State of the Art on: Machine Learning algorithms applied on RNA-seq data to classify subtypes of Colorectal Cancer

CHIARA BARBERA, CHIARA.BARBERA@MAIL.POLIMI.IT

1. Introduction to the research topic

Colorectal cancer represents the third major cause of death with respect to cancer disease [1]. This cancer is studied also in computer science since its genomic data are suitable for the application of machine learning techniques, even if this implies facing a relatively low number of samples and a huge number of genomic features. In our work, we will replicate and then **improve the CRIS (ColoRectal Intrinsic Subtype) classification** described in [2], with the **aim of developing a single-sample classifier using only RNA-seq data** [1]: given a unique cancerous sample, the algorithm should be able to assign it confidently to one of the CRIS subtypes.

Even if conferences and journals can be ranked on the basis of H5 index, in the case of genomic computing related to cancer data it is difficult to decide which journal or conference is more important, since the computational side and the biological side can be seen as complementary. Among the relevant journals, we can cite *Bioinformatics*, *Computer Methods and Programs in Biomedicine*, *IEEE/ACM Transaction on Bioinformatics and Computational Biology*, *Genome Medicine*, *Cancer Research*, *European Journal of Cancer*.

1.1. Preliminaries

Machine learning is a field of computer science that studies how to solve problems automatically. In our research, we will consider a **classification problem**: given a set of input data $D = \{d | d =< \vec{x}, c >\}$, where each sample d is made by a vector $\vec{x}$ of input features and a target c representing the class of the sample, the objective of classification is the computation of a function $f(\vec{x})$ such that $f(\vec{x}) = c \ \forall$ samples $d \in D$. Examples of classification algorithms are Logistic Regression, Naive Bayes Classifier, Random Forest and Support Vector Machines (SVM). In this field, we usually have more features than samples, due to the massive number of genomic data associated with each investigated sample. which makes a classification problem harder to solve. To cope with this issue, **feature selection methods** are applied, whose objective is to select only the input features meaningful for the task. A great advantage of this processing is reducing the possibility of overfitting, which refers to the inability of having a robust prediction on new data if the model fits too much the training data.

Rather than listing classifiers, we think that a glossary with the list of recurring biological terms would be useful for the reader, in order to better comprehend the rest of the document:

Colorectal Cancer: according to [3] the term *cancer* refers to *"a group of diseases which cause cells in the body to change and grow out of control"*. Cancers that develop in colon or rectum have common characteristics and thus they can be referred to with the unifying term *colorectal cancer* (CRC) [4]. A detailed description of the stages of CRC can be found at [5].

Genes, oncogenes and tumor suppressor genes [6]: genes are portions of a DNA molecule that form a chromosome. They are usually stored in at most two copies in the genome (**copy number** ≤ 2). Genes may go

[1] **RNA-seq**: data relative to the sequencing of RNA obtained through NGS techniques (Next Generation Sequencing)

through **mutations**, i.e. changes of one or more nucleotides into their sequence. **Oncogenes** (for example BRAF and KRAS,)[7] if mutated, contribute to the cancer development. **Tumor suppressor genes** (e.g. APC and TP53,)[7], if mutated, become unable to control the cell division process.

Genome and Epigenome: Genome is the set of all genes of an individual [6]; epigenome is referred to as the ensemble of factors that can cause an alteration of the activity of the genes without modifying the DNA sequence[8]. These factors, which are not part of DNA, can modify or mark the genome and thus change where, how and when genes are expressed [6].

Gene expression [6]: Each gene can be active or inactive, according to its possibility to be translated or not into a functional product it encodes (i.e a protein or an RNA molecule): by measuring the quantities of functional products (in general, proteins), it is possible to measure the **expression** of a gene, i.e. the representation of its level of activity.

NGS and Microarray technologies: Microarray technology uses a fluorochrome marked sample to detect which sequences of it bind to a chip of genes [6]. Next Generation Sequencing (NGS) is a high-throughput method that can process multiple sequences at once [9] and can provide both DNA-seq, RNA-seq, and ChIP-seq data (sequences where transcription factors bind). While with microarray we need to know the genes or the mutations we are looking for [10], with NGS we can perform whole genome sequencing, detect unknown changes [11] [12] and SNPs [10] (Single Nucleotide Polymorphisms, variations of a single nucleotide in a gene).

Normalization of data: process whose goal is *to remove global variation to make readings across different experiments comparable* [13]. The most frequently adopted normalization techniques for RNA-seq expression values are used to preprocess datasets and express gene level quantification in terms of well-known expression metrics such as FPKM/RPKM (Fragments/Reads Per Kilobase Million), or TPM (Transcripts Per Million) [14].

