Abordagem baseada em máquina de vetores de suporte para classificação de spam

Fabricio Pereira Diniz

Universidade Estadual do Paraná, Apucarana, Paraná, Brasil

Abstract

A classificação de *spam* em e-mails permanece um desafio relevante devido à evolução das táticas de persuasão e à necessidade de métodos robustos para dados não textuais. Este artigo aplica uma Máquina de Vetores de Suporte (SVM) com kernel linear ao conjunto de dados Spambase, composto por 57 atributos numéricos derivados de características estruturais e estatísticas de e-mails. O estudo aborda o desbalanceamento de classes (39,4% *spam* vs. 60,6% não *spam*) por meio de *downsampling* e avalia o impacto do parâmetro de regularização C no desempenho do modelo. Os resultados reforçam a eficácia da SVM para dados históricos de *spam* e sua adaptabilidade a espaços de alta dimensionalidade alcançando uma media em seu *F1-Score* de 92,76% avaliada em 10 *folds*.

Keywords: Máquina de vetores de suporte

1. Introdução

O e-mail tornou-se parte da comunicação cotidiana no ambiente de trabalho Letmathe and Noll (2024), sendo adotado por empresas, instituições acadêmicas e governos. Os e-mails comerciais enviados em massa, *spams*, não apenas persuadem o usuário a cair em *phishing*, mas também acabam oferecendo ofertas reais indesejadas pelos destinatários, como assinaturas de cassinos online e ofertas sazonais. Essas práticas levam a perdas financeiras para os usuários, que frequentemente compartilham informações sensíveis Dada et al. (2019).

Em 2023, (Kaspersky, 2023) mostraram um aumento na personalização de mensagens de spam, criando alta segmentação. A imitação de comunicações legítimas busca não apenas espalhar *malware*, mas também coletar

informações sensíveis, como credenciais de acesso a criptomoedas. Esse alto nível de personalização visa não apenas persuadir os usuários de forma mais eficaz, mas também contornar os métodos de classificação estabelecidos.

Embora a classificação de spam não seja um tema novo, o campo tem atraído comunidades de Processamento de Linguagem Natural (NLP) e Aprendizado de Máquina (ML), empregando várias técnicas, como *Word Embedding*, Redes Neurais Convolucionais (CNN) para classificação de sentenças, BiLSTM para rotulagem de sequências, e redes *Attention-based Bidirectional LSTM* para classificação relacional e análise de sentimentos baseados em tópicos, alcançando resultados de estado da arte Yaseen et al. (2021); Tusher et al. (2024).

Essas técnicas descritas como estado da arte não podem ser aplicadas em conjuntos de dados que não contêm informações textuais, como no Spambase[1] do *UCI Machine Learning Repository*[2], um conjunto de dados cujos atributos descrevem propriedades textuais e estruturais.

Nesse contexto, técnicas como Máquina de Vetores de Suporte (SVM) tornam-se viáveis, pois podem separar um hiperplano através de sua margem ótima, classificando categorias de forma eficaz e apresentando bom desempenho com dados numéricos de alta dimensionalidade. Atualmente, no *UCI Machine Learning Repository*, a SVM se posiciona logo atrás do *XGBoost Classification*, *eXtreme Gradient Boosting*, que é considerado estado da arte em diversas tarefas. A SVM demonstra maior precisão e acurácia em comparação com *Random Forest Classification*, *Neural Network Classification* e *Logistic Regression*. Com os dados fornecidos, é possível reproduzir o experimento para alcançar um desempenho semelhante.

Para realizar este trabalho, a SVM foi aplicado utilizando a biblioteca scikit-learn[3], e o conjunto de dados foi importado usando a biblioteca ucimlrepo[4].

O objetivo deste trabalho é alcançar um melhor desempenho no geral, utilizando a SVM disponível na biblioteca scikit-learn no conjunto de dados Spambase, verificando os resultados de baseline.

