Matrix Factorization-Based Data Fusion for Drug-Induced Liver Injury Prediction

Marinka Žitnik[1] and Blaž Zupan[1,2]

[1] Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
[2] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX-77030, USA

marinka.zitnik@fri.uni-lj.si, blaz.zupan@fri.uni-lj.si

Abstract. We report on a data fusion approach for prediction of outcome of drug-induced liver injury (DILI) in humans from gene expression studies as provided by the CAMDA 2013 Challenge. Our aim was to investigate if the data from all four toxicogenomics studies can be fused together to boost prediction accuracy. We show that recently proposed matrix factorization-based fusion provides an elegant framework for integration of CAMDA and related data sets. Our data fusion approach yields a high cross-validated AUC of 0.819 (in vivo assays), which is above the accuracy of standard machine learning procedures (stacked classification with feature selection). Achieved accuracy is also a substantial improvement of the highest scores on the same data sets reported in CAMDA 2012. Our data analysis shows that animal studies can be replaced with in vitro assays (AUC = 0.799) and that we can predict liver injury in humans from animal data (AUC = 0.811).

1 Introduction

Molecular biology abounds with data from sequencing, expression studies, function annotations, studies of interactions and other. These data sources are related, and analysis of one data set could benefit from inclusion of others. We have recently proposed a data fusion approach [1] that can elegantly integrate heterogeneous data sources, representing each data set in a matrix and fusing the data sets by simultaneous matrix factorization. We here report on the fusion of 29 data sets from CAMDA Challenge and related data repositories to predict DILI potential. We compare the accuracy of data fusion to that of a standard multi-classifier approach where we stack four state-of-the-art classification algorithms. We additionally investigate feature subset selection by CUR matrix decomposition [2] applied before stacking [3]. Our principal contribution is a demonstration that toxicogenomics studies can substantially benefit from data fusion.

2 Data fusion by Matrix Factorization

We use data fusion by matrix factorization [1], an intermediate data integration approach that is able to fuse heterogeneous data sources. Intermediate integration is often the preferred integration strategy [4,5,6] as it embeds the structure of the data into a predictive model and for this reason often achieves higher accuracy.

Data fusion considered 14 object types (nodes in Fig. 1, *e.g.,* drug, GO term, or drug type) and a collection of 29 data sources, each relating a pair of object types (arcs in Fig. 1, *e.g.,* gene annotations that relate genes and GO terms). In addition to FARMS-summarized expression data sets we include data on drugs available from DrugBank[3], gene annotations from Gene Ontology[4], protein-protein interactions from STRING[5], and

[3] http://www.drugbank.ca
[4] http://www.geneontology.org
[5] http://string-db.org

hematological and clinical chemistry data for each animal and array metadata information, the latter being provided by the challenge organizers. We did not use in vivo pathological findings in the fused model.

We represent the observations from a data source that relates two distinct objects types i and j in a sparse relation matrix $\mathbf{R}_{ij}$ (*e.g.,* $\mathbf{R}_{1,13}$ for annotations of genes in rat in vivo single study). A data source that provides relations between objects of the same type i is represented by a constraint matrix $\mathbf{\Theta}_i$ (*e.g.,* $\mathbf{\Theta}_{10,10}$ for DrugBank's drug interactions). Relation matrices $\mathbf{R}_{ij}$ are simultaneously factorized under constraints by $\mathbf{\Theta}_i$ [1]. The resulting system contains factors $\mathbf{S}_{ij}$ that are specific to each data source and factors $\mathbf{G}_i$ that are specific to each object type, such that each relation matrix $\mathbf{R}_{ij}$ is approximated as $\widehat{\mathbf{R}}_{ij} = \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T$. Fusion takes place due to matrix factor sharing during decomposition of relation matrices.