In this field, programming languages like Python and R are used, together with dedicated statistical and machine learning packages. For example, a list of useful libraries for Python is provided [15]: `Pandas` can be used for data extraction and preparation; `Numpy` can be used to handle multi-dimensional arrays of data; `Scikit-learn` offers several supervised and unsupervised machine learning algorithms. For R language, some useful packages are listed too [16]: `Caret` integrates training and prediction for regression and classification; `e1071` can be used to implement several machine learning algorithm; `Kernlab` focuses on kernel-based algorithms.

1.2. Research topic

This research is important because it will cope with issues of the existing classification approaches; in particular, we will start from the CRIS classification of Isella et al. [2], which emerged in their work as a valuable classification system for Colorectal cancer. The objective of our research is to improve the single-sample classifier they developed and focus only on RNA-seq data: differently from them, thus, we will not consider microarray data, to analyze the role that additional features provided by RNA-seq data may have in the identification of the subtypes. On the computer science side, other than the aforementioned problem of dataset size, we have to take into account that genomic data extracted from cancer cells are intrinsically heterogeneous and differences in their normalization procedures and/or in platforms used for data production may introduce a bias in the classification. Finally, an inherent ambiguity of the biological data may lead to the impossibility of correctly classifying a portion of the samples. It is known that an early detection diagnosis of this disease is important, since the 5-year relative survival rate is about 90% if the cancer is discovered at early stages [17]. Moreover, CRC is an heterogeneous disease: having a reliable classification system for its subtypes can allow to predict the prognosis based on the subtype and provide tailored therapies.

2. Main related works

2.1. Classification of the main related works

Since our work will start from the results of Isella *et al.* in [2], this will be our main reference paper, which will be described with particular attention. Its description will be the last in order to provide the reader with information on other works which, in some cases, have constituted the basis also for the research of Isella *et al.*. The references will be described with the main author's name (with publication year), the main results presented in the work, the machine learning techniques applied and the limitations, if any, of each research. We chose to report these reference because they show previous attempts of classification ([18], [2]), how normalization procedures have influence on the classification result ([19]) and how research on genomic data relative to cancer can be relevant for the prognosis and the chemotherapy response [20]. Finally, we cited [21] because of the comparison of the different algorithms and of the original usage of a neural network.

2.2. Brief description of the main related works

Guinney *et al.* (2015) [18]:

Main results - Reorganization of 6 different classifications in the Consensus Molecular Subtypes (CMS); prognostic and clinical differences were found within the four subtypes of the CMS.

Datasets - 18 CRC datasets, microarray and RNA-seq uniformly preprocessed and normalized;

Techniques applied - Network based approach + Markov Cluster algorithm to identify 4 main groups; 78% of the samples were significantly associated with one of the 4 types. A Random Forest approach was used for the realization of the classifier.

Limitations - 22% of the samples remained unlabeled and scattered between the four subtypes; CMS4 subtype was influenced by stroma.

Murcia *et al.* (2019) [20]:

Main results - Proof of differences in prognosis, survival rate and response to chemotherapy within the subtypes proposed by Phipps *et al.* [22] and Sinicrope *et al.* [23] in 2015.

Datasets - Data from 878 CRC patients with data on the mutation status of BRAF and KRAS genes, CIMP [2] and MSI [3] status (together with other parameters).

Techniques applied - Completed the samples with missing data related to at most two markers between CIMP, MSI and gene mutation status. They performed univariate and multivariate analysis. [22].

Limitations - Classified only 71% of samples because of either inherent ambiguity of the samples or because some molecular pathways have not been included in the classification they used. The small amount of samples from patients that respect criteria related to therapies can bring bias in the classification.

Franks *et al.* (2018) [19]:

Main results - Proposed a Feature Specific Quantile Normalization (FSQN) that removes platform bias allowing to obtain good results for classifiers trained on microarray data and tested on RNA-seq data. The performances with data normalized through FSQN were higher with respect to other normalization procedures.

[2] **CIMP** = CpG Island Methylator Phenotype. An epigenetic feature that leads to methylation at promoters of CpG island (regions of the DNA with a high density of sequences Citosine-Guanine); the methylation causes inactivation of tumour suppressor genes.

[3] **MSi** = MicroSatellite Instability. A variation in the sequence or in the copy number of small repeated segments of the genome.

Datasets - RNA-seq and microarray data from The Cancer Genome Atlas (TCGA) breast invasive carcinoma and CRC samples, the latter ones classified according to the CMS.

Techniques applied - Training on microarray data and classification on RNA-seq data; the latter ones have been treated with different normalization processes, i.e. Quantile Normalization (QN), Training Distribution Matching (TDM), Numeric Pattern Normalization (NPN), untransformed log2 and FSQN; the data have been used to test three different classifiers trained on microarray data (GLMnet, i.e. Generalized Linear Model, SVM and Random Forest)

Kim *et al.* (2019) [21]:

Main results - Development of a binary classifier (normal vs. cancerous tissue) and of a pan-cancer classifier (21 types of cancer) based on bulked RNA-seq data (i.e. taken from a group of cells) and single-cell RNA-seq data (scRNA-seq).

