Analyzing the Japanese Toxicogenomics Project Dataset with SVM and RLS Classifiers

Jari Björne, Antti Airola, Tapio Pahikkala and Tapio Salakoski
Department of Information Technology, University of Turku
Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3-5, 20520 Turku, Finland
firstname.lastname@utu.fi

1 Introduction

The TGP dataset from the Japanese Toxicogenomics Project concerns the response of rats and rat and human *in vitro* cell cultures to a number of drugs [1]. In the CAMDA 2013 challenge this dataset is utilized for analyzing drug-induced liver injury (DILI). Questions include the evaluation of the dataset to determine whether the animal model can be replaced with an *in vitro* cell culture, and whether DILI can be predicted using toxicogenomics data from animals. Both pathology data and microarray genomic expression data are provided in the challenge. We approach these questions as a machine learning task, evaluating the dataset in the context of SVM and RLS classifiers and in defining an experimental setup for automated prediction of DILI.

2 Dataset and Methods

We use the FARMS normalized version of the CAMDA dataset, intended to overcome observed cell culture effects [2]. The dataset consists of a large number of experiments, but in light of the proposed experimental question, predicting the liver injury potential of a drug, there are only 101 distinct examples, each example representing a single drug with a human DILI-concern rating, with features potentially combined from several *in vivo* or *in vitro* experiments. Of these drugs, 8 are in the "no DILI concern", 52 in the "less DILI concern" and 41 in the "most DILI concern" categories.

A *per-drug* example dataset in LibSVM format is provided as part of the CAMDA challenge, for the task of classifying drugs into "no DILI concern" vs. "most DILI concern". With only 8 examples in the "no DILI concern" class, if e.g. 10-fold cross validation were applied to the dataset, each subset would contain on average just a single example of this class, leading to potentially unstable results. We note that Pessiot et. al. [3] performed classification experiments using binary classification into "no or less DILI concern" vs. "most DILI concern", an experimental setup resulting in a more balanced class distribution.

In defining our experimental setup our primary aim was to formulate a question that would result in a larger dataset, potentially producing more reliable results. Therefore, we defined as our

per-individual experiment whether the individual animal or cell culture in a single experiment had been treated with a drug of "no or less DILI concern" or "most DILI concern". This setup is of course very close to classification on the level of drugs, but allows us to explore the classification potential of the individual variation between experiments, and provides us with a larger set of examples. To maximize available data we also combined single and repeated dose rat *in vivo* experiments. In preliminary classification studies, we observed that models trained on high drug dose experiments had best performance, and that 9 hr, 24 hr and 29 day time points for the rat *in vivo* data, as well as 8 hr for the human and 2 hr for the rat *in vitro* data had best performance. Selecting these experiments for further study, we produced datasets with "most"/"no or less" examples at 80/160 for human *in vitro*, 82/120 for rat *in vitro* and 205/215 for rat *in vivo* experiments.

There are of course strong implied dependencies between individual experiments with a single drug, between not only replicates, but potentially also time points and doses. To avoid information leaks, when selecting examples for training and testing, we always put all examples treated with the same drug into either the training or the testing set.

2.1 Features

We defined a number of feature groups to be used for classifying the data. *Pathology features* are the pathology, hematology, biochemistry and liver weight data, available for the *in vivo* rat experiments. *Array features* are the FARMS-processed, non-collapsed microarray expression values available for all experiment types. We also experimented with using *INI scaling*, multiplying the expression values with their reliability estimates ($value * (1 - INI)$) [4].

In addition to these basic features, we also explore refining the dataset with additional data on tissue specificity of gene expression. We retrieve from UniGene[1] known tissues of expression for both rat and human genes. For each tissue-specific group of expression values we define a set of statistical features (minimum, maximum, mean, median and variance) intended to give an overview of expression values. Alternatively, we also select as array features and the tissue-specific statistics only the subset of genes known to be expressed in the *liver*, based on UniGene data.

2.2 Machine learning approach

We apply two state-of-the-art machine learning algorithms: the support vector machine (SVM) and the regularized least-squares (RLS) method, also popularly known as least-squares SVM, or ridge regression [5]. The methods are closely related, and have in numerous experimental comparisons been shown to have quite similar performance. A specific advantage for RLS is the existence of efficient computational short cuts for computing cross-validation estimates. These are especially useful in the considered setting, since due to the small sample size, a central challenge for the evaluation is how to do parameter selection, and at the same time obtain a reliable estimate of the predictive performance. For RLS learning and cross-validation algorithms, we use the implementations in the RLScore[2] software package.

