Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi
Introduction

• Learning from large data
Introduction

• Learning from large data is not just about scaling-up existing algorithms.
Introduction

• Learning from large data is not just about scaling-up existing algorithms.

• Large data is best tackled by fundamentally new learning algorithms.
• Error typically reduces as data quantity increases.
• Error typically reduces as data quantity increases.
• Different algorithms have different curves.
• Algorithms that closely fit complex multivariate distributions will tend to overfit small data
• Algorithms that closely fit complex multivariate distributions will tend to overfit small data, but can better fit large data
• Algorithms that closely fit complex multivariate distributions will tend to overfit small data, but can better fit large data: *Bias-Variance Tradeoff*
Capacity to fit = model space + optimization

RMSE vs Training set size for Naïve Bayes and Logistic Regression.
Much research of questionable relevance

- Most prior research has used relatively small data.
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction

Bayesian Network Classifiers

Selective KDB Experiments

Conclusions & Future Research

Requirements

- Need algorithms that can closely fit complex multivariate distributions
• Need algorithms that can closely fit complex multivariate distributions, while being very computationally efficient.
• Need algorithms that can closely fit complex multivariate distributions, while being very computationally efficient.
  • low bias
• Need algorithms that can closely fit complex multivariate distributions, while being very computationally efficient.
  • low bias
  • few passes through data
• Need algorithms that can closely fit complex multivariate distributions, while being very computationally efficient.
  • low bias
  • few passes through data
  • out of core
State-of-the-art

- Most low bias algorithms do not scale
  - Random Forest
  - SVM
  - Deep Learning
State-of-the-art

• Most low bias algorithms do not scale
  • Random Forest
  • SVM
  • Deep Learning
and inherently in-core
State-of-the-art

- Most low bias algorithms do not scale
  - Random Forest
  - SVM
  - Deep Learning
  and inherently in-core
- Selective KDB is a scalable low-bias Bayesian Network Classifier
Introduction
Bayesian Network Classifiers
Selective KDB
Experiments
Conclusions & Future Research
Bayesian Network Classifiers

- Defined by parent relation $\pi$ and Conditional Probability Tables (CPTs)
  - $\pi$ encodes conditional independence / structure
  - CPTs encode conditional probabilities
- Classifies using $P(y \mid x) \propto P(y \mid \pi_Y) \prod P(x_i \mid \pi_i)$
- Usually makes $Y$ a parent of all $X_i$

Given $\pi$, CPTs can be learned by counting joint frequencies
- single pass
- incremental
k-Dependence Bayes (KDB)

- Two pass learning
- 1st pass, learn structure:
  - Collect counts for pairs of attributes with the class.
  - Order attributes based on MI with the class.
  - Select parents based on CMI.
    - no more than $k$ parents
    - parents must be earlier in the order

![Diagram](attachment:image.png)
k-Dependence Bayes (KDB)

- Two pass learning
- 1st pass, learn structure:
  - Collect counts for pairs of attributes with the class.
  - Order attributes based on MI with the class.
  - Select parents based on CMI.
    - no more than $k$ parents
    - parents must be earlier in the order

- 2nd pass, learn CPTs:
  - Collect statistics according to the structure learned.
k-Dependence Bayes (KDB)

- Training Space: $O(ya^2v^2 + yav^{k+1})$
- Classification Space: $O(yav^{k+1})$
- Training Time: $O(ta^2 + ya^2v^2 + tak)$
- Classification Time: $O(yak)$

$t = \text{no. of training examples}$

$a = \text{number of attributes; }$

$v = \text{average number of values; }$

$y = \text{number of classes }$
k-Dependence Bayes (KDB)

- Parameter $k$ controls bias-variance tradeoff

![Graph showing the performance of Naive Bayes and KDB with different $k$ values across varying training set sizes.](image-url)
k-Dependence Bayes (KDB)

- Parameter $k$ controls bias-variance tradeoff

- No a priori means to anticipate best $k$
k-Dependence Bayes (KDB)

- Parameter $k$ controls bias-variance tradeoff

- No a priori means to anticipate best $k$
- Spurious attributes may increase error
KDB models are nested

- $M_{i,j} = \text{KDB with } k = i \text{ using attributes 1 to } j$.

<table>
<thead>
<tr>
<th>Attributes (ordered by MI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
KDB models are nested

- $M_{i,j} = \text{KDB with } k = i \text{ using attributes } 1 \text{ to } j$.  

