CS 6604 Spring 2017

Final Term Project Report

Team: Social Communities

Team Members:
Prashant Chandrasekar
Mostafa Mohammed*
{peecee, profmdn}@vt.edu

06/28/2017
Virginia Tech
Blacksburg, VA, 24061

Instructor: Dr. Edward A. Fox

* Was part of the team for the first 3 weeks.

Social Communities Knowledge Discovery: Approaches applied to clinical study

Prashant Chandrasekar
Virginia Tech
peecee@vt.edu

ABSTRACT

In recent efforts being conducted by the Social Interactome team, to validate hypotheses of the study, we have worked to make sense of the data that has been collected during two 16-week experiments and three Amazon Mechanical Turk deployments. The complexity in the data has made it challenging to discover insights/patterns. The goal of the semester was to explore newer methods to analyze the data. Through such discovery, we can test/validate hypotheses about the data, that would provide a direction for our contextual inquiry to predict attributes and behavior of participants in the study. The report and slides highlight two possible approaches that employ statistical relational learning for structure learning and network classification. Related files include data and software used during this study; results are given from the analyses undertaken.

1 INTRODUCTION

The Social Interactome (SI) is an ongoing research study of the Addiction Recovery Research Center at VT Carilion Research Institute and the Computer Science and Statistics departments at Virginia Tech (VT). [1] The high-level aim of the study is to understand/identify properties of an online social network-based environment that aids the efforts of participants along their road to recovery from substance abuse. This clinical-trial based study involves a set of experiments where data is compared for a test and a control group. In the first two replicas of experiment 1, each group contains 128 participants, recruited from the International Quit and Recovery Registry (IQRR) and other online support groups. [2] We also recruit participants from Amazon Mechanical Turk for smaller scale experiments.

The inquiries on the data are, logically, driven by a hypothesis defined prior to each experiment. This leads to researchers computing histograms (or frequency tables), correlation coefficients, and difference in means tests (such as t-tests), as well as running survival analysis to test the hypothesis. Most of these tests are conducted on a set of 5-10 variables. Given that we are rich in information -- about the participants' demographic information, history of addiction, history of recovery, and other psychology-based scores -- that pertains to addiction and how it has impacted their life, along with textual and web analytics data from the website, it is critical that we conduct exploratory data analyses that help make sense of the data and help map the state of participants in the experiment as well as monitor change in states. This may or may not overlap with the aforementioned efforts towards testing the hypothesis of the experiment. But it would certainly help bring a 360-degree perspective on all the high quality, self-reported information to identify latent variables in the data.

2 LITERATURE REVIEW

There are many approaches that have been considered as a way forward for analyzing data for the task of prediction.

The approaches are dependent on what we're trying to predict and what information could we use to extract features. The "holy grail" in our study would be to predict if and when a participant is about to relapse. But, apart from that, we have many categorical variables that define the psychological state of the participants. Additionally, we can try to analyze the online behavior of the participants and try to find causes of particular social patterns.

During the early stages of the study, a group of undergraduate students with the help of co-PIs did some initial classification of textual content to see if participants talking about a successful event in their life, related to their efforts to recover from addiction.[1*] Using the data from the testimonials found on the International Quit and Recovery Registry, they manually labelled each paragraph in each testimonials with two labels "success" and "failure", depending on whether the content was describing a successful event or an unsuccessful event or failure. They trained many models on the data and built the test set using the actual posts of the participants of the Social Interactome study. The classification model did not perform very well because we learnt that the language used by participants in the study is different from the language in the testimonials (which were used to train the model). Therefore, a next step was to understand the textual content of the participants in more detail.

Kim Yoon developed an approach to use convolutional neural-networks for sentence classification [8]. This approach could be used for predicting if a sentence or group of sentences, written by the participants, are about a "successful event" or "unsuccessful event".

