Técnicas de Projeção para Identificação de Grupos e Comparação de Dados Multidimensionais Usando Diferentes Medidas de Similaridade

Paulo Joia[1], **Luis Gustavo Nonato**[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC) – USP, São Carlos, Brasil

{pjoia,gnonato}@icmc.usp.br

Abstract. *This work explores the potential of the projection techniques to solve problems related to:* ***clustering*** *and* ***similarity search*** *in multidimensional data. For clustering data, a local and interactive projection technique capable of projecting data with effective preservation of distances was developed and also a new clustering method, which operates in the visual space, ensuring that clusters are not fragmented during the visualization. For the similarity search, we build a family of class-specific metrics and the fuzzy set theory was used to estimate a degree of uncertainty that is embedded in the metric, increasing its precision. The results confirm the effectiveness of the developed techniques, which represent significant contributions for this investigation area.*

Resumo. *Este trabalho explora o potencial das técnicas de projeção para resolver problemas relacionados à:* ***identificação de agrupamentos*** *e* ***busca por similaridade*** *em dados multidimensionais. Para identificação de agrupamentos foi desenvolvida uma técnica de projeção local e interativa capaz de projetar dados com ótima preservação de distâncias, além de um novo método para identificação de agrupamentos, o qual opera no espaço visual, garantindo que os grupos obtidos não fiquem fragmentados durante a visualização. Para as buscas por similaridade em dados multidimensionais, uma família de métricas baseada em classes foi construída e a teoria dos conjuntos fuzzy foi usada para estimar um valor de incerteza que é transferido para a métrica, aumentando sua precisão. Os resultados confirmam a efetividade das técnicas desenvolvidas, as quais representam significativa contribuição nesta área de investigação.*

1. Introdução

Visualização de informação desempenha um papel importante na organização e exploração de dados multidimensionais, graças à capacidade de percepção visual do sistema cognitivo humano, assim como técnicas de projeção, as quais reduzem a dimensionalidade do conjunto de dados permitindo visualizar informações muitas vezes ocultas na alta dimensão. Este trabalho explora o potencial das técnicas de visualização de informação com ênfase em projeção para auxiliar na *identificação de agrupamentos* e *busca por similaridade em dados multidimensionais*.

1.1. Motivação

Consideráveis avanços têm sido observados nas técnicas de projeção nos últimos anos, com aplicações em diferentes domínios. Algumas até propõem soluções para os problemas de identificação de agrupamentos e busca por similaridade em dados multidimensionais, contudo, longe da solução ideal. Por exemplo, identificar grupos é uma tarefa

complexa e na maioria das vezes os grupos obtidos não correspondem à verdadeira natureza dos dados, normalmente organizados com base na geometria. Quanto à busca por similaridade, existem várias abordagens, todavia, poucas empregam medidas de similaridade realmente aptas a discriminar objetos de acordo com as classes existentes.

2. Contribuições

A tese de doutorado aqui apresentada, produziu resultados significativos na área de visualização de informação, com ênfase em projeção multidimensional. Resultados que podem ser comprovados pelas novas técnicas desenvolvidas e respectivas metodologias empregadas no desenvolvimento, conforme sumarizado abaixo. Para um exame mais detalhado sobre cada técnica, incluindo sua formulação matemática, algoritmo, resultados e comparações com outras técnicas, recomendamos que o leitor consulte [Joia 2015].

2.1. Técnica de Projeção Local Interativa

Local Affine Multidimensional Projection (LAMP) [Joia et al. 2011a] permite manipular pontos de controle no espaço visual de modo a organizá-los, possibilitando ao usuário guiar a projeção, porém, com uma grande vantagem sobre as demais técnicas que se apoiam em subconjunto de amostras: requer um número muito reduzido de pontos de controle como entrada, tornando-se ideal para aplicações interativas. Tem formulação matemática baseada em mapeamentos ortogonais, garantindo ótima preservação de distâncias durante a projeção multidimensional e, não depende de grafos de vizinhança para construir o mapeamento. É altamente precisa, com baixo custo computacional, apropriada para aplicações interativas envolvendo grandes volumes de dados. LAMP está atualmente entre as técnicas do estado da arte em relação à preservação de distâncias e eficiência computacional, além de permitir projetar dados explorando tanto relações globais como locais entre instâncias, de maneira efetiva. A Figura 1 mostra o potencial da LAMP ao estabelecer uma correlação visual entre conjuntos de dados, a princípio, sem qualquer conexão.

