Tree based feature induction for biomedical data

Konstantinos Pliakos[1] and Celine Vens[2]

***Abstract*— During the recent years, a great advance in both biomedical data acquisition technologies and feature extraction methods has been witnessed. Harnessing these new tools and technologies has led to an indisputable increase in the number of available biomedical datasets. Despite the efforts made so far, the representational power of features used to describe a sample in such datasets, such as a gene in gene function prediction datasets or a protein in protein interaction datasets has yet to be improved. Here, the performed study focuses on the feature representation power from a machine learning perspective.**

I. INTRODUCTION

The information contained in the data features and the ability to discriminate the samples in a dataset are of paramount importance for various machine learning tasks, such as classification or interaction prediction. The task of representing a data sample is not trivial when it comes to biomedical datasets where the features derive from laboratory experimental work. Feature extraction is often a time consuming and expensive process. The machine learning community has multiple biomedical research datasets at its disposal and indisputably, researchers should get advantage of this amount of data, but the question that arises is how representative the data features really are. For instance, in the domains of gene function prediction (e.g., [1], [2]) or interaction prediction (e.g., [3]) there are examples of widely used benchmark datasets where there is not only an issue of lacking variance between the samples, but there are samples having identical feature representation. The irrational behind this occurrence is that there are for example genes, which despite belonging to very different classes (gene functions), have exactly the same feature representation. In Table I, some examples of this issue are presented.

TABLE I

DATASETS, THE NUMBER OF SAMPLES AND THEIR UNIQUE REPRESENTATIONS.

Context	Dataset	|samples|	|unique sample representations|
Gene fct. prediction	church	3755	**2352**
(S. cerevisiae)	pheno	1591	**514**
	hom	3854	3646
	seq	3919	3913
	struc	3838	3785
(A. thaliana)	scop	9843	9415
	struc	11763	11689
Interaction prediction	drugs	1862	1779
(DPI network)	proteins	1554	**683**

[1] [2] KU Leuven Kulak, Department of Public Health and Primary Care, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium

[1]konstantinos.pliakos@kuleuven.be, [2]celine.vens@kuleuven.be

The replicated feature vectors have a large, unwanted, effect on the behaviour of learning algorithms. For instance, consider a nearest neighbor classifier. A simple 1-NN classifier that is applied to the training set is expected to yield 100% accuracy in theory. However, with replicate feature vectors, it is no longer guaranteed that an instance is mapped onto itself. For the pheno dataset from Table I, running ML-KNN with $K = 1$ on the training set yields an average precision of only 51.59%. Furthermore, in a decision tree learner, genes with replicated feature vectors will all be classified into the same leaf node, as there exists no split to separate them. In Fig. 1, the distances between the feature vectors representing the data samples of the *pheno* dataset and the distances between the corresponding label vectors are demonstrated. It can be clearly seen that samples having the same feature representation by means of distance equal to zero are characterized by different labels.

One option to deal with replicate feature vectors is to reduce the training set such that only unique feature vectors are retained. However, it is unclear which targets should be associated with these unique instances and also, removing instances will alter the data distribution. In addition, removing instances is not an option in datasets that represent a biological entity, for example all genes of an organism. To this end, the only proper option is to keep the instances, and try to add diversity to them by introducing extra features. More precisely, we want to discriminate the samples having the same or approximately the same features but much different labels. To that aim, extra features are induced by using information from the label set.

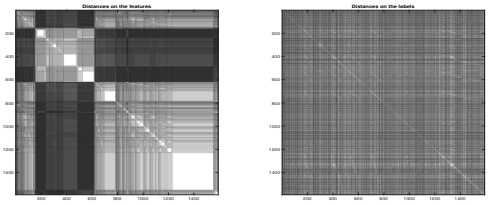

Fig. 1. Distance matrices for the features (left) and the labels (right) of the pheno dataset.

II. METHOD

Motivated by [4], we propose a method that generates a new feature set from an ensemble of trees constructed over the data. The ensemble that we use is Extremely Randomized Trees [5]. More precisely, the nodes of each decision tree of the ensemble are treated as clusters, containing all the

samples that fall into that tree node. Next, binary feature vectors are generated, where each component represents the presence or absence of a sample in a cluster (node). The new features are generated in an inductive manner. Moreover, they are label-aware (i.e., the new features are constructed using label information) as the clusters are formed by a supervised learning procedure. Different from [4], we assign a weight to each cluster (node), based on the variance of the label vectors that correspond to the samples contained in it.

