Feature Induction based on Extremely Randomized Tree Paths

Konstantinos Pliakos[1] and Celine Vens[2]

[1] KU Leuven, Campus KULAK, Department of Public Health and Primary Care,
Etienne Sabbelaan 53, Kortrijk 8500, Belgium
konstantinos.pliakos@kuleuven.be,
[2] KU Leuven, Campus KULAK, Department of Public Health and Primary Care,
Etienne Sabbelaan 53, Kortrijk 8500, Belgium
celine.vens@kuleuven.be

Abstract. The volume of data generated and collected using modern technologies grows exponentially. This vast amount of data often follows a complex structure, significantly affecting the performance of various machine learning tasks. Despite the effort made, the problem of efficiently mining and analyzing such data is still persisting. Here, a novel data mining framework for unsupervised learning tasks is proposed based on decision tree learning and ensembles of trees. The proposed approach introduces an informative feature representation and is able to handle data diversity (e.g., numerical, canonical, etc.) and complexity (e.g., graphs, networks, data containing missing values etc.). Learning is performed in an unsupervised manner, following also the inductive setup. The experimental evaluation confirms the effectiveness of the proposed approach.

Keywords: tree-ensembles, extremely randomized trees, tree-embedding, data mining

1 Introduction

Nowadays, a great advance in data acquisition and feature construction methods is witnessed. Due to modern technological advances, huge amounts of data are generated in terms of both cardinality (i.e., the number of samples) and dimensionality (i.e., the number of features that describe each sample). These data often follow more complex structures, combining information from multiple sources. Furthermore, as the volume of data grows, problems such as the existing noise in the data or the missing values in some datasets remain. To this end, methods that can handle the aforementioned issues and succeed in mining complex patterns in big datasets are indisputably needed.

During the last years, an interest was witnessed in leveraging the mining of complex patterns by mapping the data to different feature spaces. This way, the performance of machine learning algorithms was improved. Most of the developed methods were based on kernel learning [1, 2], mainly due to the very good performance of Support Vector Machines (SVMs) [3]. However, these methods are often characterized by high computational costs and limited flexibility as one should compute and handle the whole Gram matrix. Many of these kernel-based methods have also been developed in a transductive setup where test instances are available during the training phase [1].

Moreover, there are several works where new features are constructed inductively using clustering techniques or decision tree learning. An unsupervised transformation of the data, using a set of random clustering forests was proposed in [4] for visual codebook construction. In particular, the features were generated by random trees embedding. The data encoding was based only on the indices of the leaves where a data sample ends up. The approach leads to a high dimensional, sparse binary coding. Most of the recently developed feature construction methods were developed for supervised learning tasks. In [5], a feature induction method based on random forests [6] was proposed. It was based on a metric transformation that mapped the identity of the tests performed in each node of a decision tree to a feature indicator. Feature vectors were generated by concatenating all the features corresponding to each tree in the forest and they were further encoded using hashing. In [7], a label-specific feature scheme for multi-label classification was proposed. For each label, a distinct feature set was constructed by clustering the label's positive and negative instances (separately), and then calculating the distances of each instance to the obtained cluster centroids. This way, the predictive performance of a classifier trained for that specific label was increased.

Here, we focus on developing a feature representation using tree ensembles. The main goal is to leverage unsupervised machine learning tasks, such as clustering or information retrieval. Decision tree induction algorithms [8, 9] are one of the most popular data mining algorithms. They have been applied extensively in many fields such as systems biology [10] or social media analysis [11]. Among the main advantages of these methods are the interpretability of the models they produce, which makes them transparent and understandable to human experts, leveraging knowledge discovery. Other advantages include their scalability from a computational point of view and their fair predictive accuracy. Combining them with ensemble methods [12, 6] improves their predictive performance and provides state-of-the-art results.

Motivated by [5], here we propose an unsupervised framework for feature construction based on tree ensembles and specifically Extremely Randomized Trees [13], hereafter denoted as ERT. In particular, 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 (i.e., the training samples are not needed to compute the feature vector of a test sample). Different from [5], the learning procedure is performed in an unsupervised manner. In addition, the employment of dimensionality reduction techniques [14, 15] is studied and the efficiency in detecting an underlying manifold over complex data is tested. The experimental results demonstrate the effectiveness of the proposed approach.