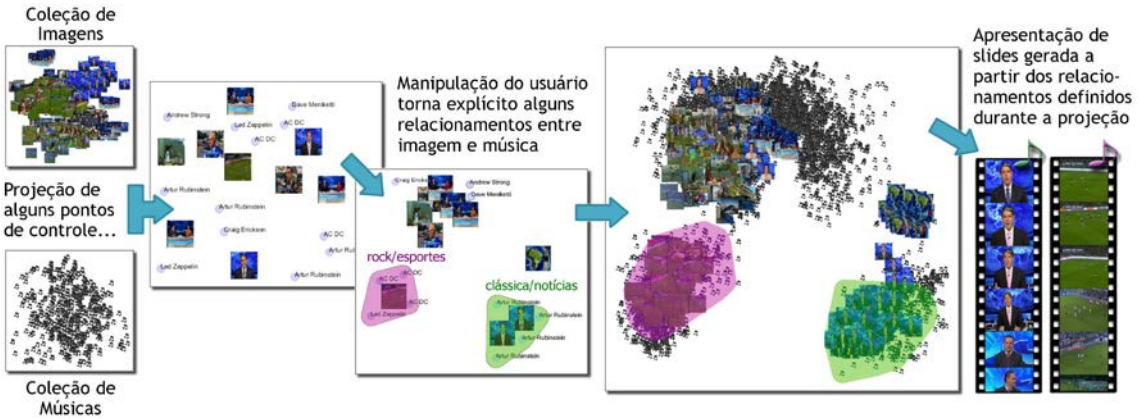

Figura 1. Utilizando a LAMP para correlacionar dados de diferentes naturezas. Inicialmente, uma projeção é criada para cada conjunto de dados, a partir de algumas amostras. A correlação entre as amostras é definida pelo usuário, agrupando objetos no espaço visual (imagens e músicas). Em seguida, os dados são projetados segundo as associações criadas pelo usuário. Por fim, as listas de objetos associados são usadas para criar uma apresentação de *slides* onde imagens e músicas são reproduzidas de forma sincronizada.

2.2. Método para Identificação de Grupos com Base em Projeção

Column Selection Method (CSM) [Joia et al. 2015], um método de visualização apoiado em projeção multidimensional que permite agrupar dados. CSM opera no espaço visual, garantindo que os grupos obtidos não fiquem fragmentados durante a visualização. É orientado por um mecanismo de amostragem determinístico, capaz de identificar instâncias representativas que correspondem a um certo padrão nos dados. O mecanismo de amostragem é baseado em decomposição matricial (SVD) e capaz de operar mesmo em conjuntos de dados desbalanceados. Além de identificar instâncias representativas, o mecanismo de amostragem pode ser ajustado para identificar os atributos mais relevantes de cada agrupamento obtido. Portanto, em um único *framework*, três tarefas são contempladas: *amostragem de dados*, *detecção de agrupamentos* e *seleção de atributos*. A Figura 2 ilustra a metáfora visual utilizada pela CSM para representar grupos e atributos, por meio de superfícies e nuvens de palavras.

