European Proceedings of
Computers and Technology
EpCT

www.europeanproceedings.com

e-ISSN: 2672-8834

DOI: 10.15405/epct.23021.47

HMMOCS 2022
International Workshop "Hybrid methods of modeling and optimization in complex systems"

CANCER PREDICTION MODELS USING GENE EXPRESSION AND LOGICAL ANALYSIS OF DATA

M. Bartosh (a), I. Masich (b)*
*Corresponding author

(a) Siberian Federal University, 79, Svobodny av., Krasnoyarsk, 660041, Russian Federation, bartoshmari@yandex.ru
(b) Siberian Federal University, 79, Svobodny av., Krasnoyarsk, 660041, Russian Federation, Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy av., Krasnoyarsk, 660037, Russian Federation, i-masich@yandex.ru

Abstract

The paper analyzes gene expression data in relation to the diagnosis and prognosis of the development of oncological diseases. The goal is to create a hybrid prediction model based on gene expression data and interpretive machine learning. The experiments were carried out on four publicly available gene expression datasets in relation to the prediction of breast and lung cancers. Data sets contain information about positive and negative observations, described by tens of thousands of attributes with gene expression data. Logical analysis of data is investigated as the main method for building a model. This method is based on combinatorics and optimization. As a result of logical analysis of data, a set of patterns is built, each of which involves only a small number of input attributes (genes). The search for a reference set of attributes, which is a step in the logical analysis of data, yields a small number of genes that have a combinatorial effect on the result. The resulting patterns have a small number of conditions and are understandable to the user. A comparison was made with other machine learning algorithms, including rule based classifiers: RIPPER, decision trees, and others. Logical analysis of data has advantages both in terms of classification accuracy and result interpretability, and therefore provides greater confidence in the recognition result.

Keywords: Logical analysis of data, interpretable machine learning, gene expression, cancer prediction

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

1. Introduction

Cancer is a large group of diseases that occur when normal cells turn into cancer cells. Cancer is a genetic disease caused by changes in the genes responsible for controlling cell growth and proliferation.

The behaviour of normal cells follows the algorithm laid down in the genes. Cancer cells ignore these rules: they spread rapidly and uncontrollably, are not programmed to die and are able to move around.

Scientists have discovered hundreds of DNA and genetic mutations that contribute to the formation, growth and spread of cancer (Schrijver et al., 2017).

Oncology is characterised by the phenomenon of metastasis. This process involves the spread of cancer cells to other parts of the body, via the blood and lymphatic systems. For example, breast cancer can metastasise and spread to the lungs.

It is worth mentioning that metastasis is a feature of malignant cancers. There are also benign cancers that can grow but do not spread to other organs.

Scientists distinguish more than 120 types of cancer. Some are tumour-free, such as leukaemia, myeloma and most types of lymphoma.

At the moment, there are several ways of treating cancer. One of them is biomarker testing for the selection of a personalised treatment. This method consists of searching for genes, proteins and other coliforms in the human body that provide information about cancer.

2. Problem Statement

There are different sets of genetic data available. The challenge is to determine a patient's predisposition to cancer and to predict the development of cancerous tumours.

This can be done using artificial intelligence (hereafter, AI) and machine learning technologies. Several AI platforms currently exist to assist in diagnosis. Such services often work with images such as MRI and CT scans or mammograms.

Machine learning technologies can work with genetic datasets, for example, to prescribe immunotherapy (Sanatkar et al., 2022). Among machine learning techniques, interpretive machine learning stands out.

3. Research Questions

There are now genetic data sets for most cancers. Among women, breast cancer is the most common. It accounts for 12.5%, according to 2020 data. In second place is lung cancer (12.2%). In men, the most common cancers are lung cancer (15.4%) and prostate cancer (15.1%) (Koul, 2022).

3.1. GSE22820 (Breast cancer)

This kit contains gene expression profiles of 176 disease-positive patients and 10 negative samples. Sample data with some attributes is presented in Table 1. The RNA of the patients is displayed in the kit. Comparison of gene expression levels allows differentiation of classes based on gene intensity. Each

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

observation contains information on 33580 genes (Feltes et al., 2019; Krishnan et al., 2016; Kumaran et al., 2017; Liu et al., 2011; Pandya et al., 2016; Wuest et al., 2015).

Table 1. GSE22820 dataset

samples	type	NM_004900	AA085955	NM_014616	AK092846	NM_001539
GSM564920_1	primary_breast_cancer	7.64210136995268	4.80302240648272	7.93966585147253	5.79012706345684	12.1626034560638
GSM563922_3	primary_breast_cancer	4.75895429681992	5.1505727893952	6.56899813819261	5.35249740369706	12.0672644033371
GSM563923_4	primary_breast_cancer	6.42325438655315	4.40895476438672	8.668738795381	4.97965304445194	11.7047309785621
GSM564105_186	normal	6.8765956678364	4.48879000210762	8.70501433307991	4.97997959958908	11.2328609531273

3.2. GSE42568 (Breast Cancer)

This kit contains gene expression profiles of 121 patients, of whom 104 with susceptible breast cancer and 17 with absent disease. The observation is a sequence of 54676 genes (Table 2).

