Seleção de Features Para Predição da Síndrome do Ovário Policístico utilizando o algoritmo Particle Swarm Optimization

Keterly Geovana Gouveia Silva
Instituto de Geociências e Ciências Exatas
Universidade Estadual Paulista (UNESP)
Rio Claro - SP, Brasil
keterly.silva@unesp.br

Fabricio Aparecido Breve
Instituto de Geociências e Ciências Exatas
Universidade Estadual Paulista (UNESP)
Rio Claro - SP, Brasil
fabricio.breve@unesp.br

***Resumo*—Síndrome do Ovário Policístico (SOP) é uma endocrinopatia que atinge mulheres em idade fértil, sendo o diagnóstico precoce considerado de extrema importância. É crescente o número de propostas que utilizam técnicas de aprendizado de máquina para a predição de eventos de saúde e, se torna necessário, escolher os melhores atributos que determinam um padrão. Este trabalho propõe o estudo do uso do algoritmo Particle Swarm Optimization (PSO) para a seleção de atributos em um conjunto de dados sobre a SOP, utilizando Random Forest, Logistic Regression e Multilayer Perceptron.**

Área: Inteligência Computacional

I. Introdução

A Síndrome do Ovário Policístico (SOP) é considerada uma das endocrinopatias mais comuns em mulheres em idade reprodutiva. A SOP é uma das causadoras do aumento do nível de hormônios masculinos no corpo feminino, também podendo ser caracterizada pelos eventos de irregularidade menstrual, acne, alopecia, aumento do tamanho dos ovários e surgimento de cistos ovarianos [1].

Atualmente, na busca pelo diagnóstico precoce de doenças físicas e mentais, estudos propõe o uso de Aprendizado de Máquina para a detecção. No trabalho de [2], é proposto o diagnóstico da SOP com a utilização de diferentes tipos de classificadores como o modelo hibrido de floresta aleatória e regressão logística, chegando a 91,01% de precisão. Neste cenário, é comum a contribuição de profissionais da área para a escolha de características que melhor indicam a ocorrência de uma patologia. Em outros casos, os atributos são escolhidos através da análise da correlação entre variáveis.

Diante disso, este trabalho estuda a utilização de um algoritmo de inteligência de enxames, o Particle Swarm Optimization (PSO), para a criação de um método de seleção de atributos. O estudo propõe aplicar o algoritmo para avaliar atributos de um conjunto de dados relativo à Síndrome do Ovário Policístico e determinar, com base nos resultados de acurácia, as melhores variáveis para modelos de predição específicos. Além disso, objetiva-se testar diferentes classificadores previamente configurados para avaliação da seleção, bem como uma comparação entre os resultados obtidos com o uso de métodos existentes em bibliotecas comumente utilizadas.

XII Workshop do Programa de Pós-Graduação em Ciência da Computação da Unesp (WPPGCC 2023), Unesp, 21 e 22 de novembro de 2023.

II. Conceitos e Técnicas

O estudo de seleção de atributos para detecção de Sindrome do Ovário Policístico, utilizou o Particle Swarm Optimization como exemplo de algoritmo inspirado pela natureza. Da categoria de algoritmos de inteligência de enxames, o Particle Swarm Optimization (PSO) foi proposto em 1995 por James Kennedy e Russell Elberhart para otimização de problemas de domínio contínuo. O PSO surgiu da observação do comportamento social de espécies de pássaros e cardumes de peixes [3].

O algoritmo Particle Swarm Optimization é aplicado para a resolução de diferentes problemas, sendo eles simples, complexos e até mesmo não lineares. Como exemplos de aplicações, o PSO é útil para otimização de pesos em treinamento de redes neurais, otimização de processos, seleção de atributos para problemas de reconhecimento de padrão e para exemplos simples e didáticos como minimização de funções.

Neste estudo, técnicas de Aprendizado de Máquina foram aplicadas a medida em que os classificadores foram testados. De modo geral, o Aprendizado de Máquina utiliza teorias da estatística para a criação de modelos preditivos e descritivos, permitindo que programas de computadores aprendam por meio da experiência e executem tarefas que não são possíveis através da programação clássica [4].

III. Metodologia de Desenvolvimento

A. Seleção e manipulação conjunto de dados

Um conjunto de dados disponibilizado publicamente no repositório Kaggle [5], intitulado PCOS Dataset, foi utilizado para este experimento, contendo inicialmente 43 atributos e 541 padrões de entrada. Após exploração de dados, observou-se a necessidade de remoção de linhas com caracteres alfabéticos, conversão de colunas não numéricas em numéricas, bem como a ausência de valores nulos.

Neste processo colunas que apresentaram o número de série e o número de arquivo do paciente foram removidas, uma vez que tais dados eram irrelevantes. Após a limpeza e conversões necessárias, as amostras também foram normalizadas pelo método min-max. O conjunto final somou 538 exemplos, sendo 362 exemplos de mulheres sem SOP e 176 exemplos de mulheres com a síndrome. Os dados também foram divididos em 70% para treino e 30% para teste.

