Sobre o pré-processamento de dados de espectrometria de massa: instruções e ferramentas para tratamento e uso em modelos de aprendizado de máquina para classificação de biomarcadores

D. A. S. Mendes[1], **R. J. Rangel**[1], **A. B. Pavan**[1], **A. F. Tavares da Silva**[1], **A. F. Oliveira**[1], **I. N. Drummond**[1], **G. de Assis Mello**[1]

[1] *GREMLING* – Grupo de Pesquisa e Estudos em Machine Learning
Universidade Federal de Itajubá (UNIFEI)
Caixa Postal 50 – CEP: 37500 903 – Itajubá – MG – Brazil

alan@unifei.edu.br

Abstract. *This work presents a discussion on the importance and types of mass spectrometry data pre-processing techniques when used in machine learning models to classify biomarkers. A survey of computational tools that perform such analysis was carried out and a dataset was organized.*

Resumo. *Neste trabalho é apresentada uma discussão sobre a importância e os tipos de técnicas de pré-processamento de dados de espectrometria de massa, quando são usados em modelos de aprendizado de máquina para classificar biomarcadores. Um levantamento de ferramentas computacionais que realizam tais tratamentos foi realizado e um dataset foi organizado.*

1. Introdução

Desde Dezembro de 2019, no município de Wuhan (Hubei, China), quando o primeiro caso de COVID-19 foi comunicado pela comissão de saúde do município à Organização Mundial de Saúde – OMS [1], o vírus rapidamente se espalhou pelo mundo. Com isso, vários grupos de pesquisa vem buscando formas de identificar as causas que levam as pessoas infectadas a óbito e como diagnosticar a presença do coronavírus. A seleção e classificação de biomarcadores em dados de espectrometria de massa (EM) empregando modelos de aprendizado de máquina (AM) é uma proposta promissora [Liebal et al. 2020] nessa direção. Recentemente, [Delafiori and Navarro 2021] propuseram a criação de um teste de diagnóstico de COVID-19 que utiliza AM na análise de dados de EM.

Essas técnicas também tem sido aplicadas, tanto no desenvolvimento de sensores metabólicos [Dias-Audibert et al. 2020], quanto no diagnóstico de outras doenças, tais como Dengue e o Zika vírus [Melo 2018].

Em todos esses importantes trabalhos mencionados foram utilizadas diferentes técnicas de pré-processamento (PP) dos dados brutos de EM para sua adequada inserção nos algoritmos de AM. Contudo, segundo investigou-se, tal assunto é muito pouco discutido em trabalhos que são essencialmente da área da saúde.

[1] https://www.who.int/

Neste contexto, este trabalho apresenta uma discussão sobre a importância e os tipos de PP de dados de EM e uma indicação de quais ferramentas podem ser utilizadas para realizar este tratamento.

2. Tipos de pré-processamento (PP)

Em geral, os equipamentos de EM podem gerar diferentes formatos de arquivos de saída de dados. Um dos mais populares é o .RAW, proveniente dos instrumentos científicos da Thermo Fisher Scientific (para análise no Xcalibur Software). Outros, de formato aberto, igualmente populares são o .mzML e o .mzXML.

A etapa de PP consiste na análise e manipulação de vetores de dados que estão contidos nos arquivos de saída e cujas componentes são, geralmente, a razão massa/carga (m/z), a intensidade e o tempo. Os algoritmos de PP identificam, alinham e normalizam os picos de intensidade medidos e filtram ruídos por meio de manipulações matemáticas e estatísticas dos dados. Por isso, é fundamental que essa etapa seja feita de maneira criteriosa, evitando assim, o comprometimento dos dados experimentais e a inserção de informações espúrias.

As seções 2.1, 2.2, 2.3 e 2.4 descrevem 4 técnicas de PP comumente empregadas na análise de dados de EM.

2.1. Filtragem dos dados (F)

A filtragem dos dados visa a redução da quantidade de ruídos provenientes da operação do instrumento de medição. Para tanto é interessante obter alguns arquivos de dados apenas com o ruído gerado pelo equipamento, ou seja, sem a presença da amostra. Isso será importante para a definição do limite inferior durante o processo de operação matemática de filtragem. Dentre os *softwares* mais comuns para esta técnica podemos citar o MZmine, Analyst , OpenMS e o XCMS [Li and Gaquerel 2021, Pérez-Cova et al. 2021].

2.2. Alinhamento (A)

Durante a operação de aquisição dos dados é interessante a utilização de uma solução calibrante conhecida, juntamente com sua amostra, em cada análise. Tal solução vai permitir um ajuste da exatidão de m/z dos íons, detectados em diferentes intervalos de peso molecular [Chaerkady et al. 2021]. Se tal prática não é adotada os picos de intensidade podem aparecer desalinhados ou deslocados podendo gerar classificações incorretas dos metabólitos. As principais causas dessas variações são devidas a mudanças de temperatura, variações no pH, degradação da coluna cromatográfica dentre outros. O alinhamento pode ser realizado utilizado uma referência ou durante a detecção dos picos, estimando a variabilidade dos sinais. Alguns *softwares*, já citados, possuem essa função, como é o caso do MZmine e o XCMS.

2.3. Detecção dos picos (D)

O objetivo desse processo é localizar de forma robusta os picos verdadeiros referentes aos metabólitos da amostra. Tal processo permite excluir picos falsos, principalmente os provenientes de ruídos do equipamento, reduzindo a complexidade dos dados, antes de sua análise. Dentre as bibliotecas e funções que realizam essa técnica podemos citar o "find_peaks"do Scipy, que encontra picos dentro de um sinal com base nas propriedades de pico, e o F-score do MZmine, que detecta os picos através da construção do cromatograma, que cria uma lista de massas e da deconvolução dos picos [Leier et al. 2020].

