Seleção de Genes em Dados de Expressão Gênica por meio de um Ensemble Baseado em Grafos

Alexssander F. Cândido[1], Aline Martins Dias[2], Luiz C. B. Torres[1,2]

[1]Departamento de Computação e Sistemas
Universidade Federal de Ouro Preto (UFOP)
João Monlevade – MG – Brazil

[2]Programa de Pós-Graduação em Ciência da Computação,
Universidade Federal de Ouro Preto (UFOP),
35400-000 – Ouro Preto – MG – Brazil

{alexssander.candido, aline.md}@aluno.ufop.edu.br

luiz.torres@ufop.edu.br

Abstract. *Gene expression datasets present high dimensionality and small sample sizes, making the identification of stable genes challenging. This work proposes an ensemble framework based on the Gabriel Graph (G_G) to analyze structural recurrence of attributes across predictive models. Classifiers are generated through attribute sub-sampling and evaluated using stratified cross-validation. Gene relevance is then assessed according to recurrence among high-performing models. Experiments on biomedical datasets show competitive ROC-AUC results compared with Support Vector Machines reported in the literature. A case study on the Golub leukemia dataset reveals a subset of genes consistently associated with the best-performing models.*

Resumo. *Conjuntos de dados de expressão gênica apresentam alta dimensionalidade e poucas amostras, dificultando a identificação de genes estáveis. Este trabalho propõe um ensemble baseado no Grafo de Gabriel (G_G) para analisar a recorrência estrutural de atributos em modelos preditivos. Classificadores são gerados por subamostragem de atributos e avaliados por validação cruzada estratificada. A relevância dos genes é medida pela recorrência entre os modelos de melhor desempenho. Experimentos em conjuntos de dados biomédicos mostram resultados competitivos de ROC-AUC quando comparados a Support Vector Machines reportadas na literatura. Um estudo de caso no dataset Golub identifica genes consistentemente presentes nos modelos mais eficazes.*

1. Introdução

A classificação molecular da leucemia aguda a partir de dados de expressão gênica constitui um dos marcos fundadores da bioinformática moderna aplicada ao diagnóstico oncológico. No estudo seminal de Golub et al. [Golub et al. 1999], perfis de expressão obtidos por microarranjos foram utilizados para distinguir dois subtipos clinicamente distintos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

de leucemia: *Acute Lymphoblastic Leukemia* (ALL) e *Acute Myeloid Leukemia* (AML). Embora ambos sejam amplamente reconhecidos na prática clínica, sua diferenciação pode envolver múltiplos exames complementares e apresentar desafios diagnósticos em casos limítrofes, isto é, situações em que os perfis de expressão não apresentam separação clara entre as classes, resultando em baixa confiança preditiva. Essa distinção é crítica, uma vez que os protocolos terapêuticos diferem substancialmente entre os subtipos.

O conjunto de dados original consiste em 72 amostras de medula óssea coletadas no momento do diagnóstico, conforme descrito por Golub et al. [Golub et al. 1999], das quais 38 compõem o grupo de treinamento (27 ALL e 11 AML) e 34 compõem o grupo de teste (24 ALL e 10 AML), cada uma descrita por 7129 sondas de expressão gênica. Esse cenário caracteriza um regime de alta dimensionalidade extrema ($d \gg n$), no qual o número de atributos supera amplamente o número de instâncias, impondo desafios estatísticos e computacionais significativos.

Nesse contexto, o problema central deixa de ser apenas a classificação supervisionada e passa a envolver, primordialmente, a seleção de genes relevantes capazes de capturar assinaturas moleculares discriminativas entre ALL e AML. A redução do espaço de atributos não apenas mitiga o risco de sobreajuste, mas também favorece a interpretabilidade biológica e a identificação de potenciais biomarcadores.

Este trabalho aborda o problema sob uma perspectiva geométrica e estrutural. Propõe-se um modelo de classificação baseado no Grafo de Gabriel (G_G) [Gabriel and Sokal 1969], no qual arestas são definidas com base em critérios de proximidade que preservam relações locais entre pontos. Esse modelo é integrado a um arcabouço de *ensemble*, em que múltiplos classificadores são combinados para aumentar a robustez e a capacidade de generalização, sendo a diversidade induzida por múltiplas projeções estruturais do espaço de atributos. Diferentemente de abordagens que utilizam seleção de atributos como etapa preliminar independente, aqui a seleção de genes emerge como consequência da recorrência estrutural dos atributos nos G_G pertencentes ao estrato superior de desempenho preditivo.

