AN EXTENSIVE MULTI-LABEL ANALYSIS OF THE TOXCAST DATA SET

Jörg Wicker[1], Julian Lemke[1] and Stefan Kramer[1]
[1] Institut für Informatik/I12, Technische Universität München

The ToxCast™ data set was released by the EPA in beginning of 2009 [1]. In August 2009, the data of phase one became publicly available. The data set contains 320 chemical structures and 1633 endpoints. The chemicals are tested against the ToxRefDB, which contains 424 toxicological in vivo endpoints. The remaining features are mainly in vitro data. The main goal of the ToxCast™ data set is to break the in vitro / in vivo border and use the in vitro features to predict in vivo endpoints. Nevertheless, the data set is not easy to handle. Both the in vitro and in vivo data are of unknown and presumedly varying quality, many instances have missing values (in both in vitro and in vivo data), the structures are heterogeneous, and there is a slight skew in the class distributions. At the ToxCast™ Data Analysis Summit in May 2009, first analysis strategies were presented by partners in the project. First analyses of the data set showed that it is hard to find correlations between the in vitro data and individual in vivo endpoints. We present the results of a comprehensive multi-label classification analysis [3] of the data set. Multi-label classifiers do not predict only one class but take into account interdependencies among several classes or labels to improve the prediction. In contrast to a previous multi-label approach presented by Jeliazkova et al. [2], we use all 320 structures and not just the 160 structures with all in vivo data available. We used the Mulan multi-label library [3], as it provides several multi-label classifiers and analysis methods. The ToxCast data set consists of many missing values, as not all structures are tested in all the assays. While many learning algorithms can cope with missing values in general, the performance of many multi-label schemes should be expected to suffer, as less information on the dependencies among labels can be exploited. To alleviate this problem, we studied methods for data imputation, i.e., methods to fill in missing values in the data. Most of these methods focus on missing numeric values. For multi-label classification, a method imputing binary data is required. We applied a new method developed on the basis of multi-label classification to fill missing labels in a multi-label data set. This improves the performance and allows using multi-label classifiers which cannot handle missing values in the labels.

Data Set

- ToxCast™ project by EPA [1]
- 320 (309) Structures mapped to approximately
  - 2000 endpoints
  - 1600 in vitro endpoints and structural features
  - 400 in vivo endpoints
- Goal
  - Predict in vivo endpoints from in vitro endpoints and structural features
  - Break in vivo / in vitro border
- Previous analysis mostly focused on single endpoints (except Jeliazkova et al. [2])
- For multi-label classification: Transform in vivo endpoints to binary data

Structure	atg_ar_trans	nvs_gpcr_hm1	...	chr_mouse_adrenalgland_0_anylesion	chr_mouse_adrenalgland_0_preneoplasticlesion	...
S1	5.5	10.8	...	1	0	...
S2	1000000	3.59	...	0	1	...
...	...	...	...	...	...	...

Table 1: Schema of the ToxCast data set. First column is the name of the structure, the next block are in vitro endpoints and structural information, the last the in vivo endpoints.

Multi Label Classification

- Traditional classification: predict single class
- Multi-label classification: predict multiple interdependent binary classes [3]

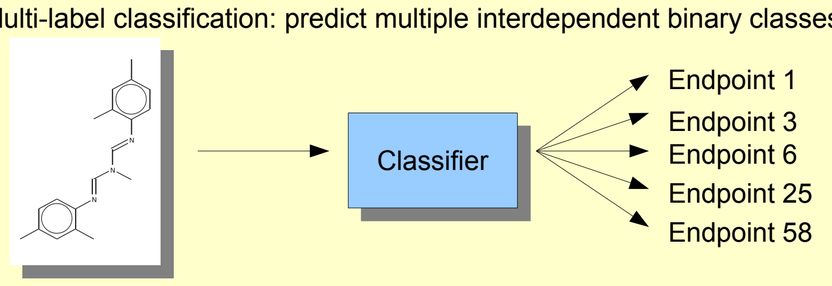

Figure 1: Multi-label Classification for ToxCast

- ToxCast:
  - 424 labels
  - Cardinality (average number of active in vivo endpoints per instance): 16.37 (very high)
  - Density (average ratio of active in vivo endpoints per instance): 0.039 (low)
  - Distinct label combinations: 308 (very high)

Missing Values

- ToxCast has many missing values in both in vitro and in vivo data
  - Missing values in Labels problem for multi-label classifiers
- Use Imputation for multi-label Classification to fill missing in vivo data
- Imputation using classifier chains
  - Learn one single label classifier for one label using only known values
  - Predict missing values using the learned model
  - Add label to features
  - Continue with next label until all missing values are filled
  - Use ensembles with each single chain using a random order

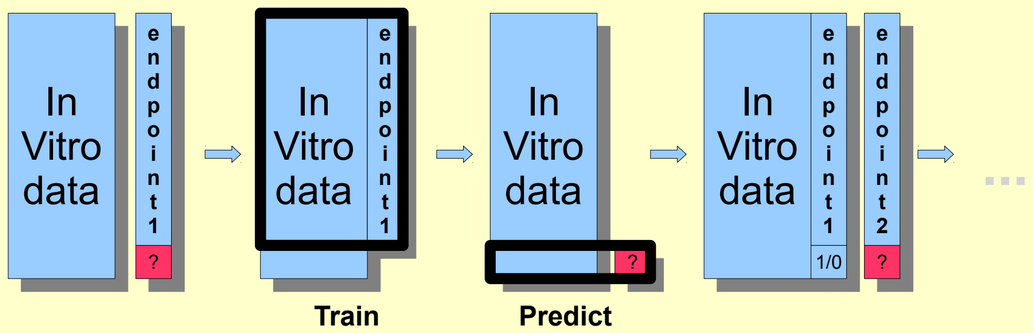

Figure 2: Imputation of missing values

Experiments

- Run 4 multi-label classifiers from the Mulan[1] workbench on the data set
- Impute missing values using different methods
- 10 fold cross validation

