Hierarchical relative expression analysis in multi-omics data classification

Marcin Czajkowski, Krzysztof Jurczuk, and Marek Kretowski

Faculty of Computer Science, Bialystok University of Technology,
Wiejska 45a, 15-351 Bialystok, Poland
{m.czajkowski, k.jurczuk, m.kretowski}@pb.edu.pl

Abstract. This study aims to develop new classifiers that can effectively integrate and analyze biomedical data obtained from various sources through high-throughput technologies. The use of explainable models is particularly important as they offer insights into the relationships and patterns within the data, which leads to a better understanding of the underlying processes.

The objective of this research is to examine the effectiveness of decision trees combined with Relative eXpression Analysis (RXA) for classifying multi-omics data. Several concepts for integrating separated data are verified, based on different pair relationships between the features. Within the study, we propose a multi-test approach that combines linked top-scoring pairs from different omics in each internal node of the hierarchical classification model. To address the significant computational challenges raised by RXA, the most time-consuming aspects are parallelized using a GPU. The proposed solution was experimentally validated using single and multi-omics datasets. The results show that the proposed concept generates more accurate and interpretable predictions than commonly used tree-based solutions.

Keywords: Relative expression analysis · Decision trees · Multi-omics data · Classification.

1 Introduction

Comprehensive multi-omics analysis refers to the simultaneous study of multiple types of omics data, such as genomics, proteomics, and metabolomics, in order to gain a more complete understanding of biological systems [10]. This type of analysis can provide a holistic view of the advanced interactions within a biological system by combining data from different omics platforms and modalities. However, multi-omics data is typically high-dimensional, making it difficult to analyze and interpret. Traditional machine learning algorithms for biomedical data tend to prioritize prediction accuracy and use complex predictive models, which can impede the discovery of new biological understanding and hinder practical applications [1]. Currently, there is a strong need for simple, interpretable models that can aid in understanding and identifying relationships between specific features, and enhance biomarker discovery.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_69

This research focuses on the development of computational methods for biomedical analysis within the field of interpretable and explainable machine learning. The main idea is not only to perform predictions efficiently, but also provide insight, which is the ultimate goal of data-driven biology. To address this challenge, we enhanced Decision Tree (DT) [11], well-known white-box approach [1], to unlock its potential in contemporary biological data analysis. We propose DTs with splits composed of several simple tests which are based on Relative eXpression Analysis (RXA) concept [7]. In the rest of this article, we will refer to such group of tests as multi-test.

In contrast to DT, Relative eXpression Analysis (RXA) was originally developed for a specific task of identifying connections among a small group of genes [9]. RXA methods focus on analyzing the relative order of gene expressions, as opposed to their raw values, which makes them robust to various factors such as methodological and technical issues, biases, and normalization procedures. The most commonly used method within RXA is top scoring pair (TSP) analysis which examines the pairwise ordering relationships between two genes within the same sample. These pairs of genes can be viewed as "biological switches" that are linked to regulatory patterns or other aspects of gene expression networks. There are several extensions of TSP within RXA, including increasing the number of pairs in the prediction model, extending the relationships to more than two genes, and hierarchical interactions between the genes. TSP and its extensions have been successful in real-world applications due to their straightforward biological interpretation. Their effectiveness has also been recognized in other fields such as proteomics and metabolomics, but they have not yet been applied for multi-omics analysis.

A simple straightforward application of the TSP solution is not possible due to the following reasons:

- TSP is designed for binary classification, however, with the use of DT it can be successfully applied for multi-class problems [3];
- high computational complexity of the RXA solution which significally limits the size of the analyzed data;
- lack of more advanced inter-gene relations especially in the context of multi-omics data.

Some of the aforementioned issues have been already addressed in various TSP extensions. In one of them called Relative eXpression Classification Tree (RXCT) [5], the top-scoring pairs are applied as a splitting rules in a top-down induced DT. The search for pairwise relationships is paralellized using the GPU which significally improves the speed of the creating the prediction model.