<table>
<thead>
<tr>
<th>Attributes (ordered by MI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

- $M_{i,j}$ is a minor extension of $M_{i,j-1}$ and $M_{i-1,j}$
KDB models are nested

- $M_{i,j} =$ KDB with $k = i$ using attributes 1 to $j$.

**Attributes (ordered by MI)**

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$M_{1,1}$</td>
<td>$M_{1,2}$</td>
<td>$M_{1,3}$</td>
<td>$M_{1,4}$</td>
</tr>
<tr>
<td>2</td>
<td>$M_{2,1}$</td>
<td>$M_{2,2}$</td>
<td>$M_{2,3}$</td>
<td>$M_{2,4}$</td>
</tr>
<tr>
<td>3</td>
<td>$M_{3,1}$</td>
<td>$M_{3,2}$</td>
<td>$M_{3,3}$</td>
<td>$M_{3,4}$</td>
</tr>
<tr>
<td>4</td>
<td>$M_{4,1}$</td>
<td>$M_{4,2}$</td>
<td>$M_{4,3}$</td>
<td>$M_{4,4}$</td>
</tr>
</tbody>
</table>

- $M_{i,j}$ is a minor extension of $M_{i,j-1}$ and $M_{i-1,j}$
KDB models are nested

- \( M_{i,j} = \text{KDB with } k = i \text{ using attributes 1 to } j \).

<table>
<thead>
<tr>
<th>Attributes (ordered by MI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

- \( M_{i,j} \) is a minor extension of \( M_{i,j-1} \) and \( M_{i-1,j} \)
KDB models are nested

- $M_{i,j} = \text{KDB with } k = i \text{ using attributes 1 to } j$.

### Attributes (ordered by MI)

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$M_{1,1}$</td>
<td>$M_{1,2}$</td>
<td>$M_{1,3}$</td>
<td>$M_{1,4}$</td>
</tr>
<tr>
<td>2</td>
<td>$M_{2,1}$</td>
<td>$M_{2,2}$</td>
<td>$M_{2,3}$</td>
<td>$M_{2,4}$</td>
</tr>
<tr>
<td>3</td>
<td>$M_{3,1}$</td>
<td>$M_{3,2}$</td>
<td>$M_{3,3}$</td>
<td>$M_{3,4}$</td>
</tr>
<tr>
<td>4</td>
<td>$M_{4,1}$</td>
<td>$M_{4,2}$</td>
<td>$M_{4,3}$</td>
<td>$M_{4,4}$</td>
</tr>
</tbody>
</table>

- $M_{i,j}$ is a minor extension of $M_{i,j-1}$ and $M_{i-1,j}$
Selective KDB

- In one extra pass select an attribute subset and best $k$ using leave-one-out CV.
Selective KDB

- In one extra pass select an attribute subset and best $k$ using leave-one-out CV.
- The full model subsumes all $k \times a$ submodels.
Selective KDB

• In one extra pass select an attribute subset and best $k$ using leave-one-out CV.
• The full model subsumes all $k \times a$ submodels.
• Very efficient selection between a large class of strong models.
Leave-one-out CV

- Very low bias estimator of out-of-sample performance
- Incremental cross-validation makes it VERY efficient for BNC
  - Collect counts from all data once
  - When classifying a hold-out object subtract it from the counts
- All $k \times a$ nested models can be evaluated with little more computation than the full KDB model
Selective KDB

- Training Space: $O(ya^2v^2 + yav^{k+1})$
- Classification Space: $O(ya^*v^{k*+1})$
- Training Time: $O(ta^2 + ya^2v^2 + tayk)$
- Classification Time: $O(ya^*k^*)$

$a = \text{number of attributes}$;
$v = \text{average number of values}$;
$y = \text{number of classes}$
$a^* = \text{number of attributes selected (}a^* \leq a)$. 
$k^* = \text{best value of } k \text{ found (}k^* \leq k_{max})$. 
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB
Experiments
Conclusions & Future Research

SelecteKDB

RMSE vs Training set size

Naïve Bayes
KDB k=1
KDB k=2
KDB k=3
KDB k=4
KDB k=5
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB
Experiments
Conclusions & Future Research

Selective KDB

- Naïve Bayes
- KDB k=1
- KDB k=2
- KDB k=3
- KDB k=4
- KDB k=5
- SKDB Only K

RMSE vs. Training set size graph.
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB Experiments
Conclusions & Future Research

