Big Data Analytics: Where Research Meets Reality

Shonali Krishnaswamy
Institute for Infocomm Research
Agency for Science, Technology and Research (A*STAR)
Previously...
Now
So much hay, so little time…
Data Analytics

- **DATA MINING**: exploration & analysis
  - of *large quantities of data*
  - by *automatic means*
  - to discover *actionable* patterns & rules
Data Analytics

Applications
Biomedical, Manufacture, Web, Transportation, Retail

Machine Learning Algorithms

DATA
Management and analysis of complex, unstructured, dynamic data
e.g. text, networks, streams, domain data (biomedical...)

INFORMATION
Extraction and representation
e.g. text extraction, ontology, privacy

KNOWLEDGE
Automated data-driven discoveries
e.g. decision trees, SVMs, NN

Value

Data Process
# Evolution of Data and Analytics

## Trend: Increasing data complexity and heterogeneity

<table>
<thead>
<tr>
<th>Data</th>
<th>Static and Structured Data</th>
<th>Unstructured data</th>
<th>Dynamic data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>Stream data, spatio-temporal data, distributed data in the cloud</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complex data</td>
<td>Graphs, fuzzy/uncertain data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multimedia data</td>
<td>Speech, image, video</td>
<td></td>
</tr>
</tbody>
</table>

## Trend: Increasing demands for data analytics

<table>
<thead>
<tr>
<th>Algo</th>
<th>Basic Analytics</th>
<th>Predictive analytics</th>
<th>Real-world data analytics</th>
<th>Decision support</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reasoning, anomaly detection, visualization, Trusted services and provenance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Multiple sources, noise, knowledge engineering, privacy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Knowledge discovery, prediction algorithms, artificial intelligence</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Online learning, distributed/cloud analytics services</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Embedded data mining, IoT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>App</th>
<th>Bioinformatics</th>
<th>Healthcare</th>
<th>Web</th>
<th>Transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Drug Discovery</td>
<td></td>
<td>Monitoring, fusion, diagnosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>User profiling, information extraction, social networking, sentiment analysis</td>
<td>Supply &amp; demand, planning, optimization, Visualization, green/sustainability</td>
</tr>
</tbody>
</table>

## Applications

<table>
<thead>
<tr>
<th>Healthcare</th>
<th>Web</th>
<th>Transportation</th>
</tr>
</thead>
</table>
Who needs to invent problems when there are so many out there?

Imbalance Correction
- From Aerospace to Telco
Churn to Mobile Activity Recognition
Automatic failure diagnostics and prognostics can be formulated as highly imbalanced time series classification tasks.
Imbalanced Time Series

Imbalance Learning Problem:
• Few (Pos) samples
• Abundant (Neg) samples
• Undesired Bias towards Neg class
• Class overlapping;
• High feature dimension with a small positive train dataset.

UCR Time Series Data:
- ECG
- Yoga (Wei-Keogh06)

Can we learn a good classifier in such a situation?
Advantages:
- Preserves the main covariance structure of Pos distribution;
- Adds protective buffer variances in the trivial eigen dimensions;
- Good for time series data classification.
Visual Comparison

- Repeating
- ADASYN
- Proposed SPO

- SMOTE
- Borderline SMOTE
- DataBoost Oversampling
# Experiment: Classification

SPO achieved remarkably better *F-Value* and *G-Mean* than several other well-known oversampling methods.

