XXIX Congresso de Iniciação Científica UNICAMP 2021

Ampliação de quantidade de padrões de matrizes classificados e reordenados pelo método Hybrid Sort de reordenação de matrizes

Matheus Alves da Silva, Thiago Gonçalves Mendes, Celmar Guimarães da Silva

Faculdade de Tecnologia, Universidade Estadual de Campinas

Resumo — Visualização de informação é uma área da informática que estuda técnicas que permitem melhorar o entendimento do usuário em problemas complexos envolvendo dados abstratos. Dentre diversas técnicas de representação visual, é de interesse para este projeto o conceito de matriz reordenável, estrutura de dados que tem como foco possibilitar a permutação de linhas e colunas de uma matriz de dados sem perder a integridade do seu conteúdo. Wilkinson e Behrisch et al. propõem nove possíveis tipos de padrões visuais de matrizes que podem evidenciar determinados comportamento nos dados. Um estudo do grupo de pesquisa teve como objetivo investigar um classificador baseado em Aprendizagem de Máquina capaz de identificar o padrão subjacente a uma matriz não reordenada com base em um vetor de características já identificado previamente na literatura. Durante este experimento, foram testadas diferentes técnicas de classificação com diferentes hiperparâmetros, dentre os quais se escolheu um classificador baseado na técnica de Random Forest. Entretanto, sua maior limitação é realizar estas classificações utilizando somente os padrões de Wilkinson – Simplex, Band, Circumplex, Equi e Block como base. O objetivo deste artigo é apresentar um classificador baseado em Aprendizagem de Máquina capaz de identificar não só os padrões propostos por Wilkinson mas também classificar os padrões propostos por Behrisch et al. - Bands, Block Diagonal, Lines/Stars e Off-Diagonal Block, utilizando as principais métricas de avaliação. Com bons resultados, o algoritmo obteve 98,6% de precisão na base de testes.

Index Terms— Aprendizado de Máquina, Matrizes Reordenáveis, Visualização de Informação.

1 INTRODUÇÃO

Visualização de informação (InfoVis) é uma área da informática que estuda técnicas para facilitar a visualização do usuário em um conjunto complexo e abstrato de dados, utilizando recursos gráficos e contribuindo na busca e identificação de padrões. Ela é muito utilizada como auxílio para as áreas da sociologia, biologia, arqueologia, antropologia, cartografia, entre outros [5].

Dentre diversas técnicas para representação visual de dados, é de interesse para este projeto o conceito de matriz reordenável. Trata-se de uma estrutura de dados que tem como foco possibilitar a permutação de linhas e colunas de uma matriz de dados sem perder a integridade do seu conteúdo, permitindo o reconhecimento de padrões não identificados anteriormente ao ser visualizada. A Figura 1 apresenta um exemplo de matriz antes e depois da reordenação.

A literatura de InfoVis define alguns tipos de padrões de matrizes, ou seja, matrizes com características específicas que evidenciam determinados comportamentos no conjunto de dados por ela mostrado. Wilkinson [9] e Behrisch et al [1] propõe nove diferentes padrões de matrizes capazes de incorporar novos *insights* nos dados de maneira visual.

Para estudos aprofundados sobre este tema, foi criada a ferramenta MRA (*Matrix Reordering Analyzer*) [7] cujo objetivo inicial era permitir realizar experimentos e comparar estaticamente algoritmos de reordenação de matrizes.

Em projetos anteriores do grupo de pesquisa, foi integrado à MRA, um classificador baseado em Aprendizagem de Máquina (*Machie Learning*) capaz de identificar o padrão de uma matriz não reordenada (ou seja, que não esteja propositadamente evidenciando algum padrão de matriz). O classificador elaborado foi integrado a um método preexistente de reordenação de matrizes (Hybrid Sort) [8] substituindo um classificador anterior que havia sido criado de maneira empírica. Sua principal limitação era realizar essa classificação com base somente nos padrões de Wilkinson – Simplex, Band, Circumplex, Equi-correlation (Equi) e Block.

Nesse contexto, este projeto teve por objetivo, implementar os padrões propostos por Behrisch (Figura 2) na ferramenta MRA investigação de um classificador baseado em Aprendizagem de Máquina (*Machine Learning*) capaz de identificar não só os padrões propostos por Wilkinson [9], mas que também seja capaz de identificar os padrões propostos por Behrisch [1].

Diferentes configurações de matrizes (tamanho, nível de ruído, variações de cada padrão) e de técnicas de Aprendizagem de Máquina foram utilizadas no processo, visando alcançar um resultado aceitável do ponto de vista científico, gerando um classificador cuja avaliação retornou melhores resultados que o classificador empírico anterior, e ainda capaz de identificar matrizes que possuem apenas ruído.

