JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Fast Prediction for Suspect Candidates from Criminal Networks

Dr. C. Krishna Priya[1]

[1]Assistant Professor, Department of Artificial Intelligence and Data Science

Central University of Andhra Pradesh, Anantapur.

G.L. Yamini[2]

[2]P.G. Student, Department of Artificial Intelligence & Data Science

Central University of Andhra Pradesh, Anantapur.

Abstract:

Machine learning approaches have been introduced to support criminal investigations in recent years. In criminal investigations, Criminal acts may be similar, and similar incidents may occur consecutively by the same offender or by the same criminal group. Among the various machine learning algorithms, network-based algorithms will be suitable to reflect such associations. In general,however, inference by network-based algorithms is slow when the size of data is large, so it is fatal in crime scenes that require urgency. And worse, the criminal network must be able to handle complex information entangled with case-to-case, person-to-person, and case-to-person connections. In this study, we propose a fast inference algorithm for a large-scale criminal network. The network we designed has a unique structure like a sandwich panel, where one side is a network of crime cases and the other side is a network of people such as victims, criminals, witnesses, etc., and the two networks are connected by relationships between the case and its corresponding people. The experimentalresults on benchmark data showed that the proposed algorithm has a fast inference time and competitiveperformance compared to the existing approaches. After performance validation, the proposed methodwas applied to the actual crime data provided by the Korean National Police to predict the suspect candidates for several cases.

Keywords: Depression Detection, Sentiment Analysis, Social Media Mining, Natural Language Processing (NLP), Sequential Deep Learning Model, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Text Classification, Emotion Detection, MentalHealth, Non-depressive Tweets.

1. Introduction

Machine learning approaches have been introduced to support criminal investigations in recent years. Crime pattern analysis [1, 2], fraud detection or traffic violation, sexual assault, and cybercrime analysis are typical examples [3, 4, 5]. In the meantime, one of the characteristics of case data is that criminal cases are related. Criminal behaviors may be similar, and similar incidents may occur consecutivelyby the same offender or by the same criminal group. Among the various machine learning algorithms, network-based algorithms will be suitable to reflect such associations.

Of the many network-based ML algorithms, the graph-based semi-supervised learning (GSSL) algorithm is one of the most popular because it is easy to use, can handle situations where data hasfew labels, and its inference is intuitive along the network structure [6, 7, 8, 9, 10, 11, 12]. However, when the size of data is large, the time complexity of network-based algorithms increases exponentially according to the size of the network, so it is inevitable that the inference speed by GSSLalso slows down. Therefore, applying GSSL to the crime scene requires the agility to achieve immediate results even in large-scale and complex networks. To satisfy the requirement, there have been algorithms we propose a fast GSSL algorithm. To define the boundaries of clusters that can be ofdifferent sizes and shapes, the proposed method employs mutual information of neighboring data points. So, it is called Neighbor Boundary Aware Semi-Supervised Learning (NBASL).

The remaining sections of the paper are organized as follows. Section 2 describes the proposed algorithm NBASL in detail. Section 3 demonstrates the comparative experiments on benchmarking datasets, and Section 4 presents the results of the practical application of criminal network analysis. Finally, we conclude in Section 5.

2. Criminal Network and Suspect Scoring Network

Construction

Criminal acts often exhibit similarities, and it is not uncommon for similar incidents to occur consecutively, either by the same offender or by a particular criminal group. In addition, in some cases, the patterns of crimes may bear resemblance, even if the offenders involved are different, resembling a copycat scenario seen in serial murder cases. Considering these factors, a network servesas an effective tool for illustrating connections between crimes and people. The criminal network we have developed captures associations between cases, individuals, and interactions between cases and individuals. This network follows a two-layered structure, with one layer representing the network of crime cases and the other layer representing the network of people. These two layers are interconnected through relationships that link specific cases with the corresponding individuals involved.

