduz os coeficientes dos atributos mais problemáticos, tornando-os menores, mas nunca iguais a zero. Todos os coeficientes continuam participando da estimativa, mas muitos se tornam pequenos e irrelevantes (MASSARON; MUELLER, 2015).

A regularização L2 reduz o impacto de atributos correlacionados, enquanto a regularização L1 tende a selecioná-los. Uma boa estratégia é misturá-los utilizando uma soma ponderada com a classe *Elastic Net* (MASSARON; MUELLER, 2015).

## 2. Metodologia

A metodologia utilizada nesse trabalho de pesquisa foi a leitura e análise bibliográfica sobre ciência de dados, estatística, matemática e aprendizado de máquina. Também foi realizada leituras de artigos científicos sobre aprendizado de máquina e inteligência artificial aplicados à área espacial, estudos sobre a linguagem de programação Python e estudos das bibliotecas pandas, numpy, scikit-learn, matplotlib, plotly etc.

Após a análise bibliográfica, foi feita a aquisição do conjunto de dados de telemetria de um satélite de coleta de dados do INPE e foi realizada a preparação (limpeza e transformação) desses dados para possibilitar a realização de experimentos com diferentes métodos e algoritmos de seleção de atributos. Cabe ressaltar que esse trabalho de pesquisa não contempla a modelagem e a validação de modelos preditivos ou classificatórios de aprendizado de máquina.

Esse trabalho de pesquisa utiliza os métodos *filter*, *wrapper* e *embedded* para seleção de atributos em um conjunto de dados de telemetria de um satélite de coleta de dados do INPE. Ele objetiva a seleção dos melhores atributos considerando como variável dependente ($y$) a telemetria TM130.

## 3. Resultados e Discussão

Os três métodos (*filter*, *wrapper* e *embedded*) foram utilizados na seleção de atributos com distintos algoritmos. Foi utilizada como variável dependente ($y$) a telemetria TM130 que possui a descrição "*Solar Sensor 1 temperature monitor*" e valores aceitáveis entre 5°C a 26°C e acurácia de ± 1°C. A Figura 1 ilustra a média, o desvio padrão e os pontos fora da curva da variável dependente ($y$) TM130.

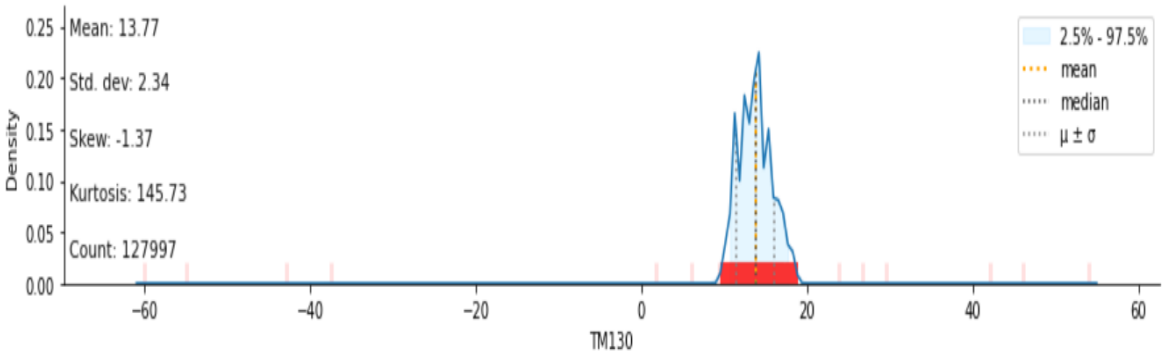


**Figura 1. Média, desvio padrão e pontos fora da curva da variável dependente (*y*) TM130.**

**Fonte: Próprio autor**

A Figura 2 ilustra como os dados da variável dependente (y) TM130 estão distribuídos. Nessa Figura é possível observar que em janeiro de 2018 há muitos dados fora dos valores aceitáveis (5°C a 26°C e acurácia de ± 1°C). O que poderia ter causado essa grande quantidade de *outliers* durante esse período? Essa é uma questão que deve ser investigada e não está no escopo desse trabalho de pesquisa.

