Non-coding RNA Sequences Identification and Classification Using a Multi-class and Multi-label Ensemble Technique

Michalis Stavridis, Aigli Korfiati, Georgios Sakellaropoulos, Seferina Mavroudi, Konstantinos Theofilatos

To cite this version:

Michalis Stavridis, Aigli Korfiati, Georgios Sakellaropoulos, Seferina Mavroudi, Konstantinos Theofilatos. Non-coding RNA Sequences Identification and Classification Using a Multi-class and Multi-label Ensemble Technique. 14th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), May 2018, Rhodes, Greece. pp.179-188, 10.1007/978-3-319-92016-0_17 . hal-01821313

HAL Id: hal-01821313

https://inria.hal.science/hal-01821313

Submitted on 22 Jun 2018

Non-coding RNA sequences identification and classification using a multi-class and multi-label ensemble technique

Michalis Stavridis[1], Aigli Korfiati[1, 2], Georgios Sakellaropoulos[1], Seferina Mavroudi[2,3] and Konstantinos Theofilatos[2, *]

[1]School of Medicine, University of Patras, Patra, Greece {michailstavridis@yahoo.com, gsak@med.upatras.gr}

[2]InSyBio Ltd, Winchester, United Kingdom {a.korfiati@insybio.com, k.theofilatos@insybio.com, s.mavroudi@insybio.com}

[3]Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece, Patra, Greece

Abstract. High throughput sequencing RNA-sequencing technologies and modern in silico techniques have expanded our knowledge on short non-coding RNAs. These sequences were initially split into various categories based on their cellular functionality and their sequential, thermodynamic and structural properties believing that their sequence can be used as an identifier to distinguish them. However, recent evidence has indicated that the same sequences can act and function as more than one type of non-coding RNAs with a striking example of mature microRNA sequences which can also be transfer RNA fragments. Most of the existing computational methods for the prediction of non-coding RNA sequences have emphasized on the prediction of only one type of noncoding RNAs and even the ones designed for multiclassification do not support multiple labeling and are thus not able to assign a sequence to more than one non-coding RNA type. In the present paper, we introduce a new multilabel- multiclass method based on the combination of multiobjective evolutionary algorithms and multi-label implementations of Random Forests to optimize the feature selection process and assign short RNA sequences to one or more non-coding RNA types. The overall methodology clearly outperformed other machine learning techniques which were used for the same purpose and it is applicable to data coming from RNA-sequencing experiments.

Keywords. Multi-label classification; non-coding RNAs; multi-objective optimization; random forests; dimensionality reduction

1 Introduction

Non-coding RNAs (ncRNAs) are RNA fragments that are not translated to proteins [1]. Small ncRNAs typical size is of 18-35 nucleotides (nt) and long-non-coding RNAs (lncRNAs) can be more than 200 nt (e.g., enhancer RNA -- eRNA) [2]. The transfer RNA (tRNA) or the ribosomal RNA (rRNA) have been the subject of several studies and this has established their well-accepted functional roles in cells. However, over the past few years, advances on sequencing biotechnologies and other improvements on experimental protocols led to the elucidation of more ncRNA categories, the most prominent being: microRNAs (miRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), small inferring RNAs (siRNAs), piwi-interacting RNAs (piRNAs), ti-RNAs, spli-RNA, tRNA fragments (tRfs) and others to be identified [3].

Recently, transcriptomics analysis using RNA-seq data has also become the state-of-the-art procedure for identifying and annotating functionally ncRNA molecules. However, since ncRNA transcripts are drastically different from mRNAs, the majority of 'general-purpose' bioinformatics programs face limitations and are not well suited for discovering effectively ncRNAs from RNA-seq [4]. To mitigate this problem, several computational methods have been tailored to the needs of ncRNA data analysis.

NcRNAs were initially split into various categories based on their cellular functionality and their sequential, thermodynamic and structural properties believing that their sequence can be used as an identifier to distinguish them. However, recent evidence has indicated that the same sequences can act and function as more than one type of ncRNAs with a striking example of mature microRNA sequences which can also be transfer RNA fragments. Most of the existing computational methods for the prediction of non-coding RNA sequences have emphasized on the prediction of only one type of ncRNAs [5] and even the ones designed for multiclassification, do not support multiple labeling and are thus not able to assign a sequence to more than one non-coding RNA types.

