UROP RESEARCH PROJECTS

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

NETWORKS CLASSIFICATION

FEATURE SELECTION

SVM SUPERVISED LEARNING

Pure Network

C=14, Linear Kernel, Split Ratio: 7:3

	BA	DD	ER	LA	REG	SW
BA	100					
DD		100				
ER			100			
LA				100		
REG					100	
SW			10.61			89.39

The confusion matrix with average classification accuracy of 98.06%. Row labels indicate the ground truth and entries identify the percentage of this type of network that was classified as each of the column labels.

Hybrid Network

C=14, Linear Kernel, Same Train & Test data

	BA	DD	ER	LA
BA	94.23	0.96		4.81
DD		100		
ER			100	
LA	1.3		1.3	97.4

Testing SVMhyb on the hybrid networks. The confusion matrix with average classification accuracy of 97.78%.10-fold cross validation is used and 90% of the dataset is selected for training in each fold, and the averages are reported.

MAX-RELEVANCE MIN-REDUNDANCY (MRMR)

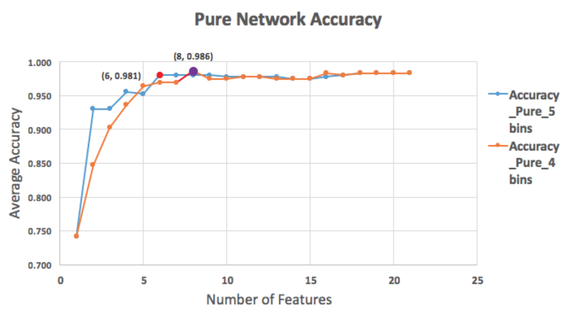
Discretize the Continuous Features

Equal Interval provides poor accuracy results, hence we inspected there existed extreme numbers. Discretize by Quantile is a good approach.

Entopy

Measure Mutual Information.

Pure Network



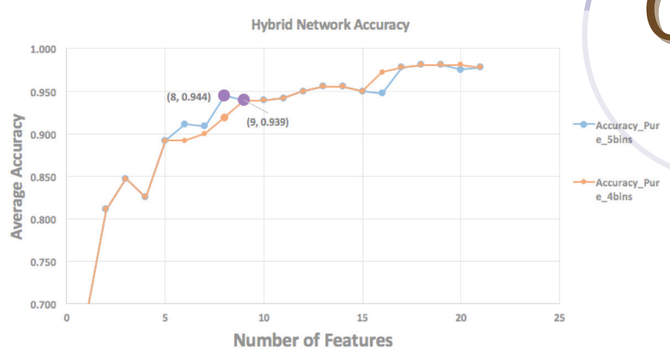
1) Four-bins discretization method had a higher accuracy than five-bins.

2) Five Bins: (0, 20%, 40%, 60%, 80%, 100%) Six features can achieve 98.06% accuracy while 21 features' accuracy is 98.33%.

3) Four Bins: (0, 25%, 50%, 75%, 100%) Eight features can achieve 98.6% accuracy.

RANK	mRMR 4_Bins	mRMR 5_Bins	SVM RFE (Weight)	Decision Tree
1	σ_{cc}	σ_{cc}	σ_{cc}	σ_{cc}
2	M_{cl}	μ_{cl}	σ_{ds}	σ_{ds}
3	M_{ds}	σ_{bc}	m_{ds}	M_{bc}
4	m_{bc}	σ_{ds}	σ_{sv}	μ_{cl}
5	μ_{cl}	σ_{cl}	M_{bc}	m_{bc}
6	M_{bc}	m_{bc}	μ_{cl}	m_{sv}
7	σ_{cl}	M_{ds}	μ_{sv}	m_{cc}
8	σ_{ds}	M_{cl}	M_{cc}	μ_{bc}
AVERAGE ACCURACY	98.61%	98.06%	98.61%	98.06%

Hybrid Network



1) Five-bins discretization method had a higher accuracy than four-bins.

2) Five Bins: (0, 20%, 40%, 60%, 80%, 100%) Eight features can achieve 94.4% accuracy while 21 features' accuracy is 97.8%.

3) Four Bins: (0, 25%, 50%, 75%, 100%) Nine features can achieve 93.9% accuracy.

Decreasing Order of Selected Features:

1) Four Bins: 'std_cc' 'mean_cl' 'M_bc' 'M_ds' 'm_sv' 'm_cl' 'std_ds' 'm_ds' 'std_cl' 'std_bc' 'm_bc' 'M_cl' 'M_cc' 'M_sv' 'diameter' 'std_sv' 'mean_sv' 'm_cc' 'mean_ds' 'mean_cc' 'mean_bc'

2) Five Bins: 'std_cc' 'mean_cl' 'M_bc' 'M_ds' 'm_sv' 'm_ds' 'm_cl' 'std_cl' 'std_ds' 'std_bc' 'm_bc' 'M_cl' 'M_cc' 'M_sv' 'diameter' 'mean_sv' 'std_sv' 'm_cc' 'mean_ds' 'mean_bc' 'mean_cc'

LDA UNSUPERVISED CLASSIFICATION

Latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora or pictures. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. LDA works well in the unsupervised learning fields, hence we implemented in the context of network classification. For unsupervised classification, we separately trained both hybrid and pure network sets without any ground-truth network type labels. We used V-measure as the evaluation of the unsupervised learning algorithm, and further compared the LDA reults against the other four methods mentioned in the paper.

V-Measure Results Comparison

Train LDA on pure/hybrid network data and test against true labels of pure/hybrid network data. Unsupervised LDA was trained only on 4 types of networks(pure/hybrid): bb, da, er and la.