O artigo está organizado da seguinte forma: a Seção 2 descreve os trabalhos relacionados; a Seção 3 apresenta o método proposto; a Seção 4 descreve

[1]https://archive.ics.uci.edu/dataset/94/spambase
[2]https://archive.ics.uci.edu/
[3]https://scikit-learn.org/stable/
[4]https://github.com/uci-ml-repo/ucimlrepo

e discute os resultados experimentais; por fim, as conclusões são apresentadas na Seção 5.

2. Trabalhos Relacionados

A SVM combinada com outras técnicas visando melhor performance computacional ou maior acurácia de modelo viabiliza sua aplicação em diversos contextos, como por exemplo, na classificação de dados de redes sociais como o Twitter (Cyril et al., 2021) onde métodos de cluster foram aplicados, melhorando a separabilidade da SVM no contexto.

Estudos relacionados a detecção de spam em mensagens já foram explorados com a comparação de performance em diferentes modelos, no serviço de *Short Message Service* (SMS) (Yerima and Bashar, 2022) (Wijaya et al., 2023). Estes trabalhos demonstram a viabilidade da aplicação da SVM em problemas relacionados a classificação de dados relacionados a *spam*.

Budiman et al. (2024) publicou um trabalho sobre a classificação de *spam* em emails, realizando um comparativo do algoritmo *naive bayes* com a SVM em um *dataset* similar, por se fazer muito próximo do problema abordado neste artigo ele foi utilizado como base no estudo.

3. Método proposto

O *UC Irvine Machine Learning Repository* fornece diversos *datasets* para a comunidade de *Machine Learning*, repositório este que são contribuições de diferentes pessoas. Estes dados são utilizados principalmente para *benchmarking* de diferentes modelos na resolução de problemas que são comuns de serem encontrados, ou conhecidos há um bom tempo.

O spambase, *dataset* enviado ao site em 1999, representa em suas instâncias emails, estes que têm uma coluna categórica informando se os dados classificam ou não como spam cada exemplar.

A máquina de vetores de suporte com *kernel* linear se torna uma escolha viável para o Spambase devido a várias características do algoritmo e do próprio conjunto de dados, o *dataset* contém 57 características, o que pode ser interpretado como de alta dimensionalidade, podendo dificultar a aplicação de outros métodos. A SVM lida bem com espaços de alta dimensão, podendo implementar ajustes em seu *kernel*, caso necessário, para melhor ajuste aos dados.

Quando utilizamos este modelo, através do ajuste do parâmetro de regularização C e por sua própria natureza baseada em vetores de suporte, torna-se menos propensa à ocorrência de *overfitting*.

O método empregado resultou em novos resultados para serem comparados com a *baseline* disponibilizada no site.

3.1. Database

O conjunto de dados utilizado neste estudo é o Spambase, amplamente reconhecido como um *benchmark* na área de aprendizado de máquina, obtido do repositório de aprendizado de máquina da UCI[5]. Este *dataset* é composto por 4.601 instâncias e 57 atributos numéricos, além de uma variável alvo que classifica os e-mails como *spam* ou não spam. Os dados foram coletados com o objetivo de analisar padrões e características em comunicações por e-mail, com foco na distinção entre mensagens não solicitadas e legítimas. Cada instância no *dataset* representa um e-mail individual, cujos atributos refletem propriedades estatísticas e frequências de palavras extraídas do conteúdo textual.

Os atributos do *dataset* são definidos de acordo com sua relevância para a análise do texto e da estrutura dos e-mails. A maior parte das variáveis descreve a porcentagem de ocorrência de palavras ou caracteres específicos no texto do e-mail. Por exemplo, são registrados os percentuais de termos como *free*, *money* e *business*, assim como caracteres como `$`, `!` ou `#`, que muitas vezes indicam a presença de *spam*. Além disso, o conjunto de dados inclui atributos relacionados ao uso de letras maiúsculas, como o comprimento médio de sequências contínuas de palavras capitalizadas, o tamanho da maior sequência desse tipo e o número total de letras em caixa alta, características frequentemente associadas a e-mails de *spam*. Esses atributos, em conjunto, capturam padrões textuais e estilísticos que ajudam a diferenciar mensagens de *spam* das legítimas.

