Nan Du

http://strideradu.github.io/•dunan00001@gmail.com•517-303-3341

Research Interest Bioinformatics • Machine Learning • Artificial Intelligence

Skills

- Intern experience on **recommender systems**, **C++** backend, **Hive/Presto**, and **Python** script
- Familiar with machine learning tools (**scikit-learn, lightGBM**, **PyTorch, TensorFlow**, **Numpy**) and achieved top rankings in Kaggle competitions with **natural language processing**, and **computer vision** topic.
- Good class performance on **algorithm design**, **machine learning**, **data mining**, and **software engineering**
- Design algorithms to analyze big biological data (**Hidden Markov Model, sequence comparison**)

Education

Ph. D., Computer Science — **May 2019 (Expected)**
Michigan State University, Michigan, USA
- GPA: 4.0/4.0

M. Sc., Condensed Matter Physics — **May 2015**
Michigan State University, Michigan, USA
- GPA: 3.6/4.0

B. Sc., Physics — **June 2011**
Fudan University, Shanghai, China

Experience

Software Engineer Ph.D. Intern, Machine Learning — **June 2018 - present**
Facebook, California, USA
- Improving Facebook Marketplace ranking by introducing new NLP features
- Implementing data pipeline for offline test in Hive and Presto
- Implementing feature fetcher in PHP and C++
- Conducting offline experiment and online AB test

Research Assistant, Bioinformatics Lab — **Jan 2015 – June 2018**
Michigan State University, Michigan, USA
- Working on developing the new algorithm for the challenge of the long erroneous sequence.
- Designed an algorithm use Hidden Markov Model combined with one popular error correction method based on the directed acyclic graph to infer the possible errors in the reads and correct it. After correction, the alignment length of protein profile can be improved 40% on low coverage dataset.
- Designed an algorithm that uses statistic method to determine the threshold to filter the raw data and chaining the data points to make the prediction of the similarity of two long error reads. We expect this method can boost 5% of the sensitivity without losing the accuracy.

Projects

Silver Medal, Avito Demand Prediction Challenge — **April 2018 – June 2018**
Kaggle Competition
- Final ranking **68 of 1917** with another teammate
- Based on user purchase history and the product information (image and post) to predict the CTR
- Extract features from customer activities, image embedding and text features and embeddings
- Combine gradient boosting method using lightGBM and neural network model implemented by Keras

Silver Medal, Toxic Comment Classification Challenge — **Jan 2018 – March 2018**
Kaggle Competition
- Final ranking **69 of 4551** with two team mates
- Classified online comment to 7 toxic comment class
- Use various neural network model based on LSTM, GRU, and Text CNN

Silver Medal, Google Cloud & YouTube-8M Video Understanding Challenge — **March 2017 – June 2017**
Kaggle Competition
- Final ranking **35 of 650** with team RandomForest
- Predict tag of video (sequence of processed feature matrix) from YouTube (video classification)
- Data processing on large YouTube video dataset (1.7TB)
- Ensemble based on variations of LSTM and MoE models

Publication

- Du N., Chen J., and Sun Y., Improving the sensitivity of detecting long read overlaps using grouped short k-mer matches, submitted to *APBC 2019*
- Pak, D., Du, N., Kim, Y., Sun, Y., & Burton, Z. F. (2018). Rooted tRNAomes and evolution of the genetic code. *Transcription*, *9*(3), 137-151.
- Du, Nan, and Sun, Yanni. "Improve homology search sensitivity of PacBio data by correcting frameshifts." *Bioinformatics* 32.17 (2016): i529-i537.
- Improve homology search sensitivity of Pacbio data by correcting frameshifts (Proceeding Talk), *15th European Conference on Computational Biology (ECCB 2016)*, The Hague, Netherlands