The outline of the paper is as follows. In Section 2, the proposed approach is described in detail. The experimental evaluation is presented in Section 3. More precisely, in Subsection 3.1, the datasets used for the experiments are described and the results obtained are shown in Subsection 3.2. Conclusions are drawn and topics of future research are discussed in Section 4.

2 Method

2.1 Learning using Extremely Randomized Trees

Decision trees are typically constructed with a top-down induction method. Starting from the root node that is associated to the complete training set, the nodes are recursively split by applying a test on one of the features. In order to find the best split, all features and their corresponding split points are considered and a split quality criterion is evaluated. In supervised learning tasks, this criterion is often information gain (classification), or variance reduction (regression). When the data contained in a node is pure w.r.t. the target, or when some other stopping criterion holds, the node becomes a leaf node and a prediction is assigned to it. This prediction is the majority class assigned to the training instances in the leaf for classification, or the average of their target values for regression. The prediction for test instances is obtained by sorting them through the tree into a leaf node. In this work, the decision tree learners employed are set in the Predictive Clustering Tree (PCT) [9] framework, adopting the hierarchical clustering view of decision trees. PCTs are constructed by maximally reducing intra-cluster variance at each split. By computing the variance over the feature set, rather than the target, PCTs can be applied to (unsupervised) clustering tasks.

Since decision trees often have a large variance, their predictive performance can be improved by having several trees returning an aggregated prediction. Such a collection of decision trees is called an ensemble, and several instances of ensembles exist. In this work, we consider the ensemble method of Extremely Randomized Trees (ERT) [13]. In an ERT ensemble, each tree is constructed by considering only a random set of split candidates at each node. More precisely, a random subset of features is picked, and for each feature, a random split point is picked. From these candidates, the candidate yielding the best value for the split criterion is chosen. ERT was shown to have a better predictive performance than the more popular Random Forests [13].

2.2 Feature construction with extremely randomized trees

A new feature set is generated by applying ERT on the initial feature set. The nodes of each tree in the ERT setting, $\mathbf{N} = \{n_1, n_2, \cdots, n_{|N|}\}$ are treated as clusters containing all the samples that fall into them traversing the tree. Let $\mathbf{X} \in \Re^{|S|\times|M|}$ be the initial feature set and $\mathbf{F} \in \Re^{|S|\times|N|}$ the induced one, where $|S|$ and $|N|$ correspond to the number of samples and the number of the induced features of the dataset, respectively. Next, the clusters $n_j \in \mathbf{N}$ are treated as features of the feature set $\mathbf{F}$. Each $f_{ij} \in \mathbf{F}$ equals to 1 if the sample $i \in \mathbf{S}$ is contained in cluster (node) $j \in \mathbf{N}$ and 0 otherwise.

At this point, it has to be noted that a similar encoding could be produced by any hierarchical clustering method. However, the employment of ERTs is proved to be beneficial. First, ERT is a tree ensemble method, and therefore it is robust to small perturbations in the data. It is also robust to non-informative or noisy features due to the implicit feature selection mechanism. This way, the generated feature representation is considered less variant to noise. Moreover, another advantage is that the tree ensembles can treat both numerical and non-numerical values without pre-processing, making the method more easily applied and robust. In addition to that, in contrast to many other methods, it offers a natural way to deal with missing values by distributing instances with a missing split value over all branches or by selecting at random one branch to follow. Other advantages of the proposed approach is that it is parameter-free and it is performed in an inductive manner. After the training, the model can handle any new data without any need of the training set. This makes the application to modern online systems as well as systems that handle large scale data feasible.

3 Experimental Evaluation

3.1 Datasets

In this section, the experimental validation of the proposed approach is presented. For evaluation purposes, some well-known datasets from UCI repository [16] were employed. Including several datasets from various fields contributes in avoiding any biased conclusions and revealing the robustness of our method. The labels contained in these datasets were used only for evaluation purposes and were not included in any part of the learning process. In Table 1, further information about the used datasets is provided. A pre-processing step was also introduced as in [5]. In particular, out of simplicity and homogeneity, the data have been whitened by normalizing all features to have zero mean and unit standard deviation. Non-binary classification tasks were transformed into binary ones by considering the major class versus all the others or by grouping the classes to two sets of balanced size. Despite the fact that tree-ensembles do not require any pre-processing of the data, in order to compare the proposed feature representation to the original one the missing values were replaced by the features's average and the nominal features in some datasets were transformed into a set of binary ones using one-hot encoding.

