Avaliação Inicial e Pré-processamento de Grandes Conjuntos de Dados: Revelação de Perspetivas nas Condições Meteorológicas e Consumo de Energia

Nuno Gomes
MEIC- Mineração de Dados em Larga Escala
Instituto Superior Engenharia de Lisboa
Lisboa, Portugal
E-mail: A18364@alunos.isel.ipl.pt

Ricardo Ramos
MEIC- Mineração de Dados em Larga Escala
Instituto Superior Engenharia de Lisboa
Lisboa, Portugal
E-mail: A46638@alunos.isel.ipl.pt

Rafael Carvalho
MEIC- Mineração de Dados em Larga Escala
Instituto Superior Engenharia de Lisboa
Lisboa, Portugal
E-mail: A47663@alunos.isel.ipl.pt

Resumo - O objetivo deste trabalho é demonstrar claramente a compreensão do problema de mineração de dados, selecionando e aplicando soluções adequadas ao longo do desenvolvimento do modelo de aprendizagem. Nesta primeira fase, procura-se descrever o problema específico de mineração de dados e enquadrá-lo no contexto geral desta área. Além disto, é caracterizado detalhadamente o conjunto de dados utilizado, explicando os critérios seguidos para sua construção. São indicados os problemas encontrados durante o pré-processamento dos dados e as soluções adotadas.

I. INTRODUÇÃO

Perante os grandes conjuntos de dados abarcando condições atmosféricas e consumos energéticos, o objetivo é extrair informações valiosas e padrões úteis. Foi realizada uma avaliação inicial para observar os conjuntos de dados, com o intuito de refinar e gerar hipóteses para a extração de informações e padrões que possam ser conclusivos em relação a um ou mais problemas formulados. Por exemplo, determinar um classificador de zonas residenciais ou industriais com base no consumo energético e/ou gerar um modelo de regressão que permita prever o consumo energético com base nas condições meteorológicas para um determinado código postal, ou mesmo antecipar aumentos de consumo com base tanto nas condições meteorológicas como nos períodos sazonais.

Após esta fase, os dados são submetidos a um pré-processamento para criar conjuntos de dados que permitam a extração das informações desejadas e treinar modelos para obter classificadores o mais robustos possível.

Num projeto de pré-processamento e preparação de dados é imperativo cumprir algumas etapas antes de tratar de qualquer tipo de tarefa de extração de características e modulação de um problema. Deve ser feita uma limpeza dos dados, removendo eventual ruído introduzido por algumas características, e corrigir observações que possam estar incompletas, ou pouco precisas. Além disto, é essencial ser possível de compreender e caracterizar o conjunto de dados disponível, descrevendo detalhadamente a sua origem, estrutura e qualidade [1].

Identificar e corrigir os problemas dos dados é crucial para garantir a qualidade e integridade dos dados analisados para o modelo a ser treinado, perspetivando uma melhor *performance* para um determinado objetivo traçado. Posteriormente, deve ser considerado o problema da maldição da dimensionalidade, e a complexidade introduzida no treino do modelo devido a este problema, pelo que devem ser aplicadas de técnicas de redução de dimensionalidade, efetuando a seleção e redução de características. A maldição de dimensionalidade demonstra que à medida que se aumenta o conjunto de *features* de um conjunto de dados, o treino de modelos de *machine learning* requer cada vez mais um maior número de amostras, mostrando-se um obstáculo na robustez do modelo. Posto isto, a redução de dimensionalidade compromete a variância dos dados originais, porém oferecendo melhores características para o treino de modelos de aprendizagem. Durante o processo da redução de dimensionalidade removem-se informações pouco relevantes ao conjunto de dados, eliminando possível ruído na fase de treino.

Neste trabalho, o processo de análise, correção, e seleção de dados para a construção de novos conjuntos de dados relevantes ao problema a modular foi realizado utilizando a linguagem de programação R.

Neste processo, foram criados os conjuntos de dados fundamentais para o desenvolvimento de modelos de aprendizagem. Ao compreender claramente o problema, caracterizar os dados e aplicar técnicas de pré-processamento adequadas, estamos a preparar o terreno para análises mais avançadas nas fases subsequentes.