More recently, the discussion of analyzing the data has been centered around modeling a structure that represents how the different types of data, that we collect from participants, relate to one another (with the structure describing the relationships between them). Naturally, one thinks of a graphical representation for the structure, and probabilistic graphical models as an approach to conduct structure learning and inference. Learning the structure of a graphical model is a research area in itself. Netrapalli, Praneeth, et al. provide a greedy approach to learning the structure of all the

[*] 1 Abigail Bartolome and Victoria Worrall were the pair of undergraduate students who worked on the classification task.

discrete variables by learning an optimal neighborhood of a node, sequentially [9]. Though computationally complex, the methodology used, and other related methods, can be used to learn the structure of the various variables that we collect from each experiment in the study. Davis and Domingos have also proposed a method for learning a Markov structure [10]. The understanding and analyses of the algorithms proposed in these two papers appear to be preliminary.

We describe two approaches that we have identified as avenues, using which we will be conducting exploratory analyses.

Approach 1: Within-Network Classification using Relational Learning

The papers by Macskassy and Provost highlight the importance of statistical relational learning approaches and how they relate to "network data" mining. [3][4][5] The research supports the notion that it is more beneficial to utilize the "relation" between entities, rather than extract the entities and attempt to classify by treating them as i.i.d. One of the main contributions of the paper is a classification model that takes advantage of relational information, in addition to attribute information for entity classification.

NetKit-SRL is a modular toolkit, also discussed in their papers, that achieves classification for attributes of entities that are connected to entities for which a class label is known. The framework combines three components -- non-relational learning, relational learning, and collective inference procedures -- through which one could understand and infer various properties of networked data based on node and edge-based properties/attributes. The non-relational learning is used to construct the priors for the process. The key challenge in using this framework is in designing and describing the various links in the graph, based on which the model is constructed. As the papers describing the toolkit suggest, link selection is as critical as feature selection in traditional classification applications. Also, the labels to be predicted have to be categorical. We tested this framework on our data. In sections 3 and 4, we describe our experiment parameters and results.

Approach 2: Learning of Markov Logic Networks

Dr. Pedro Domingos and colleagues at the University of Washington have a different approach based on statistical relational learning. [6][7] Their efforts and corresponding contributions are directed towards using relational information to build Markov Logic Networks. A Markov Logic Network, simply put, is a Markov network containing feature spaces that are predicates and formulas expressed in first-order logic. Unlike traditional learning of first-order logic using the SAT algorithm and enforcing hard-constraints on formulas (such as: if the "world" violates a formula, that world is impossible), inference and learning is done on the Markov network, wherein when a "world" violates a formula, that world is "less probable". Therefore, in a Markov Logic network, each formula is given a weight (or alternatively, a weight can be learnt). This, along with a set of constants, helps define and represent a Markov network. This approach is different from the previous approach in that we infer causal relationships between sub-categories in the data. For all the types of data that we collect about the participants, we would like to understand and discover the underlying structure of dependencies/independencies among them. These dependencies/independencies are represented using graphical models. The contribution of their efforts also includes a framework, known as Alchemy, for building Markov Logic networks, and includes examples for using these networks for various tasks such as logistic regression, entity resolution, semantic processing, etc. We plan on using the Alchemy framework for building Markov networks to learn interaction effects of the different variables in our data.

3 EXPERIMENT ON NETWORK-CLASSIFICATION

A graph was constructed to evaluate Netkit-SRL [4] and understand how the classification models in the system performed against the data. The data that was used for this experiment was derived from replicate 2 of the first 16-week experiment of the Social Interactome study. The nodes in the graph correspond to the participants of the experiment. During the experiment, participants have the opportunity to read up on relevant news stories and testimonials and to view educational modules, among other web resources. Each of these is loosely categorized based on the various substances one could be addicted to. Using this loose characterization, the edges between a pair of nodes or participants represents the stories and educational modules that they read/viewed in common. The weight of the edge represents the number of times an overlap occurred.

