CSCI 6803: Algorithms in Bioinformatics

Jéssica Pauli de Castro Bonson

Project Proposal

Goal of the Project: In this project I intend to develop a classifier for the p53 Mutants Set [1], and to identify what type of features have a bigger impact in the classification performance. The dataset is heavily imbalanced and have 5408 attributes, so it will be necessary to balance the dataset and to perform filter filtering in order to train the classifier. In previous work [3, 4, 5, 6] the authors focused mainly on achieving a good accuracy for subsets of the dataset. So the first goal is to implement a classifier that works with the full dataset. The 5408 attributes are divided between 4826 features that represent 2D electrostatic and surface data, and 581 3D distance based features. So the second goal is to determine what feature group has a bigger relevance in the data classification.

Choice of Data Set:

- What data set will you use? p53 Mutants Set [1]
- From where will you retrieve it? From the UCI Machine Learning Repository [2]
- Why is this data set interesting? The dataset contains 5408 features extracted from biophysical simulations of mutant p53 proteins, since the actual full structure of wild-type p53 is unknown [5]. p53 is a tumor suppressor protein, and mutations in this protein are strongly correlated with human cancer [5]. The dataset contains 143 active (positive/transcriptonally competent) elements and 16629 inactives (negative/cancer). The data can be used to produce a classifier so a better understanding about p53 mutations can be achieved without the cost of labor and resources involved in in-vitro experimentation, and to identify what features differentiate the inactive from the active p53 mutations. At the end, the insights obtained may lead to improvements in the cancer treatment [6].

Choice of Method:

- Set of approaches: Clean or normalize the data to remove data points with missing values; Stratified K-fold; SVM (more specifically, the Linear Support Vector Classification model(LinearSVC)). And, if necessary, I will use the SMOTE algorithm. For feature filtering and determining feature importance, I will pre-process the data using statistical analysis (e.g.: VarianceThreshold and Univariate feature selection) and them further select features using one or more of these three methods: LinearSVC, Tree-based estimators, or Genetic Algorithm (GA).

- Why are these approaches appropriate to the problem? The first problem of this dataset is that it is heavily unbalanced, so to deal with it I will use Stratified K-fold, and SVM. I will specifically use LinearSVC because I can tune it so it will give a higher importance to the positive cases, and also because it was a good performance for data with a lot of samples. Since SVM is able to work well even with a high number of dimensions, I think it is also a good choice regarding the high number of features of the dataset. SMOTE will be used in case the previous choices didn't solve the problem of unbalanced data satisfactorily. The last problem is to select the most relevant features, so the statistically insignificant features will be removed, and then one or more of the three methods will be used depending on the resultant features from the statistical analysis.
- What considerations did you take into account when choosing this approach? High-dimensionality, unbalanced classes, big size of the dataset, missing values, and feature filtering.
- Will you take advantage of existing software or libraries to implement your approach? Yes, I will use the Python library scikit-learn. If I end up using GA, I will implement it myself.

Analysis of Results:

- What form will your results take? The results from the classifier will be accuracy, sensivity, specificity, and Matthews correlation coefficient.

- Will you be comparing your results against other methods? Most of the related papers used a variance of the dataset that uses more samples and more features [3, 4, 5], so results will not be able to be completely comparable, another problem is that as far as I noticed the papers focused on subsets of the dataset, while I am interested in developing a classifier that identifies patterns for the whole dataset. I will look further for other related papers and ways to compare the results with them (e.g.: try to produce the same subsets for comparison with paper [6], since we use the exact same dataset and the same metrics).
- How will you assess the significance of your results? By now the goal is to obtain metrics results higher than random and better than a classifier that just assumes all data is negative. About the feature selection, the goal is to find the % of the original 2D and 3D features that are indeed useful for the classification task. Since I am using common metrics for 2-class classification of unbalanced data, I think I will be able to also assess the significance by comparing these metrics to other related papers.

References:

[1] http://archive.ics.uci.edu/ml/datasets/p53+Mutants

[2] http://archive.ics.uci.edu/ml/

[3] Danziger, S.A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G.W., Kaiser, P., and Lathrop, R.H. (2009) Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning, PLOS Computational Biology, 5(9), e1000498

[4] Danziger, S.A., Zeng, J., Wang, Y., Brachmann, R.K. and Lathrop, R.H. (2007) Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants, Bioinformatics, 23(13), 104-114.

[5] Danziger, S.A., Swamidass, S.J., Zeng, J., Dearth, L.R., Lu, Q., Chen, J.H., Cheng, J., Hoang, V.P., Saigo, H., Luo, R., Baldi, P., Brachmann, R.K. and Lathrop, R.H. (2006) Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 3, 114-125.

[6] Geetha Ramani, R., & Jacob, S. G. (2013). Prediction of P53 Mutants (Multiple Sites) Transcriptional Activity Based on Structural (2D&3D) Properties. *PLoS ONE*, *8*(2), e55401. doi:10.1371/journal.pone.0055401