Classificação de subtipos de câncer de mama com dados transcriptômicos estimados e limitados utilizando ensemble learning

Gabriel de Queiroz Sousa
Instituto de Computação - Universidade Federal do Amazonas
E-mail: gabriel.queiroz@icomp.ufam.edu.br
Beatriz Albuquerque Rodrigues
Instituto de Computação - Universidade Federal do Amazonas
E-mail: beatriz.albuquerque@icomp.ufam.edu.br
Fabiola Guerra Nakamura
Instituto de Computação - Universidade Federal do Amazonas
E-mail: fabiola@icomp.ufam.edu.br

1. Introdução

O câncer de mama é uma doença complexa, e sua análise molecular é crucial para um tratamento eficaz. Atualmente, os médicos classificam o câncer de mama em subtipos moleculares (como Luminal A, Luminal B, Her2, Basal) com base em dados de expressão gênica do paciente. Esses dados, frequentemente obtidos via microarrays, apresentam desafios inerentes, como a ocorrência de valores perdidos devido a fatores como poeira ou arranhões nas lâminas[1]. Tais dados omissos, somados à escassez e ao desbalanceamento das classes, podem afetar negativamente a classificação de subtipos de câncer de mama [2].

Em um trabalho anterior, que serve como baseline para esta pesquisa[1], foi proposto um método de estimação de expressões gênicas de câncer de mama com base em correlação. Nesse estudo, aplicou-se o modelo Support Vector Classifier (SVC) para classificar os subtipos de câncer de mama em um conjunto de dados com aproximadamente 120 amostras, foi observada uma considerável dificuldade em distinguir classes como Luminal A (LumA) e Luminal B (LumB) [1].

A literatura recente tem demonstrado que, mesmo com a limitação de datasets, a adoção de estratégias de ensemble pode levar a resultados satisfatórios em tarefas de classificação de subtipos de câncer[3,4,5]. Trabalhos como o de Batool e Byun (2024), por exemplo, utilizaram um classificador de votação com diferentes modelos para obter alta acurácia na classificação de câncer de mama em um dataset considerado pequeno[6].

Sendo assim, esse estudo propõe uma solução inspirada em uma estratégia que a literatura científica chama de "aprendizado em comitê" ou "ensemble learning"[4,5]. A ideia é, ao invés de utilizar um único modelo, unificar diferentes modelos, cada um com sua própria forma de assimilar os dados, assim gerando um modelo comitê diverso para auxiliar na tomada de decisão final.

2. Material e Métodos

Esta seção detalha o protocolo experimental desenvolvido para avaliar a eficácia do nosso modelo de ensemble na classificação de subtipos de câncer de mama. Apresentamos as bases de dados utilizadas, os modelos de classificação investigados e as etapas de avaliação empregadas.

2.1. Base de dados

O modelo foi treinado e testado utilizando um conjunto de dados de expressão gênica de câncer de mama que já havia sido pré-processado com a estimação de valores ausentes. Este conjunto de dados é derivado da base CPTAC (Clinical Proteomic Tumor Analysis Consortium), e filtrado com os 50 genes do PAM50. No total, o conjunto de dados possui 117 amostras, distribuídas em quatro subtipos: Basal (29 amostras), Luminal A (57 amostras), Luminal B (17 amostras) e Her2 (14 amostras)[1].

Tabela 1 – Descrição do conjunto de dados

Descrição	# de genes	Subtipos	# de amostras	Total de amostras
Cptac 2C	23122	Basal	29	
		LumA	57	117
		LumB	17	
		Her2	14	

2.2. Métodos Investigados

A pesquisa investiga o desempenho de classificadores individuais e juntos em um comitê para a tarefa de classificação dos subtipos de câncer de mama. Os classificadores foram escolhidos por relevância e desempenho em problemas de classificação.

2.2.1. Classificadores Individuais

Os classificadores individuais que compõem o comitê são:

- **Support Vector Classifier (SVC):** Escolhido por ser o modelo de referência do baseline.
- **Random Forest Classifier (RF):** Um comitê composto apenas de árvores de decisão.
- **Gradient Boosting Classifier (GB):** Um comitê sequencial, onde cada novo membro corrige os erros do membro anterior.
- **Logistic Regression (LR):** Um modelo simples e popular em problemas de classificação.