II. Caracterização do conjunto de dados disponível

A caracterização define o processo de conhecer de forma mais detalhada o conjunto fornecido, nomeadamente em termos de dimensionalidade, número de amostras, que tipo de características tem, o que cada valor representa, e identificar possíveis valores ruidosos ou em falta por forma a tomar decisões de como limpar o *dataset* para treinar um modelo com apenas informação pertinente que garanta o melhor desempenho possível [2].

Neste capítulo, o grupo descreve sucintamente o significado de cada uma das *features* dos dois conjuntos de dados. O conhecimento das variáveis com as quais tratamos permite então formular possíveis perguntas que se pretendam responder ou solucionar preparando os dados para o treino de modelos de *machine learning*.

De seguida, é feita a análise estatística procurando identificar valores ruidosos, ou em falta, e compreender o significado dos mesmos, por forma a determinar se é feita imputação dos dados, e como, ou removidas as entradas.

A. Dados de consumos de energia

Os dados fornecidos são relacionados ao consumo elétrico ativo, medido em quilowatt-hora para um código postal numa determinada hora ao longo de vários dias compreendidos entre o fim de 2022 e 2023.

i) Estrutura dos Dados:

Date/Time: Representa a data e hora em que a medição foi feita. Parece estar em formato ISO 8601, incluindo o fuso horário.

Date: Data da medição no formato DD/MM/AAAA.

Hour: Hora da medição, formato HH:mm

Zip Code: Primeiros quatro dígitos do código postal do local onde a medição foi feita.

Active Energy: Variável contínua, representa a energia elétrica ativa consumida, medida em kWh.

ii) Qualidade dos Dados:

Este conjunto de dados é composto por 5 *features*, com um registo de um total de aproximadamente 3.8 milhões medições de energia ativa consumida para os vários códigos postais compreendidos ao longo do país. Verifica-se que todos as entradas possuem um valor não nulo, ou NA, para todas as *features*, pelo que não foi necessário fazer qualquer tipo de imputação. A existência de valores que possam ser considerados *outliers* foi preservada, uma vez que um pico de consumo energético pode representar alguma ocasião especial, nomeadamente um tipo de evento espontâneo e que possa ser do interesse ser identificado por um modelo de *machine learning*, sendo este um dos objetivos definidos pelo grupo na preparação deste conjunto de dados. O consumo energético, por se tratar de uma *feature* contínua, terá de ser discretizada após eventual processo de *feature selection* ou *feature reduction* por forma a evitar ruído na aprendizagem do modelo a treinar numa fase posterior. Notou-se um atributo comum ao segundo conjunto de dados fornecido relativo às condições metereológicas em Lisboa (código postal 1000), sendo este a coluna do fuso horario (*timestamp*), que pode ser útil para unir ambos e avaliar o consumo energético dadas as condições metereológicas para um conjunto de códigos postais, perspetivando um modelo geral que possa fazer uma aproximação de consumo energético dadas as condições previstas.

Tem-se como observação adicional a existência de *features* que demonstram as mesmas características, como as do fuso horário, que podem então ser tratadas apenas por uma, reduzindo *à priori* o conjunto de dados para uma dimensionalidade menor, sem que sejam necessárias técnicas de *feature selection* ou *feature reduction*.

Desta forma, apenas por compreensão e análise dos dados deste conjunto, o grupo traça três tipos de modelos que pretende treinar e avaliar a performance numa fase posterior: Dois classificadores, o primeiro responsável por identificar eventos num determinado Zip.Code com base em picos esporádicos de consumo, e um modelo classificador de Zip.Code como residencial ou industrial. O terceiro modelo trata-se de um algoritmo baseado em regressão, que com base nas condições meteorológicas prevê o consumo energético para Lisboa.

B. Dados de condições climatéricas

O segundo conjunto de dados retrata uma série temporal de observações meteorológicas, provenientes de uma estação meteorológica local ou de uma fonte semelhante que regista informações climatéricas em Lisboa.

i) Estrutura dos Dados:

Cada linha representa uma observação de hora a hora, com os seguintes parâmetros/colunas:

name: Nome da localidade (constantemente "Lisbon" neste conjunto).

datetime: Data e hora da observação no formato "AAAA-MM-DD-HH:MM:SS".

temp: Temperatura média em graus Celsius, variável continua.

