UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA

ALEXANDRE ROSSETO LEMOS

TRABALHO COMPUTACIONAL 2

COMPUTAÇÃO NATURAL

Vitória
Março 2022

Objetivo

A solução da tarefa compreende na implementação dos algoritmos PSO e GA para treinar uma rede neural *feedforward* para classificação de dados para 3 benchmarks: Iris, Wine, e Breast Cancer.

Os algoritmos PSO e GA são utilizadas para otimização de funções. No caso deste projeto, eles foram utilizados para maximizar a função de acurácia dos modelos, dada pela equação abaixo.

$$Acurácia = \frac{V_P + V_N}{V_P + V_N + F_P + F_N}$$

Onde V_P, V_N, F_P e F_N são os valores verdadeiro positivos, verdadeiro negativos, falso positivos e falso negativos calculados pelo modelo.

Implementação

Utilização de algoritmos/bibliotecas já existentes

Para o desenvolvimento deste projeto, foi utilizada a rede neural da biblioteca *sickit-learn* (MLPClassifier), que foi utilizada para gerar as partículas/populações iniciais. Não houveram modificações realizadas na biblioteca propriamente, apenas os pesos dos modelos foram atualizados conforme os algoritmos PSO e GA eram executados e geravam novos pesos.

Os algoritmos PSO e GA foram desenvolvidos completamente para realizar a otimização dos parâmetros.

Algoritmos utilizados

Os parâmetros utilizados pelas redes neurais foram *default*, exceto pela quantidade de camadas ocultas e quantidade de neurônios em cada camada oculta (*hidden_layer_sizes*), o parâmetro de quantidade de iterações (*max_iter*) e o parâmetro que permite a rede executar um novo treinamento levando em consideração os pesos atuais dela (*warm_start),* parâmetro esse que foi utilizado para substituir os pesos da rede pelos pesos calculados pelos algoritmos PSO e GA.

Para o algoritmo PSO, foram utilizados como parâmetros w_{min} e w_{max} de 0,2 e 06 respectivamente, 50 partículas e 100 iterações. Utilizou-se também os parâmetros c_1 e c_2 com valores de 2,0 e os parâmetros r_1 e r_2 iguais a 0,729. O vetor de velocidades foi inicializado como sendo um vetor de zeros.

Para o algoritmo GA foram utilizados como parâmetros taxa de crossover de 0,9, 250 indivíduos na população, limites de mutação entre [-5, 5], taxa de mutação de 10% e 20 gerações. O torneio foi utilizado como forma de seleção dos pais, comparando dois pais por vez ($k = 2$).

Testou-se diferentes combinações de quantidade de neurônios e quantidade de neurônios em cada camada oculta, para analisar se existia alguma conclusão a ser extraída relacionando a *fitness* obtida com a estrutura da rede neural.

Dados utilizados

Os dados necessários para o desenvolvimento do projeto (Iris, Wine e Breast Cancer presentes no repositório *UCI Machine Learning Repository* da Universidade da Califórnia Irvine) foram obtidos através da biblioteca *scikit-learn*. Cada conjunto de dados possui características diferentes que serão explicitadas nesta sessão.

O conjunto de dados Iris possui informações a respeito de diferentes plantas do gênero Iris, com 150 amostras, quatro colunas de características das pétalas e uma coluna de rótulo que informa qual subgrupo a amostra pertence (existem três possíveis categorias). O conjunto de dados possui 50 amostras em cada uma das três classes, sendo totalmente balanceado.

O conjunto de dados Breast Cancer contém informações a respeito de imagens dos núcleos celulares extraídos de pacientes, como raio, textura, perímetro e área, por exemplo. Também possui campo de identificação do paciente e uma coluna de rótulo que informa se o tumor é benigno ou maligno (duas

classes). O conjunto de dados possui 357 amostras pertencentes à classe de câncer benigno e 212 amostras pertencentes à classe de câncer maligno.

O conjunto de dados Wine contém informações sobre análises químicas realizadas em 178 vinhos da mesma região, porém de três localidades diferentes (três classes). Ele possui características como concentração de álcool, magnésio e intensidade da cor do vinho, por exemplo. O conjunto de dados possui 59 amostras pertencentes à classe 1, 71 pertencentes à classe 2 e 48 pertencentes à classe 3.

Resultados

Devido à natureza aleatória dos algoritmos utilizados, cada algoritmo, inclusive o *baseline*, foi executado 15 vezes. Os resultados médios de cada modelo estão explicitados na tabela a seguir, para cada estrutura da rede neural utilizada.

Com uma camada oculta com 10 neurônios.

Modelo	Conjunto de dados	Fitness	Tempo de Execução [s]	Média de iterações até convergir	Diferença com o Baseline (Fitness)	Diferença do Baseline (Tempo em segundos)
MLP com PSO	Wine	0,907407	20,471066	6,333333	0,112345	19,99907
	Breast Cancer	0,956725	23,565971	6,533333	0,030409	22,65087
	Iris	0,939259	19,816867	4,466667	0,302222	19,38587
MLP com GA	Wine	0,791358	4,266797	1,200000	-0,003704	0,728
	Breast Cancer	0,767901	4,358063	1,066667	-0,158415	0,15157
	Iris	0,751852	4,303000	1,266667	0,114815	0,835667
Baseline	Wine	0,795062	0,472000	Não convergiu	0	0
	Breast Cancer	0,926316	0,915097	Não convergiu	0	0
	Iris	0,637037	0,431000	Não convergiu	0	0

Com uma camada oculta com 20 neurônios.