	Pure	1% Hybrid	5% Hybrid	10% Hybrid	20% Hybrid	30% Hybrid
K-Means	0.580	0.239	0.224	0.198	0.164	0.097
Spectral Clustering	0.742	0.638	0.378	0.443	0.191	0.137
GMM	0.729	0.352	0.278	0.227	0.224	0.076
DBSCAN	0.801	0.513	0.477	0.417	0.272	0.146
LDA (All Features)	0.737	0. 182	0.281	0.282	0.180	0.191
# of Learned Clusters	6	4	5	5	5	10
LDA (Selected Features)	0.689	0.376	0.367	0.326	0.260	0.226
# of Learned Clusters	4	5	4	5	7	4

Discussion

LDA clustering with all 21 features on pure network data ranks in the second place among five algorithms. However, LDA has a worse performance in hybrid network classification with all features.The low V-measure score may indicate the overfitting of too many learned features.

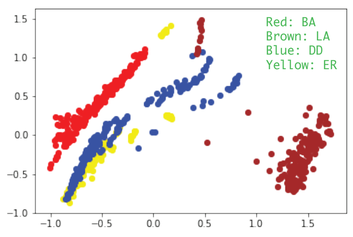
Later, we trained LDA with the top eight features selected from the mRMR method. The V-measure scores had significant increment especially in hybrid network clustering. Moreover, LDA V-scores are more stable over different level of contaminated networks, and it outperforms the other four in 30% hybrid network data.

Feature Selected

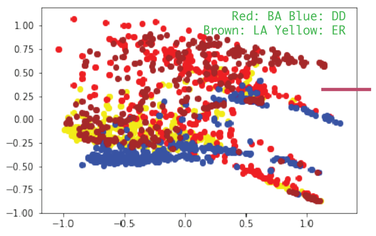
Hybrid Networks: Use the selected featuresa as following: 'std_cc' 'mean_cl' 'M_bc' 'M_ds' 'm_sv' 'm_cl' 'std_ds' 'm_ds' 'std_cl'

2D Projected Visualisation

Ground truth of pure network data

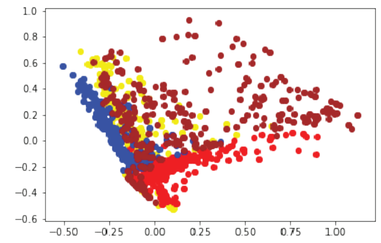


Ground truth of all hybrid networks

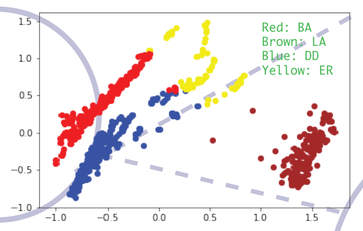


The actual data cannot be well distinguished by the given features as shown in the figure.

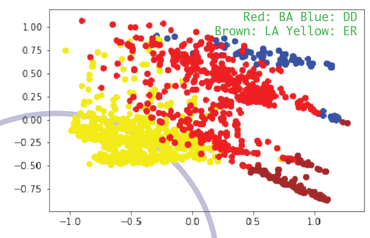
Ground truth of all hybrid networks with selected feature



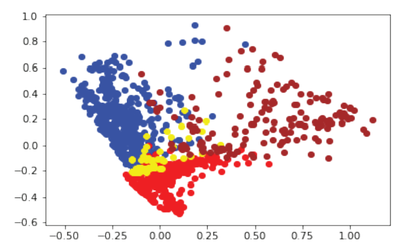
Clustering result of pure networks



Clustering result of all hybrid networks



Clustering result of all hybrid networks with selected features



INTRODUCTION

Network science has largely followed a scientific approach, namely focusing on developing new statistics with which to characterize networks and new (in-silico) experiments to generate observations. In the past decades, studying systems as networks has provided tremendous insight into various domains such as the organization and robustness of some of our most critical infrastructures, including the properties of the Internet; the competition and financial stability of global markets through the ties between corporations; the role of transcription factors in coordinating protein regulatory networks; extracting functional clusters within the brain; the rich interactions present in both online and offline social interactions, such as political affiliation, dynamics of the blogosphere, and the spread of happiness; and extinction in ecosystems.

Significant interest lies in understanding the causal network structures that identify and differentiate networks. Establishing that two seemingly different types of networks have an underlying similarity can help us study and understand one network by investigating the other. Importantly, networks that are seemingly very different may exhibit such fundamental similarities. Interest in determining these connections between network types has motivated several significant contributions to network science.

In this work, we studied the classification of different types of networks from a data-driven perspective. A set of twenty-one network properties were selected as our features of interests. These features includes network diameter; mean, standard deviation, maximum and minimum value of betweenness centrality, clustering coefficient, closeness centrality, degree distribution and singular values. These are the most representative network statistics. The raw data contains both pure and contaminated (or 'hybrid') networks, which fall in to six categories, BA, DD, ER, LA, REG and SW. Features were extracted from these networks and were normalized. Usually, the pure networks are easier to classify while a low rate of contamination can largely influence the accuracy.

FUTURE WORK

Feature selection

In current studies the features are mostly representing the global structure of the network. More local features that describe microstructures could be added in for a better representation of the network, such as the number of triangles, quadrangles or pentagons inside the network connection.

LDA

The parameter of the LDA model could be further tuned to test for the accuracy. Especially for hybrid network, which can be seen from the graph that DD and ER are difficult to differentiate.
Moreover, various posterior inferences could be tested out.

Other unsupervised models

In general, the accuracy of unsupervised models are lower than supervised model. However, there could still be other unsupervised models that could produce higher score than the current one.

INSTRUCTOR|Prof.Selin Damla Ahipasaoglu|STUDENTS|Yiran Wang (ESD 2017)|Ziqi Guo (ESD 2017)