Mutual Information Independence Model using Kernel Density Estimation for Segmenting and Labeling Sequential Data

ZHOU GuoDong, YANG LingPeng, SU Jian, JI DongHong

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{zhougd, lpyang, sujian, dhji}@i2r.a-star.edu.sg

Abstract. This paper proposes a Mutual Information Independence Model (MIIM) to segment and label sequential data. MIIM overcomes the strong context independent assumption in traditional generative HMMs by assuming a novel pairwise mutual information independence. As a result, MIIM separately models the long state dependence in its state transition model in a generative way and the observation dependence in its output model in a discriminative way. In addition, a variable-length pairwise mutual information-based modeling approach and a kNN algorithm using kernel density estimation are proposed to capture the long state dependence and the observation dependence respectively. The evaluation on shallow parsing shows that MIIM can effectively capture the long context dependence to segment and label sequential data. It is interesting to note that using kernel density estimation leads to increased performance over using a classifier-based approach.

1 Introduction

A Hidden Markov Model (HMM) is a model where a sequence of observations is generated in addition to the Markov state sequence. It is a latent variable model in the sense that only the observation sequence is known while the state sequence remains "hidden". In recent years, HMMs have enjoyed great success in many tagging applications, most notably part-of-speech (POS) tagging [1,2,3] and named entity recognition [4,5]. Moreover, there have been also efforts to extend the use of HMMs to word sense disambiguation [6] and shallow/full parsing [7,8,9].

Given an observation sequence $O_1^n = o_1 o_2 \cdots o_n$, the goal of a HMM is to find a stochastic optimal state sequence $S_1^n = s_1 s_2 \cdots s_n$ that maximizes $P(S_1^n, O_1^n)$:

$$S^* = \arg\max_{S_1^n} \log P(S_1^n, O_1^n) = \arg\max_{S_1^n} \{\log P(S_1^n) + \log P(O_1^n \mid S_1^n) \quad (1)$$

Traditionally, HMM segments and labels sequential data in a generative way by making a context independent assumption that successive observations are independent given the corresponding individual state [10]:

$$P(O_1^n \mid S_1^n) = \prod_{i=1}^{n} P(o_i \mid s_i) \quad (2)$$