This experiment was designed to test the hypothesis of homophily influencing recovery. Homophily suggests that people with common backgrounds express shared interests. We are testing the reverse of this: Given shared interest, can we predict the homophily measure? We consider three types of shared characteristic: 1) Addiction, 2) Education level, 3) Income. These specific measures were selected based on the analysis of survey responses on IQRR, where people reported these factors, along with Age and Gender, as the most important characteristics for deciding whether to add someone as a friend to their social network. These measures were added as attributes to the nodes of the graph.

Table 1 describes the frequency table of substances that the participants identify as the primary substance that they are recovering from. As is shown, the reported substances are heavily skewed. This isn't the ideal case, as we would want an equal proportion of representation for all classes (in this case substance that participants are recovering from) when training a classification model. As will be seen later, this will have an effect on the performance of the classifier.

Table 1: Frequency table for substances that participants have reported as their primary addiction (that they are recovering from)

Primary Substance In Recovery From	Frequency
Alcohol	139
Opioids	41
Stimulants	30

Cocaine	18
Prescription Pain Killers	17
Cannabis	7
Other	1
Nicotine	1

Tables 2 and 3 similarly show the frequency of income and education level, respectively, as reported by the participants. As mentioned above, our task is to run the network classification algorithm/model with these three variables (addiction, income and education) as class labels.

Table 2: Frequency table for income that participants have reported

Income	Frequency
< 30,000	101
< 50,000	45
< 70,000	38
< 150,000	26
< 90,000	22
> 150,000	13
Other	11

Table 3: Frequency table for education level that participants have reported

Education	Frequency
Other	89
Bachelors	63
Masters	37
Associate Degree	33
Diploma	24
Doctoral	7
High School	3

The experiment can be run with many configurations. Table 4 gives some of the parameters and possible options for each of them.

Table 5: Accuracy of the classification task for modeling to predict addiction substance for various combinations. We did not configure on the local classifier and we used uniform-priors for all.

Relational Classifier/Collective Inference Method	Relax-ation Labeling	Gibbs Sampling	Iterative Classification
Weight Vote Relational Neighbor	0.36601	0.37908	0.39216
Class-Distributional Relational Neighbor	0.1586	0.22222	0.18954

4 EXPERIMENT RESULT

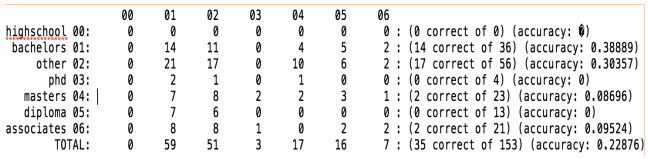
Table 5 shows the accuracy for the task of predicting addiction of participants, given the graph. Figures 1, 2 and 3 show the confusion matrix for the prediction task for addiction, education level, and income.

Table 4: Parameters that are part of the NetKit learning package. These parameters can be set in the command line, when making a call to the framework for execution

Parameter	Options
Non-relational classifier	1) Class priors; 2) Uniform priors; 3) Weka classifiers to learn priors
Relational classifier	1) Weighted-vote relational neighbor; 2) Class-distributional relational neighbor; 3) Network-only multinomial Bayes classifier with Markov Random Field estimation
Collective Inference	1) Relaxation labeling; 2) Iterative classification; 3) Gibbs sampling
Runs	Parameter for k-fold classification
Sample	The split percentage for training set. (Remainder would be for testing)