O restante deste texto está organizado da seguinte forma: a Seção 2 apresenta os materiais e métodos utilizado nesta pesquisa, abordando os padrões de matrizes com mais detalhes, a existência do padrão noise, a extração dos vetores de características das matrizes, a geração de amostras e o processo de modelagem; a Seção 3 apresenta os resultados finais obtidos neste projeto; a Seção 4 aborda

a discussão sobre a técnica de geração de amostras e os próximos passos; e a Seção 5 apresenta a conclusão do projeto.

2 MATERIAIS E MÉTODOS

Esta seção aborda os materiais e métodos necessários para construção do classificador.

2.1 PADRÕES DE MATRIZES

Wilkinson [9] e Behrisch [1] propõem nove padrões subjacentes a matrizes com objetivo de revelar potenciais *insights* no conjunto de dados de maneira inovadora. Estes padrões possuem características individuais e podem ser aplicados em diferentes tipos de problemas Behrisch et al. [1]:

Padrão Band: Padrão em formato de banda que possui valores mais altos em torno da diagonal principal da matriz. Esses valores diminuem quanto mais estiverem perto do canto superior direito ou canto inferior esquerdo da matriz.

Padrão Bands: Padrão de banda que consiste em linhas paralelas à diagonal da matriz. Uma banda indica um caminho, que representa uma sequência um nó para outro. Este padrão pode ser encontrado, por exemplo, nos caminhos possíveis de transmissão de informações em uma rede social ou uma sequência de reações em redes biológicas.

Padrão Block: O padrão Block divide o conjunto de colunas em k-grupos, e o conjunto de linhas em grupos 2k. Esses grupos definem uma supermatriz com k × 2k supercélulas. Os valores de uma supercélula podem variar entre 0 ou 1.

Padrão Block Diagonal: Padrão em formato de bloco na diagonal da matriz (pelo menos 2 x 2 células). Este bloco na diagonal se refere a uma forte conexão entre os nós do bloco. Este padrão pode ser encontrado, por exemplo, em redes sociais representando grupos de amigos em comum.

Padrão Circumplex: Padrão semelhante ao padrão de banda, mas que pode representar uns relacionamentos fortes que acontece de forma circular.

Padrão Equi: Padrão com valores que aumentam na direção de cima para baixo, tendo os valores em cada linha iguais (exceto quando há presença de ruído).

Padrão Lines/Stars: Padrão em formato de estrela na matriz, em uma linha horizontal e uma vertical. No entanto, uma linha não precisa abranger toda a matriz. Um padrão em formato de estrela representa um nó altamente conectado em uma rede. Este padrão pode ser encontrado, por exemplo, nas relações de uma pessoa famosa em uma rede de amizades.

Padrão Off-Diagonal Block: Padrão em formato de blocos nos cantos da matriz, sem tocar a diagonal principal (canto superior esquerdo e canto direito inferior seguindo a definição de Behrisch et Al. [1]. Este padrão pode ser encontrado, em redes de 2 modos, por exemplo, uma relação entre pessoas e cidades que já visitaram.

Padrão Simplex: Padrão que tende a aumentar a partir do canto inferior esquerdo da matriz até o superior direito. A diagonal principal divide o padrão em duas partes: a região inferior esquerda, com os valores mais baixos da matriz, e a superior direita, com os valores mais altos.

Os padrões de Behrisch foram integrados na ferramenta MRA juntamente com os padrões de Wilkinson, conforme apresentados na Figura 3.

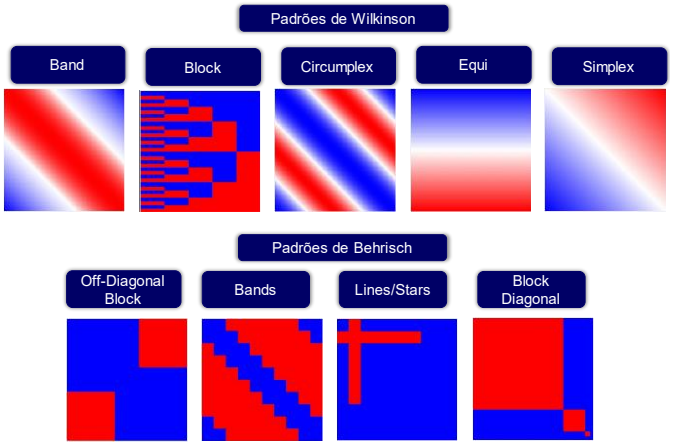

Figura 3 – Exemplo de padrões de matrizes propostos por Wilkinson e Behrisch integrados na ferramenta MRA.

