Classificação Multiclasse da Qualidade do Ar em São Paulo: Uma Abordagem de Mineração de Dados

Andrei Inoue Hirata
UNESP
andreihirata@unesp.br

RESUMO

A poluição do ar é um desafio crítico para grandes centros urbanos, impactando a saúde pública e o meio ambiente. Este trabalho explora a aplicação de técnicas de mineração de dados para a classificação multiclasse da qualidade do ar em São Paulo, utilizando dados horários da estação de monitoramento da CETESB. O objetivo é comparar diferentes algoritmos de classificação e estratégias de decomposição de problemas multiclasse. O estudo é dividido em duas partes: (1) a comparação de desempenho entre os modelos Árvore de Decisão, Bagging, AdaBoost e Random Forest; e (2) a análise das abordagens de decomposição One-vs-Rest (OVR) e One-vs-One (OVO) usando uma Árvore de Decisão como classificador base. Os resultados, avaliados pela métrica F1-Score, indicam que o Random Forest obteve o melhor desempenho na classificação direta, e que a abordagem OVR superou a OVO na estratégia de decomposição. O trabalho conclui com uma otimização de hiperparâmetros do modelo Random Forest, demonstrando uma melhora adicional em sua performance.

Palavra Chaves

Mineração de Dados, Classificação Multiclasse, Qualidade do Ar, Cidades Inteligentes, Random Forest, Decomposição OVO/OVR

1. INTRODUÇÃO

A classificação é uma das tarefas fundamentais da mineração de dados, cujo objetivo é construir um modelo a partir de dados rotulados para prever a classe de novas instâncias. Enquanto problemas binários são comuns, muitos cenários do mundo real, como o monitoramento ambiental em Cidades Inteligentes, envolvem múltiplas categorias, caracterizando um problema de classificação multiclasse [2].

Este estudo de caso foca na previsão da qualidade do ar em São Paulo no ano de 2024. Utilizamos dados horários da Companhia Ambiental do Estado de São Paulo (CETESB) [1] para classificar o Índice de Qualidade do Ar (IQA) em múltiplas categorias. A análise se divide em duas investigações principais: a comparação direta de algoritmos de aprendizado supervisionado e a avaliação de técnicas de decomposição que transformam o problema multiclasse em múltiplos problemas binários. [1]

2. METODOLOGIA

O processo metodológico envolveu a preparação dos dados, a implementação dos modelos e a avaliação de desempenho.

2.1 Dataset e Pré-processamento

Utilizamos os dados horários da estação Pinheiros da CETESB no ano de 2024, contendo medições do poluente Partículas Inaláveis (MP10).

- **Limpeza e Preparação:** O arquivo de dados foi lido, tratando o cabeçalho específico do formato da CETESB. A coluna de data/hora foi convertida para o formato *datetime* para extração de features temporais (mês, dia da semana, hora).
- **Criação da Variável Alvo:** Com base nas faixas oficiais da CETESB para o poluente MP10, criamos a variável alvo 'IQA_cat' com 4 classes: 'Boa' ($<= 50 \mu g/m^3$), 'Moderada' (> 50 e $<= 100$), 'Ruim' (> 100 e $<= 150$), e 'Muito Ruim' (> 150).
- **Imputação:** Valores ausentes na coluna de feature 'mp10' foram preenchidos usando a interpolação linear, uma técnica adequada para séries temporais.
- **Validação:** Utilizamos a estratégia de validação *Hold-out*, dividindo os dados em 70% para treinamento e 30% para teste, garantindo que a ordem cronológica fosse preservada.

2.2 Avaliação de Modelos

Todos os modelos foram avaliados usando a métrica F1-Score (ponderado), que é robusta para datasets com classes desbalanceadas, pois calcula a média harmônica entre precisão e revocação.

[1] O código-fonte, o dataset e os resultados gerados neste trabalho estão disponíveis em: `https://github.com/gitandreihirata/Unesp_Doutorado/tree/main/Mineracao%20de%20Dados/Exercicios/Ex2`

3. RESULTADOS E DISCUSSÃO

Os experimentos foram divididos em duas partes, conforme solicitado no escopo do trabalho.

3.1 Parte 1: Comparação de Algoritmos Multiclasse

Nesta seção, comparamos quatro algoritmos capazes de lidar nativamente com problemas multiclasse: Árvore de Decisão, Bagging, AdaBoost e Random Forest. Os resultados de desempenho no conjunto de teste são apresentados na Figura 2.