For each one of the aforementioned short non-coding RNA types, a variety of computational methods exists for their prediction [6]. These are split into the computational methods which are designed to be specific to one type of short ncRNAs and the ones which can predict more than one type of short ncRNAs [7]. The first category of methods presents increased classification performances but their applicability is limited as they cannot be used to analyze on a single run a large transcriptomic dataset and to predict different types of short ncRNAs. Moreover, most of the methods belonging to the second category exclude from their analysis significant types of short ncRNAs, while others are based on data mining in existing repositories and thus they cannot extend the current knowledge on short ncRNAs. Finally, most of these methods use the same features for all the different types of short ncRNAs. Another important drawback of existing methodologies is that they consider the different types of non-coding RNAs as separate forcing every RNA sequence to be classified in only one type of non-coding RNAs. However, as already mentioned, recent evidence has proven that this does not hold since tRNA fragments can have the same sequence as miRNAs [8]. If we add this fact to the previously known one that several types of non-coding RNAs are generated from pruning other types of non-coding RNAs, as in the case of pre-miRNAs and mature miRNAs, then a need for a multilabel computational method is raised to treat these data effectively.

In the present paper, we introduce a new multilabel, multiclass method called Multi-label GARF, which combines multi-objective evolutionary algorithms with multi-label Random Forest implementations. In particular, it uses a Pareto-based multi-objective optimization to select the optimal subset of features to be used as inputs, to select the most suitable implementation of Random Forests for every dataset and to optimize its parameters. This optimization process is being guided by 7 fitness functions which are evaluating the solutions based on the classification performance of the Random Forest models extracted from them, and based on their simplicity in terms of the number of selected inputs and their number of random trees. The multi-objective optimization framework, by design avoids local optimal solutions promoting the optimal exploration of the search space.

For the problem of classifying the RNA sequences to non-coding RNA types, for the purposes of the present manuscript, a new dataset was constructed with pre-miRNAs, mature miRNAs, snoRNAs, tRNAs, tRFs, rRNAs, pseudo hairpins and random RNA sequences. All these sequences were pairwise compared to locate similar sequences and for some of them multiple labels were assigned by this process. For all the sequences of the dataset, 58 sequential, thermodynamical and structural features were calculated including most significant features from existing non-coding RNA classification methods. The proposed solution was applied on this dataset and its performance was compared with existing state-of-the-art multi-label methods. Multi-label GARF significantly outperformed other methods in terms of classification performance on the testing dataset. Its performance surpassed 60% of the very strict multi-label accuracy metric which considers a sample to be classified correctly only if all its labels have been predicted correctly. It is noteworthy than none of the random RNA sequences were assigned to any of the non-coding RNA types and this makes the final predictive models suitable for screening RNA-seq reads for non-coding RNA sequences.

2 Materials and Methods

RFAM database [9] was used to download mature miRNA (1865), pre-miRNA (2547), tRNA (12522), rRNA (25723) and snoRNA (12522) sequences. Moreover, tRNA fragments (tRFs) were downloaded from MINTbase [10] and tRFdb [11]. In order to train and test effectively the machine learning classifiers, the random undersampling method was applied randomly selecting only 1865 sequences of each category to be included in the final dataset since this is the plurality of the minority class. The produced dataset was extended with 1865 pseudo hairpin sequences constructed following the method described in [12] and 1865 random RNA-sequences of lengths from 20 to 200.

These sequences were pairwise compared to identify equal or similar sequences allowing total 2 mismatches belonging to more than one categories or sequences which can be found within other sequences. This analysis was conducted to assign multiple labels to these types of sequences.

As a next step for all of these sequences 58 structural, sequential and thermodynamic features were calculated using InSyBio ncRNAseq tool [13]. The final dataset was split in training and testing set with 2/3 of its sequences being assigned to the training set and the remaining 1/3 to the test set. This split was conducted randomly but reassuring that 1:1 proportion is maintained for every class in the dataset.