<table>
<thead>
<tr>
<th>Eval. metric</th>
<th>Dataset</th>
<th>Oversampling Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>REP</td>
</tr>
<tr>
<td><em>F-Value</em></td>
<td>Adiac</td>
<td>.375</td>
</tr>
<tr>
<td></td>
<td>S-Leaf</td>
<td>.716</td>
</tr>
<tr>
<td></td>
<td>Wafer</td>
<td>.962</td>
</tr>
<tr>
<td></td>
<td>FaceAll</td>
<td>.935</td>
</tr>
<tr>
<td></td>
<td>Yoga</td>
<td>.710</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.740</td>
</tr>
<tr>
<td><em>G-Mean</em></td>
<td>Adiac</td>
<td>.480</td>
</tr>
<tr>
<td></td>
<td>S-Leaf</td>
<td>.800</td>
</tr>
<tr>
<td></td>
<td>Wafer</td>
<td>.965</td>
</tr>
<tr>
<td></td>
<td>FaceAll</td>
<td>.950</td>
</tr>
<tr>
<td></td>
<td>Yoga</td>
<td>.741</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.787</td>
</tr>
</tbody>
</table>
Experiments: Comparison with SoA for Time Series Classification

<table>
<thead>
<tr>
<th>Eval. metric</th>
<th>Dataset</th>
<th>Learning methods</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Easy</td>
<td>Bal.</td>
<td>1NN</td>
<td>1NN DTW</td>
<td>SPO</td>
</tr>
<tr>
<td>F-Value</td>
<td>Adiac</td>
<td>.534</td>
<td>.348</td>
<td>.800</td>
<td>.917</td>
<td>.963</td>
</tr>
<tr>
<td></td>
<td>S-Leaf</td>
<td>.521</td>
<td>.578</td>
<td>.716</td>
<td>.429</td>
<td>.796</td>
</tr>
<tr>
<td></td>
<td>Wafer</td>
<td>.795</td>
<td>.954</td>
<td>.949</td>
<td>.857</td>
<td>.982</td>
</tr>
<tr>
<td></td>
<td>FaceAll</td>
<td>.741</td>
<td>.625</td>
<td>.802</td>
<td>.959</td>
<td>.936</td>
</tr>
<tr>
<td></td>
<td>Yoga</td>
<td>.356</td>
<td>.689</td>
<td>.652</td>
<td>.710</td>
<td>.702</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.589</td>
<td>.639</td>
<td>.784</td>
<td>.766</td>
<td>.876</td>
</tr>
<tr>
<td>G-Mean</td>
<td>Adiac</td>
<td>.782</td>
<td>.897</td>
<td>.875</td>
<td>.920</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td>S-Leaf</td>
<td>.712</td>
<td>.898</td>
<td>.798</td>
<td>.572</td>
<td>.898</td>
</tr>
<tr>
<td></td>
<td>Wafer</td>
<td>.817</td>
<td>.970</td>
<td>.953</td>
<td>.870</td>
<td>.984</td>
</tr>
<tr>
<td></td>
<td>FaceAll</td>
<td>.792</td>
<td>.918</td>
<td>.983</td>
<td>.985</td>
<td>.957</td>
</tr>
<tr>
<td></td>
<td>Yoga</td>
<td>.464</td>
<td>.688</td>
<td>.695</td>
<td>.741</td>
<td>.735</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.713</td>
<td>.874</td>
<td>.860</td>
<td>.810</td>
<td>.914</td>
</tr>
</tbody>
</table>

*SPO achieves better overall performance than other methods. The second best is 1NN or 1NN-DTW.*

*Easy: EasyEnsemble;*

*Bal.: BalanceCascade;*

*1NN: One nearest neighbor classifier with Euclidean distance measure;*

*1NN-DTW: One nearest neighbor with dynamic time warping distance measure.*
Publications

• Hong Cao, Xiao-Li Li, Yew-Kwong Woon, See-Kiong Ng, "SPO: Structure Preserving Oversampling for Imbalanced Time Series Classification", IEEE International Conference on Data Mining (ICDM 2011), Vancouver, Canada.

