Ensembles for the Discovery of Compact Structures in Data

Madalina Fiterau
Carnegie Mellon University
School Of Computer Science
Machine Learning Department

Gatsby Unit, UCL
9th of March 2015
The Big Data Paradox

Introduction
The Big Data Paradox

- Heterogeneous
- Highly sparse
- Unlabeled
- Non-standard
- Multi-source
- Noisy
The Big Data Paradox

Introduction

Compact Patterns
Talk Outline

Informative Projection Recovery (IPR)
- Projection Retrieval as a combinatorial problem
- Optimization procedure for IPR
- RIPR for classification, clustering, regression, active learning

Applications to Data Diagnostics
- Pattern Discovery in Clinical Data
- Finding Gaps in Training Data for Radiation Threat Detection

Back-propagation Forests
- Learning Fuzzy Decision Trees Using Back-propagation
- Potential Extensions of BP Forests
Informative Projection Recovery
Considerable effort expended on building *complex models* from *vast* amounts of data, not enough to make models *comprehensible*.

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**
Sparse Predictive Structures

Heterogeneous data

Informative Projection Recovery
Sparse Predictive Structures

Learning Global Models

Issue: different features are relevant in different parts of the input space.
Learning Local Models

Issue: insufficient training data in the neighborhood or the sample.
Compact Partitioning Models
Compact Partitioning Models

Split on Y

Split on X

Split on X, Y

Informative Projection Recovery
- Select low-d subspaces which allow confident classification

- Clinical data example: vital signs and derived features

\[ \text{RIPR} = \text{Regression-based Informative Projection Retrieval} \][1]

Dataset Assumptions

- Only a small subset of the projections have useful structure
- Projections are complementary, dealing with different samples

---

- Engineered data - unintentionally introduced artifacts usually show in low-dimensional patterns
- Clinical data - multiple sub-models reflect specifics of particular conditions and patient characteristics
RIPR Framework

1. Query: $X$
2. Selector: $g(X)$
3. Projections: $\pi_1(X)$, $\pi_2(X)$, $\pi_3(X)$
4. Solvers: $\tau_1(\pi_1(X))$, $\tau_2(\pi_2(X))$, $\tau_3(\pi_3(X))$
5. Context:

---

Informative Projection Recovery
Dual-Objective Training Process

1. Data is split across informative projections
2. Train solvers using data assigned to each projection
RIPR Model

Model Components

- **P** - set of axis-aligned sub-spaces, max. d features
- **T** - set of solvers trained on each of the projections in **P**
- **g** – determines the projection/solver for a point \( x \), \( (\pi_g(x), \tau_g(x)) \)
- \( \ell(\tau_g(x)(\pi_g(x)(x)), y) \) represents the model loss at point \( x \)

Dataset \( X = \{x_1 \ldots x_n\} \in \mathcal{X}^n \), where \( x_i \in \mathcal{X} \subseteq \mathbb{R}^m \)

\[
\mathcal{M}_d = \{ \Pi = \{ \pi; \pi \in \Pi, |\pi| \leq d \}, T = \{ \tau; \tau_i \in \mathcal{T}, \tau_i : \pi_i(\mathcal{X}) \rightarrow \mathcal{Y} \ \forall i = 1 \ldots |\Pi| \}, g \in \{ f : \mathcal{X} \rightarrow \{1 \ldots |\Pi|\} \} \}.
\]

Small set of projections

Target model

Selection function

Solvers
RIPR Objective Function

Model Components

- P - set of axis-aligned sub-spaces, max. d features
- T - set of solvers trained on each of the projections in P
- g – determines the projection/solver for a point x, \( (\pi_g(x), \tau_g(x)) \)
- \( \ell(\tau_g(x)(\pi_g(x)(x)), y) \) represents the model loss at point x

\[
M^* = \arg \min_{M \in \mathcal{M}_d} \mathbb{E}_{x, y} \left[ y \neq h_g(x)(\pi_g(x)(x)) \right]
\]

