MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

ANÁLISE DA CORRELAÇÃO ENTRE ÍNDICE DE CINTILAÇÃO E O CONTEÚDO ELETRONICO TOTAL DA IONOSFERA NO PICO DA ANOMALIA MAGNÉTICA.

Pedro Alexandre dos Santos

Monografia para Exame de Qualificação do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Stephan Stephany, aprovada em 11 de dezembro de 2018.

URL do documento original:
<http://urlib.net/xx/yy>

INPE
São José dos Campos
2018

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3945-6923/6921
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

ANÁLISE DA CORRELAÇÃO ENTRE ÍNDICE DE CINTILAÇÃO E O CONTEÚDO ELETRONICO TOTAL DA IONOSFERA NO PICO DA ANOMALIA MAGNÉTICA.

Pedro Alexandre dos Santos

Monografia para Exame de Qualificação do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Stephan Stephany, aprovada em 11 de dezembro de 2018.

URL do documento original:
<http://urlib.net/xx/yy>

INPE
São José dos Campos
2018

Informar aqui sobre marca registrada (a modificação desta linha deve ser feita no arquivo publicacao.tex).

RESUMO

A ionosfera é uma camada da atmosfera que se estende de aproximadamente 60 km a 100 km de altitude. Esta camada influi nos sinais de radiofrequência transmitidos por satélites para a superfície terrestre, sendo composta por gases ionizados principalmente pela radiação solar e elétrons livres. Variabilidades no fluxo solar geram alterações nos campos elétrico e magnético no espaço e, consequentemente, no campo magnético da Terra causando flutuações na ionização e, portanto, na quantidade de elétrons livres na ionosfera, alterando a transmissão de sinais de radiofrequência. Dentre as várias pertubações ionosféricas, há a anomalia magnética equatorial, que consiste na formação de uma região com alta densidade de elétrons ente 15 e 20 graus magnéticos ao norte e sul do equador. Entretanto, essa anomalia é mais significativa no Brasil, especificamente no Vale do Paraíba, estado de São Paulo. Os sinais de radiofrequência dos sistemas de navegação por satélites (GNSS - Global Navigation Satellite System) são afetados por flutuações da ionização, que constituem o fenômeno de cintilação ionosférica, que pode ser medidas pelo índices S4. A cintilação afeta sinais GNSS, especialmente no pico da anomalia, afetando a navegação aérea e outras atividades humanas que dependem de sinais recebidos de satélites. Por outro lado, a quantidade de elétrons livres na ionosfera pode ser medida pelo conteúdo eletrônico total vertical (VTEC), sendo que regiões com baixos valores de VTEC em relação à sua vizinhança, caracterizam as denominadas de bolhas ionosféricas, associadas às cintilações. Dada a existência de redes de estações de medição que provém valores de S4 e VTEC, este trabalho busca correlacionar os valores destas variáveis, bem como analisar sua evolução espaço-temporal por meio de técnicas de mineração e visualização de dados, considerando como estudo de caso a cidade de São José dos Campos.

Palavras-chave: Cintilação Ionosférica. Bolha Ionosférica. Índice S4. VTEC. Mineração de dados. GNSS. Vale do Paraíba.

ESCREVER O TÍTULO EM INGLÊS PARA PUBLICAÇÕES ESCRITAS EM PORTUGUÊS E EM PORTUGUÊS PARA PUBLICAÇÕES ESCRITAS EM INGLÊS

ABSTRACT

The ionosphere is a atmosphere layer which extends from about 60 Km to 1,000 Km altitude. This layer influences in the radio frequency signals transmitted by satellites to the terrestrial surface. It is composed of ionized gases and free electrons generated by the solar radiation. Variations in the solar flux generates electrical and magnetic field fluctuations in the space, and consequently, in the magnetic field of Earth causing perturbations in the ionization, and therefore, in the quantity of free electrons in the ionosphere, changing the transmission of the radio frequency signals. Among the various ionospheric disturbances, there is the equatorial anomaly, which consists in the formation of a high density electrons region between 15 and 20 degrees north and south of the magnetic equator. This anomaly is more significant in Brazil, particularly in the Vale do Paraíba, in the state of São Paulo. The radio frequency signals from Global Navigation Satellite System (GNSS) are affected by ionization fluctuations, which constitute the ionospheric scintillation phenomenon, that can be measured by the S4 indices. Scintillation affects GNSS signals, especially at the peak of the anomaly, affecting air navigation and other human activities that rely on signals received from satellites. On the other hand, the amount of free electrons in the ionosphere can be measured by the vertical total electronic content (VTEC), and regions with low VTEC values in relation to their neighborhood, characterize the so called ionospheric bubbles associated with scintillation. Given the existence of measuring stations that provided values of S4 and VTEC, this works seeks to correlate the values of these variables, as well as to analyze their spatial-temporal evolution through mining techniques and data visualization, considering as a case study the city of São José dos Campos.

Keywords: Ionospheric Scintillation. S4 Index. Ionospheric Bubble. VTEC. Data Mining. GNSS. Vale do Paraíba.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

GPS	–	Sistema de Posicionamento Global
OACI	–	Organização da Aviação Civil Internacional
GNSS	–	Sistema de Navegação Global por Satélite
EIA	–	Anomalia da Ionização Equatorial
VTEC	–	Conteúdo eletrônico total vertical
AACGM	–	Altitude Adjusted Corrected Geomagnetic Coordinates
KDD	–	Descoberta de Conhecimento em Base de dados
CART	–	Classification and Regression Trees
SVM	–	Support Vector Machine

SUMÁRIO

Pág.

1 IONOSFERA

A ionosfera é uma região ionizada da alta atmosfera, estendendo-se de 60 até 10.000 km de altitude, assim, engloba partes da mesosfera, termosfera, e exosfera. Esta camada constitui-se de íons e elétrons livres criados primariamente por processo de fotoionização, e alguma porção de gás neutro. A fotoionização ionosférica consiste de um processo físico-químico, onde alguma espécies químicas presentes na atmosfera ganham ou perdem elétrons decorrentes da absorção de radiação solar predominantemente nas faixas mais altas do ultravioleta e raios-X (Rishbeth; Garriott, 1969; NEGRETI, 2012). A ionização, também, pode ocorrer devido a colisões com partículas altamente energéticas, provindas do meio solar ou galácticas, o que é mais facilmente observado em altas latitudes, e na região da Anomalia Magnética do Atlântico Sul.

A ionosfera pode ser divida em regiões, faixas de altitudes, as quais se diferenciam pelos processos físicos e químicos que governam o comportamento daquela faixa, além disso, este mesmos processos podem variar devido a quantidade de radiação solar recebida, logo, vai-se observar diferenças entre a noite e o dia. Ao longo da noite, a camada F é a única que apresenta uma ionização significativa, enquanto as camadas E e D apresentam um valor extremamente baixo de ionização. Durante o dia, a camada D e E se tornam mais ionizadas, assim como a camada F, que se divide em duas regiões, F1 que é mais fracamente ionizada, e F2 que é mais intensamente ionizada. A camada F2 existe durante a noite e durante o dia, sendo a principal responsável pela reflexão e refração dos sinais de rádio.

A camada D é a mais interna, estando entre 60 e 90 km acima da superfície da Terra. Sua ionização é devido a radiação do hidrogênio ionizado na série de Lyman-alpha no comprimento de onda de 121.6nm ionizando o óxido nítrico, NO, presente na camada. Além disso, raios X altamente energéticos, com comprimento de onda inferior de 1 nm podem ionizar as moléculas de N_2 e de O_2. A camada D apresenta alta taxa de recombinação, de modo que existem mais moléculas neutras do que íons. Tem uma taxa de absorção considerável para ondas de radio de média e alta baixa frequências, e baixas frequências apresentam elevada atenuação, principalmente, devido a absorção de energia pelo elétrons livres, o que aumenta suas chances de colisão. Este efeito desaparece durante à noite, devido a uma menor ionização. Pode apresentar valores elevados de ionização em altas latitudes em decorrência de erupções solares com grandes quantidades de matéria hadrônica, prótons, em sua maioria, com uma duração de 24 à 48 horas.

A camada E é a intermediária e está situada entre 90 e 150 km acima da superfície da Terra. A ionização decorre principalmente devido ao espalhamento de raio-X leve (entre 1 e 10 nm) e ultravioleta distante (UV) provindos do Sol com moléculas de oxigênio. A estrutura vertical da camada E é determinada em sua maior parte pela competição entre efeitos de ionização e de recombinação. É importante pela presença de correntes elétricas que nela fluem e interagem com o campo magnético (KIRCHHOFF, 1991). A noite, a camada E quase desaparece, pois sua fonte primaria de ionização não está presente.

