Training the max-margin sequence model with relaxed slack variables

Lingfeng Niu, Yong Shi
Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences
niulf@lsec.cc.ac.cn, yshi@gucas.ac.cn
Jianmin Wu[*]
Yahoo! Bejing Research & Development Center
jianmin@yahoo-inc.com

Extended abstract:
Sequence models are widely used in many applications such as named entity recognition, online optical character recognition, part-of-speech tagging, etc.. Although the sequence data can be labeled by predicting the label of each node in the sequence independently, this method does not use the label dependency hidden in the sequence, which is valuable for improving the accuracy of sequence classification. Several methods are already invented to utilize the label dependency relationship, such as Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), Conditional Random Fields (CRFs), Hidden Markov Support Vector Machines (HM-SVMs), Maximum Margin Markov Networks (M³Ns), Kernel Conditional Graphical Models (CGMs), etc.. Because the size of the label space for sequence data is exponential to the length of the sequence, the training processes of CRFs, M³Ns as well as HM-SVMs are complex and time consuming. Exploring the exponential label space for each of the training samples is the main challenge for the sequence and structured learning problem.

   Inspired by the work of CGMs, we propose a solution to simplify the training of sequence models in this work. By relaxing of the slack variables, in this paper, we propose a new approach to train the max-margin based sequence model by relaxing the slack variables. With the canonical feature mapping definition, we solve the relaxed problem by training a multiclass classification Support Vector Machine. Compared with the state-of-the-art solutions for sequence data, our new method has the following advantages: first, the sequence training problem is transferred into the widely studied multiclass classification problem, which has solid theoretical background on generalization analysis and mature off-the-shelf software packages for training; second, the new algorithm has cheaper computational cost than the algorithms of training sequence models directly; last but not least, nonlinear

---

[*] corresponding author. Tel: 010-82159306

feature space can be easily explored by employing kernels. Experimental results on the task of named entity recognition, information extraction and handwritten letter recognition with the public data sets illustrate the efficiency and effectiveness of our method.