• Hong Cao, Xiao-Li Li, Yew-Kwong Woon and See-Kiong Ng, "Integrated Oversampling for Imbalanced Time Series Classification", IEEE Transactions on Knowledge and Data Engineering (TKDE), 2013
EU OPPORTUNITY Mobile Activity Recognition
Best in International Benchmarking Competition
First Place in All Four Challenge Tasks

9 participating teams from 6 countries in this challenge

I2R Teams

- Filling missing value
- Normalization
- Structure preserving oversampling
- SVM and 1NN
- Fusion and smoothing
- Decision Trees
- HMM

Output

Ubicomp 2012: An Integrated Framework for Human Activity Classification
H Cao, MN Nguyen, C Phua, S Krishnaswamy, XL Li
PAKDD 2012 Churn Prediction Challenge

- **Tasks**: Predict customer churns and win-backs for a large telco within the next 3 months
  - Predict top-5% of existing customers who will churn
  - Predict top-5% loyal customers who will not churn
  - Predict top-5% of ex-customers, who will come back soon
  - Identify the driving factors for churning or not churning
- **Data**: Customer transactional data from a major telco in Malaysia covering one whole year’s data in 2011.

**Mobile Customer Transactional Data**
- Year 2011’s data.
- Over 1 million customers.
- Over 30 million records for customer profile, service profiles, billing account, etc.

**I2R algorithm**

**Winner** of this competition out of 8 practitioners, 12 acad, 26 student teams

**Predicted top-5% near-future**
- Churners;
- Win-backs
- Loyal customers

**Ranked**
- Driving factors for churning
Background

• Telco industry (giant, high-penetration) is a dynamic and competitive industry with new technology and telco services rolling out in fast paces.

• Engaging, managing customer relation and maintaining a large customer base are crucial.
  – Monthly churn rate can be 2.2% (Wei & Chiu, 2002)
  – Acquiring a new client costs 5 to 6 times more than retaining an existing customer. (Verbeke et al., 2012)

• Timely churn prediction helps.
  – Retaining the existing customers by focusing on the limited CRM resource to the right consumer group and at the right time.
  – Attracting the ex-customers backs at the right time.
  – Identifying the key driving factors on churning or non-churning.
Big Transactional Data

• **1.2 Million customers** in year **2011** of a large Telco of Malaysia

• Six big tables
  – Customer profile (1.2M records)
  – Service Profile (1.4M records) for SME
  – Service Request/Compliant (2.7M records)
  – Billing Account (15M records)
  – Broadband Usage (1.3M+1.3M records)
  – Voice Usage (8.2M records)

*Challenge*: Can we turn such **transactional records** into insights and **predictive model** for near-future churning and comeback behaviors?
Relevant Research on Churn Management

• Existing Data Mining techniques for churn management
  – Decision Tree
  – Neural Network
  – Fuzzy data mining
  – Bayesian Belief Network
  – Profit-driven data mining
  – ....

• But seldom they tell the churning time
Technical Challenges

- **Predict Label and Time**: We need to take into account the timeline.
- **Scalability**: Huge datasets with large numbers of records and many attributes.
- **Dynamics**: Non-churner can turn into churner in a short time.
Our Approach

- Data preprocessing and time-line based data segmentation
- Feature engineering: design feature descriptors for all relevant factors
- Imbalance correction (oversampling)
- Benchmarking a large number of machine learning tools for achieving the best precision.
## Results for Consumers

<table>
<thead>
<tr>
<th></th>
<th>ADTree</th>
<th>Decision Stump</th>
<th>RepTree</th>
<th>J48</th>
<th>NaiveBayes</th>
<th>TreeLMT</th>
<th>Random Forest</th>
<th>Bagging + Decison Stump</th>
<th>Bagging + Simple Cart</th>
<th>SimpleCart</th>
<th>Clas.Via.Regession</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prec_1.</strong></td>
<td>75.8</td>
<td>59.7</td>
<td>75.8</td>
<td>69.2</td>
<td>74.0</td>
<td>77.7</td>
<td>75.7</td>
<td>86.6</td>
<td>74.5</td>
<td>77.4</td>
<td>78.6</td>
</tr>
<tr>
<td><strong>Prec_0</strong></td>
<td>70.8</td>
<td>69.1</td>
<td>71.9</td>
<td>72.3</td>
<td>69.7</td>
<td>59.7</td>
<td>72.7</td>
<td>65.4</td>
<td>73.6</td>
<td>74.5</td>
<td>71.0</td>
</tr>
<tr>
<td><strong>Accu.</strong></td>
<td>73.0</td>
<td>71.3</td>
<td>74.5</td>
<td>72.3</td>
<td>72.3</td>
<td>65.4</td>
<td>73.6</td>
<td>74.5</td>
<td>74.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prec_1: Precision for churner prediction; Prec_0: Precision for non-churner prediction
Accu.: Accuracy rate for classification of churners and non-churners

