Relatório da 3 Avaliação - Tópicos Especiais em Computação

Marcos Antonio e Vandirleya Barbosa

[1]Universidade Federal de Piauí (UFPI)

marcos.brito, vandirleya.barbosa@ufpi.edu.br

Resumo. *Este estudo realiza uma análise detalhada e comparativa do desempenho de diversos algoritmos classificadores, especificamente ID3, Random Forest, SVM e MLP, na tarefa crítica de reconhecimento de padrões, aplicando-os à base de dados Global Data on Sustainable Energy. Ao longo dos experimentos, foram considerados quatro cenários distintos, cada um explorando diferentes combinações de normalização dos dados e seleção de características, com o objetivo de identificar as condições ideais para cada algoritmo. As métricas de desempenho, que incluem acurácia, precisão, recall, AUC (Área Sob a Curva ROC), e Kappa, foram rigorosamente calculadas para fornecer uma visão abrangente da eficácia de cada classificador. Os resultados mostram que o algoritmo Random Forest obteve o melhor desempenho geral, destacando-se particularmente nos cenários com normalização e seleção de características e sem normalização e seleção de características, com valores elevados em todas as métricas. O Random Forest demonstrou uma alta acurácia e estabilidade em comparação com os outros algoritmos, sugerindo sua eficácia superior em tarefas de reconhecimento de padrões quando as condições de normalização e seleção de características são aplicadas.*

Abstract. *This study conducts a detailed and comparative analysis of the performance of various classification algorithms, specifically ID3, Random Forest, SVM, and MLP, in the critical task of pattern recognition, applying them to the Global Data on Sustainable Energy dataset. Throughout the experiments, four distinct scenarios were considered, each exploring different combinations of data normalization and feature selection, with the goal of identifying the optimal conditions for each algorithm. Performance metrics, including accuracy, precision, recall, AUC (Area Under the ROC Curve), and Kappa, were rigorously calculated to provide a comprehensive view of each classifier's effectiveness. The results show that the Random Forest algorithm achieved the best overall performance, particularly excelling in scenarios with both normalization and feature selection and without normalization and feature selection, with high values across all metrics. Random Forest demonstrated high accuracy and stability compared to the other algorithms, suggesting its superior effectiveness in pattern recognition tasks when normalization and feature selection conditions are applied.*

1. Introdução

O reconhecimento de padrões é um aspecto fundamental na ciência de dados e aprendizado de máquina, desempenhando um papel crucial em diversas aplicações, desde a análise de imagens e reconhecimento de voz até a previsão de comportamentos e decisões.

O objetivo principal desse estudo é avaliar e comparar o desempenho de diferentes algoritmos classificadores na tarefa de reconhecimento de padrões, utilizando a base de dados "Global Data on Sustainable Energy". Esta base de dados contém informações detalhadas sobre o consumo e produção de energia sustentável global, o que torna a tarefa de reconhecimento de padrões particularmente relevante para entender tendências e padrões na área de energia sustentável.

Neste contexto, os algoritmos de aprendizado de máquina selecionados para esta análise incluem o ID3 (Iterative Dichotomiser 3), Random Forest, Support Vector Machines (SVM) e Multi-Layer Perceptron (MLP). Cada um desses algoritmos possui características distintas e métodos de classificação variados, o que pode influenciar significativamente o desempenho na tarefa de reconhecimento de padrões. O algoritmo ID3 (Iterative Dichotomiser 3) é um método clássico de árvore de decisão que utiliza a medida de entropia para construir a árvore. A principal vantagem do ID3 é sua simplicidade e a capacidade de lidar com dados categóricos diretamente, sem necessidade de pré-processamento extensivo. No entanto, o ID3 pode ser suscetível a overfitting, especialmente com conjuntos de dados complexos ou muito grandes, devido à sua tendência de criar árvores muito profundas.

Random Forest é um algoritmo de aprendizado de máquina baseado em conjuntos que combina várias árvores de decisão para melhorar a precisão e robustez do modelo. Cada árvore é treinada em um subconjunto diferente dos dados e utiliza uma amostra aleatória de características, o que ajuda a reduzir a variabilidade e melhorar a generalização do modelo. A principal vantagem do Random Forest é sua capacidade de manejar grandes volumes de dados e características sem necessidade de ajuste extensivo de parâmetros, além de sua robustez contra overfitting. O Support Vector Machine (SVM) é um classificador poderoso que busca encontrar o hiperplano que maximiza a margem entre diferentes classes no espaço de características. Utilizando funções de kernel, o SVM pode lidar com problemas de classificação não linear, mapeando os dados para um espaço de dimensões mais altas onde um hiperplano linear pode ser encontrado. A força do SVM está na sua capacidade de lidar com conjuntos de dados complexos e encontrar separações eficazes, embora possa ser sensível a parâmetros de configuração e ao tamanho dos dados.