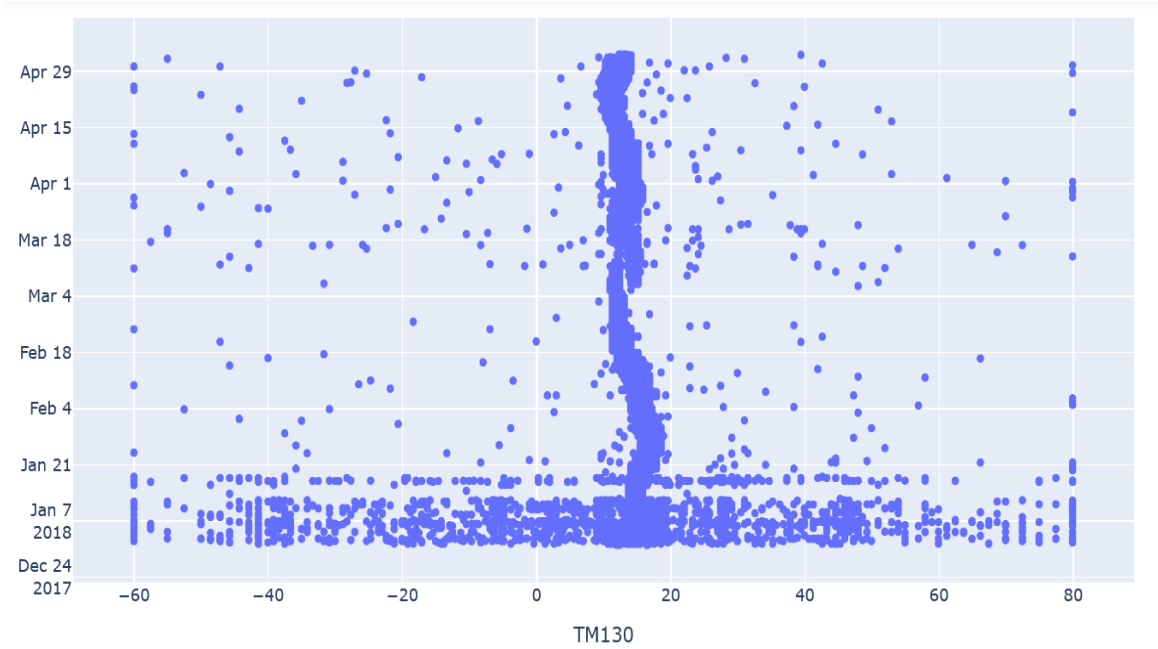


**Figura 2. Distribuição dos dados da variável dependente (*y*) TM130.**

**Fonte: Próprio autor**

Os métodos de seleção de atributos contidos na Tabela 1 foram executados em um conjunto de dados com 133 atributos e 127.997 amostras (de 2018-01-01 00:08:08.581000 até 2018-

05-03 05:24:24.657000) em um notebook Dell Latitude com processado Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50 GHz e 08GB de memória RAM.

Além de considerar o método e o algoritmo de seleção de atributos é importante também considerar o valor de k (número de variáveis que serão selecionadas) e o tempo despendido na execução dos algoritmos. A Tabela 1 mostra os métodos utilizados, o tempo de execução e os atributos (telemetrias) que foram selecionadas durante o processo de seleção de atributos.

**Tabela 1. Seleção de atributos.**

**Fonte: Próprio autor**

| Variável dependente (y) | Método | Estimador | k | Tempo de execução | Atributos selecionados |
|---|---|---|---|---|---|
| **TM130** | ***Univariate Feature Selection*** | ***SelectKBest*** | **10** | **6min 25s** | **'TM114', 'TM117', 'TM118', 'TM119', 'TM124', 'TM128', 'TM129', 'TM131', 'TM132', 'TM133'** |
| **TM130** | ***SelectFromModel*** | ***Random Forest Regressor*** | **10** | **7min 42s** | **'TM114', 'TM115', 'TM116', 'TM117', 'TM118', 'TM119', 'TM128', 'TM129', 'TM131', 'TM132'** |

(Continua)

