CMSC 635 Project

Spring 2018
(Group work; 25 pts)

The project asks you to develop, evaluate and compare models for the prediction of proteins that interact with DNA and RNA using a provided dataset. Your model must classify a given protein sequence into one of four outcomes, i.e., interacts with DNA (DNA), interacts with RNA (RNA), interacts with both DNA and RNA (DRNA), and does not interact with DNA or RNA (nonDRNA). Although each group will solve the same task, the corresponding designs should be unique, i.e., collaboration between groups is not allowed.

Datasets

Two datasets are/will be provided:

- *training.csv* (*training dataset*) that includes 391 DNA proteins, 523 RNA proteins, 22 DRNA proteins, and 7859 nonDRNA proteins, for the total of 8795 proteins.
- *test.csv* (*blind test dataset*) that includes 8795 proteins, with similar proportions between the four classes of proteins. This is an independent test set, which means that entire design procedure (including feature generation, feature selection, parameterization and selection of classifiers, etc.) should be completed using only the training dataset. The test dataset should be used to evaluate your system only once. This dataset will be posted on the class web site 4 days before the project submission deadline and it will **not** include the annotation of the outcomes. You will have to predict the outcomes and the instructor will process and assess these predictions.

The training dataset is provided in the comma-separated format where each protein is represented by:

- an amino acid sequence
- the outcome/class encoded as DNA, RNA, DRNA, and nonDRNA

The test dataset is in the same format as the training dataset, except that the outcome is not provided.

Evaluation of Predictions

You are required to perform the 5-fold cross validation when using the *training dataset*. This cross validation divides the training dataset into 5 random, equal-size subsets, where one subset is used to test the prediction model and the remaining four to train/develop the prediction model; this is repeated 5 times, each time using a different subset as the test set. Consequently, this test results in predicting every sequence in the training dataset. This test procedure is supported by RapidMiner.

For each of the four outcomes you will convert the dataset into a binary problem, i.e., a given outcome (positive outcome) vs. all other outcomes (negative outcomes). For example, all proteins that are labeled as "material production failed" will be considered as positive, and the remaining proteins (purification failed, crystallization failed, and crystallizable) as negative. Next, for each of the four outcomes you will compute the following measures

$$Sensitivity = SENS = 100*TP / (TP + FN)$$
$$Specificity = SPEC = 100*TN / (TN + FP)$$
$$PredictiveACC = 100* (TP+TN) / (TP+FP+TN+FN)$$
$$MCC = (TP*TN – FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]$$

where TP is the number of true positives (correctly predicted positive outcomes), FP denotes false positives (negative outcomes that were predicted as positives), TN denotes true negatives (correctly predicted negative outcomes), FN stands for false negatives (positive outcomes that were predicted as negatives). You will also compute:

$$averageMCC = (MCC_{DNA} + MCC_{RNA} + MCC_{DRNA} + MCC_{nonDRNA})/4$$

accuracy = 100*TP_{all} / (number of all protein in the dataset)

where MCC_{DNA}, MCC_{RNA}, MCC_{DRNA}, and $MCC_{nonDRNA}$ denote the MCC values when using the DNA, RNA, DRNA, and nonDRNA outcomes as the positives, and TP_{all} is the number of correctly predicted outcomes (DNA proteins predicted as DNA proteins, RNA proteins predicted as RNA proteins, etc.). These measures can be computed based on the confusion matrix. You should **round the values** to one digit after the decimal point when reporting the accuracy, sensitivities, and specificity and to three digits after the decimal point when reporting MCC. **You report should also include the confusion matrix for your final solution**.

You must also provide and summarize predictions on the *blind test dataset*. To do that you will compute your model using the entire training dataset (using the same design, i.e., features, values of parameters, etc., as in your best 5 fold cross validation result) and you will use this model to predict sequences from the blind test dataset. In your report, you must discuss the corresponding results on both the training and blind test dataset; on the blind test dataset you can summarize your results by explaining and comparing how many proteins were predicted with a given outcome.

Design

You need to **design** your predictive model to maximize its predictive performance **evaluated based on averageMCC using the 5 fold cross validation on the training dataset**. The design may consider:

- Use of different features to encode the input protein sequence. The data mining algorithms require a rectangular dataset with a fixed size and structure of the feature vector for each object (protein). Thus, you will need to convert the input protein sequences (that have variable length) into a fixed set of (numerical) features.
- Selection of a subset of the input features. This could potentially speed up computation of the model, remove weak/noisy features, and reduce overfitting. Feel free to combine results of multiple feature selection methods.
- Selection of the classification algorithm that you will use to compute your model from among many algorithms that are available in RapidMiner.
- Parametrization of the selected classification algorithm(s). This involves setting values of their key parameters.
- Consideration of how to perform the prediction. There are at least two alternatives: use one model to predict all 4 classes vs. use 4 models to predict each of the four classes. In the latter case, you will have to combine the four results to select one "best" result for each protein. The advantage of the second approach is that you can choose different subsets of features and different classification algorithms and their parameters for each outcome/class.
- Build a system with multiple models that are used together. For instance, you could use multiple models that predict all 4 classes and combine their results together to generate one prediction. Check the methods in RapidMiner at Operators → Modeling → Predictive → Ensembles.