O Multi-Layer Perceptron (MLP) é uma rede neural feedforward composta por múltiplas camadas de neurônios, cada uma conectada às camadas adjacentes. O MLP é capaz de modelar relações não lineares complexas entre características e classes, utilizando funções de ativação não lineares e técnicas de retropropagação para ajustar os pesos da rede. Sua principal vantagem é a flexibilidade e capacidade de aprender representações complexas dos dados, mas pode exigir um volume significativo de dados e tempo de treinamento para atingir um desempenho ideal. O presente estudo realiza uma análise comparativa entre esses algoritmos em quatro cenários experimentais distintos, variando as técnicas de normalização dos dados e a seleção de características. As métricas de desempenho, incluindo acurácia, precisão, recall, AUC (Área Sob a Curva ROC) e Kappa, serão utilizadas para avaliar a eficácia de cada algoritmo em identificar padrões relevantes no conjunto de dados. O objetivo é determinar quais condições e algoritmos oferecem o melhor desempenho para a tarefa de reconhecimento de padrões em energia sustentável, fornecendo informções valiosas para aplicações práticas e futuras pesquisas na área.

2. Desenvolvimento

Esta seção descreve a metodologia adotada para a elaboração deste relatório. Serão detalhados o uso do conjunto de dados e os passos seguidos para a análise e obtenção dos resultados. Isso inclui a seleção das variáveis de interesse, a aplicação de técnicas de pré-processamento e estatísticas específicas e a seleção dos gráficos para interpretação dos padrões identificados.

2.1. Metodologia

Para a realização da análise comparativa dos algoritmos classificadores, utilizou-se o conjunto de dados "Global Data on Sustainable Energy", que contém informações sobre consumo e produção de energia sustentável. A metodologia adotada para esta análise envolve as seguintes etapas: Os dados foram inicialmente carregados e processados utilizando a biblioteca `pandas` do Python. Colunas desnecessárias, como 'Entity', 'Year', 'Latitude' e 'Longitude', foram removidas para simplificar a análise. A base de dados foi então tratada para lidar com valores ausentes, empregando o `SimpleImputer` da biblioteca `sklearn` para preencher os valores NaN com a média das colunas.

A coluna de destino, "Access to electricity (% of population)", foi utilizada para criar uma variável categórica, classificando os dados em três categorias: 'Baixa', 'Média' e 'Alta', com base no percentual de acesso à eletricidade. Esta transformação categórica permite a aplicação dos algoritmos de classificação. Os dados foram divididos em conjuntos de treino, validação e teste. Inicialmente, o conjunto de dados foi dividido em treino (80%) e teste (20%) usando a função `train_test_split`. Posteriormente, o conjunto de teste foi subdividido em teste e validação, cada um com 20% dos dados originais.

A metodologia adotou quatro cenários experimentais distintos para avaliar o impacto das técnicas de pré-processamento e seleção de características:

- **Cenário 1: Sem normalização e sem seleção de características** - Utilização direta dos dados brutos.
- **Cenário 2: Com normalização, sem seleção de características** - Aplicação de normalização com `StandardScaler` para padronizar as características.
- **Cenário 3: Com seleção de características, sem normalização** - Aplicação de seleção de características com `SelectKBest` para escolher as 10 melhores características com base na análise de variância.
- **Cenário 4: Com normalização e seleção de características** - Aplicação simultânea de normalização e seleção de características.

Os algoritmos a serem avaliados incluem:

- **ID3 (DecisionTreeClassifier)** - Algoritmo de árvore de decisão.
- **Random Forest (RandomForestClassifier)** - Algoritmo baseado em ensembles de árvores de decisão.
- **Support Vector Machine (SVC)** - Classificador que encontra o hiperplano ótimo para separar as classes.
- **Multi-Layer Perceptron (MLPClassifier)** - Rede neural com múltiplas camadas de neurônios.

Cada modelo foi treinado e avaliado em cada cenário, e as métricas de desempenho foram calculadas, incluindo acurácia, precisão, recall, AUC (Área Sob a Curva ROC) e Kappa. A função `train_and_evaluate_model` foi utilizada para ajustar e avaliar cada modelo, e os resultados foram armazenados em um DataFrame para análise. Em seguida para garantir a robustez dos resultados, a validação cruzada com 5 folds foi realizada. Cada algoritmo foi avaliado em cada fold, repetindo os quatro cenários descritos anteriormente. As métricas obtidas foram agregadas e analisadas para fornecer uma visão abrangente do desempenho dos algoritmos.