## Win-Back Prediction:

<table>
<thead>
<tr>
<th></th>
<th>ADTree</th>
<th>Decision Stump</th>
<th>RepTree</th>
<th>J48</th>
<th>NaiveBayes</th>
<th>TreeLMT</th>
<th>Random Forest</th>
<th>Bagging + BF Tree</th>
<th>Bagging + LaddTree</th>
<th>SimpleCart</th>
<th>Clas.Via.Regession</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prec_1.</strong></td>
<td>67.7</td>
<td>59.8</td>
<td>66.0</td>
<td>63.2</td>
<td>64.6</td>
<td>68.1</td>
<td>67.3</td>
<td>65.3</td>
<td>67.1</td>
<td>67.8</td>
<td>68.8</td>
</tr>
<tr>
<td><strong>Prec_0</strong></td>
<td>66.0</td>
<td>64.7</td>
<td>65.5</td>
<td>62.7</td>
<td>63.9</td>
<td>67.7</td>
<td>63.2</td>
<td>66.8</td>
<td>67.2</td>
<td>69.6</td>
<td>67.7</td>
</tr>
<tr>
<td><strong>Accu.</strong></td>
<td>66.8</td>
<td>60.7</td>
<td>65.8</td>
<td>63.0</td>
<td>64.2</td>
<td>67.9</td>
<td>65.0</td>
<td>66.0</td>
<td>67.2</td>
<td>68.2</td>
<td></td>
</tr>
</tbody>
</table>

Prec_1: Precision for win-back prediction; Prec_0: Precision for churner prediction
Accu.: Accuracy rate for classification of win-backs and churners
## Results for SMEs for Voice Services Only

### Churn Prediction:

<table>
<thead>
<tr>
<th></th>
<th>ADTree</th>
<th>Decision Stump</th>
<th>FT</th>
<th>J48</th>
<th>J48graft</th>
<th>LADTree</th>
<th>NBTree</th>
<th>RandoForest</th>
<th>RandoTree</th>
<th>RepTree</th>
<th>Simple Cart</th>
<th>Clas.Via. Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec_1</td>
<td>80.3</td>
<td>93.4</td>
<td>80.3</td>
<td>78.1</td>
<td>78.3</td>
<td>79.5</td>
<td>74.3</td>
<td>83.9</td>
<td>76.3</td>
<td>78.7</td>
<td>77.0</td>
<td>81.8</td>
</tr>
<tr>
<td>Prec_0</td>
<td>71.0</td>
<td>63.9</td>
<td>71</td>
<td>76.5</td>
<td>76.6</td>
<td>72.4</td>
<td>73.0</td>
<td>79.9</td>
<td>73.7</td>
<td>73.1</td>
<td>75.5</td>
<td>75.3</td>
</tr>
<tr>
<td>Accu</td>
<td>74.8</td>
<td>71.1</td>
<td>74.8</td>
<td>77.3</td>
<td>77.4</td>
<td>75.5</td>
<td>73.6</td>
<td>81.8</td>
<td>74.8</td>
<td>75.6</td>
<td>76.2</td>
<td>77.1</td>
</tr>
</tbody>
</table>

**Prec_1**: Precision for churner prediction; **Prec_0**: Precision for non-churner prediction  
**Accu**: Accuracy rate for classification of churners and non-churners