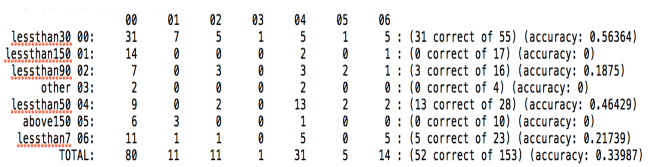
	00	01	02	03	04	05	06	07	08	09	
stimulants 00:	1	8	3	0	1	1	0	0	0	0	: (1 correct of 14) (accuracy: 0.07143)
alcohol 01:	7	45	18	0	6	1	1	5	0	0	: (45 correct of 83) (accuracy: 0.54217)
opioids 02:	4	11	5	0	2	0	0	1	0	0	: (5 correct of 23) (accuracy: 0.21739)
dissociative 03:	0	0	0	0	0	0	0	0	0	0	: (0 correct of 0) (accuracy: 0)
cocaine 04:	1	10	1	0	1	1	0	0	0	0	: (1 correct of 14) (accuracy: 0.07143)
prescription 05:	0	0	5	0	0	4	0	0	1	0	: (4 correct of 10) (accuracy: 0.4)
tranquilizers 06:	0	1	0	0	0	0	0	0	0	0	: (0 correct of 1) (accuracy: 0)
cannabis 07:	0	5	1	0	0	0	0	0	0	0	: (0 correct of 6) (accuracy: 0)
Other 08:	0	1	0	0	0	0	0	0	0	0	: (0 correct of 1) (accuracy: 0)
nicotine 09:	0	0	1	0	0	0	0	0	0	0	: (0 correct of 1) (accuracy: 0)
TOTAL:	13	81	34	0	10	7	1	6	1	0	: (56 correct of 153) (accuracy: 0.36601)

Figure 1: Confusion matrix for predicting the addiction substance of the participants.



	00	01	02	03	04	05	06	
highschool	00:	0	0	0	0	0	0	0 : (0 correct of 0) (accuracy: 0)
bachelors	01:	0	14	11	0	4	5	2 : (14 correct of 36) (accuracy: 0.38889)
other	02:	0	21	17	0	10	6	2 : (17 correct of 56) (accuracy: 0.30357)
phd	03:	0	2	1	0	1	0	0 : (0 correct of 4) (accuracy: 0)
masters	04:	0	7	8	2	2	3	1 : (2 correct of 23) (accuracy: 0.08696)
diploma	05:	0	7	6	0	0	0	0 : (0 correct of 13) (accuracy: 0)
associates	06:	0	8	8	1	0	2	2 : (2 correct of 21) (accuracy: 0.09524)
TOTAL:		0	59	51	3	17	16	7 : (35 correct of 153) (accuracy: 0.22876)

Figure 2: Confusion matrix for predicting the education level of the participants.



	00	01	02	03	04	05	06	
lessthan30	00:	31	7	5	1	5	1	5 : (31 correct of 55) (accuracy: 0.56364)
lessthan150	01:	14	0	0	0	2	0	1 : (0 correct of 17) (accuracy: 0)
lessthan90	02:	7	0	3	0	3	2	1 : (3 correct of 16) (accuracy: 0.1875)
other	03:	2	0	0	0	2	0	0 : (0 correct of 4) (accuracy: 0)
lessthan50	04:	9	0	2	0	13	2	2 : (13 correct of 28) (accuracy: 0.46429)
above150	05:	6	3	0	0	1	0	0 : (0 correct of 10) (accuracy: 0)
lessthan7	06:	11	1	1	0	5	0	5 : (5 correct of 23) (accuracy: 0.21739)
TOTAL:		80	11	11	1	31	5	14 : (52 correct of 153) (accuracy: 0.33987)

Figure 3: Confusion matrix for predicting the income level of the participants.

The accuracy from this set of experiments is poor. Firstly, accuracy isn't necessarily the best way to show the performance of classifier, but that information, augmented with the confusion matrix, gives us an idea of the "learning" (or lack thereof) achieved by the non-relational classifier. For all three predicted variables, the confusion matrix shows heavy influence from the prior. This could be because the experiment setup did not include learning priors using a non-relational classifier, as suggested by the authors of the approach. Additionally, from Table 5, we see that the choice of the collective inference model doesn't effect the performance as much as expected. (The expectation comes from the results of experiments described by the authors, in their work [4][5].) Secondly, the approach described by the authors expects the data to be at a larger scale. This suggests that we should use more than 3 attributes to describe each node. Furthermore, many more types of edges, and weighting mechanisms, need to be defined, and selected, to truly represent relationships between the nodes.

