Analiza mikromacierzy DNA w predykcji występowania przerzutów raka piersi

Stanisław Wilczyński

Promotor: dr. Jan Chorowski

Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Informatyki

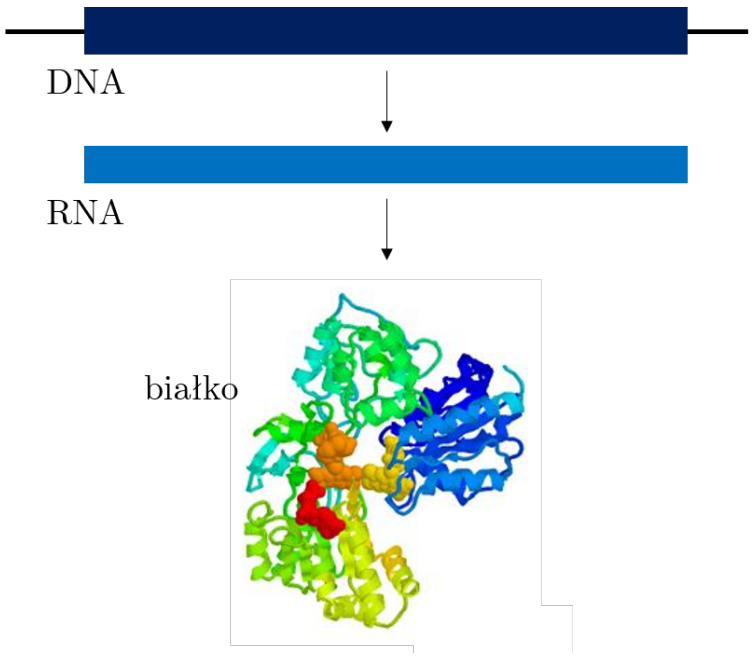
20.09.2019

Wstęp

- Mnóstwo prac poświęconych klasyfikacji guzów i predykcji rozwoju choroby
- **Cel:** redukcja liczby pacjentów wysłanych niepotrzebnie na chemioterapię
- Mikromacierze jako innowacyjna metoda badań medycznych dostarczająca ogromne ilości danych
- Interesujący, mniej popularny obszar klasyfikacji ($n \ll p$)
- Analiza i porównanie nowatorskich oraz standardowych metod używanych w literaturze
- Kontynuacja badań z pracy licencjackiej

Biologiczny kontekst

- DNA złożone z genów ulega ekspresji tworząc białka
- Poziom ekspresji - ilość produktu
- Mikromacierze DNA - zróżnicowane technologie mierzenia poziomu ekspresji
- Guzy i przerzuty
- Użyte dane - połączone 10 zbiorów z Gene Expression Omnibus (969×12179)



Rysunek: Schemat ekspresji genów

Wstępna analiza danych (EDA)

- Badanie podstawowych statystyk (skośność + kurtoza)
- Analiza rozkładów poszczególnych zmiennych
- Wpływ różnych rodzajów normalizacji
- Studium korelacji

Wstępna analiza danych (EDA)

- Badanie podstawowych statystyk (skośność + kurtoza)
- Analiza rozkładów poszczególnych zmiennych
- Wpływ różnych rodzajów normalizacji
- Studium korelacji

Wniosek

Wystarczająca zgodność z rozkładem normalnym, brak potrzeby dalszego preprocessingu

Wykorzystane metody

Regularyzowane klasyfikatory:

- RF
- LR
- RDA
- NSC
- SVC
- **RLR**

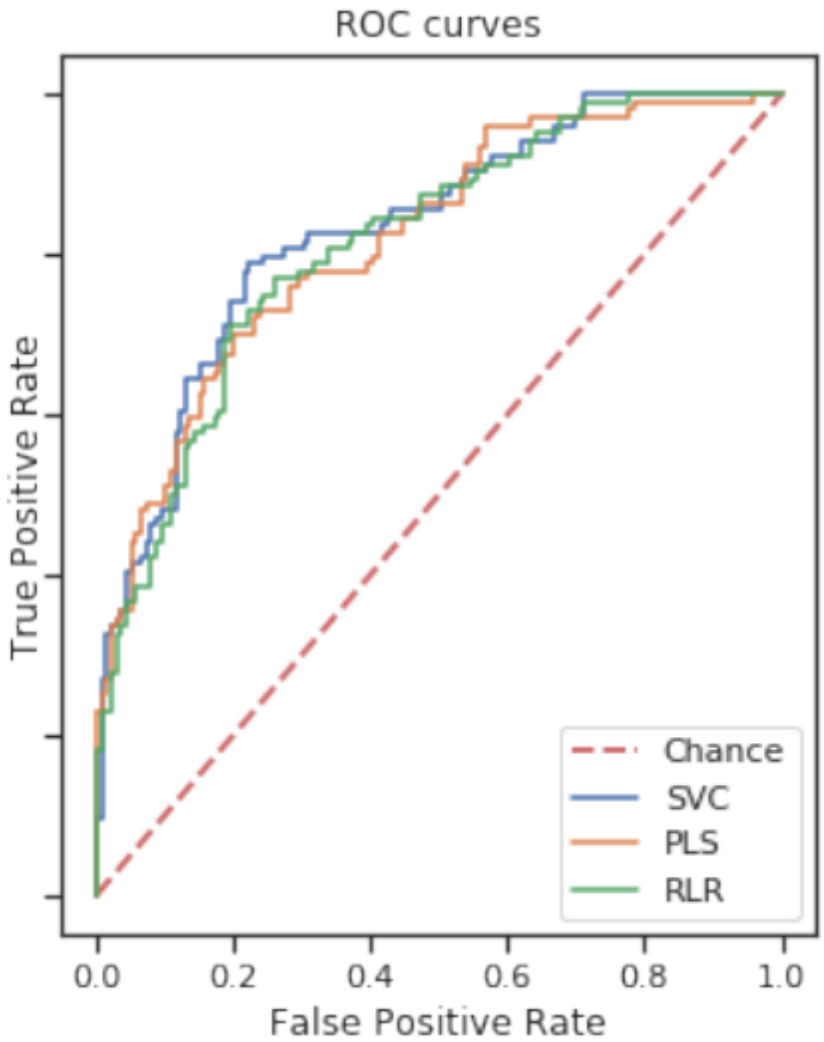
Ekstrakcja zmiennych:

- Bez nadzoru
  - PCA
  - SPCA
  - **MLCC**
- Z nadzorem
  - PLS
  - FDNN

Selekcja zmiennych:

- SAM
- RFE

Modele wybrane za pomocą NCV



Rysunek: Krzywe ROC dla PLS, SVC i RLR

Analiza wyników

- Recall - $\frac{TP}{TP+FN}$
- Precision - $\frac{TP}{TP+FP}$
- ROC AUC - wartość recall uśredniona po progach FPR

	ROC AUC	Prec (0.95)	Prec (0.8)	Acc (0.9)
PLS	0.818	0.556	0.596	0.650
SVC	0.831	0.515	0.689	
RLR	0.815	0.527	0.641	
PCA-AE-Ada[1]	0.714			
Protein Net[1]				0.558

Tabela: Metryki dla wybranych metod, (wartość recall)

[1] Dane literaturowe

Analiza wyników

- Recall - $\frac{TP}{TP+FN}$
- Precision - $\frac{TP}{TP+FP}$
- ROC AUC - wartość recall uśredniona po progach FPR

	ROC AUC	Prec (0.95)	Prec (0.8)	Acc (0.9)
PLS	0.818	0.556	0.596	0.650
SVC	0.831	0.515	0.689	
RLR	0.815	0.527	0.641	
PCA-AE-Ada[1]	0.714			
Protein Net[1]				0.558

Tabela: Metryki dla wybranych metod, (wartość recall)