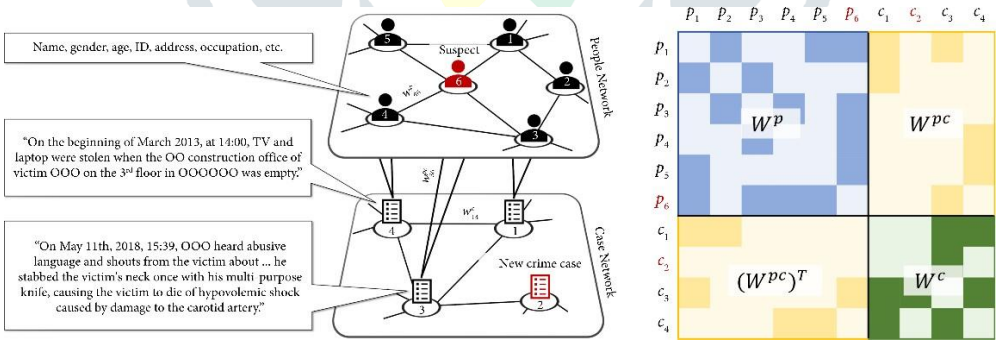

Fig 1. The schematic picture of the criminal network. The upper layer is the people network and the lower layer is the case network. The corresponding weight matrices are represented as and in the figure on the right, respectively. Two networks are connected by person-to-case edges denoted by in the weight matrix. A person node has demographic information of an individual while a case node hasinformation on crime reports (texts are translated from Korean to English. Names or titles in texts are de-identified. All the meaningless symbols were removed).

3. Clustering Analysis

Given a criminal network, to explore the organization and structure of criminals, it is necessary to analyze its clustering or partition. In the later part, we will find that the change of the networks' structure will reveal much more information than expected. Now, the most popular method includes hierarchy clustering [14, 15] and block modeling. However, always some approaches need us to determine the number of clusters beforehand, while others directly merge different clusters into one whole cluster with a strict structure. Inversely, criminal activities in the real world have become more and more complex and huge, so it is nearly impossible for a few leaders to govern the whole organization; thus, the emergence of criminal subnetworks is inevitable. Besides, the structure, partition, and even number of these subclusters are unknown before investigation. In this way, the effectiveness of hierarchy clustering and modeling may fail.

Here we design an optimal function to avoid the above problems as the following:

Here, is an experimental parameter, which varies from 0 to and determines the average size of subnetworks. The bigger the value is, the more sub-networks will generate, but the average size of sub-networks will be smaller. The value cannot be determined theoretically; it is the result of fitting according to actual cases. represents the number of subclusters. represents all the possible partitions of clusters. represents the subnetwork's sum of inner edges' weight after the and are determined.

This optimal function is based on the density of each subnetwork. If the partition is determined, then the density is defined as below: represents the number of nodes of the subnetwork.

From the definition of density, we can find that density increases when the ratio of the inner interaction of the subnetwork dominates the whole interaction of the members of the subnetworks, which reveals the information that these members are intensively connected. We multiply all the subnetworks' densities to get the optimal function:

It reflects the overall tight degree of all the subnetworks. The bigger the value of the optimal function is, the more intensive each subnetwork's inner structure is, and the maxima of the optimal function determines the best clustering method of the network.

4. Existing Approaches

Utilizing network analysis techniques to identify suspicious individuals based on their connections within criminal networks. This involves analyzing the topology, centrality measures, and community structures of criminal networks to pinpoint potential suspects. Applying SNA methods to analyze communication patterns, interactions, and relationships among individuals in criminal networks. Suspect candidates can be identified based on their proximity to known criminals or their involvementin suspicious activities. Employing machine learning algorithms to predict suspect candidates by learning patterns and characteristics from historical crime data. This approach may involve feature

engineering, classification, and anomaly detection techniques to identify individuals with high likelihood of criminal involvement.

Leveraging predictive policing strategies to forecast future criminal activities and identify suspect candidates preemptively. This involves analyzing crime hotspots, temporal patterns, and other relevantfactors to prioritize law enforcement efforts and allocate resources effectively. Conducting behavioral analysis of individuals within criminal networks to identify suspicious activities or deviations from normal behavior. This may involve monitoring digital footprints, financial transactions, or other behavioral indicators to flag potential suspects. Integrating heterogeneous data sources, including criminal records, surveillance footage, social media activities, and other sources of information, to generate comprehensive profiles of suspect candidates. Data fusion techniques enable the synthesis of disparate data types to enhance predictive accuracy and reliability. Utilizing graph-based algorithms tomodel criminal networks as graphs and identify suspect candidates based on graph properties, such as node centrality, clustering coefficients, and community detection. Graph-based methods offer a scalable and efficient approach to analyzing complex criminal networks. Employing entity linkage techniques to connect disparate data points and identify individuals across multiple datasets. This involves resolving identity ambiguities and linking related entities to uncover hidden connections and identify potential suspect candidates.