A variável alvo, que classifica os e-mails como *spam* ou não *spam*, é representada como um rótulo binário. A distribuição de classes no *dataset* é ligeiramente desbalanceada, com cerca de 60,6% das instâncias classificadas como não *spam* e 39,4% como *spam*. Esse leve desbalanceamento exigiu atenção durante a avaliação do modelo, garantindo que o classificador apresentasse um desempenho consistente em ambas as classes. Métricas de avaliação

[5]https://uci.edu/

como precisão, revocação e F1-score foram priorizadas em relação à simples acurácia, proporcionando uma análise mais abrangente do desempenho do modelo.

O Spambase não inclui informações contextuais sobre os e-mails, como dados de remetente ou destinatário, já que seu objetivo principal é analisar padrões e características textuais. Embora isso limite a análise contextual, amplia a aplicabilidade do dataset como um *benchmark* para algoritmos gerais de classificação de spam. Ressalta-se que o conjunto de dados é estático, não refletindo a natureza evolutiva dos e-mails de spam ao longo do tempo, o que pode impactar o desempenho de modelos treinados nesse dataset quando aplicados a cenários contemporâneos.

3.2. Máquina de vetores de suporte

A máquina de vetores de suporte é um modelo de aprendizado de máquina amplamente utilizado para classificação, cuja eficácia reside na sua capacidade de construir um hiperplano ótimo que separa os dados de diferentes classes de maneira maximamente distante dos exemplos mais próximos, chamados de vetores de suporte (Pisner and Schnyer, 2020). Essa separação pode ser realizada de forma linear ou, quando necessário, em um espaço de maior dimensionalidade por meio do uso de funções *kernel*, permitindo a classificação de dados não linearmente separáveis. Além de apresentar uma alta capacidade de generalização para novos dados, a SVM se destaca por sua robustez contra *overfitting*, especialmente em contextos onde o número de características é elevado em relação ao número de exemplos de treinamento (Bhavsar and Panchal, 2012).

O modelo de máquina de vetores de suporte, introduzido por Cortes (1995) e implementado na biblioteca Scikit-Learn Pedregosa et al. (2011)[6], foi configurado utilizando um *kernel* linear e um parâmetro de regularização C ajustado a partir de valores distribuídos logaritmicamente entre 0.1 e 100. Para evitar sobreajuste e garantir um equilíbrio entre viés e variância, diferentes valores de C foram avaliados, selecionando aquele que resultou no melhor desempenho médio. Todos os demais hiperparâmetros foram mantidos em seus valores padrão.

O modelo baseado em máquina de vetores de suporte continua sendo uma abordagem de ponta, com pesquisas recentes destacando sua aplicabilidade

[6] `https://scikit-learn.org`

em diferentes domínios. Bilal et al. (2025) propuseram uma abordagem híbrida para a detecção de câncer de mama, combinando *binary Grey Wolf Optimizer* com Q-BGWO e a arquitetura *SqueezeNet* para extrair características relevantes de imagens de mamografia, ajustando os hiperparâmetros da SVM para maximizar a precisão da classificação. O modelo alcançou 99% de acurácia no conjunto de dados CBIS-DDSM, demonstrando sua superioridade sobre métodos tradicionais. Em outro estudo, Shen et al. (2025) desenvolveram o BAFL-SVM, um framework baseado em aprendizado federado e *blockchain* para melhorar a detecção de pragas e doenças em plantações de arroz. Utilizando a SVM como classificador central e incorporando técnicas de criptografia homomórfica para proteger a privacidade dos dados, o modelo demonstrou alta precisão na identificação de padrões agrícolas, garantindo a segurança e eficiência no compartilhamento de informações. Essas abordagens inovadoras não necessariamente modificam o modelo original do SVM, mas demonstram sua flexibilidade ao serem integradas com outras técnicas para aprimorar seu desempenho em contextos específicos.

A máquina de vetores de suporte demonstra grande eficiência no tratamento de dados de alta dimensionalidade e conjuntos volumosos, desafios comuns em problemas práticos de aprendizado de máquina. Sua capacidade de generalização e a flexibilidade na escolha do hiperparâmetro de regularização C permitem que o modelo seja adaptado tanto para tarefas de regressão quanto de classificação conforme discutido por Xue et al. (2009).