For the initial SVM experiments we applied the SVM^{multiclass} support vector machine[3] [6]

[1] `http://www.ncbi.nlm.nih.gov/unigene`

[2] `http://www.tucs.fi/RLScore/`

[3] `http://svmlight.joachims.org/svm_multiclass.html`

with a linear kernel. Experiments with 10-fold cross-validation proved very unstable, likely due to the small number of examples in each subset, so we adopted a 5-fold cross-validation approach, where two fifths were used to train the classifier, two fifth's to optimize the parameters (opt set) and one fifth to evaluate performance (test set). All examples for the same drug were always placed in the same fold.

To maximize the use of the available data we took advantage of the cross-validation capabilities of the RLScore package. Following the recommendations of [7] we apply a leave-pair-out cross-validation scheme, defined as follows:

$$\frac{1}{|I_+||I_-|}\sum_{i\in I_+}\sum_{j\in I_-} H(f_{\overline{\{i,j\}}}(x_i) - f_{\overline{\{i,j\}}}(x_j)),$$

where $f_{\overline{\{i,j\}}}$ denotes a classifier trained with the whole data set except the i-th and j-th training examples, and $I_+ \subset I$ and $I_- \subset I$ denote the indices of the positive and negative instances in the whole data set Z, respectively. We enumerate all the drug-pair combinations, and on each round of cross-validation leave as test examples all data points corresponding to these two drug pairs. The setup guarantees that we have no information leak between training and test data, since all data points corresponding to same drug are always in the same fold. Further, as shown by [7], the method makes maximal use of the available data, producing an almost unbiased estimate of the AUC, with lower variance than alternative approaches. We perform nested cross-validation, with an inner leave-pair-out loop used for parameter selection, and an outer one for performance estimation.

3 Results and Discussion

In Table 1 we present RLS leave-pair-out classification results for the preprocessed *per-drug* TGP data prepared by the CAMDA organizers. We notice considerable variance in the results: while the best performance achieved on high-dose level at 24 h is 0.71, the lowest one is 0.23 AUC at 2 h on low dosage, which is much worse than a random classifier would be expected to perform (0.5). Therefore we consider it unclear how much predictive power the learned models really have, or if the detected patterns are just due to random chance.

In Table 2 are shown the results for our *per-individual* approach to the CAMDA dataset, testing both SVM and RLS classifiers. We again notice considerable variance on the results. While the RLS cross-validation presents the most efficient way of utilizing the available data for training, the 5-fold SVM cross-validation should result in relatively similar results for truly reliable predictions. We notice the two experiments provide similar results mostly on the rat *in vivo* data, where performance is also the highest. The rat and human *in vitro* datasets show considerably lower performance, with human results slightly more promising. In our experimental setup, the use of INI values did not have much impact on performance. The direct use of the pathology data as features resulted in very unstable models for the *in vivo* data. We note that the UniGene tissue specific expression statistics show some potential on the *in vitro* datasets, achieving on occasion relatively high performance with a much smaller number of features, but due to the variance of the dataset, these observations should be considered highly speculative. Overall, the use of the *in vivo* expression data as features resulted in the most stable models.

Table 1: RLS nested leave-pair-out cross-validation results on preprocessed TGP per-drug data.

Dose	AUC (2 h)	AUC (8 h)	AUC (24 h)	AUC (all timepoints)
All				0.60
Low	0.23	0.48	0.57	
Middle	0.61	0.61	0.63	
High	0.46	0.49	0.71	

Table 2: RLS nested leave-pair-out cross-validation and SVM 5-fold cross-validation results (parameter optimization set and test set) on per-individual TGP data. Results over AUC 0.6 are shown in bold and under AUC 0.5 in italics.