Selective KDB

Naïve Bayes
KDB k=1
KDB k=2
KDB k=3
KDB k=4
KDB k=5
SKDB Only K
SKDB k=5 Only Atts

RMSE vs. Training set size graph
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB
Experiments
Conclusions & Future Research

Selective KDB

- Naïve Bayes
- KDB k=1
- KDB k=2
- KDB k=3
- KDB k=4
- KDB k=5
- SKDB Only K
- SKDB k=5 Only Atts
- SKDB

RMSE vs Training set size

Graph showing the performance of different classifiers (Naïve Bayes, KDB k=1 to k=5, SKDB Only K, SKDB k=5 Only Atts, SKDB) across various training set sizes. The graph compares the Root Mean Square Error (RMSE) for each classifier as the training set size increases from 0 to 1,000,000.
## 16 datasets (165K-54M examples)

<table>
<thead>
<tr>
<th>Name</th>
<th>No. of cases (million)</th>
<th>No. of Atts.</th>
<th>No. of Classes</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>localization</td>
<td>0.165</td>
<td>5</td>
<td>11</td>
<td>11MB</td>
</tr>
<tr>
<td>census-income</td>
<td>0.299</td>
<td>41</td>
<td>2</td>
<td>136MB</td>
</tr>
<tr>
<td>USPSExtended</td>
<td>0.342</td>
<td>676</td>
<td>2</td>
<td>603MB</td>
</tr>
<tr>
<td>MITFaceSetA</td>
<td>0.474</td>
<td>361</td>
<td>2</td>
<td>584MB</td>
</tr>
<tr>
<td>MITFaceSetB</td>
<td>0.489</td>
<td>361</td>
<td>2</td>
<td>603MB</td>
</tr>
<tr>
<td>MSDYear-Prediction</td>
<td>0.515</td>
<td>90</td>
<td>90</td>
<td>601MB</td>
</tr>
<tr>
<td>covtype</td>
<td>0.581</td>
<td>54</td>
<td>7</td>
<td>72MB</td>
</tr>
<tr>
<td>MITFaceSetC</td>
<td>0.839</td>
<td>361</td>
<td>2</td>
<td>1.1GB</td>
</tr>
<tr>
<td>poker-hand</td>
<td>1.025</td>
<td>10</td>
<td>10</td>
<td>24MB</td>
</tr>
<tr>
<td>uscensus1990</td>
<td>2.458</td>
<td>67</td>
<td>4</td>
<td>325MB</td>
</tr>
<tr>
<td>PAMAP2</td>
<td>3.851</td>
<td>54</td>
<td>19</td>
<td>1.7GB</td>
</tr>
<tr>
<td>kddcup</td>
<td>5.210</td>
<td>41</td>
<td>40</td>
<td>754MB</td>
</tr>
<tr>
<td>linkage</td>
<td>5.749</td>
<td>11</td>
<td>2</td>
<td>251MB</td>
</tr>
<tr>
<td>mnist8ms</td>
<td>8.100</td>
<td>784</td>
<td>10</td>
<td>19GB</td>
</tr>
<tr>
<td>satellite</td>
<td>8.705</td>
<td>138</td>
<td>24</td>
<td>3.6GB</td>
</tr>
<tr>
<td>splice</td>
<td>54.628</td>
<td>141</td>
<td>2</td>
<td>7.3GB</td>
</tr>
</tbody>
</table>

s sparse format.
Numeric attributes

- 5 bin equal frequency discretisation
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB Experiments
Conclusions & Future Research

Out-of-core BNCs

RMSE

KDB5 vs SKDB

best KDB vs SKDB
Out-of-core BNCs

- Training Time Comparisons

![Training Time Comparisons Graph]

- Datasets:
  - localization
  - Census-income
  - Poker-hand
  - donation
  - covtype
  - uscensus
  - MITFaceSetA
  - MITFaceSetB
  - UPSSextended
  - MITFaceSetC
  - kddcup

- Classifiers:
  - NB
  - TAN
  - AODE
  - KDB k=5
  - SKDB
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction

Bayesian Network Classifiers

Selective KDB Experiments

Conclusions & Future Research

Out-of-core BNCs

• Classification Time Comparisons

Datasets:
- localization
- Poker-hand
- Census-income
- covtype
- uscensus
- MITFaceSetA
- MITFaceSetB
- MITFaceSetC
- kddcup

Time (seconds)