A camada F se estende de 150 a mais de 500 km acima da superfície da Terra. Apresenta a maior concentração de elétrons, portanto, sinais que são capazes de penetrar até esta subcamada são capazes de escapar para o espaço. Predominam, nesta, a ionização de átomos de oxigênio por meio de radiação solar no espectro do extremo ultravioleta, entre, 10 e 100 nm. A camada é subdivida em duas regiões, a F2 que está presente durante o dia e a noite, e a F1 que aparece somente durante o dia.

A subcamada F2 se inicia aproximadamente a 300 km de altitude, englobando toda a região superior da ionosfera, inclusive a região de pico da densidade de elétrons. Este máximo no perfil vertical de ionização decorre do balanço entre os processos de transporte de plasma e os processos físico-químicos. Acima deste pico, a ionosfera se encontra em equilíbrio difusivo, ou seja, o plasma se distribui com a sua própria escala de altura. A presença do campo magnético contribui para a distribuição da ionização.

1.1 Anomalias na ionosfera

A ionosfera apresenta várias anomalias, ou seja, várias irregularidades na distribuição de elétrons. Este trabalho tem interesse na anomalia equatorial. Esta aparece aproximadamente entre 15 e 20 graus de latitude magnética, tanto no hemisfério norte, quanto no hemisfério sul, na camada F2. Consiste na formação de uma região de alta densidade eletrônica, e é uma anomalia, pois a densidade de plasma deveria ser maior em regiões equatoriais, e não em latitudes magnéticas mais altas.

Sua origem decorre da deriva vertical do plasma da camada F na região equatorial: o processo de ionização da camada F faz surgir um campo elétrico, apontando para leste, enquanto o campo magnético aponta para o norte, considerando então $\vec{E} \times \vec{B}$, tem-se o surgimento de uma força perpendicular ao campo magnético e ao campo elétrico, o que neste caso, aponta para cima, deslocando o plasmas para regiões de

mais alta altitudes. Agora, quando em altas altitudes, o plasma for efeito gravitacional e diferença de pressão é trazido de volta à altitudes mais baixas, porém este movimento de descida é mais eficiente ao longo das linhas de campo magnético, levando a um aumento na densidade de plasma em regiões de médias latitudes.

A distribuição do plasma também pode ser alterada pela ação de outras variáveis, como o vento. O Vale do Paraíba, no estado de São Paulo, encontra-se na região da anomalia equatorial, mais especificamente no pico da anomalia, ou seja, na região onde a densidade de plasma, em altas altitudes, atinge seu valor máximo.

As anomalias que surgem na ionosfera apresentam um certo nível de organização devido às linhas de campo magnético da Terra. Isto ocorre pois partículas ionizadas ou carregas podem ser mover livremente ao longo das linhas magnéticas mas não entre elas. Assim, o estudo do campo magnético da Terra se faz relevante para um grande número de aplicações, (LAUNDAL; RICHMOND, 2017). Um resultado imediato deste estudo é que o norte geográfico e magnético não coincidem, e que simplificadamente o campo magnético poderia ser descrito por um dipolo magnético, com centro comum ao da Terra, porém inclinado em relação a linha que liga o norte e sul geográfico. Atualmente, existe vários sistemas de coordenadas magnéticas cujo proposito dependem da região, da aplicação e da faixa de altitude de interesse, para uma revisão entre os sistemas mais comuns consulte a referência (LAUNDAL; RICHMOND, 2017). Para este trabalho foi adotado o sistema AACGM, pois é mais adequado à altura ionosférica, contudo ele pertence a classe de sistemas não-ortogonais.

A cintilação ionosférica é uma variação rápida de amplitude e fase em sinais de radio frequência quando estes atravessam irregularidades no plasma ionosférico, como uma bolha de plasma, que é de particular interesse para este trabalho. Neste trabalho se realiza um estudo sobre algumas variáveis relacionadas com o efeito de cintilação ionosférica e com a estrutura de bolha no plasma.

1.2 Algumas variáveis importantes para o estudo da cintilação ionosférica

Existem várias quantidades que são necessárias para um estudo completo da dinâmica ionosférica como, por exemplo, medidas do fluxo de radiação solar, do campo magnético da Terra, da composição da atmosfera. Contudo, este trabalho foca em duas quantidades principais o VTEC e o índice S4, pois são as que fornecem mais informação à respeito do fenômeno de cintilação ionosférica, e as anomalias que a causam.

O Total Eletronic Content (TEC), ou conteúdo total de elétrons em português, é uma quantidade descritiva utilizada para avaliar a densidade de elétrons no plasma ionosférico. É o número total de elétrons integrado ao longo da trajetória entre um transmissor, no espaço, e um receptor, na Terra, em uma seção unitária (HOFMANN-WELLENHOF et al., 2013). Por sua definição o TEC é uma quantidade que depende da trajetória, seu cálculo fica mais claro, ao dizer que é a integral de um densidade de elétrons dependente de posição ao longo de uma trajetória que atravessa a ionosfera:

$$\mathrm{TEC} = \int n_e(s)ds, \tag{1.1}$$

onde ds especifica o elemento de integração na trajetória. Geralmente é reportada em unidades de TEC (TECU), definido por 1 $TECU = 10^{16}$ elétrons/m^2. É importante para determinar a cintilação e os atrasos de fase e de grupo em ondas de rádio no meio. VTEC, ou conteúdo total de elétrons vertical é projeção do TEC, ao longo de uma linha normal a superfície da Terra, em outras palavras, ela fornece o conteúdo de elétrons ao longo da normal para cada ponto da superfície da Terra.

O índice S4 é utilizado para medir, avaliar, a cintilação ionosférica. Corresponde ao desvio padrão da intensidade do sinal de GPS de um minuto de dados, coletados com 50 amostras por segundo:

$$S_4^2 = \frac{\langle I^2 \rangle - \langle I \rangle^2}{\langle I \rangle^2}. \tag{1.2}$$

1.3 Bolhas de Plasma

As bolhas ionosféricas podem ser definidas como regiões de baixa densidade de plasma ionosférico quando comparadas com a sua vizinhança. Utilizando medidas de VTEC é possível definir essa diferença como 30-50 TECU (TAKAHASHI et al., 2016).

São originadas na região equatorial, após a rápida elevação do plasma, devido a anomalia equatorial, isto é, o plasma ao acender cria regiões de baixa densidade. Após sua formação podem evoluir para altas altitudes (centenas de quilômetros), estendendo-se ao longo das linhas de campo magnético (milhares de quilômetros) nas direções norte-sul, alcançado em torno de 20 graus de latitude magnética.

2 MINERAÇÃO DE DADOS

A mineração de dados é uma área da ciência da computação que permite inferir conhecimento a partir de uma massa de dados específica de um fenômeno ou evento qualquer.

Em particular, a mineração de dados contempla a inferência de modelos direcionados por dados, ou seja, modelos derivados unicamente a partir de conjuntos de dados de entrada e de saída específicos de um dado fenômeno ou evento. Utilizam tipicamente um enfoque denominado aprendizado de máquina (machine learning), que basicamente é uma forma de estatística aplicada com enfase no uso de computadores para estimar funções, em geral, complicadas em um conjunto de dados.

A mineração de dados é frequentemente confundida com o processo mais amplo de descoberta de conhecimento em base de dados (KDD- Knowledge Discovery in Databases), a qual é definida como a extração de padrões e desenvolvimento de representações associados ao conhecimento de um processo ou fenômeno a partir de um conjunto de dados associados. A mineração de dados é apenas uma etapa deste processo, sendo definida como a extração de padrões em um conjunto de dados. As diferenças ficam mais claras as sumarizando as etapas contidas em um processo de KDD:

a) **Limpeza de dados**: remoção de dados com ruído, inconsistentes e incompletos;

b) **Integração de dados**: combinação de múltiplas fontes de dados, por meio de operações como união e intersecção de tabelas;

c) **Seleção de dados**: os dados considerados relevantes ao processo são extraídos do banco de dados;

d) **Transformação de dados**: os dados são transformados e consolidados em uma forma mais apropriada para mineração, sendo por exemplo discretizados, normalizados, agrupados;

e) **Mineração de dados**: é a etapa do KDD associada com a aplicação de algoritmos estatísticos, de reconhecimento de padrões ou de inteligência computacional, como por exemplo, redes neurais, árvores de decisão, visando a extração de padrões, juntamente com a avaliação dos modelos desenvolvidos, por exemplo, medir o desempenho de um sistema de classificação;

f) **Apresentação do conhecimento**: técnicas de visualização e representação dos dados são usadas para exibir o conhecimento extraído aos usuários.

