Discriminative Parameter Learning for Bayesian Networks

Jiang Su, Harry Zhang,
Charles X. Ling, Stan Matwin
University of Ottawa

Introduction

- Learning Bayesian networks includes structure and parameter learning
- Parameter learning is an inner loop of structure learning
- An **efficient** and **effective** parameter learning method is required in Bayesian network learning

Introduction

- The traditional parameter learning method is Frequency Estimate (FE)
- The objective function of FE is likelihood
- The objective function of classifiers should be discriminative (accuracy, conditional likelihood, etc)

Related Works

- Extended Logistic Regression (ELR) performs better than FE{Greiner2002}
  - use FE to learn plug-in parameters
  - conjugate gradient and line search
  - cross tuning

- Gradient descent methods are computationally expensive in structure learning {Friedman1997, Grossman2004}

Frequency Estimate

- An example:

Smoke	Gender
N	F
N	F
Y	F

$$\hat{P}(Smoke = N|Gender = F) = \frac{N_{Smoke=N,Gender=F}}{N_{Gender=F}} = \frac{2}{3}$$

Algorithm 1 Frequency Estimate

1. **For** each training instance e^t, $t = 1, 2, \cdot, \cdot, \cdot$.
   - **For** each attribute
     - Let $f_i^{t+1}=f_i^t+1$.

Frequency Estimate

Smoke	Gender
N	F
N	F
Y	F

Δf_i

Smoke	Gender
1	1
1	1
1	1

- Frequency information in data:
  - The frequency of Smoke=N or Y equals to the frequency of Gender=F
  - The frequency of Smoke=N is not greater than the frequency of Gender=F
  - ……
- Frequency information offers constraints during parameter learning

Discriminative Frequency Estimate

- Idea: **discriminatively** count the **frequencies** in data
- Example:

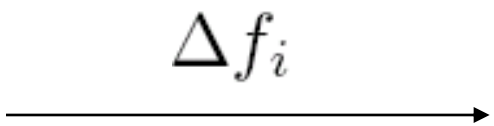
$$\hat{P}(Smoke = N|Gender = F) = \frac{\sum_{t=1}^{N_{Smoke=N,Gender=F}} error(e^t)}{\sum_{t=1}^{N_{Gender=F}} error(e^t)}$$

Algorithm 2 Discriminative Frequency Estimate

1. **For** each training instance e^t, $t = 1, 2, \cdot, \cdot, \cdot$.
   - Compute $error(e^t)$
   - **For** each attribute
     - Let $f_i^{t+1} = f_i^t + error(e^t)$.

Discriminative Frequency Estimate

Smoke	Gender
N	F
N	F
Y	F



Smoke	Gender
0.5	0.5
1	1
0.7	0.7

- Frequency information in data:
  - The frequency of Smoke=N or Y equals to the frequency of Gender=F
  - The frequency of Smoke=N is not greater than the frequency of Gender=F