5. Proposed Method

We begin by representing criminal networks as graphs, where nodes represent individuals andedges represent connections or interactions between them. Various attributes such as criminalrecords, social connections, and demographic information are encoded as node features. Additionally, temporal information and contextual data, such as crime locations and timestamps, are incorporated into the graph representation. We propose a graph neural

network architecture designed to learn latent representations of individuals within the criminal network. The GNN model iteratively aggregates information from neighboringnodes to capture structural patterns and identify suspicious individuals. Specifically, we employ graph convolutional layers followed by pooling and aggregation layers to extract hierarchical representations of criminal network nodes.

Following the graph neural network encoding, we apply anomaly detection techniques to identify outlier nodes or anomalous patterns within the criminal network. Anomalous nodes are indicative of individuals with unusual behavior or connections that warrant further investigation. We employ unsupervised anomaly detection algorithms such as Isolation Forest, Local Outlier Factor (LOF), or One-Class SVM to flag suspect candidates based on deviations from normal network behavior. The identified suspect candidates are ranked and prioritized based on their anomaly scores or likelihood of criminal involvement. Individuals with the highest anomaly scores are considered top-priority suspects and recommended for further scrutiny by law enforcement authorities. Additional contextual information, such as past criminal records or intelligence reports, may be integrated to refine the ranking and prioritize suspects based on their risk level.

We evaluate the performance of our proposed method using real-world criminal network dataand benchmark

datasets. Metrics such as precision, recall, and F1-score are employed to assess the effectiveness and efficiency of suspect prediction. Comparative analysis with existing methods and sensitivity analysis on hyperparameters are conducted to validate the robustness and generalizability of our approach.

Variables		Descriptions
Person	Basic info	Nationality, gender (female or male), age, height, weight, address, and occupation
	History	History of criminal records, possession of a firearm, and sexual assault or not
	Class	Suspect/victim/witness
Crime case	Basic info	Time of incidence, location, type of crime, and arrested or not
	Case Summary	Crime summary report (text)
	Type	Battery, assault, drug, theft, traffic violation, disorderly conduct, and financial crime

Table 1. Crime data description.

Among the crime case data variables, the summary report is unstructured data in text format. Since most of the information about the incident is included in this report, preprocessing of the text is vital. To process text data, nouns, and verbs were extracted using theKorean morphology analyzer KoNLPy [35]. It is known as one of the best analyzers among open-source Korean morphology analyzers. Extracted words from the reports are converted into vectors through TF-IDF [33]. By using TF-IDF, the influence of unnecessarily repeated words can be reduced to some extent, and important information can be highlighted. Figure 1 shows an example of a part of the case report, and more details are described in Table 1.