The average age of the patients is 58 years, with only 20 of them being less than 50 years old at the time of diagnosis. Cancer size ranged from 0.6 to 8.0 cm and was divided into 3 classes:

- tumors measuring less than 2 cm, which amounted to 18 observations;
- tumors measuring 2 - 5 cm corresponding to 83 observations;
- tumors larger than 5 cm were observed in 3 patients.

The tumours were also classified according to type: invasive ductal carcinoma (82 observations), invasive lobular carcinoma (17 observations) and tumours of a special type (5 observations). During the course of the disease, 59 patients developed axillary lymph node metastases. 69 women who underwent surgery received estrogen-suppressing tamoxifen. 50 patients received adriamycin as adjuvant chemotherapy. For 9 women, information about the treatment received is unknown. The longest follow-up period was 3026 days (> 8 years), and the average follow-up was 1887 days (> 5 years) (Clarke et al., 2013; Feltes et al., 2019).

Table 2. GSE42568 dataset

samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at
191	normal	7.94422547688724	5.25693845600489	4.93463002670381	6.60842556084795	2.48428964316359
194	normal	8.88403691067497	5.33157540543122	4.9048321795713	7.20400599343956	2.74970120303935
310	tumoral	10.0278438131529	5.85311025769218	4.90416913535735	6.59378388264941	2.67204979821597
311	tumoral	9.29546818685965	5.58137539611906	4.99097593040023	6.55083610004225	2.54460438498333

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

3.3. GSE18842 (Lung cancer)

This kit contains 91 cases of non-small cell lung cancer (Table 3). The aims of this study are:

- to establish gene signatures in primary adeno- and squamous cell carcinomas;
- to identify differentially expressed gene sequences depending on the stage of disease;
- identify sequences that are significant for tumour progression.

The observation consists of 54676 genes (Feltes et al., 2019; Sanchez-Palencia et al., 2010).

Table 3. GSE18842 dataset

samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at
6947	tumoral	11.0737628604903	8.43113304937739	5.90357477452566	6.89779224534877	3.45763443170828
6949	tumoral	10.9837223698108	8.83455752113108	9.06444259220682	7.27456152478921	3.51932170859168
7035	normal	9.97493674075779	7.10414716143154	8.18893534554987	7.14837101091024	3.36495783166931
7037	normal	8.45008996426443	7.08409859386128	8.18535432285567	7.31864954061628	3.64044193964678

3.4. GSE7670 (Lung cancer)

This kit contains a total of 66 normal lung tumour samples at early and advanced stages (Table 4). It contains:

- 27 paired samples from post-operative patients;
- an adjacent normal lung tissue mixture;
- tissue mixture of lung adenocarcinoma;
- 7 lung cancer cell lines.

Each patient is described by a sequence of 22284 genes (Chen et al., 2009; Feltes et al., 2019; Su et al., 2007).

Table 4. GSE7670 dataset

samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at
811	normal	9.46199409777702	5.59132341292466	6.25163449980152	7.89667667920585	3.38626959357819
813	normal	9.26361768067406	5.7590963160393	6.55558973180007	7.61068493370437	3.58771059783809
862	adenocarcinoma	10.1421329431078	6.52326213670261	8.13811857045264	7.82614327297554	3.70498914430554
864	adenocarcinoma	10.2903949922857	6.82044430110137	6.871090435516	7.67430770010589	3.5108310593542

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

4. Purpose of the Study

Extensive research on cancer prediction based on gene expression has resulted in a wealth of data. These data include expression values of tens of thousands of genes. The use of machine learning methods allows you to create models to support decision making in predicting the development of cancer. But in addition to the prediction and recognition itself, problems of this type require justification and interpretation of the results. The aim of the study is to build rule-based classifiers using interpretable machine learning. Of greatest interest from this point of view is the logical analysis of data, with which you can create compact classifiers based on a small number of input attributes. For the problems considered in this work, the goal is to build prediction systems based on several genes that have a combinatorial effect on the result.

5. Research Methods

This study used interpretive machine learning algorithms such as Repeated Incremental Pruning to Produce Error Reduction or RIPPER (Cohen, 1995), Decision Tree (Zhifang & Yi, 2020), Naive Bayesian Classifier (Ou et al., 2022), Random Forest (Meenal et al., 2021), k Nearest Neighbours or kNN (Lujano et al., 2022), Logical Analysis of Data or LAD (Alexe et al., 2006; Lyutikova, 2022).