In order to prove the efficiency of the proposed feature representation approach on more complex data structures, 2 interaction prediction datasets were also introduced [17]. They are interaction datasets representing homogeneous biological networks. In particular:

- **Metabolic network (MN)**[18]. This dataset consists of 668 S. cerivisiae enzymes and the predicted values are the catalysation of succesive reactions

Table 1. The datasets used in the evaluation procedure.

Dataset	Nb of Instances	Nb of Features
Pima Indians diabetes	768	8
Ecoli	336	7
Glass identification	163	9
Haberman's survival	306	3
Ionosphere	351	34
Iris	150	4
Libras movement	192	90
Robot Execution Failures (Lp5)	164	90
Mammographic mass	961	4(14)
Sonar	208	60
Spectf Heart	267	44
Statlog (Vehicle)	846	18
Breast cancer (orig.)	699	9
Breast cancer (diag.)	569	30
Wine	178	13
Breast cancer (prog.)	198	32

between two enzymes. In total, there are 2782 catalysations out of the $668^2 = 446224$ pairs of enzymes. Originally, 325 features are used for the predictions. They are a set of expression data, phylogenetic profiles and localization data.

- **Protein-protein interaction network (PPI)** [19]. It contains 2438 interactions between 984 S. cerivisiae proteins. The input features are also a set of expression data, phylogenetic profiles and localization data.

3.2 Results

Although we target unsupervised learning tasks, we first use a supervised set-up for the evaluation of the proposed feature construction technique. In particular, we use the class labels as ground truth and check the predictive performance of a k-NN classifier, trained using the induced features generated by the proposed approach, denoted as Extremely Randomized Clustering tree Paths (ERCP). The underlying idea is that instances with the same class should get a similar feature representation, even though that class information is not used in the construction of the features. We compare the performance with k-NN employed on the original feature set. Furthermore, totally random trees embedding [4] was also used in comparison. It was employed as an unsupervised transformation of the data, using a forest of Extremely Random Clustering trees (ERC). Similar to our approach, ERT was also chosen as the base estimator. As for k-NN, the 3 nearest neighbors were considered ($k = 3$).

The number of trees used in the ensembles for all the compared methods was set equal to 300, as it is generally an acceptable number for these tasks. In addition, at that number, the Gram matrix induced on the new features converged in the supervised setting [5]. The number of the features selected as splitting candidates 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 set equal to 3. The evaluation was performed in a 10-fold cross validation (10 CV) scheme.

The evaluation measures that were employed were the common accuracy and the area under the receiver operating characteristic curve (AUROC). A ROC curve is defined as the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. Alternatively, the true-positive rate is known as sensitivity and the false-positive rate as (1 - specificity).

Table 2. *AUROC* measures for the compared approaches.

Data	*original*	*ERC*	*ERCP*
Pima Indians diabetes	*0.767	0.726	0.731
Ecoli	*0.966	0.965	0.965
Glass identification	0.805	0.823	*0.871
Haberman's survival	0.629	0.609	*0.630
Ionosphere	0.897	0.937	*0.957
Iris	*1	*1	*1
Libras movement	0.753	*0.801	0.735
Robot Execution Failures (Lp5)	0.915	0.886	*0.968
Mammographic mass	0.791	*0.795	0.791
Sonar	0.718	0.713	*0.734
Spectf Heart	0.707	0.748	*0.779
Statlog (Vehicle)	0.981	*0.986	0.971
Breast cancer (orig.)	0.982	*0.983	*0.983
Breast cancer (diag.)	0.984	*0.985	0.977
Wine	0.970	*0.991	0.973
Breast cancer (prog.)	0.503	0.546	*0.590
Average	0.836	0.844	*0.854
Nb wins	3	7	9

As it is reflected in Tables 2 and 3 the proposed method *ERCP* outperforms *ERC* in terms of both *AUROC* and accuracy. For each dataset, the best result is indicated with $*$. Furthermore, both tree-based ensemble methods succeed in generating a better feature representation set than the original one. More precisely, the average *AUROC* results for *ERCP* and *ERC* are 0.854 and 0.844 respectively. On the original set the average drops to 0.836. When it comes to accuracy the same behavior was witnessed as the rates are 0.839, 0.822, and 0.817 for the *ERCP*, *ERC*, and the original set respectively.

