Research Statement

Arunkumar Bagavathi

Research Objectives

Can millions of hospital records forecast hospital readmissions? Can social media give insights on online extremism? These questions motivate my research and I solve these problems using **data mining, social network analysis, cloud computing, and applied machine learning**. Information fusion creates a larger impact on forecasting and prediction models in recent times, challenging the data science community to analyze big data and extract patterns. For example, cost of surgery and diagnosis supporting patient’s hospital record and surrogate information of a topic from web and social media forums. My research focus is to design a deep learning framework to learn multiple activities in social media to enhance network based machine learning tasks like prediction, forecasting, and clustering. My work also leverage the use of cloud tools to frame a collection of scalable data agnostic models to glean actionable recommendations from massive data sources. For example, my multiplex model can identify user communities in the social media from their various activities and my action mining model can impart personalized recommendations for hospitals to handle patients to improve care and reduce readmission rates.

Action Mining using Cloud Frameworks

**Motivation:** The world wide web and social media creates an era of big data. The availability of such diverse data sources challenge the data science community in modeling scalable dynamical process. Examples include, improving customer satisfaction based on the feedback collected from in and around USA and decreasing patients readmission in US hospitals by analyzing medical and taxpayer records. And, these processes are crucial in supplying recommendations for analysts to undertake necessary actions. Recently, there is an increased traction towards collecting big data from supplementary data sources to refine prediction or forecasting tasks. However, these works proceed with a laborious process of analyzing numerous results to take necessary actions and it demands a knowledge mining model will recommend the required outcomes. Crafting these recommendation models for large datasets pose interesting challenges and questions that motivate my research: How to optimize these models to handle big datasets? How to measure the veracity of the resulting recommendations?

**Scalable action mining algorithms:** Action mining is one of the knowledge extraction methods, which supplies an actionable plan to the end user to support their decisions in gaining profit or increasing productivity. However, the literature showed efficacy of such action mining methods only on small to average sized databases. In the advent of big data and massive data available in web and social media, there exists powerful cloud based tools and frameworks, like Apache Spark, to clean, process, and extract knowledge from the data. These frameworks help researchers to store and analyze the given massive data using distributed/cloud computing and storage. We introduced multiple algorithms [1,4] to mine actionable patterns from the big data using such popular cloud computing frameworks.
Contributions:
- We framed multiple ideas to uncover actionable knowledge from the big data, where each idea can be used according to the given needs. For example, we crafted algorithms that extract both complete and approximated set of actionable patterns from the data.
- The main objective in our research is to find an optimal data distribution that minimize the loss of knowledge. We proposed a random data split model, which partition the given data by records and mine an approximated set of actionable recommendations. We also proposed a vertical data split model, for data with high-dimensional feature space, which finds optimal feature combinations to maximize the actionable pattern outcome. In our latest work [1], we incorporated the use of information granules to make approximations in sampling the feature space.
- Our methods shows efficiency in execution times compared to the state-of-the-art methods of extracting actionable patterns for larger datasets. In our latest work, we also showed that the information granules help in making better approximation in feature selection and our algorithm extracts a complete set of actionable patterns.

Low cost action mining methods: Actionable knowledge extracted from any action mining methods are prone to incur some cost to users, since the recommendations include improvements or changes to undertake. A cost can be of any forms like money, time, energy, and power. Actionable patterns are interesting to users, only if the recommendations are cost efficient or affordable to them. To extract such effective patterns, we introduced a notion of Action Graph - a distributed network representation of actionable knowledge and its properties, to search and mine low cost actionable patterns. We used a cloud based graph framework called Spark GraphX to accomplish distributed graph operations.

Contributions:
- We created distributed versions of graph search algorithms like Dijkstra’s Shortest path algorithm, Breadth First Search and Depth First Search for action graphs to search cost effective actionable patterns [3].
- We also introduced matrix factorization based post pruning approach to study the overlaps of actions and prune the large quantity of knowledge from actionable patterns.

Multiplex Network Embeddings

Figure 1. An example multiplex network with three layers of networks(G₁, G₂, and G₃). Each layer can be either different social media forums like Twitter, Facebook, and Reddit, or different activities like replies, mentions, and retweets in a single social media forum (Example. Twitter)

Motivation: The activity of millions of humans on social networks provide great potential to address persisting scientific problems and it inspires the design of future social computing
applications. As given in Figure 1, these activities emerge as network structures with many forms and modalities in the names of multiplex or composite or complex networks. Such multiplex networks advance the research in network modelling and diffusion processes like contagion dynamics and influence modelling. These networks provide the machine learning models an ability to gain insights from various connectivity patterns of nodes and subgraphs. Recent advancements in word and document embedding methods had created a significant research thrust on graph/network embedding methods. The low-dimensional feature vector space from such models showcase their benefits in plethora of applications like detecting anomalies and forecasting relationships in social and other networks. The availability of complex networks and network embedding models motivate my research to utilize multiplex networks characteristics for enhancing the network models.