Assim, o foco deste estudo recai sobre o conjunto de dados de leucemia previamente descrito, tratado não apenas como *benchmark*, mas como objeto central de investigação. A hipótese subjacente é que a recorrência topológica de genes em G_G estruturalmente estáveis pode revelar subconjuntos discriminativos, conciliando desempenho preditivo e consistência estrutural.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico; a Seção 3 descreve a metodologia proposta; a Seção 4 discute os resultados experimentais; e a Seção 5 apresenta as conclusões.

2. Referencial Teórico

2.1. Seleção de Genes na Classificação de Leucemias

O problema introduzido por Golub et al. [Golub et al. 1999] tornou-se paradigma em bioinformática por explicitar a tensão entre desempenho preditivo e estabilidade de seleção. Em contextos de microarranjos, milhares de genes são mensurados simultaneamente, enquanto o número de amostras permanece limitado. Tal configuração intensifica a variância dos estimadores e amplia o espaço de hipóteses compatíveis com os dados.

Nesse regime, múltiplos subconjuntos de genes podem apresentar desempenho comparável, porém com baixa sobreposição entre si. Essa instabilidade compromete a reprodutibilidade e dificulta a interpretação biológica dos resultados. A seleção de genes passa, portanto, a exigir não apenas poder discriminativo, mas também consistência sob perturbações amostrais.

O conjunto proposto nesse contexto consolidou-se como referência justamente por tornar explícita essa problemática: distinguir ALL de AML com alta acurácia é possível, conforme demonstrado por Golub et al. [Golub et al. 1999]; identificar subconjuntos gênicos estruturalmente estáveis é substancialmente mais desafiador.

2.2. Grafo de Gabriel

O Grafo de Gabriel (G_G) foi introduzido por Gabriel e Sokal [Gabriel and Sokal 1969] como ferramenta para análise de variação geográfica em estudos biológicos. Trata-se de um grafo geométrico definido sobre um conjunto de pontos em um espaço métrico, cuja estrutura captura relações locais de vizinhança com base em um critério puramente geométrico.

Considere um conjunto de pontos $V = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$. Dois vértices x_i e x_j são conectados por uma aresta se, e somente se, a hiperesfera cujo diâmetro é o segmento $\overline{x_i x_j}$ não contém qualquer outro ponto do conjunto em seu interior. Formalmente, a aresta (x_i, x_j) pertence ao G_G se:

$$\delta(x_i, x_j)^2 \leq \delta(x_i, x_k)^2 + \delta(x_j, x_k)^2, \quad \forall x_k \in V,\ k \neq i, j, \tag{1}$$

onde $\delta(\cdot, \cdot)$ representa a distância Euclidiana.

Geometricamente, essa condição equivale a exigir que o círculo (ou hiperesfera em dimensão superior) de diâmetro $\overline{x_i x_j}$ esteja vazio. Esse critério garante que apenas vizinhos geométricos diretos sejam conectados, preservando a estrutura local da distribuição dos dados.

Entre suas propriedades destacam-se a preservação de fronteiras naturais, a adaptação à densidade local e, em duas dimensões, a planaridade. A construção ingênua do G_G possui complexidade $\mathcal{O}(n^3)$, embora otimizações reduzam o custo prático, conforme discutido por Fernandes et al. [Fernandes et al. 2024].

2.3. Classificadores de Margem Larga Baseados em G_G

Uma aplicação relevante do G_G no contexto de aprendizado supervisionado foi proposta por Torres et al. [Torres et al. 2015], que introduziram um classificador de margem larga baseado na estrutura do G_G.

Nesse modelo, após a construção do G_G, identificam-se as chamadas *Support Edges* (SEs), isto é, arestas que conectam vértices pertencentes a classes distintas. Cada SE induz um hiperplano local, posicionado no ponto médio da aresta e ortogonal ao vetor que conecta seus extremos.

Esses hiperplanos atuam como classificadores locais de margem máxima na região correspondente do espaço de atributos. A decisão global pode ser obtida por meio da combinação das decisões locais, frequentemente ponderadas por funções de distância. A ausência de otimização iterativa complexa e a baixa dependência de hiperparâmetros tornam essa abordagem particularmente atrativa em regimes de alta dimensionalidade.

2.4. Diversificação Estrutural via *Ensemble Learning*

Métodos de *Ensemble Learning* [Kunapuli 2023] baseiam-se na construção de múltiplos modelos base e na combinação de suas predições com o objetivo de reduzir erro de generalização.