[1] Dane literaturowe

Przewaga selekcji nad ekstrakcją

- Środowisko medyczne
  - Brak akceptacji *czarnych skrzynek*, jasne uzasadnienie diagnozy
- Dodatkowe atuty
  - Wskazanie konkretnych biomarkerów
  - Odkrycie zależności pomiędzy genami, a nie binarna diagnoza

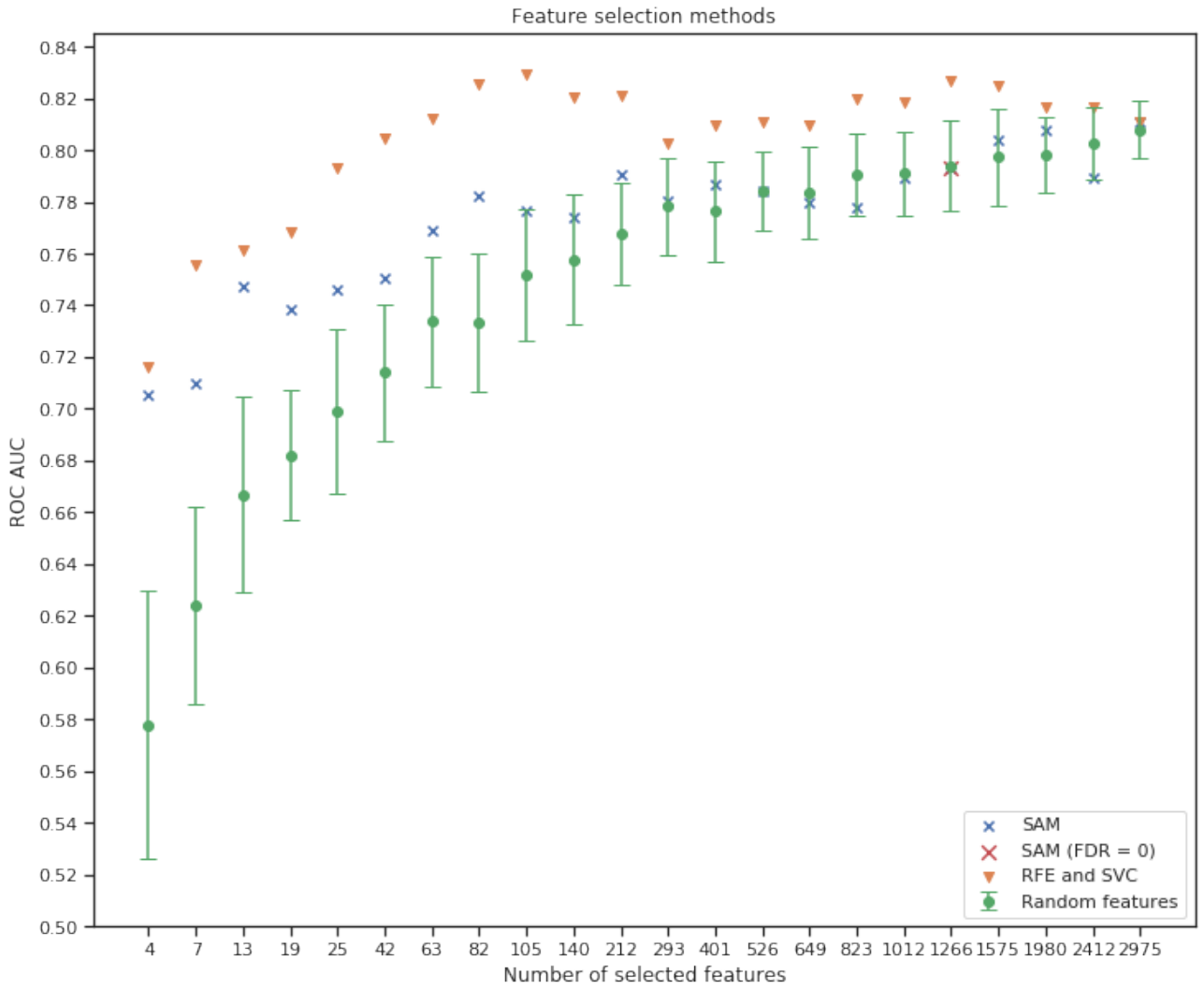
Przewaga selekcji nad ekstrakcją

- Środowisko medyczne
  - Brak akceptacji *czarnych skrzynek*, jasne uzasadnienie diagnozy
- Dodatkowe atuty
  - Wskazanie konkretnych biomarkerów
  - Odkrycie zależności pomiędzy genami, a nie binarna diagnoza

Wniosek

Pomimo mniejszych wartości metryk klasyfikacji, selekcja może być bardziej odpowiednia dla zastosowań medycznych

Selekcja zmiennych



Rysunek: ROC AUC dla metod selekcji zmiennych

Wnioski

- Wybrana metoda - RFE z użyciem SVC (105 zmiennych)
  - Wysokie ROC AUC
  - Zalety selekcji przeważają
  - **Wada:** brak zgodności w wybieranych zmiennych $\Rightarrow$ konieczność sprawdzenia przez specjalistę
- Konieczność współpracy statystyków i biologów:
  - Różnorodność (małych) zbiorów danych i wyników
  - Brak dostępnego kodu źródłowego, wyniki niereprodukowalne
  - Użycie tylko prostych modeli
  - Brak wsparcia profesjonalnych statystyków
  - Ewentualne uwzględnienie dodatkowych biomarkerów
  - Weryfikacja wyników przez biologów
- Narzędzie wspomagające diagnozę

Bibliografia I

Charu C. Aggarwal.

Data classification : algorithms and applications.
Chapman and Hall/CRC, 1st edition, 2014.

Christophe Ambroise and Geoffrey J McLachlan.

Selection bias in gene extraction on the basis of microarray gene-expression data.
Proceedings of the National Academy of Sciences of the United States of America, 99(10):6562–6, may 2002.

Sako Arts.