2.2.2. Aprendizado em Comitê

Para esse problema, foi utilizado o ***VotingClassifier***, que combina as previsões dos classificadores individuais para uma decisão final. Nessa pesquisa, o comitê utilizou duas estratégias de votação, com pesos iguais, sendo eles:

- **Hard Voting:** A classe final é determinada pela maioria das inferências geradas pelos classificadores individuais.
- **Soft Voting:** A classe final é determinada através da média das probabilidades fornecidas pelos classificadores individuais.

2.3. Etapas de Experimentos Desenvolvidos

2.3.1. Estrutura de Avaliação

Para garantir a confiança nos resultados e evitar o viés de partição dos dados, utilizamos a Validação Cruzada K-Fold, com um total de 5 divisões (folds) para as 117 amostras disponíveis no conjunto de dados.

2.3.2. Métricas de Avaliação

Para avaliar o desempenho dos modelos, utilizamos as seguintes métricas:

- **Acurácia:** A proporção de previsões corretas sobre o total de instâncias.
- **Precisão:** A capacidade do modelo de evitar falsos positivos.
- **Recall:** A capacidade do modelo de identificar corretamente todas as instâncias positivas.
- **F1-Score:** Uma média harmônica entre precisão e recall, é especialmente útil para avaliação de conjuntos de dados com classes desbalanceadas.

3. Resultados e Discussão

Esta seção apresenta os resultados obtidos com os classificadores individuais e com o modelo de "aprendizado em comitê" proposto, seguida por uma análise detalhada do desempenho e das implicações dos achados.

3.1. Resultados

O desempenho dos classificadores individuais (Random Forest, Gradient Boosting, Logistic Regression e Support Vector Classification) e do VotingClassifier, em duas configurações de soft voting e hard voting (com e sem o Gradient Boosting), é sumarizado na Tabela 2. As métricas de acurácia, precisão, recall e FI-Score foram calculadas como a média dos 5 folds da validação cruzada.

Tabela 2 – Resultados da classificação com classificadores individuais e ensemble

Métrica	Random Forest	Gradient Boosting	Logistic Regression	Support Vector Classification	Voting soft + GB	Voting soft	Voting hard + GB	Voting hard
Acurácia	0.932	0.829	0.889	0.906	**0.966**	0.914	0.922	0.906
Precisão	0.873	0.778	0.89	0.917	**0.957**	0.9	0.88	0.896
Recall	0.917	0.791	0.913	0.891	**0.976**	0.91	0.949	0.901
F1-Score	0.885	0.76	0.875	0.891	**0.96**	0.894	0.897	0.885

3.2. Discussão

A análise dos resultados demonstra que o modelo de "aprendizado em comitê" superou significativamente o desempenho dos classificadores individuais, confirmando a hipótese de que a combinação de modelos pode trazer maior acurácia em cenários de dados limitados e estimados.

Inicialmente, havia uma preocupação quanto ao risco de sobreajuste (overfitting) devido ao dataset pequeno e ao desbalanceamento das classes. Contudo, a consistência dos resultados observada através da validação cruzada K-Fold indica que o ensemble resolveu as dificuldades de classificação entre Luminal A e B, generalizando de forma eficaz para todas as classes.

Ao analisar as matrizes de confusão, é notável que os classificadores individuais apresentaram facilidade em classificar corretamente as classes Basal e Her2. Em contrapartida, todos os modelos individuais demonstraram dificuldade considerável em distinguir as classes Luminal A e Luminal B. Essa confusão é evidente pelo número de falsos positivos e falsos negativos entre essas duas classes nas matrizes.

A superioridade do soft voting sobre o hard voting é aparente. No soft voting, uma média das probabilidades geradas por cada modelo é calculada, permitindo que classificadores com alta confiança em suas previsões impactem mais a decisão final. O cenário de soft voting com o Gradient Boosting atingiu a acurácia de 0.966 e F1-Score de 0.96, superando todos os modelos individuais e as configurações de hard voting. O comitê foi capaz de acertar as instâncias que os modelos individuais não conseguiram devido à diversidade dos modelos, pois cada classificador aprende padrões nos dados de forma diferente.