6. Result

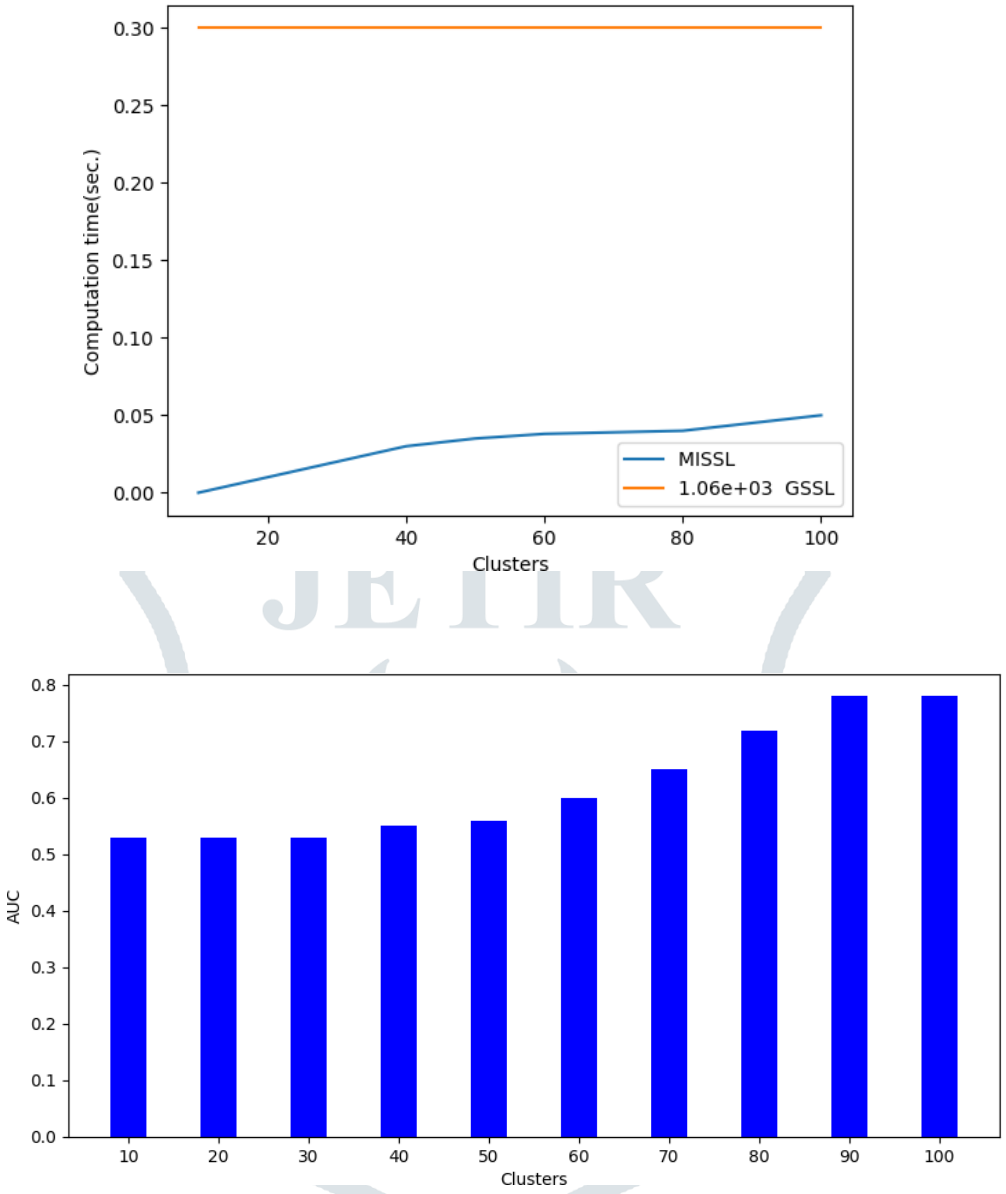

Figure 2: The computation time (line chart) and AUC performance (bar chart) of MISSL while varying the number of clusters. The performance of GSSL is shown by red dotted lines.

