Aspects of Semi-Supervised and Active Learning in Conditional Random Fields

Nataliya Sokolovska

LRI, CNRS UMR 8623 & INRIA Saclay,
University Paris Sud, Orsay, France

ECML 2011, Paper ID 243

Outline

Motivation

- Problem of sequence labeling (text, biological data, audio data, etc.)
  - Natural Language Processing
  - Data with sequential underlying structure

$$\Downarrow$$

Model of Conditional Random Fields

- Cheap unlabeled data vs. expensive labeled data
  - Exploit unlabeled data $\Rightarrow$ Semi-Supervised Learning
  - Choose instances of high training quality $\Rightarrow$ Active Learning

Problem of Sequence Labeling: formalizations

Given N independent labeled sequences $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$, where

- $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_{T_i}^{(i)})$ denotes an input sequence
- $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_{T_i}^{(i)})$ is an output sequence
- T_i is a length of sequences $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$

Problem of Sequence Labeling: formalizations

Given N independent labeled sequences $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$, where

- $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_{T_i}^{(i)})$ denotes an input sequence
- $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_{T_i}^{(i)})$ is an output sequence
- T_i is a length of sequences $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$

The aim is to minimize the negated conditional maximum likelihood

$$\ell(\mathcal{D}; \theta) = -\sum_{i=1}^{N} \log p_\theta(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}) + \rho_2 \|\theta\|^2$$

with respect to the parameter θ.

Model of Conditional Random Fields

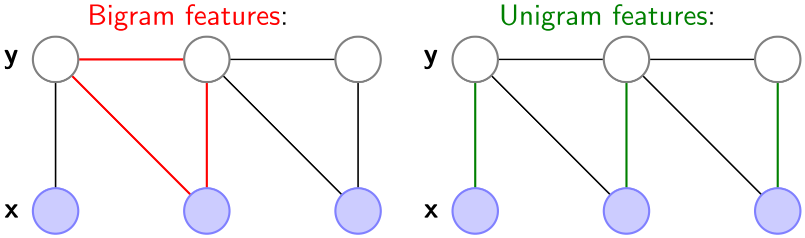
Conditional Random Fields (*Lafferty, McCallum, Pereira, 2001*) are based on the discriminative probabilistic model

$$p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) = \frac{1}{Z_\theta(\mathbf{x}^{(i)})} \exp\left\{ \sum_{t=1}^{T_i} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}^{(i)}, y_t^{(i)}, x_t^{(i)}) \right\},$$

- $\{f_k\}_{1 \leq k \leq K}$ is an arbitrary set of feature functions
- $\{\theta_k\}_{1 \leq k \leq K}$ are real-valued parameters, associated with the feature functions
- the normalization factor

$$Z_\theta(\mathbf{x}^{(i)}) = \sum_{(y',y) \in \mathcal{Y}^2} \exp\left\{ \sum_{t=1}^{T_i} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}^{(i)}, y_t^{(i)}, x_t^{(i)}) \right\}.$$

Feature Functions



$$\sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) = \sum_{X \in \mathcal{X}} \left(\sum_{y \in Y, x \in X} \mu_{y,x} \mathbb{1}\{y_t = y, x_t = x\} + \sum_{(y',y) \in Y^2, x \in X} \lambda_{y',y,x} \mathbb{1}\{y_{t-1} = y', y_t = y, x_t = x\} \right).$$

We get $|X| \cdot |Y| + |X| \cdot |Y|^2$ to estimate.

Application: Phonetization task (NetTalk Corpus)

Phonetization task: 20 000 English words and their transcriptions

$$X = \{\text{letters}\},\ |X| = 26,$$
$$Y = \{\text{phonemes}\},\ |Y| = 53.$$

Ex. apple - [' æ p l]

Training corpus – 16 000 sequences

Application: Named-Entity Recognition Task (CoNLL 2003)

Predict a sequence of labels given 3 aligned sequences of observations.

Word	Part of Speech	Syntactic Tag	Label
Slovenia	NNP	I-NP	I-LOC
and	CC	I-NP	O
Poland	NNP	I-NP	I-LOC
target	NN	I-NP	O
EU	NNP	I-INTJ	I-ORG
,	,	O	O
NATO	NNP	I-NP	I-ORG
membership	NN	I-NP	O
.	.	O	O

Training corpus – 15 000 sequences

Semi-Supervised Probabilistic Criterion

$\{X_i, Y_i\}_{i=1}^n$ are observations and their labels

Let $g(y|x;\theta)$ be the conditional probability function, parameterized by θ. Then the standard conditional maximum likelihood estimator is defined by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i|X_i;\theta),$$

where $\ell(y|x;\theta) = -\log g(y|x;\theta)$ denotes the negated conditional log-likelihood function.

The asymptotically optimal semi-supervised estimator $\hat{\theta}_n^s$ proposed by *Sokolovska et al., 2008* is defined by

$$\hat{\theta}_n^s = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}} \ell(Y_i|X_i;\theta),$$

where $q(x)$ is the marginal probability of observations.

