A Concise but Effective Discriminative Pattern-Based Predictive Model for Personalized Medicine

Jingbo Shang, Wenzhu Tong, Jinfeng Xiao, Jian Peng, Jiawei Han
University of Illinois, Urbana-Champaign
Mar 4, 2016
Presented by Jingbo Shang
Outline

- Motivation
- Related Work
- DPMed: Methodology
- Experimental Results
- Discussion and Future Work
Motivation: Personalized Medicine

- Personalized Medicine is very popular and desired
Motivation: Discriminative Patterns

- Trade-offs
  - Complex Models v.s. Simple Models
  - Accuracy v.s. Explanations
- Discriminative Patterns-based
  - Bridge Complex and Simple Models
  - Comparable Accuracy
  - Concise Model and thus Interpretable
Outline

- Motivation
- Related Work
- DPMed: Methodology
- Experimental Results
- Discussion and Future Work
## Related Work

- **Complex Models**
  - Multi-tree based Models: (Depth Limited) Random Forest
  - Accurate but NOT Interpretable
- **Discriminative Patterns-based Classification**
- **DDPMine**
- **DPClass and its variant DPReg for regression**
- **Comparably accurate and interpretable but NO Personalization**

### Table 1: Compared methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>DPMed</th>
<th>DPClass</th>
<th>DDPMine</th>
<th>LRF</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretable</td>
<td>✓</td>
<td>✓</td>
<td>somehow</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Personalized</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Scalable</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Related Work: DPClass

- Our previous work
  - Jingbo Shang, Wenzhu Tong, Jian Peng, and Jiawei Han, *DPClass: An Effective but Concise Discriminative Patterns-Based Classification Framework*, In Proceedings of 2016 SIAM International Conference on Data Mining (SDM 2016), 2016
DPClass

Training Dataset

Multiple Tree-based Model

Top-k Discriminative Patterns

Testing Dataset

Discriminative Patterns Generation

Linear Model Training

Efficient Testing

Possible discriminative patterns:
- A non-leaf node & a discriminate pattern
- A selected discriminative pattern
- A non-selected discriminative pattern

Compressed Model

Efficient Testing

Linear Model Training

Possible discriminative patterns:
- A non-leaf node & a discriminate pattern
- A selected discriminative pattern
- A non-selected discriminative pattern

\[
0.8 + 0.5 * b - 1 * g + 2.1 * f - 0.7 * j
\]
DPClass I: The Specific Task

- We discuss binary classification here
- \( N \) training instances \((x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\)
- \( \forall 1 \leq i \leq N, y_i \in \{+1, -1\} \)
- \( x_i \) is the feature vector of \( i \)-th instance
  - Both numeric (continuous) and categorical (discrete) variables are acceptable
DPClass II: Discriminative Patterns Generation

- Random Forest
  - Maximize the randomness
    - Random features
    - Random partitions
    - Random instances (bootstrap)
- Prefix paths are treated as discriminative patterns
DPClass II: Discriminative Patterns Generation

- Parameters
  - # of trees = $T$
  - loss function = information gain
  - depth $\leq d$
  - support $\geq \sigma$ (based on bootstrapped instances)

- We admit all prefix of these tree-paths as patterns
  - # of leaves $\leq \min \left\{ 2^d, \frac{N}{\sigma} \right\} \cdot T$
  - # of candidate patterns $\leq \min \left\{ 2^d, \frac{N}{\sigma} \right\} \cdot T \cdot d$

- Assume $T = 100$, # of candidate pattern $\sim 10^4$
DPClass III: Top-$k$ Selection

- Similar to Feature Selections
- Select a top-$k$ discriminative patterns

Implementation
- Forward Selection (Greedy)
- LASSO (GLMNET)
DPClass IV: Generalized Linear Models

- Binary Feature Representations (bag-of-patterns)
- Logistic Regression for binary classifications
Outline

- Motivation
- Related Work
- DPMed: Methodology
- Experimental Results
- Discussion and Future Work
DPMed: Overview