The methodology of logical analysis of data has its origins in the 1986 work of Peter L. Hammer. This algorithm consists of several steps:

- Binarization of features.
- Construction of a reference set of attributes.
- Finding logical patterns.
- Construction of a classifier (solver function) on the pattern axis.

LAD is a binary interpretable classification algorithm based on combinatorics, logic and optimization. This combination allows one to explore the entire dataset without exception, focus on the classification power of gene combinations and extract new information about the role of genes and their combinations (Alexe et al., 2006).

For further analysis of the data, the traits need to be binary, so the binarization step is important in the case of other types of traits.

A greedy algorithm was used to find the reference feature set.

A pattern (or rule) is a term that covers at least one observation from a class and does not cover any observation from another class. Prime patterns are patterns that, when any literal is removed, cover observations from different classes. A minterm is a pattern that covers a single observation from a class and contains all literals from an observation.

There are several ways of finding patterns:

- Enumeration approach for searching for pattern with some properties.
- Optimization model and heuristic algorithms (for example, greedy pattern search algorithms).

We have used the latter approach in our studies.

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

6. Findings

For dataset GSE22820, 62 positive patterns and 1 negative pattern were generated (e.g., for positive NM_006928 ≤ 9.69615683152957 and NM_013989 ≤ 12.4291841944343 and NM_002666 ≤ 11.40701162971345 and XM_295309 ≤ 5.18905687566618, for negative NM_002666 > 11.40701162971345). The algorithm has allocated 14 cut points.

For dataset GSE42568, 3 positive patterns and 12 negative patterns were generated (e.g., for positive 206030_at > 6.0795822056367, 1555741_at > 4.236602141589765, for negative 206030_at ≤ 6.0795822056367 and 1555741_at ≤ 4.236602141589765 and 1553033_at ≤ 3.993169149187495). The algorithm has allocated 4 cut points.

For dataset GSE18842, 7 positive patterns and 58 negative patterns were generated (e.g., for positive 205064_at > 5.891860443655585 and 1556589_at > 6.118819468067455, for negative AFFX-HUMRGE/M10098_3_at > 10.165519828262891 and 1556589_at ≤ 6.118819468067455 and 210081_at > 7.289193563391985). The algorithm has allocated 10 cut points.

For dataset GSE7670, 58 positive patterns and 6 negative patterns were generated (e.g., for positive 205725_at > 7.482708619698739 and 201883_s_at > 8.77148410809596, for negative 205725_at ≤ 7.482708619698739 and 216510_x_at > 7.592556909814875). The algorithm has allocated 17 cut points (see Table 5).

Table 5. Results of cancer prognosis

Method	GSE22820	GSE42568	DSE18842	GSE7670
RIPPER	100%	95.6522%	100%	90%
Decision Tree	96.4286%	100%	100%	80%
Naive Bayesian Classifier	100%	95.6522%	22.2222%	70%
Random Forest	96.4286%	100%	100%	90%
kNN	92.8571%	8.6957%	72.2222%	50%
LAD	100%	100%	97.7778%	84.3137%

7. Conclusion

The results show that logical analysis of data is as accurate as, and sometimes better than, other known machine learning algorithms. The use of logical analysis of data made it possible to select a small number of genes from information on several tens of thousands of genes, which are sufficient to distinguish between positive and negative observations. Based on the data on the expression of these genes, compact classifiers were built, consisting of several patterns. The resulting classifiers have high accuracy and good interpretability. The application of the studied method seems promising for solving problems of this type. With the help of modern technology, incurable diseases such as cancer can be effectively tackled.

Acknowledgments

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

References

Alexe, G., Alexe, S., Axelrod, D. E., Bonates, T. O., Lozina, I. I., Reiss, M., & Hammer, P. L. (2006). Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Research, 8*(4), 1-20. https://doi.org/10.1186/bcr1512

Chen, C.-H., Lai, J.-M., Chou, T.-Y., Chen, C.-Y., Su, L.-J., Lee, Y.-C., Cheng, T.-S., Hong, Y.-R., Chou, C.-K., Whang-Peng, J., Wu, Y.-C., & Huang, C.-Y. F. (2009). VEGFA Upregulates FLJ10540 and Modulates Migration and Invasion of Lung Cancer via PI3K/AKT Pathway. *PLoS ONE, 4*(4), e5052. https://doi.org/10.1371/journal.pone.0005052

Clarke, C., Madden, S. F., Doolan, P., Aherne, S. T., Joyce, H., O'Driscoll, L., Gallagher, W. M., Hennessy, B. T., Moriarty, M., Crown, J., Kennedy, S., & Clynes, M. (2013). Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis, 34*(10), 2300-2308. https://doi.org/10.1093/carcin/bgt208