2.2. Resultados

Os resultados das experimentações foram salvos em um arquivo CSV para facilitar a análise posterior. O arquivo contém as métricas de desempenho para cada cenário e algoritmo, permitindo a comparação direta dos resultados obtidos. Esta metodologia proporciona uma avaliação detalhada e comparativa dos algoritmos de classificação, considerando diferentes técnicas de pré-processamento e seleção de características, com o objetivo de identificar o desempenho ótimo em tarefas de reconhecimento de padrões em energia sustentável.

A Tabela 1 apresenta o cenário 1, onde não houve normalização nem seleção de características, o RandomForestClassifier destacou-se com uma Accuracy de 97.26%, Precision de 97.31%, Recall de 97.26%, ROC AUC de 99.72% e Kappa de 93.06%. Esse desempenho superior sugere que, mesmo sem pré-processamento, o RandomForestClassifier é altamente eficaz na classificação dos dados. O DecisionTreeClassifier também apresentou resultados sólidos com Accuracy de 95.75%, Precision de 95.90%, Recall de 95.75%, ROC AUC de 93.95% e Kappa de 89.27%, indicando um bom desempenho geral. Em contraste, o SVC e o MLPClassifier mostraram um desempenho significativamente inferior, com Accuracy de 75.89% e 76.16%, respectivamente, e baixos valores de ROC AUC e Kappa, refletindo dificuldades na discriminação entre classes e baixa concordância com a classificação aleatória.

Tabela 1. Cenário 1: Sem Normalização e Sem Seleção de Características

Algoritmo	Accuracy	Precision	Recall	ROC AUC	Kappa
DecisionTreeClassifier	0.9575	0.9590	0.9575	0.9395	0.8927
RandomForestClassifier	0.9726	0.9731	0.9726	0.9972	0.9306
SVC	0.7589	0.5759	0.7589	0.4985	0.0000
MLPClassifier	0.7616	0.6276	0.7616	0.5124	0.0290

A Tabela 2 apresenta o Cenário 2, que incluiu normalização dos dados, o RandomForestClassifier continuou a demonstrar um desempenho notável, com Accuracy de 97.39%, Precision de 97.43%, Recall de 97.39%, ROC AUC de 99.73% e Kappa de 93.41%. A normalização parece ter contribuído para uma ligeira melhoria na capacidade de discriminação do modelo. O SVC obteve uma melhoria significativa, com Accuracy de 90.82%, Precision de 90.71%, Recall de 90.82%, ROC AUC de 97.64% e Kappa de 76.66%, sugerindo que a normalização ajudou a melhorar a performance deste modelo. O MLPClassifier também se beneficiou, apresentando Accuracy de 95.48%, Precision de 95.51%, Recall de 95.48%, ROC AUC de 99.19% e Kappa de 88.58%, mostrando que a normalização contribuiu para uma maior estabilidade e precisão.

Tabela 2. Cenário 2: Com Normalização e Sem Seleção de Características

Algoritmo	Accuracy	Precision	Recall	ROC AUC	Kappa
DecisionTreeClassifier	0.9575	0.9590	0.9575	0.9395	0.8927
RandomForestClassifier	0.9739	0.9743	0.9739	0.9973	0.9341
SVC	0.9082	0.9071	0.9082	0.9764	0.7666
MLPClassifier	0.9548	0.9551	0.9548	0.9919	0.8858

A Tabela 3 apresenta o Cenário 3, com seleção de características aplicada e sem normalização, o RandomForestClassifier manteve a melhor performance com Accuracy de 97.53%, Precision de 97.59%, Recall de 97.53%, ROC AUC de 99.75% e Kappa de 93.79%, indicando que a seleção de características não prejudicou seu desempenho. O DecisionTreeClassifier apresentou uma leve redução na Accuracy e na Precision, mas manteve boas métricas de ROC AUC e Kappa, sugerindo que a seleção de características teve pouco impacto negativo. O SVC continuou a mostrar resultados baixos com Accuracy de 73.69% e ROC AUC de 61.57%, indicando que a seleção de características não trouxe melhorias significativas para este modelo. O MLPClassifier apresentou uma melhora notável com Accuracy de 89.45%, Precision de 88.97%, Recall de 89.45%, ROC AUC de 96.02% e Kappa de 72.24%, indicando que a seleção de características teve um impacto positivo.