5 DISCUSSION AND FUTURE WORK

For the experiments above, we have not run the non-relational (local) classifier component in the workflow. It would be the logical next step to see if learning the priors from the local classifier model has some effect on the overall classification accuracy. Even so, the local classifier estimates the class probabilities, based on the attributes of the nodes. So far, we have only attached 2 additional attributes to the nodes: education level and income. This won't be sufficient to predict the class probabilities for addiction. We can further explore the performance by adding more attributes, each of which would serve as features for the local classifier.

Our future work also includes building a probabilistic model for the variables, using the Alchemy system, discussed previously. Using the approaches described by the system, we can learn the structure and the weights of the edges for the Markov Logic Network that describe the relationships between the various observations that we have collected from the participants of the study. Since the learning is solely based on the observations, the structure and the weights will help validate/test our hypotheses on how the observations relate to one another. Furthermore, the weights of edges between classes of observations, or factors, will exhibit the strength of dependencies/independencies between factors. For example, one of the hypotheses of the study suggests that there is a direct and positive relationship between engagement and relapse. If we build a network with these two variables as factors, we can learn if the data suggests a relationship between them and, if so, how they are related. The information from the structure of the Markov Logic Network could be used when constructing the node attributes and defining the edges (and their weights), for the network classification task described previously. From the structure, if we learn that a set of variables includes great predictors of another variable, this information could be encoded into the graph as either weighted edges or could be added to the node's attribute set (to be used as a feature for the non-relational classifier).

Further steps will depend on our results from the different experiments that we can execute using permutations and combinations of both these approaches.

6 CONCLUSION

Through an initial set of experiments, we have been able to explore within-network classification methods. The next steps would be to explore the NetKit-SRL toolkit further and test various other graph representations for prediction, and to compare/contrast the performance. After building a prior class probability distribution using homophily features and initial information about the participants, we can identify possibility of signals in the data, which could be represented as factors in Markov networks. After that, we can employ Alchemy to learn the structure of the model and the weights of the factors. Resulting discoveries/insights (or lack thereof) would be helpful in our efforts to design future studies and/or conduct analyses.

7 ACKNOWLEDGEMENTS

This work was in part funded through NIH Grant 1R01DA039456-01, The Social Interactome of Recovery: Social Media as Therapy Development. We would like to thank Dr. Edward Fox for his guidance in the project.

8 REFERENCES

[1] Addiction Recovery Research Center, The Social Interactome, https://quitandrecovery.org/the-social-interactome/, 2015/02/08

[2] Addiction Recovery Research Center, International Quit & Recovery Registry, https://quitandrecovery.org/, 2014/10/12

[3] Macskassy, Sofus A., and Foster Provost. A simple relational classifier. New York Univ NY Stern School of Business, 2003.

[4] Macskassy, S. A., & Provost, F. (2005). NetKit-SRL: A Toolkit for Network Learning and Inference. In North American Association for Computational Social and Organizational Science (NAACSOS) Conference.

[5] Macskassy, Sofus A. "Relational classifiers in a non-relational world: Using homophily to create relations." Machine Learning and Applications and Workshops (ICMLA) 2011, 10th International Conference on. Vol. 1. IEEE, 2011.

[6] Domingos, Pedro and Richardson, Matthew (2007). Markov Logic: A Unifying Framework for Statistical Relational Learning. In L. Getoor and B. Taskar (eds.), Introduction to Statistical Relational Learning (pp. 339-371), 2007. Cambridge, MA: MIT Press.

[7] Richardson, Matt and Domingos, Pedro (2006). Markov Logic Networks. Machine Learning, 62, 107-136, 2006.

[8] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

[9] Netrapalli, Praneeth, et al. "Greedy learning of Markov network structure." Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on. IEEE, 2010.

[10] Davis, Jesse, and Pedro Domingos. "Bottom-up learning of Markov network structure." Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.