Semi-Supervised Probabilistic Criterion Applied to CRF

The semi-supervised criterion applied to the conditional random fields criterion, referred later to as weighted CRF, takes the form:

$$C(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} -q(\mathbf{x}) \frac{1}{N_{\mathbf{x}}} \log p_\theta(\mathbf{y}|\mathbf{x}),$$

where $N_{\mathbf{x}}$ is the number of times a sequence $\mathbf{x}$ has been observed in the training corpus, and $p_\theta(\mathbf{y}|\mathbf{x})$ is defined

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) \right\}.$$

.

Semi-Supervised Criterion: Simulated Data

Artificial data simulated by a hidden Markov Model (first order);
A – the state transition probabilities, B – the observation probabilities matrix.

$$q(\mathbf{x}) = \sum_Y p(\mathbf{x}, \mathbf{y}) = \sum_Y p(y_1) b_{y_1}(x_1) a_{x_1,x_2} b_{y_2}(x_2) \ldots a_{x_{T-1},x_T} b_{y_T}(x_T).$$

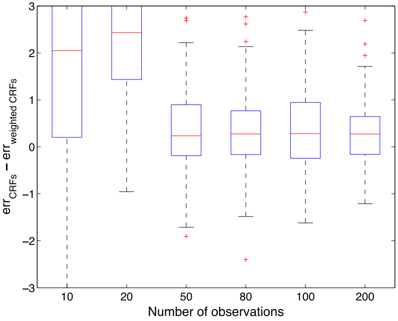


Figure: Simulated data. Difference of error rates of standard and weighted CRF by marginal probability. Weighted CRF performs better if n is small.

Approximation of Marginal Probability of Observations

We follow the idea of *n*-grams linguistic models:

$$q(\mathbf{x}) = q(x_1, \ldots, x_T) = \prod_t p(x_t | x_{t-1}, x_{t-2}, x_{t-3}),$$

where

$$p(x_t | x_{t-1}, x_{t-2}, x_{t-3}) \approx C(x_t, x_{t-1}, x_{t-2}, x_{t-3}) / C(x_{t-1}, x_{t-2}, x_{t-3}),$$

$C(\cdot)$ means counts.

For the realistic data sets:

- NetTalk: *n*-grams model, $n = 3$;
- CoNLL 2003: *n*-grams model, $n = 2$;
  $p(\mathbf{x}) = p(\mathbf{x}_{\text{word}}) p(\mathbf{x}_{\text{POS tag}}) p(\mathbf{x}_{\text{synt. tag}})$.

Motivation for Pool-Based Active Learning

Quota Sampling instead of Stratified Sampling

Intuition: rare events are not less important than frequent ones

Use quota sampling to select training instances efficiently:

- Candidates for training are sorted according to their marginal probabilities
- Get *n* frequency groups of training points
- Choose (randomly) one training instance per frequency group

Active Learning: random sampling vs. quota sampling CoNLL 2003

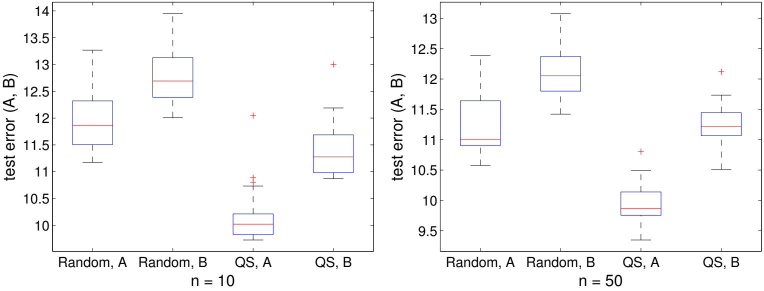
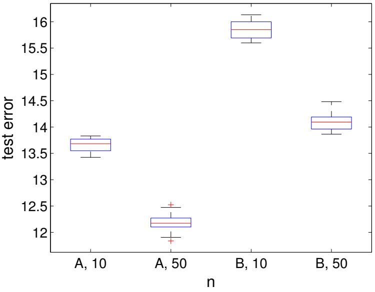


Figure: CoNLL 2003 data set. Comparison of error rates (for test A and test B sets) while training on $n = 10$ and $n = 50$ sequences. Active learning based on marginal probability (QS on the boxplots) is much more efficient than arbitrary choice of observations for training. Quota sampling outperforms random sampling.

Active Learning: FuSAL/Fully Supervised Active Learning, *(Tomanek et al., 2009)*, CoNLL 2003

m – number of examples selected within one loop
$\mathcal{D}_l$ – set of labeled instances
$\mathcal{D}_u$ – set of unlabeled instances
$u_\theta(\mathbf{x})$ – utility function

while stopping criterion is not met **do**
  train model M using D_l
  estimate $u_\theta(\mathbf{x}_i)\ \forall \mathbf{x}_i \in \mathcal{D}_u$
  choose m examples whose $u_\theta(\mathbf{x})$ is maximal
  get labels for the m chosen instances
  move the m labeled examples from $\mathcal{D}_u$ to $\mathcal{D}_l$
end while



Conclusions and Perspectives

- Conclusions
  - If the number of observations is small, state-of-the-art methods are not stable
  - The quota-based active learning outperforms state-of-the art methods on real data sets
  - Application of the semi-supervised criterion is problematic (marginal probability approximation)
- Perspectives
  - Approximation of marginal probability of structured data (graphical models)
  - Theoretical analysis of the pool-based active learning method
  - Theoretical analysis of the non-asymptotic case of the semi-supervised criterion