Seja $h_1(x), h_2(x), \ldots, h_M(x)$ um conjunto de classificadores base. O modelo *ensemble* é definido como:

$$H(x) = \mathcal{A}\big(h_1(x), h_2(x), \ldots, h_M(x)\big), \tag{2}$$

onde $\mathcal{A}(\cdot)$ representa um operador de agregação, como votação majoritária ou média ponderada. A eficácia dessa estratégia depende da diversidade entre os modelos individuais.

Duas estratégias clássicas promovem tal diversidade: *bagging*, baseado em amostragem com reposição, e *random subspaces*, baseado na subamostragem de atributos. Em grafos geométricos, pequenas variações em instâncias ou atributos modificam relações de proximidade, produzindo estruturas distintas de conectividade.

No contexto do G_G, pequenas variações nas instâncias ou nos atributos alteram as relações de proximidade entre os pontos, resultando em diferentes estruturas de conectividade e, consequentemente, diferentes fronteiras de decisão. Essa sensibilidade estrutural torna o G_G naturalmente adequado à construção de *ensembles* diversificados.

3. Metodologia

3.1. Visão Geral

A metodologia proposta baseia-se na construção de classificadores estruturais fundamentados no Grafo de Gabriel (G_G) [Gabriel and Sokal 1969], posteriormente integrados a um arcabouço de *ensemble learning*, com o objetivo de induzir diversidade estrutural e reduzir a variância do modelo. As subseções seguintes formalizam a construção do classificador estrutural e o procedimento de seleção de genes. A Figura 1 ilustra o fluxo geral do método.

3.2. Construção e Representação Computacional do Grafo

Dado um conjunto de dados rotulado $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}$, com $y_i \in \{-1, +1\}$ e $\mathbf{x}_i \in \mathbb{R}^d$, constrói-se o Grafo de Gabriel $G_G = (V, E)$, onde $V = \{\mathbf{x}_i \mid i = 1, \ldots, N\}$ e $E \subseteq V \times V$ é definido pelo critério geométrico estabelecido na Eq. 1. O grafo resultante é representado por uma matriz de adjacência binária simétrica

$$\mathbf{A} \in \{0, 1\}^{N \times N}, \tag{3}$$

com

$$A_{ij} = \begin{cases} 1 & \text{se } (\mathbf{x}_i, \mathbf{x}_j) \in E, \\ 0 & \text{caso contrário,} \end{cases} \quad i, j = 1, \ldots, N. \tag{4}$$

Utiliza-se a implementação vetorizada proposta por Fernandes et al. [Fernandes et al. 2024], garantindo viabilidade computacional.

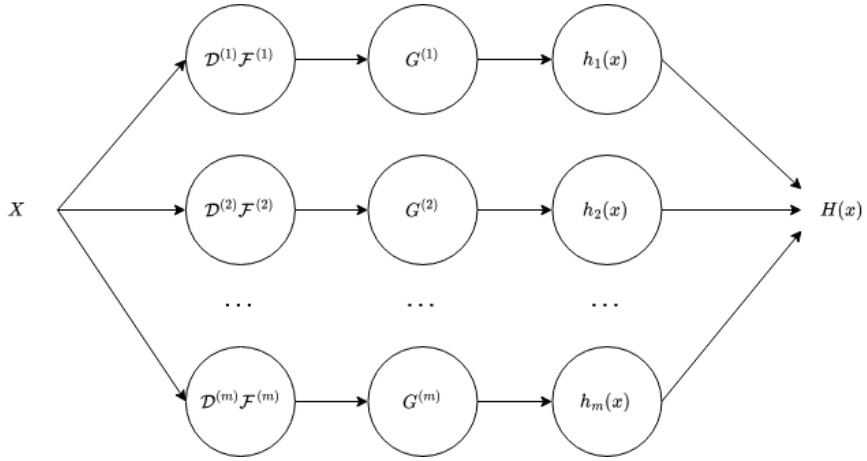

Figura 1. Fluxograma do *ensemble* baseado no Grafo de Gabriel (G_G). A partir de reamostragem de instâncias e projeções em subconjuntos de atributos, induzem-se múltiplos grafos geométricos e seus respectivos classificadores estruturais h_m. Cada modelo produz uma distribuição de probabilidade, e a consolidação preditiva é realizada por *soft-voting*, resultando no estimador final $H(\mathbf{x})$. No experimento de seleção gênica, analisa-se adicionalmente a recorrência estrutural entre modelos de alto desempenho.

3.3. Classificação Baseada em Vizinhança

A classificação de uma instância não rotulada $\mathbf{x}$ baseia-se na estrutura local induzida pelo Grafo de Gabriel (G_G). A decisão depende exclusivamente dos vértices rotulados diretamente adjacentes a $\mathbf{x}$ dentro do G_G.

