ANÁLISE DE MÉTODOS PARA REDUÇÃO DE DIMENSIONALIDADE DE BANCO DE DADOS RELACIONAIS DE IMAGENS DE SENSORIAMENTO REMOTO APLICADO À AGRICULTURA.

Alessandra Leachenski (PROVIC) e-mail: alessandraleachenski@gmail.com, Silvia Ribeiro Mantuani(Co-autora) e-mail: silviamantuani@gmail.com, Maria Salete Marcon Gomes Vaz (Orientadora) e-mail: salete@uepg.br.

Universidade Estadual de Ponta Grossa/Departamento de Informática

Área: Ciência da Computação, Subárea: Metodologia e Técnicas da Computação.

Palavras-chave: Banco de dados relacional, Sensoriamento remoto, agricultura.

Resumo

O processamento de imagens de sensoriamento remoto é um processo complexo e que resulta em uma enorme quantidade de dados de difícil conversão em informação útil para a pesquisa cientifica, no entanto, essas informações podem ser refinadas utilizando métodos de redução de dimensionalidade quando são armazenadas em bancos de dados relacionais. O objetivo desse artigo é analisar métodos para redução de dimensionalidade de bancos de dados relacional e aplicar esses métodos de redução em bancos de dados onde foram armazenados estes tipos de informação. Neste estudo foram estudados três métodos de redução e por meio do software R foram aplicadas funções de redução de dimensionalidade em um banco de dados, obtido por meio da segmentação da imagem LandSat- OLI 8. Após a comparação das características de cada método foi definido qual o melhor método para se utilizar. O método que proporcionou melhor redução de dimensionalidade foi a Análise de Componentes Principais, por apresentar redução de mais de 50% do banco de dados.

Introdução

A obtenção de informações ou novos conhecimentos de banco de dados relacional, gerados pelo processamento de imagens de sensoriamento remoto é complexo, principalmente, devido ao grande volume de dados armazenado nos mesmos (BINGHAM; MANNILA, 2001).

A grande quantidade de dados armazenados gera a necessidade de desenvolver novos métodos e ferramentas que possam transformar, de maneira inteligente e automática, os dados processados em informações úteis e conhecimento (KUMAR; STEINBACH; TAN, 2009).

Para a escolha de métodos para redução de dimensionalidade devem ser avaliadas as características que os métodos possuem e quais proporcionarão melhor desempenho para a redução de dimensionalidade. (VAN DER MAATEN; POSTMA; VAN DEN HERIK, 2009).

Este artigo tem por objetivo analisar métodos para redução de dimensionalidade de bancos de dados relacional, aplicar esses métodos em

imagens obtidas por sensoriamento remoto e analisar qual método gera uma redução maior na dimensionalidade dos dados.

Materiais e Métodos

Neste trabalho foi utilizado um banco de dados obtido por meio da segmentação da imagem LandSat- OLI 8, obtida no dia 7 de julho de 2015. Os dados foram cedidos pelo Departamento de Geografia da Universidade Estadual de Ponta Grossa - UEPG.

Utilizou-se de três tipos de métodos de redução. O primeiro foi à análise dos componentes principais, para seleção dos descritores mais relevantes. O segundo método foi o escalonamento multidimensional, obtendo-se os dados em forma bidimensional. E o terceiro, e último método, foi o método *Locally Linear Embedding*.

Foi utilizado o Software R, com implementação de *scripts* com as funções dos métodos. As bibliotecas utilizadas foram *stats*, *Mvar.pt* e *lle*.

Para a aplicação do método de Componentes Principais é necessário preparar os dados para análise, deixando os descritores na mesma escala, ou seja, realizar a normalização dos dados, nessa pesquisa a normalização dos dados é resultante da média e desvio padrão. Após a normalização, o processo da Análise de Componentes principais pôde ser aplicado a matriz de dados obtendo-se assim um novo conjunto de dados com os componentes principais.

Para o método de Escalonamento Multidimensional ser aplicado é necessário medir as dissimilaridades entre os dados, nessa pesquisa as dissimilaridades são obtidas através da distância Euclidiana, a mais popular para esse tipo de aplicação. Com as dissimilaridades calculadas o método pôde ser aplicado, resultado em um gráfico bidimensional.