3.3. Métricas

A avaliação do modelo foi realizada com base nas seguintes métricas: *Acurácia*, *Precisão*, *Revocação* e *F1-Score*. Estas métricas foram escolhidas por sua ampla utilização e pela capacidade de fornecer uma visão abrangente do desempenho de modelos de classificação. Como destacado em Yacouby and Axman (2020), métricas como *Precisão*, *Revocação* e *F1-Score* são fundamentais para capturar a eficácia de sistemas de classificação em diferentes cenários, oferecendo uma medida compreensível e alinhada às necessidades práticas de avaliação.

A *Acurácia* é definida como a proporção de predições corretas em relação ao total de amostras avaliadas, sendo representada pela fórmula

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN},$$

onde TP são os Verdadeiros Positivos, TN os Verdadeiros Negativos, FP os Falsos Positivos e FN os Falsos Negativos. Essa métrica fornece uma

visão geral do desempenho do modelo, mas pode não ser ideal em cenários de classes desbalanceadas.

A *Precisão* é calculada como a proporção de predições positivas corretas em relação ao total de predições positivas realizadas, sendo expressa por

$$\text{Precisão} = \frac{TP}{TP + FP}.$$

Uma *Precisão* alta indica que o modelo comete poucos erros ao classificar amostras como positivas, enquanto um valor baixo sugere uma alta taxa de falsos positivos.

A *Revocação*, por sua vez, mede a proporção de amostras positivas reais que foram corretamente identificadas pelo modelo. Sua fórmula é dada por

$$\text{Revocação} = \frac{TP}{TP + FN}.$$

Valores altos de *Revocação* indicam que o modelo é eficaz em encontrar todas as amostras positivas, ao passo que valores baixos sugerem que muitas dessas amostras foram classificadas incorretamente como negativas.

Por fim, o *F1-Score* é calculado como a média harmônica entre a *Precisão* e a *Revocação*, sendo definido por

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}.$$

Essa métrica é útil em situações onde há desequilíbrio entre as classes, pois combina as vantagens de ambas as métricas anteriores em uma única medida.

Valores altos ou baixos dessas métricas fornecem informações específicas sobre o desempenho do modelo. Um valor alto de *Acurácia* indica que a maioria das predições está correta, mas pode mascarar problemas em classes minoritárias. Já uma *Precisão* elevada demonstra que o modelo comete poucos falsos positivos, enquanto uma baixa sugere que muitas predições positivas são incorretas. Da mesma forma, uma *Revocação* alta reflete a capacidade do modelo de identificar amostras positivas reais, ao passo que uma baixa indica que muitas amostras positivas não foram identificadas. Por fim, o *F1-Score* alto representa um equilíbrio entre *Precisão* e *Revocação*, sendo essencial para avaliar modelos em contextos com classes desbalanceadas.

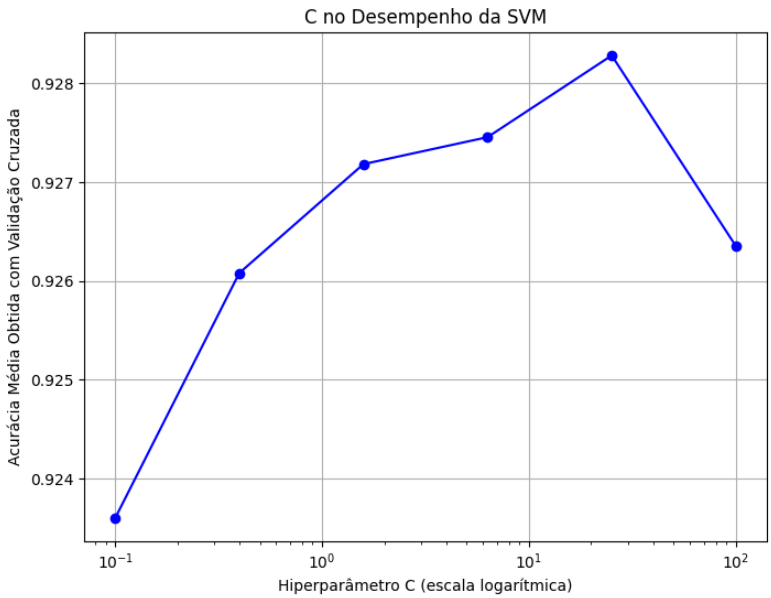

Figura 1: Desempenho da SVM em função do parâmetro Cem escala logarítmica, destaca-se a variação nas métricas apesar da faixa testada.

