Avaliação de técnicas de remoção de sobreposição em projeções multidimensionais

Wilson Estécio Marcílio Júnior, Danilo Medeiros Eler
Universidade Estadual Paulista – UNESP
Presidente Prudente – SP, Brazil
wilson_jr@outlook.com, danilome@gmail.com

***Resumo*—A Visualização busca fornecer entendimento sobre conjuntos de dados por meio de representações gráficas. Para dados multidimensionais, representações gráficas podem ser geradas após a aplicação de técnicas de projeção multidimensional, no entanto, o processo de redução de dimensionalidade impõe sobreposição nos marcadores que representam as instâncias de dados, dificultando o processo de análise. Este trabalho apresenta uma avaliação de quatro técnicas de remoção de sobreposição mediante métricas que consideram relações de vizinhança, classe e similaridade.**

***Abstract*—Visualization aims at providing knowledge about data sets using visual representations. For multidimensional data, visual representations can be generated after application of multidimensional projection techniques, however, the dimensionality reduction process imposes overlaps among the markers used to represent data instances, impairing the analysis process. This work presents an evaluation of four overlap removal techniques through metrics that consider neighborhood, class and similarity relations.**

I. Introdução

Atualmente, existe uma grande preocupação com a dificuldade em analisar grandes conjuntos de dados. A Visualização busca fornecer o entendimento aos analistas por meio de representações gráficas geradas de acordo com esses dados. Para dados multidimensionais, uma classe de técnicas amplamente utilizada são as projeções multidimensionais [1]–[3]. Uma técnica de projeção multidimensional mapeia dados de alta dimensão ($\mathbb{R}^m$) para dados de baixa dimensão ($\mathbb{R}^p$), com o objetivo de manter as relações de similaridade e vizinhança tanto quanto possível, isto é, se determinadas instâncias são similares no espaço multidimensional tais instâncias também devem ser similares no espaço projetado.

Uma característica inerente do processo de redução multidimensional é a sobreposição dos marcadores utilizados para representar as instâncias dos conjuntos de dados. Esse problema de sobreposição, que é intensificado pelo aumento da dimensionalidade e tamanho do conjunto, prejudica a análise efetiva das visualizações geradas por projeções multidimensionais, visto que diversas instâncias ficam obstruídas. Neste trabalho, é apresentada a avaliação de quatro técnicas de remoção de sobreposição mediante cinco métricas de análise que consideram relações de similaridade, vizinhança e classe, assim como os tempos de execução e as características mais importantes de cada técnica.

O restante deste artigo está organizado da seguinte maneira. Na Seção II são apresentadas as técnicas de remoção de sobreposição avaliadas. Na Seção III são apresentados os resultados da avaliação. Por fim, uma conclusão é apresentada na Seção IV.

II. Técnicas de Remoção de Sobreposição

As técnicas de remoção de sobreposição consideram diferentes abordagens para diminuir a sobreposição dos marcadores, utilizando tanto estratégias simples quanto algoritmos que minimizam funcionais de energia. A técnica ***ProjSnippet*** [4] utiliza um funcional de energia construído por dois componentes, um que codifica as relações de vizinhança e outro que codifica a sobreposição entre as instâncias. Na técnica *RWordle* [5], primeiramente as instâncias são ordenadas de duas maneiras possíveis: *Linear Ordering (**RWordle-L**)*, considerando uma linha de varredura especificada por um ângulo α; e *Concentric Ordering (**RWordle-C**)*, considerando a distância para o centro geométrico da projeção. Conforme as instâncias são projetadas, as sobreposições são removidas por meio de um processo que move cada instância em espiral, até encontrar uma posição livre. A técnica ***PRISM*** [6] opera minimizando uma função de *stress* que tem por objetivo remover a sobreposição de cada par de nós conectado por uma aresta em um grafo de proximidades. Por fim, a técnica ***VPSC*** [7] constrói um conjunto de restrições de sobreposição e as remove adicionando as instâncias em blocos que possam contê-las sem a sobreposição.