We apply data fusion to infer relations between drugs and DILI potential, respectively. This relation, encoded in a target matrix $\mathbf{R}_{10,14}$, is observed in the context of all other data sources. Matrix $\mathbf{R}_{10,14} \in \mathbb{R}^{131\times 3}$ is a $[0,1]$-matrix that is only partially observed. Its entries indicate drugs' degree of membership to the three DILI severity classes, which are "No concern DILI", "Less concern DILI" and "Most concern DILI", respectively. We aim to predict the unobserved entries in $\mathbf{R}_{10,14}$ by reconstructing them through matrix factorization. The DILI severity of p-th drug is determined as $\arg\max_i \widehat{\mathbf{R}}_{10,14}(p,i)$.

3 Multi-Classifier Approach and Feature Subset Selection by CUR Matrix Decomposition

We use FARMS-summarized gene expression data for the four toxicogenomics studies that were provided by the organizers of the challenge [7]. We employ CUR matrix decomposition [2] to identify a small set of information carrying genes. CUR matrix decomposition in an unsupervised manner approximates target matrix $\mathbf{A}$ as $\mathbf{A} \approx \mathbf{CUR}$, where $\mathbf{C}$ and $\mathbf{R}$ are low-dimensional matrix factors that contain a subset of columns and rows from $\mathbf{A}$, respectively. The advantage of CUR decomposition over some well known low-rank matrix decompositions such as principal component analysis (PCA) or singular value decomposition (SVD) is its explicit representation in terms of a small number of actual columns and rows of target data matrix. The CUR decomposition-selected features correspond to original gene expression profiles instead of their linear combinations as with PCA and SVD. We then apply several state-of-the-art classifiers to predict the DILI concern in human from the matrix factor $\mathbf{C}$ obtained for each toxicogenomics study separately. We use gradient tree boosting with multinomial deviance as a loss function to model the three classes of DILI severity, random forests, support vector machine with polynomial kernel. Individual predictions are ensembled through stacking with logistic regression [3].

4 Results and Discussion

The performance of proposed inference approaches was estimated through 10-fold cross-validation. Feature subset selection for multi-classifier approach was performed on training data sets. Parameters of the classification and matrix decomposition algorithms, such as the number of iterations and the sizes of the constituent trees in gradient tree boosting, were estimated through internal cross-validation on the training data.

In our first experiment we considered the DILI prediction problem for each study separately and pursued a multi-classifier approach (Table 1). Feature subset selection by CUR