In addition to k-NN, k-means was employed extending the evaluation of the proposed method to a clustering setting. The number of clusters was set equal to 2. The evaluation metric that was used was the adjusted Rand index, measuring

Table 3. Accuracy results for the compared approaches.

Data	*original*	ERC	$ERCP$
Pima Indians diabetes	*0.743	0.713	0.733
Ecoli	*0.961	*0.961	0.931
Glass identification	0.754	0.760	*0.791
Haberman's survival	*0.696	0.673	0.693
Ionosphere	0.841	0.904	*0.932
Iris	0.993	*1	*1
Libras movement	0.680	*0.746	0.682
Robot Execution Failures (Lp5)	0.799	0.751	*0.940
Mammographic mass	0.753	*0.755	0.753
Sonar	0.638	0.630	*0.720
Spectf Heart	0.722	0.737	*0.775
Statlog (Vehicle)	0.943	*0.957	0.925
Breast cancer (orig.)	0.960	*0.966	0.963
Breast cancer (diag.)	*0.964	*0.964	0.958
Wine	0.949	0.961	*0.962
Breast cancer (prog.)	*0.677	0.676	0.665
Average	0.817	0.822	*0.839
Nb wins	5	7	7

the similarity between the ground truth class assignments and the clustering algorithm assignments. As it is reflected in Table 4, the proposed method $ERCP$ slightly outperforms the other comparing methods.

In Figs. 1 and 2, a visualization of PPI and MN datasets is displayed by projecting the data in a 2-dimensional (2D) space using PCA. Other linear or nonlinear techniques such as the t-SNE [20] could have been used but the common PCA was chosen out of simplicity. For visualization purposes, the multi-label setting corresponding to PPI or MN datasets was projected in a single vector (1D) using PCA and different colors were assigned to its elements. As reflected in the Figs. 1 and 2, the generated data distribution after applying PCA to the original data fails to detect any underlying manifold and it is similar to a common random projection, especially for the MN dataset. The outcome is not significantly different in case of ERC. However, the application of PCA to the $ERCP$-induced feature space leads to a more informative distribution. To this end, it can be deduced that $ERCP$ succeeds in providing a more informative feature representation for complex datasets.

4 Conclusions and Future work

In this paper, we proposed an efficient feature representation framework based on tree ensembles for unsupervised learning tasks. In particular, we employed the setting of Extremely Randomized Trees in an unsupervised manner, transforming the data into a high-dimensional, very sparse feature space. The proposed