[1] http://mulan.sourceforge.net

Results

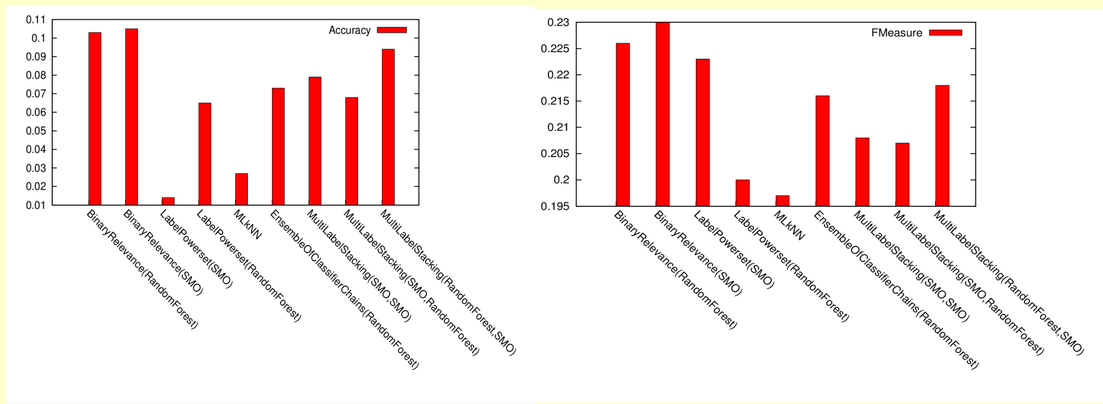

Figure 3: Accuracy and FMeasure for selected Multi Label Algorithms. Binary Relevance is the baseline using every single label as independent class

- Using simple multi-label Algorithms show no benefit over single label classifiers.
- Multiple Problems:
  - Label Powerset suffers from many unique label combinations and Missing Labels
  - MLkNN suffers from high dimension and Missing Labels
  - Ensemble of Classifier Chains suffers from small number of instances
  - Multi-label Stacking suffers from missing values

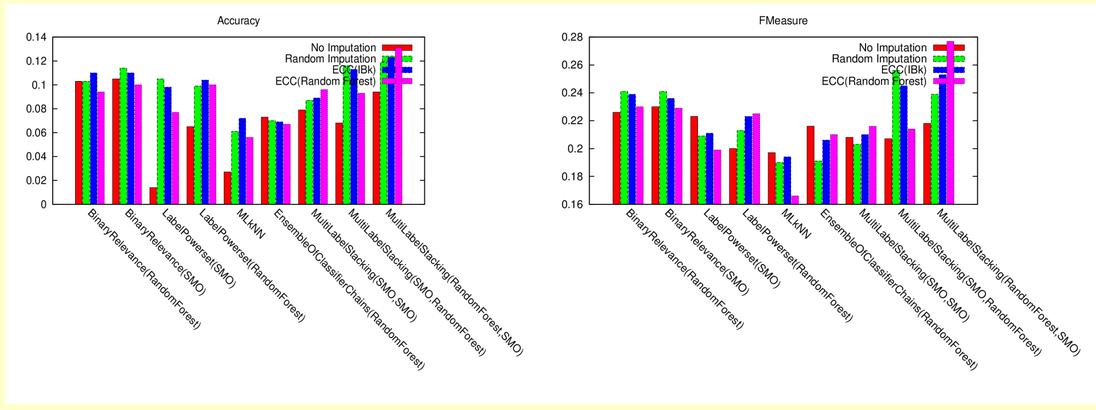

Figure 4: Accuracy and FMeasure for selected Multi Label Algorithms using different imputation methods

- Using imputation, the performance clearly improves when using multi-label classification methods
- Note that the multi-label classifiers used are still rather simple
  - No meta algorithms like HOMER used
- Imputation using classifier chains works better than Random imputation
- Certain multi-label classifiers use own imputation methods, so no benefit from using imputation
  - Most classifiers use Random imputation
- Note that even small changes in performance can be caused by larger change in one single label

Conclusions and future work

- Plain Multi Label classification on ToxCast suffers from many problems
- Multi Label classification with data imputation works on ToxCast
  - Still not very good performance
- Future Work:
  - Apply more sophisticated Multi-Label classifiers
  - Run extensive tests for data imputation and Multi Label Classification
  - Examine performance on single labels

References

[1] Judson, R. et al. (2010) "In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization - The ToxCast Project", Environmental Health Perspectives, in press (published online in December 2009).

[2] Jeliazkova, N., Jeliazkov V. (2009) "Hierarchical Multi-Label Classification of ToxCast Datasets", ToxCast Data Analysis Summit, US EPA, Research Triangle Park, NC, May, 14-15, 2009.

[3] Tsoumakas, G., Katakis, I., Vlahavas, I. (2010) "Mining Multi-Label Data", Data Mining and Knowledge Discovery Handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd edition, 2010.