Expected loss for task solver trained on projection assigned to point
Starting point: the loss matrix

Loss estimators

Projections

Samples

HIGH LOSS

LOW LOSS

- low loss
- moderate loss
- high loss

Informative Projection Recovery
Starting point: the loss matrix

Loss estimators

Projections

Samples

- low loss
- moderate loss
- high loss

Informative Projection Recovery
Matrix of Loss Estimators (L)

Data Points

Projections

1
2
3
4
5
6
7

optimal

nearly optimal

Penalty – limits

# of projections
The Optimization Procedure

Matrix of Loss Estimators (L)

Data Points

Projections

1 2 3 4 5 6 7

optimal

nearly optimal

some points use suboptimal projections
The Optimization Procedure

Matrix of Loss Estimators (L)

- $L_{ij}$ is the loss for sample $i$ at projection $j$
- For each point $i$, let $T_i$ be the lowest loss over the projections $T_i = \min L_{ij}$
- $B$ binary selection matrix
- $B_{ij}$ is 1 if projection $j$ is to be used to solve point $i$ and 0 otherwise

Target Loss (T)

- Convex program: $B = \min_B \|T - L \odot B\|_1 + \text{reg } (B)$

where $L \odot B \xrightarrow{\text{def}} \sum_{j=1}^{m} L_{.,j} B_{.,j}$

Informative Projection Recovery
The Optimization Procedure

Matrix of Loss Estimators (L)

Projections

Data Points

Target Loss (T)

- Convex program: \( B = \min_B \| T - L \odot B \|_1 + \text{reg} (B) \)

where \( L \odot B \overset{\text{def}}{=} \sum_{j=1}^{m} L_{.,j} B_{.,j} \)

IPR problem - solved through this regression
- RIPR learns the binary selection matrix B in a manner resembling the adaptive lasso

- **Iterative procedure**
  - Initialize selection matrix B
  - Compute multiplier $\delta$ inv. prop. with projection popularity
  - Use penalty $|B\delta|_1 \rightarrow$ new B
RIPR can solve the following tasks\cite{2}:

- (Semi-supervised) classification
- Clustering
- Regression
- Active learning

Loss matrix computed differently for each task

*RIPR can solve any learning task for which the risk can be decomposed using consistent loss estimators.*

- Neighbor-based estimator for conditional entropy*

\[ H(Y | \pi_j(X); g(X) = j) \]

- For unlabeled samples, assume label with lowest loss

*Based on the divergence estimator by Poczos and Schneider, “On the estimation of alpha-divergences” (AISTATS 2011)
## Classification Results

### Comparison of Classification Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Features</th>
<th># Instances</th>
<th>K-NN</th>
<th>RIPPED K-NN</th>
<th># RIPR projections</th>
<th>#features in projection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Tissue</td>
<td>10</td>
<td>106</td>
<td>1.000</td>
<td>1.000</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Cell</td>
<td>6</td>
<td>200</td>
<td>0.707</td>
<td>0.7640</td>
<td>4</td>
<td>{1,2,2,2}</td>
</tr>
<tr>
<td>Mini BOONE</td>
<td>50</td>
<td>130065</td>
<td>0.790</td>
<td>0.740</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Nuclear Threat</td>
<td>50</td>
<td>200</td>
<td>0.7788</td>
<td>0.7807</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>SPAM</td>
<td>57</td>
<td>4601</td>
<td>0.7680</td>
<td>0.7680</td>
<td>5</td>
<td>{1,2,3,3,3}</td>
</tr>
<tr>
<td>Vowel</td>
<td>10</td>
<td>528</td>
<td>0.984</td>
<td>0.984</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>
Loss Estimators: Clustering

- Density-based clustering
- Loss is lower for high density areas
- Negative KL divergence to uniform

Informative Projection Recovery
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.
Clustering Evaluation Metrics

DISTORTION (goodness-of-fit)

LOG CLUSTER VOLUME (compactness)