III. Avaliação

Para avaliação das técnicas descritas anteriormente, foram utilizadas quatro métricas de análise amplamente empregadas na literatura para avaliação de projeções multidimensionais, além de uma métrica utilizada para avaliação da modificação do *layout* inicial. Essas métricas são úteis para verificar se os algoritmos mantêm as relações de classe, similaridade e vizinhança impostas pelas técnicas de projeção:

- A métrica *Similaridade de Layout* (**SL**) [6] verifica o quanto as estruturas foram modificadas no processo de remoção de sobreposição. Quanto menor for o valor para a métrica **SL**, melhor a técnica de remoção de sobreposição;
- A métrica *Stress* [8] calcula o quanto de informação foi perdida durante a projeção. Quanto menor for o valor de *Stress*, melhor será a técnica de projeção;

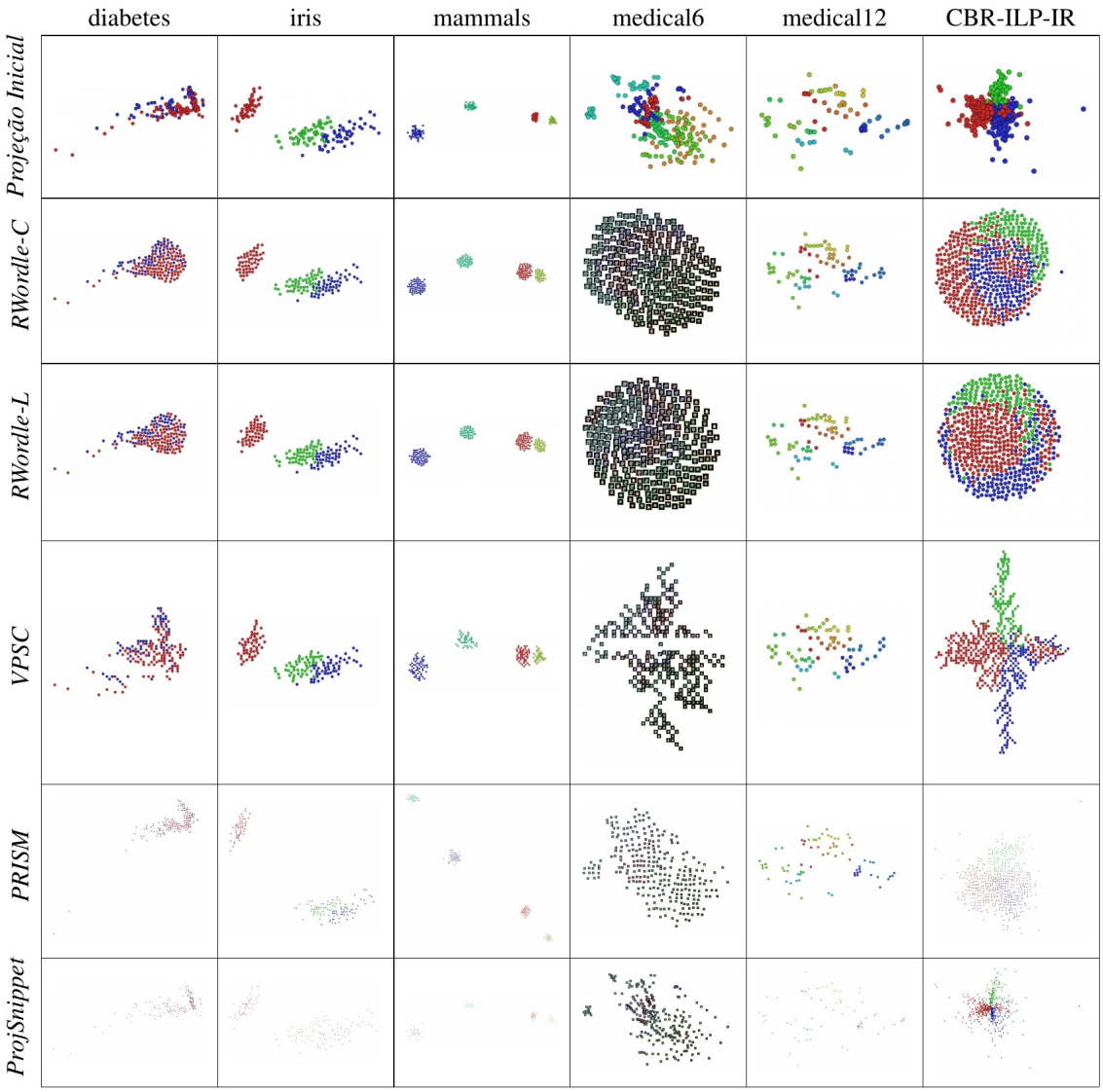

Figura 1. Projeções iniciais dos conjuntos de dados e *layouts* após a remoção de sobreposição.

- A métrica *Coeficiente de Silhueta* (**CS**) [9] é utilizada para interpretar a consistência dos agrupamentos de dados, com valor variando entre -1 a 1. Melhor qualidade é indicada por valores próximos de 1;
- A métrica *Neighborhood Hit* (**NH**) [2] é utilizada para avaliar a percepção de classes em uma projeção, isto é, quão bem uma projeção pode separar as classes;