O processo de KDD é iterativo, como pode ser visto na Figura 2.1 no sentido em que etapas podem ser revistas e reexecutadas em função dos resultados obtidos.

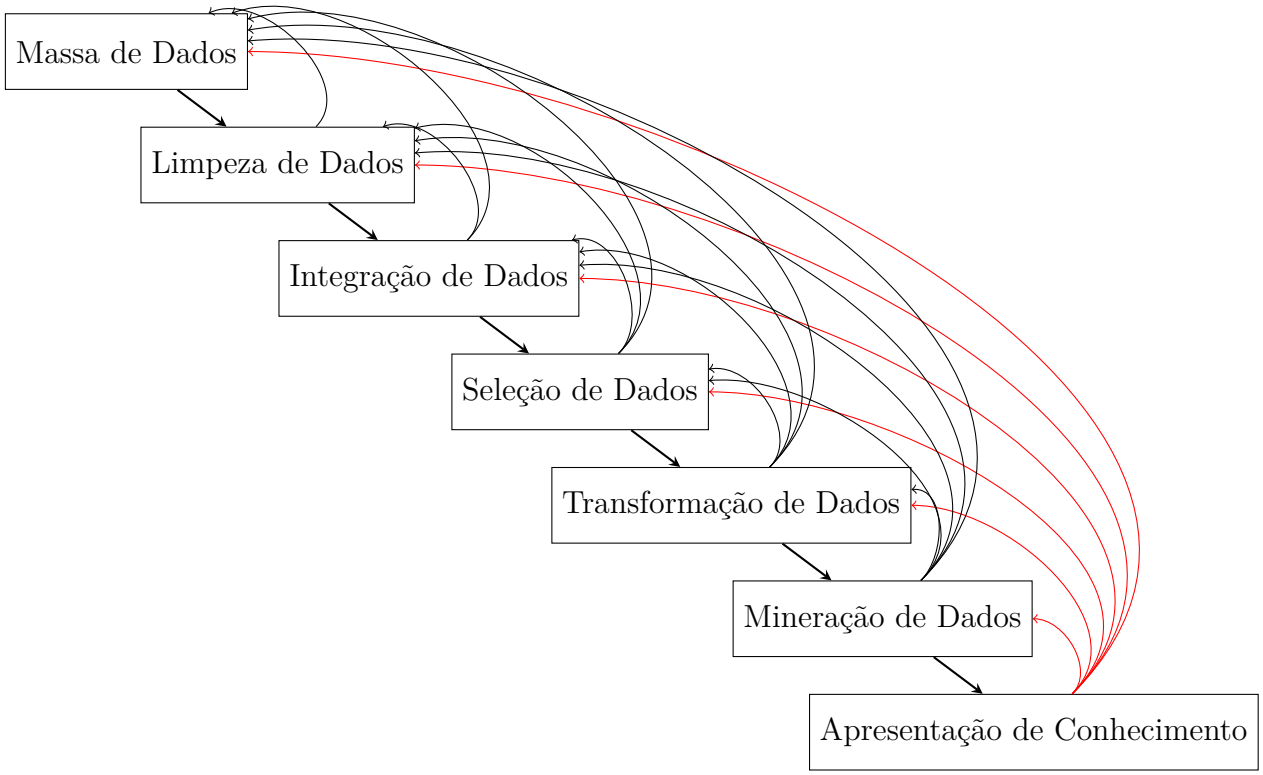

Figura 2.1 - Passos de um processo de KDD. Fonte: Próprio Autor.

2.1 Aprendizagem de máquina

A aprendizagem de máquina faz uso de algoritmos que são capazes de aprender a partir dos dados, que são ainda agrupados pelo termo **algoritmo de aprendizado de máquina**. Note que esta definição é incompleta, uma vez que não se definiu o que significa aprender a partir dos dados. Assim, considere uma tupla (P, D, M) onde T corresponde à tarefa, D ao conjunto de dados (domínio) e M a métrica, diz-se que um algoritmo aprende com relação ao problema P, ao conjunto de dados D e a métrica M, se a sua performance no tratamento do problema P, avaliado segundo a métrica M melhora conforme itera sobre os dados.

Basicamente, pode ser visto como o processo de estimar (ajustar) uma função segundo algum objetivo (métrica) a um conjunto de dados. Também é adequado definir o termo função, portanto, sejam X e Y dois conjuntos não vazios, que podem ou não ser iguais, e uma regra, ou conjunto de regras, f que assina cada elemento x de X a um único elemento y em Y, um **função** consiste de uma tupla formada por (X, Y, f). Assim, termos como modelo, estimador, preditor são sinônimos do termo função, pois são entendidos como uma realização do algoritmo de aprendizagem para um conjunto de dados, ou seja, um conjunto de regras inferidas dos dados pelo algoritmo.

Um problema P pode ser descrito pela maneira trata um amostra $(\mathbf{x}, y)$ ou $(\mathbf{x})$ do domínio D, isto é, o que ela faz ou como processa uma instância. Esta, por sua vez, corresponde a um conjunto de características, atributos ou variáveis que foram quantificadas de algum objeto ou evento que se deseja estudar. Quando uma amostra é da forma $(\mathbf{x}, y)$, diz-se que a instância apresenta um rótulo ou resposta indicado por y, enquanto amostras da forma $(\mathbf{x})$ são ditas não rotuladas. O termo $\mathbf{x}$ também pode ser denominado de atributos de informação, variáveis preditoras, entres outras, enquanto o termo y como como atributo de decisão. Os dados também podem ser numéricos, quantitativos, ou qualitativos. Estes últimos são dados categóricos, que correspondem à atribuição de rótulos que os qualificam.

Uma definição formal de um problema seria um tanto trabalhosa e foge do escopo deste trabalho, entretanto, é possível iniciar a construção uma imagem mental deste por meio de exemplos:

- Um problema de classificação: neste tipo de problema o objetivo é determinar uma função f que mapeia cada amostra $\mathbf{x}$ a uma categoria j, em um conjunto de n categorias $\{0, .., n-1\}$;
- Um problema de regressão: neste tipo de problema o objetivo é determinar uma função f que dado uma instância $\mathbf{x}$ prediga um valor numérico.

Um segundo aspecto com relação ao problema se refere a maneira pela qual os dados são tratados pelo algoritmos. Neste contexto, os algoritmos de aprendizado de máquina são agrupados basicamente em dois grandes grupos: os supervisionados e os não-supervisionados. Na última década apareceram variações, tais como, algoritmos auto-supervisionados, ou semi-supervisionados, que não são abordados aqui. No caso dos supervisionados, um algoritmo é ajustado na fase de treinamento a partir de instâncias conhecidas, ou seja, de atributos preditores e os correspondentes atributos

respostas. Estes algoritmos são utilizados para tarefas como classificação, predição, regressão. Por outro lado, os algoritmos não-supervisionados buscam inferir conhecimento ao agrupar os dados em conjuntos distintos (sem o conhecimento prévio de seus rótulos) ou ao reduzir a dimensionalidade dos dados, ou ao apresentá-los numa forma distinta ou ao extrair padrões, de forma a permitir uma melhor análise do fenômeno ou evento de estudo.

O grande desafio das técnicas de aprendizagem de máquina é que estas precisam apresentar bom desempenho em amostras de dados diferentes das usadas no treinamento. Assim, define-se o conceito de generalização que é a capacidade de apresentar um bom desempenho em amostras de dados não observadas previamente. Usualmente, antes da fase de treinamento de um algoritmo de aprendizagem de máquina, o domínio D é particionado em dois subconjuntos, um de treinamento e um de validação. O subconjunto de treinamento é utilizado na fase de ajuste do modelo, isto é, o treinamento, e a ele por meio de uma métrica, pode-se associar um erro de treinamento, que basicamente, mensura quão bens o modelo se ajusta aos dados deste subconjunto. Note que o erro de treinamento pode não fornecer um indicativo de quão bem o modelo se ajusta a dados não observados. Neste ponto, a aprendizagem de máquina começa a ser separar das abordagens usuais de ajuste de funções, pois define o erro de generalização, ou erro de validação, que deve fornecer informações sobre como o modelo se comportar para novas amostras.

O erro de generalização é avaliado mensurando o desempenho do modelo no subconjunto de validação. Em um primeiro analise, não fica claro como é possível a performance no subconjunto de teste apenas observando os dados de treinamento. Assim, neste ponto, faz-se necessário adotar alguns pressupostos a respeito dos dados: as amostras são independentes entre si; o subconjunto de treinamento e o de teste são identicamente distribuídos, e amostrados da mesma distribuição de probabilidade. Logo, pode-se concluir que o valor estimado de ambos os erros devem ser iguais. Todavia, em um problema real de aprendizagem de máquina tais pressupostos não são completamente verdadeiros, e diversas abordagens são adotadas de forma a melhor se aproximar destes. Portanto, em geral, espera-se que o erro esperado de validação seja maior ou igual ao de treino.

