資工三 B00902064 宋昊恩

Paper Studying Report

1. Introduction:

i. **Title:**
Machine Learning Paradigms for Speech Recognition: An Overview

ii. **Author:**
Li Deng, Fellow, IEEE, and Xiao Li, Member, IEEE

iii. **Publication:**
IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 5, MAY 2013 (p.1060 ~ p.1089)

2. Motivation:

i. I am studying in Machine Learning related field for about one year. I am interested in utilizing these machine-learning-oriented algorithms in our daily life. Digital Speech Recognition is a well-known and long-studied field, which is strongly related to machine learning. That is also my main incentive to enroll in this course.

ii. I had no idea when I started deciding my final project topic in lecture slides. Finally, I found this paper was on the last page which was exactly I looked for.

iii. I will briefly introduce these algorithms or mechanisms categorized by subject at the former part of this report; some personal understanding and thought to these ideas will be at the latter part.

iv. References for further background information are added as external links.

3. Machine learning and digital speech recognition:

i. Machine learning field used to have some theoretical assumptions before deriving an algorithm, and few of them put experiments on real-life data.

ii. Digital speech recognition filed, in contrast, usually derived new idea or algorithm driven by real-life problems.

iii. Machine learning filed can be categorized as generative and discriminative learning in terms of loss function; supervised, semi-supervised. unsupervised, and active learning in terms of training data; single-task learning, adaptive learning, and multi-task learning in terms of the relationship between source and target data distribution. Another field which is gaining in popularity area is called deep learning.

iv. Digital speech recognition field can be categorized as feature extraction, frame-based acoustic modeling, pronunciation modeling and language modeling and hypothesis search. This paper focuses on acoustic modeling.

v. Transferring acoustic data sequence to word sequence can be viewed as a structured classification problem in machine learning.

vi. Acoustic data usually have variable length of features, which are different from what Machine Learning Problem used to be. Even two output word sequences are identical; they may have distinct length of features.

vii. Speech is used to be viewed as two dimensions, spatial and temporal one.

4. Generative learning:

i. This kind of loss function focuses on joint probability distribution. If we use x as input train data, y as label, λ as model parameter, joint probability distribution is $p(x, y; \lambda)$.
Its discriminant function is $d_y(x; \lambda) = \ln\big(p(x, y; \lambda)\big) = \ln(p(x|y; \lambda)p(y; \lambda))$, that is to say, it models how data was generated in order to do categorization.

ii. There are lots of models using this kind of loss function, for example, naïve Bayes classifier[1] with no-dependency assumption, Gaussian mixture model (GMM)[2], dynamic Bayesian network (DBN)[3] with directed acyclic graph, Markov random field (MRF)[4] with undirected edges and trajectory, segment models. Furthermore, there is one specific Markov random field model called restricted Boltzmann machine (RBM)[5], which becomes more popular nowadays.

iii. To link with acoustic models, for example, teacher has taught us Baum-Welch algorithm[6] which is used in GMM-HMM model.

5. Discriminative learning:

i. This kind of loss function focuses on conditional probability distribution. If we use x as input train data, y as label, λ as model parameter, conditional probability distribution is $p(y|x; \lambda)$.
Its discriminant function is $d_y(x; \lambda) = \ln(p(y|x; \lambda)$, that is to say, it does the categorization directly.

[1] http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2] http://en.wikipedia.org/wiki/Mixture_model

[3] http://en.wikipedia.org/wiki/Dynamic_Bayesian_network

[4] http://en.wikipedia.org/wiki/Markov_random_field

[5] http://en.wikipedia.org/wiki/Restricted_Boltzmann_machine

[6] http://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm

- **ii.** There are two major types of models use this loss function, one is called Bayesian Minimum Risk (BMR)[7] classifiers, which focuses on probability conditional distribution, another one focuses more on decision margin, for example, support vector machine (SVM)[8] is used to maximize the margin between labels.
- **iii.** To link with acoustic models, multi-layer perceptron (MLPs)[9]'s conditional probability output can be directly used in HMM. Similar to MLP-HMM, SVM-HMM can give even better performance.