feelslike: Sensação térmica em graus Celsius, variável continua

dew: Ponto de orvalho em graus Celsius, variável continua

humidity: Humidade relativa em percentagem representada por variável contínua.

precip: Precipitação em milímetros, variável continua

precipprob: Probabilidade de precipitação em percentagem, variável continua.

preciptype: Tipo de precipitação (por exemplo, "rain" para chuva).

snow: Quantidade de neve em milímetros, variável continua

snowdepth: Profundidade da neve em milímetros, variável continua

windgust: Rajada máxima do vento em quilómetros por hora.

windspeed: Velocidade do vento em quilómetros por hora medida a 10m do solo, variável continua

winddir: Direção do vento em graus, variável discreta

sealevelpressure: Pressão atmosférica ao nível do mar em milibares, variável continua

cloudcover: Cobertura de nuvens no céu, em percentagem, variável continua

visibility: Visibilidade em quilómetros, variável continua
solarradiation: Radiação solar em Watts por metro quadrado, variável continua
solarenergy: Energia solar em Joules por metro quadrado, variável continua
uvindex: Índice UV, variável discreta
severerisk: Risco severo, variável discreta
conditions: Condições meteorológicas resumidas.
icon: Ícone representando as condições meteorológicas.
stations: Lista de estações meteorológicas.

ii) Qualidade dos Dados:

O conjunto de dados original é fornecido pela Visual Crossing Weather [3] e apresenta uma dimensionalidade de 24 *features* com 9504 observações. Da análise estatística de observações, notamos que para a *feature 'severerisk'* estão em falta 8760 valores, representativo de aproximadamente 92% das observações têm valores em falta. Desta forma, é afastada a ideia de remover estas observações, uma vez que a perda de informação seria consideravelmente grande, reduzindo o *dataset* a apenas 8% da dimensão original. Em simultâneo, outras *features* apresentam valores em falta, sendo elas *'preciptype', 'solarradiation'* e *'uvindex'*.

Da fonte dos dados, destacamos um parágrafo onde é relatado que observações com valores de *features* em falta não são o mesmo que o valor 0, isto é, representam valores desconhecidos ou não disponíveis na altura da medida.

Determinada a necessidade de realizar a imputação, foi necessário definir um critério por forma a reduzir ao máximo o *bias* introduzido pela mesma. É então pertinente realizar a análise dos valores da média e desvio padrão destas *features*, onde se sabe, de boa prática, que a imputação de valores com a média da *feature* produz bons resultados e menor *bias* caso o desvio padrão seja menor, o que apenas se verificou para a *feature 'solarradiation'*. Para a *feature 'preciptype'*, tomou-se a decisão de imputar o valor '*unknown'*, indicativo que não existe informação acerca da precipitação no momento da medição da amostra.

Atendendo ao elevado desvio padrão da *feature 'uvindex'*, a decisão de imputação passou por utilizar a mediana, o que pode resultar na introdução de *bias* para este valor. A fase de limpeza dos dados termina com a imputação do valor 0 para a *feature 'severerrisk'* uma vez que ao observar registamos que ou tem valor 10, ou é omisso. Desta forma, o grupo contactou com três diferentes metodologias de imputação, compreendendo os cenários onde cada uma é mais vantajosa que outras, e toma consciência da introdução do enviesamento destes valores no treino do modelo. É, no entanto, importante mencionar, que alguns modelos lidam com dados esparsos e valores em falta, pelo que numa fase de treino de modelos devem ser considerados conjuntos de dados com e sem imputação.

III. Soluções adotadas e modelos a estudar

Resolvido o problema de características com valores em falta, a análise de ambos o conjunto de dados forneceu alguma informação válida e pertinente para a construção os três modelos mencionados anteriormente. Numa fase inicial, o grupo procurou analisar individualmente o conjunto de dados de consumo energético, procurando encontrar informação relevante que pudesse ser tratada e transformada para treinar um modelo com base nestes dados. Começando pelos códigos postais, com auxílio a dados externos indicadores de consumo industrial e doméstico, provenientes da PorData [4], o grupo considerou interessante a preparação de um *subset* dos dados originais, onde se categoriza cada código postal escolhido como Residencial ou Industrial, na perspetiva de treinar um classificador supervisionado, que possa prever o tipo de zona com base nos consumos energéticos ao longo do dia. Para suportar esta classificação, traçaram-se dois gráficos comparativos de consumo energético médio ao longo de 24h para zonas classificadas como residenciais e industriais, onde, depois de normalização, se verificou que em zonas industriais o consumo energético muito raramente desce abaixo de 70% do seu consumo máximo diário, ao contrário das zonas residenciais onde a tendência das horas "mortas" e a madrugada demonstra uma forte descida do consumo energético. Este novo conjunto de dados apresenta uma dimensão de 240480 tuplos com 8 *features* cada um.