K-means Model

Ripped K-means Model
Clustering on Artificial Data

PERCENTAGE REDUCTION IN SUM OF CLUSTER LOG VOLUMES

Q = NUMBER OF INFORMATIVE PROJECTIONS
K = NUMBER OF CLUSTERS ON EACH PROJECTION

COMPRESSION IS REDUCED AS MORE CLUSTERS/PROJECTIONS ARE ADDED

NOTE: THE K-MEANS AND RIPR MODELS HAVE THE NUMBER OF CLUSTERS.
# Clustering on 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></td>
<td>RIPR</td>
<td>Kmeans</td>
<td>RIPR</td>
<td>Kmeans</td>
</tr>
<tr>
<td>Seeds</td>
<td>16</td>
<td>107</td>
<td>90.73</td>
<td>3.33</td>
</tr>
<tr>
<td>Libras</td>
<td>9</td>
<td>265</td>
<td>98.54</td>
<td>-2.52</td>
</tr>
<tr>
<td>MiniBOONE</td>
<td>125</td>
<td>1,154,704</td>
<td>99.99</td>
<td>104.23</td>
</tr>
<tr>
<td>Cell</td>
<td>40,877</td>
<td>8,181,327</td>
<td>99.78</td>
<td>23.75</td>
</tr>
<tr>
<td>Concrete</td>
<td>1,370</td>
<td>55,594</td>
<td>98.01</td>
<td>21.39</td>
</tr>
</tbody>
</table>

LOWER IS BETTER. RIPR MODELS ALWAYS HAVE A SMALLER TOTAL VOLUME.
Loss Estimators: Regression

- Estimates error in point neighborhood

\[
\hat{\ell}_{reg}(\pi_i(x), \tau_i(\pi_i(x))) = (\hat{\tau}(\pi_i(x)) - y)^2 \quad \hat{\ell}_{reg} \to 0
\]

\[
\hat{\tau}_i(\pi_i(x)) = \frac{\sum_{i=1}^{k} w(i) y(i)}{\sum_{i=1}^{k} w(i)}, \quad \text{where } w(i) = \frac{1}{||x - x(i)||_2}
\]
Regression on Artificial Data

ACCURACY OF RIPPED SVM COMPARED TO ACCURACY OF STANDARD SVM
- THE NUMBER OF INFORMATIVE PROJECTIONS: 2-10 (OUT OF 45)
- PERCENTAGE OF NOISY SAMPLES: 0-50% (OUT OF 1600)

<table>
<thead>
<tr>
<th>NOISY SAMPLES</th>
<th>IP #</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>0.19</td>
<td>0.23</td>
<td>0.45</td>
<td>0.20</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>6.25%</td>
<td>0.53</td>
<td>1.24</td>
<td>0.57</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>12.5%</td>
<td>0.52</td>
<td>0.68</td>
<td>2.76</td>
<td>0.49</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>1.58</td>
<td>1.17</td>
<td>0.82</td>
<td>0.94</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>1.33</td>
<td>6.33</td>
<td>1.23</td>
<td>0.76</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Problem: how to select the appropriate projection for a specific query $x$?

Solution: select the projection in the learned subset $P$ for which the estimated loss is the lowest.

$$(k^*, y^*) = \arg\min_{(k \in \{1 \ldots |P|\}, y \in \mathcal{Y})} \ell(\tau_k(\pi_k(x), y))$$
Active Learning with RIPR

- Scoring functions
  - Uncertainty sampling
  - Query by committee
  - Information gain (best performance)
  - Low conditional entropy
- Clinical application: framework requests *half of the labels* requested by random forests with active learning

active exploration: only relevant features $\rightarrow$ fast rates!
Talk Outline

Informative Projection Recovery (IPR)
- Projection Retrieval as a combinatorial problem
- Optimization procedure for IPR
- RIPR for classification, clustering, regression, active learning

Applications to Data Diagnostics
- Pattern Discovery in Clinical Data
- Finding Gaps in Training Data for Radiation Threat Detection

Back-propagation Forests
- Learning Fuzzy Decision Trees Using Back-propagation
- Potential Extensions of BP Forests
Applications to Data Diagnostics

Healthcare Alert Prediction
Nuclear Threat Detection
Learning from Multiple Datasets
Artifacts in Clinical Alerts

- **Hear Rate**: <40 or >140
- **Respiratory Rate**: <8 or >36
- **Systolic Blood Pressure**: <80 or >200
- **Diastolic Blood Pressure**: >110
- **SPO₂**: <85%

Features computed from time series include common statistics of each VS: mean, stdev, min, max, range of values, duty cycle ...

