Research Plan

The field I will focus on in master and doctor course (if possible) is how to develop more meaningful models using the combination of machine learning algorithms and apply these models to the field of NLP. By now I have been learning basic knowledge of statistical machine learning and deep learning though I’m working in other fields, however I’m still a newbie in machine learning so I need to take much more time in my study and research.

Basically, machine learning itself often uses four strategies to learn models. They are supervised learning (e.g.: KNN), unsupervised learning (e.g.: cluster (K Means), association), semi-supervised learning (mixture algorithms) and reinforcement learning.

The first three machine learning strategies above mainly help us solve two main tasks: classification and regression. In which classification is more used in real world than regression, many different items can be recognized by the classifier like Bayes classifier, decision tree, linear classifier, linear SVM, logistic regression classifier (sigmoid function, softmax function), perceptron(neural network), and others. However, only using algorithms above is definitely not enough for us to create a perfect model because some problems like overfitting (often happens in linear classifier, sometimes tiny change in training dataset may occur different models) and over computational load may occur when we are training model. We should take actions to reduce these issues, for example, in linear classifier we often add L1 or L2 regularization item. We often reduce dimensions of vector using algorithms like LDA and PCA to reduce computational load. To make calculation of lost function optimization faster, we use algorithm like gradient descent in non-convex optimization and use algorithm like Newton’s method in convex optimization.

Currently, the most widely-used technology is Deep Learning which is based on NN. Along with improvement of hardware, deep learning is more efficient than other statistical algorithms in most time. In my opinion, deep learning is an online algorithm which can improve itself along with the streaming dataset, sometimes we may combine DL with Reinforcement learning in graph shortest path finding. For example, one research I’m very interested in is to use deep learning to help us improve A\* heuristic function in shortest path finding algorithms, for example, we can use some graphs and their known heuristic algorithms as training dataset to learn the best way of path finding, we can even consider to use reinforcement learning to improve time complexity of A\* though it’s not very efficient because both of RL and A\* are based on greedy strategy.(Reinforcement learning uses a rewarding strategy which is based on Markov decision process. In Markov chain, following state is only depend on the current state and weight of closer state is more than farther state. Upon this point, reinforcement learning is sometimes similar with greedy strategy I think.)

As we can see, most of algorithms listed above are mainly based on supervised learning, why? In my opinion, machine learning is an application driven technology, a learning strategy we choose is depend on the filed we would like to apply our models to. For example, in application field of image recognition, audio recognition and text recognition, it’s a trend that using supervised learning is better. So if we can come out with more ideas of the application of machine learning model, unsupervised learning or reinforcement learning may be more useful.

In the field of NLP, there’re also some specific technologies. For example, how can we transform the word into vectors so that we can these text data can be trained by machine learning algorithms, currently, word embedding(word2vec) is a most used technology. In addition, there’re many sub-fields in NLP, for example, word segmentation (e.g.: Chinese and Japanese), Spelling Correction, Sentence parsing (e.g.: Minimum edit distance), Part-of-speech tagging, text categorization, information extraction, Sentiment analysis, machine translation, Q&A and others. Among these fields, I think machine translation and Q&A is most difficult to realize so I would like to focus on these two fields in master course.

In addition, I think currently the application of NLP technology is very limited. We may develop some more NLP applications in many more fields in the future. For example, how to create a DSL(Domain Specific Language) which can recognize commands a part of which are written as natural language. Currently, there’re some research on how to transform web page image to JavaScript code automatically using Deep Learning. It’s very similar with what I want to research and I think it’s very meaningful because it can help programmers relieve much stress.