The proposed designed and implemented method is a hybrid method which solves on parallel the problems of dimensionality reduction and multiple labels classification. In specific the proposed method, Multi-Label GARF, is an ensemble dimensionality reduction technique which utilizes a multiobjective evolutionary algorithm [14] for the identification of the optimal feature subset to be used as input to the classifiers as well as for the selection of the most suitable type of Random Forests classifier [15] to be used and its optimal parameters.

The multi-objective evolutionary framework used for this implementation was based on the multi-objective evolutionary algorithm initially applied in [16] for the optimization of the preprocessing pipeline for the analysis of Mass Spectrometry data. This is based on a Pareto-optimization technique to allow fast convergence to good exploration properties and effective handling of the contradictory goals of minimizing the number of used features, maximizing the accuracy of the classifiers and minimizing the complexity of the classifier to achieve better generalization properties.

The algorithm starts by randomly initializing a first population of solutions which are represented as float vectors. These vectors consist of i) float variables for the parameter number of random trees to be used and the minimum number of samples assigned per leaf to control the splitting process on Random Forests, ii) a float variable to choose between the two alternative multi-label Random Forest algorithms (a value greater than 0.5 indicates the selection of method 2 and a value less than 0.5 indicates the selection of method 1) and 58 float variables for deciding if a feature will be selected as input or not (values greater than 0.5 forces a feature to be used as input). The float vectors of the population are initialized randomly with values from the normal distribution with mean equal to *min_value+ (max_value-min_value)/2* and variance *(max_value-min_value)/2*, where *max_value* is the maximum allowed value of an optimization variable and *min_value* is its minimum allowed value.

For the variation of the population of solutions in order to create new solutions, crossover and mutation operators are sequentially applied. Regarding the crossover operators, two crossover methods are used with probabilities which are also provided by the user (45% probability for two-point crossover operator, 45% probability for arithmetic crossover and 10% for not applying crossover operator were used for this implementation). For the mutation, the Gaussian mutation operator is applied since it is the most suitable operator for the float representation scheme which is adopted in the proposed algorithm with mutation probability 1%.

To better handle the multiple objectives of the current problem, the selection process was based on a multi-objective optimization method. The first step is to calculate the number of Pareto frontiers (sets of solutions where no solution is better than other solutions in the same set to all optimization goals). To calculate the Pareto frontiers, the efficient and fast solution described in [17] was used. An initial fitness value is then assigned to every solution equal to the reverse order of the parent front to which it has been assigned by the previous step. Next, the method calculates solution niches by grouping together solutions according to their similarity. The fitness values of every solution in a given niche are divided by the variable m (average similarity of every solution) which is calculated by performing pairwise comparisons of all solutions of the niche calculating their geometrical distances and calculating the mean pairwise distance of them. These fitness values are then tuned according to the number of solutions belonging to each niche. Roulette Wheel Selection is then used to select the population of the next generation. The best solution passes as it is to the next generation.

The total fitness value is calculated with the weighted sum of the optimization goals with the weights pre-defined by the user. The specific fitness functions (FF) which were used for multi-label GARF algorithm were the following:

$FF1 = \frac{1}{1 + \text{Number of selected features}}$
$FF2 = \text{Classification Accuracy}$
$FF3 = \text{Hamming Loss Classification Metric}$
$FF4 = \frac{1}{1 + \text{Number of trees used by Random Forests}}$
$FF5 = \frac{\text{Average number of samples per node Split in Random Trees}}{\text{Total Number of samples}}$
$FF6 = \text{Recall Classification Metric}$
$FF7 = \text{Precision Classification Metric}$

Fitness function 1 aims at the minimization of the selected features in order to increase the interpretability of the classifier. Fitness functions 4 and 5 were used to promote solutions which lead to simpler models to present better generalization properties. The rest of the fitness functions are employed to increase the algorithm's classification performance.

The algorithm was terminated when the population of solutions is deemed as converged (the similarity among the solutions surpasses a predefined threshold) or when the maximum number of generations is reached.

The multi label Random Forest classifiers implementation was based on sklearn python library [18] using two different approaches regarding the function used to measure the quality of a split in the process of generating the trees of random forests. The first method is called Gini and uses the Gini impurity function [19] and the second method uses the information gain function [20].
The overall pipeline of the multi-label GARF is depicted in Figure 1.

