Feature Selection Algorithm Recommendation for Gene Expression data with Meta Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Feature selection is an important step in gene expression data analysis. However, many feature selection methods exist and a costly experimentation is usually needed to determine the most suitable one for a given problem. This paper presents the application of gradient boosting and neural network techniques for the construction of meta-models that can recommend rankings of {feature selection - classification} algorithm pairs for new gene expression classification problems through the usage of learning-to-rank and collaborative filtering approaches. Results in a corpus of 60 public datasets show the superiority of these techniques in producing more useful rankings in relation to classical meta-models.

1 Motivation and research problem

The current gene expression profiling technologies have generated great hopes for the construction of early diagnosis and prognosis systems for cancer and other diseases. However, one of the major difficulties for achieving clinically acceptable accuracies is the high dimensionality of the feature space (genes) relative to the number of samples [1], which is usually handled by applying feature selection techniques. However, many feature selection methods exist and none is clearly superior in the domain of gene expression data [1], which entails high experimentation times and computational resources. In order to circumvent this, a Meta-Learning (MtL) approach is proposed, aiming to construct predictive models (metamodels) that relate characteristics of datasets (metafeatures) to performance of algorithms so they can be used to recommend algorithms for unseen problems.

2 Technical contribution

This work follows the general MtL scheme in [2] (Fig. 1). The data characterization module extracts metafeatures that describe each dataset from the training repository. The evaluation module assess the performance scores of each considered algorithm for each dataset, which are then converted to rankings. A machine learning algorithm can then be used to induce a metamodel with the metafeatures as input variables and the ranking as target variables, so that it can be used at testing time to predict a ranking of algorithms for a new problem, based on the metafeatures of the input dataset.

The first evaluated method is the LightGBM (LGBM) algorithm [3], an ensemble of gradient boosting decision trees, with the LambdaRank pairwise loss function, which has shown successful results in real world ranking problems [4], because it allows to optimize the Normalized Discounted Cumulative Gain (NDCG) metric. In order to find the best configuration of parameters of the estimators and the training procedure, hyperparameter tuning is performed with Bayesian Optimization [5].

As an alternative ranking metamodel we propose a neural network architecture inspired on a matrix factorization approach, which is widely used in collaborative filtering. The architecture is illustrated in Fig. 1, where the feature selection method is transformed to a dense representation through an

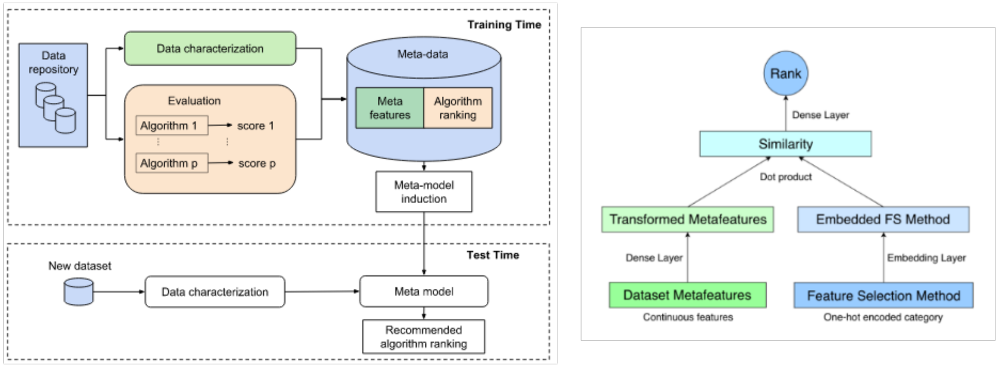

Figure 1: (Left) Metalearning for algorithm recommendation. (Right) Neural network architecture.

embedding layer, while the metafeatures, already in a continuous representation, are transformed to the same latent space through a dense layer with a nonlinear activation. Once in the latent space, both representations are combined with a dot product to ensure that they share the same latent space. As a baseline we also evaluate the classic K-Nearest Neighbors (KNN) algorithm as a ranking recommendation method [6] and the average ranking.

For the experiments of this work we used a collection of 60 public gene expression datasets derived from different cancer-related studies. Each dataset was evaluated with every combination of 4 feature selection algorithms and 3 classification methods. The average Gmean was used as a score of each combination, which was used to construct the target ranking. As metafeatures we used 12 common statistics and based on information theory measures [2] which we expanded to 51 using the framework for systematic development of metafeatures proposed by [7].