- NB
- TAN
- AODE
- KDB
- SKDB

- k=4
- k=4
- k=2
- k=4.8
- k=5
- k=4
- k=3
- k=4
- k=5
- k=5
- k=4.5

- 0,1
- 1
- 10
- 100
- 1000
Out-of-core SGD

- SGD in Vowpal Wabbit (VW):
  - Squared and logistic function.
  - Quadratic features (best results).
  - Different number of passes (3 or 10).
  - Discrete attributes into binary features.
  - One-against-all for multiclass classification.
Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB
Experiments
Conclusions & Future Research

Out-of-core SGD

VW (logistic) - RMSE
VW (squared) - 0-1 Loss

<table>
<thead>
<tr>
<th>VW</th>
<th>RMSE (logistic)</th>
<th>0-1 Loss (squared)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective KDB</td>
<td>(7+2)-0-7</td>
<td>8-1-7</td>
</tr>
</tbody>
</table>
Out-of-core SGD

- Training Time Comparisons

![Graph showing comparisons of training times for different datasets and algorithms (VWSF, VWLF, SKDB).]
Out-of-core SGD

- Classification Time Comparisons

Datasets:
- localization
- Poker-hand
- Census-income
- donation
- covtype
- uscensus
- MITFaceSetA
- UPSExtended
- MITFaceSetB
- MITFaceSetC
- kddcup
In-core BayesNet and RF

BayesNet

RF

$\times = \text{sampled dataset}$

<table>
<thead>
<tr>
<th>$k$-selective KDB</th>
<th>BayesNet</th>
<th>RF (Num)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-3-3</td>
<td>5-0-6</td>
<td></td>
</tr>
</tbody>
</table>
In-core BayesNet and RF

- Training Time Comparisons

![Graph showing training time comparisons for different datasets.]
In-core BayesNet and RF

Classification Time Comparisons

Datasets

- Localization
- Census-income
- Donation
- Covtype
- MITFaceSetA
- MITFaceSetB
- USPSExtended

Time (seconds)

- BayesNet
- RF
- SKDB
Global comparisons

- **Cumulative Ranking**

- Selective KDB copes well with high-dimensional datasets and datasets with more than 1 million points.
- RF performs better on datasets with small number of attributes.
- VW has advantage for sparse numeric data.
Global comparisons

- Ranking

<table>
<thead>
<tr>
<th>Learners</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>7.63</td>
</tr>
<tr>
<td>AODE</td>
<td>6.81</td>
</tr>
<tr>
<td>TAN</td>
<td>5.88</td>
</tr>
<tr>
<td>BayesNet</td>
<td>3.47</td>
</tr>
<tr>
<td>VWLF</td>
<td>3.44</td>
</tr>
<tr>
<td>KDB (best k)</td>
<td>3.34</td>
</tr>
<tr>
<td>RF (Num)</td>
<td>2.91</td>
</tr>
<tr>
<td>SKDB</td>
<td>2.53</td>
</tr>
</tbody>
</table>

- Selective KDB Experiments

Conclusions & Future Research
Global comparisons (Time)

Training

Test

Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction

Bayesian Network Classifiers

Selective KDB

Experiments

Conclusions & Future Research
Stepping back

- The key trick is nested evaluation of a large class of count-based models
- Works also with 2-pass selective evaluation of AnDE models
- Combines the efficiency of simple generative techniques with the power of discriminative learning
- Can use any loss function that is a function of each $\hat{P}(y \mid x)$
Future Research

- Numeric attributes
  - it seems that there should be something better than discretisation.
  - this remains elusive ...

- Overfitting avoidance

- Increase number of alternative models considered

- Explore other forms of nested BNCs

- Two pass Selective KDB
  - sample small test set in second pass

- Single pass generative/discriminative learning
  - initially learn a simple generative model
  - collect discriminative statistics and refine the model when there is sufficient evidence to make a choice
    - refinements might be attribute selection, structural refinement or weighting
  - repeat
Satellite learning curves
Satellite learning curves

Scalable Bayesian Network Classifiers

Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction

Bayesian Network Classifiers

Selective KDB Experiments

Conclusions & Future Research
Satellite learning curves

Scalable Bayesian Network Classifiers
Geoff Webb
Ana Martinez
Nayyar Zaidi

Introduction
Bayesian Network Classifiers
Selective KDB Experiments
Conclusions & Future Research
Conclusions

• Large data calls for fundamentally new learning algorithms
• We are pioneering a new generation of theoretically well-founded algorithms that are both scalable to very large data and capable of exploiting the fine detail inherent therein