DPClass: Effective but Concise Discriminative Patterns-Based Classification

Jingbo Shang, Wenzhu Tong, Jian Peng, Jiawei Han
University of Illinois, Urbana-Champaign
May 6, 2016
Presented by Jingbo Shang
Outline

- Motivation: Why Discriminative Patterns based?
- DPClass: Methodology
- Experimental Results
- Discussion and Future Work
Why Discriminative Patterns based?

- Single Feature vs. Combinations of Features
  - A single feature sometimes means nothing
  - Combinations of Features are more meaningful
    - Example: Xor Problem: *not linearly separable* using single features
- Mining semantically meaningful patterns
  - Construct **high-order** interactions in features
  - **Compress** the predictive model
Classification: Why Not Use Tree-based Models?

- Single-Tree Models
  - Example: Decision Tree/Boosted Tree
  - Sensitive to training instances → Overfitting

- Multiple-Tree Models
  - Example: Random Forest
  - Tree-independent: the growth and pruning of different decision trees are independent
  - Model size could be very large → Slow online prediction
  - Uninterpretable
Classification: Why Not Use PatClass/DDPMine?

- Frequent Patterns vs. Discriminative Patterns
  - *Frequent* does not imply *discriminative*
  - The number of frequent patterns might be very large
- PatClass/DDPMine
  - May generate a large but useless pool of frequent patterns
Outline

- Motivation: Why Discriminative Patterns based?
- DPClass: Methodology
- Experimental Results
- Discussion and Future Work
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
Definitions

- **Pattern**
  - A conjunctive clause of conditions on specific feature dimensions
  - i.e., \((x_{i1} < v_1) \land (x_{i2} \geq v_2) \land \ldots \land (x_{im} \geq v_m)\)

- **Discriminative patterns**
  - have **strong signals on the classification task** given by the labels of data
  - E.g., a pattern with very high information gain on the training data
DPClass: Compatible Discriminative Patterns for Linear Models

Training Dataset

Multiple Tree-based Model

Top-k Discriminative Patterns

Testing Dataset

Compressed Model

Efficient Testing

0.8 + 0.5 * b - 1 * g + 2.1 * f - 0.7 * j

A non-leaf node & a discriminate pattern
A selected discriminative pattern
A non-selected discriminative pattern
What Kind of Patterns Are of Top-$k$ Patterns?

- Some effects of different patterns may have a large portion of overlaps, e.g., $(v_0 \cap v_1 \cap v_2)$ and $(v_0 \cap v_1 \cap v_2 \cap v_3)$

- **Top-$k$ patterns** is a size-$k$ subset of discriminative patterns, which have the best performance, which is the accuracy in classification tasks, based on the training data.
DPClass II: Discriminative Patterns Generation

- Random Forest
  - Maximize the randomness
    - Random features
    - Random partitions
    - Random instances (bootstrap)
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-$k$ Patterns Selection

- Select a $k$-set of discriminative patterns, which can achieve the best training accuracy.

- Implementation
  - Forward Selection (Greedy)
  - LASSO (GLMNET)

**Algorithm 3: Top-$k$ Pattern Selection: Forward**

```
Require: $n$ training instances $(x_i, y_i)$, a set of discriminative patterns $\mathcal{P}$ and $k$
Return: Top-$k$ discriminative patterns set $\mathcal{P}_k$ and a generalized linear model $f(\cdot)$
$\mathcal{P}_k \leftarrow \emptyset$
for $t = 1$ to $k$ do
  for each pattern $p$ in $\mathcal{P}$ do
    $x' \leftarrow$ construct pattern space $(x, \mathcal{P}_k \cup \{p\})$
    $g(\cdot) \leftarrow$ a generalized linear model [25] on $(x'_i, y_i)$
    $acc_p \leftarrow g(\cdot)$’s training accuracy
    $\mathcal{P}_k \leftarrow \mathcal{P}_k \cup \{\arg\max_p acc_p\}$
  end for
  $x' \leftarrow$ construct pattern space $(x, \mathcal{P}_k)$
  $f(\cdot) \leftarrow$ a generalized linear model on $(x'_i, y_i)$
return $\mathcal{P}_k, f(\cdot)$
```
DPClass III: Top-$k$ Patterns Selection

- Select a $k$-set of discriminative patterns, which can achieve the best training accuracy.

