

Relatório 2 - Inteligência Computacional

Jhúlia Graziella de Souza Rodrigues

¹Bacharelado em Ciência da Computação
Universidade Federal do Uberlândia (UFU)
Uberlândia (MG)

jhuliagraziella@gmail.com

Resumo. *Este trabalho é um relatório referente à implementação de um Algoritmo Genético para Data Mining sobre uma base de dados dermatológicos a fim de reunir sintomas comuns de diferentes classes de doenças.*

Palavras-Chave. *Algoritmo Genético, Data Mining*

1. Introdução

De acordo com [Pozo et al.], Computação Evolutiva (CE) é um ramo de pesquisa originado da Inteligência Artificial que propõe um paradigma para solução de problemas inspirado na Seleção Natural (Darwin, 1859).

Algoritmo Genético (AG) é uma técnica da Computação Evolutiva para encontrar soluções aproximadas em problemas de otimização e busca, geralmente quando não se sabe muito bem por onde começar a busca mas é possível definir uma função heurística para avaliar resultados.

De acordo com [Wikipédia 2017], Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados. Isso permite que sejam realizadas estimativas através dos dados obtidos.

O objetivo deste trabalho é implementar um AG semelhante ao utilizado nos artigos de [Rulliam et al. 2000] e [Andrade and Grossa] e fazer testes sobre este algoritmo e sobre algumas variações dele.

2. Método

A representação do indivíduo e os operadores genéticos utilizados nos trabalhos citados foram mantidos. O cromossomo é representado por um conjunto de 34 genes correspondentes aos atributos na base de dados: incontinência de melanina, eritema, idade etc. Cada gene do cromossomo é constituído de três elementos: peso, operador e valor.

- Peso: valor real entre 0 e 1, indica se esse atributo será utilizado na regra ou não;
- Operador: operador lógico =, \neq , \geq ou $<$;
- Valor: valor numérico do campo.

A população é um conjunto de 50 indivíduos, conforme definido nos trabalhos-base. O algoritmo foi executado 10 vezes para cada uma das classes, iniciando com uma população gerada aleatoriamente e prosseguindo de acordo com os passos da Figura 1.

A seleção foi feita utilizando-se um torneio estocástico de tamanho 3: 3 indivíduos são selecionados pelo método da roleta e, depois, competem entre si pelo método do torneio para serem selecionados.

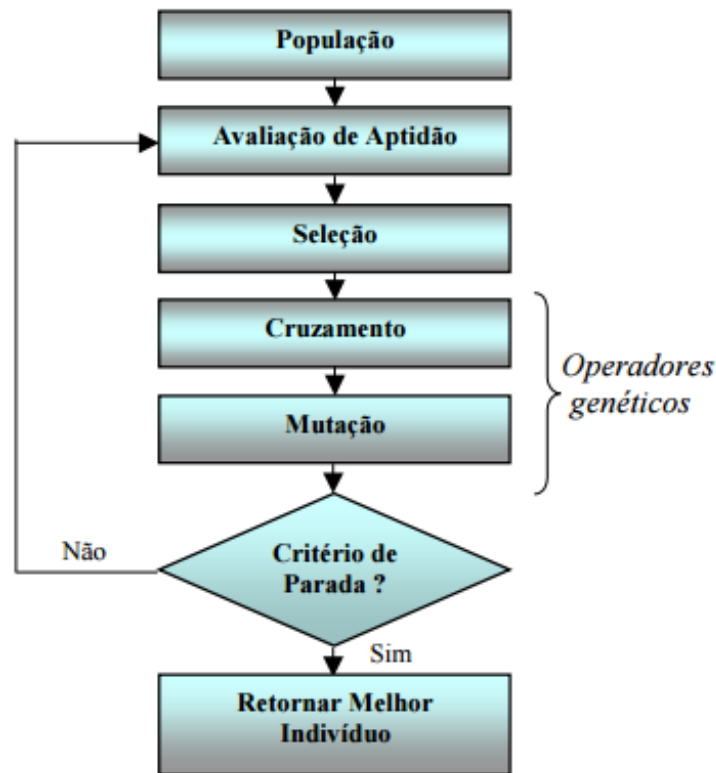


Figura 1. Etapas comuns de um algoritmo genético. Figura da [Pozo et al.]

O crossover foi baseado no crossover de dois pontos: dois pontos no intervalo [0, 33] são selecionados e todos os genes entre esses pontos são trocados. A taxa de crossover é de 100%.

A mutação é realizada sobre todos os campos de todos os genes, com probabilidade de 30% de ocorrer em cada gene.

O método de reinserção escolhido foi como o de [Andrade and Grossa]: a população foi substituída pelos melhores entre os pais e filhos.

A condição de parada do algoritmo é a quantidade máxima de gerações, predefinida como 50.

Para calcular a aptidão de um indivíduo foi utilizada a função $A = Sp * Se$, onde $Sp = \frac{Tn}{Tn + Fp}$ se refere à sensibilidade e $Se = \frac{Tp}{Tp + Fn}$ se refere à especificidade do indivíduo, calculados de acordo com a quantidade de falsos positivos (Fp), falsos negativos (Fn), verdadeiros positivos (Tp) e verdadeiros negativos (Tn) encontrados na base de dados.

3. Resultados

Os melhores resultados obtidos após as 10 execuções do AG sobre cada uma das classes estão presentes na tabela 3.

