Classificação de moléculas Biodegradáveis

André F. Guimarães[1], João P.S. Mendes[1], Maria E.D. Camargos[1], Nicolas V. Galindo[1]

[1]Departamento de Computação – Universidade Federal de Ouro Preto (UFOP)

andre.guimaraes@aluno.ufop.edu.br, joao.siqueira@aluno.ufop.edu.br

maria.camargos@aluno.ufop.edu.br, nicolas.galindo@aluno.ufop.edu.br

Abstract. *This meta-article describes the application and evaluation of algorithms to classify the biodegradation of a molecule based on attributes of the chemical structure[Mansouri et al. 2013]. The database that will be used is QSAR biodegradation Data Set present in UCI Machine Learning Repository. From these data, genetic algorithm runs combined with classification algorithms, present in the sklearn library, were performed to search for the most relevant attributes. At the end, a comparison of the performance of the classification of molecules with and without attribute selection was performed.*

Resumo. *Este meta-artigo descreve a aplicação e avaliação de algoritmos para classificar a biodegradação de uma moléculas a partir de atributos da estrutura química[Mansouri et al. 2013]. A base de dados que será utilizada é QSAR biodegradation Data Set[1] presente no UCI Machine Learning Repository. A partir destes dados, foram realizadas execuções de algoritmo genético combinado com algoritmos de classificação, presentes na biblioteca sklearn, para buscar os atributos mais relevantes. Ao final, foi realizado uma comparação de desempenho da classificação das moléculas com e sem seleção de atributos.*

1. Algoritmos de Classificação

Para classificação do conjunto de dados, foram realizadas execuções considerando três algoritmos: Regressão Logística, Árvore de Decisão, Árvore Aleatória.

1.1. Regressão Logística

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomador por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binarias[de Azevedo Gonzales 2018].

Para este trabalho, utilizamos a uma classe importada da biblioteca sklearn chamada LogisticRegression. Através desta classe foi possível criar os modelos a partir de um conjunto de dados e rótulos. A execução do algoritmo teve como configuração: C=10.

1.2. Árvore de Decisão

A árvore de decisão, ou decision trees, está entre os métodos mais comuns aplicados ao aprendizado de máquina. Tais algoritmos subdividem progressivamente os dados em conjuntos cada vez menores e mais específicos, em termos de seus atributos, até atingirem um

[1]Disponível em: http://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation

tamanho simplificado o bastante para ser rotulado. Para isso é necessário treinar o modelo com dados previamente rotulados, de modo a aplicá-lo a dados novos.[Casali 2020]

Para este trabalho, utilizamos a uma classe importada da biblioteca sklearn chamada tree. A execução do algoritmo seguiu as configurações padrões oferecidas pela biblioteca sklearn.

1.3. Árvore Aleatória

Uma floresta aleatória é uma técnica de aprendizado de máquina usada para resolver problemas de regressão e classificação. Ele utiliza aprendizagem por conjunto, que é uma técnica que combina muitos classificadores para fornecer soluções para problemas complexos. Um algoritmo de floresta aleatório consiste em muitas árvores de decisão. [Mbaabu 2020]

Para este trabalho, utilizamos a uma classe importada da biblioteca sklearn.ensemble chamada RandomForestClassifier. A execução do algoritmo seguiu as configurações padrões oferecidas pela biblioteca sklearn.

2. Algoritmo Genético

O Algoritmo Genético representa uma classe de algoritmos de otimização que empregam mecanismo de pesquisa probabilístico de soluções, baseado no processo de evolução biológico, combinando aspectos da mecânica da genética e da seleção natural de indivíduos.[Holland 1975][De Jong 1975]

Para aplicação em modelos de classificação, através de gerações são ranqueados as melhores acurácias através de um torneio entre o conjunto de modelos. Os vencedores continuam para próxima geração e os perdedores dão espaço para uma nova população a partir dos vencedores.

A execução do algoritmo genético pode considerar uma taxa de mutação para buscar novos caminhos. Para isto, é realizada uma verificação de sorte a cada geração de um novo individuo, para verificar se este sofreu mutação.

3. Base de dados

Para este trabalho, foi utilizado como base de dados o conjunto de dados *Qsar biodegradation data set*. Este é um conjunto de dados contendo valores para 41 atributos (descritores moleculares) usados para classificar 1055 produtos químicos em 2 classes: biodegradáveis e não biodegradáveis[Mansouri et al. 2013].

O conjunto de dados original possui classes desbalanceadas com 699 observações não biodegradáveis e 352 biodegradáveis e, por isso, foi realizado uma seleção, aleatoriamente, de 400 exemplos da classe majoritária.