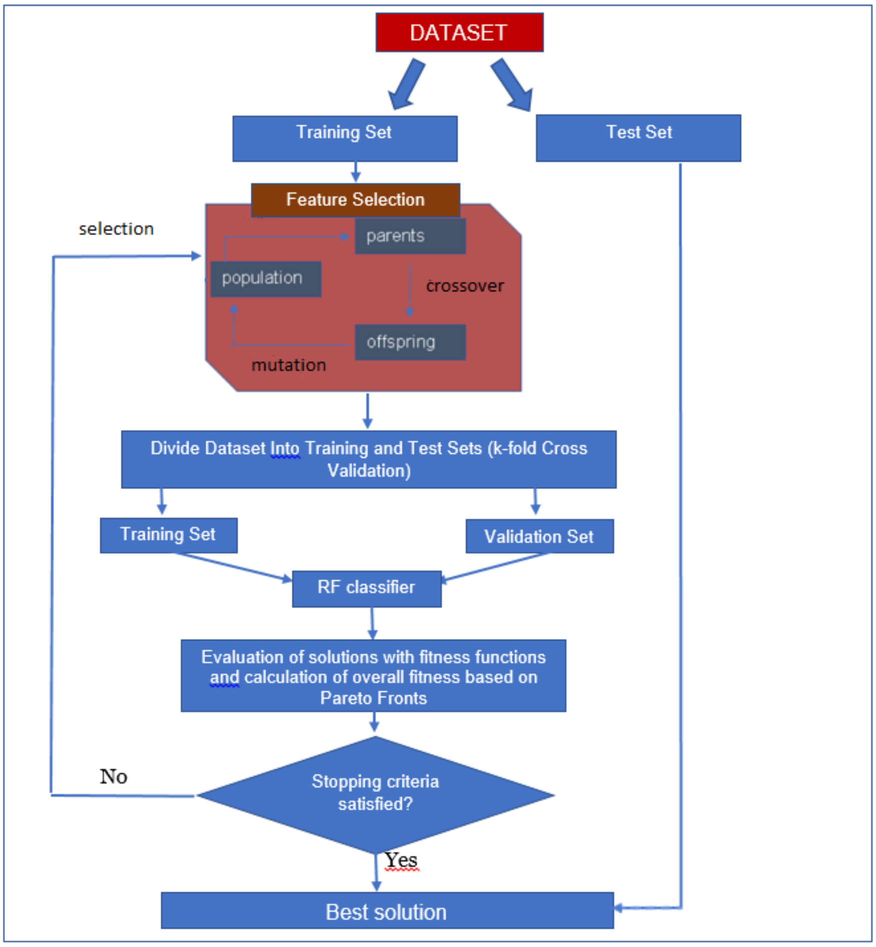
DATASET
Training Set
Test Set
Feature Selection
selection
parents
population
crossover
offspring
mutation
Divide Dataset Into Training and Test Sets (k-fold Cross Validation)
Training Set
Validation Set
RF classifier
Evaluation of solutions with fitness functions and calculation of overall fitness based on Pareto Fronts
No
Stopping criteria satisfied?
Yes
Best solution

Fig. 1. Flowchart of Proposed Method: Multi-Label GARF

3 Experimental Results

All features of the dataset were arithmetically normalized to the interval [-1, 1] and missing values were imputed using KNN-Impute method [21]. Then the Multi-label GARF and other multi-label methods were used to train multi-label classification models using the training dataset. The trained classification models were applied on the testing dataset and Table 1 presents the Accuracy and Hamming-Loss metric for every model.

The state of the art multi-label models which were used for comparison reasons were the Random Forest implementation of sklearn library, the extra trees algorithm [22], the multi-label KNN [23], a Decision Tree algorithm [24] and the Binary Relevance algorithm [25]. For all these algorithms, all features were used as inputs and they were tested using their default parameters.