2.4. Normalização (N)

A normalização permite uma comparação quantitativa, removendo variações não desejadas entre as amostras. Isso garante que, nas análises multivariadas, a comparação entre sinais de duas amostras preparadas em concentrações distintas possam ser realizados. O MZMine é um dos *softwares* mais utilizados para esta função.

Desta forma, pode-se observar o quão importante é a etapa de PP dos dados e como ela interfere diretamente na qualidade do dado que será inserido nos modelos de AM, afetando seu desempenho de maneira significativa.

3. Discussões

A apresentação e o detalhamento das técnicas de PP dos dados de EM mostram-se cruciais para o entendimento e reprodução dos resultados obtidos dos modelos de AM de diversos trabalhos que abordam o tema.

Após uma busca na literatura foi realizada uma compilação dos *softwares* mais utilizados no PP de dados de EM. Na Tabela 1 foram indicadas uma descrição breve, o tipo e as principais funções disponíveis em cada *software*.

Tabela 1. Principais softwares para tratamento de dados de EM.

Software	Descrição	Tipo	TP
MZMine 2	Software de código aberto, baseado nas ferramentas originais do MZmine descritas na publicação Bioinformatics de 2006.	Livre	F-A-D-N
XCMS	Software de bioinformática, criado pelo Laboratório Siuzdak da Scripps Research.	Livre	F-A-D-N
OpenMS	Software de código aberto projetado para a análise de dados MS.	Livre	F-A-D-N

Dentre os *softwares* analisados pode-se apontar, como uma opção bastante acessível, o MZmine já que é um *software open source* e possui diversos tutoriais sobre seu funcionamento e operação. O MZmine fornece diferentes algoritmos para detecção de picos: o algoritmo de limiar recursivo reduz o número de falsos positivos evitando a detecção de ruído; o algoritmo de transformação *Wavelet* é adequado para filtragem de dados com ruído; o algoritmo de "Massa exata"supõe espectros de alta qualidade e determina o centro de cada m/z utilizando a largura a meia altura.

Um *dataset* mais detalhado com outros *softwares* utilizados no tratamento de dados de EM pode ser encontrado no repositório *Github* do grupo de pesquisa – Gremling Research Group

Com esse levantamento realizado espera-se lançar alguma luz e apontar algumas estratégias de abordagem ao problema da apresentação e discussão do PP de dados de EM em trabalhos que aplicam modelos de AM na área de saúde.

4. Considerações Finais

A análise de dados de EM revela-se complexa e delicada devido, especialmente, à natureza do aparato experimental e das características dos materiais estudados. Essa di-

ficuldade vem sendo gradualmente vencida com o advento dos algoritmos de AM que são usados para selecionar, classificar e catalogar substâncias de interesse para a área da saúde. Contudo, sem um bom PP dos dados a análise executada por esses algoritmos fica extremamente prejudicada afetando as conclusões e a reprodução dos resultados obtidos.

Observa-se ser premente uma melhor apresentação e discussão das técnicas de PP utilizadas em todo trabalho, na área da saúde, que lida com dados de EM como entrada em modelos de AM. Uma última questão ainda merece consideração: Existiria alguma técnica ou algoritmo alternativo ao PP que permita o uso direto de dados brutos de EM? Uma possível resposta, que ainda pretende-se investigar, pode surgir da análise do emprego de redes neurais profundas. Esse procedimento pode ser promissor especialmente na tarefa de classificação de biomarcadores.

Referências

Chaerkady, R., Zhou, Y., Delmar, J. A., Weng, S. H. S., Wang, J., Awasthi, S., Sims, D., Bowen, M. A., Yu, W., Cazares, L. H., Sims, G. P., and Hess, S. (2021). Characterization of citrullination sites in neutrophils and mast cells activated by ionomycin via integration of mass spectrometry and machine learning. *Journal of Proteome Research*.

Delafiori, J. and Navarro, e. (2021). Covid-19 automated diagnosis and risk assessment through metabolomics and machine learning. *Analytical Chemistry*, 93(4):2471–2479. PMID: 33471512.

Dias-Audibert, F. L., Navarro, L. C., de Oliveira, D. N., Delafiori, J., Melo, C. F. O. R., Guerreiro, T. M., Rosa, F. T., Petenuci, D. L., Watanabe, M. A. E., Velloso, L. A., Rocha, A. R., and Catharino, R. R. (2020). Combining machine learning and metabolomics to identify weight gain biomarkers. *Frontiers in Bioengineering and Biotechnology*, 8:6.

Leier, H. C., Weinstein, J. B., Kyle, J. E., Lee, J.-Y., Bramer, L. M., Stratton, K. G., Kempthorne, D., Navratil, A. R., Tafesse, E. G., Hornemann, T., Messer, W. B., Dennis, E. A., Metz, T. O., Barklis, E., and Tafesse, F. G. (2020). A global lipid map defines a network essential for zika virus replication. *Nature Communications*, 11(1).

Li, D. and Gaquerel, E. (2021). Next-generation mass spectrometry metabolomics revives the functional analysis of plant metabolic diversity. *Annual Review of Plant Biology*, 72(1).

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., and Blank, L. M. (2020). Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites*, 10(6).

Melo, C. F. O. R. e. a. (2018). A machine learning application based in random forest for integrating mass spectrometry-based metabolomic data: A simple screening method for patients with zika virus. *Frontiers in Bioengineering and Biotechnology*, 6:31.

Pérez-Cova, M., Bedia, C., Stoll, D. R., Tauler, R., and Jaumot, J. (2021). MSroi: a preprocessing tool for mass spectrometry-based studies. *Chemometrics and Intelligent Laboratory Systems*, page 104333.