EndNote: Feature-based classification of networks

Ian Barnett[*],

University of Pennsylvania, Department of Biostatistics, Philadelphia, 19104, USA

Nishant Malik[*],

Rochester Institute of Technology, Rochester, 14623, USA

Marieke L. Kuijjer,

Dana Farber Cancer Instutute, Boston, 02115, USA

Peter J. Mucha,

University of North Carolina, Department of Mathematics, Chapel Hill, 27599, USA

Jukka-Pekka Onnela

Harvard University, Department of Biostatistics, Boston, 02115, USA

Abstract

Network representations of systems from various scientific and societal domains are neither completely random nor fully regular, but instead appear to contain recurring structural features. These features tend to be shared by networks belonging to the same broad class, such as the class of social networks or the class of biological networks. Within each such class, networks describing similar systems tend to have similar features. This occurs presumably because networks representing similar systems would be expected to be generated by a shared set of domain specific mechanisms, and it should therefore be possible to classify networks based on their features at various structural levels. Here we describe and demonstrate a new hybrid approach that combines manual selection of network features of potential interest with existing automated classification methods. In particular, selecting well-known network features that have been studied extensively in social network analysis and network science literature, and then classifying networks on the basis of these features using methods such as random forest, which is known to handle the type of feature collinearity that arises in this setting, we find that our approach is able to achieve both higher accuracy and greater interpretability in shorter computation time than other methods.

Keywords

social and biological networks; network classification; random forest

---

(ibarnett@pennmedicine.upenn.edu, (617)495-1000).

[*]These authors contributed equally.

1 Introduction

Past work in the area of network classification has primarily focused on distinguishing networks from different categories using two different approaches. In the first approach, network classification is carried out by examining certain specific structural features and investigating whether networks belonging to the same category are similar across one or more dimensions as defined by these features (Onnela et al., 2012; Pennacchiotti & Popescu, 2011; Ralaivola et al., 2005; Richiardi et al., 2011). In other words, in this approach the investigator manually chooses the structural characteristics of interest and more or less manually (informally) determines the regions of the feature space that correspond to different classes. These methods are scalable to large networks and yield results that are easily interpreted in terms of the characteristics of interest, but in practice they tend to lead to suboptimal classification accuracy. In the second approach, network classification is done by using very flexible machine learning classifiers that, when presented with a network as an input, classify its category or class as an output (Borgwardt & Kriegel, 2005; Gärtner et al., 2003; Horváth et al., 2004; Kashima et al., 2003; Kondor et al., 2009; Niepert et al., 2016; Ramon & Gärtner, 2003; Shervashidze & Borgwardt, 2009; Thoma et al., 2010; Yanardag & Vishwanathan, 2015). To somewhat oversimplify, the first approach relies on manual feature specification followed by a manual and subjective selection of a classification system, whereas the second approach is its opposite, relying on automated feature detection followed by automated and objective classification. While the latter approach can yield very accurate class predictions and is capable of identifying important yet unseen network characteristics, its computational cost typically scales poorly and the potentially opaque nature of the methodology may make it difficult to interpret the obtained results. Note that even though we draw a distinction between manual and automatic feature selection, automatic methods, such as those that rely on convolutional neural networks, also require some measure of subjectivity when it comes to choosing the network architecture of the classifier, which can have a large impact on its performance (Saxe et al., 2011).

This paper presents a third, hybrid approach to the network classification problem. We first specify network features of interest manually and then use existing automatic methods, such as random forests, to carry out the classification task using these features. In other words, our approach uses manual feature selection followed by automated classification. This approach enables one to leverage domain specific knowledge to specify a much broader set of relevant features. These features might be based on some standard network characteristics, such as vertex degree, betweenness centrality, or motif counts, but they can also incorporate nodal attributes, such as the sex or age of a person in a social network. It is possible to incorporate even richer information, such as data related to the functional or dynamic state of the nodes and edges. For example, in the context of network epidemiology, the frequency with which a node changes state from susceptible to infected in a contact network in the course of a spreading process could be used as a predictor. Since most classifiers assign importance scores to the features that are used in classification, the resulting organization of networks can be readily interpreted in terms of these features and their importances.