Regarding the Multi-label GARF algorithm, after extensive testing in the training set, a population of 50 individuals was used and a maximum number of 200 generations was set for the termination criterion. Further increase in the number of generations did not improve the results. Mutation Probability was set to 1%. Moreover, during its training phase internal 5-fold cross validation was applied using the training dataset. In order to take into account, the stochastic nature of this method we run the experiments 25 times and the results of Table 1 are the mean values together with their standard deviation (SD). The goal significances for the fitness functions 2, 3, 6, 7 were set to 1.5 and for the other goals to 1. By this way we are setting for the final ordering of the solutions, classification performance to be twice as significant as the goals of minimizing the classification model's complexity.

Table 1. Comparative results on the classification of non-coding RNA sequences using the proposed technique and other existing machine learning techniques compatible with multi-label datasets.

Method	Accuracy (mean - SD)	Hamming-Loss Metric (mean - SD)
Multi-label GARF	**61% - 0.3461**	**0.051 – 0.002529**
Random Forest	28% - 0.0002	0.010 – 0.000061
Extra Trees	12% - 0.0002	0.009 – 0.000062
Multi Label KNN	1% - 0.0008	0.010 – 0.000006
Decision Tree	29% - 0.0004	0.010 – 0.000007
Binary Relevance	22% - 0.0005	0.012 - 0.000005

The best solution uncovered by Multi-label GARF method among all runs used the Gini metric for splitting criterion on the random tree nodes, 8637 number of random trees, 3 as minimum number of samples assigned per leaf and 27 features were selected. Moreover, a closest examination on the results indicated that none of the random RNA sequences was assigned to any of the non-coding RNA types by the multi-label GARF method.

Multi-label GARF significantly outperformed the other examined methods in terms of classification performance on the testing dataset. Its performance was over 60% of the very strict multi-label accuracy metric which considers a sample to be classified correctly only if all its labels have been predicted correctly. None of the other examined methods presented accuracy more than 30%.

Finally, it is noteworthy than none of the random RNA sequences were assigned to any of the non-coding RNA types and this makes the final predictive models suitable for screening RNA-seq reads for non-coding RNA sequences.

4 Discussion

Predicting and classifying short ncRNAs is of crucial importance for systems biology and translational medicine to: i) allow the prediction of new ncRNA molecules for human or other not well studied organisms (including the human microbiome), ii) ease the analysis of high throughput sequencing experiments by allowing for the efficient identification and quantification of non-coding RNAs without the need of analyzing them separately by modifying samples with specialized libraries and iii) enable the understanding of ncRNAs functionality. To the best of our knowledge, most existing methods are offering prediction of very limited number of non-coding RNA categories with most of them emphasizing on miRNAs. Moreover, all existing methods do not take into account the fact that some sequences may be assigned to more than one ncRNA categories while some ncRNA sequences are being generated from pruning other ncRNAs.

In the present work, we have attempted to overcome these difficulties by proposing a unified multi-label multi-classification algorithmic framework which combines multi-label Random Forests with a multi-objective optimization algorithm to find the optimal classification model with the minimum number of inputs. To test the proposed solution, we generated the first multi-label dataset for non-coding RNAs mining information from several databases and compared the proposed solution with other state-of-the-art multi-label classification models. The multi-label GARF clearly outperformed all other methods in both classification metrics used. Moreover, on this specific dataset none random RNA sequence was assigned to any type of non-coding RNAs. Furthermore, the additional labels which are assigned by it in some of the sequences which maintained the strict metric of multi-label accuracy in approximately 60% should be further explored in the future as meaningful information could reside on them about other functionalities that could be performed by the same sequences or parts of them.

The fact that the proposed technique was able to train models that can predict non-coding RNAs within longer RNA sequences makes it appropriate for being applied directly on the reads extracted from deep RNA-sequencing experimental techniques. As an example, in the case of identifying miRNAs from RNA-sequencing, a common problem is whether to search for pre-miRNAs or for mature miRNAs since both types of RNA sequences can be detected in a biosample. The extracted predictive models from the present paper can solve efficiently this problem as well as the problem that non-coding RNAs are most of the time only a subset of a read being generated by sequencing technologies. Thus, an interesting future direction includes testing and validating the performance of the proposed method directly on sequences exported from RNA-sequencing experiments. Moreover, despite the promising results of the proposed model, its performance can be further improved by including even more features as potential inputs (e.g. existence of clover structure). Finally, the current implementation can be expanded to allow for the prediction of other types of non-coding RNAs such as scRNAs and diRNAs.