- A métrica *Neighborhood Preservation* (**NP**) [10] é utilizada para avaliar uma projeção segundo a preservação de vizinhança. Assim como para a métrica **NH**, o valor da métrica **NP** varia de 0 a 1, com melhor preservação de vizinhança indicada por valores próximos de 1;
- Por fim, são apresentados os tempos de execução para cada algoritmo.

Vale ressaltar que o resultado para as métricas **NH** e **NP** são calculados considerando vizinhanças de 1 a 30.

Os testes foram efetuados mediante seis conjuntos de dados: *diabetes*, que contém 169 instâncias de pacientes descritos por 8 atributos; *iris*, que contém 150 instâncias de flores Iris descritas por 4 atributos; *Mammals*, que contém 183 instâncias de mamíferos descritos por 72 atributos; *Medical6* e *Medical12*, que contém 270 e 72 imagens de MRI, respectivamente, e são descritos por 28 atributos; e *CBR-ILP-IR*, que contém 574 documentos descritos por 43 termos. As projeções dos conjuntos de dados, realizadas pela técnica *IDMAP* [1], e a aplicação das técnicas de remoção de sobreposição podem ser visualizadas na Figura 1.

A. Similaridade de Layout

Na Figura 2 são apresentados os resultados para a métrica **SL**. É possível perceber que as técnicas *PRISM* e *ProjSnippet* desempenham o melhor papel em preservar as estruturas. O ótimo resultado dado pela técnica *ProjSnippet* também é

devido ao fato de que a técnica não consegue remover toda sobreposição, tornando o *layout* gerado mais similar com a projeção inicial.

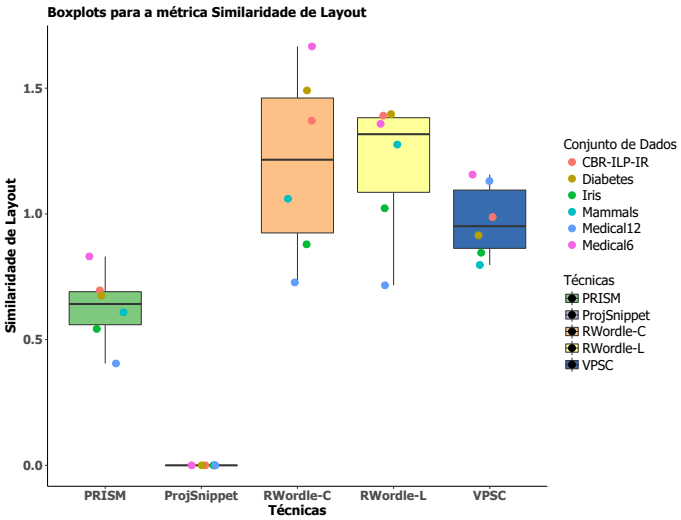

Figura 2. Boxplots para a métrica Similaridade de *Layout*.

