

UNIVERSIDADE FEDERAL DE MINAS GERAIS



PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA
ELÉTRICA

Novos métodos nebulosos de seleção de características

Autor:

Vinícius C. Ferraz

today

1 Introdução

Sistemas de predição, regressão ou classificação costumam ser divididos nas seguintes etapas: extração dos dados; **seleção das variáveis** de entrada ou características; divisão dos dados em treino, teste e validação; ajuste dos coeficientes do modelo de acordo com dados reais treino e teste; validação ou implantação, em que a saída estimada não poderá imediatamente ser comparada com a desejada. Queremos otimizar a seleção de características, de forma a melhorar a acurácia, a incerteza e o custo computacional do classificador/regressor, sem *overfitting* nem *underfitting*. Pearson [1901] introduziu o conceito de correlações entre cada entrada e saída, das quais podemos excluir as de módulo baixo. Há também casos em que se duas entradas estão altamente correlacionadas, o desempenho aumenta ao excluir uma delas. Os algoritmos de seleção de características (FSA) atuais são divididos em três grandes áreas: *filter* — e.g., chi square (Liu and Seti-ono [1995]) —, *wrapper* — e.g., (Akaike's) sequential feature selection; recursive feature elimination (RFE) — e *embedded* — e.g., lasso (Chong and Jun [2005]), elastic net e ridge regression.

Suponha que cada variável seja uma coluna de uma matriz e que cada amostra seja uma linha da mesma. Suponha ainda que temos as colunas de entrada x_i , $1 \leq i \leq n$ e uma coluna de saída y .

No método dos Mínimos Quadrados Recursivo (Aguirre [2007]), os pesos são os regressores, provindos da regressão linear $X\theta = Y \Rightarrow \theta = (X^\top X)^{-1}X^\top Y$. Os algoritmos do tipo *wrapper* são os mais lentos, por adicionarem ou excluírem colunas uma a uma. O critério de Akaike (Aguirre [2007]) consiste em começar por todas as n variáveis com n candidatas a retirada, dentre as quais é selecionado um modelo com $n - 1$ variáveis, por exemplo, por erro mínimo. Esse processo é repetido para gerar modelos de $n - 2, \dots, 2, 1$ variáveis, cada uma com seu Akaike Information Criterion (AIC). A inclinação do gráfico $n \times AIC$ fica eventualmente maior ou igual que zero, indicando que acrescentar variáveis não é interessante. Temos, assim, na saída do FSA um subconjunto das n entradas.

Muitos dos FSAs do tipo *filter* utilizam o conceito de *score* e *ranking*. A cada entrada, associamos um peso w_i não negativo e de soma unitária: $\sum_i w_i = 1 = 100\%$. Por meio de ordenação decrescente, os maiores pesos são considerados, e as variáveis de menor contribuição percentual são descartadas. O amplamente utilizado PCA (principal component analysis), que foi primeiramente estudado por Jolliffe [2002], altera os valores originais

das entradas, por isso está em uma quarta grande área: feature extraction algorithms.

Li et al. [2018] divide os algoritmos em sete áreas: information theoretical based — e.g., FCBF (Yu and Liu [2003]); similarity based — e.g., fisher score (He et al. [2005] e Duda et al. [2012]); sparse learning based — e.g., ls_l21 (Liu et al. [2012]); statistical based — e.g., f score (Wright [1965]); streaming — e.g., alpha investing (Zhou et al. [2005]); structure — e.g., group fs e tree fs; e wrapper — e.g. decision tree backward e forward (Guyon and Elisseeff [2003]).

Yamakawa [1992] propôs um modelo YNFN (neurofuzzy network) que dá pesos a cada variável por função de pertinência. Normalizando a média de seus módulos, conseguimos um score e um ranking das variáveis que estão sendo treinadas. Esse processo pode facilmente ser convertido para evolutivo (Angelov and Yager [2011]), em que não é necessário armazenar nenhum dado além do atual. De forma análoga ao YNFN, existem modelos de regressores que podem facilmente ser convertidos em FSAs, tais como os granulares: Leite [2012] exhibe os algoritmos FBeM e eGNN. Leite et al. [2019] exibem o algoritmo eOGS e Garcia et al. [2019] exibem o algoritmo eFGP.

Com alguns algoritmos, é-nos possível fazer um subscore por amostra (em alguns casos duas a duas linhas — para haver uma partição em dois conjuntos de interseção vazia; no FCBF, o mínimo encontrado foi de sete em sete registros, a fim de não haver divisão por entropia igual a zero), e tomar por score a soma normalizada dos subscores. Uma desvantagem seria desconsiderar relações entre as tuplas.

