Informative Projection Recovery for Classification, Clustering and Regression

MADALINA FITERAU
ARTUR DUBRAWSKI

ICMLA 2013
Motivation

1. NEED COMPACT MODELS TO ENABLE ANALYSIS AND VISUALIZATION
2. LEVERAGING EXISTING STRUCTURE IN DATA → HIGH PERFORMANCE
3. COMPACT ENSEMBLES OF COMPLEMENTARY LOW-D SOLVERS

BORDER CONTROL

DIAGNOSTICS

VEHICLE CHECKS
Presentation Roadmap

- Informative Projection Retrieval
- RIPR* Framework Overview
  * Regression-based Informative Projection Retrieval
- The Optimization Procedure
- Applicability to Learning Tasks
- Performance Evaluation
- Medical Application Case Study
Projection Retrieval for a Learning Task

- problem of selecting low-d (2D, 3D) subspaces
- s.t. queries are resolved with high-confidence
- models perform the task with low expected risk

example: features represent vital signs and derived features; considering only the duty cycles of the signals might be sufficient

RIPR = Regression-based Informative Projection Retrieval*

* A generalization of our prior work in “Projection Retrieval for Classification”, NIPS 2012
RIPR Target Datasets

- Most of the features are redundant (non-informative)
- There exists one or several sets of features with structure
- The ‘tidy’ part of the set may span only part of the points
- Jointly, the sets of projections handle all data

Clinical Data - several sub-models, corresponding to underlying conditions and patient characteristics

Human-engineered datasets - corrupted with artifacts which can be identified as low-dimensional patterns

Aspect of most projections

IP for blue/red group

IP for light blue/purple group
A Dual-Objective Training Process

1. Data is split across informative projections

2. Each projection has a solver trained using only the data assigned to that projection
RIPR Framework

\[ g(X) \to \pi_1(X) \to \tau_1(\pi_1(X)) \]
\[ g(X) \to \pi_2(X) \to \tau_2(\pi_2(X)) \]
\[ g(X) \to \pi_3(X) \to \tau_3(\pi_3(X)) \]

CONTEXT SOLVERS PROJECTIONS SELECTOR QUERY
RIPR Model

Model components:
- Set of $d$-dimensional, axis-aligned sub-spaces of the original feature space $P \in \Pi$
- Each projection has an assigned solver of the task $T$; the solvers are selected from some solver class $\mathcal{T}$
- A selection function $g$, which yields, for a query point $x$, the projection/solver pair $(\pi_g(x), \tau_g(x))$ for the point;
- $\ell(\tau_g(x)(\pi_g(x)), y)$ represents the model loss at point $x$
RIPR Objective Function

Model components:
• Set of $d$-dimensional, axis-aligned sub-spaces of the original feature space $P \in \Pi$
• Each projection has an assigned solver of the task $T$; the solvers are selected from some solver class $\mathcal{T}$
• A selection function $g$, which yields, for a query point $x$, the projection/solver pair $(\pi_{g(x)}, \tau_{g(x)})$ for the point;
• $\ell(\tau_{g(x)}(\pi_{g(x)}), y)$ represents the model loss at point $x$

Minimization:

$$M^* = \underset{M \in \mathcal{M}_d}{\text{argmin}} \mathbb{E}_X \ell(\tau_{g(x)}(\pi_{g(x)}), y)$$

Expected loss for task solver trained on projection assigned to point
Presentation Roadmap

• Informative Projection Retrieval

• RIPR Framework Overview
  o The Optimization Procedure
  o Applicability to Learning Tasks
  o Performance Evaluation
  o Medical Application Case Study
Starting point: the loss matrix

<table>
<thead>
<tr>
<th>Loss estimators</th>
<th>Projections</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Samples</td>
<td></td>
</tr>
<tr>
<td>HIGH LOSS</td>
<td>LOW LOSS</td>
</tr>
</tbody>
</table>

- low loss
- moderate loss
- high loss
Starting point: the loss matrix

Loss estimators

Projections

Samples

- low loss
- moderate loss
- high loss
The Optimization Procedure

We introduce a penalty over # of columns to limit the # of projections in the model.
The Optimization Procedure

Suboptimal projections will be used for some of the points.

Matrix of Loss Estimators (L)

- optimal
- nearly optimal
- Suboptimal projections will be used for some of the points
Regression for Informative Projection Recovery (RIPR)

RIPR learns a binary selection matrix $B$ in a manner resembling the adaptive lasso.