Note na Figura 2 que, como os conjunto de dados *Diabetes*, *CBR-ILP-IR* e *Medical6* apresentam alta taxa de sobreposição, tais conjuntos prejudicam o resultado das técnicas.

B. Stress

Para a métrica *Stress* (ver Figura 3), a técnica *ProjSnippet* novamente apresentou resultados muito similares aos da projeção. No entanto, os resultados para as técnicas restantes também são satisfatórios. A projeção inicial também foi inserida para efeitos de comparação.

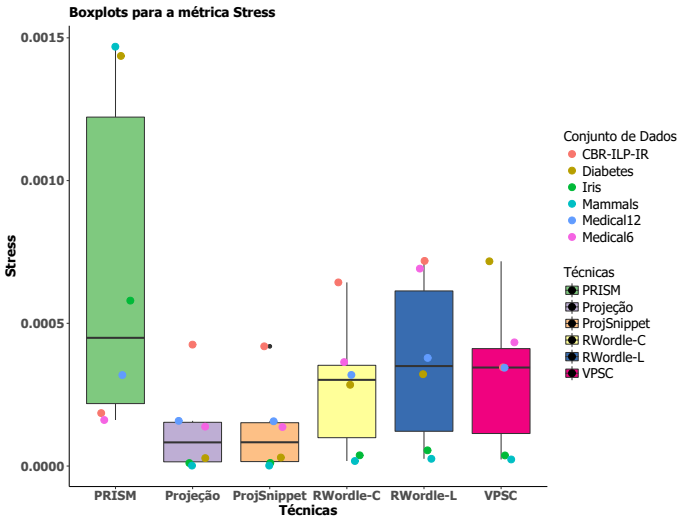

Figura 3. Boxplots para a métrica *Stress*.

C. Coeficiente de Silhueta

Considerando a métrica **CS**, as técnicas *PRISM*, *ProjSnippet* e *VPSC* apresentaram resultados muito similares, visto que conseguem preservar as estruturas de classes da projeção, como pode ser verificado na Figura 4.

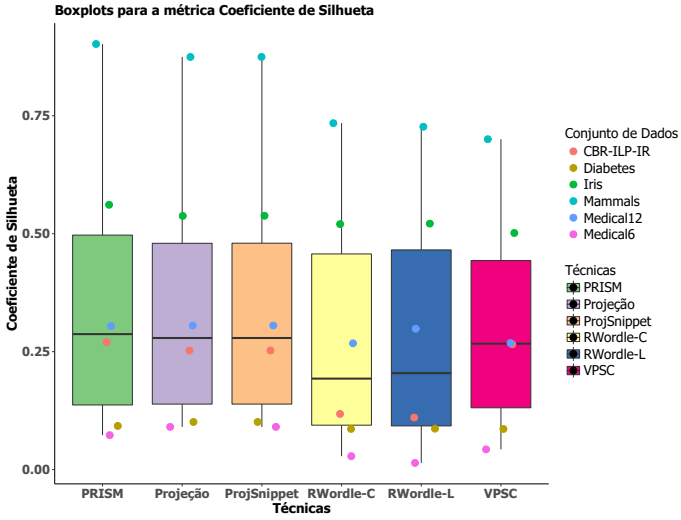

Figura 4. Boxplots para a métrica Coeficiente de Silhueta.

Enquanto que para todos os conjuntos de dados o comportamento é similar, as técnicas *RWordle-C* e *RWordle-L* apresentaram resultados inferiores principalmente para o conjunto de dados *CBR-ILP-IR*, já que não conseguem manter a separação de classes.

D. Neighborhood Hit

Para a métrica **NH**, pode-se notar na Figura 5 que as técnicas *PRISM* e *ProjSnippet* apresentaram os melhores resultados, mais similares aos da projeção.