In this study, we use the RXCT solution [5] as a baseline and extend the RXA concept to the multi-omics analysis. In the proposed solution called Relative Multi-test Classification Tree (RMCT), we introduce the multi-test splitting rule [4] which can be viewed as a collection composed of multiple pairwise comparisons. The feature space of each pair of attributes is limited to a single omic and preserve the clarity in interpretation by not mixing "apples and oranges". The general idea is to use multi-tests in which top-scoring-pairs are tightly linked

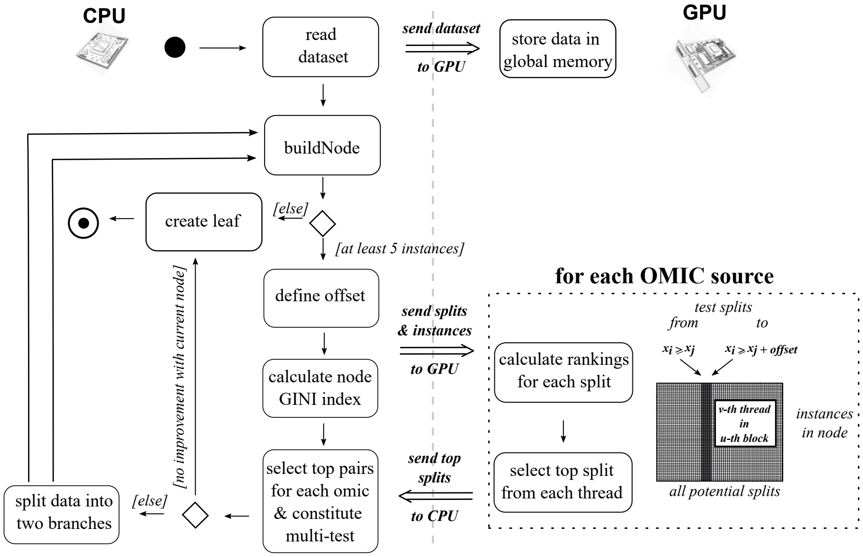

CPU
read dataset
send dataset
to GPU
store data in global memory
GPU
buildNode
create leaf
[else]
[at least 5 instances]
define offset
send splits & instances
to GPU
for each OMIC source
test splits
from
to
$x_i \geqslant x_j$
$x_i \geqslant x_j$ + offset
calculate rankings for each split
v-th thread in u-th block
instances in node
all potential splits
calculate node GINI index
[no improvement with current node]
select top pairs for each omic & constitute multi-test
send top splits
to CPU
select top split from each thread
split data into two branches
[else]

Fig. 1. General flowchart of a RMCT algorithm.

and could participate in a common pathway. The measure of similarity between the pairwise unit tests pairs that compose multi-test is the number of observations routed in the same way. It can be viewed as a mix of horizontal and vertical integrations of the multi-omics sources in each internal node of the tree.

Experimental validation of RMCT was performed on six single and two multi-omics datasets from three omics sources: gene expression, DNA methylation and miRNA. In order to evaluate the performance of the proposed concept accuracy and F1 weighted score were used.

2 Relative Multi-test Classification Tree

The new RMCT solution utilizes the RXCT algorithm as a foundation. In this section some basic steps such as DT induction and TSP creation are briefly mentioned. Next, we focus on our contribution and highlight the differences between RMCT and the RXCT system.

2.1 Overview

The general flowchart of our GPU-accelerated RMCT is illustrated in Fig. 1. It can be seen that the DT induction is run in a sequential manner on a CPU, and the most time-consuming operation is performed in parallel on a GPU. This ensures that the parallelization does not alter the behavior of the original algorithm.

DOI: 10.1007/978-3-031-36021-3_69

The overall structure of the proposed solution is based on a typical top-down induced binary classification tree. The greedy search starts with the root node, where the locally optimal split (multi-test) is searched. Then the training instances are redirected to the newly created nodes and this process is repeated for each node till there is a noticable improvement in the Gini index (default 5%). We propose using a two-stage scoring method to enhance performance. Initially, a screening and scoring process is done using the GPU, and then the top results are further evaluated by the CPU.

As it is illustrated in Figure 1, the data is initially transferred from the CPU's main memory to the GPU's device memory, allowing each thread block to access it. This process is done only once before beginning the tree induction, as later on only the indexes of instances that are present in a calculated node are sent. The CPU launches GPU functions, known as kernels, separately for each omic source. Each thread on the device is given an equal amount of relations, referred to as offset, to compute. The algorithm scores all possible pairwise relations from a single source and uses the Gini index to calculate the scores. This is done to make the algorithm suitable for multi-class analysis. After all the thread blocks finish their calculations, the results are transferred from the GPU's memory to the CPU's memory.

2.2 Constituting the multi-test

The final tree node split is built on the CPU on the basis of the results from the GPU computation. We have studied three ways of constituting multi-test by taking into account two most common strategies [15] of integrating multi-omics data (see Fig. 2.) and our own hybrid approach.

(i) The horizontal integration treats each type of omics measurements equally. It can be viewed as a direct extension from single-omics data analysis to integrative analysis, where important associations between multi-level omics measurements have been identified simultaneously in a joint model. We realize horizontal integration by building the multi-test split with top-scoring-pair from each single-omic source.