Tabela 3. Cenário 3: Com Seleção de Características e Sem Normalização

Algoritmo	Accuracy	Precision	Recall	ROC AUC	Kappa
DecisionTreeClassifier	0.9548	0.9560	0.9548	0.9369	0.8858
RandomForestClassifier	0.9753	0.9759	0.9753	0.9975	0.9379
SVC	0.7369	0.5431	0.7369	0.6157	0.0000
MLPClassifier	0.8945	0.8897	0.8945	0.9602	0.7224

Finalmente, A Tabela 4 apresenta o Cenário 4, onde foram aplicadas tanto a seleção de características quanto a normalização, o RandomForestClassifier continuou a ser o melhor modelo, com Accuracy de 97.53%, Precision de 97.59%, Recall de 97.53%, ROC AUC de 99.75% e Kappa de 93.79%. A combinação de normalização e seleção de características aprimorou ainda mais seu desempenho. O SVC também se beneficiou da combinação de técnicas, apresentando Accuracy de 89.59%, Precision de 89.11%, Recall de 89.59%, ROC AUC de 97.05% e Kappa de 74.47%. O MLPClassifier teve um desempenho sólido com Accuracy de 95.07%, Precision de 95.06%, Recall de 95.07%, ROC AUC de 99.03% e Kappa de 88.25%, mostrando que tanto a normalização quanto a seleção de características contribuíram para uma maior estabilidade e precisão.

Tabela 4. Cenário 4: Com Seleção de Características e Com Normalização

Algoritmo	Accuracy	Precision	Recall	ROC AUC	Kappa
DecisionTreeClassifier	0.9521	0.9533	0.9521	0.9319	0.8786
RandomForestClassifier	0.9753	0.9759	0.9753	0.9975	0.9379
SVC	0.8959	0.8911	0.8959	0.9705	0.7447
MLPClassifier	0.9507	0.9506	0.9507	0.9903	0.8825

2.3. Análise Comparativa

Nos quatro cenários analisados, o RandomForestClassifier se destacou como o melhor algoritmo, apresentando consistentemente o melhor desempenho em todas as métricas principais (Accuracy, Precision, Recall, ROC AUC e Kappa). No cenário 1 O RandomForestClassifier obteve a maior Accuracy (97.26%) e ROC AUC (99.72%). Já o SVC e o MLPClassifier mostraram um desempenho bem inferior. O cenário 2 a normalização melhorou a performance do SVC e do MLPClassifier. No entanto, o RandomForestClassifier ainda se destacou com a melhor Accuracy (97.39%) e ROC AUC (99.73%). No cenário 3 com a seleção de características, o RandomForestClassifier manteve o melhor desempenho, com Accuracy de 97.53% e ROC AUC de 99.75%. O MLPClassifier também melhorou, mas ainda ficou atrás do RandomForest. Por fim, o cenário 4 combinando normalização e seleção de características, o RandomForestClassifier continuou sendo o melhor, com Accuracy de 97.53% e ROC AUC de 99.75%. O SVC e o MLPClassifier mostraram melhorias, mas não alcançaram o desempenho do RandomForest.

3. Conclusão

Este trabalho propôs a análise detalhada dos algoritmos DecisionTreeClassifier, RandomForestClassifier, SVC e MLPClassifier utilizando o conjunto de dados *Global Data on Sustainable Energy*, com o objetivo de avaliar o desempenho preditivo e a aplicabilidade desses modelos na predição de padrões em energia sustentável. A análise revelou que o RandomForestClassifier consistentemente superou os demais algoritmos em todos os cenários experimentais. Com alta acurácia e valores elevados de ROC AUC, o RandomForestClassifier se destacou como o algoritmo mais eficaz para a tarefa, demonstrando uma robustez superior tanto com quanto sem normalização e seleção de características. Embora o MLPClassifier tenha apresentado melhorias significativas quando combinado com normalização e seleção de características, ele não conseguiu alcançar o desempenho do RandomForestClassifier. O SVC, por sua vez, mostrou um desempenho inferior, especialmente sem normalização e seleção de características. Assim, os resultados confirmam que o RandomForestClassifier é o mais adequado para a predição de padrões no contexto de energia sustentável, oferecendo a maior precisão e confiabilidade para a análise dos dados. O estudo evidenciou a importância de escolher o algoritmo apropriado para maximizar a performance preditiva, com o RandomForestClassifier se destacando como a escolha ideal para o conjunto de dados em questão.