Seja $G_G = (V, E)$ o Grafo de Gabriel construído sobre o conjunto de instâncias rotuladas $\mathcal{D}$. Para classificar uma nova instância $\mathbf{x}$, considera-se sua inserção em V, estabelecendo-se as arestas $(\mathbf{x}, \mathbf{x}_j)$ que satisfazem a Eq. 1. Define-se a vizinhança de $\mathbf{x}$ como $\mathcal{N}(\mathbf{x}) = \{\mathbf{x}_j \in V : (\mathbf{x}, \mathbf{x}_j) \in E\}$.

Para cada classe $c \in \{-1, +1\}$, define-se $\mathcal{N}_c(\mathbf{x}) = \{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}) : y_j = c\}$, $n_c(\mathbf{x}) = |\mathcal{N}_c(\mathbf{x})|$ e $D_c(\mathbf{x}) = \sum_{\mathbf{x}_j \in \mathcal{N}_c(\mathbf{x})} \delta(\mathbf{x}, \mathbf{x}_j)$, onde $n_c(\mathbf{x})$ representa o suporte estrutural local e $D_c(\mathbf{x})$ a dispersão geométrica agregada da classe c.

O escore discriminativo é definido por

$$S_c(\mathbf{x}) = \frac{n_c(\mathbf{x})}{D_c(\mathbf{x})}. \tag{5}$$

A regra de decisão é dada por

$$\hat{y}(\mathbf{x}) = \arg \max_{c \in \{-1,+1\}} S_c(\mathbf{x}). \tag{6}$$

Para interpretação probabilística, normaliza-se

$$P(c \mid \mathbf{x}) = \frac{S_c(\mathbf{x})}{\sum_{c' \in \{-1,+1\}} S_{c'}(\mathbf{x})}. \tag{7}$$

Caso $\mathcal{N}(\mathbf{x}) = \varnothing$, assume-se distribuição uniforme sobre $\{-1, +1\}$.

Diferentemente do classificador k-Nearest Neighbors (k-NN) [Hastie et al. 2009], no qual a vizinhança é definida como o conjunto dos k pontos mais próximos, resultando em cardinalidade fixa, no modelo baseado no G_G a cardinalidade de $\mathcal{N}(\mathbf{x})$ emerge do critério geométrico da Eq. 1, resultando em grau variável e adaptação à densidade local.

3.4. Ensemble Baseado em Grafos de Gabriel

Com o objetivo de reduzir variância estrutural e aumentar a robustez preditiva, o classificador baseado no Grafo de Gabriel (G_G) é estendido para um modelo de *ensemble*, no qual a diversidade entre os modelos base é induzida por mecanismos de reamostragem aplicados tanto no espaço de instâncias quanto no espaço de atributos. O *ensemble* é composto por M grafos $\{G_G^{(m)} \mid m = 1, \ldots, M\}$, cada um induzido a partir de subconjuntos distintos do conjunto original $\mathcal{D}$.

3.4.1. Bootstrap de Instâncias

Para cada $m \in \{1, \ldots, M\}$, gera-se um subconjunto

$$\mathcal{D}^{(m)} = \text{Bootstrap}_p(\mathcal{D}), \tag{8}$$

onde $\text{Bootstrap}_p(\cdot)$ denota amostragem com reposição de uma fração $p \in (0, 1]$ de instâncias.

3.4.2. Subamostragem de Atributos

Além da reamostragem de instâncias, seleciona-se um subconjunto de atributos

$$\mathcal{F}^{(m)} \subseteq \{1, \ldots, d\}, \quad |\mathcal{F}^{(m)}| = \alpha d, \tag{9}$$

onde $\alpha \in (0, 1]$ representa a fração de atributos utilizados.

Três estratégias são avaliadas:

1. **All**: utiliza-se o conjunto completo de atributos, isto é, $\mathcal{F}^{(m)} = \{1, \ldots, d\}$;
2. **Fixed**: seleciona-se um subconjunto de cardinalidade fixa αd,, amostrado sem reposição;
3. **Var**: selecionam-se subconjuntos de atributos com cardinalidade variável, amostrada aleatoriamente no intervalo $[2, d]$.

3.4.3. Construção dos Modelos Base

Cada par $(\mathcal{D}^{(m)}, \mathcal{F}^{(m)})$ define um espaço projetado específico, sobre o qual é construído

$$G_G^{(m)} = GG\big(\mathcal{D}^{(m)}[\mathcal{F}^{(m)}]\big). \tag{10}$$

O grafo resultante induz um classificador base h_m. A diversidade estrutural do *ensemble* decorre exclusivamente das diferenças em $(\mathcal{D}^{(m)}, \mathcal{F}^{(m)})$.