O primeiro passo para a aplicação do método *Locally Linear Embedding* é encontrar os vizinhos mais próximos de cada descritor no conjunto de dados, após isso é necessário expressar cada descritor com uma combinação linear dos outros descritores e encontrar as coordenadas de cada objeto no espaço p-dimensional. O método foi aplicado com diferentes números de vizinhos mais próximos, sendo escolhido o *layout* que apresentou o melhor resultado visualmente (oito vizinhos).

Resultados e Discussão

Com base no resultado obtido pela técnica de Componentes Principais, foi encontrado o nível de relevância dos autovetores e seus respectivos descritores, permitindo que 70 (setenta) descritores fossem analisados a partir das trinta e três primeiras Componentes Principais, com mais de 90% de variabilidade. As Componentes Principais não selecionadas do trinta e quatro à setenta representaram menos que 6% da variância total.

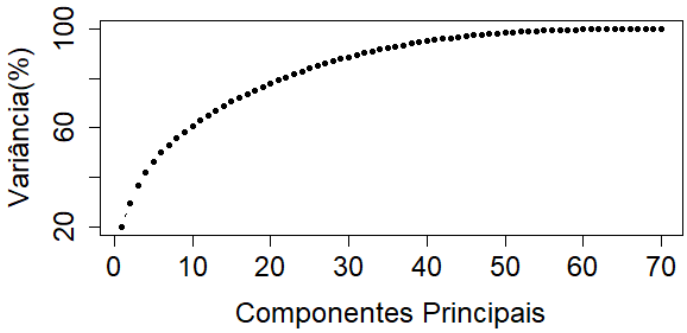


Figura 1- Variância dos componentes principais
Fonte: As autoras.

O gráfico na Figura 1 apresenta a variância para cada componente, onde é possível perceber que às trinta e três primeiras componentes garantem a variabilidade de 90,93%, desta forma, não é necessário analisar toda a informação, mas somente os descritores de variabilidade relevantes. Com este método, o número de Componentes Principais foi reduzido de setenta descritores para trinta e três descritores, formando um novo conjunto de dados menor que 50% do tamanho do conjunto de dados original.

O resultado obtido através do método de Escalonamento multidimensional está representado, de forma bidimensional, no gráfico da Figura 2 que representa a disposição espacial dos descritores em função de seu grau de semelhança.

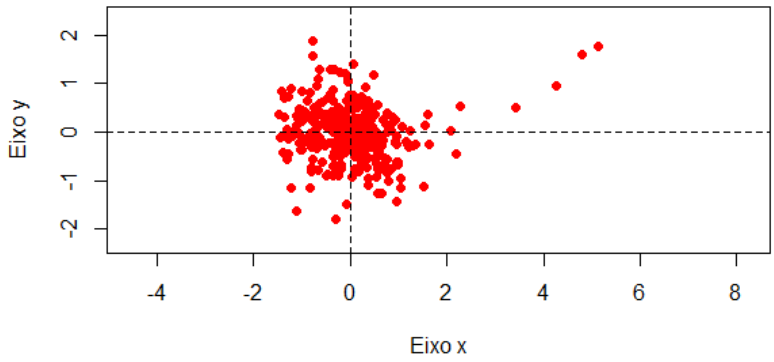


Figura 2- Projeção do Escalonamento Multidimensional
Fonte: As autoras

É possível perceber que aproximadamente quarenta descritores possuem dissimilaridades em relação aos demais e os outros trinta descritores possuem um alto grau de similaridade, ou seja, podem ser representados através de descritores semelhantes. Sendo assim, pode-se reduzir o banco em aproximadamente 43% do tamanho do conjunto de dados original.

A figura 3 apresenta um exemplo de projeção multidimensional usando o método *Locally Linear Embedding* (LLE), as coordenadas desses espaços de incorporação estão relacionadas a atributos significativos do conjunto de descritores.

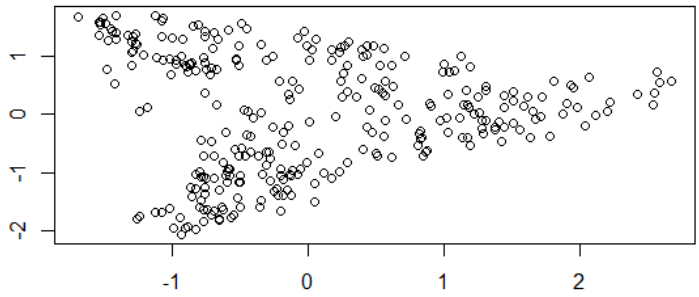


Figura 3- Mapeamento de descritores por meio de LLE
Fonte: As autoras

Por meio desse método não é possível perceber o quanto o conjunto de dados original foi reduzido, mas é possível notar, comparando a Figura 3 com a Figura 4, que a quantidade de descritores no espaço de incorporação não é tão reduzida em relação à quantidade de descritores iniciais.