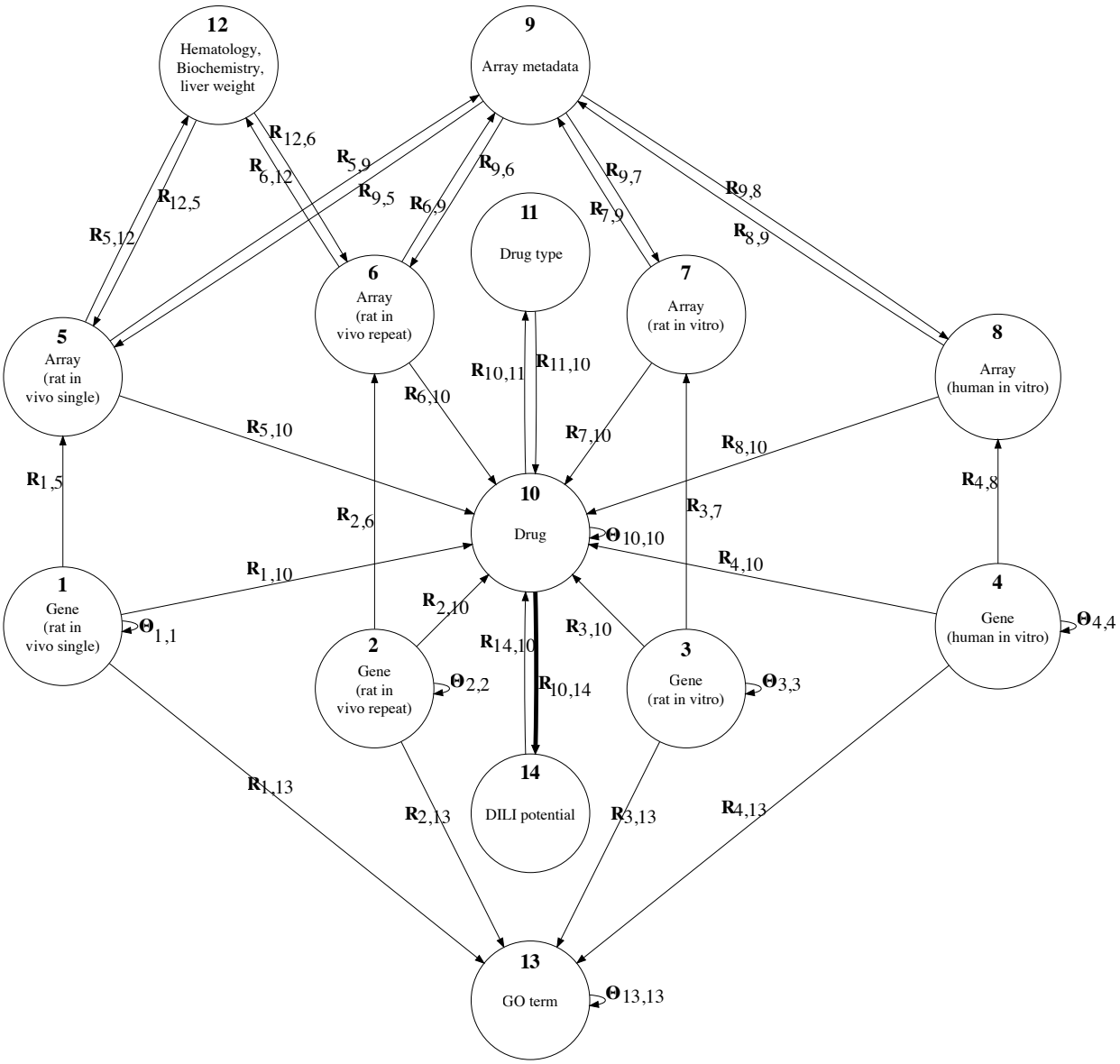

Fig. 1: Fused data sources. Nodes represent 14 object types. Arcs denote data sources that relate objects of different type (relation matrices, $\mathbf{R}_{ij}$) or objects of the same type (constraints, $\mathbf{\Theta}_i$) for a total of 29 matrices-data sources. Bold arc ($\mathbf{R}_{10,14}, \mathbf{R}_{14,10} = \mathbf{R}_{10,14}^T$) represents relation between drugs and DILI potential that we try to augment. Fused data sources include gene annotations that are encoded in $\{0,1\}$-matrices $\mathbf{R}_{1,13}$, $\mathbf{R}_{2,13}$, $\mathbf{R}_{3,13}$ and $\mathbf{R}_{4,13}$, expression profiles ($\mathbf{R}_{1,5}$, $\mathbf{R}_{2,6}$, $\mathbf{R}_{3,7}$, $\mathbf{R}_{4,8}$), hematology, body weight and clinical chemistry data for each rat ($\mathbf{R}_{5,12}$, $\mathbf{R}_{6,12}$, $\mathbf{R}_{12,5} = \mathbf{R}_{5,12}^T$, $\mathbf{R}_{12,6} = \mathbf{R}_{6,12}^T$), array metadata information such as dose level, dosage time and sacrifice time ($\mathbf{R}_{5,9}$, $\mathbf{R}_{6,9}$, $\mathbf{R}_{7,9}$, $\mathbf{R}_{8,9}$, $\mathbf{R}_{9,5} = \mathbf{R}_{5,9}^T$, $\mathbf{R}_{9,6} = \mathbf{R}_{6,9}^T$, $\mathbf{R}_{9,7} = \mathbf{R}_{7,9}^T$, $\mathbf{R}_{9,8} = \mathbf{R}_{8,9}^T$), drug targets ($\mathbf{R}_{1,10}$, $\mathbf{R}_{2,10}$, $\mathbf{R}_{3,10}$, $\mathbf{R}_{4,10}$), indication of medical drugs tested with arrays ($\mathbf{R}_{5,10}$, $\mathbf{R}_{6,10}$, $\mathbf{R}_{7,10}$, $\mathbf{R}_{8,10}$), structure and categorization of drugs ($\mathbf{R}_{10,11}$, $\mathbf{R}_{11,10} = \mathbf{R}_{10,11}^T$). Constraint matrices encode protein-protein interactions ($\mathbf{\Theta}_{1,1}$, $\mathbf{\Theta}_{2,2}$, $\mathbf{\Theta}_{3,3}$, $\mathbf{\Theta}_{4,4}$), drug interactions ($\mathbf{\Theta}_{10,10}$) and semantic structure of Gene Ontology graph ($\mathbf{\Theta}_{13,13}$).