Além dos nove padrões de Wilkinson e Behrisch, também foi adicionado o padrão noise (já estudado anteriormente pelo grupo de pesquisa) para a lista de classes do classificador. Este padrão representa matrizes acima de 90% de ruído, e foi construído com objetivo de evitar resultados enviesados.

2.2 EXTRAÇÃO DOS VETORES DE CARACTERÍSTICAS

Todas as amostras utilizadas para construção do classificador foram compostas por quatro pares de características extraídos previamente das matrizes. A matriz inicialmente é normalizada em valores de 0 e 1 e posteriormente é realizado o cálculo de quatro conjuntos de vetores:

- **maxRows:** valores máximos de cada linha;
- **minRows:** valores mínimos de cada linha;
- **maxColumns:** valores máximos de cada coluna;
- **minColumns:** valores mínimos de cada coluna.

Com estes recursos, geramos uma regressão linear e extraímos todos os coeficientes angulares (a) e lineares (b) de cada conjunto, resultando nos seguintes vetores de características: [aminRows, amaxRows, aminColumns, amaxColumns, bminRows, bmaxRows, bminColumns, bmaxColumns].

Além dessas características, também verificamos a necessidade de incorporar ao vetor de características novas medidas relacionadas aos padrões de Behrisch et al., listadas a seguir:

- **isStars(M):** Calcula a divisão do número de linhas com mais da metade das células acima de 0.5 pelo total de linhas.
- **avgDeviationOfQtyHighValues(M):** Calcula a quantidade de valores maiores que 0.5 na linha i de T, onde T é a matriz triangular superior de M.
- **lowValues(M):** Calcula a quantidade relativa de valores baixos – menores que 0.5.

Estas medidas permitem identificar com melhor precisão os padrões propostos por Behrisch et al [1].

2.3 GERAÇÃO DE AMOSTRAS

Para construção do modelo, selecionamos as amostras utilizando a técnica estatística de sobreamostragem (*oversampling*) para o

balanceamento das classes. O conjunto total de amostras utilizadas foram de 57.600 amostras para base de treino e 14.400 amostras para a base de testes (representando aproximadamente 25% da quantidade de amostras totais utilizadas na base de treino). Para cada classe, foram geradas 6.400 amostras por padrão na base de treino e 1.600 amostras por padrão na base de teste.

2.4 Processo de Modelagem

Após a geração das amostras, utilizamos a ferramenta WEKA [3] para construção do modelo. Também utilizamos uma ramificação da ferramenta, chamada de Auto-WEKA [4] que permite encontrar a melhor técnica de classificação para nosso conjunto de dados.

O WEKA considera os seguintes candidatos para classificadores (o número entre parênteses indica o número de possibilidades para hiperparâmetros):

- **Learners:** BayesNet (2), DecisionStump (0), DecisionTable (4), GaussianProcesses (10), IBk (5), J48 (9), JRip (4), KStar (3), LinearRegression (3), LMT (9), Logistic (1), M5P (4), M5Rules (4), MultilayerPerceptron (8), NaiveBayes (2), NaiveBayesMultinomial (0), OneR (1), PART (4), RandomForest (7), RandomTree (11), REPTree (6), SGD (5), SimpleLinearRegression (0), SimpleLogistic (5), SMO (11), SMOreg (13), VotedPerceptron (3), ZeroR (0);
- **Ensemble Methods**: Stacking (2), Vote (2);
- **Meta-Methods:** LWL (5), AdaBoostM1 (6), AdditiveRegression (4), AttributeSelectedClassifier (2), Bagging (4), RandomCommittee (2), RandomSubSpace (3);
- **Attribute Selection Methods**: BestFirst (2), GreedyStepwise (4).

Os resultados foram medidos utilizando as seguintes métricas de avaliação de classificação encontradas na literatura:

Matriz de Confusão: Tabela que apresenta os erros e acertos do modelo em comparação ao resultado esperado.

Acurácia: Representa a proporção de acertos do modelo. Calcula-se utilizando o número total de observações que o modelo acertou e divide-se pelo número total de amostras previstas.

Recall: Para uma dada classe A, indica quantas amostras são realmente A, dentre as quais foram classificadas como A.

Precisão: Representa dentre todas as situações de classe A como valor esperado, quantas foram classificadas como A.

F-Measure: Representa média harmônica entre precisão e recall. Seu objetivo é trazer um número único que determine a qualidade do modelo.