Dimensionality Reduction of Gene Expression Data.
Mater's thesis, Eindhoven Univeristy of Technology, 2018.

Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva.
NCBI GEO: archive for functional genomics data sets—update.
Nucleic Acids Research, 41(D1):D991–D995, nov 2012.

Michaela Bayerlová, Kerstin Menck, Florian Klemm, Alexander Wolff, Tobias Pukrop, Claudia Binder, Tim Beißbarth, and Annalen Bleckmann.
Ror2 Signaling and Its Relevance in Breast Cancer Progression.
Frontiers in Oncology, 7:135, jun 2017.

James Bergstra and Yoshua Bengio.

Random Search for Hyper-Parameter Optimization.
Journal of Machine Learning Research, 13(Feb):281–305, 2012.

Bibliografia II

B.M. M Bolstad, R.A A Irizarry, M. Astrand, and T.P. P Speed.

A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias.
Bioinformatics, 19(2):185–193, jan 2003.

A.-L. Boulesteix and K. Strimmer.

Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.
Briefings in Bioinformatics, 8(1):32–44, may 2006.

Leo Breiman.

Random Forests.
Machine Learning, 45(1):5–32, 2001.

Terence A. Brown.

Gene cloning and DNA analysis : an introduction.
Wiley-Blackwell, 2016.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, and Gaël Varoquaux.

API design for machine learning software: experiences from the scikit-learn project.
sep 2013.

Gavin C. Cawley and Nicola L. C. Talbot.

On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation.
Journal of Machine Learning Research, 11(Jul):2079–2107, 2010.

J. J. Chen, C.-A. Tsai, H. Moon, H. Ahn, J. J. Young, and C.-H. Chen.

Decision threshold adjustment in class prediction.
SAR and QSAR in Environmental Research, 17(3):337–352, jun 2006.

Bibliografia III

Han Yu Chuang, Eunjung Lee, Yu Tsueng Liu, Doheon Lee, and Trey Ideker.