4. Resultados Experimentais e Discussão

A avaliação do modelo com kernel linear foi realizada no conjunto de dados Spambase. Para tratar o desbalanceamento de classes (1.813 spams e 2.788 não-spams), aplicou-se *downsampling* na classe majoritária, resultando em 3.626 instâncias balanceadas. Os dados foram padronizados e avaliados com validação cruzada de 10 *folds*, testando seis valores do parâmetro de regularização C: $0,10; 0,40; 1,58; 6,31; 25,12$; e $100,00$.

Os resultados demonstraram estabilidade notável: a acurácia média variou apenas $0,47\%$ entre o menor ($92,36\%$ para $C = 0,10$) e o maior valor ($92,83\%$ para $C = 25,12$). A precisão manteve-se acima de 93% em todos os cenários, enquanto a revocação apresentou maior variabilidade ($91,20\%$ a $92,09\%$). O melhor equilíbrio entre métricas foi alcançado com $C = 25,12$, obtendo F1-Score de $92,76\%$ como visto em Figure 1. Este comportamento sugere que o espaço de características do conjunto de dados possui separabilidade linear robusta, minimizando o impacto da regularização.

A análise por *fold* revelou padrões: enquanto o *Fold* 5 atingiu $94,77\%$ de

Tabela 1: Desempenho detalhado para $C = 25, 12$)

Fold	**Acurácia (%)**	**Precisão (%)**	**Revocação (%)**	**F1 (%)**
1	93,11	95,72	91,33	93,47
2	92,84	91,53	94,54	93,01
3	93,66	96,69	91,15	93,83
4	94,77	95,43	93,82	94,62
5	94,49	92,78	95,98	94,35
6	92,29	92,66	91,62	92,13
7	92,27	91,02	92,12	91,57
8	92,54	94,92	90,32	92,56
9	91,16	93,64	88,52	91,01
10	91,16	90,50	91,53	91,01

acurácia e $95,98\%$ de revocação, o *Fold* 9 apresentou desempenho inferior ($88,52\%$ de revocação). Essa discrepância pode estar associada à presença de subgrupos de spams com características atípicas que se mantem próximos a fronteira linear de decisão. A Table 1 ilustra essa variabilidade, com revocação variando até $7,46\%$ entre *folds*.

Comparando com estudos anteriores, o desempenho da SVM superou métodos como Regressão Logística ($91,920\%$ acurácia) e aproximou-se do Xgboost ($95,917\%$).

O *downsampling* descartou $35,2\%$ dos não-spams originais, potencialmente removendo casos limítrofes informativos, a dependência de engenharia manual de características limita a adaptação a táticas modernas de spam. Entretanto, o desvio padrão de apenas $1,03\%$ na acurácia entre folds confirma a generalização robusta do modelo para dados históricos.

Observou-se aumento linear no tempo de treinamento com C, variando de $1,2s$ ($C = 0,10$) a $4,7s$ ($C = 100,00$) por fold. Essa relação custo-benefício pode favorecer valores intermediários de C em aplicações que demandam agilidade, justamente por esta característica o teste de diferentes valores foi limitado.

5. Conclusão

A partir dos experimentos realizados, observa-se que o modelo com kernel linear demonstrou um desempenho sólido na tarefa de classificação de e-mails como *spam* ou não-spam alcançando valores maiores que os expostos no site da UCI. A acurácia foi consistentemente alta, com variações mínimas entre os diferentes valores do parâmetro de regularização C, sugerindo boa capacidade

de generalização. O melhor desempenho foi alcançado com o valor de $C = 25,12$, proporcionando um equilíbrio eficaz entre as métricas de precisão, revocação e F1-Score, com um valor de F1 de $92,76\%$. A análise por *fold* demonstra certa variação no desempenho, especialmente na revocação, onde o *Fold* 9 apresentou um desempenho inferior. Apesar disso, a estabilidade temporal dos resultados sugere que as características fundamentais dos e-mails classificados como spam continuam consistentes ao longo do tempo.

