Análise Multivariada de uma Clusterização K-Means e Modelagem com base na Regressão Linear Múltipla

Décio Miranda Filho - RA: 236087

27 de setembro de 2023

1 Introdução

A estatística multivariada é uma área que propicia analisar assuntos distintos entre múltiplas variáveis. O composto enzima 5-lipoxigenase (5-LOX) desempenha um papel crucial nas etapas iniciais da produção de mediadores inflamatórios. A disfunção do 5-LOX está fortemente associada a diversas condições inflamatórias, como asma, doença pulmonar obstrutiva crônica (DPOC), artrite, psoríase e aterosclerose. Por conta dessa importância há um interesse crescente em descobrir novos inibidores do 5-LOX que sejam mais seguros e eficazes. Com isso, esse trabalho baseou-se no estudo de Andrada et al., 2015 com a aplicação dos métodos de clusterização e de regressão multivatriada sobre o conjunto de dados representante dessa enzima.

Neste estudo analisado fora propostos uma análise abrangente de Relação Quantitativa Estrutura-Atividade $QSAR$ para uma série de compostos com atividade inibitória do 5-LOX. Este método QSAR aproveita-se da técnica de regressão ou classificação para entender a premissa entre estrutura e atividade do efeito dessas moléculas. Consoante, o artigo citado aproveita-se de duas abordagens estatísticas impresicndíveis e é com base nelas que o presente trabalho pretende analisar.

Mediante isso, planejou-se utilizar duas técnicas estatísticas essenciais: a clusterização K-means e a regressão multivariada. A clusterização foi utilizada para auxiliar na identificação de clusters ou grupos de compostos com características estruturais semelhantes, permitindo uma abordagem mais precisa na seleção de conjuntos de treinamento e teste. Já a regressão multivariada, por sua vez, será fundamental para compreender as relações entre as propriedades moleculares e a atividade inibitória do composto multivariadamente.

Dessa forma, o estuydo buscou não apenas desenvolver um modelo preditivo, mas também explorar as vantagens da clusterização K-means e da regressão multivariada para aprofundar nossa compreensão das relações entre a estrutura molecular e a atividade biológica dos inibidores do 5-LOX.

2 Descrição Matemática do K-Means e da Regressão Multivariada

2.1 Clusterização K-Means

O algoritmo K-Means é uma técnica amplamente utilizada em análise de dados para particionar um conjunto de dados em grupos distintos, chamados de clusters. A ideia central por trás do K-Means é agrupar pontos de dados com base em sua proximidade espacial em um espaço de características multidimensional sendo eficiente e adequado para grandes volumes de dados.

Dado um conjunto de dados $\mathbf{X}$ com n observações, onde cada observação é representada por um vetor $\mathbf{x}_i$ de m características, o objetivo do K-Means é encontrar k centróides $(\mu_1, \mu_2, \ldots, \mu_k)$ de tal forma que a soma das distâncias quadráticas entre os pontos de dados e os centróides mais próximos seja minimizada.

A função objetivo do K-Means pode ser expressa como:

$$J(\mathbf{X}, C) = \sum_{i=1}^{n} \sum_{j=1}^{k} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

Onde:

- J é a função objetivo a ser minimizada.
- $\mathbf{X}$ é o conjunto de dados.
- n é o número de observações.
- k é o número de clusters desejado.
- C é o conjunto de centróides $(\mu_1, \mu_2, \ldots, \mu_k)$.
- $\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$ é a distância Euclidiana ao quadrado entre um ponto de dados $\mathbf{x}_i$ e um centróide $\boldsymbol{\mu}_j$.

Começamos selecionando aleatoriamente k pontos de dados do conjunto $\mathbf{X}$ como os centróides iniciais.

Cada ponto de dados é atribuído ao cluster representado pelo centróide mais próximo. Isso é feito calculando as distâncias entre cada ponto de dados e todos os centróides e atribuindo o ponto ao cluster do centróide mais próximo. Após a atribuição inicial, os centróides de cada cluster são recalculados como a média dos pontos de dados atribuídos a esse cluster:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

Onde C_j é o conjunto de pontos de dados atribuídos ao cluster j.Os passos 2 e 3 são repetidos iterativamente até que os centróides não se movam significativamente ou um número máximo de iterações seja atingido.O resultado final é um conjunto de k clusters, onde cada cluster é representado por um centróide, e cada ponto de dados pertence a um dos clusters.

2.2 Regressão Multivariada

A regressão multivariada é uma técnica usada para modelar relações complexas entre múltiplas variáveis independentes e uma variável dependente. É uma extensão natural da regressão linear simples, na qual se permite que mais de uma variável independente influencie a variável dependente. Este método é amplamente utilizado em ciências sociais, econômicas, biológicas e muitos outros campos para entender e prever fenômenos multidimensionais.