B. Construção do selecionador de atributos

A classe PSOFeatureSelection teve como exemplo de responsabilidade, o cálculo de aptidão e atualização de velocidade, aspetos básicos da implementação do algoritmo de otimização por enxame de partículas. A seleção se baseou não apenas no conjunto de dados, mas também no modelo e configuração a ser utilizada na predição. Por este motivo, a função de aptidão foi calculada executando o modelo escolhido e avaliando a acurácia para uma amostra de 10 atributos aleatórios a cada execução.

O retorno dos melhores atributos para a predição se deu após ordenando-os em ordem crescente de valores, sendo os últimos 10 atributos considerados os mais aptos.

Nesta etapa também foram definidos os valores para cada um dos parâmetros necessários na execução do PSO: max_it = 100 (número máximo de iterações), Ac1 = 0,01 (constante de aceleração), Ac2 = 1 (constante de aceleração), v_max = 2 (valor mínimo para velocidade), v_min = -2 (valor máximo para velocidade), n = 80 (número de particulas), num_features = 41 (número de atributos) e n_neighbors = 3 (número de vizinhos).

C. Definição dos classificadores

RandomForestClassifier, LogisticRegression e MLPClassifier foram escolhidos para este trabalho. Para fins de comparação, 10 atributos também foram selecionados através do método SelectKBest, da biblioteca scikit-learn.

IV. Resultados

Esta seção apresenta os resultados experimentais da utilização do algoritmo PSO para seleção de atributos. Os algoritmos foram configurados e testados utilizando atributos selecionados pelo PSO. Considerando que o algoritmo de enxame de particulas é um algoritmo estocástico, ou seja, que varia o resultado em cada execução, cada modelo de classificação executou a classe PSOFeatureSelection 10 vezes.

Em cada uma das execuções, além de diferentes valores para as métricas calculadas, o selecionador retornou atributos diferentes. A Tabela I apresenta recall, precisão e acurácia média após 10 execuções de cada um dos diferentes classificadores utilizados neste trabalho.

Posteriormente, os mesmos mantiveram suas configurações e foram treinados com atributos selecionados pelo método SelectKBest. O método, que utiliza teste estatístico para a obtenção das melhores variáveis, não apresenta o comportamento de atributos diferentes a cada vez que é chamado. Sendo assim, apenas uma execução foi realizada.

Tabela I
SELEÇÃO DE ATRIBUTOS COM PSO - MÉTRICAS

Classificador	Acurácia	Precisão	Recall
RandomForestClassifier	0,86790	0,81767	0,75961
LogisticRegression	0,88518	0,81632	0,76923
MLPClassifier	0,88148	0,83217	0,79038

A Tabela II apresenta recall, precisão e acurácia média após a classificação.

Tabela II
SELEÇÃO DE ATRIBUTOS COM SELECTKBEST - MÉTRICAS

Classificador	Acurácia	Precisão	Recall
RandomForestClassifier	0,89506	0,85714	0,80769
LogisticRegression	0,88888	0,88636	0,75
MLPClassifier	0,90740	0,87755	0,82692

SelectKBest obteve melhores valores em quase todas as métricas quando comparado aos resultados da seleção utilizando PSO. No entanto, a métrica recall apresentou valor ligeiramente maior no modelo de regressão logística testados com os atributos retornados pelo algoritmo de enxame de particulas.

De modo geral, quando a diferente entre os valores são calculados, retornam diferença máxima de aproximadamente 0,027 para acurácia, 0,070 para precisão e 0,048 para recall. Os valores foram obtidos comparando RandomForestClassification para acurácia, LogisticRegression para precisão e RandomForestClassification para recall.

V. Considerações Finais

Este trabalho teve como objetivo realizar um estudo sobre o desempenho do algoritmo PSO para a seleção de atributos em um conjunto de dados sobre a Síndrome do Ovário Policístico, utilizando acurácia para o cálculo da aptidão das soluções propostas.

Os experimentos demonstraram que o elemento de aleatoriedade do PSO gerou maior impacto na lista de atributos selecionados, quando comparado ao uso específico de modelos de predição. Mesmo com resultados inferiores a utilização de métodos empregados em bibliotecas, também demonstrou que é possível realizar a seleção com desempenho considerável.

Para trabalhos futuros, deseja-se adaptar a seleção de atributos com PSO para utilizar, além da acurácia, outras métricas importantes na medição de desempenho do modelo, além de etapas de validação cruzada.

Referências

[1] H. H. G. de Moura, D. L. M. Costa, E. BagatinIII, and C. T. S. and; Mônica Manela-Azulay, *Síndrome do ovário policístico: abordagem dermatológica*. Sociedade Brasileira de Dermatologia, 2011, vol. 86.

[2] S. Bharati, P. Podder, and M. R. H. M. , *Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms*, 2020.

[3] A. B. de Souza Serapião, *Fundamentos de otimização por inteligência de enxames: uma visão geral*. Sociedade Brasileira de Automática, 2009.

[4] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[5] V. Thakre, *PCOS Dataset*. https://www.kaggle.com/datasets/shreyasvedpathak/pcos-dataset/, 2020.