Finalmente, determina-se quão bem um algoritmo desempenha seu papel em relação a um conjunto de dados, avaliando sua habilidade de reduzir o erro de treinamento, mantendo a diferença entre o erro de treinamento e o de validação pequenos. Esta habilidade leva a duas dificuldades extremas opostas, o sobreajuste (overfitting) e

sobajuste (underfitting). O primeiro ocorre quando a diferença entre os erros de treinamento e validação divergem; o segundo ocorre quando o modelo não é capaz de reduzir suficientemente o erro de treinamento.

A possibilidade de um modelo sofrer de sobreajuste ou sobajuste pode ser quantificada, teoricamente, em termos de sua capacidade, inclusive a tendência de um ou outro pode ser controlada modificando a capacidade do modelo. De uma maneira informal, pode-se dizer que a capacidade está associado ajustar a habilidade do modelo se ajustar a um grande número de funções. Dois pontos práticos podem ser obtidos dessa construção, modelos de baixa capacidade podem não se ajustar de maneira adequada ao subconjunto de treinamento, enquanto modelos com grande capacidade podem se ajustar de maneira excessiva memorizando o conjunto de treinamento, e assim, não generalizando. Estas colocações devem levantar a necessidade de uma escolha adequada de algoritmo e ajuste de seus parâmetros (ver-se-á alguns exemplos posteriormente) frente aos dados a serem tratados.

2.2 Métricas

As métricas podem ser agrupadas segundo a abordagem do problema, isto é, existem métricas que são adequadas para problemas de classificação, existem métricas que são adequadas para problemas de regressão, e analogamente para diversos outros problemas em mineração de dados e aprendizado de máquina.

2.2.1 Classificação

Primeiramente, é necessário definir o conceito de matriz de confusão, uma vez que a partir dela diversas outras métricas são mais facilmente definidas.

Admita um problema de classificação com n classes, então a matriz de confusão será uma tabela(matriz) onde cada linha representa instâncias em uma classe predita, enquanto cada coluna representa instâncias em uma classe verdadeira, ou inverso. Assim, dado a i-ésima linha e a j-ésima coluna com x elementos, diz-se que x elementos da j-ésima classe foram preditos como da i-ésima classe. A Tabela 2.1 ilustra esta configuração, e por exemplo, pode-se dizer que h elementos da Classe 1 foram previstos como sendo da Classe 2, com a segunda coluna pela terceira linha.

Enquanto na representação inversa (transposta) diz se que x elementos da i-ésima classe foram preditos como da j-ésima classe. A Tabela 2.2 apresenta esta configuração, e se pode dizer, por exemplo, que h elementos da Classe 1 foram previstos como sendo da Classe 2, com a segunda linha pela terceira coluna.

		Verdadeiro			
		Classe 0	Classe 1	Classe 2	Total
Predito	Classe 0	a	b	c	$a+b+c$
	Classe 1	d	e	f	$d+e+f$
	Classe 2	g	h	i	$g+h+i$
	Total	$a+d+g$	$b+e+h$	$c+f+i$	N

Tabela 2.1 - Exemplo de matriz de confusão, onde as colunas estão associadas as instâncias verdadeiras e as linhas as instâncias preditas. Fonte: próprio autor.

		Predito			
		Classe 0	Classe 1	Classe 2	Total
Verdadeiro	Classe 0	a	d	g	$a+d+g$
	Classe 1	b	e	h	$b+e+h$
	Classe 2	c	f	i	$c+f+i$
	Total	$a+b+c$	$d+e+f$	$g+h+i$	N

Tabela 2.2 - Exemplo de matriz de confusão, onde as colunas estão associadas as instâncias preditas e as linhas as instâncias verdadeiras. Fonte: próprio autor.

Uma vez que ambas as representações são equivalentes, isto é, vão fornecer a mesma informação, a escolha por uma ou outra, ficar a cargo do usuário quando responde peguntas como: qual tenho mais afinidade? Qual acho mais simples? Neste, trabalho adotar-se-á a matriz de confusão com o verdadeiro correspondendo as linhas, e o predito as colunas, 2.2.

Em posse da matriz de confusão, pode-se definir:

a) **Verdadeiro Positivo (VP):** é o número de elementos que pertencendo à classe i foram corretamente classificados como i. Para a Classe 0 é $VP_0 = a$. Generalizando, para uma classe i qualquer:

$$VP_i = A_{ii}; \tag{2.1}$$

b) **Falso Positivo (FP):** é o número de elementos que não pertencem à classe i, mas que foram classificados como pertencendo à i. Para a Classe 0 é $FP_0 = b + c$. Generalizando, para uma classe i qualquer:

$$FP_i = \sum_{j=0, j\neq i}^{n-1} A_{ji}; \tag{2.2}$$

c) **Verdadeiro Negativo (VN):** é o número de elementos que não pertencem à classe i e foram classificados como não pertencendo à i. Para a Classe 0 é $VN_0 = e + h + i + f$. Generalizando, para uma classe i qualquer:

$$VN_i = \sum_{j=l=0, j\neq i, l\neq i}^{n-1} A_{jl}; \tag{2.3}$$

d) **Falso Negativo (FN):** é o número de elementos que pertencendo à classe i, foram erroneamente classificados como não pertencendo à i. Para a Classe 1 é $FN_0 = d + g$. Generalizando, para uma classe i qualquer:

$$FN_i = \sum_{j=0, j\neq i}^{n-1} A_{ij}; \tag{2.4}$$

onde na expressões (2.1)-(2.4), A é a matriz de confusão, n é o número de classes, $i, j, l \in 0, ..., n-1$, e os exemplos são dados com relação a Tabela 2.2.

Utilizando as definições de Verdadeiro Positivo e suas correlatas é possível estabelecer um enorme conjunto de métricas, como por exemplo, acurácia, precisão, especificidade entre outras. Neste texto, apresentar-se-á a definição das duas primeiras, assim como o porquê da adoção de uma em relação a outra.

A **acurárcia** (acc) é uma das métricas mais tradicionais para o tratamento de problemas de classificação. E é definida como o número total de predições corretas sobre o número total de elementos. Seja A uma matriz de confusão associada a um problema com n classes, então a acurácia é dada por:

$$acc = \frac{\sum_{i=0}^{n-1} A_{ii}}{\sum_{i=0, j=0}^{n-1} A_{ij}}. \tag{2.5}$$

Agora, considere um problema com 3 classes, e que o número de elementos da Classe 1 e 2 combinados seja x. Admita ainda que o classificador acerte todos as instâncias da Classe 0, porém erre todas as demais duas classes, e que a Classe 0 sozinha tenha x elementos, então a acurácia será 1/2 ou 0,5. Considere agora que a Classe 0 tenha $2x$ instâncias, neste caso a acurácia será 2/3 ou 0,6. Um aumento no número de elementos da Classe 0 leva um aumento da acurácia, entretanto as Classes 1 e 2 são erroneamente classificadas, portanto neste caso, a acurácia não é capaz de avaliar o classificador.

Conjuntos de dados onde o número de elementos por classes são diferentes por ordens de grandezas, como o exemplo estabelecido no parágrafo anterior. são ditos não balanceados e acurácia como definida é incapaz de avaliar classificadores desenvolvido em cima destes dados. Portanto, é necessário definir uma nova métrica, neste caso a precisão, assim seja A a matriz de confusão já apresentada anteriormente, a precisão por classe é dada por:

$$precision_i = \frac{VP_i}{VP_i + FP_i}. \tag{2.6}$$

A precisão para uma dada classe por ser entendida como a proporção de resultados verdadeiros que são definitivamente verdadeiros. E a precisão fica dada por uma média aritmética sobre a precisão de cada classe. Esta é a métrica adotada nos problemas de classificação discutidos nesta proposta.

2.2.2 Regressão

O **erro quadrático médio** (mse) é definido por

$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.7}$$

onde n é o número de instâncias a serem preditas (avaliadas), y_i a variável alvo para amostra i e $\hat{y}_i$ o valor predito para a amostra i. Esta medida é uma das mais aplicadas, não somente no contexto, de problemas de regressão, mas também nos problemas de otimização em aprendizagem de máquina que necessitem de funções diferenciáveis. Por ser definida em termos de uma operação de média, existem distribuições dos dados a serem preditos para os quais essa métrica é incapaz de avaliar de maneira satisfatória o desempenho do regressor (distribuições altamente localizadas) em uma analogia com o problema de classificação com classes não balanceadas.