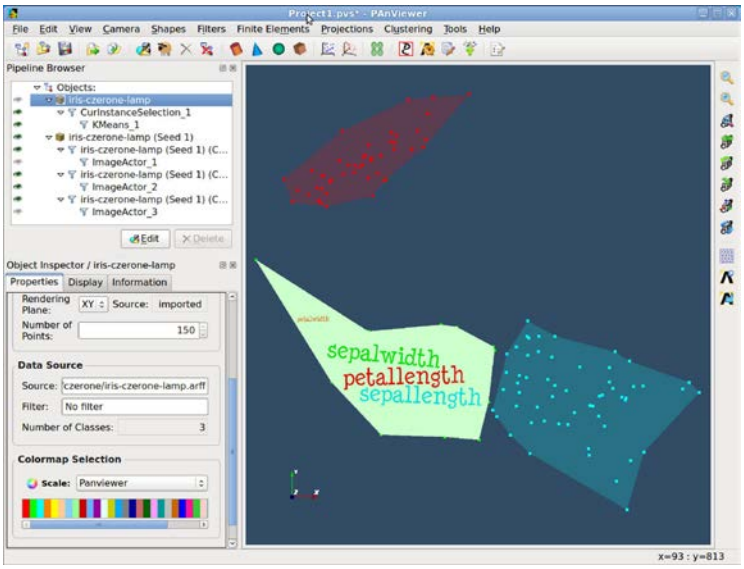

Figura 2. Metáfora visual utilizada pela CSM para representar grupos e atributos.

2.3. Família de Métricas Classes-Específicas

Muitas técnicas propõem medidas de similaridade para comparar dados multidimensionais, mas nenhuma diretamente relacionada às classes de objetos existentes no conjunto de dados. A *Class-Specific Multidimensional Projection* (CSMP) [Joia et al. 2012] é uma técnica de projeção baseada em uma família de métricas específicas por classe para projetar e comparar dados multidimensionais. As métricas são obtidas pela seleção dos atributos que melhor representam cada classe do conjunto de dados, de modo a minimizar a dissimilaridade entre pares de objetos pertencentes à mesma classe e, ao mesmo tempo, maximizá-la para objetos pertencentes a classes distintas. As métricas classes-específicas são avaliadas no contexto de recuperação de imagens por conteúdo para encontrar imagens similares a uma dada imagem de consulta. A lista de imagens similares pode ser retornada pelo sistema ou selecionada diretamente pelo usuário, a partir do *layout* da projeção, conforme exemplificado na Figura 3.

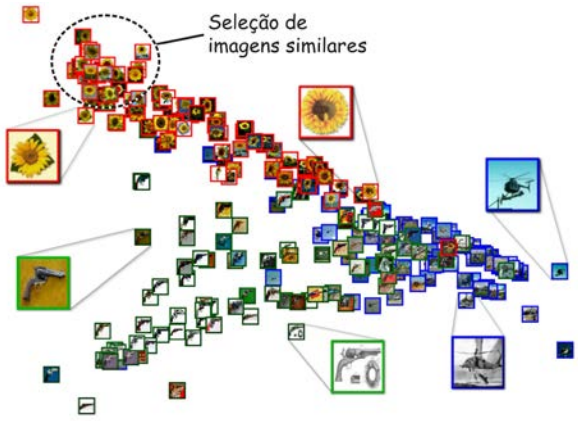

Figura 3. Seleção de imagens similares com o uso da CSMP.

2.4. Cálculo de Incerteza na Família de Métricas Classes-Específicas

Com o intuito de aumentar a precisão da família de métricas classes-específicas empregada na CSMP, uma nova técnica denominada *Class-Specific with Weight Image Retrieval* (CSWIRe) foi desenvolvida. Nesta abordagem, o usuário constrói um modelo a partir de um subconjunto de imagens, denominado "*modelo de classes*". A seguir, um classificador é aplicado sobre este modelo, retornando as melhores características e pesos que representam cada classe do modelo. Utilizando a teoria dos conjuntos *fuzzy*, um valor de incerteza é então calculado e associado à resposta do classificador para derivar uma família de métricas classes-específicas com pesos utilizada para comparar imagens com maior precisão. A Figura 4 ilustra o processo de recuperação de imagens da CSWIRe, utilizando uma interface gráfica apropriada.