Our approach to the network classification problem is scalable and easily interpretable. For example, the computational complexity of classifying n networks using graph kernels, a technique that falls under the second (fully automated) approach, is $O(n^2)$, while the computational complexity of random forests, used in our approach, is $O(n \log n)$. The approach also leads to remarkably high classification accuracy as we demonstrate by discerning different days of the week in unipartite social communication networks, by distinguishing between different tumor body sites in bipartite biological transcription factor-gene regulatory networks, and by testing the methodology on a collection of network classification benchmarks. It is usually not clear *a priori* in our approach what the best network features might be for classification as they are expected to depend on the domain of the network. In cases where there is no domain-specific knowledge of the pertinent features to include, we recommend liberal inclusion of a wide variety of features, letting the classifier determine which features are pertinent. Not all classifiers are equally suited to the task. In particular, many network properties are related to one another and their collinearity can cause problems when they are used as predictors. This calls for a classifier that handles collinear predictors well.

2   Method and data description

We studied three different types of networks. First, to demonstrate classification on social networks, we constructed and performed classification of daily communication networks using call detail records from the largest telecom operator (57% market share) in a European country. This identified social patterns corresponding with certain days of the week. Second, to demonstrate classification on biological networks, we used regulatory networks from tumor cells from patients with either lung (lung adenocarcinoma), brain (glioblastoma multiforme), or ovarian (ovarian serous cystadenocarcinoma) cancer. For each sample, we constructed a bipartite network of 10,903 genes and 113 transcription factors with edge weights corresponding to the strength of regulation between them (Kuijjer et al., 2015). Third, we investigated a variety of network classification benchmarks, including internet-based ego-centric movie actor networks constructed from the Internet Movie Database (IMDb). For these benchmarks, we used network features to classify the forum thread networks by their topic and the actor networks by the movie genre.

Within each of these families of networks, we performed classification by a two-step process (Fig. S1): in the first step, we select and calculate network features that are potentially pertinent to the classification problem for each network; in the second step, we train and test a classifier built upon these features. Importantly, because features are first selected manually based on available information and then further refined by the classifier, the set of features used for accurate classification varies depending on the family of networks of interest. The common network features that we use in each setting included average degree, global clustering coefficient, degree assortativity, and network size, alongside specialized context-specific network features detailed in the Supplementary Materials.

After selecting the network features, the second step requires choosing the appropriate method for classification. We used three popular classification techniques: k-means, k-nearest-neighbors (KNN), and random forests. The k-means and KNN can have difficulty

when features are high dimensional and strongly collinear. On the other hand, random forests account for correltaed features and use a combination of multiple rectangular regions which allow for more flexibility with feature space partitioning. For each approach, the classifier is trained on a randomly selected subset of the networks, and classification accuracy of each approach is tested using the remaining networks. This randomization is performed repeatedly and the average across randomizations is the classification accuracy.

3 Results and discussion

None of the classifiers we used had difficulty separating weekends from weekdays in the phone-based social / communication networks, with all methods achieving greater than 95% prediction accuracy (Fig. 1). In the random forest classifier, the fraction of edges connecting individuals residing in the same zip code was the most important feature based on mean decrease impurity being 4.5 times more important than the average feature used in classification (Fig. S2). On the weekends, there was a clear increase in the proportion of ties that connect people from the same zip code. The second most important feature was network size, reflecting the marked decrease in the number of phones used on the weekend as compared to weekdays. The average age difference over all network edges, which quantifies age-based network assortativity, was approximately one year greater on weekends compared to weekdays, leading to this feature being third in importance. This near perfect accuracy by different classifiers indicates that the features for weekdays and weekends are easily distinguished from one another.

Distinguishing tumor types based on their regulatory networks proved to be a more difficult task. The random forest classifier had an overall prediction accuracy of 68% compared to the 62% prediction accuracy of the KNN classifier using the same set of features. The most important feature in the random forest classification of the tumor samples was degree assortativity in the projected gene-gene unipartite network, at 1.6 times as important as the average feature. However, in contrast to the phone-based social network where a subset of the features were clearly driving the results, in this case there was a more uniform contribution from all selected features. As seen in Fig. 2(a), one of the lung tumor tissue samples was classified alongside the ovarian tumor tissue samples. This demonstrates how classification can be used to identify outliers to be checked for potential mislabeling.