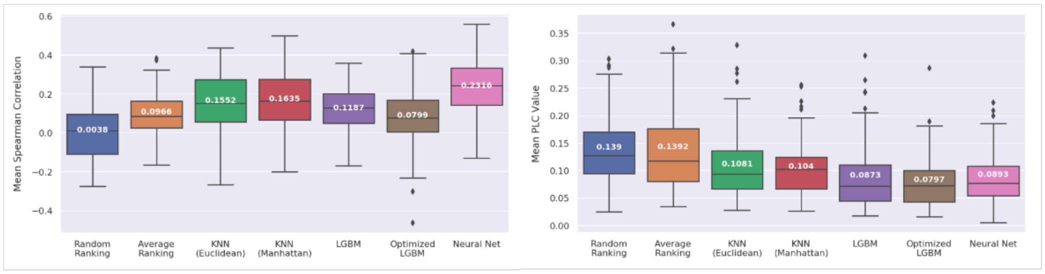

Figure 2: Spearman correlation and PLC results of different metamodel induction methods. The white labels display the mean value.

For evaluation, we use the Spearman correlation coefficient [2], which assesses the overall proximity of the estimated ranking w.r.t. the ideal ranking, and the performance loss curve (PLC) metric [8], which evaluates how useful the ranking is (in terms of accuracy) if the algorithms are evaluated in the ranking order. We observe that the neural network metamodel presents the best Spearman scores (Fig. 2), followed by KNN and LGBM metamodels, while the LGBM and neural network metamodels tend to offer similar PLC scores. However, the LGBM optimized version improves slightly the average PLC results. It is important to note that the PLC metric is a more useful measure for the intended task than the Spearman coefficient, since it gives us an idea of how much we can gain or lose in accuracy if we follow the recommended ranking to build the classifiers. The Spearman coefficient evaluates the overall proximity of the inferred ranking to the ideal one, giving the same weight to errors in the higher or lower part of the ranking and without worrying about the predictive accuracy of the base-level models.

As part of the ongoing work, we are currently evaluating more ranking approaches, such as CofiRank [9], as well as combining the two studied approaches by using the neural collaborative filtering model with pairwise loss functions, such as WARP [10]. Furthermore, we are aiming to evaluate the effectiveness of these approaches with general domain datasets, such as the StatLog [11], OpenML [12] and AutoML [13] dataset repositories.

References

[1] Abhishek Bhola and Shailendra Singh. Gene selection using high dimensional gene expression data: An appraisal. *Curr. Bioi.*, 13:225–233, 2018.

[2] Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, and Ricardo Vilalta. *Metalearning: Applications to Data Mining*. Springer, 1 edition, 2008.

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Adv. in Neural Inf. Proces. Systems 30*, pages 3146–3154. 2017.

[4] Fajie Yuan, Guibing Guo, Joemon M. Jose, Long Chen, Haitao Yu, and Weinan Zhang. Lambdafm: Learning optimal ranking with factorization machines using lambda surrogates. In *25th ACM CIKM '16*, pages 227–236, 2016.

[5] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2960–2968, 2012.

[6] Bruno Feres de Souza, Carlos Soares, and André C.de Carvalho. Meta-learning approach to gene expression data classification. *Int. J. of Intel. Comp. and Cyber.*, 2(2):285–303, 2009.

[7] Fábio Pinto, Carlos Soares, editor="Bailey James Mendes-Moreira, João", Latifur Khan, Takashi Washio, Gill Dobbie, Joshua Zhexue Huang, and Ruili Wang. "towards automatic generation of metafeatures. In *PAKDD 2016*, pages 215–226, 2016.

[8] Salisu Mamman Abdulrahman, Pavel Brazdil, Jan N. Van Rijn, and Joaquin Vanschoren. Algorithm selection via meta-learning and sample-based active testing. In *MetaSel'15*, pages 55–66, 2015.

[9] Markus Weimer, Alexandros Karatzoglou, Quoc V Le, and Alex J Smola. Cofi rank-maximum margin matrix factorization for collaborative ranking. In *Advances in neural information processing systems*, pages 1593–1600, 2008.

[10] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI'11*, pages 2764–2770, 2011.

[11] Ross D. King, Cao Feng, and Alistair Sutherland. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal*, 9(3):289–333, 1995.

[12] Mustafa Mısır and Michèle Sebag. Alors: An algorithm recommender system. *Artificial Intelligence*, 244:291–314, 2017.

[13] Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, et al. Analysis of the automl challenge series 2015–2018. In *Automated Machine Learning*, pages 177–219. Springer, 2019.