**Simple random walk based embedding:** Similar to word embedding algorithms, network embedding models demand node/edge context. One simple approach is few epochs of random walks in the given network to collect node context. Random walks in multiplex networks has to consider connectivity patterns all network layers. To accomplish such a task, each edge of a node is assigned a transition probability $P_i$ to move between nodes in the same network layer, and each node is assigned a switch probability $P_s$ to transfer between network layers. With aggregated node contexts from random walks, a machine learning model like skip-gram model learns such context using Stochastic Gradient Descent and produce embedding of each node.

**Contributions:**
- We proposed a multiplex random walk model by assigning a simple transition and switch probabilities, that follows uniform distribution to collect node context from the multiplex network
- We prove the quality of the node embeddings of multiplex networks from our methods using a network construction task. We trained a skip-gram model to generate embeddings of nodes from all but one network using the node sequences given by our multiplex random walk model. We then used the learned embeddings to reconstruct the left out network.
- Our model showed performance improvement of at least 11% accuracy, and faster [2] than the state-of-the-art random walk based network embedding methods for all the four real world multiplex networks.

**Scalable multiplex network embeddings:** All real world networks have dynamic nature, creating or destroying new nodes and edges over time. The drawback of our previous method is the execution of few iterations of random walks on all nodes in the network. This consumes longer execution time compared to model training and also the parameter tuning for random walks is complex. In this ongoing work, we are designing a batch process in which, we acquire embeddings of most influential nodes, and other nodes learn their representation from their neighborhood.

**Contributions:**
- We are exploring multiple methods to get influential nodes in multiplex networks. Few of the methods are PageRank, betweenness centrality, degree centrality, and HITS. Similarly, we are experimenting multiple neighborhood aggregation methods like max pooling, mean aggregator, and LSTM aggregator to learn node embeddings from influential nodes. We are designing an autoencoder model to learn node context and obtain node embeddings.
- Since the representation learning is using neighborhood aggregation, we claim that that new nodes can also learn representations from their neighbors. We are also working on solving the problem of multiplex network community detection using node embeddings.
Vision and Future Directions:

All my research has more open questions to explore in the near future. A Few of them include:

- The existing complex network embedding are only based on their structural patterns. However, the representation learning framework can produce better results if it incorporates surrogate information of the network. I will conduct a wide range research on adding opinion of users, dynamic communities, and other textual features of actors for learning node and subgraph representations.
- Apart from representation learning, I will develop techniques to analyze dynamic evolution of network groups in the multiplex setup. Using these analyses, I will conduct research on network oriented data mining like study of contagion spread, community detection from heterogeneous networks, and online extremism.
- I will extend the scope of action mining to leverage multiplex network representation learning tasks. Action patterns can be an alternative to optimize parameters for machine learning algorithms and it forms a generic framework that learn the network properties to recommend parameters to use. Also, I will extend the research to make actionable recommendations, that follows game mechanisms on how to attain specific goals in the future based on the present scenario. This has applications in viral marketing and immunization spread.

Scalable computation over massive data is always the ultimate goal in my research and it has applicational impact across social science, cyber security, business, and medical domains. Although, all my research is based on data mining, I will constantly collaborate with cross domain experts, like my work in online extremism [5] and health informatics [6], to study the potency of my methods in their research areas. I plan to continue my current collaborations with professors at UNCC (Dr. Bojan Cukic, Dr. Zbigiew Ras, and Dr. Siddharth Krishnan) to explore new dimensions in data science, with my main focus based on network analysis and scalable data mining. I also plan to write grants with my colleagues and students to funding agencies like NSF, DARPA, and IARPA to contribute innovations to the data science community and broaden my knowledge.

References:

5. Phillips, M.; Bagavathi, A.; E. Reid, S.; Valasik, M.; and Krishnan, S., "The daily use of Gab is climbing: Which talker might become as violent as the Pittsburgh Synagogue Gunman?", in Washington Post Monkey Cage, November 29, 2018