NOTE 1: Ensure that you perform all design activities (e.g., feature selection) using the 5-fold cross validation on the training dataset. Otherwise you could overfit this dataset and your results on the test dataset could suffer.

NOTE 2: In your report, you should clearly indicate **one** best set of results, which must be selected based on the cross validation results on the training datasets. Moreover, these results should be compared with your intermediate results (earlier designs, other alternatives, etc.) and with results of the existing method, see Table 1, to justify your design choices. **In your write up, report your results by adding them into Table 1 to make it easy to compare the various results; indicate which result is the best/final**. You are not expected to necessarily outperform the baseline results from Table 1, but you should provide a convincing argument why and how your method is good/competitive.

Table 1. Formatting of the predictive results based on the 5-fold cross validation on the training dataset (this table is available in the Blackboard).

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	8.9				
	Specificity	99.9				
	PredictiveACC	95.6				
	MCC	**0.255**				
RNA	*Sensitivity*	44.1				
	Specificity	99.0				
	PredictiveACC	96.0				
	MCC	**0.549**				
DRNA	*Sensitivity*	0.0				
	Specificity	100.0				
	PredictiveACC	99.7				
	MCC	**0**				
nonDRNA	*Sensitivity*	99.1				
	Specificity	28.7				
	PredictiveACC	91.6				
	MCC	**0.443**				
averageMCC		**0.312**				
accuracy		91.4				

Deliverables

Each group shall provide the following two deliverables:

1. **Report** that consists of:
   - **Cover page** that gives the class number and title, date of your submission, name of your group and names of all team members.
   - **Description of the design of the prediction system**. You should briefly explain your features; how and which were selected; which classification algorithms and their parameters you tried and why and which you have chosen; and which other design options you considered and applied.
   - **Results** (see *Evaluation of Predictions* section). You must organize the results in a table using the format of Table 1. Using this format, compare your best results with the results from earlier/alternative designs and with the results shown in Table 1.
   - **Conclusions**. This is a **very important part** of your report. You should comment on the quality of your results and compare them against the results from Table 1. Also, describe your experience in this project, and explain advantages and disadvantages of your method and why you think your results are good or bad.
2. **Predictions on the *blind test* dataset**. These predictions should be submitted via email to lkurgan@vcu.edu as a text file named with the name of your group, where each row provides prediction for a given "blind" protein. The format should be as follows:

   DNA
   DNA
   RNA
   nonDRNA
   …

   where DNA, RNA, DRNA and nonDRNA are the predicted outcomes for the protein from the same row in the *test.csv* file. The instructor will use these results to evaluate your method on the blind test dataset against the true classes, and these results will be forwarded to you as part of the evaluation of your project.

Marking

The evaluation of the project report and predictions constitutes 25% of the final mark from the course and includes the following three parts:

1. 30% for the quality of the report
2. 20% for the quality of the design of the prediction method
3. 50% for the quality of the predictions measured using the 5 fold cross validation on the training dataset and on the blind test dataset.

NOTE 3: The *averageMCC* is of the primary importance for part 3. This is the main quality measure that will be use to rank and evaluate all submitted solutions, but the conclusions **must discuss** the other quality indices as well. MCC that is high(er) relative to other submissions or results in Table 1 is not necessary to receive a full mark. However, it is important to secure and explain the MCC value of your best design that is higher relative to your own alternative solutions, and to highlight and explain advantages provided by your best design when compared to the results in Table 1 and your other alternatives.

NOTE 4: **Bonuses of 20%, 15%, and 10%** of the achieved project mark will be given to the submissions that obtain the highest, the second highest and the third highest average value of *averageMCC* on the blind test dataset. In case of a tie, the winner will be decided based on the higher value of the *accuracy*.

Deadlines and Delivery

Submission of the reports and the predictions is due on April 30 (Monday), 2018, before 3:30pm. The report should be delivered as a hard copy in the classroom and predictions should be send by email to lkurgan@vcu.edu.

Final Notes

- Remember to fill in and sign the team project contracts and return them to the instructor (in person) by the next class (April 4, 2018 at 3:30pm) at the latest.
- Do not cheat (e.g., do not inflate or "tweak" the results). It is better to report honest results than to get caught cheating. In the latter case you are risking receiving 0 marks for the project.
- Always copy the email communications to yourself so you can prove that it was sent.
- Contact the instructor immediately if any problems occur.