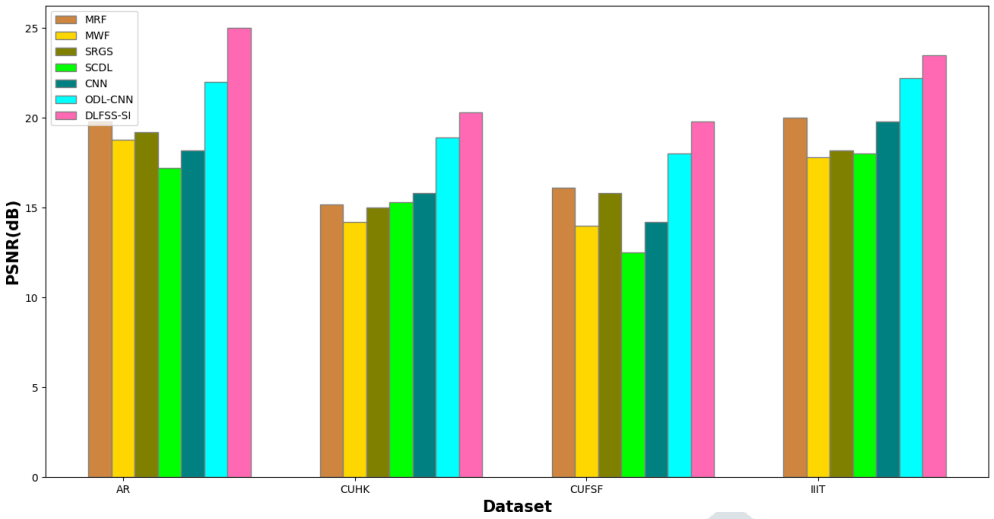

Figure 3: PSNR analysis of DLFSS-SI model

Fig. 3 investigates the PSNR analysis of the DLFSS-SI model with existing methods. On the applied AR dataset, the MWF, SCDL, and CNN models have showcased inferior results with the minimum PSNR of 18.74, 17.19, and 18.23 dB. Followed by, certainly better results are achieved by the MRF, SRGS, and ODL-CNN models with the PSNR values of 19.84, 19.13, and 21.98 dB respectively. However, the presented DLFSS-SI model has depicted effective results with a maximum PSNR of 24.86 dB. Next to that, on the applied CUHK dataset, the MWF,SRGS, and MRF models have demonstrated inferior outcomes with the minimum PSNR of 14.41, 14.79, and 15.07 dB. Next, certainly, optimal outcomes are attained by the SCDL, CNN, and ODL-CNN model with the PSNR values of 15.14, 15.64, and 18.64 dB correspondingly. However, the proposed DLFSS-SI model has depicted effective results with the highest PSNR of 20.65 dB. Along with that, on the applied CUFSF dataset, the SCDL, MWF,and CNN models have showcased inferior results with the minimum PSNR of 12.4, 14.34, and 14.36 dB. Afterward, certainly, better results are reached by the SRGS, MRF, and ODL- CNN model with the PSNR values of 15.34, 15.72, and 18.07 dB correspondingly. The projected DLFSS-SI model has showcased effective results with the maximum PSNR of 19.56 dB. Furthermore, on the applied IIIT dataset, the MWF, SCDL, and SRGS methods have showcased inferior results with the minimum PSNR of 17.2, 18.33, and 18.46 dB. Followed by, certainly better results are obtained by the MRF, CNN, and ODL-CNN models with PSNR values of 19.26, 19.62, and 21.74 dB respectively. However, the proposed DLFSS- SI model has outperformed efficient results with the highest PSNR of 23.53 dB.

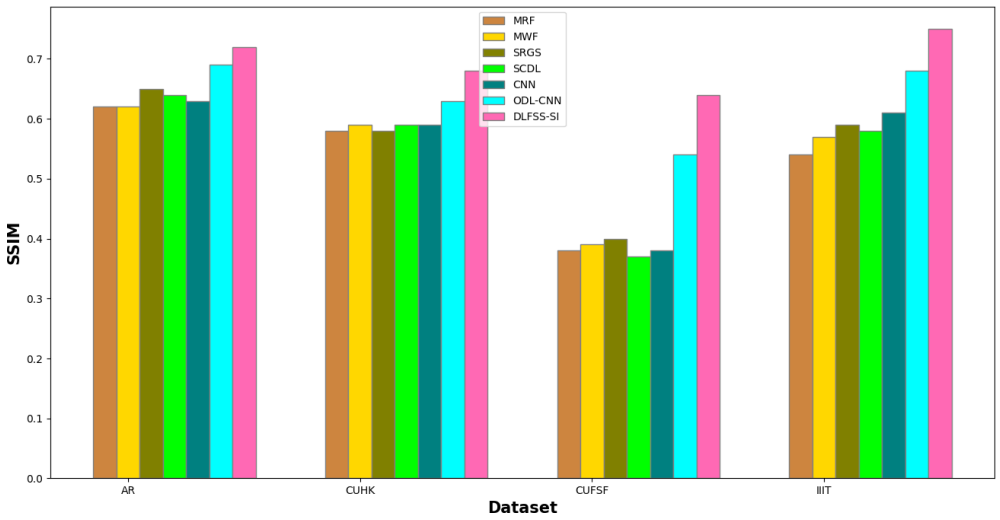

Figure 4: SSIM analysis of DLFSS-SI model

7. Conclusion

To make SSL practical in large-scale networks, we developed an algorithm based on anchor networks and mutual information. The proposed NBASL is scalable and has fast inference time. NBASL performed close to the reference method on the benchmark datasets and proved to be faster on large-scale networks. The suspect candidate scoring model for criminal network is developed with NBASL. The predicted results show the validity and efficacy of the model. The framework of suspect candidate scoring introduces a new way of analyzing the crime data and the results shed new light onnetwork-based machine learning approaches to social network analysis.