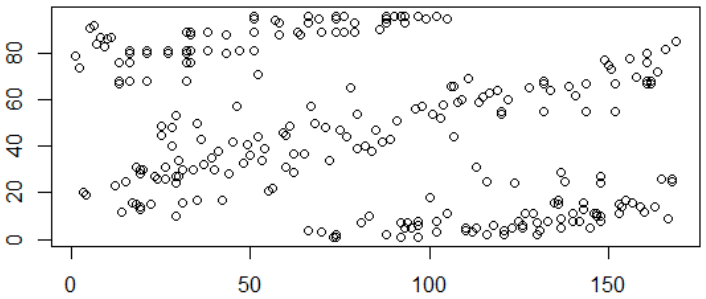


Figura 4- Descritores iniciais no espaço de incorporação
Fonte: As autoras

Conclusões

Existem diversos métodos para redução de dimensionalidade de bancos de dados relacionais. O método de Análise de Componentes Principais é um método clássico que concentra a informação presente em uma coleção de descritores iniciais, em uma nova coleção de descritores, sem perder informação relevante. (JOLLIFFE, 2002). O novo conjunto de dados, com dimensão reduzida, é formado pelas componentes principais.

O método de Escalonamento Multidimensional é uma técnica para a análise de dados de similaridade ou dissimilaridade de um conjunto de descritores. O escalonamento multidimensional tenta modelar os dados como distâncias entre pontos em um espaço geométrico, de forma que permita descobrir as similaridades entre cada descritor pertencente ao conjunto de dados.

O *Locally Linear Embedding* é um método de redução de dimensionalidade baseado em intuições geométricas simples. Essencialmente, ele tenta reduzir a dimensão dos dados enquanto preserva a relação entre pontos de dados vizinhos. A ideia básica é a de minimização global do erro de reconstrução do conjunto de todas as vizinhanças locais no conjunto de dados (RIDDER; DUIN, 2002).

Dentre os três métodos escolhidos e aplicados, podemos observar que o melhor foi o de Análise de Componentes Principais, que reduziu o banco de dados em aproximadamente 50%, enquanto o Escalonamento Multidimensional apresentou

a redução, do banco de dados, em apenas 43%. Com o método *Locally Linear Embedding* não pôde se obter um resultado aproximado sobre o quanto o conjunto inicial de dados foi reduzido, um resultado preciso poderia ser obtido se esse método fosse combinado com outro método de redução, dessa forma, sozinho, ele não se torna eficiente.

Agradecimentos

Agradecemos ao Departamento de Geografia da Universidade Estadual de Ponta Grossa – UEPG por ceder a base de dados.

Referências

BINGHAM E, MANNILA H. **Random projection in dimensionality reduction: applications to image and text data.** In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA, USA; 2001. p. 245–50.

BORG, I.; GROENEN,P.J.F. **Modern Multidimensional Scaling: Theory and Applications**. New York: Springer. 2ª Edição. 2005.

COX, T.; COX, M. **Multidimensional Scaling. Chapman & Hall, London**, 1994.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariada Statistical Analysis**. 6ª edição. Nova Jersey: Pearson/Prentice Hall, 2007. 794 p.

JOLLIFFE, I. T. **Principal Component Analysis**. 2ª edição. New York: Springer, 2002. 487 p.

KUMAR, V., STEINBACH, M., TAN, P-N. **Introdução ao Data Mining – Mineração de dados. Ciência Moderna**, 1ª Edição. 2009.

RIDDER DICK; DUIN, P. W. ROBERT. **Locally linear embedding for classification.** Pattern Recognition Group Technical Report Series, 2002.

SAUL, L. K., ROWEIS, S. T. **Think Globally, Fit Locally: Unsupervised Learning of Nonlinear Manifolds**. Journal of Machine Learning Research 4: 119-155, 2003.

VAN DER MAATEN L.; POSTMA J.; VAN DEN HERIK J. **Dimensionality Reduction: A Comparative Review**. Maastricht University, 2008.

VARINI, C.et al. **Visual Exploration of Pathology Images by a discrete Wavelet Transform preprocessed Locally Liner Embedding**. 2006.