(ii) The hierarchical integration incorporates the prior knowledge of the regulatory relationship among different platforms of omics data. The integration methods are developed to more closely reflect the nature of multidimensional data. Alike in horizontal integration, we use multi-test with each pair from single-omic source. However, instead of looking at the highest scoring pairs, we focus on ones that are associated with the same class (surrogate test). This way, the multi-test stores hierarchical interactions between each source and becomes a collection of rules with similar patterns.

(iii) The proposed hybrid method simplifies the guidelines in (ii) by eliminating the requirement to utilize steam from every source. Even though the number of pairs in the multi-test stays the same, we permit the possibility that in certain parts of the tree the data may be split using pairs from only 2 or 1 data sources. This way the hybrid approach can also work with the single-omic data. The attributes that make up the pairs cannot be repeated, and

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_69

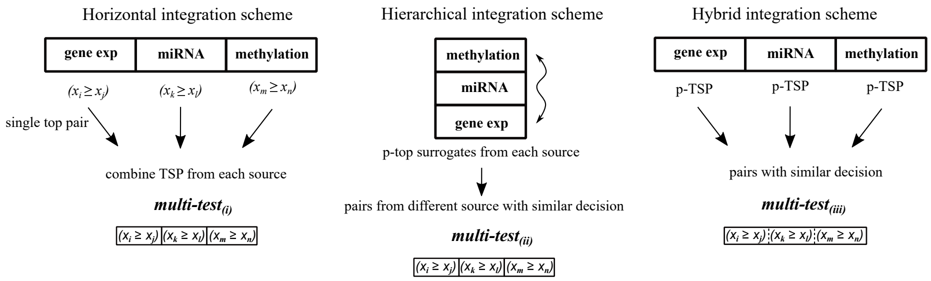

Horizontal integration scheme
gene exp
miRNA
methylation
$(x_i \geq x_j)$
$(x_k \geq x_l)$
$(x_m \geq x_n)$
single top pair
combine TSP from each source
multi-test$_{(i)}$
$(x_i \geq x_j)$ $(x_k \geq x_l)$ $(x_m \geq x_n)$
Hierarchical integration scheme
methylation
miRNA
gene exp
p-top surrogates from each source
pairs from different source with similar decision
multi-test$_{(ii)}$
$(x_i \geq x_j)$ $(x_k \geq x_l)$ $(x_m \geq x_n)$
Hybrid integration scheme
gene exp
miRNA
methylation
p-TSP
p-TSP
p-TSP
pairs with similar decision
multi-test$_{(iii)}$
$(x_i \geq x_j)$ $(x_k \geq x_l)$ $(x_m \geq x_n)$

Fig. 2. Integration schemes of multi-omics data to constitute a split in internal node.

similarity between the pairs remains the primary criterion for including them in the multi-test.

Finally, the splitting criterion is guided by a majority voting mechanism in which all pair components of the multi-test have the same weight.

3 Experiments

In this section, we experimentally validate the proposed RMCT approach and confront its results with popular counterparts.

3.1 Datasets and setup

In our experiments we used datasets from Multi-Omics Cancer Benchmark TCGA Preprocessed Data repository [12]. From the list of datasets we have selected two: Glioblastoma with 4 classes (Classical: 71 instances, Mesenchymal: 84, Neural: 47, Proneural: 72) and Sarcoma with 5 classes (DDLPS: 71, LMS: 105, MFH: 29, MFS: 21, UPS: 21) as they have the largest number of patients and clinical data with defined class labels. Each dataset consists of three types of omics data: gene expression, DNA methylation, and miRNA expression. We also perform single-omics analysis in which algorithms are tested on each data source seperately. Due to the performance reasons, the Relief-F [13] feature selection was applied and the number of selected genes was arbitrarily limited to the top 500 for each omic, thus 1500 attributes for multi-omic datasets.

We evaluate the performance of the proposed RMCT solution with three multi-test variants denoted as $RMCT_i$, $RMCT_{ii}$, $RMCT_{iii}$ (see Section 2.2) against:

- RXCT [5], the predecessor of the RMCT approach;
- C4.5 [14]: popular state-of-the-art tree learner (Weka implementation [8] under name J48);
- JRip: rule learner – repeated incremental pruning to produce error reduction (RIPPER) [2].

