Classificação Automática de Textos Utilizando o Algoritmo KNN e Suas Variações

Vinicius Henrique Otti Masson RA 102678
Handrey Hartung Toppa RA 105883

Resumo

Este trabalho apresenta uma abordagem para a classificação automática de textos utilizando o algoritmo dos k-vizinhos mais próximos (KNN) e suas variações: o KNN Invertido (kINN) e o KNN Simétrico (kSNN). A proposta visa avaliar a eficácia desses métodos em corpora textuais amplamente utilizados, como Reuters, 20 Newsgroups e Ohsumed (conjuntos de dados de referência amplamente utilizados na área de mineração de texto e aprendizado de máquina). Para isso, são aplicadas técnicas de pré-processamento textual, vetorização com TF-IDF e geração de novas características a partir de matrizes de similaridade. Os resultados obtidos indicam que as variações kINN e kSNN superam o KNN tradicional em determinadas coleções, enquanto o KNN demonstra maior estabilidade frente à variação do parâmetro K. O estudo também explora a aplicação de SVM como método comparativo, evidenciando ganhos estatisticamente significativos com a geração de características. Este trabalho contribui para o avanço da mineração de texto e para o aprimoramento de métodos supervisionados de classificação.

Palavras-chave: Aprendizado Supervisionado, KNN, Classificação de Textos, TF-IDF, Mineração de Texto, Geração de Características.

1 Introdução

O crescimento exponencial de dados textuais, impulsionado por redes sociais, e-mails e artigos científicos, tem gerado uma demanda crescente por métodos automáticos de análise e categorização. A classificação manual de grandes volumes de texto torna-se inviável, abrindo espaço para técnicas de aprendizado de máquina.

Entre os diversos algoritmos supervisionados, o KNN (K-Nearest Neighbors) destaca-se por sua simplicidade e eficácia em diferentes contextos. Este trabalho tem como objetivo aplicar e comparar variações do KNN na tarefa de classificação automática de textos, avaliando o impacto de diferentes estratégias de ponderação e reciprocidade entre vizinhos.

A estrutura deste artigo é organizada da seguinte forma: a Seção 2 apresenta a fundamentação teórica, a Seção 3 descreve os métodos empregados, e a Seção 4 discute os resultados obtidos e as conclusões.

2 Fundamentação Teórica

2.1 Classificação Automática de Textos (CAT)

A Classificação Automática de Textos (CAT) tem como objetivo atribuir categorias a do-

cumentos não rotulados com base em padrões aprendidos a partir de exemplos previamente classificados. Suas aplicações incluem filtragem de spam, análise de sentimentos e recomendação de conteúdo.

O principal desafio da CAT está na representação textual e na escolha de algoritmos que lidem bem com alta dimensionalidade e ambiguidade semântica.

2.2 Representação de Textos

A etapa de representação textual é crucial para o desempenho da classificação. Os métodos mais comuns incluem:

- **Bag of Words (BoW)**: Representa documentos como vetores de frequência de termos.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Pondera os termos de acordo com sua frequência e importância no corpus.

Além disso, o artigo propõe gerar novas características a partir de matrizes de similaridade entre documentos, enriquecendo a representação vetorial e melhorando a separabilidade entre classes.

2.3 Aprendizado Supervisionado

O aprendizado supervisionado consiste em treinar um modelo com exemplos rotulados para que ele possa classificar novos casos. As métricas de avaliação incluem acurácia, precisão, recall e F1-score.

Neste estudo, compara-se o desempenho do KNN com o SVM, considerado um dos métodos mais robustos da literatura. Os resultados mostram que a geração de características pode melhorar o desempenho de ambos os algoritmos.

2.4 Algoritmo KNN

O KNN classifica um documento com base em seus K vizinhos mais próximos no espaço vetorial. Apesar de simples, o método é sensível à escolha de K e à dimensionalidade dos dados. Entretanto, o KNN tradicional apresenta maior estabilidade frente à variação desse parâmetro quando comparado às suas variações.

2.5 Variações do KNN

kINN (Inverse KNN) Pondera os vizinhos de forma inversamente proporcional à distância, de modo que vizinhos mais próximos tenham maior peso. Essa abordagem melhora a eficácia em coleções como Reuters e Ohsumed.

kSNN (Symmetric KNN) Considera a reciprocidade entre documentos — se A é vizinho de B, então B também deve ser vizinho de A. Essa estratégia aumenta a robustez da classificação, evitando relações unilaterais e reforçando padrões de similaridade bidirecional.

2.6 Critérios de Seleção e Geração de Termos

A proposta também inclui critérios para seleção dos melhores vizinhos com base em medidas de similaridade, gerando novas características que enriquecem a representação textual. Essa abordagem foi testada com SVM e apresentou ganhos estatisticamente significativos, comprovando a eficácia da técnica.

3 Conclusão

A classificação automática de textos continua sendo um dos desafios mais relevantes da mineração de dados. O uso do KNN e suas variações demonstra que, mesmo algoritmos

simples, quando adaptados e combinados com boas representações vetoriais, podem atingir resultados competitivos.

As variações kINN e kSNN apresentaram ganhos relevantes em determinadas bases de dados, especialmente quando aplicadas em conjunto com técnicas de geração de características. Esses resultados reforçam a importância da experimentação e da customização de algoritmos supervisionados para diferentes domínios textuais.

Futuras pesquisas podem explorar a integração dessas técnicas com representações mais modernas, como *embeddings* baseados em redes neurais (Word2Vec, BERT), para aprimorar ainda mais a qualidade da classificação.

GitHub Public:

https://github.com/Viniciushom/Projeto-I.A-II—Classifica-o-Autom-tica-de-Textos-Utilizando-o-Algoritmo-KNN-e-Suas-Varia-es