Network-based classification of breast cancer metastasis.

Molecular Systems Biology, 3(140):1–10, 2007.

Emily Clough and Tanya Barrett.

The Gene Expression Omnibus Database.

Methods in molecular biology (Clifton, N.J.), 1418:93–110, 2016.

Aaron Defazio, Francis R Bach, and Simon Lacoste-Julien.

SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives.

CoRR, abs/1407.0, jul 2014.

Sandrine Dudoit, Jane Fridlyand, and Terence P Speed.

Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.

Journal of the American Statistical Association, 97(457):77–87, mar 2002.

Liat Ein-Dor, Or Zuk, and Eytan Domany.

Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.

Proceedings of the National Academy of Sciences of the United States of America, 103(15):5923–8, apr 2006.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin.

LIBLINEAR: A Library for Large Linear Classification.

Journal of Machine Learning Research, 9(Aug):1871–1874, 2008.

Jerome H. Friedman.

Regularized Discriminant Analysis.

Journal of the American Statistical Association, 84(405):165, mar 1989.

Bibliografia IV


M. Gaasenbeek, E. S. Lander, D. K. Slonim, M. L. Loh, P. Tamayo, C. Huard, H. Coller, J. P. Mesirov, T. R. Golub, C. D. Bloomfield, J. R. Downing, and M. A. Caligiuri.
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.
Science, 286(5439):531–537, 1999.


Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.


Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik.
Gene Selection for Cancer Classification using Support Vector Machines.
Machine Learning, 46(1/3):389–422, 2002.


Xiyi Hang and Fang-Xiang Wu.
Sparse Representation for Classification of Tumors Using Gene Expression Data.
Journal of Biomedicine and Biotechnology, 2009:1–6, 2009.


Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning.
Springer Series in Statistics. Springer New York, New York, NY, 2009.


Zena M. Hira and Duncan F. Gillies.
A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.
Advances in Bioinformatics, 2015(1):1–13, 2015.


John D. Hunter.
Matplotlib: A 2D Graphics Environment.
Computing in Science & Engineering, 9(3):90–95, 2007.

Bibliografia V

R. A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed.
Exploration, normalization, and summaries of high density oligonucleotide array probe level data.
Biostatistics, 4(2):249–264, apr 2003.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: With Applications in R.
Springer Publishing Company, Incorporated, 2014.

I. T. Jolliffe.
Principal component analysis.
Springer, 2002.

Yunchuan Kong and Tianwei Yu.
A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification.
Scientific Reports, 8(1):16477, dec 2018.

Olivier Ledoit and Michael Wolf.
Honey, I Shrunk the Sample Covariance Matrix.
The Journal of Portfolio Management, 30(4):110–119, jul 2004.

George Lee, Carlos Rodriguez, and Anant Madabhushi.
An Empirical Comparison of Dimensionality Reduction Methods for Classifying Gene and Protein Expression Datasets.
In *Bioinformatics Research and Applications*, pages 170–181. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Bibliografia VI

Ying Lu and Jiawei Han.

Cancer classification using gene expression data.
Information Systems, 28(4):243–268, jun 2003.

Kevin P. Murphy.

Machine learning : a probabilistic perspective.
MIT Press, 2012.

National Human Genome Research Institute.

Fact Sheets about Genomics.
https://www.genome.gov/about-genomics/fact-sheets, 2015.

D. V. Nguyen and D. M. Rocke.

Tumor classification by partial least squares using microarray gene expression data.
Bioinformatics, 18(1):39–50, jan 2002.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer.

Automatic differentiation in PyTorch.
In *NIPS-W*, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay.

Scikit-learn: Machine Learning in Python.
Journal of Machine Learning Research, 12(Oct):2825–2830, 2011.

Sebastian Raschka.

Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.
nov 2018.

Bibliografia VII

Sebastian Raschka and Vahid Mirjalili.

Python Machine Learning, 2nd Ed.
Packt Publishing, Birmingham, UK, 2 edition, 2017.