Comparisons

The Δf_i matrix from different algorithms

FE

Smoke	Gender
1	1
1	1
1	1

DFE

Smoke	Gender
0.5	0.5
1	1
0.7	0.7

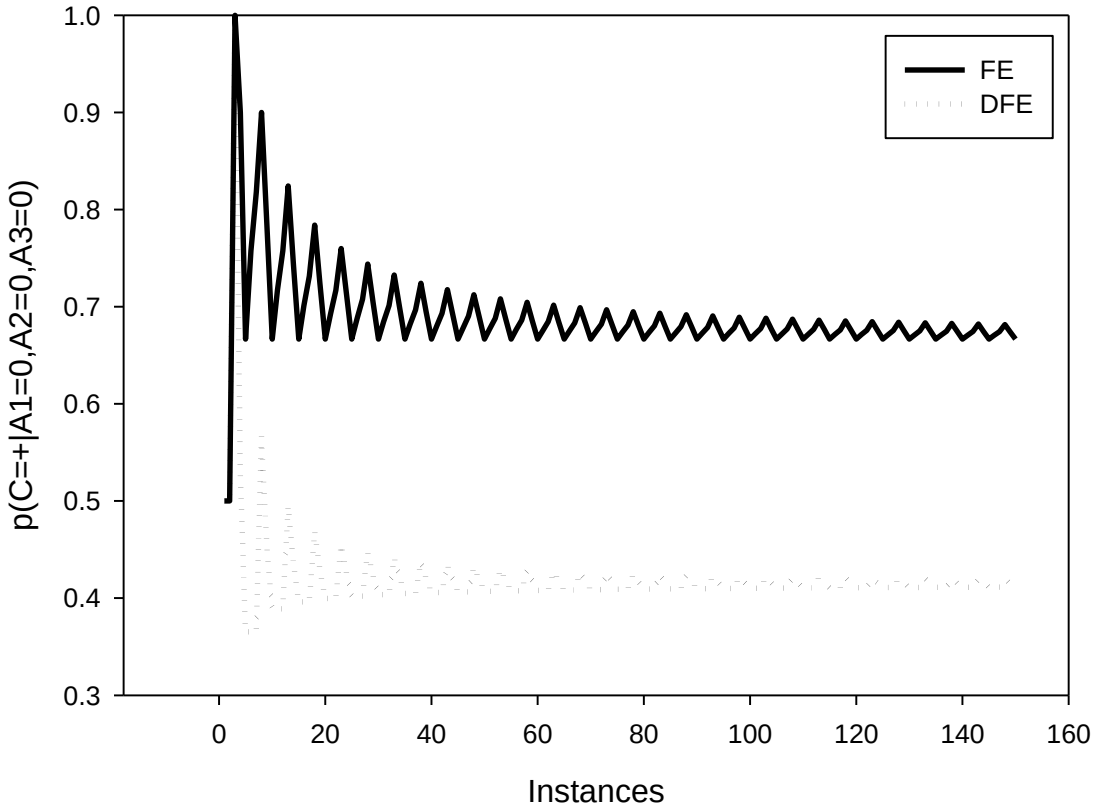
Gradient Descent

Smoke	Gender
0.1	0.8
0.5	0.1
0.7	0.3

Discriminative Frequency Estimate

Example: a dataset with 3 training instances, and 1 test instance

- The predictions from FE and DFE are influenced by the frequency information in data
- DFE converges slightly slower than FE



Experimental Setup

- 33 UCI datasets (2 classes, discretization, missing value)

- Parameter learning methods:
  - FE: frequency estimate
  - DFE: discriminative frequency estimate
  - ELR: a gradient descent method {Greiner2002}
  - Ada: use Ada boosting method to generate a set of Bayes classifiers{Freund96}
- Structure learning methods:
  - HGC: hill-climbing search algorithm (2 parents)

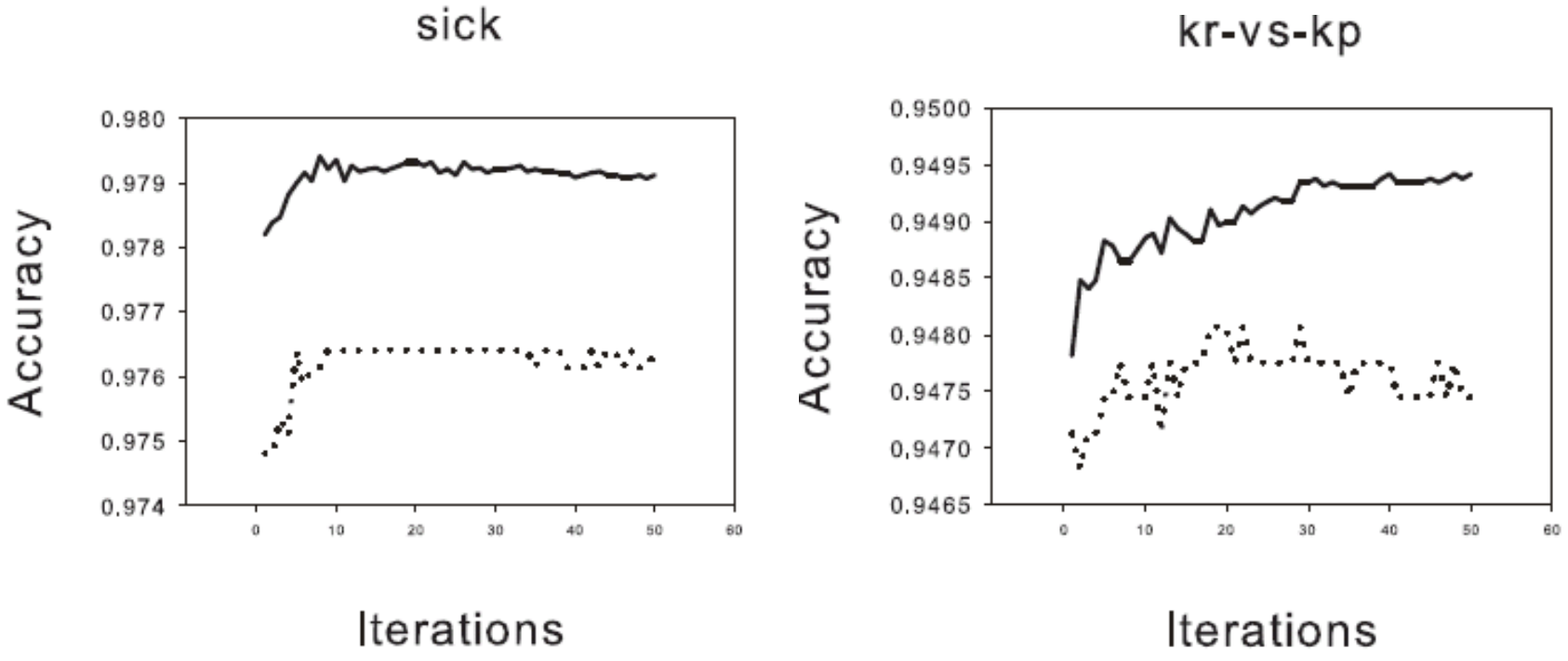
Experiments-accuracy

- DFE performs competitively with ELR, and both of them are better than FE and Ada
- Structure learning improves the performance of Bayes classifiers.(HGC+FE>NB+FE)
- NB+ELR=HGC+FE, HGC+DFE=NB+DFE