Species	Feature Groups	Features	SVM(opt)	SVM(test)	RLS
human *in vitro*	INI, array	9011	**0.59 ± 0.06**	**0.53 ± 0.07**	0.54
human *in vitro*	INI, array, unigene	9479	**0.60 ± 0.07**	*0.50 ± 0.09*	*0.40*
human *in vitro*	INI, array, unigene(liver)	8049	**0.60 ± 0.06**	**0.52 ± 0.09**	0.53
human *in vitro*	INI, unigene	471	**0.55 ± 0.06**	**0.53 ± 0.07**	0.57
human *in vitro*	array	18980	**0.60 ± 0.07**	*0.49 ± 0.07*	0.54
human *in vitro*	array, unigene	19448	**0.60 ± 0.06**	*0.48 ± 0.08*	*0.40*
human *in vitro*	array, unigene(liver)	12093	**0.60 ± 0.06**	*0.49 ± 0.07*	0.53
human *in vitro*	unigene	471	*0.54 ± 0.05*	*0.52 ± 0.05*	**0.65**
human *in vitro*	unigene(liver)	15	*0.53 ± 0.05*	*0.49 ± 0.01*	0.53
rat *in vitro*	INI, array	7950	0.54 ± 0.03	*0.53 ± 0.06*	0.55
rat *in vitro*	INI, array, unigene	8130	0.54 ± 0.03	*0.52 ± 0.05*	0.53
rat *in vitro*	INI, array, unigene(liver)	4752	0.56 ± 0.03	*0.51 ± 0.06*	0.51
rat *in vitro*	INI, unigene	183	0.55 ± 0.02	*0.53 ± 0.06*	0.56
rat *in vitro*	array	12080	0.54 ± 0.03	*0.53 ± 0.04*	0.55
rat *in vitro*	array, unigene	12260	0.54 ± 0.03	**0.54 ± 0.06**	0.51
rat *in vitro*	array, unigene(liver)	5533	0.56 ± 0.03	*0.52 ± 0.06*	0.51
rat *in vitro*	unigene	183	0.55 ± 0.02	*0.54 ± 0.05*	0.58
rat *in vitro*	unigene(liver)	9	0.50 ± 0.00	0.50 ± 0.00	*0.36*
rat *in vivo*	INI, array	6753	**0.60 ± 0.06**	**0.58 ± 0.04**	**0.61**
rat *in vivo*	INI, array, unigene	6933	**0.58 ± 0.03**	**0.58 ± 0.08**	**0.61**
rat *in vivo*	INI, array, unigene(liver)	4299	**0.57 ± 0.03**	0.55 ± 0.03	**0.60**
rat *in vivo*	INI, unigene	187	**0.59 ± 0.03**	*0.51 ± 0.05*	0.55
rat *in vivo*	array	12084	**0.60 ± 0.06**	**0.61 ± 0.09**	**0.60**
rat *in vivo*	array, pathology	12189	**0.62 ± 0.11**	*0.37 ± 0.11*	*0.49*
rat *in vivo*	array, unigene	12264	**0.59 ± 0.04**	**0.59 ± 0.06**	**0.61**
rat *in vivo*	array, unigene(liver)	5537	**0.58 ± 0.03**	0.56 ± 0.03	**0.60**
rat *in vivo*	pathology	112	**0.62 ± 0.10**	*0.42 ± 0.04*	*0.41*
rat *in vivo*	unigene	187	**0.59 ± 0.03**	*0.50 ± 0.03*	0.55
rat *in vivo*	unigene(liver)	13	0.52 ± 0.01	*0.48 ± 0.02*	0.50

4 Conclusions

We find it notable that mostly the largest drug doses produced data that could be classified the best, possibly indicating that the DILI-related gene expression response is rather faint, pointing to the need for an experimental setup strong enough to produce unambiguous data.

Our Python-based experimental software is built on publicly available tools, depending only on open source classifiers. We will also provide all of our code under an open source license, hopefully useful for further research on the topic.

In testing various feature representations, we observed potential value on refining the expression data with external databases such as UniGene. However, most importantly, performing a large set of experiments with somewhat related feature representations and different classifiers highlighted the disturbingly large variance in classification performance. In understanding the potential of the TGP dataset for building predictive models we therefore consider it highly important that all experimental results are carefully compared and evaluated.

References

[1] T. Uehara, A. Ono, T. Maruyama, I. Kato, H. Yamada, Y. Ohno, and T. Urushidani, "The Japanese toxicogenomics project: Application of toxicogenomics," *Molecular Nutrition Food Research*, vol. 54, no. 2, pp. 218–227, 2010.

[2] S. Hochreiter, D.-A. Clevert, and K. Obermayer, "A new summarization method for affymetrix probe level data," *Bioinformatics*, vol. 22, no. 8, pp. 943–949, 2006.

[3] J.-F. Pessiot, P. S. Wong, T. Maruyama, R. Morioka, S. Aburatani, M. Tanaka, and W. Fujibuchi, "The impact of collapsing data on microarray analysis and DILI prediction," *Systems Biomedicine*, vol. 1, no. 3, pp. 1–7, 2013.

[4] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. Ghlmann, "I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data," *Bioinformatics*, vol. 23, no. 21, pp. 2897–2902, 2007.

[5] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, pp. 1–50, April 2000.

[6] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research (JMLR)*, vol. 6(Sep), pp. 1453–1484, 2005.

[7] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski, "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve," *Computational Statistics & Data Analysis*, vol. 55, pp. 1828–1844, April 2011.