Evento	Salão UFRGS 2020: FEIRA DE INOVAÇÃO TECNOLÓGICA DA UFRGS - FINOVA
Ano	2020
Local	Virtual
Título	BioSelector: um framework para descoberta de genes candidatos a biomarcadores de doenças
Autor	FELIPE COLOMBELLI
Orientador	MARIANA RECAMONDE MENDOZA GUERREIRO

RESUMO

TÍTULO DO PROJETO: BioSelector: um framework para descoberta de genes candidatos a biomarcadores de doenças

Aluno: Felipe Colombelli
Orientadora: Mariana Recamonde Mendoza

RESUMO DAS ATIVIDADES DESENVOLVIDAS PELO BOLSISTA

Muitas técnicas de Aprendizado de Máquina vêm sendo exploradas para realizar seleção de atributos em dados com alta dimensionalidade, especialmente em dados biológicos como os de expressão gênica. Particularmente, investigamos a seleção de atributos pelo paradigma de aprendizado por ensemble. Ensembles se apoiam na ideia de que a agregação de diferentes opiniões fornece mais conhecimento do que uma opinião especialista isolada. Há duas principais formas de se buscar diversidade de opinião para montar um Ensemble: através da perturbação de funções e através da perturbação de dados. Este projeto de pesquisa visou propor uma nova abordagem unificando as duas formas de busca por diversidade em um único Ensemble (aqui chamado de híbrido), comparando-o com as demais estratégias da literatura. Como estudo de caso, investigamos a seleção de atributos em câncer de mama, visando identificar novos candidatos a biomarcadores.

A validação e comparação das abordagens foi baseada em duas métricas que fornecem indicativos de descoberta de informação: capacidade preditiva, medida pela área abaixo da curva ROC (ROC-AUC), e estabilidade, medida através do *Kuncheva Index*. Utilizamos ainda, para cada experimento, um processo de validação cruzada estratificada com subamostragem da classe predominante.

Além disso, normalizamos diferentes conjuntos de dados de diferentes plataformas através da *Nonparanormal Distribution*, visando uma validação ainda maior pela separação dos conjuntos utilizados em treino e teste. As altas ROC-AUCs obtidas indicam que nossa abordagem consegue aprender e generalizar sobre o problema, não se limitando apenas a um aprendizado orientado ao conjunto de dados provenientes da mesma fonte.

Toda a montagem dos experimentos foi automatizada a nível de código e este foi utilizado para a construção do BioSelector, um framework que oferece uma interface gráfica amigável para a construção flexível de experimentos análogos aos nossos, com cinco opções de seletores e a possibilidade de adicionar novos algoritmos personalizados pelos usuários.