O modelo Random Forest apresentou o melhor desempenho inicial. O AdaBoost teve um resultado notavelmente inferior, o que pode ser atribuído à sua sensibilidade a ruídos ou à necessidade de um ajuste mais fino. A matriz de confusão para o modelo Random Forest (Figura 1) detalha sua performance, mostrando que a maioria dos erros ocorre entre as classes adjacentes 'Moderada' e 'Ruim'.

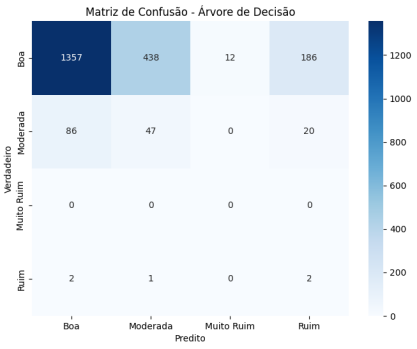

Figure 1: Matriz de Confusão para o modelo Random Forest.

3.1.1 Otimização de Hiperparâmetros

Para o Random Forest, realizamos uma busca em grade ('GridSearchCV') para otimizar os hiperparâmetros 'n_estimators' e 'max_depth'. A melhor combinação encontrada ('n_estimators=200', 'max_depth=20') elevou o F1-Score no conjunto de teste, demonstrando uma melhora marginal, porém positiva, em sua capacidade preditiva.

3.2 Parte 2: Decomposição em Problemas Binários

Nesta seção, avaliamos se a transformação do problema multiclasse em múltiplos problemas binários melhora o desempenho, utilizando uma Árvore de Decisão como classificador base. Implementamos as estratégias One-vs-Rest (OVR) e One-vs-One (OVO).

Os resultados completos, incluindo a comparação de todas as abordagens, são consolidados na Figura 2.

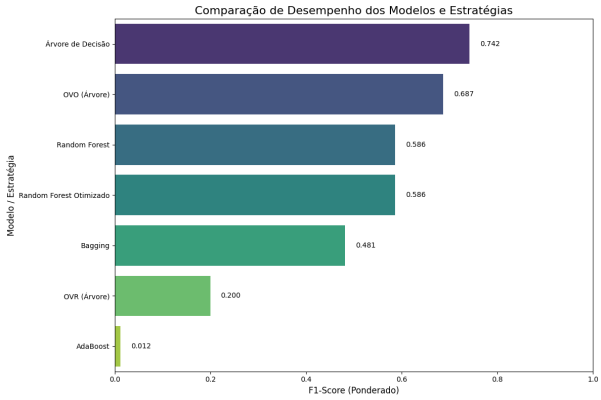

Figure 2: Comparação de Desempenho de todos os Modelos e Estratégias.

A abordagem One-vs-Rest (OVR) apresentou um resultado ligeiramente melhor que a Árvore de Decisão padrão, enquanto a estratégia OVO teve o pior desempenho entre todas. Isso pode ocorrer em datasets onde os classificadores OVO, treinados com menos dados (apenas duas classes por vez), não conseguem generalizar tão bem quanto os modelos OVR ou os modelos multiclasse nativos.

4. CONCLUSÃO

Este trabalho demonstrou a aplicação de um pipeline de mineração de dados para um problema de classificação multiclasse no contexto de cidades inteligentes. A análise comparativa revelou que, para o dataset de qualidade do ar da CETESB, o algoritmo de ensemble Random Forest foi o mais eficaz, especialmente após a otimização de hiperparâmetros.

Adicionalmente, ao investigar estratégias de decomposição, a abordagem One-vs-Rest (OVR) se mostrou uma alternativa viável, superando a abordagem OVO e a Árvore de Decisão padrão.

Conclui-se que tanto algoritmos de ensemble robustos quanto estratégias de decomposição são ferramentas poderosas para abordar problemas complexos de classificação multiclasse, fornecendo uma base sólida para a criação de modelos preditivos aplicáveis a desafios do mundo real.

5. REFERENCES

[1] Companhia Ambiental do Estado de São Paulo (CETESB). (2025). *QUALAR: Consulta de Dados Horários*. Acessado em 6 de outubro de 2025, de `https://qualar.cetesb.sp.gov.br/qualar/home.do`.

[2] Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining (2nd ed.)*. Pearson.