Deste modo faz necessário avaliar uma métrica que seja capaz de mensurar de maneira mais adequada erros em regiões menos localizadas da distribuição, para tal adotou-se o **erro absoluto máximo** (mae) definido por:

$$mae = \max|y_i - \hat{y}_i|, \tag{2.8}$$

onde i vária entre 1 e n, com este sendo o úmero de instâncias a serem preditas e y_i

e $\hat{y}_i$ sendo respectivamente a variável alvo e o valor predito para a amostra i.

2.3 Algoritmo: Árvores

Métodos baseados em árvores particionam os espaço de atributos preditores em um conjunto de retângulos, e então aproxima cada um destes por uma simples função, por exemplo, uma constante. Existem diferentes algoritmos baseados em árvores, CART, C4.5, C4, neste texto discutir-se-á a CART.

Considere um problema de regressão com duas variáveis preditoras x_1 e x_2 e uma resposta contínua y, cada um destes tomando valores entre $[0, 1]$. Cada partição do espaço de atributos modela y por uma constante diferente. Se as linhas de partições forem definidas definidas de maneira arbitrária, apesar de apresentarem simples descrições, tal como $x_2 = c$, a região resultante pode apresentar uma descrição de difícil tratamento. Assim, é interessante restringir a maneira pela o qual as partições são criadas. A limitação mais natural a ser adotada é que as partições sejam definidas recursivamente dividindo o espaço em duas sub-regiões (recursão binária), em outras palavras, inicialmente o espaço de atributos é divido em duas regiões, neste caso, adota-se modelar a resposta y pelo valor médio em cada região. Então, uma destas regiões ou ambas são divididas em mais duas regiões. Tal procedimento é aplicado até que alguma condição de parada seja alcançada.

A divisão do espaço é realizada segundo algum critério, tal que se deseja obter o melhor ajuste. Finalmente, o correspondente modelo de regressão para a variável y fica definido por:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}, \tag{2.9}$$

onde R_m indica a m-ésima região, x é um elemento do espaço de atributos, c_m é a constante que aproxima y na m-ésima região, M é o número de regiões, e I é uma função identidade definida por

$$I(x \in R_m) = \begin{cases} 0 & sex \in R_m, \\ 1 & sex \in R_m. \end{cases} \tag{2.10}$$

para este exemplo, $M = 2$ e x é da forma (x_1, x_2). Um importante ganho advindo da restrição em recursão binária, é que a representação da árvore passa a ter uma

simples interpretação, onde uma simples condição em cada nó, define os subnós.

O processo de construção de uma árvore depende se o problema a ser tratado é de regressão ou classificação, o que será discutido nas próximas subseções. E alguns dos elementos da estrutura árvore binárias discutidas apresentam nomes, e condições sobre elas que são

- um nó pode ser de dois tipos, **decisão**, neste caso fica associado um subconjunto, e uma regra, que particiona esta em duas novas sub-regiões, ou **folha**;

- um nó somente pode dividir uma região em duas sub-regiões ou em nenhuma, no primeiro caso diz-se que ele é um pai, que gera dois filhos, e cada filho está associado a um ramo, no segundo caso ele é uma folha;

- folha é um nó associado ao final da árvore, portanto o conjunto associado a ele não pode ser dividido, em outras palavras, ele não pode ter filhos, e assim, a cada folha fica associado uma partição e uma constante;

2.3.1 Árvores de Regressão

Admita que o conjunto de dados apresente p atributos preditores e um atributo resposta, e um total de N instâncias, ou seja, é da forma $(\mathbf{x_i}, \mathbf{y_i})|\mathbf{i} \in \{\mathbf{1}, ..., \mathbf{N}\}$, com $\mathbf{x_i} = (\mathbf{x_{i,1}}, \mathbf{x_{i,2}}, ..., \mathbf{x_{i,p}})$. O papel do algoritmo é operar sobre este conjunto de dados e automaticamente decidir em qual variável dividir, qual o ponto (valor) para dividir, e a forma que a árvore deverá assumir. Considere inicialmente que o espaço esteja particionado em M regiões $R_1, ..., R_M$, e que a reposta seja ajustada por uma constante c_m em cada região, (2.9). Ir-se-á adotar como critério para a divisão minimização do erro quadrático médio (2.7), isto é, o mse em cada sub-região, traduzido na expressão

$$min \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2. \tag{2.11}$$

Avaliando a expressão (2.11) com relação aos paramêtros c_m, isto é, derivando a expressão com relação a c_m e igualando a zero, obtém-se

$$\sum_{i=1}^{N}(y_i - f(x_i))frac\partial(f(x_i))\partial c_l = \quad 0 \tag{2.12}$$

$$\sum_{i=1}^{N}(y_i - \sum_{m=1}^{M} c_m I(x \in R_m))I(x \in R_l) = \quad 0 \tag{2.13}$$

$$\sum_{i=1}^{N}(y_i)I(x \in R_l) - \sum_{i=1}^{N}\sum_{m=1}^{M} c_m I(x \in R_m)I(x \in R_l) = \quad 0 \tag{2.14}$$

$$\sum_{i=1}^{N}(y_i)I(x \in R_l) - c_l \# R_l = \quad 0, \tag{2.15}$$

ou

$$\hat{c}_l = \frac{1}{\# R_l} \sum_{i|x_i \in R_l} y_i, \tag{2.16}$$

onde $\# R_l$ indica a cardinalidade da partição R_l, isto é, o número de elementos da região R_l, finalmente o lado direito da expressão (2.16) corresponde a média sobre os valores de y_i que estão na região R_l. Adotar-se-á a seguinte notação para simplificar a expressão (2.16) $\hat{c}_m = \mu(y_i|x_i \in R_l)$.

Encontrar a melhor partição que minimiza o mse de maneira geral como definido acima é um problema quase intratável computacional, entretanto, é possível adotar-se uma abordagem gulosa, que lida com um problema semelhante. Comece com o conjunto de todos os dados, considere dividir este utilizando a variável j das p variáveis, em um ponto s dos N pontos, definindo-se assim duas regiões

$$R_1(j, s) = \{x \,|\, x_j \leq s\} \tag{2.17}$$

$$R_2(j, s) = \{x \,|\, x_j > s\}. \tag{2.18}$$

Então, procura-se a variável j e ponto s que resolvem o problema de minimização

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]. \tag{2.19}$$

Note entretanto que já foi resolvido um problema similar para a minimização da parte interna $(y_i - c_1)$ para qualquer que seja j e s:

$$\hat{c}_1 = \mu(y_i|x_i \in R_1(j,s)) \quad \text{e} \quad \hat{c}_2 = \mu(y_i|x_i \in R_2(j,s)). \tag{2.20}$$

Assim, para cada variável de divisão j, a determinação do ponto de divisão s pode ser realizada rapidamente, varrendo todos os possíveis valores, por exemplo, tal que encontra o melhor par (j, s) para cada região seja tratável. Uma vez encontrado este par, particionasse a região em duas sub-regiões, e aplica-se o processo para cada uma das duas sub-regiões. Este processo é repetido para cada nova sub-região até que algum critério de parada seja alcançado.

O tamanho da árvore governa a complexidade do modelo, e um valor ótimo deve ser escolhido segundo os dados. Assim, existem diferentes critérios para dividir ou não um nó, por exemplo, divida o nó somente se o decréscimo no mse foi maior que um determinado valor.

2.3.2 Árvores de Decisão

No caso da variável resposta estar associada a um problema de classificação, as modificações no algoritmo se resumem a modificação do critério de divisão do nó e os critérios de parada. Defini-se a medida de impureza de um nó m com relação a uma árvore T, para um problema de regressão, por

$$Q_m(T) = \frac{1}{\#R_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2, \tag{2.21}$$

note que quanto mais $Q_m(T)$ tende a 0, mais se diz que o nó é puro. Observando também a relação de impureza fica claro que ela apresenta relação fundamental com o critério de divisão de um nó, entretanto para um problema de classificação, tal relação não é adequada. Assim, é necessário definir ao menos uma medida para este tipo de problema. Seja m um nó de um árvore em um problema de classificação, representando uma região m com N_m instâncias ($N_m = \#R_m$), define-se

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m I(y_i = k)}, \tag{2.22}$$

a proporção de elementos da k-ésima classe no nó m. A classe associada ao nó m denotada por $k(m)$ é $k(m) = \max_k \hat{p}_{mk}$, ou seja, a classe que apresenta o maior número de elementos neste nó. Assim, pode-se definir algumas medidas de impureza:

- **Erro de Classificação**

$$\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}; \tag{2.23}$$

- **Indice Gini**

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}); \tag{2.24}$$

- **Entropia Cruzada**

$$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{2.25}$$

Finalmente, o critério de divisão fica definido como maximizar o nível de pureza ao dividir um nó. Considere um nó m com N_m elementos, tal que cada elemento tenha p variáveis preditoras e um variável resposta que indica a classe da amostra. Assim, de maneira gulosa, busca-se uma variável j e um ponto de divisão s tal que a soma da medida de impureza de cada sub-região $Q_{x_j \leq s}(T)$ e $Q_{x_j > s}(T)$ seja mínima, ou

$$\min_{j,s} \left[Q_{x_j \leq s}(T) + Q_{x_j > s}(T) \right]. \tag{2.26}$$

A escolha da medida de impureza depende do algoritmo, a CART utiliza o índice Gini.