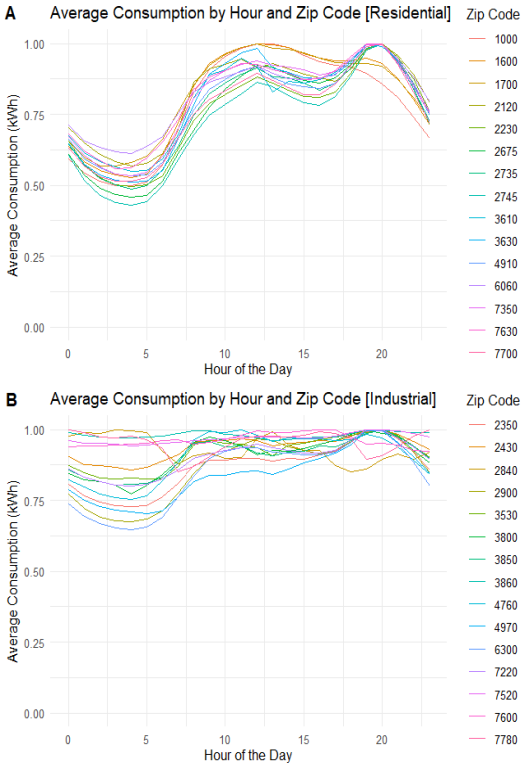

Figura 1 - Consumo médio por hora, residencial VS industrial

Ainda no *dataset* do consumo energético, o grupo achou relevante, embora ainda sem objetivo definido, preparar um conjunto de dados com a evolução do consumo médio energético ao longo dos diferentes dias da semana para zonas industriais e residenciais, onde se verificaram valores menos acentuados aos fins de semana.

Como trabalho ainda por desenvolver, o grupo pretende preparar dados por forma a poder identificar a existência ou

não de um evento esporádico ao identificar picos de consumo energético inesperados. Numa fase inicial, o etiquetamento de haver ou não evento é feito com base em cenários reais para um determinado código postal, por exemplo, para um jogo de futebol com grande afluência espera-se um aumento súbito do consumo de energia ativa. Este tipo de problemas envolve a dependência temporal, uma vez que para ser feita a identificação deve-se ter em conta os consumos de horários anteriores. Para tal, estima-se que o *dataset* a preparar seja constituído por *features* referentes ao código postal, consumo energético ativo, data (separada em dia, mês, ano) e hora.

Numa segunda etapa, a fim de implementar um modelo de regressão com o objetivo de prever consumos energéticos com base em condições meteorológicas na localidade de Lisboa, o grupo uniu os dois *datasets* fornecidos, uma vez que os dados disponíveis se focam para as condições meteorológicas no código postal 1000. A interseção dos dois conjuntos de dados foi feita segundo o atributo comum, *timestamp,* para o código postal referente a Lisboa (1000) garantindo que as condições meteorológicas e o consumo ativo são correlacionáveis e precisas.

O resultado da união dos *datasets* evidencia a necessidade da eliminação de *features* repetidas e redundantes ao problema, tais como Zip.Code, name, datetime, stations, icon. É também necessário o tratamento da *feature* '*date'*, convertida em quatro novas *features* 'Day_of_Week', '*day', 'month', 'year'*, tornando desnecessária a *feature 'datetime'* razão pela qual foi anteriormente eliminada.

IV. Redução de dimensionalidade e discretização

No tópico de redução de dimensionalidade, para o primeiro conjunto preparado de consumos energéticos por zona, e com perspetiva de treinar um classificador, o grupo realizou uma transformação das variáveis categóricas: Na *class label*, as zonas industriais e residenciais foram transformadas em alternativas binárias, valor "1" ou "2", respetivamente. Transformou-se também a característica de *timestamp* em quatro *features* cada uma representativa do dia, mês, ano, e hora registada, pelo que terminámos com um conjunto de dados de dimensionalidade n =240480, d = 8. Atendendo ao número de características, em que duas são únicas, nomeadamente o código postal e a *class label*, o grupo evidencia a baixa dimensionalidade, e deixa em aberto o estudo da *performance* de um modelo treinado com uma redução deste conjunto original, face à desvantagem da perda de informação relevante consoante o *threshold* definido.