3.4.4. Estratégia de Consolidação

Sejam $h_1, \ldots, h_M$ os classificadores base induzidos a partir dos grafos $G_G^{(1)}, \ldots, G_G^{(M)}$. Cada modelo h_m fornece, para uma instância $\mathbf{x}$, uma estimativa posterior $P^{(m)}(c \mid \mathbf{x})$, obtida conforme a Eq. 7.

A consolidação do *ensemble* é realizada por *soft-voting*, isto é, pela média aritmética das distribuições posteriores produzidas pelos modelos $\{h_m(\mathbf{x}) \mid m = 1, \ldots, M\}$. Define-se, para cada classe $c \in \{-1, +1\}$,

$$H_c(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} P^{(m)}(c \mid \mathbf{x}), \tag{11}$$

onde $H(\mathbf{x}) = \big(H_{-1}(\mathbf{x}), H_{+1}(\mathbf{x})\big)$ corresponde ao posterior médio do *ensemble*.

A predição final é obtida por

$$\hat{y}(\mathbf{x}) = \arg \max_{c \in \{-1,+1\}} H_c(\mathbf{x}). \tag{12}$$

Escopo da Análise Estrutural O critério baseado em percentil superior de desempenho, introduzido na Seção 3.7, é empregado exclusivamente para análise de recorrência estrutural no conjunto de Golub et al. [Golub et al. 1999]. Tal procedimento não altera o protocolo de consolidação preditiva do *ensemble*, que permanece baseado na agregação probabilística definida nas Eqs. 11–12, sendo utilizado apenas para identificar modelos de alto desempenho com fins de análise de estabilidade estrutural, sem modificar a regra de decisão na avaliação comparativa.

3.5. Reprodutibilidade

A reprodutibilidade foi assegurada por meio de um esquema determinístico de geração de sementes aplicado aos mecanismos de *bagging* e subamostragem de atributos.

Para cada modelo $m \in \{1, \ldots, M\}$, a semente foi definida por

$$\text{seed}_m = \text{SHA1}(random\_state \parallel m)$$

onde $\parallel$ denota concatenação de strings.

Os primeiros 32 bits do hash resultante são interpretados como um inteiro não negativo, utilizado como semente da respectiva iteração.

Esse procedimento garante:

1. Independência determinística entre grafos do *ensemble*;
2. Consistência reprodutível entre execuções independentes;
3. Estabilidade mesmo para diferentes tamanhos de *ensemble*.

Dessa forma, a aleatoriedade inerente à reamostragem é controlada sem comprometer a diversidade estrutural induzida no *ensemble*.

Os experimentos foram conduzidos em ambiente Python 3.11.9, com utilização das bibliotecas NumPy, Pandas e PyTorch, com suporte a aceleração por GPU via CUDA quando disponível. Esse ambiente, aliado ao controle determinístico de sementes, assegura a reprodutibilidade integral dos resultados.

3.6. Configuração Experimental

A configuração experimental operacionaliza integralmente a metodologia proposta nas subseções anteriores, avaliando o *ensemble* baseado em Grafos de Gabriel (G_G) sob diferentes regimes de reamostragem estrutural.

Os experimentos envolveram conjuntos biomédicos do *UCI Machine Learning Repository* [Dua and Graff 2025], bem como os conjuntos de expressão gênica *Golub* [Golub et al. 1999] e *BcrHess* [Hess et al. 2006], caracterizados por alta dimensionalidade e reduzido número de amostras.

Para garantir relevância estatística, adotou-se validação cruzada estratificada com $K = 10$ *folds*. Em cada *fold*, o *ensemble* foi induzido a partir do subconjunto de treinamento, incluindo *bootstrap* de instâncias, subamostragem de atributos e construção dos G_G, conforme descrito anteriormente. O conjunto de teste foi utilizado exclusivamente para inferência e avaliação, assegurando separação estrita entre indução e teste.

A análise investigou combinações entre três eixos estruturais:

1. Número de modelos no *ensemble* $M \in \{5, 9, 15\}$;
2. Fração de instâncias por *bootstrap* $p \in \{0.8, 0.6, 0.4\}$;
3. Subamostragem de atributos: All, Var e Fixed.

A redução progressiva da fração p à medida que M aumenta foi adotada como mecanismo de controle do custo computacional, mantendo simultaneamente diversidade na configuração topológica dos grafos.

