KU LEUVEN kulak

"Towards better predictions"

Improving interaction prediction exploiting background information

Konstantinos Pliakos and Celine Vens*

KU Leuven Kulak, Faculty of Medicine, Department of Public Health and Primary Care

**konstantinos.pliakos@kuleuven.be*

Abstract

Here, a new prediction scheme is proposed that is based on supervised learning using Random Forest (RF) extended by Kernel Principal Component Analysis (KPCA).

Problem statement

- **Interaction data** are characterized by **two sets of objects**, each described by its own set of features.
- In **interaction prediction**, the goal is to predict the interactions between both object sets, a process also referred to as **network inference**.

Importance

- The experimental methods for identifying such interactions are both expensive and time-consuming.
- Increase in the amount of the available research data.
- More complex feature representations of the data.
- Efficiency and accuracy are still open problems.

Proposed approach

- A straight-forward, supervised approach is proposed where interactions are predicted using multi-output **tree-based ensemble** methods and **KernelPCA**.
- KPCA is employed as an effective feature extraction technique, generating a more discriminative feature set from the original one.

Particular advantages:

- The power of the new feature set can be boosted by providing a suitable kernel.
- The possible existence of a non-linear manifold can be exploited.
- Using dimensionality reduction the existing noise in the dataset is discarded.
- The proposed scheme follows the inductive setup.
- Tree-based ensembles are applied on the concatenation of the new feature set and the original one. This way, the performance is enhanced and the **interpretability**, by means of feature ranking, provided by the tree-based methods, is kept.

Results

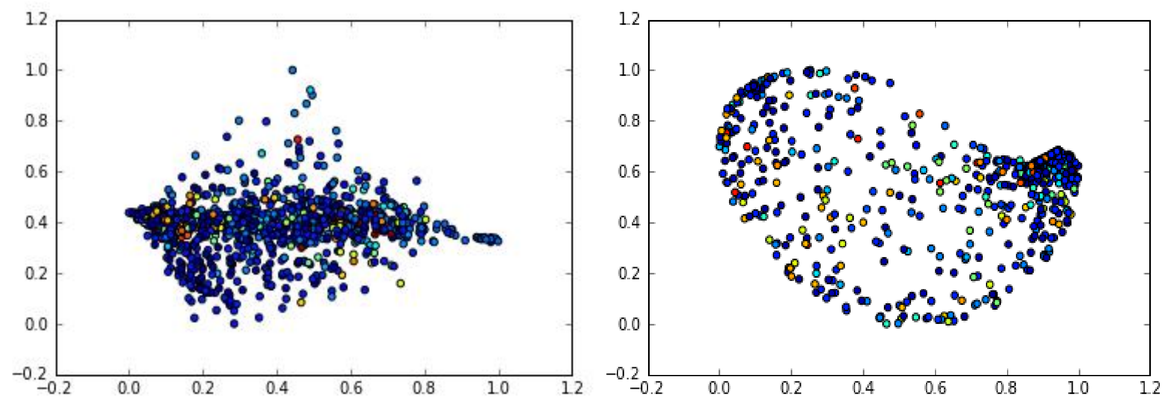
- Random Forest (RF) on the original features in comparison to RF on the extended feature set (ERF).

Data	AUPR	AUROC
	RF/ERF	RF/ERF
PPI (protein-protein)	0.17/0.18	0.80/0.81
MN (enzymes)	0.13/0.13	0.76/0.76
ERN (TF-genes)	0.40/0.42	0.85/0.86
ERN2 (genes-TF)	0.09/0.09	0.65/0.66
DPI (drug-protein)	0.13/0.13	0.75/0.75
DPI2 (protein-drug)	0.03/0.04	0.60/0.62
SRN (TF-genes)	0.20/0.20	0.82/0.82
SRN2 (genes-TF)	0.02/0.02	0.52/0.52

***TABLE 1**. AUPR and AUROC measures for the compared approaches.*

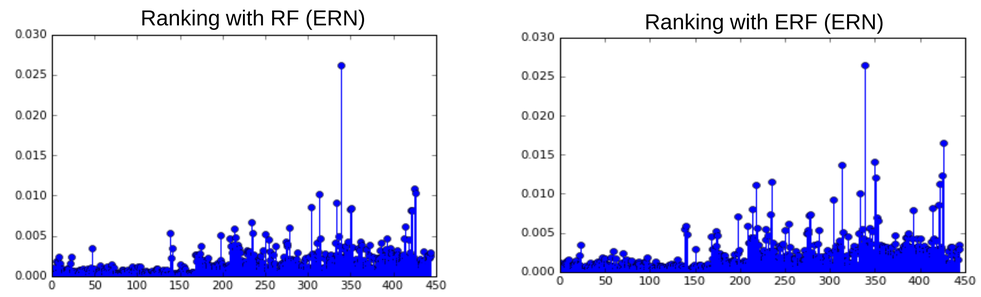
Data Visualization

Visualization of the ERN data distribution in the 2D space.

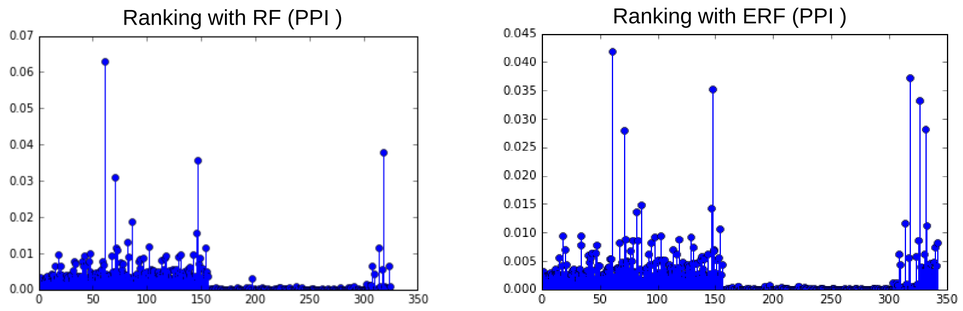


- On the left, the linear projection of the data is displayed using PCA.
- On the right, the non-linear projection of the data is demonstrated, using KernelPCA (rbf kernel).
- It is shown that by applying KPCA the data distribution is spread. Thus, there is potentially room for improvement in the performance of a predictor.

Feature ranking on the ERN and PPI datasets.



- It is shown that the ranking on the original features is not lost.



- It is shown that even if the results are not significantly improved the new features are still selected.

Conclusion

A promising interaction prediction approach was proposed, based on multi-output tree-based ensembles and further improved by including a more informative feature set in the learning procedure in concatenation with the original one.

Future research

- Employment of data-specific kernels in order to boost the performance.

REFERENCES

- Schrynemackers, M. *et al*. (2015). Classifying pairs with trees for supervised biological network inference. Mol. BioSys. 11, 2116-2125.
- You, Z.-H. *et al*. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC bioinformatics, 14, S10.
- Yu, H. *et al*. (2012). A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PloS one, 7, e37608.