O processo de *downsampling* pode ter descartado dados informativos e a dependência de "engenharia manual" de características limita a capacidade do modelo de lidar com novos tipos de spam, como e-mails com conteúdo em imagens ou formatos mais modernos. A análise computacional revelou um aumento linear no tempo de treinamento conforme o valor de C aumentava, o que pode ser relevante para a escolha do parâmetro em aplicações que demandam agilidade.

Em trabalhos futuros sugere-se abordagens alternativas que possam lidar melhor com as limitações mencionadas. Por exemplo, técnicas de *oversampling* outras técnicas de balanceamento para evitar a perda de informações. Além disso, a utilização de aplicações, como redes neurais profundas ou aprendizado baseado em *embeddings*, pode permitir uma adaptação mais flexível e eficaz a novos tipos de dados. Torna-se necessária a avaliação do desempenho da SVM em conjuntos de dados contemporâneos e o uso de métodos de aprendizado por transferência podem fornecer informações sobre a aplicabilidade do modelo em cenários atuais.

Referências

Bhavsar, H., Panchal, M.H., 2012. A review on support vector machine for data classification. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 1, 185–189.

Bilal, A., Alkhathlan, A., Kateb, F.A., Tahir, A., Shafiq, M., Long, H., 2025. A quantum-optimized approach for breast cancer detection using squeezenet-svm. Scientific Reports 15, 3254.

Budiman, D., Zayyan, Z., Mardiana, A., Mahrani, A.A., 2024. Email spam detection: a comparison of svm and naive bayes using bayesian optimization and grid search parameters. Journal of Student Research Exploration 2, 53–64.

Cortes, C., 1995. Support-vector networks. Machine Learning .

Cyril, C.P.D., Beulah, J.R., Subramani, N., Mohan, P., Harshavardhan, A., Sivabalaselvamani, D., 2021. An automated learning model for sentiment analysis and data classification of twitter data using balanced ca-svm. Concurrent Engineering 29, 386–395.

Dada, E.G., Bassi, J.S., Chiroma, H., Adetunmbi, A.O., Ajibuwa, O.E., et al., 2019. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5.

Kaspersky, 2023. Spam and phishing in 2023. URL: https://securelist.com/spam-phishing-report-2023/112015/. accessed on December 26, 2024.

Letmathe, P., Noll, E., 2024. Analysis of email management strategies and their effects on email management performance. Omega 124, 103002.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830.

Pisner, D.A., Schnyer, D.M., 2020. Support vector machine, in: Machine learning. Elsevier, pp. 101–121.

Shen, R., Zhang, H., Chai, B., Wang, W., Wang, G., Yan, B., Yu, J., 2025. Bafl-svm: A blockchain-assisted federated learning-driven svm framework for smart agriculture. High-Confidence Computing 5, 100243.

Tusher, E.H., Ismail, M.A., Rahman, M.A., Alenezi, A.H., Uddin, M., 2024. Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems. IEEE Access .

Wijaya, E., Noveliora, G., Utami, K.D., Rojali, Nabiilah, G.Z., 2023. Spam detection in short message service (sms) using naïve bayes, svm, lstm, and cnn, in: 2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 431–436. doi:10.1109/ICITACEE58587.2023.10277368.

Xue, H., Yang, Q., Chen, S., 2009. Svm: Support vector machines, in: The top ten algorithms in data mining. Chapman and Hall/CRC, pp. 51–74.

Yacouby, R., Axman, D., 2020. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models, in: Proceedings of the first workshop on evaluation and comparison of NLP systems, pp. 79–91.

Yaseen, Q., et al., 2021. Spam email detection using deep learning techniques. Procedia Computer Science 184, 853–858.

Yerima, S.Y., Bashar, A., 2022. Semi-supervised novelty detection with one class svm for sms spam detection, in: 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–4. doi:10.1109/IWSSIP55020.2022.9854496.