2.4 Técnicas de Conjunto

2.5 Extreme Gradient Boosting: XGBOOST

A explicação aqui apresentada sobre o algoritmo XGBOOST é baseada no trabalho de (CHEN; GUESTRIN, 2016). Considere um conjunto de dados com N amostras e p atributos $\mathcal{D}(\mathbf{x}_i, y_i)$ onde $\mathbf{x} \in \mathbb{R}^N$ e $y_i \in \mathbb{R}$, um modelo baseado em um conjunto de árvores utiliza K funções aditivas (árvores) para predizer, estimar, a variável resposta:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F}, \tag{2.27}$$

onde $\mathcal{F} = \{f(\mathbf{x}) = c_{q(\mathbf{x})}\}(q : \mathbb{R}^N \rightarrow M, c \in \mathbb{R}^M)$ é o espaço de árvores de regressão (CART), q representa a estrutura de cada árvore que mapeia uma amostra na correspondente folha, M é o número de folhas na árvore. Cada função f_k corresponde a uma estrutura de árvore completa e independente q e folhas com valor c. Assim, para uma dada amostra, utiliza-se as regras de decisões nas árvores dadas por q para classificar em uma folha, e calcular o valor de predição final somando os valores nas correspondentes folhas, dadas por c. A Figura 2.2

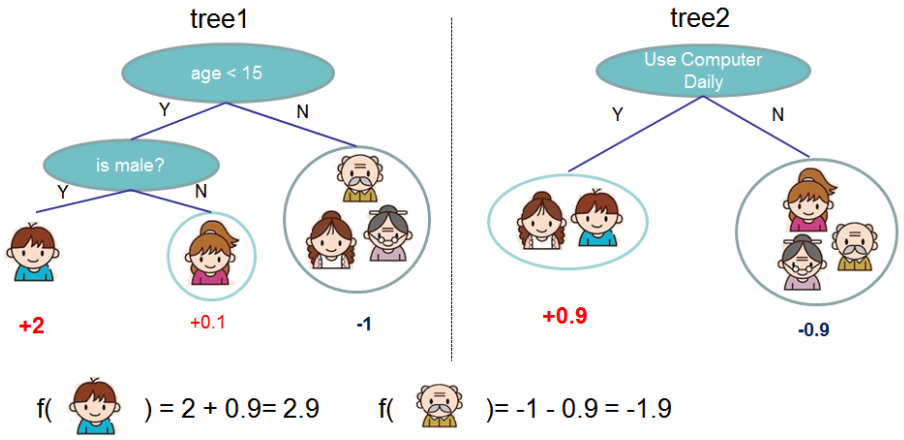

Figura 2.2 - Modelo de conjunto de árvores. O resultado final para um dada amostra é a soma das predições para cada árvore. Fonte: Adaptado de (CHEN; GUESTRIN, 2016).

O processo de aprendizado deste algoritmo, consiste em determinar as funções f_k, para tal, minimiza-se a seguinte função objetivo regularizada:

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega f_k, \tag{2.28}$$

onde

$$\Omega(f) = \lambda T + \frac{1}{2}\lambda ||\mathbf{c}||^2. \tag{2.29}$$

l é uma função de erro convexa diferenciável que mede a diferença entre o valor predito $\hat{y}_i$ e o valor reposta y_i. O segundo termo penaliza a complexidade do modelo. Sua adição suaviza os coeficientes c aprendidos, o que por sua vez reduz a proba-

bilidade de sobreajuste. Note que quando $\gamma = 0$, o algoritmo se torna boosting de árvores com gradiente tradicional.

O modelo de conjunto de árvores na equação (2.27) inclui funções não analíticas, e o problema de minimização (2.28) incluí funções com parâmetros e, portanto, não pode ser minimizado tradicionalmente utilizando técnicas de otimização em espaços Euclideanos. Assim, uma abordagem aditiva, análoga ao método guloso adotado na construção de árvores de regressão (CART), é utilizado. Seja, $\hat{y}_i^{(t)}$ a predição da i-ésima instância na t-ésima interação, adiciona-se f_t de forma a minimizar o seguinte objetivo

$$\mathcal{L}^t = \sum_{1}^{K} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t). \tag{2.30}$$

Utilizando-se de uma aproximação de segunda ordem para o termo $\mathcal{L}^t$:

$$\mathcal{L}^t \approx \sum_{1}^{K} \left[l(y_i, \hat{y}_i^{(t-1)}) + \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}^{(t-1)}} f_t(\mathbf{x}_i) + \frac{1}{2} \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}^{(t-1)^2}} f_t^2(\mathbf{x}_i) \right] + \Omega(f_t), \tag{2.31}$$

cuja a notação pode ser simplificada definindo:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}^{(t-1)}} \quad \text{e} \quad h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}^{(t-1)^2}}, \tag{2.32}$$

que são respectivamente os gradientes de primeira e segunda ordem para a função de erro. Descartando termos constantes e utilizando as simplificações definidas em (2.32), obtém-se a seguinte função objetivo simplificada no tempo t

$$\tilde{\mathcal{L}}^t = \sum_{i=1}^{K} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t). \tag{2.33}$$

Expandindo a equação (2.33), usando que $f_t(x_i) = \sum_{j=1}^{T} c_j I(x_i \in R_j)$ e que $f_t(x_i)^2 = \sum_{j=1}^{T} c_j^2 I(x_i \in R_j)$:

$$\tilde{\mathcal{L}}^t = \sum_{i=1}^{K} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} c_j^2 \tag{2.34}$$

$$= \sum_{i=1}^{K} \left[g_i \sum_{j=1}^{T} c_j I(x_i \in R_j) + \frac{1}{2} h_i \sum_{j=1}^{T} c_j^2 I(x_i \in R_j) \right] + \lambda T + \frac{1}{2}\lambda \sum_{j=1}^{T} c_j^2, \tag{2.35}$$

trocando i por j e j por i no primeiro termo da expressão (2.34) e agrupando-se os termos, tem-se:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} \left[c_j \sum_{i|\mathbf{x}_i \in R_j} g_i + \frac{1}{2} c_j^2 \left(\sum_{i|\mathbf{x}_i \in R_j} h_i + \lambda \right) \right] + \gamma T. \tag{2.36}$$

Tomando a variação da equação (2.36) com relação a c_j, encontra-se que o valor ótimo de $\hat{c}_j$ para a folha j é dado porquê

$$\hat{c}_j = -\frac{\sum_{i|\mathbf{x}_i \in R_j} g_i}{\lambda + \sum_{i|\mathbf{x}_i \in R_j} h_i}, \tag{2.37}$$

e o correspondente valor ótimo por

$$\tilde{\mathcal{L}}^{(t)} == -\frac{1}{2} \sum_{j=1}^{T} \frac{\sum_{i|\mathbf{x}_i \in R_j} g_i}{\lambda + \sum_{i|\mathbf{x}_i \in R_j} h_i} + \gamma T. \tag{2.38}$$

A equação (2.38) pode ser utilizada para medir a qualidade de uma árvore. Seu valor é semelhante a uma medida de impureza, entretanto é derivada de um largo conjunto de funções objetivos. Assim, como aconteceu para a árvore de decisão, é impossível avaliar todas as possíveis árvores, por isso, um algoritmo guloso que inicia com uma única folha (se a árvore tiver apenas um único nó, o original, pela definição de nó folha, ele será uma folha) e adiciona ramos à árvore é utilizado. Considere que R_E e R_D sejam respectivamente o nó da esquerda e da direita após a divisão. Definindo $R = R_E \cup R_D$, a redução de erro (análoga à redução de impureza) após dividir os nós é dador por

$$\tilde{\mathcal{L}}_{split} = \frac{1}{2} \left[\frac{\sum_{i|\mathbf{x}_i \in R_E} g_i}{\lambda + \sum_{i|\mathbf{x}_i \in R_E} h_i} + \frac{\sum_{i|\mathbf{x}_i \in R_D} g_i}{\lambda + \sum_{i|\mathbf{x}_i \in R_D} h_i} - \frac{\sum_{i|\mathbf{x}_i \in R} g_i}{\lambda + \sum_{i|\mathbf{x}_i \in R} h_i} \right] - \gamma. \tag{2.39}$$

A fórmula (2.39) é usada na prática para avaliar os candidatos de divisão.O algoritmo de busca guloso é apresentado em 1. A Figura 2.3 apresenta a estrutura do processo de cálculo da divisão.