DOI: 10.1007/978-3-031-36021-3_69

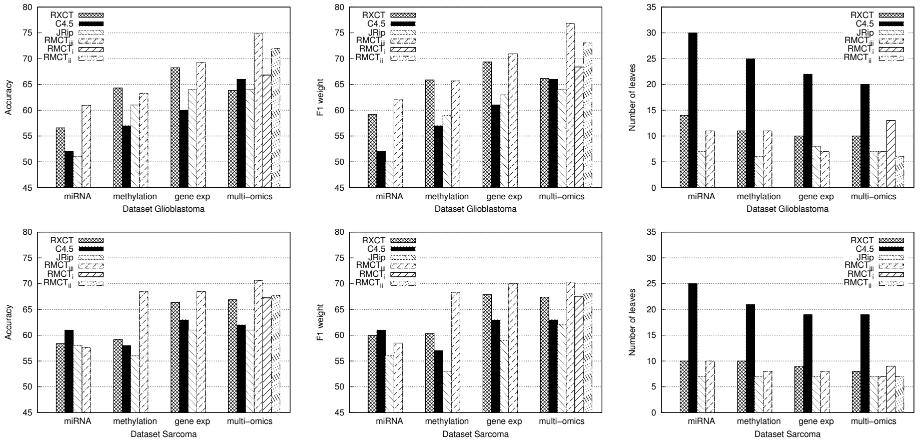


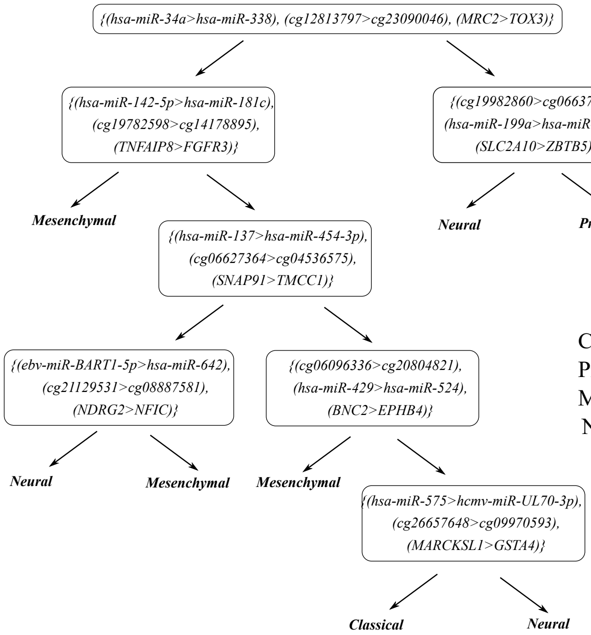
Fig. 3. The performance of the algorithms in terms of accuracy, F1 weight and size on the multi-omics Glioblastoma and Sarcoma databases and their single-omic sets.

A standard 10-fold cross-validation technique is used to evaluate the performance of the proposed solutions. The evaluation is based on accuracy, F1 weight score (a modified version of the F1 macro algorithm that takes into account the imbalance in the samples) and the number of tree leaves/rules. Due to the multi-class nature of the datasets, it was not possible to compare the RMCT with other TSP-family solutions. However, in previous research [5], we found that the RXCT algorithm outperformed other popular TSP-family algorithms on 8 real-life cancer-related datasets that concern binary classification. It should be noted that in case of single-omic data, results are only provided for $RMCT_{iii}$.

3.2 Results

Figure 3 presents a summary of the classification performance for the proposed solution, the RMCT, and its competitors for both multi-omics datasets and their individual components. The results show that RMCT outperforms the predecessor RXCT algorithm and popular white box classifiers such as the decision tree C4.5 and JRip learner. Analysis of the results using the Friedman test revealed statistically significant differences between the algorithms ($p\text{-value} < 0.05$) in terms of accuracy. According to Dunn's multiple comparison test [6], the RMCT with a hybrid multi-test variant ($RMCT_{iii}$) was able to significantly outperform the other algorithms on both multi-omics datasets and most of the single-omic sets. On average, most of the tested classifiers showed improved results when using multi-omics data. However, in some cases, using single-omics data resulted in more distinct patterns that distinguished between classes. In all cases, the classification models built on gene expression datasets were more accurate than those built on miRNA data. The highest prediction performance on all datasets was achieved by the RMCT algorithm with a hybrid multi-test variant ($RMCT_{iii}$).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_69


{(hsa-miR-34a>hsa-miR-338), (cg12813797>cg23090046), (MRC2>TOX3)}
{(hsa-miR-142-5p>hsa-miR-181c), (cg19782598>cg14178895), (TNFAIP8>FGFR3)}
{(cg19982860>cg06637774), (hsa-miR-199a>hsa-miR-590), (SLC2A10>ZBTB5)}
Mesenchymal
Neural
Proneural
{(hsa-miR-137>hsa-miR-454-3p), (cg06627364>cg04536575), (SNAP91>TMCC1)}
{(ebv-miR-BART1-5p>hsa-miR-642), (cg21129531>cg08887581), (NDRG2>NFIC)}
{(cg06096336>cg20804821), (hsa-miR-429>hsa-miR-524), (BNC2>EPHB4)}
Neural
Mesenchymal
Mesenchymal
{(hsa-miR-575>hcmv-miR-UL70-3p), (cg26657648>cg09970593), (MARCKSL1>GSTA4)}
Classical
Neural