Todos os atributos foram previamente normalizados por padronização z-score, utilizando transformação linear, para a faixa $[-1, 1]$, preservando a consistência geométrica das distâncias que governam a construção do G_G.

O desempenho foi quantificado por meio da área sob a curva Receiver Operating Characteristic (ROC-AUC) [Fawcett 2006]. A curva ROC representa a relação entre taxa de verdadeiros positivos e taxa de falsos positivos para diferentes limiares de decisão, sendo a área sob essa curva uma medida robusta de desempenho classificatório independente de limiar. Reporta-se a média obtida ao longo dos $K = 10$ *folds*.

3.7. Seleção Estrutural de Genes no Conjunto Golub

A seleção de genes é conduzida como etapa analítica complementar ao protocolo preditivo, sendo aplicada exclusivamente ao conjunto Golub, caracterizado por alta dimensionalidade e reduzido número de amostras.

A avaliação preditiva é realizada por consolidação probabilística via *soft-voting*, conforme definido na Eq. 11. Entretanto, para fins de análise estrutural, adota-se um critério adicional, fundamentado na recorrência estrutural de genes sob múltiplas projeções associadas a alto desempenho preditivo.

Seja $\mathcal{M}^{(k)} = \{1, \ldots, M\}$ o conjunto de modelos induzidos no k-ésimo *fold*. Define-se o limiar de alto desempenho como

$$Q_{1-\tau}^{(k)} = \text{Quantile}_{1-\tau}\left(\text{ROC-AUC}(h_m^{(k)}) : m \in \mathcal{M}^{(k)}\right), \tag{13}$$

onde $\tau \in (0, 1)$ representa a fração superior considerada.

O conjunto de modelos selecionados no *fold* k é então dado por

$$\mathcal{M}_\tau^{(k)} = \left\{ m \in \mathcal{M}^{(k)} \ : \ \text{ROC-AUC}(h_m^{(k)}) \geq Q_{1-\tau}^{(k)} \right\}, \tag{14}$$

assegurando-se adicionalmente um número mínimo pré-definido de modelos, caso o percentil produza cardinalidade inferior ao limite estabelecido. Cada modelo $m \in \mathcal{M}_\tau^{(k)}$ está associado a um subconjunto de genes $\mathcal{F}^{(m,k)}$.

Define-se a recorrência estrutural intra-*fold* de um gene j como

$$r_j^{(k)} = \frac{1}{|\mathcal{M}_\tau^{(k)}|} \sum_{m \in \mathcal{M}_\tau^{(k)}} \mathbf{1}\left[j \in \mathcal{F}^{(m,k)}\right], \tag{15}$$

a qual representa a frequência relativa com que o gene participa das projeções estruturais de alto desempenho no *fold* considerado.

A recorrência estrutural global é então definida como

$$r_j = \frac{1}{K} \sum_{k=1}^{K} r_j^{(k)}, \tag{16}$$

onde K corresponde ao número de *folds* da validação cruzada.

A métrica $r_j \in [0, 1]$ quantifica a estabilidade estrutural do gene sob múltiplas reamostragens de instâncias e projeções no espaço de atributos, sendo interpretável como a frequência média com que o gene integra modelos estruturalmente robustos ao longo da validação cruzada.

Os genes são finalmente ranqueados em ordem decrescente de r_j, produzindo o conjunto de genes estruturalmente mais recorrentes.

Esse procedimento é empregado exclusivamente para análise de relevância gênica, não alterando o protocolo de avaliação preditiva, que permanece baseado na agregação probabilística do *ensemble*.

4. Resultados e Discussão

4.1. Avaliação Preditiva

A Tabela 1 apresenta o desempenho preditivo do *ensemble* baseado no Grafo de Gabriel (G_G) em comparação às variantes de Support Vector Machines (SVM) com kernels RBF e polinomial reportadas por Torres et al. [Torres et al. 2015], fundamentadas na formulação de margens máximas proposta por Cortes and Vapnik [Cortes and Vapnik 1995]. Conforme já descrito, a avaliação foi conduzida exclusivamente sobre conjuntos de dados biomédicos.

As métricas reportadas correspondem à média aritmética e ao desvio-padrão da ROC-AUC obtidos ao longo dos $K = 10$ *folds* de validação cruzada estratificada. A métrica ROC-AUC foi adotada por ser amplamente utilizada em problemas de classificação biomédica, particularmente em cenários com possível desbalanceamento entre classes. Para cada conjunto de dados e para cada estratégia de subamostragem (*Var*, *All* e *Fixed*), a tabela apresenta apenas a configuração de hiperparâmetros (M, p) que produziu o maior valor médio de ROC-AUC. A coluna Ni/Nd indica, respectivamente, o número de instâncias (N_i) e o número de atributos (N_d) de cada conjunto de dados.