- Iterative procedure
  - Initialize selection matrix $B$
  - Compute multiplier $\delta$ inversely proportional with projection popularity
  - Use penalty $|B\delta|_1 \rightarrow \text{new } B$
The RIPR Algorithm

1. Compute loss matrix $L$, target $T$

2. Estimate selection matrix $B$
   
   $min_B \| T - L \otimes BJ_{|\Pi|,1} \|_2^2 + \lambda \sum_{k=1}^{|\Pi|} |B_k|_1$

3. Compute multiplier $\delta$ inversely proportional with utility
   
   $\delta_k = |B_k|_1$, \hspace{1cm} $\delta = 1 - \delta/|\delta|_1$

4. Obtain new selection matrix $B$ penalizing $B\delta$
   
   $min_B \| T - L \otimes BJ_{|\Pi|,1} \|_2^2 + \lambda |B\delta|_1$

   where $L_{ij} \otimes B_{ij} = L_{ij}B_{ij}$

ITERATE UNTIL CONVERGENCE
Applicability to Learning Tasks

We show how RIPR can solve the following tasks:

- Classification
- Semi-supervised classification
- Clustering
- Regression

The matrix of loss estimators is computed differently for each of these tasks.

The generality of the method does not stop here: RIPR can solve any learning task for which the risk can be decomposed using consistent loss estimators.
Neighbor-based estimator for conditional entropy*:

\[
\widehat{H}(Y|X \in \mathcal{A}) \propto \frac{1}{n} \sum_{i=1}^{n} I[x_i \in \mathcal{A}] \left( \frac{n - 1}{n} \left( \frac{\text{dist}_{k+1}(x_i, X_{y_i})}{\text{dist}_k(x_i, X_{-y_i})} \right)^{\text{dim}(X)} \right)^{1-\alpha}
\]

For a projection \( \pi \), the estimator is \( \widehat{H}(Y|\pi(X); g(X) \rightarrow \pi) \).

The optimal model can be computed through the minimization:

\[
\widehat{M} = \arg \min_{M \in M_d} \sum_{\pi_j \in \Pi} \sum_{i=1}^{n} I[g(x_i) \rightarrow \pi_j] \left( \frac{\text{dist}_{k+1}(\pi_j(x_i), \pi_j(X_{y_i}))}{\text{dist}_k(\pi_j(x_i), \pi_j(X_{-y_i}))} \right)^{\text{dim}(\pi_j)(1-\alpha)}
\]

\(B_{ij}\) - selection matrix

\(L_{ij}\) - local entropy contributions

\[
T_i = \min_j L_{ij}
\]

Based on the divergence estimator by Poczos and Schneider, “On the estimation of alpha-divergences” (AISTATS 2011)
Loss Estimators: Semi-supervised Classification

- For labeled samples: same as for classification

- For unlabeled samples:
  - Consider all possible label assignments
  - Assume the most ‘confident’ label (with smallest loss)

*Equivalent to*

- Penalizing unlabeled samples proportional to how ambivalent they are to the label assigned
Loss Estimators: Semi-supervised Classification

- For labeled samples: same as for classification
- For unlabeled samples:
  - Consider all possible label assignments
  - Assume the most ‘confident’ label (with smallest loss)
  
  **Equivalent to**

  - Penalizing unlabeled samples proportional to how ambivalent they are to the label assigned

\[
R_{ssc} \left( X \in \mathcal{A}(\pi_j) \right) = \sum_{x_i \in \text{labeled}} \left( \frac{\text{dist}_{k+1}(\pi_j(x_i), \pi_j(X_{y_i}))}{\text{dist}_k(\pi_j(x_i), \pi_j(X_{\neg y_i}))} \right)^{\dim(\pi_j)(1-\alpha)} + \\
\sum_{x_i \in \text{unlabeled}} \min_{\gamma \in y} \left( \frac{\text{dist}_{k+1}(\pi_j(x_i), \pi_j(X_{\gamma}))}{\text{dist}_k(\pi_j(x_i), \pi_j(X_{\neg \gamma}))} \right)^{\dim(\pi_j)(1-\alpha)}
\]
Entropy Estimators for Clustering

- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering
Entropy Estimators for Clustering

- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering
- We focus on density-based clustering
- The loss is lower for densely packed regions
- We eliminate dimensionality issues by considering negative KL divergence to uniform on the same space

POOR

GOOD
Entropy Estimators for Clustering

0 Point-wise estimators are problematic for clustering
0 An ensemble view of the data is typically required
0 It is unknown which data should be assigned to which projection prior to clustering
0 We focus on density-based clustering
0 The loss is lower for densely packed regions
0 We eliminate dimensionality issues by considering negative KL divergence to uniform on the same space*

\[ R_{clustering} \left( X \in \mathcal{A}(\pi_j) \right) = \rightarrow -KL(\pi_j(X), \pi_j(\text{Unif})) \]

\[ \ell(\tau_j(\pi_j(x))) = \left( \frac{\text{dist}(\pi_j(x), \pi_j(X))}{\text{dist}(\pi_j(x), \pi_j(U))} \right)^{\text{dim}(\pi_j)(1-\alpha)} \]