Learn Top-3 Patterns

Detect Patient Clusters

Group 1

Group 2

Group 3

Learn Top-1 Patterns

Local Patterns

Construct Bag-of-patterns

Global Patterns

Global Patterns

Local Patterns

Train Generalized Linear Model

Personalized Medicine

Patient

Symptom

Pattern
Top-\(k\) Pattern Generation

Algorithm 1: Learning Top-\(k\) Patterns

Require: A set of patients \(\mathcal{S}\), the loss function \(l\), the number of trees \(T\), the depth threshold \(D\), minimum tree bag size \(\sigma\), parameter \(k\), and a small value \(\epsilon\)

Return: Top-\(k\) patterns \(\mathcal{P}_k\)

\(\mathcal{P} \leftarrow \emptyset, \; x' \leftarrow 0, \; \mathcal{P}_k \leftarrow \emptyset\)

for \(t = 1\) to \(T\) do

| Build a random decision tree \(\mathcal{V}\) based on the patients in \(\mathcal{S}\) using the loss function \(l\) with the maximum depth \(D\) and the minimum tree bag size \(\sigma\). |

| for each non-leaf node \(u\) do |

| \(\mathcal{P} \leftarrow \mathcal{P} \cup \{\text{the path: root} \rightarrow u\}\) |

for each patient \(i\) in \(\mathcal{S}\) do

| for each pattern \(p\) in \(\mathcal{P}\) do |

| if patient \(i\) satisfies pattern \(p\) then |

| \(x_{ip}' \leftarrow 1\) |

Top-\(k\) Patterns \(\mathcal{P}_k \leftarrow \text{top-}k\text{ features solved by LASSO as Equation}\ [1]\)

return \(\mathcal{P}_k\)
Diagnose Group Detection

- Bag of patterns
- LDA
Outline

- Motivation
- Related Work
- DPMed: Methodology
- Experimental Results
- Discussion and Future Work
Table 2: Test Accuracy in Classification Tasks and Test RMSE in Regression Tasks. DPClass and DDPMine are designed to attack classification problems. DPRreg is the regression variant of DPClass. DDPMine is too slow to get results on high dimensional datasets. Utilizing \(K_g = 20\) and \(K_l = 5\) in classification tasks and \(K_g = 30\) and \(K_l = 10\) in regression tasks, in most of the time, DPMed achieves the best performance among interpretable models (DPMed, DPClass, DPRreg, and DDPMine) and comparable accuracy as the complex models (RF and LRF). The number of equivalent patterns adopted in RF and LRF is at least 100 times more.
Comments:

• not very sensitive to the number of clusters
• there should not be a large number of clusters
  • the insufficient number of patient instances
  • tens of patterns are not enough to distinguish patients into too many clusters
# of Patterns

(a) Classification Tasks  
(b) Regression Tasks

Figure 5: Vary number of global patterns

(a) Classification Tasks  
(b) Regression Tasks

Figure 6: Vary number of local patterns

Comments:
- pretty robust when increasing $K_g$
- too large $K_l$ leads to overfitting
# of trees

(a) Classification Tasks   (b) Regression Tasks

Figure 7: Vary number of trees

Comments:
• more trees $\rightarrow$ more diverse patterns
• too many Trees $\rightarrow$ overfitting
Application: ALS

- DREAM-Phil Bowen ALS Prediction Prize4Life Challenge
  - A training set of 918 patients
  - A leaderboard set of 279 patients (not used)
  - A validation set of 627 patients
- More than 1000 participants, 37 unique algorithms
Application: ALS

- Comparable RMSE

- DPmed
  - Team 1: Bayesian trees
  - Team 2: Random forest
  - Team 3: Random forest
  - Team 4: Random forest
  - Team 5: Random forest
  - Team 6: Nonpar. regression

- Baseline: Support vector regr.
- Team 7: Prediction of mean
- Team 8: Linear regression
- Team 9: Multivariate regr.
- Team 10: Linear regression
Figure 9: Overall Important Variables and Patient Cluster Specific Important Variables. Highlighted are the novel markers we discovered.
Outline

- Motivation
- Related Work
- DPMed: Methodology
- Experimental Results
- Discussion and Future Work
Future Work

- More real world clinical data
- how to empirically or theoretically determine the optimal values of parameters?
  - # of diagnosis-stratified patient clusters
  - # of global/local patterns