Datasets - 7398 cancer samples and 640 normal samples from 21 tumours and normal tissues in TCGA data have been used for the training.

Techniques applied - Compared the performances of a Neural Network (NN), a linear Support Vector Machines (L-SVM), a radial basis function Support Vector Machine (RBF-SVM), a k-Nearest Neighbours (kNN) and a Random Forest (RF) classifier trained on different sizes of gene sets. Through 10-fold cross-validation, the NN with 300 selected genes resulted to be the best performing in both cases (binary and pan-cancer classifier). Performances were measured with accuracy (from 0 to 1) and Matthew Correlation Coefficient (MCC, from -1 to 1).

Limitations - Some samples have been confused by the pan-cancer classifier, probably because they belong to cancers that developed into near regions of the body.

Isella *et al.* (2017) [2]: our main reference work, on which we provide more details with respect to the others.

Main results - After observing that the subtype CMS4 described in [18] was susceptible of stromal[4] influence, they developed a ColoRectal Intrinsic Subtypes classification (CRIS) that distinguishes different subtypes of CRC (CRIS-A, CRIS-B, CRIS-C, CRIS-D, CRIS-E). This partitioning provides insights on drug sensitivity and prognosis. They developed a dataset-oriented classifier based on Nearest Template Prediction (NTP) and attempted to define a single-sample classifier based on k Top Scoring Pairs (k-TSP); the performances of the k-TSP, however, were suboptimal.

Datasets - For the NTP classifier, the authors used Patient Derived Xenografts (PDX) microarray data, obtained by human metastatic samples transplanted and grown into mices; this type of data has been used both for the identification of the CRIS subtypes and for the training and testing phase of the NTP. To test the classifier, two other independent datasets have been used too: the first with RNA-seq data (from TCGA, made by 450 samples), the second with microarray data. In both cases, good performances have been achieved. Finally, other 14 independent datasets for additional validation were considered. 624 samples from Microarray and RNA-seq data were used for the k-TSP classifier too.

Techniques applied - A consensus based Non-negative Matrix Factorization was applied on Patient Derived Xenografts expression data with varying number of clusters K, each one representing a subtype. The optimal subdivision was reached with K = 5. On this data, 565 genes have been selected as significant for the distinction of subtypes.

The CRIS dataset-oriented classifier is an implementation of the Nearest Template Prediction (NTP) algorithm working on the datasets normalized with z-score. For each subtype, a gene expression template has been defined and, for each sample, the cosine similarity with the five templates is computed. The assigned subtype is chosen considering the highest correlation between the centroids and the z-scores of the samples.

[4] **stroma**: tissue composed of cells that serve as a matrix in which the other cells are embedded [24]

The authors developed also a single-sample classifier based on the k-TSP (k Top Scoring Pairs) algorithm, whose performances, however, resulted to be suboptimal: when comparing two classes, each pair of the TSP algorithm votes to assign a sample to a specific class if the expression of gene A is higher than the one of gene B, or to the other class otherwise. This paradigm has been applied to compare each subtype with the others (10 possible combinations), considering a variable number of non-overlapping genes (highest performances with 80 genes). The subtype was assigned by highest proportion of votes according to the results of the pairwise comparisons. Since the results were poor with respect to the NTP classifier, the latter one was tested again with the reduced set of 80 genes and the similarity of the result increased.

Limitations - The robustness of the classifier (NTP) must be checked both with respect to the features selected and to the samples set. In the first case, this happens because the subtypes were not identified on RNA-seq data, which provide a bigger set of valuable features. In the second case, this happens because the z-score used to normalize the data influences the classification, since it depends on the samples included in the dataset. If the classification is performed on different subsets of the dataset, the result of the normalization changes: this leads to a different classification for some of the samples (which thus cannot be classified univocally). Another limitation lies in the suboptimal performances of the single-sample k-TSP classifier, probably due to the fact that the selected pairs of genes must be in common both to microarray and RNA-seq data to obtain a cross-platform classification.

2.3. Discussion

In this final section we describe overall limits and open issues that emerged from the researches listed above.

First of all, being the data intrinsically heterogeneous, it is difficult to define a stable classification system for subtypes of CRC: the analysis of new genomic features may lead to the identification of different subtypes, as [18] and [2] proved. Secondly, it may happen that a portion of samples remains unclassified ([18], [21], [2]), due to inherent ambiguity of the samples. [20] had to face the availability of only a limited number of samples with desired therapy characteristics: having few samples may lead to a biased model, because the whole population may not be well represented. Another issue regards the heterogeneity in platforms used to retrieve the data, as well as in normalization procedures (which can lead to incomparable data). Finally, in the work of [2] the authors had to sacrifice the performances for the single-sample classifiers in order to use both RNA-seq and microarray data. They showed also that the dataset composition could affect the result when the z-score is used to normalize the data.