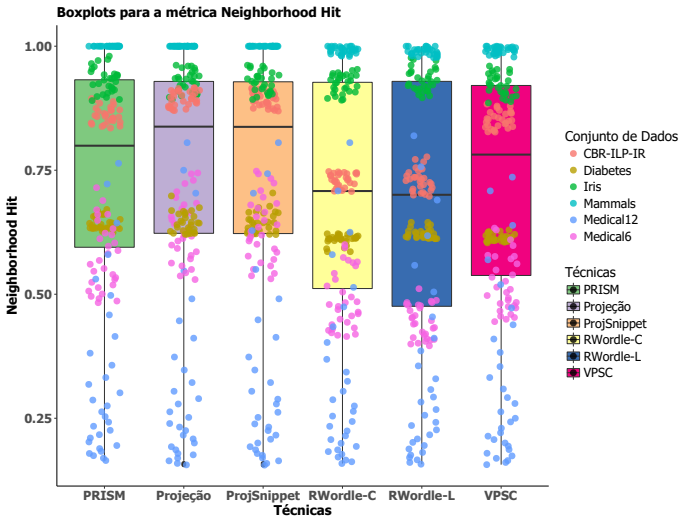

Figura 5. Boxplots para a métrica *Neighborhood Hit*.

Enquanto os melhores resultados são apresentados para os conjuntos de dados com separação de classes bem definida, as técnicas *RWordle-C* e *RWordle-L* diminuem consideravelmente a qualidade da projeção para conjuntos mais problemáticos, em que classes estão próximas uma das outras.

E. Neighborhood Preservation

Na Figura 6 são apresentados os resultados para a métrica **NP**. Os conjuntos de dados mais problemáticos, isto é, aqueles em que há alta taxa de sobreposição, diminuem o desempenho das técnicas. Novamente, as técnicas *PRISM* e *ProjSnippet* apresentaram resultados similares aos da projeção inicial.

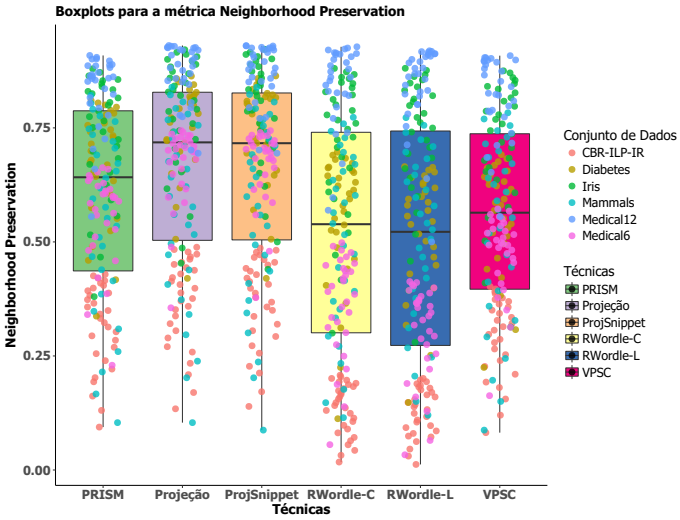

Figura 6. Boxplots para a métrica *Neighborhood Preservation*.

F. Tempo de execução

Considerando o tempo de execução (ver Figura 7) , os conjuntos de dados *Medical6* e *CBR-ILP-IR* foram responsáveis pela maior variação de tempo dos algoritmos. Enquanto as técnicas *PRISM*, *RWordle-C* e *RWordle-L* não apresentaram escalabilidade no tempo de execução, as técnicas *ProjSnippet* e *VPSC* conseguiram manter bom desempenho.