6. Supervised, unsupervised, semi-supervised and active learning:

- **i.** Supervised learning means that all the input data have a label.
- **ii.** Unsupervised learning means that all the input data do not have a label.
- **iii.** Semi-supervised learning means that some input data have a label, but some do not.
- **iv.** Active learning means that it wants to label the minimum number of unlabeled data to gain most significant improvement.

7. Transfer learning:

- **i.** A growing learning mechanism used in machine learning field, but not so popular in digital speech recognition field. Its concept is to transfer the data from one well-known category to one which has less data. There are two types of transfer learning, homogeneous or heterogeneous one.
- **ii.** Adaptive learning is one form of transfer learning, which transfers knowledge in a sequential manner.
- **iii.** Multi-task learning means to learn multi tasks simultaneously.
- **iv.** Audio-visual ASR[10], multi-lingual and cross lingual ASR, pronunciation learning and detected-based ASR may use this kind of learning algorithm.

8. Deep learning:

- **i.** It exploits unsupervised learning for pattern classification. It gained so much popularity nowadays mainly because the improved power of hardware and the

[7] Vaibhava Goel, William J Byrne. Minimum Bayes-risk automatic speech recognition, Computer Speech & Language, Volume 14, Issue 2, April 2000, Pages 115 – 135

[8] http://en.wikipedia.org/wiki/Support_vector_machine

[9] http://en.wikipedia.org/wiki/Multilayer_perceptron

[10] Automatic speech recognition, http://en.wikipedia.org/wiki/Speech_recognition

usage of GPU heterogeneous parallelism. Both of them make high complexity calculation becomes possible.

ii. Deep belief network (DBNs)[11] is a kind of deep generative model, which uses greedy, layer-to-layer learning algorithms to optimize parameter weight.

9. Some ideas:

i. In my opinion, active learning is one of the most important machine learning techniques used in DSP field. As technology advances, requirement for auto-acoustic-response system or auto-acoustic-recognition system increases rapidly. For example, in future, a system may auto detect and recognize one's sound after acquiring few decisive question to gain pronunciation.

ii. Another topic that I am interesting is on-line learning. Basically all lectures we learned in this course taught us how to analyze information, extract features, and build model off-line. However, in real life, it is not the case. We need to update our model immediately when we gain any further information.

iii. One related topic to **ii.** is what I am now working on, an on-line missing imputation algorithm[12]. In real life, feature missing event commonly happens, it can be caused by device failure, software error, or even hardware congenital limitation, and thus it is difficult for us to do on-line model updating. I want to derive an on-line algorithm to impute these missing features in time.

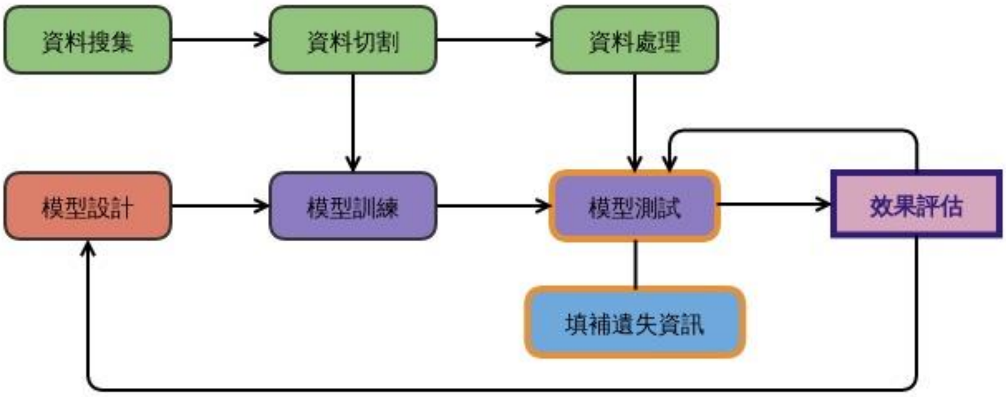

Figure 1: *missing feature imputing pseudo procedure*

Possible solutions include RBM model or some simplified graphical model. After deriving an efficient algorithm to solve this problem, we will be more confident to build up a real-time updating acoustic model to solve on-line speaking recognition problem.

[11] http://en.wikipedia.org/wiki/Deep_belief_network

[12] 103 學年度國科會大專生研究計畫 - 給定模型下即時分類遺失部份資訊的資料 (宋昊恩)