On Learning Parsimonious Predictive Models for High Dimensional Data Sets

Xue Bai \textsuperscript{12} and Rema Padman \textsuperscript{1}

\textsuperscript{1} The H. John Heinz III School of Public Policy and Management
\textsuperscript{2} School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213
\{xb, rpadman\}@cmu.edu
Outline

- Motivation
- Problem definition
- Methodology
- Evaluation
- Contributions and conclusion
Motivation

• High Dimensional Data sets with relatively few cases arise in
  – Health care
  – Text mining
  – Internet marketing

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>#Variables</th>
<th>#Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Diagnosis</td>
<td>779</td>
<td>326</td>
</tr>
<tr>
<td>Sentiment Extraction</td>
<td>7,716</td>
<td>1,400</td>
</tr>
<tr>
<td>Patronage Prediction</td>
<td>235</td>
<td>2,936</td>
</tr>
</tbody>
</table>
Internet Movie Database: Movie reviews
http://www.cs.cornell.edu/people/pabo/movie-review-data/

reviewed by

Extract the main text of the reviews using MixUp (Cohen, 2004)

Vector representation for the reviews

I, Robot → ["aah"=2,…, "awesome"=1, … ]

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>#Variables</th>
<th>#Samples</th>
<th>Variable Types ($X_i$)</th>
<th>Target Variable ($Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer opinions</td>
<td>34,000</td>
<td>1,400</td>
<td>discrete</td>
<td>Positive / Negative</td>
</tr>
</tbody>
</table>
Prostate Cancer Database (American Association for Cancer Research)

http://cancerres.aacrjournals.org/misc/terms.shtml

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>#Variables</th>
<th>#Samples</th>
<th>Variable Types ($X_i$)</th>
<th>Target Variable ($Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer diagnosis</td>
<td>779</td>
<td>326</td>
<td>discretized</td>
<td>Yes / No</td>
</tr>
</tbody>
</table>
Importance and Challenges

• Challenges
  – dimensional difficulties
  – limited data

• Importance
  – the impact of the complexity of a model (or simply the number of variables) on
    • the speed of computation
    • the quality of decisions
    • operational costs
    • understandability and user acceptance of the decision model
Problem definition

• To build effective prediction models from high dimensional and insufficient data
The goal

• To find a parsimonious *Markov Blanket Directed Acyclic Graph* (MB DAG) that:
  – Has substantially fewer predictor variables
  – Encodes the conditional dependence relations
  – Provides good prediction
Methodology

- **TS/MB - Tabu Search enhanced Markov Blanket procedure:**
  - **1\textsuperscript{st} stage:** Learning the structure of an initial MB DAG using *conditional independence tests* (Spirtes, Glymour and Scheines, 1993) and *causal reasoning* (Pearl, 1988, Spirtes, Glymour and Scheines, 2000)
  - **2\textsuperscript{nd} stage:** Using *Tabu Search* (TS) heuristic (Glover, 1989) to improve the predictive structure of the MB DAG
A Bayesian Network

–A graphical representation of the joint probability distribution $P$ of a set of random variables $S$
A Bayesian Network

• Markov factorization of $P$ according to $S$

$$p(X) = \prod_{i=1}^{n} p_i(X_i|pa_i)$$
• **Markov Blanket**
  
  – *MB*: a smallest subset of variables in $S$ such that $Y$ is independent of $S \setminus MB$, conditional on the variables in $MB$
  
  – Equals the set of parents of $Y$, children of $Y$, and the parents of children of $Y$

---

*Figure 1.* The Bayesian Network $(S, P)$  
*Figure 2.* The Markov Blanket for $Y$
Advantage of Markov Blanket Search

Figure 1. The Bayesian Network \((S, P)\)

\[
p(Y, X_1, ..., X_6) = p(Y|X_1) \cdot p(X_4|X_2, Y) \cdot p(X_5|X_3, X_4, Y) \cdot p(X_2|X_1) \cdot p(X_3|X_1) \cdot p(X_6|X_4) \cdot p(X_1)
\]

Figure 2. The Markov Blanket for \(Y\)

\[
p(Y|X_1, ..., X_6) = C' \cdot p(Y|X_1) \cdot p(X_4|X_2, Y) \cdot p(X_5|X_3, X_4, Y)
\]
Representation and Background Knowledge (cont.)