Our research will tackle all the described issues focusing on RNA-seq data only to exploit additional gene/features, with the aim of overcoming the discussed limitations in order to provide a step forward towards the clinical application of a cancer subtype classifier.

References

[1] "Key statistics for colorectal cancer." https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html. Accessed: 2020-03-022.

[2] C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, "Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer," *Nature Communications*, vol. 8, p. 15107, 2017.

[3] "Glossary: definitions and phonetic pronounciations." https://www.cancer.org/cancer/glossary.html. Accessed: 2020-03-022.

[4] "What is colorectal cancer?." https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html. Accessed: 2020-03-022.

[5] "Colorectal cancer stages." https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html. Accessed: 2020-03-022.

[6] "Talking glossary of genetic terms." https://www.genome.gov/genetics-glossary. Accessed: 2020-03-027.

[7] "Oncokb cancer gene list." https://www.oncokb.org/cancerGenes. Accessed: 2020-03-031.

[8] "What is epigenetics?." https://ghr.nlm.nih.gov/primer/howgeneswork/epigenome. Accessed: 2020-03-031.

[9] "Definition of next-generation sequencing." https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/next-generation-sequencing. Accessed: 2020-03-024.

[10] "Technologies in molecular diagnostics: microarrays and ngs." https://www.gelifesciences.com/en/us/news-center/trends-in-molecular-diagnostics-part-1-10001. Accessed: 2020-03-029.

[11] W. Z, G. M, and S. M., "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, 2009.

[12] B. Wilhelm and J. Landry, "Rna-seq-quantitative measurement of expression through massively parallel rna-sequencing," *Methods (San Diego, Calif.)*, vol. 48, no. 3, pp. 249–257, 2009.

[13] M. Ghandi and M. Beer, "Group normalization for genomic data," *PloS One*, vol. 7, p. e38695, 2012.

[14] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for rna-seq data analysis," vol. 17, pp. 13–13, 2016.

[15] "Best python libraries for machine learning." https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/. Accessed: 2020-03-030.

[16] "The 20 best r machine learning packages in 2020." https://www.ubuntupit.com/best-r-machine-learning-packages/. Accessed: 2020-03-030.

[17] "Can colorectal polyps and cancer be found early?." https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/detection.html. Accessed: 2020-03-022.

[18] J. Guinney, R. Dienstmann, X. Wang, A. Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B. Bot, J. Morris, I. Simon, S. Gerster, E. Fessler, F. De Sousa E Melo, E. Missiaglia, H. Ramay, D. Barras, and S. Tejpar, "The consensus molecular subtypes of colorectal cancer," *Nature Medicine*, vol. 21, pp. 1350–1356, 2015.

[19] J. Franks, G. Cai, and M. Whitfield, "Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data," *Bioinformatics (Oxford, England)*, vol. 34, p. 1868–1874, 2018.

[20] O. Murcia, M. Juárez, M. Rodríguez-Soler, E. Hernández-Illán, M. Giner-Calabuig, M. Alustiza, C. Egoavil, A. Castillejo, C. Alenda, V. Barberá, C. Mangas-Sanjuan, A. Yuste, L. Bujanda, J. Clofent, M. Andreu, A. Castells, X. Llor, P. Zapater, and R. Jover, "Colorectal cancer molecular classification using braf, kras, microsatellite instability and cimp status: Prognostic implications and response to chemotherapy," *Plos One*, vol. 13, p. e0203051, 2018.

[21] B.-H. Kim, K. Yu, and P. Lee, "Cancer classification of single cell gene expression data by neural network," *Bioinformatics (Oxford, England)*, vol. 36, p. 1–7, 2019.

[22] A. Phipps, P. Limburg, J. Baron, A. Burnett-Hartman, D. Weisenberger, P. Laird, F. Sinicrope, C. Rosty, D. Buchanan, J. Potter, and P. Newcomb, "Association between molecular subtypes of colorectal cancer and patient survival," *Gastroenterology*, vol. 148, pp. 77–87, 2015.

[23] F. Sinicrope, Q. Shi, T. Smyrk, S. Thibodeau, R. Dienstmann, J. Guinney, B. Bot, S. Tejpar, M. Delorenzi, R. Goldberg, M. Mahoney, D. Sargent, and S. Alberts, "Molecular markers identify subtypes of stage iii colon cancer associated with patient outcomes," *Gastroenterology*, vol. 148, pp. 88–99, 2015.

[24] "Tissue | definition, types & facts | britannica." `https://www.britannica.com/science/tissue#ref163759`. Accessed: 2020-03-031.