Mark Schmidt, Nicolas Le Roux, Francis Bach, Nicolas Le Roux, and Francis Bach.

Minimizing finite sums with the stochastic average gradient.
Mathematical Programming, 162(1-2):83–112, sep 2013.

Haipeng Shen and Jianhua Z. Huang.

Sparse principal component analysis via regularized low rank matrix approximation.
Journal of Multivariate Analysis, 99(6):1015–1034, jul 2008.

Jinlong Shi and Zhigang Luo.

Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples.
Computers in Biology and Medicine, 40(8):723–732, 2010.

Piotr Sobczyk, Malgorzata Małgorzata Bogdan, and Julie Josse.

Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood.
Journal of Computational and Graphical Statistics, 26(4):826–839, jun 2017.

Piotr Sobczyk, Stanislaw Wilczynski, Julie Josse, and Malgorzata Bogdan.

varclust: Variables Clustering, 2019.

C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu.

Breast cancer classification and prognosis based on gene expression profiles from a population-based study.
Proceedings of the National Academy of Sciences, 100(18):10393–10398, 2003.

Bibliografia VIII

Alexander Statnikov, D Hardin, and Constantin Aliferis.

Using SVM Weight-Based Methods to Identify Causally Relevant and Non-Causally Relevant Variables.
Sign, 1, 2006.

R Tibshirani, Michael J Seo, G Chu, Balasubramanian Narasimhan, and Jun Li.

samr: SAM: Significance Analysis of Microarrays, 2018.

Robert Tibshirani.

Regression Shrinkage and Selection via the Lasso.
Journal of the Royal Statistical Society, 58:267–288, 1996.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu.

Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays.
Statistical Science, 18(1):104–117, feb 2003.

Laura Toloși and Thomas Lengauer.

Classification with correlated features: unreliability of feature ranking and solutions.
Bioinformatics, 27(14):1986–1994, jul 2011.

V G Tusher, R Tibshirani, and G Chu.

Significance analysis of microarrays applied to the ionizing radiation response.
Proceedings of the National Academy of Sciences of the United States of America, 98(9):5116–21, apr 2001.

Laura J. Van't Veer, Hongyue Dai, Marc J. Van de Vijver, Yudong D. He, Augustinus A M Hart, Mao Mao, Hans L. Peterse, Karin Van Der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend.

Gene expression profiling predicts clinical outcome of breast cancer.
Nature, 415(6871):530–536, 2002.

Bibliografia IX

Sudhir Varma and Richard Simon.

Bias in error estimation when using cross-validation for model selection.
BMC bioinformatics, 7:91, feb 2006.

Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els MJJ Berns, David Atkins, and John A Foekens.

Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.
The Lancet, 365(9460):671–679, feb 2005.

Britta Weigelt, Johannes L. Peterse, and Laura J. Van't Veer.

Breast cancer metastasis: Markers and models.
Nature Reviews Cancer, 5(8):591–602, 2005.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins.

Predicting the clinical status of human breast cancer by using gene expression profiles.
Proceedings of the National Academy of Sciences, 98(20):11462–11467, sep 2001.

Stanisław Wilczyński.

Reduction of dimensionality by sparse subspace clustering.
Bachelor's thesis, University of Wrocław, 2017.

Dejun Zhang, Lu Zou, Xionghui Zhou, and Fazhi He.

Integrating Feature Selection and Feature Extraction Methods with Deep Learning to Predict Clinical Outcome of Breast Cancer.
IEEE Access, 6:28936–28944, may 2018.

Bibliografia X

Shunpu Zhang.

A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance.
BMC bioinformatics, 8:230, jun 2007.

J. Zhu and Trevor Hastie.

Classification of gene microarrays by penalized logistic regression.
Biostatistics, 5(3):427–443, jul 2004.

Hui Zou, Trevor Hastie, and Robert Tibshirani.

Sparse Principal Component Analysis.
Journal of Computational and Graphical Statistics, 15(2):265–286, jun 2006.