Algorithm 1 Algoritmo Guloso Exato de Busca de Divisão

gain $\leftarrow 0$
G $\leftarrow \sum_{i \in R} g_i$
H $\leftarrow \sum_{i \in R} h_i$
for k=1 to N **do**
    G_L $\leftarrow 0$
    H_L $\leftarrow 0$
    for j em R ordenado **do**
        G_L $\leftarrow$ G_L $+ g_j$
        H_L $\leftarrow$ H_L $+ h_j$
        G_R $\leftarrow$ G $- G_L$
        H_R $\leftarrow$ H $- H_L$
        value $\leftarrow \max(\text{value}, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda})$

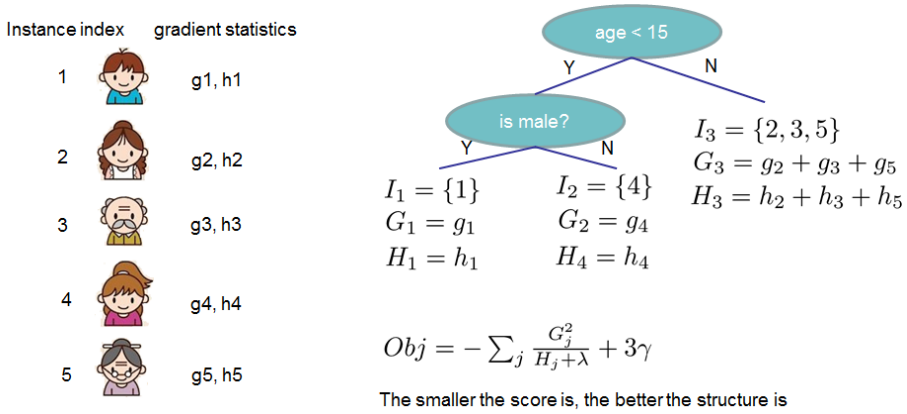

Figura 2.3 - Estrutura do cálculo de divisão.Apenas é necessário soma o gradiente e o gradiente de segunda ordem da função de erro em cada nó, então aplicar a fórmula (2.39) para obter a medida de qualidade. Fonte: Adaptado de (CHEN; GUESTRIN, 2016).

O algoritmo combina várias outras técnicas de forma a tratar o problema de sobreajuste e generalizar os resultados, entretanto foge do escopo desta proposta apresentar uma discussão sobre elas. Todavia, algumas destas serão utilizadas neste trabalho, e quando o feito, uma breve introdução a estas será realizada. Finalmente, os prin-

cipais elementos do algoritmo XGBOOST foram discutidos nesta seção.

3 REZENDE

3.1 Revisão

O trabalho (REZENDE, 2009) foi pioneiro na utilização de modelos direcionados por dados para a exploração do problema de cintilação ionosférica, entretanto este trabalho não apresenta uma abordagem completamente reprodutível, visto ausência dos códigos e da base de dados. O objetivo desta revisão é entender as dificuldades deste trabalho, assim como sua incompletude em relação a proposta a ser estabelecida nessa dissertação.

3.1.1 Principais Pontos

A previsão de cintilação de curto prazo (uma hora) foi realizada utilizando **amostras com intervalos de 5 minutos**, entretanto alguns dos atributos, como já discutido, apresentam resolução inferior, sendo portanto interpolados, ou copiados de sua vizinhança, enquanto outros apresentam resolução mais altas, e neste caso precisaram ser agrupados, segundo algum critério, por exemplo, por uma operação de máximo.

Os atributos foram:

- **Hm_Eq** representa a hora no equador (em São Luis), em intervalos de 5 minutos;
- **Vel_Der** é a velocidade máxima de deriva vertial do plasma medida no equador entre as 17LT e 19LT (20UT e 22UT), com resolução de um valor por dia;
- **Kp** é a média do índice Kp definido pela expressão:

$$\sum_{i=1}^{n}\sum_{j=1}^{i}\frac{Kp_j}{ni}, \tag{3.1}$$

  onde Kp_1 corresponde ao Kp medido entre 14-17LT, Kp_2 entre 11-14LT, até o valor medido entre 5-8LT, o valor de n é 4. Este valor apresenta resolução diária;
- **F10.7** é o fluxo solar;
- **S4_Eq** é o maior valor do índice S4 medido no equador, em um período de 5 minutos;

- **S4_PA_tempo_real** é o maior valor do índice S4 medido no pico da anomalia (São José dos Campos), em um período de 5 minutos;
- **S4_PA** é o S4 estimado com uma hora de antecedência para o pico da anomalia, com resolução de 5 minutos.

As primeiras 5 variáveis correspondem aos atributos preditores, enquanto a variável S4_PA corresponde ao atributo reposta. Os atributos acima passam por algum pré-processamento antes e depois da seleção das amostras:

- **S4**, antes da seleção, somente são utilizados os valores medidos para satélites com ângulo superior a 30 graus;
- **S4**, depois da seleção, a grande variabilidade destes dados leva a adoção de um filtro passa baixa, realizado por meio de uma suavização com média móvel, com 15 pontos;
- **Kp**, depois da seleção, somente mantidos amostras com valores de Kp inferiores a 3, pois valores maiores que este caracterizam forte pertubação magnética, associadas a eventos extremos como tempestades magnéticas, e nesta configuração a predição se torna inviável;
- **Dados**, depois da seleção, devem compreender o período entre as 18-23LT (21-01UT) de forma a predizer os valores no intervalo 19-24LT (22-02UT).

Ao final, a base de dados deste trabalho apresentava um total de 80 dias de dados com 4680 amostras, coletadas entre 2000 e 2002. Este trabalho também realizou testes com predições com 1 dia de antecedência, entretanto não apresentaramm um resultado tão significativo e, portanto, não serão discutidos.

Restam definir dois elementos para que o problema fica completamente definido, as métricas, e os modelos. Uma vez que as métricas ficam restringidas segundo os modelos, ir-se-á estabelecer estes primeiros:

- agrupamento por expectation-maximization implementado no ambiente Weka;
- regras de associação utilizando o algoritmo apriori, onde neste caso os atributos foram discretizados;

- regressão, utilizando árvores CART com a estratégia de ensemble bagging, implementadas pelo autor, análogo à Floresta Randômica.

As métricas foram erro quadrático médio, e índice de correlação de Pearson para o problema de regressão, e inspeção para as demais. A aplicação de agrupamento permitiu concluir que se tratava de um problema altamente não linear, as regras de associação geram conclusões que já eram bem conhecidas da literatura do problema na área de aeronomia. Finalmente, os resultados mais interessantes foram estabelecidos pelo problema de regressão, com erro quadrático médio de 0.05 com correlação de Pearson de 0.985.

O trabalho concluí se apresentando como uma abordagem inédita para o problema da predição da cintilação ionosférica.

3.1.2 Análise

A análise consistiu em levantar e sintetizar alguns pontos que levam a uma definição parcialmente incompleta do problema, ou talvez errônea do problema:

a) uma vez que o código e a base de dados não é disponibilizada de maneira pública e nem devidamente documentada, a reprodução somente é possível em partes, visto que serão utilizados algoritmos que muito se assemelham, mas que devido a diferença devem levar a resultados diferentes;

b) do ponto de vista de implementação não foi definido uma representação para a variável **Hm_Eq** que pode ser então representada em segundos, minutos, entre outras opções, observar entretanto que isto não deveria levar a diferença significativas no resultado final;

c) a variável Kp é medida por padrão nos intervalos 00-03UT, 03-06UT, 06-09UT, 09-12UT, 12-15UT, 15-18UT, 18-21UT e 21-24UT, e não nos intervalos utilizados por, sendo portanto necessário definir como é feito este mapeamento, o que não está presente no texto;

d) a definição de como a média móvel é aplicada na quantidade S4 está um tanto incompleta, isto é, dado o i-ésimo elemento de um vetor, com um tamanho de janela de 15 pontos, ela poderia ser aplicada levando-se em consideração: os 14 pontos anteriores, $\{i-14, i-13, ..., i-1, i\}$; ou os 7 pontos anteriores e 7 posteriores, $\{i-7, ..., i-1, i, i+1, ..., i+7\}$, denominada de forma central; entre outras combinações;

e) levando em consideração que a forma centrada tenha sido adotada, a janela de 15 pontos exigirá que 7 pontos do futuro sejam conhecidos, neste caso, dado um instante t seriam necessários 35 minutos de dados a frete deste para o cálculo da média e, portanto, na verdade, o resultado não seria previsto com uma hora de antecedência, mas sim 25 minutos;

f) a métrica, erro quadrático médio, pode não contemplar o problema. Para entender tal proposição, considere a adaptação deste problema para uma classificação, ficará evidente que eventos com altos valores de cintilação são mais raros e, portanto, ter-se iá, um problema de classificação não balanceado. Retornando, a regressão, pode-se ocorrer do modelo predizer muito bem valores baixos, que então irão mascarar os efeitos de erros em valores mais altos, visto que irão predominar no processo de cálculo de média.