Cohen, W. W. (1995). Fast effective rule induction. *In Machine learning proceedings 1995* (pp. 115-123). Morgan Kaufmann. https://doi.org/10.1016/b978-1-55860-377-6.50023-2

Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology, 26*(4), 376-386. https://doi.org/10.1089/cmb.2018.0238

Koul, P. A. (2022). Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019. A Systematic Analysis for the Global Burden of Disease Study 2019. *JAMA Oncol., 8*(3), 420-444. https://doi.org/10.1001/jamaoncol.2021.6987

Krishnan, P., Ghosh, S., Wang, B., & Heyns M., Li, D., Mackey, John R., Kovalchuk, O., & Damaraju, S. (2016). Genome-wide profiling of transfer RNAs and their role as novel prognostic markers for breast cancer. *Scientific Reports, 6,* 32843. https://doi.org/10.1038/srep32843

Kumaran, M., Cass, C. E., Graham, K., Mackey, J. R., Hubaux, R., Lam, W., Yasui, Y., & Damaraju, S. (2017). Germline copy number variations are associated with breast cancer risk and prognosis. *Scientific Reports, 7,* 14621. https://doi.org/10.1038/s41598-017-14799-7

Liu, R.-Z., Graham, K., Glubrecht, D. D., Germain, D. R., Mackey, J. R., & Godbout, R. (2011). Association of FABP5 Expression with Poor Survival in Triple-Negative Breast Cancer. *The American Journal of Pathology, 178*(3), 997-1008. https://doi.org/10.1016/j.ajpath.2010.11.075

Lujano, E., Lujano, R., Huamani, J. C., & Lujano, A. (2022) Hydrological modeling based on the KNN algorithm: an application for the forecast of daily flows of the Ramis river, Peru. *Tecnología y ciencias del agua.* https://doi.org/10.24850/j-tyca-14-2-5

Lyutikova, L. A. (2022). Analysis of the results of using logical methods and neural networks in diagnostic tasks. *Procedia Computer Science, 213,* 580-587. https://doi.org/10.1016/j.procs.2022.11.107

Meenal, R., Michael, P. A., Pamela, D., & Rajasekaran, E. (2021). Weather prediction using random forest machine learning model. *Indonesian Journal of Electrical Engineering and Computer Science, 22*(2), 1208. https://doi.org/10.11591/ijeecs.v22.i2.pp1208-1215

Ou, G., He, Y., Fournier-Viger, P., & Huang, J. Z. (2022). A Novel Mixed-Attribute Fusion-Based Naive Bayesian Classifier. *Applied Sciences, 12*(20), 10443. https://doi.org/10.3390/app122010443

Pandya, V., Glubrecht, D., Vos, L., & Hanson, J. (2016). The pro-apoptotic paradox: the BH3-only protein Bcl-2 interacting killer (Bik) is prognostic for unfavorable outcomes in breast cancer. *Oncotarget, 7*, 33272-33285. https://doi.org/10.18632/oncotarget.8924

Sanatkar, S. A., Heidari, A., & Rezaei, N. (2022). Cancer Immunotherapy: Diverse Approaches and Obstacles. *Current pharmaceutical design, 28*(29), 2387–2403. https://doi.org/10.2174/1381612828666220728160519

Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J. A., Pedraza, V., Boyero, L., Rosell, R., & Fárez-Vidal, M. E. (2010). Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer, 129*(2), 355-364. https://doi.org/10.1002/ijc.25704

https://doi.org/10.15405/epct.23021.47
Corresponding Author: I. Masich
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

Schrijver, W. A. M. E., van Diest, P. J., Moelans, C. B., & Dutch Distant Breast Cancer Metastases Consortium. (2017). Unravelling site-specific breast cancer metastasis: a microRNA expression profiling study. *Oncotarget, 8*(2), 3111-3123. https://doi.org/10.18632/oncotarget.13623

Su, L. J., Chang, C. W., Wu, Y. C., & Chen, K. C. (2007). Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics, 8*(140), 1-12. https://doi.org/10.1186/1471-2164-8-140

Wuest, M., Kuchar, M., Sharma, S. K., Richter, S., Hamann, I., Wang, M., Vos, L., Mackey, John R., Wuest, F., & Löser, R. (2015). Targeting lysyl oxidase for molecular imaging in breast cancer. *Breast Cancer Research, 17*(1), 1-15. https://doi.org/10.1186/s13058-015-0609-9

Zhifang, S., & Yi, L. (2020). Optimization of Decision Tree Machine Learning Strategy in Data Analysis. *Journal of Physics: Conference Series, 1693*(1), 012219. https://doi.org/10.1088/1742-6596/1693/1/012219