Results
- accuracy: 93.43%
- F1 weight: 93.58
- size: 8 leaves

Confusion matrix:

		(A)	(B)	(C)	(D)
Classical	(A)	65	0	0	3
Proneural	(B)	0	67	0	2
Mesenchymal	(C)	6	3	82	0
Neural	(D)	0	2	2	42

Fig. 4. An example multi-test DT induced by $RMCT$ for the Glioblastoma multi-omics dataset together with prediction results and a confusion matrix.

The complexity comparison (see Figure 3), shows that the C4.5 algorithm generates larger trees than the other tested solutions, despite all being considered interpretable. When the classifier representation is inadequate, the true relationship is only partially captured or with the inclusion of uninformative features, making it difficult to understand and interpret the output model. This is often reflected in the increased size of the generated model, rather than a decrease in accuracy.

Due to the limited space in the paper, the authors only briefly mention that they examined the rules generated by the $RMCT$ and their biological relevance in the TCGA research network [12]. They found that on average, 25% of the features used in the models (especially in the upper parts of the tree) were directly related to the analyzed cancer, and an additional 30-40% were discussed in several papers in the medical literature. An example DT induced by the $RMCT$ for Glioblastoma multi-omic dataset is shown in Figure 4. However, these are preliminary results, and further work is planned with biologists to better understand the gene-gene relationships generated by the $RMCT$.

4 Conclusions

The presented research explores the use of a decision tree and relative expression analysis for classification of multi-omics data. Three data integration methods, referred to as multi-test, were proposed and validated for determining the split in the tree nodes. Results from initial experiments on both single and multi-omics datasets indicate that the proposed method, RMCT, is able to identify various patterns and improve accuracy compared to other solutions. Future research aims to expand the use of integrated multi-omics analysis on proteomic and metabolomic data for pathway analysis and more specialized classification.

Acknowledgments This project was funded by the Polish National Science Centre and allocated on the basis of decision 2019/33/B/ST6/02386 from BUT founded by Polish Ministry of Science and Higher Education (first and second author). The third author was supported by the grant WZ/WI-IIT/4/2023 from BUT founded by Polish Ministry of Science and Higher Education.

References

1. Chen, X., Wang, M., Zhang, H.: The use of classification trees in bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge 55–63 (2011)
2. Cohen, WW.: Fast Effective Rule Induction. In: ICML95, Morgan Kaufmann, San Francisco, CA, USA, 115–123 (1995)
3. Czajkowski M., Kretowski, M.: Top Scoring Pair Decision Tree for Gene Expression Data Analysis. Advances in Experimental Medicine and Biology 696, 27–35 (2011)
4. Czajkowski M., Kretowski M.: Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. Expert Syst. Appl. 137, 392–404 (2019)
5. Czajkowski M., Jurczuk K., Kretowski M., Relative Expression Classification Tree. A Preliminary GPU-based Implementation. In: PPAM'19, LNCS 12043 359–369 (2020)
6. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7 1–30 (2006)
7. Eddy, JA., Sung, J, Geman D., Price, ND.: Relative expression analysis for molecular cancer diagnosis and prognosis. Technol Cancer Res Treat. 9(2) (2010)
8. Frank E., Hall, MA., Witten IH.: The WEKA Workbench. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann (2016)
9. Geman, D., d'Avignon, C., Naiman, DQ., Winslow, RL.: Classifying gene expression profiles from pairwise mRNA comparisons. Statistical Applications in Genetics and Molecular Biology 3(19) (2004)
10. Huang, S., Chaudhary, K., and Garmire, L. X., More Is Better: Recent Progress in Multi-Omics Data Integration Methods, Front. Genet. 8 (2017)
11. Kotsiantis, S.B.: Decision trees: A recent overview. Artificial Intelligence Review 39(4), 261–283 (2013)
12. Multi-Omics Cancer Benchmark TCGA Preprocessed Data repository [*http : //acgt.cs.tau.ac.il/multiomic_benchmark/download.html*]
13. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning, 53(1–2), 23–69 (2003)
14. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, USA (1993)
15. Wu, C., Zhou, F., et al.: A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High-Throughput 8(1) (2019)