Random forests similarly outperformed KNN in the benchmark classification problems, with an average prediction accuracy margin of 3% across the six benchmarks in Table 1. Moreover, and remarkably, our hybrid approach of manually selecting features and using random forests to automatically select their importance also outperformed three recently developed and significantly more complicated and computationally-intensive approaches to graph classification, namely graph kernels (Kondor & Lafferty, 2002), deep graph kernels (Yanardag & Vishwanathan, 2015), and convolutional neural networks (Fukushima, 1980) (results for each reported in Niepert et al., 2016). Given the value of domain specific knowledge for selecting and interpreting prospective features of importance, and given that random forests are easily trained on large data sets and allow for easy interpretation of results, our hybrid approach combining manual specification of features followed by

automated classification on the selected features, appears to have a significant advantage in terms of precision of classification, cost of computation, and ease of interpretation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

Bibliography

Borgwardt Karsten M, & Kriegel Hans-Peter. (2005). Shortest-path kernels on graphs. *Pages* 8–*pp of*: Fifth ieee international conference on data mining (icdm'05) IEEE.

Fukushima Kunihiko. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193–202. [PubMed: 7370364]

Gärtner Thomas, Flach Peter, & Wrobel Stefan. (2003). On graph kernels: Hardness results and efficient alternatives. *Pages* 129–143 *of*: Learning theory and kernel machines Springer.

Horváth Tamás, Gärtner Thomas, & Wrobel Stefan. (2004). Cyclic pattern kernels for predictive graph mining. *Pages* 158–167 *of*: Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining ACM.

Kashima Hisashi, Tsuda Koji, & Inokuchi Akihiro. (2003). Marginalized kernels between labeled graphs. *Pages* 321–328 *of*: Icml, vol. 3.

Kondor Risi, Shervashidze Nino, & Borgwardt Karsten M. (2009). The graphlet spectrum. *Pages* 529–536 *of*: Proceedings of the 26th annual international conference on machine learning ACM.

Kondor Risi Imre, & Lafferty John. (2002). Diffusion kernels on graphs and other discrete input spaces. *Pages* 315–322 *of*: Icml, vol. 2.

Kuijjer Marieke Lydia, Tung Matthew, Yuan GuoCheng, Quackenbush John, & Glass Kimberly. (2015). Estimating sample-specific regulatory networks. arxiv preprint arxiv:1505.06440.

Niepert Mathias, Ahmed Mohamed, & Kutzkov Konstantin. (2016). Learning convolutional neural networks for graphs. *arxiv preprint arxiv:1605.05273*.

Onnela Jukka-Pekka, Fenn Daniel J, Reid Stephen, Porter Mason A, Mucha Peter J, Fricker Mark D, & Jones Nick S. (2012). Taxonomies of networks from community structure. Physical review e, 86(3), 036104.

Pennacchiotti Marco, & Popescu Ana-Maria. (2011). A machine learning approach to twitter user classification. Icwsm, 11(1), 281–288.

Ralaivola Liva, Swamidass Sanjay J, Saigo Hiroto, & Baldi Pierre. (2005). Graph kernels for chemical informatics. Neural networks, 18(8), 1093–1110. [PubMed: 16157471]

Ramon Jan, & Gärtner Thomas. (2003). Expressivity versus efficiency of graph kernels. *Pages* 65–74 *of*: First international workshop on mining graphs, trees and sequences Citeseer.

Richiardi Jonas, Achard Sophie, Bullmore Edward, & Van De Ville Dimitri. (2011). Classifying connectivity graphs using graph and vertex attributes. *Pages* 45–48 *of*: Pattern recognition in neuroimaging (prni), 2011 international workshop on IEEE.

Saxe Andrew M, Koh Pang Wei, Chen Zhenghao, Bhand Maneesh, Suresh Bipin, & Ng Andrew Y. (2011). On random weights and unsupervised feature learning. *Pages* 1089–1096 *of*: Icml.

Shervashidze Nino, & Borgwardt Karsten M. (2009). Fast subtree kernels on graphs. *Pages* 1660–1668 *of*: Advances in neural information processing systems.

Thoma Marisa, Cheng Hong, Gretton Arthur, Han Jiawei, Kriegel Hans-Peter, Smola Alex, Song Le, Yu Philip S, Yan Xifeng, & Borgwardt Karsten M. (2010). Discriminative frequent subgraph mining with optimality guarantees. Statistical analysis and data mining, 3(5), 302–318.