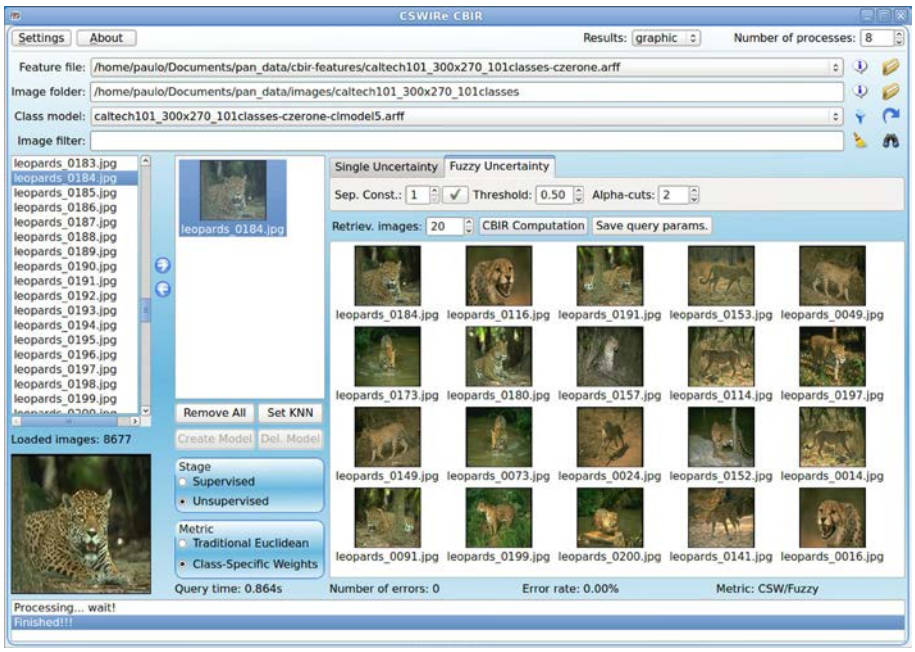

Figura 4. Interface gráfica da CSWIRe, mostrando o processo de recuperação de imagens por conteúdo.

3. Outras Produções Relevantes

Além das contribuições apresentadas anteriormente, merece destaque o trabalho desenvolvido em colaboração com o grupo de pesquisa da *Universidade de Calgary*[1], visando a exploração de espaços multidimensionais via projeção inversa, com uso da técnica intitulada *inverse-LAMP* (iLAMP). Esta técnica executa a projeção inversa através de mapeamentos locais afins que preservam a distância entre as novas amostras de modo preciso, já que ela segue os mesmos preceitos da LAMP. Desse modo, o usuário pode interativamente criar instâncias no conjunto de dados original, gerando assim, dados multidimensionais sintéticos além dos já existentes na disposição inicial. Para maiores detalhes sobre iLAMP, consulte [Dos Santos Amorim et al. 2012].

Para atender os requisitos do projeto, foram desenvolvidas algumas ferramentas computacionais. *Projection Analyzer* (PAn), um conjunto de bibliotecas de alto desempenho em ANSI C (versão inicial disponível em `http://sites.google.com/site/paulojoiafilho/tools`). Também foi implementado um módulo em Python para facilitar a execução das tarefas, que além de reutilizar o código em C, permite integração com pacotes de matemática numérica conhecidos. As interfaces gráficas mostradas nas Figuras 2 e 4, por exemplo, foram desenvolvidas a partir destas ferramentas.

4. Publicações, Prêmios e Estado da Arte

A técnica LAMP [Joia et al. 2011a] destaca-se como uma das técnicas de projeção mais precisas da atualidade. O recente trabalho apresentado por [Fadel et al. 2015] comprova este fato ao comparar várias técnicas de projeção (*stress* e tempo computacional), posicionando a LAMP como uma das técnicas de projeção do *estado da arte* em relação à preservação de distâncias e custo computacional. Além de um número considerável de `citações` em trabalhos científicos, LAMP também recebeu menção honrosa na *Conferência IEEE Information Visualization*, uma das principais dessa área de pesquisa do mundo (mais informações em `http://www.usp.br/agen/?p=79480`).