Modelo	Conjunto de dados	Fitness	Tempo de Execução [s]	Média de iterações até convergir	Diferença com o Baseline (Fitness)	Diferença do Baseline (Tempo em segundos)
MLP com PSO	Wine	0,950617	21,251320	6,600000	0,032098	20,74582
	Breast Cancer	0,956335	25,389879	7,133333	0,006238	24,35288
	Iris	0,906667	20,538162	5,000000	0,114074	20,06416
MLP com GA	Wine	0,762963	4,839955	1,133333	-0,15556	0,627831
	Breast Cancer	0,756790	4,734652	1,000000	-0,19331	-0,037
	Iris	0,776543	4,732744	1,133333	-0,01605	0,659333
Baseline	Wine	0,918519	0,505502	Não convergiu	0	0
	Breast Cancer	0,950097	1,037001	Não convergiu	0	0
	Iris	0,792593	0,474000	Não convergiu	0	0

Com uma camada oculta com 50 neurônios.

Modelo	Conjunto de dados	Fitness	Tempo de Execução [s]	Média de iterações até convergir	Diferença com o Baseline (Fitness)	Diferença do Baseline (Tempo)
MLP com PSO	Wine	0,941975	24,641784	5,133333	-0,03457	23,97965
	Breast Cancer	0,941975	24,641784	5,133333	-0,01553	21,40362
	Iris	0,909630	22,917643	4,400000	0,06963	22,32619
MLP com GA	Wine	0,795062	6,444937	1,133333	-0,18148	0,471194
	Breast Cancer	0,812346	6,436041	1,000000	-0,14516	-2,23816
	Iris	0,765432	6,411600	1,066667	-0,07457	0,475214
Baseline	Wine	0,976543	0,662139	Não convergiu	0	0
	Breast Cancer	0,957505	3,238161	Não convergiu	0	0
	Iris	0,840000	0,591453	Não convergiu	0	0

Com duas camadas ocultas com 10 neurônios em cada.

Modelo	Conjunto de dados	Fitness	Tempo de Execução [s]	Média de iterações até convergir	Diferença com o Baseline (Fitness)	Diferença do Baseline (Tempo)
MLP com PSO	Wine	0,908642	23,442171	6,666667	0,053086	22,83017
	Breast Cancer	0,953216	27,952029	6,866667	0,017154	26,68941
	Iris	0,920000	23,446600	5,333333	0,201481	22,8306
MLP com GA	Wine	0,717284	5,266202	1,200000	-0,13827	0,588001
	Breast Cancer	0,692593	5,281016	1,133333	-0,24347	-0,12929
	Iris	0,700000	5,232893	1,266667	-0,01852	0,650667
Baseline	Wine	0,855556	0,611999	Não convergiu	0	0
	Breast Cancer	0,936062	1,262618	Não convergiu	0	0
	Iris	0,718519	0,616000	Não convergiu	0	0

Com três camadas ocultas com 10 neurônios em cada.

Modelo	Conjunto de dados	Fitness	Tempo de Execução [s]	Média de iterações até convergir	Diferença com o Baseline (Fitness)	Diferença do Baseline (Tempo)
MLP com PSO	Wine	0,907407	26,605162	7,066667	0,082716	25,82516
	Breast Cancer	0,939571	32,802013	7,466667	-0,00234	31,20561
	Iris	0,891852	26,744142	4,333333	0,134815	26,00247
MLP com GA	Wine	0,677778	6,203926	1,066667	-0,14691	0,286668
	Breast Cancer	0,681481	6,154042	1,000000	-0,26043	-0,5964
	Iris	0,679012	6,023373	1,000000	-0,07803	0,258326
Baseline	Wine	0,824691	0,779999	Não convergiu	0	0
	Breast Cancer	0,941910	1,596404	Não convergiu	0	0
	Iris	0,757037	0,741674	Não convergiu	0	0

Conclusão

Pode-se concluir que, em geral, o modelo híbrido de PSO com MLP obteve uma performance melhor que o modelo *baseline* e o modelo GA com MLP desenvolvido quando se analisa apenas a *fitness*. Essa diferença tende a diminuir conforme a complexidade da rede neural aumenta. É possível perceber que os modelos com PSO e GA obtiveram uma convergência bem mais rápida que o modelo *baseline***.**

Vale ressaltar que o modelo baseline provavelmente conta com uma codificação mais performática do que a desenvolvida para os algoritmos PSO e GA neste trabalho, uma vez que o modelo do scikit-learn conta com a colaboração de milhares de desenvolvedores, então é possível que o tempo de execução para os modelos híbridos desenvolvidos possa ser otimizado.

Para o modelo MLP com PSO, conclui-se que ele obteve uma performance muito boa, obtendo uma performance melhor que o modelo *baseline* em praticamente todas as experiências. Para o modelo MLP com GA, conclui-se que o modelo não obteve uma performance tão boa quanto à do outro modelo híbrido, visto que em média sua *fitness* ficou em torno de 70%, o que é um valor abaixo do alcançado pelo modelo de referência. Observa-se que os modelos com GA convergiram muito rapidamente, sendo possível que tenham ficado presos em máximos locais, não sendo possível assim a obtenção de resultados mais satisfatórios.