matrix decomposition substantially reduced the number of features. For instance and as averaged across cross-validation folds, only about 300 features were used for training the prediction models in human in vitro study instead of original 18,988 features included by FARMS summarization. Solid performance of multi-classifier approach was not surprising [8,9], yet the substantial improvement of the AUC scores from CAMDA 2012 was.

Notice that we did not reimplement the procedures from [10], so the comparison of AUC scores is only indicative as they were obtained on different data samples chosen by cross validation. Yet the relatively large gains in AUC by our methods do provide evidence for improvements in prediction performance.

Notice also comparable performance of data preprocessing by CUR factorization and PCA. As CUR performs feature selection rather than feature transformation, it could be a preferable procedure to identify gene biomarkers.

Table 2 reports on 10-fold cross-validated accuracy for seven data fusion configurations that considered various subsets of the complete fusion model in Figure 1. The model inferred from all assays used an entire collection of data sources from Figure 1. Other models considered only selected toxicogenomics studies and associated non-expression data. For instance, fusion of in vivo assays omitted all data sets from in vitro studies (object types 3, 4, 7, and 8).

Data fusion surpassed the accuracy of multi-classifier approach to predict DILI potential in humans (Table 2). The most accurate model was inferred by fusing in vivo assays, which scored AUC of 0.819. It is surprising that in vivo assays, which relied on animal model, performed better than human assays, as we aim at predicting DILI potential in humans. However, last year's participants Pessiot *et al.*, 2012 [10] similarly observed that using in vivo animal data was more informative than using in vitro data from humans. Their AUC scores obtained by linear support vector machine classifier and inferred from separate toxicogenomics studies were substantially lower than those reported by our fusion-based approach. Also, fusion-based model inferred from animal assays (these are three studies, two in vivo and one in vitro study) outperformed model obtained by fusing human assays only (one human in vitro study), where the first achieved AUC of 0.811 and the latter AUC of 0.792. One might expect that administration of drugs to animal models would fail to identify the risk of liver injury for drugs prescribed to human due to differences in metabolic pathways and the current lack of suitable animal models that reproduce the human risk factors [11]. Our results do not confirm this hypothesis, although differences in performance are small and further investigations seem worthwhile pursuing.

Machine learning method	human in vitro	rat in vitro	rat in vivo single	rat in vivo repeated
Log. reg. stack. (RF, MD GBT, LR, SVM) w. PCA	0.741	0.765	0.748	0.761
Log. reg. stack. (RF, MD GBT, LR, SVM) w. CUR	0.758	0.755	0.764	0.778
Pessiot *et al.*, 2012 [10]	0.59	0.58	0.67	0.66
Clevert *et al.*, 2012 [12]		0.26*		

Table 1: Predictive performance of multi-classifier approach for DILI potential prediction with and without CUR dimensionality reduction. Reported are 10-fold cross-validated AUC scores. Acronyms: RF - random forests [13], MD GBT - multinomial deviance gradient boosting trees [14], LR - logistic regression, SVM - support vector machine (polynomial third degree kernel). CAMDA 2012 scores are from Pessiot *et al.* [10] and Clevert *et al.* [12] who used different cross-validation indices and data preprocessing. *Clevert *et al.* [12] reported the error rate and not AUC score.

Fused data	AUC
In vivo assays	0.819
All in vitro assays	0.790
Human in vitro assays	0.793
Animal in vitro assays	0.799
Animal assays	0.811
Human assays	0.792
All assays	0.810