Foi levantada a hipótese de que os modelos individuais podem ter classificado erroneamente as mesmas instâncias. A análise de que o ensemble conseguiu lidar bem com essas classes onde os modelos isolados encontraram dificuldades sugere que a combinação foi eficaz em corrigir esses erros. Possivelmente, esses elementos representam casos de fronteira que a diversidade do comitê de modelos conseguiu resolver de forma eficaz.

4. Conclusões

Neste trabalho, investigamos a aplicação de um modelo de "aprendizado em comitê" baseado em votação para aprimorar a classificação de subtipos de câncer de mama em datasets de expressão gênica que são incompletos e de tamanho limitado. Demonstramos que, embora os classificadores individuais enfrentassem dificuldades na distinção de certas classes, como Luminal A e Luminal B, a combinação estratégica desses modelos resultou em melhorias significativas de desempenho.

O VotingClassifier, particularmente na modalidade soft voting, apresentou acurácia e F1-Score superiores, validando a premissa de que a diversidade de modelos pode mitigar vieses individuais e otimizar a capacidade de generalização em cenários desafiadores.

Como trabalhos futuros, pretendemos explorar outras abordagens como a utilização de modelos de deep learning no ensemble, utilização de transfer learning, testar diferentes datasets e avaliar outro tipo de ensemble como o stacking.

Palavras-Chave: expressão gênica; votingClassifier; dados limitados; câncer de mama; machine learning;

Agradecimentos (Item Não obrigatório)

Gostaria de agradecer à minha orientadora pelas oportunidades providas, que foram fundamentais para o meu desenvolvimento como pesquisador, e pelo compartilhamento de conhecimento nesse curto período juntos. Também gostaria de agradecer aos professores de pós-graduação pelos ensinamentos, que tornaram possível o desenvolvimento dessa pesquisa e que muito agregaram à minha compreensão da área, ao laboratório de bioinformática dos professores Nakamura e ao programa de pós-graduação por permitir o acesso a oportunidades como essa e outras.

5.Referências

1. Rodrigues BA, Neto RM, Nakamura FF, Nakamura EF. Um método de Estimação de Expressões Gênicas de Câncer de Mama com Base em Correlação. Em: Seminário Integrado de Software e Hardware (SEMISH) [Internet]. SBC; 2023 [citado 22 de agosto de 2025]. p. 107–18. Disponível em: https://sol.sbc.org.br/index.php/semish/article/view/25066

2. Adedigba AP, Adeshinat SA, Aibinu AM. Deep learning-based mammogram classification using small dataset. Em: 2019 15th international conference on electronics, computer and computation (ICECCO) [Internet]. IEEE; 2019 [citado 22 de julho de 2025]. p. 1–6. Disponível em: https://ieeexplore.ieee.org/abstract/document/9043186/

3. Albashish D, Almansour N, Abdullah A, Mustafa HM, AlSayyed MR, Alrashdan O. Design an Ensemble Pretrained Deep Learning Model for Classification of Melanoma Skin Cancer Images. Em: 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA) [Internet]. IEEE; 2025 [citado 22 de julho de 2025]. p. 01–7. Disponível em: https://ieeexplore.ieee.org/abstract/document/11013331/

4. Cao-Van K, Minh TC, Minh LG, Quyen TTB, Tan HM. Soft-Voting Ensemble Model: An Efficient Learning Approach for Predictive Prostate Cancer Risk. Vietnam J Comp Sci. novembro de 2024;11(04):531–52.

5. Azad M, Nehal TH, Moshkov M. A novel ensemble learning method using majority based voting of multiple selective decision trees. Computing [Internet]. janeiro de 2025 [citado 22 de julho de 2025];107(1). Disponível em: https://link.springer.com/10.1007/s00607-024-01394-8

6. Batool A, Byun YC. Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm. IEEE Access. 2024;12:12869–82.