1.1 Relevância do Tema

Zhao and Liu [2011] lista aplicações relevantes em muitas áreas, como por exemplo baixo número de linhas ou baixo número de colunas; e no processamento de dados em *streaming* ou big data (dados em larga escala, como por exemplo algumas tabelas do banco de dados do Google). Há casos em que os cálculos em batelada tornam-se inviáveis. Pretendemos contribuir, por exemplo, no Projeto Genoma, em formas melhores de se fazer predição das contribuições de cada gene em um fenótipo, ou doença.

Vapnik [1999] apresenta o método Support Vector Machine (SVM), que é tão utilizado quanto o PCA. Quando temos muitas variáveis, seja da ordem de 700 a 10^7 (vide Zhao and Liu [2011]), muitos FSAs — como exemplo, os métodos SVM forward e SVM backward — devem ser descartados por incluírem as variáveis uma a uma, ou excluírem-nas lentamente.

2 Objetivos

2.1 Gerais

- Mostrar que YNFN-FSA é comparável a outros FSAs — lineares ou não — tanto em batelada quanto evolutivamente, em critérios tais quais acurácia, incerteza, quantidade ótima de colunas e complexidade computacional.
- Comparar os scores do YNFN com outros diversos selecionados. Em cada algoritmo, as variáveis contribuem de uma forma particular, utilizando problemas de otimização distintos.
- Com o solver certo, o problema de otimização certo, seja evolucionário, seja determinístico, configurará uma melhoria.

2.2 Específicos

- Comparar os métodos de otimização propostos por [Liu et al. \[2009\]](#) entre si e com o YNFN-FSA.
- Propor adaptações evolutivas a FSAs em batelada. Verificar se tais propostas são comparáveis ao YNFN.
- Aplicar teoria granular na minimização dos erros. Cada algoritmo de classificação ou regressão terá uma proposta de conversão para FSA.
- Adicionar incerteza ao método YNFN com conjuntos fuzzy tipo 2, e também aos dados de entrada.
- Granular o espaço das características, i.e., granular a contribuição das características com o tempo/linhas.

3 Revisão da Literatura

[Dash and Liu \[1997\]](#) apresentam um FSA evolucionário. [Gauchi and Chagnon \[2001\]](#) comparam 20 métodos de seleção baseados em diferentes critérios de avaliação, incluindo ajuste do modelo e capacidade de predição. Dentre os métodos, destaca-se o algoritmo

genético. [Stein et al. \[2013\]](#) utilizaram stepwise, forward, backward e propuseram o método de regressão/AIC. [Golub and Van Loan \[2013\]](#) apresentam o cálculo de ERR, análogo ao AIC, mas baseado em ortogonalidade e decomposição QR, i.e., o algoritmo de Gram-Schmidt.

[Jang et al. \[1997\]](#) explica pedagogicamente o método NFN e os fundamentos de conjuntos nebulosos tipo 2, os quais são intervalares em [Wu \[2012\]](#). [Silva et al. \[2012\]](#) encontram o α ótimo, que é o multiplicador a cada iteração do método do gradiente. Além disso, propõem o modelo eNFN-AFS, que é evolutivo com seleção de variáveis. [Chong and Jun \[2005\]](#) apresentam os métodos PLS VIP e PLS Beta de seleção de variáveis. [Zhao and Liu \[2011\]](#) apresentam vários FSAs, 35 deles implementados por [Li et al. \[2018\]](#), em que se destacaram: ls.l21, low variance ([Pedregosa et al. \[2011\]](#)), fisher score, f score, chi square, gini index ([Brown et al. \[2012\]](#); [Breiman \[1984\]](#); [Peng et al. \[2005\]](#); e [Deng and Runger \[2013\]](#)) e FCBF.

4 Metodologia

Pretendemos responder às questões: qual é a melhor forma de selecionar variáveis através de pesos? Quais são as vantagens e desvantagens de se utilizar métodos nebulosos, como o YNFN, encapsulado em um FSA?

Utilizaremos Matlab e Python na implementação. Começaremos pelo método de [Pearson \[1901\]](#).

Utilizando intervalos $[0.95x_i, 1.05x_i]$ e análise intervalar ([Moore et al. \[2009\]](#)), compararemos as incertezas propagadas.

Estudaremos as constantes ou hiperparâmetros, como por exemplo, após a ordem decrescente, costuma-se calcular a soma acumulada das contribuições, e descartar do ranking a primeira vez que atingir o limite $c = 0.90$.

Selecionaremos as bases de dados; serão feitas comparações primeiro matemáticas, depois computacionais.

5 Cronograma

O projeto de mestrado será realizado em um período de três semestres através das atividades descritas na Tabela 1.