Tabela 1. Comparação de desempenho em termos de ROC-AUC médio $\pm$ desvio-padrão entre o método baseado em G_G (melhores configurações para cada estratégia de subamostragem) e SVMs reportadas por [Torres et al. 2015]. O melhor resultado médio por conjunto está em negrito.

Dataset	Var	All	Fixed	SVM-RBF	SVM-Poly	Ni/Nd
BcrHess	0.90 ± 0.07	0.86 ± 0.07	$\mathbf{0.90 \pm 0.06}$	0.76 ± 0.11	0.77 ± 0.15	133 / 30
B. Cancer W.P.	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.97 ± 0.01	0.96 ± 0.03	683 / 9
Fertility	0.62 ± 0.38	0.64 ± 0.38	$\mathbf{0.66 \pm 0.25}$	0.50 ± 0.00	0.50 ± 0.00	100 / 9
Golub	$\mathbf{0.86 \pm 0.16}$	0.80 ± 0.25	0.85 ± 0.17	0.80 ± 0.16	0.78 ± 0.17	72 / 50
Haberman's S.	$\mathbf{0.63 \pm 0.13}$	0.62 ± 0.12	0.60 ± 0.09	0.52 ± 0.06	0.50 ± 0.02	306 / 3
ILPD	0.68 ± 0.10	0.59 ± 0.05	$\mathbf{0.68 \pm 0.06}$	0.49 ± 0.02	0.50 ± 0.00	579 / 10
Liver Disorders	0.65 ± 0.09	0.55 ± 0.12	0.66 ± 0.10	0.67 ± 0.05	$\mathbf{0.72 \pm 0.07}$	345 / 6
P. Ind. Diabetes	0.74 ± 0.07	0.74 ± 0.05	$\mathbf{0.74 \pm 0.04}$	0.71 ± 0.05	0.71 ± 0.07	768 / 8
Parkinsons	$\mathbf{0.98 \pm 0.03}$	0.95 ± 0.09	0.96 ± 0.04	0.77 ± 0.11	0.81 ± 0.12	195 / 22
Stalog Heart	$\mathbf{0.85 \pm 0.08}$	0.83 ± 0.05	0.84 ± 0.07	0.83 ± 0.07	0.83 ± 0.07	270 / 13

Considerando o maior valor médio de ROC-AUC por conjunto, estratégias baseadas em G_G obtêm o melhor resultado em 9 dos 10 conjuntos avaliados, enquanto SVM-Poly apresenta o maior valor em um conjunto.

Nos conjuntos BcrHess e Golub, observam-se maiores valores médios de ROC-AUC para ao menos uma das estratégias baseadas em G_G em comparação às variantes de SVM reportadas. Em BcrHess, as estratégias *var* e *fixed* atingem 0.90 ± 0.07 e 0.90 ± 0.06, respectivamente, superando SVM-RBF (0.76 ± 0.11) e SVM-Poly (0.77 ± 0.15). No conjunto Golub, a estratégia *var* obtém 0.86 ± 0.16, em contraste com 0.80 ± 0.16 (SVM-RBF) e 0.78 ± 0.17 (SVM-Poly). O desvio-padrão relativamente elevado nesse conjunto é consistente com o regime $d \gg n$, no qual pequenas variações na composição das amostras de treinamento podem produzir variações substanciais na separabilidade entre classes.

Em conjuntos com menor dimensionalidade relativa, como P. Ind. Diabetes e Stalog Heart, as médias de ROC-AUC situam-se em faixas próximas entre os métodos avaliados. No conjunto Liver Disorders, a maior média observada corresponde à SVM-Poly (0.72 ± 0.07).

De forma geral, os resultados indicam desempenho competitivo do *ensemble* baseado em G_G em múltiplos cenários, com variação relativa dependente das características geométricas e estatísticas de cada base.

4.2. Seleção Estrutural de Genes no Conjunto Golub

A análise estrutural foi conduzida sobre o conjunto de Golub et al. [Golub et al. 1999], composto por $n = 72$ amostras e $d = 7129$ genes. O protocolo experimental ado-

tou validação cruzada estratificada com $K = 10$ *folds* e, em cada *fold*, $M = 99$ *bootstraps*. Cada modelo foi construído utilizando a totalidade das instâncias ($p = 1$) e subamostragem fixa de atributos, na qual o subconjunto de genes possui cardinalidade $|\mathcal{F}| = \alpha d = 50$.