Para o caso geral, onde temos um conjunto de dados com n observações, cada uma com p variáveis independentes $(X_1, X_2, \ldots, X_p)$ e uma única variável dependente Y. A regressão multivariada assume o seguinte modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Onde:

- Y é a variável dependente que queremos prever.
- β_0 é o intercepto.
- $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes de regressão que indicam o efeito das variáveis independentes em Y.
- $X_1, X_2, \ldots, X_p$ são as variáveis independentes.
- ε é o termo de erro que representa a variação não explicada por nosso modelo.

O objetivo da regressão multivariada é estimar os coeficientes β_i de tal forma que o modelo se ajuste melhor aos dados.

Os coeficientes β_i são geralmente estimados usando o método dos mínimos quadrados ordinários (OLS). A ideia é minimizar a soma dos quadrados dos resíduos, ou seja, as diferenças entre os valores observados e os valores previstos pelo modelo. Isso é formulado como um problema de otimização:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}))^2$$

A solução para este problema é dada pela fórmula dos mínimos quadrados:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Onde:

- $\hat{\beta}$ é o vetor de coeficientes estimados.
- X é a matriz de dados independentes, onde cada linha corresponde a uma observação e cada coluna corresponde a uma variável independente.
- Y é o vetor de valores da variável dependente.

Uma vez que os coeficientes são estimados, é importante avaliar a qualidade do modelo. Isso é geralmente feito calculando-se várias estatísticas, como o coeficiente de determinação (R^2), que indica a proporção da variância em Y que é explicada pelas variáveis independentes.

Ainda outras estatísticas, como o teste F, podem ser usadas para verificar a significância global do modelo, enquanto os testes t podem avaliar a significância individual dos coeficientes de regressão.

3 Objetivo e Descrição dos Dados

3.1 Conjunto de Dados

Esse trabalho analisou o estudo, que utilizou um conjunto de dados composto por 58 derivados de 5 características da molécula 5-LOX. A atividade inibitória é representada por valores, indicando a concentração necessária para inibir 50% da atividade da 5-LOX. Os valores de IC50 foram transformados em log(1/IC50) e serviram como variável dependente nas investigações. A partir disso, nota-se um cuidado no preparo para avaliação das amostras.

Assim,a clusterização K-means foi utilizada para dividir o conjunto de dados em conjuntos de treinamento e teste com base nos descritores moleculares. Segundo o trabalho,foram testados diferentes valores (2, 3 e 4) para o parâmetro k. O objetivo era explorar a possibilidade de dividir o conjunto de dados em 2, 3 ou 4 clusters. Portanto, a clusterização serviu como uma forma de agrupamento de dados para um pré-processamento da regressão multivariada, a qual será analisada.

O artigo ilustra que o conjunto de dados foi dividido em conjuntos de treinamento e teste, com 80% dos dados usados para treinamento e 20% para teste. Foram exploradas 31 combinações diferentes de conjuntos de treinamento e teste. Todos os modelos QSAR foram desenvolvidos usando o Método de Substituição (RM) para seleção de descritores moleculares.

Para validar os modelos, foi utilizado um conjunto de teste separado que não fez parte do conjunto de treinamento. Além disso, o modelo QSAR ótimo foi validado usando a validação cruzada e y-randomização. Foram realizadas 10.000 simulações de y-randomização.

Assim, entende-se que os dados foaram obtidos a partir das moléculas químicas, realizou-se uma análise de clusterização K-means e houve um preparo entre treino e teste para que a etapa seguinte fosse realizada.

4 Resultados e Discussão

Antes de conduzir um estudo foi essencial ter um conjunto de dados com uma distribuição ampla e simétrica dos valores de atividade biológica em torno da média. Graças a isso, o trabalho realizou uma análise da dispersão dos dados.O estudo em questão considerou um valor limite de $\pm 1{,}5$ vezes o desvio padrão, seis valores saíram desse intervalo . Esses compostos foram identificados como outliers, e o conjunto de dados foi reduzido para 52 compostos (n=52)

4.1 Análise de Clusterização K-means

O método de clusterização K-means foi usado para explorar a possibilidade de dividir o conjunto de dados em clusters. O conjunto de dados foi dividido em dois, três e quatro clusters (k = 2, 3 e 4). Os valores de silhueta variam de +1 (indicando clusters bem separados) a -1 (pontos provavelmente classificados incorretamente), com 0 representando pontos que não foram atribuídos a nenhum cluster.