Este dados foram separado em conjunto dois conjunto de dados: Dado e Rótulo. Em seguida, foi realizada a normalização de todos os valores presentes no conjunto Dado para que todos os atributos compreendessem entre zero e um. Para o conjunto de Rótulos, as campos de texto foram substituídos, respectivamente,de ”NRB”e ”RB”para ”0”e ”1”.

4. Metodologia

O trabalho teve como proposta um algoritmo que fosse capaz de descobrir os melhores atributos que classificam se uma molécula é biodegradável ou não através de busca genética e algoritmos de classificação.

Os algoritmos de classificação considerados foram: Regressão logística, Árvore de decisão e Árvore aleatória.

A execução da busca através de algoritmo genético pelo melhor modelo tiveram alguns parâmetros, sendo eles: Número de atributos; Probabilidade de mutação; Tamanho da população; Número de gerações.

Para este trabalho, consideramos modelos classificando usando no máximo 5, 10 e 15 atributos. A taxa de mutação escolhida foi de 20%. O tamanho da população, para execução do algoritmo foi de 50, 100 e 200. Por fim, executamos 100, 250, 500 e 1000 gerações através do algoritmo genético.

O *fitness* considerado para o torneio através do algoritmo genético foi a acurácia do modelo. O torneio do algoritmo genético foi realizado como ”um contra um”aleatório. O ponto de corte para realização do *crossover* entre os modelos vencedores foi exatamente o meio, caso o vetor tenha tamanho impar, a primeira parte da divisão possuiu um atributo a mais.

Foi observado que os modelos poderiam ser gerados com atributos repetidos após a execução do *crossover*. Apesar disto, este trabalho não fazer qualquer tratamento neste sentido, pois estes modelos perdiam para outros modelos que não foram gerados com atributos repetidos e eram eliminados na próxima geração.

5. Resultados

A partir de testes feito neste trabalho, foi verificado a acurácia e F-measure dos modelos obtidos por busca genética.

5.1. Regressão Logística

O modelo de classificação por Regressão Logística, considerando apenas 5 atributos, apresentou melhora na acurácia do modelo em 2,65% com relação a classificação utilizando todo o conjunto de dados. Este aumento se estendeu quando considerado 10 atributos e 15 atributos, sendo, respectivamente, 5,29% e 7,49% de aumento. A Tabela 1 relaciona o melhor resultado sobre todas as execuções considerando 5, 10 e 15 atributos e, por fim, é apresentado o resultado da classificação a partir de todo o conjunto de dados.

Tabela 1. Execução do algoritmo para regressão logística

Regressão Logística		
k	Acurácia	F-measure
5	88.11%	87,56%
10	90.75%	90,32%
15	92.95%	92,86%
41	85,46%	85,20%

5.2. Árvore de Decisão

O modelo de classificação por Árvore de Decisão, considerando apenas 5 atributos apresentou, melhora na acurácia do modelo em 3,53% com relação a classificação utilizando todo o conjunto de dados. Este aumento se estendeu quando considerado 10 atributos e 15 atributos, sendo, respectivamente, 9,7% e 8,81% de aumento. A Tabela 2 relaciona o melhor resultado sobre todas as execuções considerando 5, 10 e 15 atributos e, por fim, é apresentado o resultado da classificação a partir de todo o conjunto de dados.

Tabela 2. Execução do algoritmo para árvore de decisão

Árvore de decisão		
k	Acurácia	F-measure
5	81,50%	80,56%
10	87,67%	87,16%
15	86,78%	86,36%
41	77,97%	78,26%

5.3. Árvore Aleatória

O modelo de classificação por Árvore de Aleatória, considerando apenas 5 atributos apresentou, melhora na acurácia do modelo em 2,2% com relação a classificação utilizando todo o conjunto de dados. Este aumento se estendeu quando considerado 10 atributos e 15 atributos, sendo, respectivamente, 5,73% e 7,49% de aumento. A Tabela 3 relaciona o melhor resultado sobre todas as execuções considerando 5, 10 e 15 atributos e, por fim, é apresentado o resultado da classificação a partir de todo o conjunto de dados.

Tabela 3. Execução do algoritmo para árvore aleatória

Árvore aleatória		
k	Acurácia	F-measure
5	86,78%	85,51%
10	90,31%	89,81%
15	92,07%	91,82%
41	84,58%	83,72%

5.4. Frequência das Moléculas

A partir dos atributos obtidos pelas execuções acima, foi feita uma análise de frequência como apresentado pela Figura 1. Portanto, foi possível verificar a frequência de todos os 41 atributos escolhidos e quais são as mais relevantes. Logo, nota-se que as 5 moléculas mais relevantes são: [1, 3, 34, 40, 41]. Estes atributos são apresentados na Tabela 4.

