Evento	Salão UFRGS 2018: SIC - XXX SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2018
Local	Campus do Vale - UFRGS
Título	Desenvolvimento de um Banco de Dados Curado para Pesquisas em Aprendizado de Máquina
Autor	EDUARDO BASSANI CHANDELIER
Orientador	MARCIO DORN

Universidade Federal do Rio Grande do Sul

Desenvolvimento de um Banco de Dados Curado para Pesquisas em Aprendizado de Máquina

Eduardo Bassani Chandelier e Márcio Dorn

O câncer é a maior causa de mortes naturais, levando a óbito mais de 8 milhões de pessoas anualmente. A geração de uma enorme quantidade de dados genômicos e transcriptômicos, revolucionou a medicina do século XXI, auxiliando no entendimento de doenças como o câncer. Por outro lado, essa grande quantidade de dados trouxe um novo problema: a dificuldade de extração de dados biológicos relevantes. Neste sentido, técnicas de bioinformática tem se tornado a principal ferramenta para processamento e análise de dados ômicos. Entretanto, mesmo as técnicas padrão da bioinformática necessitam de alto custo computacional. Neste contexto se encaixa o Aprendizado de Máquina *(AM)*, um conjunto de técnicas computacionais que possibilitam o computador a aprender utilizando dados como *input*, e com isso gerando resultados com baixo custo computacional. Um crescente esforço tem sido desenvolvido nos últimos tempos na aplicação do AM para a biologia do câncer. Todavia, a maioria desses trabalhos não disponibiliza os dados utilizados, não são padronizados, acurados ou são antigos - além de não informarem como foram tratados antes das análises. Desta forma, visando otimizar, padronizar e auxiliar os estudos de AM, aplicados ao estudo do câncer, é observada a necessidade de criação de um banco de dados ômico de câncer de qualidade. Sendo assim, o objetivo deste trabalho é criar um banco de dados de microarranjo que seja padronizado, curado e em formatos que possam ser usados por algoritmos de AM. Adicionalmente, técnicas de AM serão desenvolvidas e aplicadas para analisar estes dados.

Os dados foram coletados da plataforma NCBI-GEO (https://www.ncbi.nlm.nih.gov/gds/), utilizando um rigoroso filtro de busca e tratados com *background correction*, normalizados e manualmente curados por sondas irrelevantes. Foram gerados 104 arquivos brutos de 14 tipos de câncer diferentes, cada arquivo contendo de 6 a 288 amostras. Os dados foram empregados para gerar arquivos em formatos conhecidos e de fácil processamento por software de AM, como CSV, ARFF, GCT e CLS. Este banco estará disponível através do site (http://sbcb.inf.ufrgs.br/home) para a utilização pública.

Junto aos dados estarão disponíveis as descrições completas e *benchmarks* dos mesmos, para fins de validação ou comparação. Os *benchmarks* serão rodados em cada arquivo gerado, que são compostos por tabela onde a primeira dimensão é o paciente, ou amostra, e a segunda é a lista de aproximadamente 30-40 mil genes, onde as células da tabela são valores que representam a expressão gênica. Já entre as técnicas a serem usados para verificar a qualidade dos dados estão algoritmos de Seleção, como *Principal Components* (PCA), e de classificação, como *Support Vector Machine* (SVM), serão utilizadas validações utilizando técnicas como *cross-validation* para separar os dados em *n* subconjuntos e rodar o algoritmo *n* vezes, usando para cada uma delas um dos subconjunto para validação, ou comparar com métodos de *base line.* Futuramente será o desenvolvimento de técnicas que utilizem o banco de dados e a comparação com os resultados do *benchmarks* presentes no mesmo. Como existe uma série de *frameworks* para criar algoritmos de AM em GPU (*Graphics Processing Unit*), devido a sua alta capacidade de processamento paralelo, será investigado quais as ferramentas cada um disponibiliza para decidir qual é o mais adequados para o problema, para então decidir quais algoritmos podem ser usados.