Yanardag Pinar, & Vishwanathan SVN. (2015). Deep graph kernels. *Pages* 1365–1374 *of*: Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining ACM.
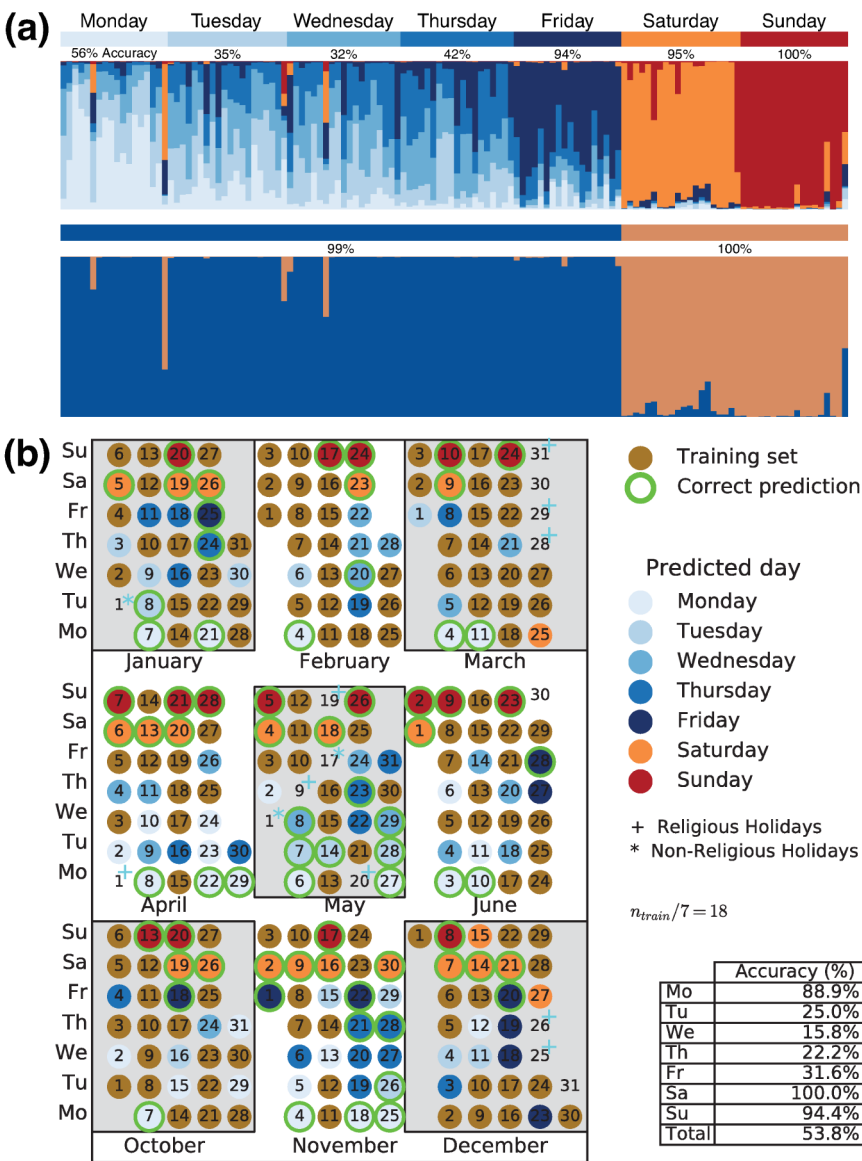
Fig. 1.
(a) **Random forest classification of days of the week**: Using odd-numbered days of the data set for training, the classification of each even day is displayed as a column. The performance of the 7-day classifier is displayed in the top row with the binary weekend/weekday classifier in the bottom row. Each column represents the color-coded probabilities of a day being classified as a day of the corresponding color. In the top row, a day is correctly classified if that day has the largest classification probability. For the bottom row, the larger of the two binary classification probabilities is used to guide the classification. All nationally-recognized holidays were removed from both the training and testing data sets as they would be expected to have unusual social dynamics. (b) **KNN classification of days of the week**: This visualizes a single realization of classification of days of the week using KNN, where n_{train} is the total number of days used for the training set, which included equal number of days of each day of the week.
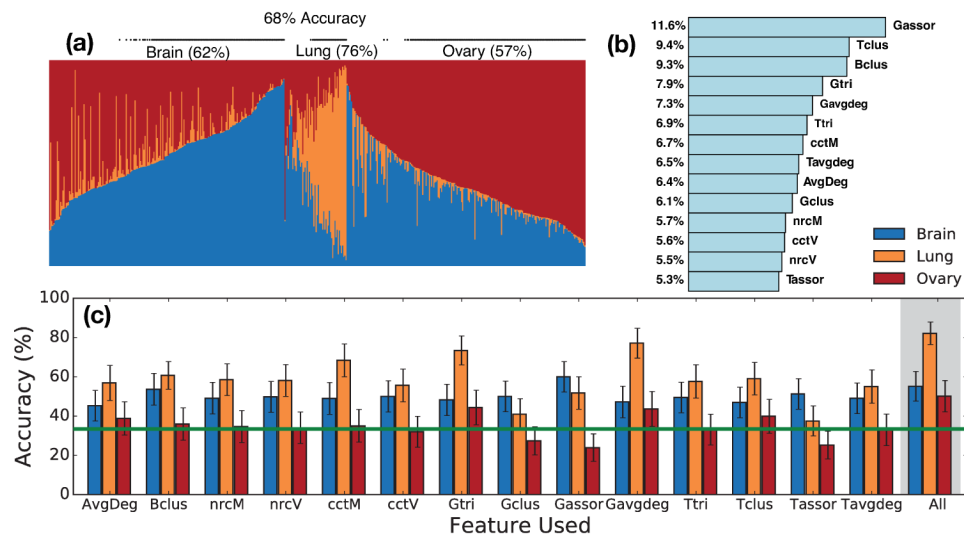
Fig. 2.