- *Tabu Search* (Glover, 1989)
  - guide traditional local search methods to escape local optima
  - the use of adaptive memory
Methodology

- Two stage algorithm:

  1. 1\textsuperscript{st} stage: Learning an initial MB DAG

  \begin{verbatim}
  InitialMBsearch {
    1.1 Finding the adjacency structure
    1.2 Pruning the graph
    1.3 Orienting edges
    1.4 Transforming into an MB DAG
  }
  \end{verbatim}
Algorithm: 1.1 finding adjacencies
Algorithm: 1.2 pruning redundant nodes and edges
Algorithm: 1.3 orienting edges
Algorithm: 1.4 pruning the graph into an MB DAG
Algorithm: 2\textsuperscript{nd} Stage – Tabu Search in Markov Blanket Space

- **Search space**
  - The space of Markov Blanket DAGs

- **Neighborhood**
  - The set of new Markov Blanket DAGs that can be constructed via one feasible move from the current Markov Blanket DAG
Algorithm: 2nd Stage – Tabu Search in Markov Blanket Space

**Moves**
edge deletion: (a) to (b),
edge addition: (b) to (c),
edge reversal: (c) to (d),
edge reversal with node pruning: (d) to (e).
Special features of Tabu Search

- **Tabu List**
  - A list of most recent moves

- **Tabu Tenure**
  - 7 for all kinds of moves

- Why Tabu Search?
  - Non-improving solution
  - Revisit
Evaluations in Three Different Domains

- Case study I, II and III
  
  I. **Text mining domain** (Movie reviewers’ opinion classification from text data)
  
  II. **Health care domain** (Prostate Cancer diagnosis from serum protein data)
  
  III. **Electronic commerce domain** (Customers’ response prediction from DealTime data)
Design of Experimental Parameters

- Design of Experimental Parameters:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Data-splits (training/testing)</th>
<th>Scoring Criteria</th>
<th>Starting Solution</th>
<th>Depth of search</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configurations</td>
<td>90/10, 80/20</td>
<td>AUC, Accuracy</td>
<td>Type I, Type II</td>
<td>1,2,3</td>
<td>0.001, 0.005, 0.01, 0.05</td>
</tr>
</tbody>
</table>
Case I - Movie reviewer opinions classification

- **Data source**: Internet Movie Database
  - Originally selected and studied by Pang et al., 2002

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>#Variables</th>
<th>#Samples</th>
<th>Variable Types ($X_i$)</th>
<th>Target Variable (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer opinions</td>
<td>7,716</td>
<td>1,400</td>
<td>binary</td>
<td>Positive / Negative</td>
</tr>
</tbody>
</table>
Computational Result I (Movie review data)

Average performance when using the full set of variables as input

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC (%)</th>
<th>Accuracy (%)</th>
<th>#Original Features</th>
<th># Selected Features</th>
<th>Size Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB</td>
<td>71.24</td>
<td>65.00</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>TS/MB</td>
<td>96.85</td>
<td>87.52</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>82.61</td>
<td>66.22</td>
<td>7,716</td>
<td>7,716</td>
<td>0%</td>
</tr>
<tr>
<td>SVM + TFIDF</td>
<td>81.32</td>
<td>84.07</td>
<td>7,716</td>
<td>7,716</td>
<td>0%</td>
</tr>
<tr>
<td>Voted perceptron</td>
<td>77.09</td>
<td>70.00</td>
<td>7,716</td>
<td>7,716</td>
<td>0%</td>
</tr>
<tr>
<td>Max. entropy</td>
<td>75.79</td>
<td>79.43</td>
<td>7,716</td>
<td>7,716</td>
<td>0%</td>
</tr>
</tbody>
</table>
Computational Result II (Movie review data)

Average performance when using the same number of variables as identified by TS/MB as input for all the other classifiers, selected by information gain criterion