Período	Atividades
Semestre 2022/2	* Cursadas as disciplinas restantes.
Semestre 2023/1	* Revisar literatura; * Pesquisar algoritmos; * Obter os primeiros resultados.
Semestre 2023/2	* Implementar mais algoritmos; * Preparar artigo para congresso.

Tabela 1: Cronograma de tarefas

Referências

- [1] L. Aguirre. *Introdução à Identificação de Sistemas – Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. Editora UFMG, 2007. ISBN 9788570415844. URL <https://books.google.com.br/books?id=f9IwE7PhOfYC>.
- [2] P. Angelov and R. Yager. Simplified fuzzy rule-based systems using non-parametric antecedents and relative data density. In *2011 IEEE Workshop on Evolving and Adaptive Intelligent Systems (EAIS)*, pages 62–69. IEEE, 2011.
- [3] L. Breiman. *Classification and Regression Trees*. Wadsworth Statistics/Probability series. Chapman & Hall, 1984. URL <https://books.google.com.br/books?id=gCOZZwEACAAJ>.
- [4] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13:27–66, 2012.
- [5] I.-G. Chong and C.-H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, 78(1-2):103–112, 2005.
- [6] M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.
- [7] H. Deng and G. Runger. Gene selection with guided regularized random forest. *Pattern recognition*, 46(12):3483–3489, 2013.

- [8] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2012. ISBN 9781118586006. URL <https://books.google.com.br/books?id=Br33IRC3PkQC>.
- [9] C. Garcia, D. Leite, and I. Škrjanc. Incremental missing-data imputation for evolving fuzzy granular prediction. *IEEE transactions on fuzzy systems*, 28(10):2348–2362, 2019.
- [10] J.-P. Gauchi and P. Chagnon. Comparison of selection methods of explanatory variables in pls regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, 58(2):171–193, 2001.
- [11] G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.
- [12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
- [13] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *Advances in neural information processing systems*, 18, 2005.
- [14] J. Jang, C. Sun, and E. Mizutani. *Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. MATLAB curriculum series. Prentice Hall, 1997. ISBN 9780132610667. URL <https://books.google.com.br/books?id=vN5QAAAAMAAJ>.
- [15] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002. ISBN 9780387954424. URL <https://books.google.com.br/books?id=TtVF-ao4fI8C>.
- [16] D. Leite. *Evolving granular systems*. PhD thesis, 2012.
- [17] D. Leite, F. Gomide, and I. Škrjanc. Multiobjective optimization of fully autonomous evolving fuzzy granular models. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2019.
- [18] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2018. URL <https://github.com/jundongli/scikit-feature>.

- [19] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE international conference on tools with artificial intelligence*, pages 388–391. IEEE, 1995.
- [20] J. Liu, S. Ji, J. Ye, et al. Slep: Sparse learning with efficient projections. *Arizona State University*, 6(491):7, 2009. URL <http://www.public.asu.edu/~jye02/Software/SLEP>.
- [21] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv preprint arXiv:1205.2631*, 2012.
- [22] R. Moore, R. Kearfott, and M. Cloud. *Introduction to Interval Analysis*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2009. ISBN 9780898717716. URL <https://books.google.com.br/books?id=tT7yKbqfEwC>.
- [23] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901.
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [25] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.
- [26] A. M. Silva, W. M. Caminhas, A. P. Lemos, and F. Gomide. Evolving neural fuzzy network with adaptive feature selection. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 440–445. IEEE, 2012.
- [27] M. Stein, M. J. Anzanello, V. L. Cervo, and A. Kahmann. Seleção das variáveis de processo mais relevantes para predição dos níveis de sucata em um processo do setor metal-mecânico. *Semana de Engenharia de Produção Sul Americana (13.: 2013 jun. 9-11: Gramado, RS). Anais..[recurso eletrônico]. Porto Alegre: FEEng, 2013.*, 2013.
- [28] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

- [29] S. Wright. The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution*, pages 395–420, 1965.
- [30] D. Wu. Approaches for reducing the computational cost of interval type-2 fuzzy logic systems: overview and comparisons. *IEEE Transactions on Fuzzy Systems*, 21(1): 80–99, 2012.
- [31] T. Yamakawa. A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. In *Proc. of the 2nd Int. Conf. on Fuzzy Logic & Neural Networks*, pages 477–483, 1992.
- [32] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.
- [33] Z. Zhao and H. Liu. *Spectral Feature Selection for Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2011. ISBN 9781439862100. URL <https://books.google.com.br/books?id=VfDRBQAAQBAJ>.
- [34] J. Zhou, D. Foster, R. Stine, and L. Ungar. Streaming feature selection using alpha-investing. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 384–393, 2005.