O gráfico para dois clusters sugere que a maioria dos pontos em cada cluster tem valores de silhueta altos, indicando boa separação.Já aa divisão em três clusters resulta em um cluster com valores de silhueta positivos altos (cluster 2) e dois clusters (clusters 1 e 3) com muitos valores de silhueta baixos e alguns valores negativos, sugerindo uma separação fraca. Com isso, quando o conjunto de dados é dividido em quatro clusters, os resultados permanecem sub-ótimos, com dois clusters mostrando principalmente valores de silhueta baixos e alguns valores negativos. Com base nesses resultados, fica evidente que o conjunto de dados naturalmente forma dois clusters distintos, com compostos dos clusters 1 e 2. Com a dada descrição percebe-se a importância de analisar detalhadatamente a divisão e como é feita a clusterização.

Fora também aplicada uma análise de discriminante em que cinco funções discriminantes diferentes foram analisadas usando de 1 a 5 variáveis independentes para minimizar o desvio padrão da função. O valor R^2 para este modelo é de 0,971, indicando uma forte capacidade preditiva.

O descritor topológico piID (ID) é um número ponderado de caminho que leva em consideração múltiplas ligações na molécula.

Esses resultados destacam a eficácia do método do piID como um descritor molecular chave na criação de uma função discriminante que separa com sucesso os compostos em clusters com base em suas características estruturais.

4.2 Desenvolvimento do Modelo

Os modelos citados anteriormente foram categorizados em três séries:

Série 1: Testados com todos os compostos do cluster 1 como conjunto de teste e todos os compostos do cluster 2 no conjunto de treinamento. Série 2: Testados com todos os compostos do cluster 2 e três compostos do cluster 1 no conjunto de teste, e o conjunto de treinamento consistia apenas de compostos do cluster 1. Série 3: Utilizou um conjunto de teste com 20% dos compostos de ambos os clusters, com 80% atribuídos ao conjunto de treinamento. Os valores médios dos parâmetros estatísticos indicaram que a Série 3 tinha a maior capacidade preditiva. As Séries 1 e 2 tinham suas limitações, seja na calibração ou validação, devido à distribuição de compostos entre os conjuntos de treinamento e teste. A Série 3 demonstrou melhores parâmetros estatísticos.

Todos os modelos desenvolvidos cumpriram a diretriz geral em QSAR que sugere ter pelo menos seis ou sete pontos de dados por descritor. As Séries 1 e 3 apresentaram modelos QSAR mais simples com alta capacidade preditiva.

O Modelo 23 da Série 3 foi selecionado como o modelo QSAR mais preditivo. Este modelo exibiu valores elevados de parâmetros de calibração e validação (Rtrein = 0.811 e Rteste = 0.801, respectivamente).

$$
\begin{aligned}
Rtrein &= 0.811 & R^2trein &= 0.658 \\
Strain &= 0.307 & Rloo &= 0.746 \\
Sloo &= 0.352 & Rteste &= 0.801 \\
R^2teste &= 0.643 & Steste &= 0.333 \\
Rl20\%o &= 0.645 & Sl20\%o &= 0.441 \\
Srand &= 0.400 &
\end{aligned}
$$

Os resultados de validação mostraram que o modelo QSAR desenvolvido é preditivo, com coeficientes de regressão superando o valor aceito e demonstrando uma boa correlação entre os valores de atividade previstos e experimentais.

Os valores previstos de log(1/IC50) para os compostos estão listados, que mostra uma forte correlação entre os valores de atividade previstos e experimentais para compostos nos conjuntos de treinamento e teste. A validação também indica a qualidade do modelo.

5 Conclusão

Em conclusão, este estudo desenvolveu com sucesso um modelo QSAR com boa capacidade preditiva que pode ser valioso para prever a atividade de inibição da 5-LOX de seus novos derivados. As principais descobertas e conclusões deste trabalho são as seguintes: Foi estabelecido um modelo QSAR robusto, demonstrando uma forte capacidade de prever a atividade de inibição da 5-LOX de compostos. Esses descritores fornecem insights sobre as características estruturais e eletrônicas das moléculas que influenciam sua atividade biológica.

Sobre o método K-means foi empregado para criar conjuntos de treinamento e teste representativos. Essa abordagem foi interessante para a seleção de dados demonstrou a estabilidade e confiabilidade do modelo, em contraste com os métodos aleatórios de divisão de dados.Além disso, a modelagem com a regressão múltipla mostorou-se eficaz. Com isso, a estabilidade e validação os parâmetros estatísticos do modelo indicam conformidade entre os resultados de validação. As informações fornecidas neste estudo podem servir como base para pesquisas futuras no desenvolvimento de potenciais inibidores da 5-LOX. O modelo QSAR estabelecido pode ser aplicado para prever a bioatividade de novos compostos, auxiliando no projeto e descoberta de novos agentes farmacêuticos.