In addition, the approach is further extended in order to tackle the issue of replicates. It is mainly based on harnessing additional information from the label set, adding extra features to the induced feature set. In particular, the label set associated to each instance i is represented as a binary vector. Then, an unsupervised clustering tree is learned on this binary representation. Afterwards, each node of this tree is regarded as a cluster and the clusters that are visited by the instance are mapped into binary features. Finally, these extra label-based feature vectors $\mathbf{L_i} = \{l_{i_1}, l_{i_2}, \cdots, l_{i_{|z|}}\}$ are concatenated to the original induced feature set, $\mathbf{F_i}$. Each feature vector $\mathbf{L_i} \in \mathbf{L}$ is multiplied by a global weight $\omega \geq 0$. Thus, the total induced feature vector corresponds to the concatenation $\mathbf{F} + \omega\mathbf{L}$. Note that the vectors $\mathbf{L_i}$ will be different for instances with replicate feature vectors but different label sets, because the instances will now reach different leaf nodes through the label-based splits. The outlined procedure is only applicable to the training instances. For the test instances, there is no knowledge of the labels, and thus they can not be classified by the clustering tree. Therefore, the new features for a test instance i are computed by averaging the vectors that correspond to the k nearest neighbors of i present in the training set.

Finally, by applying dimensionality reduction using principal component analysis (PCA) the set becomes computationally more efficient.

III. Experiments

For evaluation purposes, some multi-output prediction datasets were employed that were used also in [3]. They are interaction datasets representing homogeneous or heterogeneous biological networks.

The two proposed approaches were validated separately. First, experiments were conducted validating the robust ERT-based approach, coined as Extremely randomized tree Feature Induction (EFI), on the datasets MN and PPI. It was tested with and without applying PCA. For clarity purposes, the induced features were also validated both in concatenation to the original set (EFI_{joint}) and alone. Next, the extended EFI (EEFI) was validated on two datasets (pheno and DPI) containing replicates. The heterogeneous interaction dataset DPI was split into two multi-label classification tasks, one for each feature set.

In order to compare the original and induced feature spaces, we applied an ERT to both of them and evaluate their predictive performance under a five-fold cross validation scheme. The micro-averaged precision-recall and ROC curve were used as evaluation measures. The number of trees in the ERT was set to 100, and the number of split points considered was set equal to the square root of the number of features. All trees were unpruned, and the minimal number of instances a leaf has to cover was tuned using internal cross validation, using values from the following set: $\{1,3,5,10,30,50,100\}$. For the ERT applied to the initial and induced feature sets, this value was fixed to 3. In the PCA dimensionality reduction, the number of dimensions retained was equal to the square root of the number of original features.

TABLE II

$AUPRC/AUROC$ measures for all compared approaches.

Data	$original$	EFI	$EFI + PCA$	EFI_{joint}
MN	$0.27/0.83$	$0.33/0.83$	$0.32/0.82$	$0.30/0.84$
PPI	$0.21/0.84$	$0.21/0.84$	$0.19/0.82$	$0.22/0.84$

TABLE III

$AUPRC/AUROC$ measures for all compared approaches.

Data	$original$	$EEFI$
pheno	$0.16/0.83$	$0.17/0.85$
DPI1	$0.14/0.77$	$0.14/0.77$
DPI2	$0.03/0.60$	$0.03/0.61$

In Table II, the predictive performance of the EFI method is measured. The proposed features (column EFI) are generally slightly more powerful than the original ones.

In Table III, the $EEFI$ was evaluated using 2 datasets suffering from replicate feature vectors. It was shown that further harnessing information from the labels discriminates the data samples and the proposed approach showed a slight performance improvement. The number of unique feature vectors on pheno increased from 514 to 1523, avoiding noise addition or overfitting by simply adding for example the labels as extra features.

IV. Conclusions

A major point in this work was to highlight the issue of lacking variance in data representations used in biomedical data and inform the scientific community about the way it affects machine learning. Furthermore, a promising feature induction approach based on tree ensembles was proposed in order to handle that problem. Indisputably, there is room for improvement of the method.

References

[1] Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Machine Learning **73**(2), 185–214 (2008)

[2] Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., Džeroski, S.: Predicting gene function using hierarchical multi-label decision tree ensembles. BMC bioinformatics **11**(1), 1 (2010)

[3] Schrynemackers, M., Wehenkel, L., Babu, M.M., Geurts, P.: Classifying pairs with trees for supervised biological network inference. Molecular BioSystems **11**(8), 2116–2125 (2015)

[4] Vens, C., Costa, F.: Random forest based feature induction. In: IEEE 11th International Conference on Data Mining (ICDM), pp. 744–753 (2011)

[5] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**(1), 3–42 (2006)