* some scaling issues remain
Low-d Clustering: Why it Works

K-Means model projected on (known) informative features

The hidden structure in data is clearly revealed by the RIPR model.
Low-d Clustering: Why it Works

K-Means model projected on (known) informative features

The hidden structure in data is clearly revealed by the RIPR model.
## Loss/Risk for common Learning Tasks

<table>
<thead>
<tr>
<th>Learning Task</th>
<th>Loss/Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification*</td>
<td>Classification error approximated by conditional entropy</td>
</tr>
<tr>
<td></td>
<td>[ R_{cls}(X) = \mathbb{E}_X [y \neq h_g(x)(\pi_g(x)(x)) ] \approx H(y</td>
</tr>
<tr>
<td>Semi-supervised classification</td>
<td>Conditional entropy for labeled samples plus best case entropy over label assignments for unlabeled samples</td>
</tr>
<tr>
<td></td>
<td>[ R_{ssc}(X) = R_{cls}(X) + \min_{\gamma \in Y} H_x^{unlabeled}(\gamma</td>
</tr>
<tr>
<td>Clustering</td>
<td>Negative divergence between distribution of data and a uniform distribution on the same sample space</td>
</tr>
<tr>
<td></td>
<td>[ R_{clustering} = -KL(\pi_g(x)(x)</td>
</tr>
<tr>
<td>Regression</td>
<td>Mean squared error</td>
</tr>
<tr>
<td></td>
<td>[ R_{reg}(X) = \mathbb{E}_X [(y - h_g(x)(\pi_g(x)(x)))^2] ]</td>
</tr>
</tbody>
</table>

* The object of prior work: “Projection Retrieval for Classification”, NIPS 2012
Assigning a Projection to a Query

Problem: how to select the appropriate projection for a specific query $q$?

Solution: select the projection in $P$ for which the estimated loss* at $q$ is smallest.

\[
(k, \hat{y}) = \arg\min_{(k,y)} \hat{\ell}(\tau_k(\pi_k), y)
\]
where $k \in \{1 \ldots |P|\}$

*For clustering, the loss estimator is computed considering the cluster assignments determined during learning.
Presentation Roadmap

• Informative Projection Retrieval
• RIPR Framework Overview
• The Optimization Procedure
• Applicability to Learning Tasks
  o Performance Evaluation
  o Medical Application Case Study
Semi-supervised classification - artificial data -

Dataset contains 3 informative projections, 3000 labeled points.

RIPR correctly recovers the projections for all settings tested. Leveraging this structure, RIPR achieves higher accuracy.
Clustering
- evaluation metrics -

DISTORTION – mean distance to cluster centers

LOG CLUSTER VOLUME

K-means Model

Ripped K-means Model
Clustering  
- artificial data -

<table>
<thead>
<tr>
<th>Settings</th>
<th>Distortion</th>
<th>Log Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RIPR</td>
<td>Kmeans</td>
</tr>
<tr>
<td>Q = 2</td>
<td>K = 2</td>
<td>865</td>
</tr>
<tr>
<td>Q = 2</td>
<td>K = 3</td>
<td>622</td>
</tr>
<tr>
<td>Q = 2</td>
<td>K = 5</td>
<td>440</td>
</tr>
<tr>
<td>Q = 2</td>
<td>K = 7</td>
<td>375</td>
</tr>
<tr>
<td>Q = 3</td>
<td>K = 2</td>
<td>1,344</td>
</tr>
<tr>
<td>Q = 3</td>
<td>K = 3</td>
<td>872</td>
</tr>
<tr>
<td>Q = 3</td>
<td>K = 5</td>
<td>648</td>
</tr>
<tr>
<td>Q = 3</td>
<td>K = 7</td>
<td>530</td>
</tr>
<tr>
<td>Q = 5</td>
<td>K = 2</td>
<td>2,683</td>
</tr>
<tr>
<td>Q = 5</td>
<td>K = 3</td>
<td>1,484</td>
</tr>
<tr>
<td>Q = 5</td>
<td>K = 5</td>
<td>1,065</td>
</tr>
<tr>
<td>Q = 5</td>
<td>K = 7</td>
<td>842</td>
</tr>
<tr>
<td>Q = 7</td>
<td>K = 2</td>
<td>4,621</td>
</tr>
<tr>
<td>Q = 7</td>
<td>K = 3</td>
<td>2,174</td>
</tr>
<tr>
<td>Q = 7</td>
<td>K = 5</td>
<td>1,480</td>
</tr>
<tr>
<td>Q = 7</td>
<td>K = 7</td>
<td>1,238</td>
</tr>
</tbody>
</table>