- Implementation
  - Forward Selection (Greedy)
  - LASSO (GLMNET)

---

**Algorithm 4: Top-$k$ Pattern Selection: LASSO**

*Require:* $n$ training instances $(x_i, y_i)$, a set of discriminative patterns $\mathcal{P}$, $k$, and a small value $\epsilon$  
*Return:* Top-$k$ discriminative patterns $P_i$ and a generalized linear model $f(\cdot)$

1. $\mathcal{P}_k \leftarrow \emptyset$
2. $l \leftarrow 0$, $r \leftarrow +\infty$
3. $x' \leftarrow$ construct pattern space $(x, \mathcal{P})$
4. While $l + \epsilon < r$
   - $\lambda \leftarrow (l + r)/2$
   - $w \leftarrow \arg\min_w \mathcal{L} = \sum_{i} l(x'_i w, y_i) + \lambda \cdot \|w\|_1$
   - If non-zero weighted patterns $\leq k$ then
     - $\mathcal{P}_k \leftarrow \{p | p$'s weight is non-zero}$
     - $r \leftarrow \lambda$
   - Else
     - $l \leftarrow \lambda$
5. $x' \leftarrow$ construct pattern space $(x, \mathcal{P}_k)$
6. $f(\cdot) \leftarrow$ a generalized linear model on $(x'_i, y_i)$
7. Return $\mathcal{P}_k$, $f(\cdot)$
Outline

- Motivation: Why Discriminative Patterns based?
- DPClass: Methodology
- Experimental Results
- Discussion and Future Work
Experiments: Synthetic Experiment

- For each patient, we have several uniformly sampled features as the following
  - Age (A): Positive Integers no more than 60
  - Gender (G): Male or Female
  - Lab Test 1 (LT1): Categorical values from (A, B, O, AB)
  - Lab Test 2 (LT2): Continuous values in [0..1]

- The positive label of the hypo-disease will be given when at least one of the following rules holds
  - (age > 18) and (gender = Male) and (LT1 = AB) and (LT2 ≥ 0.6)
  - (age > 18) and (gender = Female) and (LT1 = O) and (LT2 ≥ 0.5)
  - (age ≤ 18) and (LT2 ≥ 0.9)
Experiments: Synthetic Experiment

- $10^5$ random patients in train (0.1% noise), $5 \times 10^4$ random patients in test
- 99.99% Accuracy
- Discovered Top-3 Patterns:
  - $(\text{age} > 18)$ and $(\text{gender} = \text{Female})$ and $(\text{LT1} = 0)$ and $(\text{LT2} \geq 0.496)$
  - $(\text{age} \leq 18)$ and $(\text{LT2} \geq 0.900)$
  - $(\text{age} > 18)$ and $(\text{gender} = \text{Male})$ and $(\text{LT1} = \text{AB})$ and $(\text{LT2} \geq 0.601)$
Experiments: Real World Datasets

- **RF**: Random Forest
- **DDPMine**: Discriminative Pattern-based Classification
- **DPClass**: Top-20 Discriminative Patterns
  - **DPClass-F**: DPClass using Forward Selection
  - **DPClass-L**: DPClass using Lasso

Table 2: Test Accuracy on UCI Machine Learning Datasets tested in DDPMine. DDPMine outperforms decision tree and support vector machine on all these datasets [2, 3]. RF refers the random forest without any constraints.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>adult</th>
<th>hypo</th>
<th>sick</th>
<th>crx</th>
<th>sonar</th>
<th>chess</th>
<th>namao</th>
<th>musk</th>
<th>madelon</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPClass-F</td>
<td>85.66%</td>
<td>99.58%</td>
<td>98.35%</td>
<td>89.35%</td>
<td>85.29%</td>
<td>92.25%</td>
<td>97.17%</td>
<td>95.92%</td>
<td>74.50%</td>
</tr>
<tr>
<td>DPClass-L</td>
<td>84.33%</td>
<td>99.28%</td>
<td>98.87%</td>
<td>87.96%</td>
<td>83.82%</td>
<td>92.05%</td>
<td>96.94%</td>
<td>95.71%</td>
<td>76.00%</td>
</tr>
<tr>
<td>RF</td>
<td>85.45%</td>
<td>97.22%</td>
<td>94.03%</td>
<td>89.35%</td>
<td>83.82%</td>
<td>94.22%</td>
<td>97.86%</td>
<td>96.60%</td>
<td>56.50%</td>
</tr>
<tr>
<td>DDPMine</td>
<td>83.42%</td>
<td>92.69%</td>
<td>93.82%</td>
<td>87.96%</td>
<td>73.53%</td>
<td>90.04%</td>
<td>96.83%</td>
<td>93.29%</td>
<td>59.83%</td>
</tr>
</tbody>
</table>
Experiments: Train/Test Accuracy vs. top-K

- Training and testing accuracies are almost overlapped.
- A small number (e.g., 20) of discriminative patterns are good enough.
Experiments: Train/Test Accuracy vs. #trees

- Training and testing accuracies are almost overlapped
- A small number (e.g., 20) of random decision trees are good enough
Outline

- Motivation: Why Discriminative Patterns based?
- DPClass: Methodology
- Experimental Results
- Discussion and Future Work
Conclusions

- DPClass can compress the model and thus the online prediction is extremely fast
- DPClass have comparable performance as before
  - Even better in experiments
- DPClass can learn the interpretable patterns
  - Shown in the synthetic experiment
Future Work

- Extend DPClass to DPLearn
- Task Oriented Discriminate Patterns Learning
  - Classification
  - Multi-class classification
  - Regression
  - Survival Analysis