Em resumo, a pesquisa analisada contribuiu sumariante para a nossa compreensão da da atividade preditiva bem como o uso das técnicas multivariadas sobremaneira utilizadas.

6 Críticas, Vantagens e Desvantagens

O modelo discutido apresenta várias vantagens e desvantagens: **Vantagens:**

O modelo demonstrou uma alta capacidade de predição, conforme evidenciado pelos altos valores de R^2 e outros parâmetros de validação. Isso significa que ele pode ser usado com confiança para prever a atividade biológica de novos compostos. O modelo é relativamente simples, usando apenas quatro descritores moleculares para prever a atividade biológica. Isso o torna fácil de entender e interpretar, o que é crucial em pesquisa farmacêutica e química medicinal. Ademais, O modelo foi submetido a várias formas de validação, incluindo leave-one-out, leave-more-out, conjunto de teste e y-randomização. Essa validação rigorosa aumenta a confiabilidade do modelo.

Desvantagens:

Embora o modelo cumpra a diretriz geral de ter pelo menos seis ou sete pontos de dados por descritor, o conjunto de dados ainda é relativamente pequeno. Um conjunto de dados maior e mais diversificado pode melhorar a robustez do modelo. A abordagem atual utiliza clusters fixos (K-means) para dividir o conjunto de dados. Esses clusters são determinados pela estrutura dos compostos, o que pode não refletir totalmente a diversidade biológica subjacente.

6.1 Melhorias Possíveis

A Coleta de mais dados experimentais de atividade biológica para uma variedade maior de compostos pode melhorar a robustez do modelo e sua capacidade de generalização. Considerar a inclusão de mais descritores moleculares, como descritores físico-químicos e propriedades conformacionais, pode ajudar a capturar nuances adicionais nas relações entre estrutura e atividade. Realizar validação externa com conjuntos de dados independentes pode verificar ainda mais a capacidade de predição do modelo em diferentes condições.Além da Análise Discriminante Linear, a exploração de outros algoritmos técnias multivariadas, como modelos de regressão não linear, redes neurais e métodos baseados em árvores, pode melhorar a capacidade do modelo.

Acredita-se que essas e outras melhorias potenciais podem contribuir para aprimorar ainda mais a qualidade e a aplicabilidade.

7 Exemplo Simulado da Técnica

Com n=52 e p = 5 realizou-se as análises estatísticas a devida clusterização com K-Means. Com o número de clusters igual a 4 realizou-se a regressão multivariada e análise inferencial dos resultados a partir do númeor de clusters. O algoritmo K-Means é eficaz na clusterização de dados, contudo sensível à inicialização dos centróides iniciais, o que pode levar a diferentes soluções. Portanto, sempre recomenda inicializar com diferentes valores e escolher aquele cujos resultados obtidos são os mais

Tabela 1: Resumo da regressão

Categoria	**Estimate**	**Std. Error**	**t value**
(Intercept)	9.5870	19.4279	0.493
r_a	2.4609	0.4016	6.128
r_b	1.9443	1.0116	1.922
r_c	0.5444	0.9317	0.584
r_d	6.9539	1.7442	3.987

Tabela 2: Residuals

Quantile	**Value**
Min	-12.6761
1Q	-3.4370
Median	0.0723
3Q	3.8241
Max	10.3304

consistentes. Seguem os principais resultados obtidos com a aplicação das técnicas do dataset, para mais detalhes olhar o documento .R.

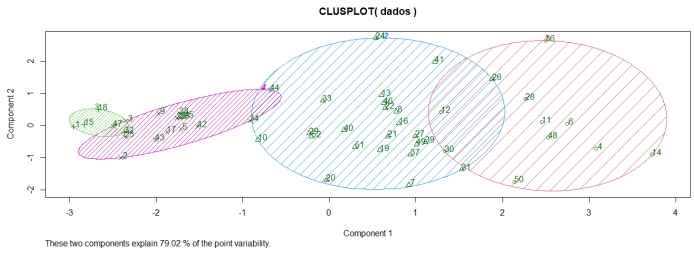

Figura 1: Plot Gráfico entre as 2 componentes

De acordo com os resultados da análise de regressão:

Nota-se coeficiente de interceptação (*Intercept*) é igual a 9.5870, com um erro padrão de 19.4279 e um valor t de 0.493. No entanto, o valor p associado é 0.62883, indicando que o intercepto não é estatisticamente significativo.