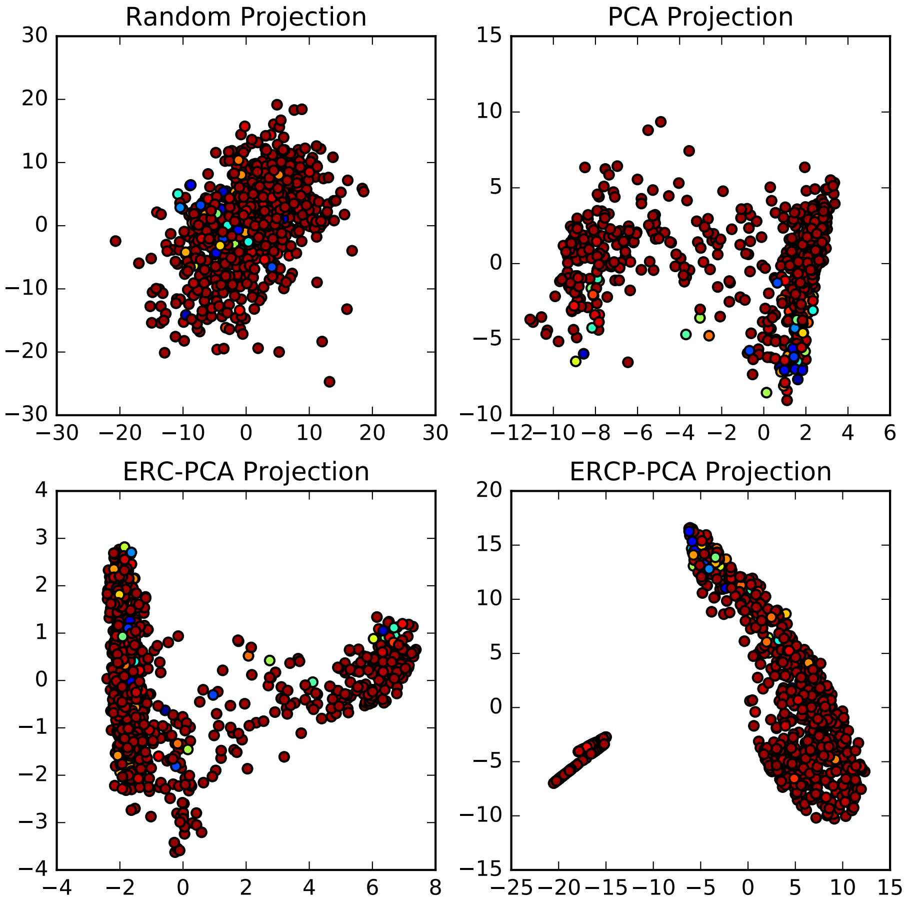

Fig. 1. PPI network data projection. Upper left a totally random projection of the data is depicted. Upper right the PCA projection of the original data is demonstrated. Down left the PCA projection of the *ERC* feature representation is displayed. Down right the PCA projection of the *ERPC* is displayed.

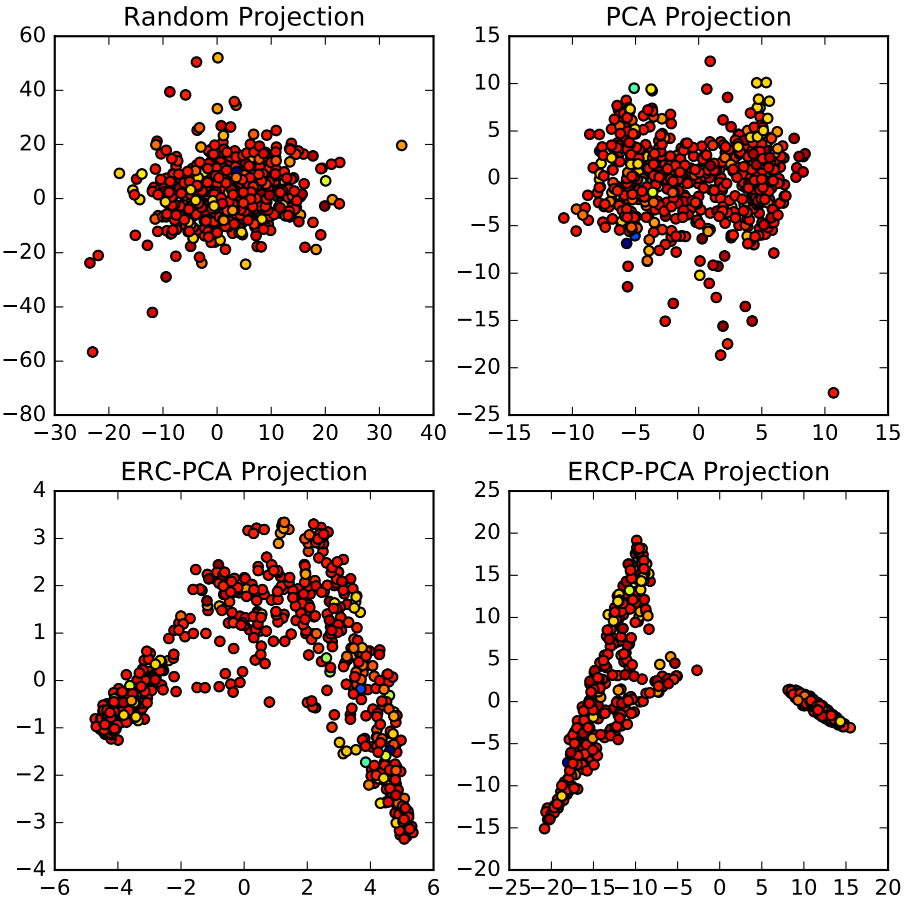

Fig. 2. MN network data projection. Upper left a totally random projection of the data is depicted. Upper right the PCA projection of the original data is demonstrated. Down left the PCA projection of the *ERC* feature representation is displayed. Down right the PCA projection of the *ERPC* is displayed.