**Tabela 2. Continuação**

| | | | | | |
|---|---|---|---|---|---|
| **TM130** | ***Recursive Feature Elimination (Wrapper)*** | ***Random Forest Regressor*** | **10** | **11h 53min 20s** | **'TM114', 'TM115', 'TM117', 'TM118', 'TM119', 'TM128', 'TM129', 'TM131', 'TM132', 'TM134'** |
| **TM130** | ***Sequential Feature Selection (Wrapper)*** | ***Random Forest Regressor*** | **10** | **2d 5h 38min 5s** | **'TM114', 'TM116', 'TM117', 'TM118', 'TM119', 'TM128', 'TM129', 'TM131', 'TM132', 'TM133'** |
| **TM130** | ***LassoCV (Embedded)*** | ***LassoCV*** | **-** | **7.14s** | **'TM027', 'TM075', 'TM107', 'TM113', 'TM114', 'TM116', 'TM117', 'TM118', TM119', 'TM120', 'TM121', TM122', 'TM126', 'TM127', TM128', 'TM129', 'TM131', TM132', 'TM134', 'TM135', TM138', 'TM141', 'TM142'** |

Na Tabela 1 é possível observar que o método *wrapper* é o método mais custoso do ponto de vista computacional e que o método *embedded* é o método com menor custo do ponto de vista computacional.

Observa-se também que as telemetrias TM114 (*MGE temperature monitor*), TM117 (*TR2 temperature monitor*), TM118 (*DEC temperature monitor*), TM128 (*DCC temperature monitor*), TM129 (*RDU temperature monitor*), TM131 (*SS2 temperature monitor*), TM132 (*CP inf. temperature monitor*) foram selecionadas pelos métodos filtro, *wrapper* e *embedded*.

## 4. Conclusão

A seleção de atributos é fundamental para a criação de um bom modelo preditivo de aprendizagem de máquina. No entanto é importante considerar o número de variáveis independentes que serão selecionadas bem como o custo computacional dessa importante fase do desenvolvido de modelos preditivos de aprendizado de máquina.

Para o conjunto de dados de telemetria de um satélite de coleta de dados tendo com variável dependente ($y$) a TM130, concluímos que o método *embedded* selecionou 23 atributos dentre os 135 atributos possíveis com um tempo de execução muito bom e que 07 telemetrias (TM114, TM117, TM118, TM128, TM129, TM131 e TM132 foram selecionadas pelos três (*filter*, *wrapper embedded*) métodos de seleção de atributos com diferentes tempos de execução.

A seleção de atributos contribui para o aumento da qualidade e para a melhoria do desempenho computacional dos modelos de aprendizado de máquina, contribui com melhorias na operação de satélites e, consequentemente, contribui para o aumento da vida útil da missão espacial.

## Referências

BOSCHETTI, A.; MASSARON, L. **Python Data Science Essentials**. Second ed. Birmingham B3 2PB, UK: Packt Publishing Ltd., 2016. 361 p. ISBN(978-1-78646-213-8).

BROWNLEE, J. **An Introduction to Feature Selection**. Disponível em: <https://machinelearningmastery.com/an-introduction-to-feature-selection/>. Acesso em: 3 jun. 2021a.

BROWNLEE, J. **Recursive Feature Elimination (RFE) for Feature Selection in Python**. Disponível em: <https://machinelearningmastery.com/rfe-feature-selection-in-python/>. Acesso em: 3 jun. 2021b.

CUESTA, H.; KUMAR, D. S. **Practical Data Analysis**. Second ed. Birmingham B3 2PB, UK: Packt Publishing Ltd., 2016. 316 p. ISBN(978-1-78528-971-2).

FILLERY, N. P.; STANTON, D. TELEMETRY, COMMAND, DATA HANDLING AND PROCESSING. In: FORTESCUE, P.; STARK, J.; SWINERD, G. (Eds.). **SPACECRAFT SYSTEMS ENGINEERING**. Third Edit ed. West Sussex, England: John Wiley & Sons, Ltd, 2003. p. 678.

MASSARON, L.; MUELLER, J. P. **Python for Data Science For Dummies**. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2015. 418 p. ISBN(978-1-118-84418-2).

NIELSEN, A. **Practical Time Series Analysis Prediction with Statistics and Machine Learning**. First Edit ed. Sebastopol, CA, USA: O'Reilly Media, Inc., 2020. 480 p. ISBN(978-1-492-04165-8).

SCIKIT-LEARN DEVELOPERS. **1.13. Feature selection**. Disponível em: <https://scikit-learn.org/stable/modules/feature_selection.html>. Acesso em: 4 jun. 2021.