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC (%)</th>
<th>Accuracy (%)</th>
<th>#Original Features</th>
<th># Selected Features</th>
<th>Size Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB</td>
<td>71.24</td>
<td>65.00</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>TS/MB</td>
<td>96.85</td>
<td>87.52</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>78.85</td>
<td>72.07</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>SVM + TFIDF</td>
<td>67.30</td>
<td>70.43</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Voted perceptron</td>
<td>78.68</td>
<td>71.71</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Max. entropy</td>
<td>68.42</td>
<td>71.93</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
</tbody>
</table>
Computational Result III (Movie review data)

Average performance when using the exact same variables as identified by TS/MB as input

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC (%)</th>
<th>Accuracy (%)</th>
<th>#Original Features</th>
<th># Selected Features</th>
<th>Size Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB</td>
<td>71.24</td>
<td>65.00</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>TS/MB</td>
<td>96.85</td>
<td>87.52</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>81.81</td>
<td>73.36</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>SVM + TFIDF</td>
<td>69.47</td>
<td>72.00</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Voted perceptron</td>
<td>80.61</td>
<td>73.93</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
<tr>
<td>Max. entropy</td>
<td>69.81</td>
<td>73.44</td>
<td>7,716</td>
<td>22</td>
<td>99.71%</td>
</tr>
</tbody>
</table>
The MB DAG learned from the Movie Review data
Early-Stage Results from Financial News Corpora

  - Mergers and acquisitions (M & A, 600 documents)
  - Finance (600 documents)
  - Mixed news (600 documents)

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>#Variables</th>
<th>#Samples</th>
<th>Variable Types (X_i)</th>
<th>Target Variable (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M &amp; A</td>
<td>10,532</td>
<td>600</td>
<td>binary</td>
<td>Positive /neutral / Negative</td>
</tr>
<tr>
<td>Finance</td>
<td>11,221</td>
<td>600</td>
<td>binary</td>
<td>Positive /neutral / Negative</td>
</tr>
<tr>
<td>Mixed news</td>
<td>15,686</td>
<td>600</td>
<td>binary</td>
<td>Positive /neutral / Negative</td>
</tr>
</tbody>
</table>
Early-Stage Results from Financial News Corpora

The average performance, with three possible opinions

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>Accuracy (%)</th>
<th>#Original Features</th>
<th># Selected Features</th>
<th>Size Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>M&amp;A</td>
<td>TS/MB</td>
<td>89.95</td>
<td>7,166</td>
<td>14</td>
<td>99.80%</td>
</tr>
<tr>
<td></td>
<td>Best competitor (SVM)</td>
<td>68.50</td>
<td>7,166</td>
<td>7,166</td>
<td>0%</td>
</tr>
<tr>
<td>Finance</td>
<td>TS/MB</td>
<td>96.31</td>
<td>7,166</td>
<td>16</td>
<td>99.78%</td>
</tr>
<tr>
<td></td>
<td>Best competitor (SVM)</td>
<td>69.00</td>
<td>7,166</td>
<td>7,166</td>
<td>0%</td>
</tr>
<tr>
<td>Mixed</td>
<td>TS/MB</td>
<td>91.97</td>
<td>7,166</td>
<td>12</td>
<td>99.83%</td>
</tr>
<tr>
<td></td>
<td>Best competitor (SVM)</td>
<td>70.20</td>
<td>7,166</td>
<td>7,166</td>
<td>0%</td>
</tr>
</tbody>
</table>
Contributions

1. Learning the Markov Blanket rather than Bayesian Network for classification
   - Reduces the search space and problem complexity
   - Theoretically correct in limit
2. Search directly for prediction score
   - Vs. criteria such as BIC
3. Tabu search among the Markov Blanket space
   - Tabu search dynamic memory is the key to improve the classification
4. The effectiveness, robustness and superiority are tested extensively on diverse real world data
   - No previous research on graphical classification models has been empirically validated
Conclusions

• **TS/MB classifier:**
  – reduces the set of predictor variables
  – encodes the conditional dependencies among features
  – provides excellent classification results

• **Theoretically correct in limit**

• **Empirically effective with finite samples**