	NB+FE	NB+ELR	NB+Ada	HGC+FE	HGC+DFE
NB+DFE	12/21/0	1/31/1	9/24/0	5/28/0	3/28/2
NB+FE		0/22/11	9/19/5	0/22/11	1/22/10
NB+ELR			4/29/0	4/27/2	2/28/3
NB+Ada				2/26/5	0/28/5
HGC+FE					0/30/3

Experiments-convergence

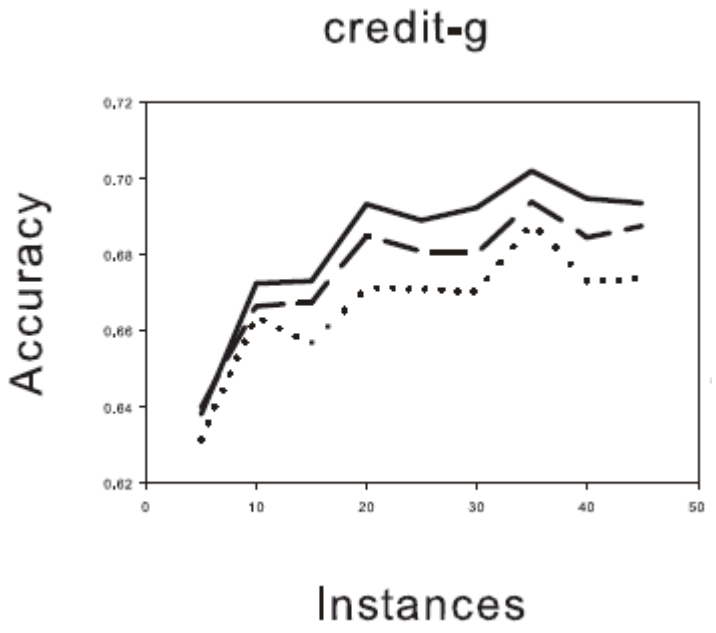
- Training Time: DFE is 250,000 times faster than ELR
- Small datasets with strong dependencies require more than 1 iteration (vowel, 200 instances, 4 iterations)
- Overfitting: training and test data accuracy are similar and increased training effort does not change the accuracy

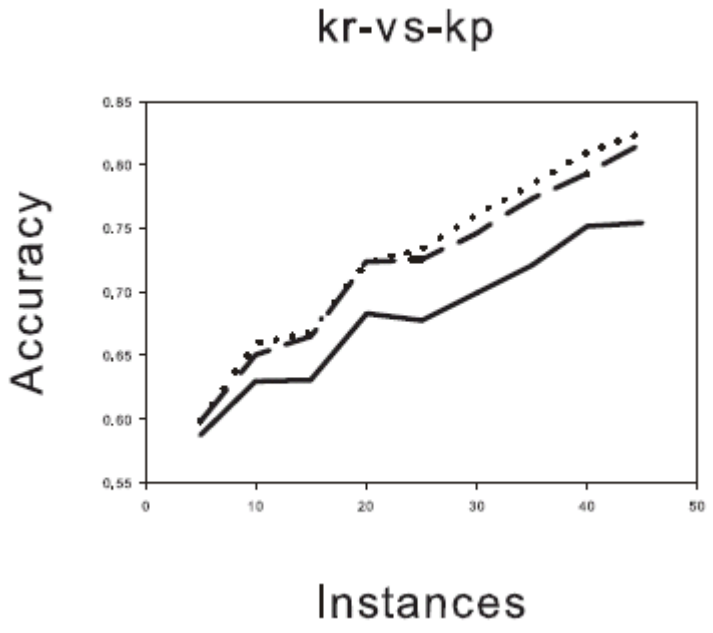


Solid: NB+DFE in training data Dotted: NB+DFE in test data

Experiments-learning curve

- Generative parameter learning does not have advantage over discriminative parameter learning in small training data





Solid: NB+FE
Dotted: NB+DFE
Dash: NB+ELR

Conclusions

- A parameter learning method for Bayesian network classifiers
  - competitive with the gradient descent method in accuracy
  - computationally efficient
  - Insensitive to the overfitting problem
  - simple to implement