As regras obtidas para cada uma das classes foram as seguintes:

- Classe 1: clubbing of the rete ridges $\neq 0$;
- Classe 2: oral mucosal involvement $\neq 3$ AND spongiosis ≥ 2 ;
- Classe 3: band-like infiltrate $\neq 0$;
- Classe 4: melanin incontinence < 2 AND spongiosis $\neq 0$;

Tabela 1. Tabela com os resultados obtidos

Classe	Fitness Treinamento	Fitness Teste
1	0.948434	0.948434
2	0.699390	0.695238
3	0.959184	1.000000
4	0.708625	0.711538
5	0.586538	0.682692
6	0.916300	1.000000

- Classe 5: scalp involvement < 1 AND exocytosis < 3 AND disappearance of the granular layer = 0;
- Classe 6: follicular horn plug ≥ 1 .

Podemos perceber que os resultados obtidos não são exatamente iguais aos dos trabalhos utilizados como referência. As diferenças na etapa de treinamento estão presentes na tabela 3.

Tabela 2. Comparação entre os resultados na etapa de treinamento

Classe	[Rulliam et al. 2000]	[Andrade and Grossa]
1	+0.024566	+0.000434
2	+0.155610	+0.058610
3	+0.019816	+0.029816
4	+0.074375	+0.108375
5	+0.413462	+0.413462
6	+0.083700	+0.078700

As diferenças na etapa de teste estão presentes na tabela 3.

Tabela 3. Comparação entre os resultados na etapa de teste

Classe	[Rulliam et al. 2000]	[Andrade and Grossa]
1	+0.024566	+0.051566
2	+0.159762	+0.151762
3	+0.000000	+0.000000
4	+0.148462	+0.076462
5	+0.317308	+0.317308
6	+0.000000	+0.000000

4. Variações

Nesta sessão serão apresentados os resultados obtidos ao aplicar o AG modificado 10 vezes para cada classe.

A primeira mudança será no método de seleção: vamos verificar se o método de seleção da roleta é melhor ou pior para este problema. A outra mudança será na taxa de mutação: vamos verificar se aumentar a taxa de mutação causa melhoras ou pioras nos resultados.

4.1. Seleção pela Roleta

Ao trocar o método de seleção para o método da roleta, obtemos os resultados apresentados na tabela em 4.1. Podemos observar que os resultados obtidos foram melhores que

os resultados obtidos previamente, utilizando o método de seleção por torneio estocástico de tamanho 3.

Tabela 4. Tabela com os resultados obtidos com a seleção por roleta

Classe	Fitness Treinamento	Fitness Teste
1	0.662651	0.763158
2	0.847317	0.876190
3	0.974490	1.000000
4	0.778555	0.750000
5	0.990385	1.000000
6	0.947137	0.991150

Utilizando esta variação, as regras obtidas pelo AG são as seguintes:

- Classe 1: munro microabcess $\neq 0$ AND spongiosis $\neq 2$;
- Classe 2: koebner phenomenon < 1 AND focal hypergranulosis $= 0$ AND spongiosis ≥ 2 ;
- Classe 3: vacuolisation and damage of basal layer $\neq 0$;
- Classe 4: oral mucosal involvement < 2 AND elongation of the rete ridges $= 0$ AND spongiform pustule $= 0$ AND saw-tooth appearance of retes < 1 AND perifollicular parakeratosis $\neq 1$;
- Classe 5: fibrosis of the papillary dermis > 1 ;
- Classe 6: follicular papules ≥ 1 .

4.2. Taxa de mutação de 50%

Após aumentar a taxa de mutação de 30% para 50%, foram obtidos os resultados apresentados na tabela em 4.2. Podemos observar que os resultados obtidos foram piores que os resultados obtidos previamente, utilizando taxa de mutação de 30%.

Tabela 5. Tabela com os resultados obtidos com taxa de mutação de 50%

Classe	Fitness Treinamento	Fitness Teste
1	0.948434	1.000000
2	0.664878	0.746667
3	0.959184	1.000000
4	0.641171	0.696833
5	0.604168	0.751131
6	0.920390	1.000000

Utilizando esta variação, as regras obtidas pelo AG são as seguintes:

- Classe 1: clubbing of the rete ridges $\neq 0$;
- Classe 2: PNL infiltrate $\neq 0$ AND exocytosis $\neq 0$ AND clubbing of the rete ridges $\neq 1$;
- Classe 3: melanin incontinence $\neq 0$;
- Classe 4: koebner phenomenon $\neq 0$ AND exocytosis > 1 ;
- Classe 5: polygonal papules ≥ 0 AND clubbing of the rete ridges $\neq 3$ AND elongation of the rete ridges $\neq 0$;
- Classe 6: fibrosis of the papillary dermis $\neq 1$ AND follicular horn plug $\neq 0$.

5. Conclusão

Podemos concluir que o poder computacional dos AG's é muito grande, podemos aplicá-los numa grande variedade de problemas, mesmo os difíceis de se modelar como este. Também nota-se que cada problema possui diferentes parâmetros ótimos: o que era melhor para o problema de criptoaritmética pode não ser o melhor para este problema, por exemplo.

Referências

- Andrade, P. S. and Grossa, P. Discovering Comprehensible Classification Rules with a Genetic Algorithm.
- Pozo, A., Cavalheiro, A., Ishida, C., Spinosa, E., and Rodrigues, E. Computação evolutiva. pages 3–7.
- Rulliam, C., Maira, G., Oliveira, B. D., Bispo, J., Consolação, R., São, C. E. P., and Sp, P. (2000). ALGORITMOS GENÉTICOS APLICADOS EM DATA MINING PARA OBTENÇÃO DE REGRAS SIMPLES E PRECISAS. pages 1–6.
- Wikipédia (2017). Mineração de dados — wikipédia, a enciclopédia livre. [Online; accessed 20-06-2017].