Table 4. Adjusted Rand index results for the compared approaches.

Data	*original*	*ERC*	*ERCP*
Pima Indians diabetes	*0.11	0.07	0.04
Ecoli	*0.62	0.59	0.58
Glass identification	0	0	0
Haberman's survival	0	0	0
Ionosphere	0.17	0.17	*0.18
Iris	1	1	1
Libras movement	0	*0.01	0
Robot Execution Failures (Lp5)	0	0	*0.08
Mammographic mass	*0.36	0.30	0.31
Sonar	0	0	0
Spectf Heart	-0.1	-0.07	*-0.05
Statlog (Vehicle)	0.15	*0.16	0.15
Breast cancer (orig.)	0.84	0.83	*0.89
Breast cancer (diag.)	0.65	*0.71	0.70
Wine	0.01	*0.11	0.08
Breast cancer (prog.)	*0.02	0	*0.02
Average	0.24	0.24	*0.25
Nb wins	4	4	5

approach is parameter-free, inductive and can handle missing values as well as complex data structures. In contrast to former similar approaches, the effectiveness of the proposed approach was investigated in a completely unsupervised manner and we took into account the whole path that corresponds to a sample traversing each tree in the ensemble. For evaluation purposes, many multivariate UCI datasets were used to avoid any biased conclusions. More complex datasets that correspond to interaction networks were also employed from the field of biomedicine. The obtained experimental results reaffirmed the efficiency of the proposed framework.

Possible topics for future research include the application of various machine learning algorithms to the generated feature representation or the development of an efficient weighing scheme, assigning a different weight to each tree-node of the ensemble. This way, the information contained in each generated feature will be distilled.

References

1. Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., Jordan, M. I.: Learning the kernel matrix with semidefinite programming. Journal of Machine learning research, 5, 27–72 (2004).
2. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge university press (2004)

3. Burges, C. J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2, 2, 121–167 (1998)
4. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. in Proc. 20th Conf. on Neural Information Processing Systems (NIPS), 985-992 (2006)
5. Vens, C., Costa, F.: Random forest based feature induction. in Proc. IEEE 11th Int. Conf. on Data Mining (ICDM), 744–753 (2011)
6. Breiman, L.: Random forests. Machine learning, 45, 1, 5–32 (2001)
7. Zhang, M., Wu, L.: LIFT: Multi-label learning with label-specific features. IEEE Trans. on Pattern Analysis and Machine Intelligence, 37, 1, 107–120 (2015)
8. Blockeel, H., De Raedt, L.: Top-down induction of first-order logical decision trees. Artificial intelligence, 101, 1, 285–297 (1998)
9. Blockeel, H., De Raedt, L., Ramon, J.: Top-Down Induction of Clustering Trees. in Proc. 15th Int. Conf. on Machine Learning, 55–63 (1998)
10. Geurts, P., Irrthum, A., Wehenkel, L.: Supervised learning with decision tree-based methods in computational and systems biology. Molecular Biosystems, 5, 12, 1593–1605 (2009)
11. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. in Proc. ACM Int. Conf. on Web Search and Data Mining, 183–194 (2008)
12. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recognition, 46, 3, 817–833 (2013)
13. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning, 63, 1, 3–42 (2006)
14. Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. in IEEE trans. on Pattern Analysis and Machine Intelligence, 29, 1, 40–51 (2007)
15. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: a comparative. Journal of Machine Learning Research, 10, 66–71 (2009)
16. Asuncion, A., Newman, D. : UCI machine learning repository. [Online] Available: http://www.ics.uci.edu/ mlearn/MLRepository.html
17. 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)
18. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., Gardner, T. S.: Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biol, 5, 1, e8 (2007)
19. Von Mering, Ch., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein–protein interactions. Nature, 417, 6887, 399–403 (2002)
20. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605 (2008)