Health alerts are some are artifacts, not true alerts
Artifacts in Clinical Alerts

The retrieved projections enable domain experts to quickly validate alert labels.

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

54% of validation data

46% of validation data
Informative Projections for SPO$_2$ alerts allow derivation of rules.

RR duty cycle $\leq 0.6$

and

HR duty cycle $\leq 0.3$

HR duty cycle $-$ SPO$_2$ duty cycle $\leq 0.2$

HR duty cycle/0.3 + RR-min/5 $\leq 1$
We applied the active learning procedure to artifact annotation.

**Annotation without Informative Projections**

**Annotation with Informative Projections**

SPO$_2$ alerts: RIPR achieves max. accuracy with 25% of the data. BP alerts: RIPR achieves max. accuracy with 50% of the data.
- Vehicles scanned at US border
- Radiation measurements
- Classify threat posed by vehicle
- Threats are rare in practice
- Training ‘threats’ are simulated
- We trained 2-D classification models
Identifying Gaps in Datasets

Joint work with Nick Gisolfi (ngisolfi@andrew.cmu.edu)

Training data is incomplete

RIPR can express training data gaps in terms of low-dimensional projections.

Additional samples requested

- Training data
- Test data
- Identified gap

Applications to Data Diagnostics
DIRECT GAP-FINDING: finding mismatches between distributions of training and testing data.

Gaps found in nuclear threat data.
DIAGNOSTIC GAP-FINDING: finding areas of the test data where the classifier behaves poorly.

Accuracy improves from 75% to 75.7% by filling the gap compared to 75.2% by randomly adding data.
Conclusions
Projects in Chronological Order

- Using Dynamic Bayes Nets for Online Vital Sign Monitoring
- Using MRFs to obtain Elevation Map of Lunar Surface from LRO LIDAR and LCROSS imagery
- Explanation-Oriented Classification via Subspace Partitioning
  ✓ Regression for Informative Projection Recovery
  ✓ Detecting Artifacts in Clinical Data via Projection Retrieval
- Feature Task Bi-clustering in Multitask Regression
- Sparsistent Additive Modeling in Multi-task Learning
  ✓ Finding Gaps in Training Data to Guide Development of a Radiation Threat Detection System
  ✓ Interpretable Active Learning in Support of Data Annotation
- Improving prediction Across Related Datasets
Informative Projections

Visualization

Compact models

Identifying Artifact Clusters

Facilitating Annotation

Customizable

Decision support

Projection 3

Nuclear Threat Classification

Artifact Detection

Artifact

ture alert

artifact

RR Alert

RR Artifact Cluster 1

RR Artifact Cluster 2

0 20 40 60 80 100 120

0 10 20 30 40 50 60 70 80 90 100

incident.riidFeatures.SNR

incident.riidFeatures.backgroundFraction

Conclusions
Collaborators:

- Artur Dubrawski, CMU, Auton Lab (advisor)
- Matt Barnes, CMU, RI and Auton Lab
- Karen Chen, CMU, Auton Lab
- Gilles Clermont, University of Pittsburgh
- Nick Gisolfi, CMU, Robotics
- Mathieu Guillaume-Bert, CMU, Auton Lab
- Peter Kontschieder, MSR Cambridge
- Marilyn Hravnak, University of Pittsburgh
- Michael R. Pinsky, University of Pittsburgh
- Donghan Wang, CMU, Auton Lab