(a) **Random forest classification of cancer types**: Each of the 483 columns represents random forest classifier probabilities as stacked bars for a tissue sample in the test set with blue, orange and red bars representing probabilities assigned to brain, lung, and ovary cancers, respectively. Each sample is then classified by the largest of these three probabilities, and correct classification is indicated by a black dot above the corresponding column. Overall, 68% of all tissue samples were correctly classified (cancer specific classification accuracies are shown in the figure). (b) **Feature importance in the tumor type classification random forest**: Feature importance is calculated from the mean decrease in tree leaf impurity over the full random forest as measured by the Gini index. Percentages are the decrease in impurity for each feature, scaled so they sum to 100%. Detailed descriptions of the variables are provided in the Supplementary Materials. (c) **KNN classification of cancer types**: Each set of bars represents accuracy of the three cancer types using a single feature. The last block, highlighted in grey, represents accuracy using all selected features. The green line represents the null rate of classification. Overall, 62% of tissue samples were correctly classified using this method. Error bars indicate the standard deviation of mean accuracy over 10,000 randomizations into training and testing sets. The features used are average degree in the bipartite network (AvgDeg), average bipartite clustering coefficient in the bipartite network (Bclus), mean closeness centrality in the bipartite network (cctM), variance of closeness centrality in the bipartite network (cctV), degree assortativity in the gene projection network (Gassor), average degree in the gene projection network (Gavgdeg), average clustering coefficient in the gene projection network (Gclus), number of triangles in the gene projection network (Gtri), mean node redundancy in the bipartite network (nrcM), variance of node redunancy in the bipartite network (nrcV), degree assortativity in the TF projection network (Tassor), average degree in the TF network (Tavgdeg), average clustering coefficient in the TF projection network (Tclus), and number of triangles in the TF projection network (Ttri).

Table 1.

Classification accuracy for benchmark social network data sets.

Results expressed as % from 10-fold cross-validation to obtain out-of-sample accuracy estimates and their standard deviations for the random forests (RF) and KNN classifiers with manual feature selection used here, compared to results for Graph Kernels (GK), Deep Graph Kernels (DGK), and convolutional neural networks (PSCN). The approach we introduce here, using manually selected network features in combination with a classifier, generally performs well; when using the RF classifier, the approach systematically outperforms others for each data set.

Data set	RF	KNN	GK	DGK	PSCN
COLLAB	**76.5±1.68**	72.69±0.80	72.84±0.28	73.09±0.25	72.60±2.15
IMDB-BINARY	**72.4±4.69**	67.03±1.90	65.87±0.98	66.96±0.56	71.00±2.29
IMDB-MULTI	**47.8±3.55**	42.40±2.70	43.89±0.38	44.55±0.52	45.23±2.84
REDDIT-BINARY	**88.7±1.99**	87.63±0.82	77.34±0.18	78.04±0.39	86.30±1.58
REDDIT-MULTI-5K	**50.9±2.07**	49.04±0.77	41.01±0.17	41.27±0.18	49.10±0.70
REDDIT-MULTI-12K	**42.7±1.28**	38.21±0.49	31.82±0.08	32.22±0.10	41.32±0.42