3.2 Reprodução

3.2.1 Original

O trabalho (REZENDE, 2009) foi parcialmente, reproduzido no contexto desta proposta, pois somente uma linha de pesquisa do original foi explorado, o problema de regressão, e cuidados adicionais foram necessários, devido a utilização de mais anos. As variáveis adotadas são:

- **ut** representa a hora em São Luiz em minutos, em intervalos de 5 minutos;
- **vhf** é a velocidade máxima de deriva vertical do plasma medida em São Luiz entre as 17LT e 18LT (20UT e 21UT), com resolução de um valor por dia;
- **ap** é a média do índice ap definido por:

$$\sum_{i=1}^{n}\sum_{j=1}^{i}\frac{ap_j}{ni}, \tag{3.2}$$

  onde ap_1 corresponde à ap15_18ut, ap_2 à ap12_15ut até ap00_03ut, mais ap21_00ut e ap18_21ut do dia anterior, totalizando um intervalo de 24 horas, com $n = 8$. Este valor apresenta resolução diária;,

- **F10.7** é o fluxo solar;

- **s4_sl** é o maior valor do índice S4 medido em São Luiz, em um período de 5 minutos;

- **s4_sj** é o maior valor do índice S4 medido em São José dos Campos, em um período de 5 minutos;

- **s4_sj_shift_1h** é o S4 estimado com uma hora de antecedência para o pico da anomalia, com resolução de 5 minutos.

E os cuidados, assim, como os pré-processamentos foram:

- A adoção do termo São Luis em relação a equador magnético foi preferida, pois esse esta em movimento, e ao longo do período coletado, assim como extensões deste período ele não estará no mesmo lugar, enquanto os dados são sempre coletados em uma estação fixa em São Luiz;

- A adoção do termo São José dos Campos em relação a pico da anomalia ocorre, pois a localização do pico depende da quantidade de radiação emitida pelo Sol em seu regime, ciclo solar. Nos anos de 2000, 2001, 2002, o pico da anomalia estava em São José dos Campos, porém nos anos de 2018 e 2019 se encontra em Presidente Prudente. Finalmente, os dados foram sempre coletados na estação em São José dos Campos;

- A altura hF não é amostrada em intervalos regulares ao longo do período de dados coletados, inicialmente, ela era amostrada em 15 min, e posteriormente passou a ser amostrado em 10 min, portanto neste trabalho se reamostrou toda a série para o intervalo de 10 min, que foi então interpolado por um spline de grau 3, até um máximo de x pontos ausentes na vizinhança do ponto a ser interpolado;

- Quanto a variável S4, primeiro, somente serão aceitas medidas cuja elevação entre a estação e o receptor sejam maiores que 30 graus; segundo, os dados de cintilação apresentam resolução temporal de 1 min, e são coletados para cada satélite acima do plano de horizonte da estação, portanto, existem vários dados por minuto, com objetivo de ficar-se somente com um dado, estes foram agrupados tomando-se o maior valor de cintilação; os dados com resolução de um minuto são interpolados por spline de ordem 3 com limite de quatro valores ausentes; quarto, os dados são suavizados por um filtro de Savitz-Goley de ordem 3 com janela de tamanho 5, o que

INTENSIDADE	S_4
Saturado	$S_4 > 1,0$
Forte	$0,6 \leq S_4 \leq 1,0$
Moderado	$0,4 \leq S_4 \leq 0,6$
Fraco	$0,2 \leq S_4 \leq 0,4$
Ausente	$S_4 \leq 0,2$

levaria a necessidade de apenas, 2 amostras de dados futuros; finalmente, os dados são reamostrados para um intervalo de 5 minutos;

- A adoção de *ap* em preferência a *kp* se deve a esta apresentar uma escala linear e, portanto ser mais condizente com operações como média.

Uma abordagem de normalização para o intervalo $[0,05,\ 0,95]$ é aplicado a todos os atributos preditores (uma para cada atributo) antes da utilização do algoritmo de aprendizagem de máquina. Este intervalo é adotado já prevendo a possibilidade de aplicação de redes neurais com função de ativação sigmoide, haja visto que esta sofre saturação para valores próximos de 0 e de 1.

Resta definir dois elementos para a serem definidos, o tipo de problema tratado e por consequência os algoritmos e as métricas a serem utilizados. Uma que vez que se trata de uma reprodução parcial, o problema de regressão foi tratado, neste caso utilizando a ferramente XGBOOST, que também consiste da utilização de árvores de decisão e regressão com algoritmos de ensemble, neste caso o boosting; as métricas adotadas foram o erro quadrático médio e o erro absoluto máximo, este último sendo capaz de lidar melhor com o desbalanceamento das amostras.

O segundo problema tratado foi de classificação, e para tal a variável **s4__sj__-shift__1h** foi discretizada utilizando a proposta estabelecida por (MUELLA, 2008), mais a adição de uma classe ausente:

Neste caso, tem-se um problema com 5 classes como já mencionado não balanceado e uma abordagem de reasmotragem foi empregada de modo balancear o números de elementos tal que todos estejam próximos da cardinalidade da classe com o maior número de elementos. Esta etapa é realizada após a normalização e o algoritmo empregado foi o ADASYN. Finalmente, adotou-se como métrica a precisão balanceada, e a ferramenta utilizada também foi o XGBOOST.

Para ambos os problemas uma busca randômica é empregado para encontrar os

melhores valores para os hiper-parâmetros.

3.2.2 Resultados

REFERÊNCIAS BIBLIOGRÁFICAS

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>. vii, 17, 18, 21

HOFMANN-WELLENHOF, B.; LICHTENEGGER, H.; COLLINS, J. **Global Positioning System: Theory and Practice**. Springer Vienna, 2013. ISBN 9783709133118. Disponível em: <https://books.google.com.br/books?id=bQntCAAAQBAJ>. 4

KIRCHHOFF, V. W. J. H. **Introdução à geofísica espacial**. [S.l.: s.n.], 1991. 152 p. ISBN 9788572330015. 2

LAUNDAL, K. M.; RICHMOND, A. D. Magnetic coordinate systems. **Space Science Reviews**, v. 206, n. 1, p. 27–59, Mar 2017. ISSN 1572-9672. Disponível em: <https://doi.org/10.1007/s11214-016-0275-y>. 3

MUELLA, M. T. de A. H. **Morfologia e dinâmica das irregularidades ionosféricas de pequena escala e imageamento ionosférico por GPS**. 383 p. INPE-15365-TDI/1393. Tese (Doutorado em Geofísica Espacial) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2008. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m18@80/2008/09.25.12.42>. 28

NEGRETI, P. M. S. **Estudo do conteúdo eletrônico total na região brasileira em períodos magneticamente perturbados**. 323 p. Sid.inpe.br/mtc-m19/2012/05.10.21.43-TDI. Tese (Doutorado em Geofísica Espacial/Ciências do Ambiente Solar-Terrestre) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2012. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/05.10.21.43>. 1

REZENDE, L. F. **Mineração de dados aplicada à análise e predição de cintilação ionosférica**. 176 p. (INPE-16080-TDI/1537). Dissertação (Mestrado em Compuatação Aplicada) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2009. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m18@80/2009/06.22.15.52>. 23, 26

Rishbeth, H.; Garriott, O. K. **Introduction to ionospheric physics**. [S.l.: s.n.], 1969. 1

TAKAHASHI, H.; WRASSE, C. M.; DENARDINI, C. M.; PáDUA, M. B.; PAULA, E. R.; COSTA, S. M. A.; OTSUKA, Y.; SHIOKAWA, K.; MONICO, J. F. G.; IVO, A.; SANT'ANNA, N. Ionospheric tec weather map over south america. **Space Weather**, v. 14, n. 11, p. 937–949, 2016. Disponível em: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016SW001474>. 4