References

1. Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. BioMed Research International, 2010.
2. Kang, W., & Friedländer, M. R. (2015). Computational pre-diction of miRNA genes from small RNA sequencing data. Frontiers in bioengineering and biotechnology, 3, 7.
3. Veneziano, D., Di Bella, S., Nigita, G., Laganà, A., Ferro, A., & Croce, C. M. (2016). Noncoding RNA: current deep sequencing data analysis approaches and challenges. Human mutation, 37(12), 1283-1298.
4. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17(1), 13.
5. Li, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q., & Shen, B. (2012). Perfor-mance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. Nucleic acids research, 40(10), 4298-4305.
6. Aghaee-Bakhtiari, S. H., Arefian, E., & Lau, P. (2017). miRandb: a resource of online services for miRNA research. Briefings in bioinformatics, bbw109. https://doi.org/10.1093/bib/bbw109
7. Washietl, S., Will, S., Hendrix, D. A., Goff, L. A., Rinn, J. L., Berger, B., & Kellis, M. (2012). Computational analysis of noncoding RNAs. Wiley Interdisciplinary Reviews: RNA, 3(6), 759-778.
8. Venkatesh, T., Suresh, P. S., & Tsutsumi, R. (2016). tRFs: miRNAs in disguise. Gene, 579(2), 133-138.
9. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic acids research, 33(suppl_1), D121-D124.
10. Pliatsika, V, Loher, P, Magee, R, Telonis, AG, Londin, E, Shigematsu, M, Kirino, Y, Rigoutsos, I. MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all

The Cancer Genome Atlas projects Nucleic Acids Res. 2017; PubMed PMID:29186503

11. Kumar, P., Mudunuri, S., Anaya, J., & Dutta, A. (2014). tRFdb: a database for transfer RNA fragments. Nucleic Acids Research (Database Issue) doi:10.1093/nar/gku1138.
12. Kleftogiannis, D., Theofilatos, K., Likothanassis, S., & Mavroudi, S. (2015). YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 12(5), 1183-1192.
13. Korfiati, A., Theofilatos, K., Alexakos, C., & Mavroudi, S. (2017). InSyBio ncRNASeq: A web tool for analyzing non-coding RNAs. EMBnet. journal, 23, e882.
14. Abraham, A., and Jain, L. (2005) Evolutionary Multiobjective Optimization. In: Abraham, A., Jain, L., and Goldberg, R., eds. Evolutionary Multiobjective Optimization: Theoretical Advances and Applications, pp. 1-6, Springer London, London
15. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
16. Corthesy, J., Theofilatos, K., Mavroudi, S., et al. (2017). An adaptive pipeline to maximize isobaric tagging data in large-scale MS-based proteomics. Journal of Proteome Research, Under Second Review at February 2018.
17. Mishra, K. K., and Harit, S. (2010) A Fast Algorithm for Finding the non Dominated Set in Multiobjective Optimization. Int. J. Comput. Appl. 1, 35-39
18. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed on December 2017)
19. Modarres, R., Gastwirth, J., (2006). A Cautionary Note on Estimating the Standard Error of the Gini Index of Inequality. Oxford Bulletin of Economics and Statistics. 68 (3): 385–390. doi:10.1111/j.1468-0084.2006.00167
20. Liu, F., Zhang, X., Ye, Y., Zhao, Y., Li, Y. (2015) MLRF: Multi-label Classification Through Random Forest with Label-Set Partition. In: Huang DS., Han K. (eds) Advanced Intelligent Computing Theories and Applications. ICIC 2015. Lecture Notes in Computer Science, vol 9227. Springer.
21. Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software, 85(11), 2541-2552.
22. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42
23. Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048
24. Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. Machine Learning, 73(2), 185
25. Luaces, O., Díez, J., Barranquero, J., del Coz, J. J., & Bahamonde, A. (2012). Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence, 1(4), 303-313.