Q = NUMBER OF INFORMATIVE PROJECTIONS
K = NUMBER OF CLUSTERS ON EACH PROJECTION
PERCENTAGE REDUCTION IN SUM OF CLUSTER VOLUME

RIPR MODELS ARE MORE COMPACT

NOTE: THE K-MEANS AND RIPR MODELS HAVE THE NUMBER OF CLUSTERS.

COMPRESSION IS REDUCED AS MORE CLUSTERS/PROJECTIONS ARE ADDED
## Clustering - UCI data -

SUM OF MEAN DISTANCES TO CLUSTER CENTERS AND LOG CLUSTER VOLUME

<table>
<thead>
<tr>
<th>UCI Dataset</th>
<th>Mean Distortion</th>
<th>% Distortion Reduction</th>
<th>Log Volume of Clusters on All Dimensions</th>
<th>% Volume Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeds</td>
<td>RIPR 16</td>
<td>Kmeans 107</td>
<td>90.73</td>
<td>RIPR 3.33</td>
</tr>
<tr>
<td>Libras</td>
<td>RIPR 9</td>
<td>Kmeans 265</td>
<td>98.54</td>
<td>RIPR -2.52</td>
</tr>
<tr>
<td>MiniBOONE</td>
<td>RIPR 125</td>
<td>Kmeans 1,154,704</td>
<td>99.99</td>
<td>RIPR 104.23</td>
</tr>
<tr>
<td>Cell</td>
<td>RIPR 40,877</td>
<td>Kmeans 8,181,327</td>
<td>99.78</td>
<td>RIPR 23.75</td>
</tr>
<tr>
<td>Concrete</td>
<td>RIPR 1,370</td>
<td>Kmeans 55,594</td>
<td>98.01</td>
<td>RIPR 21.39</td>
</tr>
</tbody>
</table>

LOWER IS BETTER. RIPR MODELS ALWAYS HAVE A SMALLER TOTAL VOLUME.
Regression
- artificial data -

ACCURACY OF RIPPED SVM COMPARED TO ACCURACY OF STANDARD SVM
- THE NUMBER OF INFORMATIVE PROJECTIONS: 2-10
- PERCENTAGE OF NOISY SAMPLES: 0-50% (OUT OF 1600)

<table>
<thead>
<tr>
<th>NOISY SAMPLES</th>
<th>MSE RIPPED-SVM</th>
<th>MSE SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.05 0.27 0.05 0.02 0.23</td>
<td>0.27 1.16 0.11 0.1 0.43</td>
</tr>
<tr>
<td>6.25%</td>
<td>0.42 1.26 0.34 1.45 0.52</td>
<td>0.8 1.02 0.6 2.99 0.94</td>
</tr>
<tr>
<td>12.5%</td>
<td>0.5 0.86 0.8 0.33 0.99</td>
<td>0.97 1.27 0.29 0.68 1.44</td>
</tr>
<tr>
<td>25%</td>
<td>0.63 1.47 1.34 1.61 0.11</td>
<td>0.4 1.26 1.64 1.71 0.08</td>
</tr>
<tr>
<td>50%</td>
<td>0.69 0.38 1.12 0.68 1.1</td>
<td>0.52 0.06 0.91 0.9 1.16</td>
</tr>
</tbody>
</table>

PRECISION AND RECALL OF THE RECOVERED PROJECTIONS

<table>
<thead>
<tr>
<th>NOISY SAMPLES</th>
<th>RIPR Precision</th>
<th>RIPR Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1 1 0.4 0.43 0.3</td>
<td>0.67 1 0.67 1 1</td>
</tr>
<tr>
<td>6.25%</td>
<td>1 0.67 0.6 0.43 0.2</td>
<td>0.67 0.67 1 1 0.67</td>
</tr>
<tr>
<td>12.5%</td>
<td>1 1 0.6 0.43 0.3</td>
<td>0.67 1 1 1 1</td>
</tr>
<tr>
<td>25%</td>
<td>1 1 0.6 0.43 0.1</td>
<td>0.67 1 1 1 0.33</td>
</tr>
<tr>
<td>50%</td>
<td>1 0.67 0.4 0.29 0.3</td>
<td>0.67 0.67 0.67 0.67 1</td>
</tr>
</tbody>
</table>
Case Study – Alert Classification
- importance of artifact adjudication -