3 Resultados

Esta seção aborda os resultados obtidos do projeto.

3.1 Métricas de Avaliação

A Tabela 1 apresenta os resultados das métricas de avaliação do classificador e a Tabela 2 apresenta sua matriz de confusão. Ambos os resultados estão relacionados a base de testes.

	Precisão	Recall	F-Measure
BANDS	0,956	0,965	0,960
BLOCKDIAGONAL	1,000	1,000	1,000
LINESTARS	0,957	0,951	0,954
OFFDIAGONALBLOCK	0,982	0,980	0,981
BAND	0,997	0,996	0,997
BLOCK	1,000	1,000	1,000
CIRCUMPLEX	0,999	0,989	0,994
EQUI	0,985	0,995	0,990
SIMPLEX	0,995	0,996	0,995

Tabela 1: Resultados das métricas de avaliação do classificador em relação a base de testes.

	BANDS	BLOCKDIAGONAL	LINESTARS	OFFDIAGONALBLOCK	BAND	BLOCK	CIRCUMPLEX	EQUI	SIMPLEX
BANDS	1544	0	56	0	0	0	0	0	0
BLOCKDIAGONAL	0	1600	0	0	0	0	0	0	0
LINESTARS	51	0	1521	28	0	0	0	0	0
OFFDIAGONALBLOCK	20	0	12	1568	0	0	0	0	0
BAND	0	0	0	0	1594	0	1	5	0
BLOCK	0	0	0	0	0	1600	0	0	0
CIRCUMPLEX	0	0	0	0	5	0	1583	12	0
EQUI	0	0	0	0	0	0	0	1592	8
SIMPLEX	0	0	0	0	0	0	0	7	1593

Tabela 2: Matriz de confusão do classificador em relação a base de testes.

4 Discussão

Observa-se nesses resultados que a classificação ocorreu de forma apropriada, com poucos casos de confusão do classificador, como apontado pela matriz de confusão da Tabela 1.

Adicionalmente, nota-se que os valores de Precisão, Recall e F-Measure ficaram muito elevados. A média dos valores de precisão é de 98,5%, indicando que a abordagem adotada é apropriada para uma classificação correta desse tipo de matrizes, de forma independente de sua ordenação.

Estudos posteriores podem considerar outros tamanhos de matrizes, bem como outras variações de parâmetros de ruídos, para averiguar um possível overfitting da solução.

4.1 Próximos Passos

4.1.1 Implementação do classificador em multi-linguagens

Atualmente o grupo de pesquisa tem se dedicado à um projeto que visa a construção desta versão do classificador na linguagem de programação JavaScript.

5 Conclusão

Apresentamos um classificador de matrizes capaz de identificar padrões de Wilkinson e Behrisch com 98,6% de precisão.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) que financiou este projeto.

Referências

[1] BEHRISCH, M. BACH, B., HENRY-RICHE, N., SCHRECK, T., FEKETE, J.-D. Matrix Reordering Methods for Table and Network Visualization. Computer Graphics Forum 35 (3), 2016, pp. 693-716.

[2] BROWNE, M.W. Circumplex Models for Correlation Matrices. Psychometrika 57(4), 1992, pp. 469-497.

[3] HALL, Mark et al. The WEKA data mining software. ACM SIGKDD Explorations Newsletter, v. 11, n. 1, p. 10–18, 2009.

[4] KOTTHOFF, Lars et al. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research, v. 18, p. 1–5, 2017.

[5] LIIV, I. Seriation and Matrix Reordering: An Historical Overview. Wiley InterScience, 2010.

[6] MEDINA, B.F. Reordenação de matrizes de dados quantitativos usando árvores PQR. 2015. 77 f. Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Tecnologia, Limeira, SP.

[7] SILVA, C. G., MELO, M. F., SILVA, F. P., MEIDANIS, J. PQR Sort - Using PQR trees for binary matrix for binary matrix reorganization. Journal of the Brazilian Computer Society. 2013. 10.1186/1678-4804-20-3.

[8] SILVA, C. G.. Hybrid Sort - A pattern-focused matrix reordering approach based on classification. In: 14th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing, 2020, (online). Proceedings of the International Conferences on Computer Graphics, Visualization, Computer Vision and Image Processing 2020, Big Data Analytics, Data Mining and Computational Intelligence 2020, and Theory and Practice in Modern Computing 2020, 2020. p. 35-43.

[9] WILKINSON, L. The Grammar of Graphics. 2. ed. [S. l.]: Springer, 2005. 693 p.