Para as variáveis independentes observa-se que coeficiente estimado para r_a é 2.4609, com um erro padrão de 0.4016 e um valor t de 6.128. O valor p é muito pequeno (1.93e-05), indicando uma forte associação entre r_a e a variável de resposta. Consoante, para r_b é 1.9443, com ume erro padrão de 1.0116 e um valor t de 1.922. O valor p é 0.07381, indicando uma possível associação entre r_b e a variável de resposta, embora não seja altamente significativa. Já o coeficiente estimado para r_c é 0.5444, com um erro padrão de 0.9317 e um valor t de 0.584. O valor p é 0.56770, indicando que r_c não é estatisticamente significativo. Por fim, O coeficiente estimado para r_d é 6.9539, com um erro padrão de 1.7442 e um valor t de 3.987. O valor p é 0.00119, sendo estatisticamente significativo

O modelo em questão tem um R-quadrado de 0.7974, o que indica que aproximadamente 79.74% da variabilidade na variável de resposta é explicada pelas variáveis independentes r_a, r_b, r_c e r_d. Isso sugere um ajuste razoavelmente bom do modelo aos dados. e assim como R-quadrado ajustado (0.7433) que leva em consideração o número de variáveis independentes no modelo também aponta um bom ajuste. A estatística F é 14.76, com 4 e 15 graus de liberdade, e o valor p associado é 4.407e-05. Isso sugere que o modelo de regressão como um todo é estatisticamente significativo.

Os resíduos do modelo apresentam uma distribuição que varia de -12.6761 a 10.3304. O resíduo mínimo e máximo são indicativos de como o modelo se ajusta aos dados.

Portanto, os resultados da regressão indicam que as variáveis r_a e r_d têm uma forte associação com a variável de resposta, enquanto r_b tem uma associação menos significativa. A variável r_c não parece ser significativamente relacionada à variável de resposta. O modelo de regressão como um

todo é estatisticamente significativo e explica uma porção significativa da variabilidade na variável de resposta.

Resumidamente, o K-Means é uma valiosa ferramenta de análise de dados para descobrir estruturas ocultas em conjuntos de dados, comumente aplicada em tarefas de agrupamento e segmentação. A regressão multivariada é uma técnica estatística poderosa que modela relações complexas entre várias variáveis independentes e uma variável dependente. É uma extensão da regressão linear simples, permitindo que várias variáveis independentes afetem a variável dependente. É amplamente empregada em diversas áreas, como ciências sociais, econômicas e biológicas, para compreender e prever fenômenos multidimensionais.

8 Considerações Finais

A regressão multivariada é uma técnica versátil que exige cuidado na escolha das variáveis independentes e na interpretação dos resultados. É uma ferramenta poderosa para modelar e compreender relações em conjuntos de dados multidimensionais, tornando-se uma pedra angular na análise de dados em várias disciplinas científicas e na tomada de decisões em negócios e pesquisa.

Aoaplicar a regressão multivariada, é importante considerar a validade das suposições do modelo e realizar uma análise adequada dos resíduos para garantir que o modelo seja apropriado para os dados em questão.

Na Estatística, a regressão multivariada desempenha um papel crucial na construção de modelos preditivos. Permite a inclusão de múltiplas variáveis explicativas, levando em consideração relações complexas nos dados. Por exemplo, em finanças, pode ser usado para prever o preço de uma ação com base em vários indicadores econômicos. Na biologia, pode ser usado para entender como várias variáveis genéticas afetam uma característica específica de um organismo.

9 Referências

Andrada, M. F., Vega-Hissi, E. G., Estrada, M. R., Garro Martinez, J. C. (2015). Application of k-means clustering, linear discriminant analysis and multivariate linear regression for the development of a predictive QSAR model on 5-lipoxygenase inhibitors. Chemometrics and Intelligent Laboratory Systems, 143, 1-7. https://doi.org/10.1016/j.chemolab.2015.03.002

Everitt, B. and Dunn, G.. Applied Multivariate Data Analysis. Segunda Edição, Wiley Sons, nova Iorque, 2001

Everitt, B.S., and T. Hothorn. 2006. A handbook of statistical analyses using R. Chapman Hall/CRC, Boca Raton, LA.

Koch, I.. Analysis of Multivariate and High Dimensional Data: Theory and Pratice. Cambridge University Press, Nova Iorque, 2014

Mardia, K.V., Kent, J.T. and Bibby, J.M.. Multivariate Analysis. Sétima Reimpressão. Academic Press, Londres, 2000