In this study, we proposed a framework for predicting suspect candidates based on the criminal network. The algorithm we employed is graph-based SSL, which may be inappropriate when networks are large and complicatedly structured. So, to put the GSSL to practical use in the criminal network, we developed an algorithm based on latent representation and mutual information. The proposed method, MISSL, shows almost similar performance to the graph-based SSL but has a much faster inference time. As an application, a criminal network is constructed from real-world crime data,and suspect candidate scoring is performed by MISSL. The predicted results show the validity and efficacy of MISSL. The framework of suspect candidate scoring introduces a novel way of analyzing

crime data and the results shed a new light on network-based machine learning approaches for social network analysis. Future efforts may identify a more efficient mechanism to optimize the hyperparameter of MISSL and the number of clusters (i.e., the number of latent dimensions) that affect performance. Empirically, a higher AUC was obtained from the increased number of clusters. From another perspective, clustering can be expensive for large-scale datasets. Therefore, further research on faster clustering methods or modeling the algorithms robust to clustering will enrich the results of our study.

Choose different perspectives to analyze problems, it will usually obtain some new feelings. To close the essence of the criminal network, we abandon the common SNA and frequency analysis or density analysis in DNA, and instead, we observe the overall structure's change of networks. Although it is hard to analyze quantitatively, we can visualize this process and observe it directly. In this way, we exemplify our superiority in touching the overall network's image, which is hard for criminals to disguise.

References:

1. Nath, Shyam Varan. "Crime pattern detection using data mining." 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops. IEEE, 2006.

2. Sardana, Divya, Shruti Marwaha, and Raj Bhatnagar. "Supervised and Unsupervised Machine Learning Methodologies for Crime Pattern Analysis." International Journal of Artificial Intelligence and Applications (IJAIA) 12.1 (2021).

3. Prabakaran, S., and Shilpa Mitra. "Survey of analysis of crime detection techniques using data mining and machine learning." Journal of Physics: Conference Series. Vol. 1000. No. 1. IOP Publishing, 2018.

4. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms toanalyze crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015): 1-12.

5. Vijayarani, S., E. Suganya, and C. Navya. "A comprehensive analysis of crime analysis using datamining techniques." International Journal of Computer Science Engineering (IJCSE) 9.1 (2020).

6. Zhu, Xiaojin, Zoubin Ghahramani, and John D. Lafferty. "Semi- supervised learning using gaussian fields and harmonic functions." Proceedings of the 20th International conference on Machine learning (ICML-03). 2003.

7. Zhou, Dengyong, and Bernhard Schölkopf. "A regularization framework for learning from graph data." ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields (SRL 2004). 2004.

8. Hoffmann, Franca, et al. "Consistency of semi-supervised learning algorithms on graphs: Probit and one-hot methods." Journal of Machine Learning Research 21 (2020): 1-55.

9. Zhou, Dengyong, et al. "Learning with local and global consistency." Advances in neural information processing systems 16 (2003).

10. Belkin, Mikhail, Partha Niyogi, and Vikas Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." Journal of machine learning research 7.11 (2006).

11. Xu, Zenglin, et al. "Discriminative semi-supervised feature selection via manifold regularization." IEEE Transactions on Neural networks 21.7 (2010): 1033-1047.

12. Zhou, Dengyong, Jiayuan Huang, and Bernhard Schölkopf. "Learning from labeled and unlabeleddata on a directed graph." Proceedings of the 22nd international conference on Machine learning. 2005.

13. Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information." Physical review E 69.6 (2004): 066138.

14. Bachman, Philip, R. Devon Hjelm, and William Buchwalter. "Learning representations by maximizing mutual information across views." Advances in neural information processing systems 32 (2019).

15. Zhu, Xiaojin, and John Lafferty. "Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning." Proceedings of the 22nd international conference on Machine learning. 2005.