4.1. Lista de Publicações

1. [Joia et al. 2011a] "Local Affine Multidimensional Projection", *IEEE Transactions on Visualization and Computer Graphics* (**Qualis A1**).
2. [Joia et al. 2011b] "Projection-based Image Retrieval using Class-Specific Metrics", *24th SIBGRAPI*, IEEE Computer Society (**Qualis B1**).
3. [Joia et al. 2012] "Class-Specific Metrics for Multidimensional Data Projection Applied to CBIR", *The Visual Computer Journal* (**Qualis B1**).
4. [Dos Santos Amorim et al. 2012] "iLAMP: Exploring High-Dimensional Spacing through Backward Multidimensional Projection", *IEEE VAST'12* (**Qualis B1**).
5. [Casaca et al. 2013] "Spectral Image Segmentation Using Image Decomposition and Inner Product-Based Metric", *Journal of Mathematical Imaging and Vision* (**Qualis A2**).
6. [Joia et al. 2015] "Uncovering Representative Groups in Multidimensional Projections", *Computer Graphics Forum* (EuroVis'15) (**Qualis A2**).
7. Joia, P.; da Silva, S. F.; Batista, J.; Nonato, L. G. "Class-specific metrics with weights and uncertainty modeling using fuzzy sets applied to content-based image retrieval", *Expert Systems with Applications*, Elsevier. Submetido, em processo de revisão (**Qualis A1**).

[1] *Interactive Reservoir Modeling and Visualization Group*, Universidade de Calgary, Alberta, Canadá.

4.2. Prêmios

1. [Joia et al. 2011a] *Local Affine Multidimensional Projection*. Honorable Mention Award - IEEE InfoVis 2011 (`http://www.cad.zju.edu.cn/home/vag/~cwf/vispapers/infovis2011.html`).
2. [Joia et al. 2011b] *Projection-based Image Retrieval using Class-Specific Metrics*. Best Paper Award - Sibgrapi 2011 (`https://www.computer.org/csdl/proceedings/sibgrapi/2011/4548/00/4548z022.pdf`).

5. Conclusões

Este trabalho apresentou soluções para problemas de *identificação de agrupamentos* e *busca por similaridade em dados multidimensionais*, utilizando técnicas de projeção. Para cada problema, duas novas técnicas foram desenvolvidas, atingindo resultados expressivos em cada uma delas, uma das quais, inclusive, equipara-se ao estado da arte em termos de precisão e eficiência computacional (LAMP).

Agradecimentos

Agradecemos ao Instituto de Ciências Matemáticas e de Computação (ICMC/USP) pelo suporte ao projeto e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela apoio financeiro (Processo #2010/07367-9).

Referências

Casaca, W., Paiva, A., Gomez-Nieto, E., Joia, P., and Nonato, L. G. (2013). Spectral Image Segmentation Using Image Decomposition and Inner Product-Based Metric. *Journal of Mathematical Imaging and Vision*, 45(3):227–238.

Dos Santos Amorim, E. P., Brazil, E. V., Daniels, J., Joia, P., Nonato, L. G., and Sousa, M. C. (2012). iLAMP: Exploring High-Dimensional Spacing through Backward Multidimensional Projection. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 53–62. IEEE.

Fadel, S. G., Fatore, F. M., Duarte, F. S. L. G., and Paulovich, F. V. (2015). LoCH: A neighborhood-based multidimensional projection technique for high-dimensional sparse spaces. *Neurocomputing*, 150(Part B):546–556.

Joia, P. (2015). *Técnicas de projeção para identificação de grupos e comparação de dados multidimensionais usando diferentes medidas de similaridade*. Tese de Doutorado, Universidade de São Paulo - USP, São Carlos, SP. Disp. em: <`http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29032016-143247`>. Acesso em: 25 maio 2016.

Joia, P., Coimbra, D., Cuminato, J. A., Paulovich, F. V., and Nonato, L. G. (2011a). Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571.

Joia, P., Gomez-Nieto, E., Batista Neto, J., Casaca, W., Botelho, G., Paiva, A., and Gustavo Nonato, L. (2012). Class-specific metrics for multidimensional data projection applied to CBIR. *The Visual Computer*, 28(10):1027–1037.

Joia, P., Gomez Nieto, E., Botelho, G., Batista Neto, J., Paiva, A., and Nonato, L. G. (2011b). Projection-based Image Retrieval using Class-Specific Metrics. In Lewiner, T. and Torres, R., editors, *24th SIBGRAPI*, pages 125–132, Maceio, AL. IEEE Computer Society.

Joia, P., Petronetto, F., and Nonato, L. G. (2015). Uncovering Representative Groups in Multidimensional Projections. *Computer Graphics Forum*, 34(3):281–290.