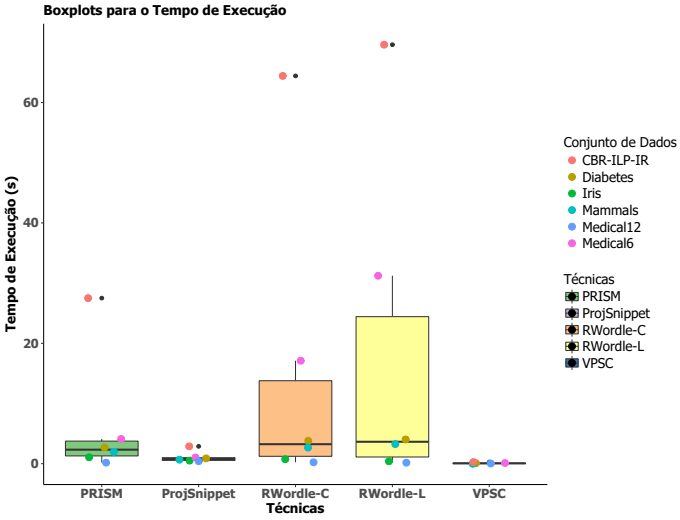

Figura 7. Boxplots para o tempo de execução.

A partir da análise, algumas considerações podem ser feitas. As técnicas *RWordle-C* e *RWordle-L* não consideram relações estruturais e de vizinhança e, com isso, seu desempenho é prejudicado em conjuntos de dados com alta taxa de sobreposição. A técnica *ProjSnippet* possui a característica de preservar estruturas, agrupamentos e a vizinhança inicial de forma efetiva, contudo, tende adicionar espaços excessivos à tela de visualização. A técnica *VPSC* gera *layouts* que preservam a estrutura geral da projeção, além de obter bom desempenho considerando o tempo de execução, contudo, apresenta desempenho mediano em relação as métricas de avaliação. Finalmente, a técnica *PRISM* apresenta bons resultados considerando as métricas utilizadas, sendo capaz de preservar as relações impostas pela projeção, mas possui dificuldade em manter o *layout* compacto.

IV. Conclusão

Técnicas de projeção são uma importante ferramenta para análise de conjuntos de dados multidimensionais, no entanto, a sobreposição dos marcadores torna o processo de análise difícil. Neste trabalho, avaliamos diferentes técnicas de remoção de sobreposição para apresentar suas características. As técnicas de remoção de sobreposição foram avaliadas segundo métricas que consideram relações de similaridade e estruturas de vizinhança e classes. As técnicas *ProjSnippet* e *PRISM*, em geral, apresentaram resultados superiores considerando as métricas utilizadas, no entanto, as diferentes características dos algoritmos podem ser exploradas de acordo com as características dos conjuntos de dados.

Agradecimentos

Os autores agradecem o suporte financeiro da agência de fomento Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – processo N^o $15/182380-9$.

Referências

[1] R. Minghim, P. F. V., and A. A. Lopes, "Content-based text mapping using multi-dimensional projections for exploration of document collections," *Visualization and Data Analysis*, 2006.

[2] F. V. Paulovich, L. G. Nonato, M. Rosane, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE Transactions on Visulization and Computer Graphics*, vol. 3, pp. 564–575, 2008.

[3] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[4] E. Gomez-Nieto, F. S. Roman, P. Pagliosa, W. Casaca, E. Helou, M. Oliveira, and L. Nonato, "Similarity preserving snippet-based visualization of web search results," *TVCG*, vol. 20, no. 3, pp. 457–470, 2014.

[5] K. Koh, B. Lee, B. Kim, and J. Seo, "Maniwordle: Providing flexible control over wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1190–1197, 2010.

[6] E. R. Gansner and Y. Hu, "Efficient, proximity-preserving node overlap removal," *Journal of Graph Algorithms and Applications*, vol. 14, no. 1, pp. 53–74, 2010.

[7] T. Dwyer, K. Marriott, and P. J. Stuckey, "Fast node overlap removal," *Proceedings of the* 13^{th} *International Conference on Graph Drawing*, pp. 153–164, 2006.

[8] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 115–129, 1964.

[9] L. Kaufman and P. J. Rousseuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Principles and Practice: Wiley-Interscience, 2005.

[10] F. V. Paulovich and R. Minghim, "Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections," *IEEE Transactions on Visulization and Computer Graphics*, vol. 14, no. 6, pp. 1229–1236, 2008.