6. Conclusão

A partir dos resultados obtidos através dos experimentos neste trabalho, pode-se concluir que todos os algoritmos tiveram uma melhora no desempenho a medida que consideramos menos atributos para o treinamento do modelos de classificação. Esta melhora foi mais significativa ao considerar até 15 atributos, sendo a única exceção o classificador por

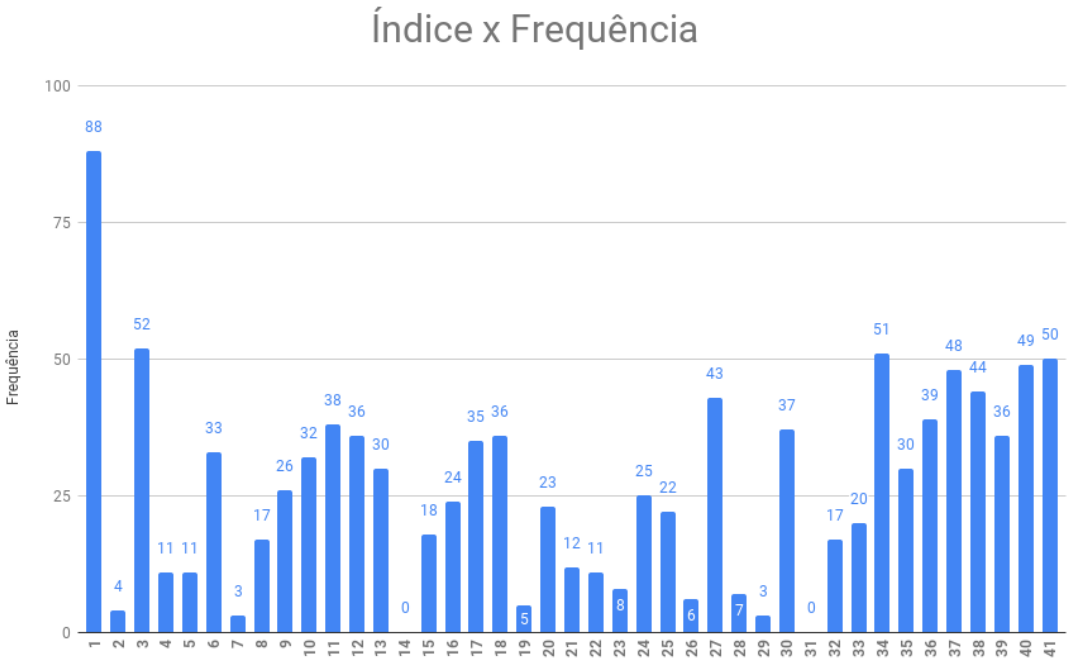

Figura 1. Índice x Frequência

Tabela 4. Tabela de atributos mais relevantes

Atributo	Frequência	Nome	Descrição
1	88	SpMax_L	Autovalor principal da matriz de Laplace
3	52	nHM	Número de átomos pesados
34	51	F02 [CN]	Frequência de C - N na distância topológica 2
40	49	nArCOOR	Número de ésteres (aromáticos)
41	50	nX	Número de átomos de halogênio

Árvore de decisão. Para os 5 melhores atributos das moléculas, obtivemos uma acurácia de até 88,11% e a medida que foi acrescentado mais atributos, a acurácia aumentou até 92,95% entre os modelos.

Foi observado que, para o algoritmo de regressão logística e árvore aleatória, ao aumentar o número de atributos, houve melhora na classificação. Apesar disto, ao considerar todos os 41 atributos, a acurácia do modelo caí significativamente. Portanto, entre o intervalo de 15 atributos e 41 atributos, poderia existir um número de atributos que permitiria uma melhor classificação do modelo. Apesar dos esforços, não foi possível seguir a diante com esta teoria.

Um outro resultado foi com relação ao tempo de execução de cada algorítimos. A execução da busca genética dos algorítimos de regressão logísticas e arvore de decisão obteve um tempo médio de 10 a 20 minutos para concluir mil gerações, enquanto o algorítimo de árvore aleatória levou cerca de 6 a 8 horas mesmo sem apresentar melhora significativa.

Referências

Casali, R. (2020). Árvore de decisão: como fazer e importância dentro da ciência de dados. Disponível em: www.digitalhouse.com/br/blog/arvore-de-decisao , Acesso em: 20 de agosto 2021.

de Azevedo Gonzales, L. (2018). Regressão logística e suas aplicações. *UFMA*.

De Jong, K. (1975). An analysis of the behavior of a class of genetic adaptive systems. *University of Michigan, USA*.

Holland, J. H. (1975). Adaptation in natural and artifcial systems. *The University of Michigan Press, Ann Arbour*.

Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., and Consonni, V. (2013). Qsar biodegradation data set.

Mbaabu, O. (2020). Introduction to random forest in machine learning. Disponível em: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/, Acesso em: 20 de agosto 2021.