O limiar de seleção foi definido pelo percentil $\tau = 90$ da distribuição de ROC-AUC em cada *fold*, sob a restrição $|\mathcal{E}_k^\tau| \geq 3$, onde $\mathcal{E}_k^\tau$ representa o subconjunto de modelos cujo desempenho pertence ao estrato superior no *fold* k. Esse valor de τ foi adotado para selecionar o estrato de modelos com maior desempenho mantendo número suficiente de grafos para estimar recorrência estrutural de forma estável.

A configuração experimental $(K, M, p, \alpha, \tau) = (10, 99, 1, 50/d, 90)$ produziu ACC $= 0.70 \pm 0.19$ e AUC $= 0.69 \pm 0.25$, correspondentes às médias e desvios-padrão obtidos ao longo dos K *folds*.

A fixação de $|\mathcal{F}| = 50$ foi adotada após experimentos preliminares com cardinalidade variável, nos quais subconjuntos extensos de genes produziram AUC $= 1$. Em regime $d \gg n$, tal comportamento é consistente com separabilidade induzida pela alta dimensionalidade. A restrição de cardinalidade limita a complexidade estrutural dos G_G e permite avaliar recorrência gênica sob dimensão controlada.

Aplicando o critério de recorrência definido na Eq. 16, obteve-se $|\{j : r_j > 0\}| = 5096$. A Tabela 2 apresenta os 11 genes com maior recorrência relativa global r_j.

Tabela 2. Genes com maior recorrência estrutural global no conjunto Golub.

Gene	r_j
X15673_s_at	0.0492
U25956_at	0.0462
X56199_at	0.0431
D00003_s_at	0.0403
U30246_at	0.0392
X01703_at	0.0391
HG2530-HT2626_at	0.0389
M80244_at	0.0386
U27109_at	0.0379
U58681_at	0.0366
U16031_at	0.0362

Embora numericamente pequenos, os valores de recorrência refletem estabilidade em um espaço de alta dimensionalidade ($d = 7129$) sob projeções aleatórias repetidas. Nesse contexto, valores acima de $r_j = 0{,}03$ já indicam seleção recorrente em modelos estruturalmente distintos e de alto desempenho.

A análise estrutural separa duas etapas: (i) avaliação preditiva dos modelos individuais via validação cruzada e (ii) quantificação da recorrência r_j restrita aos modelos do estrato superior $\bigcup_{k=1}^{K} \mathcal{E}_k^\tau$. A seleção estrutural decorre, portanto, da frequência relativa de participação dos genes nos grafos associados aos modelos de maior desempenho.

5. Conclusão

Este trabalho apresentou um arcabouço baseado em *ensemble* de classificadores construídos a partir do Grafo de Gabriel (G_G) [Gabriel and Sokal 1969] para análise estrutural

de atributos em conjuntos de dados de alta dimensionalidade. O método explora a recorrência de atributos entre modelos de alto desempenho, permitindo investigar estabilidade estrutural sem depender de um único classificador.

A avaliação experimental em diferentes conjuntos de dados biomédicos demonstrou que a abordagem apresenta desempenho preditivo competitivo em termos de ROC-AUC quando comparada a classificadores Support Vector Machines [Cortes and Vapnik 1995] reportados na literatura.

Como estudo de caso, o framework foi aplicado ao conjunto de expressão gênica de Golub et al. [Golub et al. 1999]. A análise de recorrência estrutural permitiu identificar um subconjunto de genes consistentemente associado aos modelos de melhor desempenho, evidenciando o potencial do método como ferramenta exploratória para investigação de estabilidade de atributos em dados de expressão gênica.

Como perspectivas futuras, destacam-se a análise da robustez estatística das medidas de recorrência, sua integração com métodos clássicos de seleção de atributos e a incorporação de informações biológicas externas para apoiar a interpretação dos genes identificados.

Referências

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dua, D. and Graff, C. (2025). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fernandes, J. G., Hanriot, V. M., and de Padua Braga, A. (2024). Optimizing the gabriel graph construction algorithm. In *Latinx in AI@ NeurIPS 2024*.

Gabriel, K. R. and Sokal, R. R. (1969). A new statistical approach to geographic variation analysis. *Systematic zoology*, 18(3):259–278.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). The elements of statistical learning.

Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology*, 24(26):4236–4244.

Kunapuli, G. (2023). *Ensemble methods for machine learning*. Simon and Schuster.

Torres, L., Castro, C., Coelho, F., Sill Torres, F., and Braga, A. (2015). Distance-based large margin classifier suitable for integrated circuit implementation. *Electronics Letters*, 51(24):1967–1969.