Embora a baixa dimensionalidade em termos de *features* aplicaram-se as técnicas de FS e FR. Na *Feature Selection* supervisionada, as métricas escolhidas foram o *Fisher's Ratio e Information Gain,* que tal como esperado, consideram apenas a característica do consumo ativo como determinante para classificar a zona como industrial ou residencial, uma vez que este *labeling* foi feito com base no consumo energético.

Desta forma, com base na FS, o modelo a treinar seguirá uma classificação binária muito simples, onde considera apenas se o valor do consumo ultrapassa um determinado valor para poder classificar com confiança a zona como residencial ou industrial.

Para a *Feature Reduction* aplicou-se a técnica de PCA, removendo as variáveis únicas e labels, e se obteve uma redução de 6 *features* para 5 componentes, com uma relevância de 95% dos dados originais. Desta forma, o grupo acredita que por se tratar de uma dimensão tão reduzida de *features*, a técnica do PCA pode não compensar a perda de relevância de informação. Esta afirmação será então confirmada após fase de treino dos modelos e comparação de resultados.

No conjunto de dados do consumo energético para Lisboa aplicou-se a técnica de *Feature Selection* não supervisionada e supervisionada, com a *class label* definida pela *feature "conditions"*, com o objetivo de comparar o desempenho de cada uma das métricas utilizadas. As métricas utilizadas foram o *Fisher's Ratio*, *Information Gain* e *Variance Threshold*, no entanto, para o segundo, a métrica indica que a *feature* referente ao consumo energético é das menos relevantes, que não vai de encontro ao que o grupo pretende. Para tal, foi comparado o resultado do *Fisher's Ratio* com o resultado do *Variance Threshold* e o primeiro reduziu o *dataset* em menos *features* que pareciam ser relevantes para o problema. Desta forma, o conjunto de dados final apresenta um formato de 7295 observações com 15 *features*.

Após redução da dimensionalidade dos conjuntos de dados preparados, aplicou-se a técnica de discretização de *Equal Frequency Binning*, onde se dividem os valores em *bins* com a mesma frequência de ocorrências. O número de *bins* foi definido segundo a regra de Sturges, onde:

$$\#bins = \log_2 n + 1$$

V. Conclusão

A elaboração da primeira fase do trabalho prático tinha por objetivo compreender e descrever o problema existente na mineração de dados para grandes conjuntos de dados conhecido por Maldição da Dimensionalidade, bem como interpretar e caracterizar o conjunto de dados fornecido para formular problemas e realizar o pré-processamento dos dados por forma a treinar modelos de aprendizagem automática que respondam aos mesmos. Durante a elaboração do trabalho foram compreendidos conceitos como os critérios na imputação de dados e de que forma estes podem afetar a aprendizagem do modelo. Em simultâneo, colocaram-se em prática os conhecimentos adquiridos referentes às técnicas de redução de dimensionalidade como *Feature Selection* supervisionada e *Feature Reduction* com base na técnica PCA. Destaca-se também a importância da discretização dos dados para reduzir eventuais ruídos que possam afetar o desempenho dos modelos a desenvolver na próxima etapa do trabalho prático.

O grupo considera os objetivos principais concluídos, no entanto destaca alguns aspetos a melhorar, nomeadamente a utilização de técnicas de FR como o SVD que ainda podem ser aplicadas, e também a preparação do *dataset* para identificação de eventos numa determinada zona, com base nos picos dos consumos energéticos comparativamente aos horários anteriores.

REFERENCES

[1] https://www.scalablepath.com/data-science/data-preprocessing-phaseI.

[2] https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Quantitative_Research_Methods_for_Political_Science_Public_Policy_and_Public_Administration_(Jenkins-Smith_et_al.)/03%3A_Exploring_and_Visualizing_Data/3.01%3A_Characterizing_Data

[3] https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/

[4] https://www.pordata.pt/