- Intensive Care Unit vital sign monitoring system
- Alerts are raised when patient health status deteriorates
- One alert is issued every 90s

- A significant amount of alerts are artifacts
- Frequent alerts cause alarm fatigue in medical staff
- Quality of care diminished unless artifacts are identified
Case Study – Alert Classification - vital sign data processing -

- Each alert is associated with the first abnormal vital sign:
  - Heart Rate (HR), Respiratory Rate (RR)
  - Systolic (SBP) and Diastolic (DBP) Blood Pressure
  - Peripheral arterial oxygen saturation (SpO2)

- 812 of the samples were labeled by clinicians (~10%)

- Extracted temporal features and derived metrics:
  - Vitals collected during the alert event
  - Data starting 4 minutes before alert onset
  - Moving window statistics
  - Metrics such as duty cycle
  - Data collected for each vital independently
## Case Study – Alert Classification - performance -

<table>
<thead>
<tr>
<th>Alarm Type</th>
<th>RR</th>
<th>BP</th>
<th>SPO&lt;sub&gt;2&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2D</td>
<td>2D</td>
<td>3D</td>
</tr>
<tr>
<td></td>
<td>2D</td>
<td>3D</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.98</td>
<td>0.833</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td>0.911</td>
<td>0.9151</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.979</td>
<td>0.858</td>
<td>0.896</td>
</tr>
<tr>
<td></td>
<td>0.929</td>
<td>0.9176</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.991</td>
<td>0.93</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>0.945</td>
<td>0.9957</td>
<td></td>
</tr>
</tbody>
</table>
Case Study – Alert Classification
- RIPR model for blood pressure -

**RIPR identifies interpretable projections which adjudicate alerts.**

<table>
<thead>
<tr>
<th>Alarm Type</th>
<th>RR</th>
<th>BP</th>
<th>SPO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2D</td>
<td>2D</td>
<td>3D</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.98</td>
<td>0.833</td>
<td>0.885</td>
</tr>
<tr>
<td>Precision</td>
<td>0.979</td>
<td>0.858</td>
<td>0.896</td>
</tr>
<tr>
<td>Recall</td>
<td>0.991</td>
<td>0.93</td>
<td>0.958</td>
</tr>
</tbody>
</table>

*Accuracy, Precision, Recall are calculated against true alerts.

**SPO₂-duty-cycle**
- 46% of validation data
- *duty cycle = number of readings over time units: a low value indicates high sparseness*

**RR-diff1-max**
- 54% of validation data

**RR-diff1-min**
- true alert
- artifact
Case Study – Alert Classification - utility of RIPR models -

- The model selects HR duty cycle as the most important dimension in RR artifact classification, validating expect intuition
- Uncommon RR artifacts are classified as true alerts
- The RR signals are irregular
- Such cases can be identified through using variance of signal (new features added)
- RIPR model pointed out some mislabeled alerts
Case Study – Alert Classification - deriving rules -

<table>
<thead>
<tr>
<th>Alarm Type</th>
<th>RR</th>
<th>BP</th>
<th>SPO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2D</td>
<td>2D</td>
<td>3D</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.98</td>
<td>0.833</td>
<td>0.885</td>
</tr>
<tr>
<td>Precision</td>
<td>0.979</td>
<td>0.858</td>
<td>0.896</td>
</tr>
<tr>
<td>Recall</td>
<td>0.991</td>
<td>0.93</td>
<td>0.958</td>
</tr>
</tbody>
</table>

*data density = number of readings over time units: a low value indicates high sparseness*
Case Study – Alert Classification - deriving rules -

\[
\text{RR-duty-cycle}^* \leq 0.6 \quad \text{and} \quad \text{HR-duty-cycle} \leq 0.25
\]

\[
\text{HR-data-density} - \text{SPO}_2\text{-data-density} \leq 0.2
\]

\[
\text{HR-data-density}/0.3 + \text{RR-min}/5 \leq 1
\]
Summary

- Informative Projection Retrieval is relevant to many applications requiring interaction with human users.
- We generalized RIPR, our solution to the IPR problem, to a wide range of learning tasks (classification, regression, clustering).
- RIPR expresses loss through divergence estimators:
  - Semi-supervised models: penalize unlabeled data that cannot be confidently assigned to a class.
  - Clustering models: favor high data density.
- RIPR models are compact and well-performing in practice:
  - IPs accurately recovered.
  - Often more accurate than classifiers trained on all features.
- Overall, RIPR contributes to the improvement of the quality of care for ICU.