### Win-Back Prediction:

<table>
<thead>
<tr>
<th></th>
<th>ADTree</th>
<th>Decision Stump</th>
<th>FT</th>
<th>J48</th>
<th>J48graft</th>
<th>LADTree</th>
<th>NBTree</th>
<th>RandoForest</th>
<th>RandoTree</th>
<th>RepTree</th>
<th>Simple Cart</th>
<th>Clas.Via. Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec_1</td>
<td>73.5</td>
<td>74</td>
<td>67.5</td>
<td>70.6</td>
<td>71.1</td>
<td>68.8</td>
<td>75.3</td>
<td>74.2</td>
<td>71.9</td>
<td>72.4</td>
<td>70.2</td>
<td></td>
</tr>
<tr>
<td>Prec_0</td>
<td>67.2</td>
<td>62.7</td>
<td>67.9</td>
<td>70.5</td>
<td>70.8</td>
<td>65.6</td>
<td>68.8</td>
<td>72.7</td>
<td>70.6</td>
<td>70.1</td>
<td>71.4</td>
<td></td>
</tr>
<tr>
<td>Accu</td>
<td>69.8</td>
<td>66.6</td>
<td>67.7</td>
<td>70.5</td>
<td>70.9</td>
<td>69.8</td>
<td>68.8</td>
<td>73.4</td>
<td>71.2</td>
<td>71.2</td>
<td>70.8</td>
<td></td>
</tr>
</tbody>
</table>

**Prec_1**: Precision for churner prediction; **Prec_0**: Precision for non-churner prediction  
**Accu**: Accuracy rate for classification of churners and non-churners
Who needs to invent problems when there are so many out there?

Mobile and Ubiquitous Data Stream Mining
- Energy Analytics
Situation-Aware Adaptive Data Stream Processing

- Visualization
- Mining
- Adaptation
- Situation Inference
- Context Engine
- Sensory layer
Resource-Aware Adaptive Learning Algorithms

• Clusterers
  – Light-Weight Clustering
  – RA-Cluster and DRA-Cluster
  – RA-VFKM

• Change Detection
  – CHANGE-DETECT

• Classifiers
  – Light-Weight Class (LWC)

• Frequent Items and Associations
  – LWF (Light-Weight Frequent Items)
  – HiCoRE (Highly Correlated Energy-Efficient Rules)

• Time-Series Analysis – RA-SAX and RA-HOT SAX

• Visualisations
  – Clutter-Adaptive, Energy-efficient Visualiser for Mobile Data Analysis
Situation-Aware Adaptive Data Stream Processing
Wattzup – Context-Aware Smart Energy Management for Auto Demand Response

Collaborative Research Project with IBM

Combine advanced data mining / real-time stream analytics as well as rich context information to improve predictive accuracy of appliance usage through NILM (Non-Intrusive Load Monitoring)

1. Receive Appliance Usage Signals
2. Perform on-board learning and analysis and inform central system
3. Receive “Situation” info. from the central system
4. Activate smart control of appliances
Wattalyzer: Advanced Sensing & Real-Time Distributed Analytics for Condition Monitoring

Visualization

Realtime Streaming & Near-Realtime Analytics Platform

Distributed Data Store / Sandbox Platform

DIS-CU
DISU-HFCT
DISU-UHF
DISU-T (°C)
DISU-ETC

DIS – Distributed Intelligent Sensing
CU – Communication Unit
DISU – Distributed Intelligent Sensing Unit
PD – Partial Discharge
HFCT – High Frequency Current Transformer
UHF – Ultra High Frequency

HFCT PD Sensors
UHF PD Sensors
Temperature Sensors
Other Sensors
Who needs to invent problems when there are so many out there?

Location Data and Moving Objects – Transportation Analytics
Transportation Analytics

- Free-taxi during the time window
- Pick up during the time window
- Demand/Supply Ratio
Transportation Analytics

Jingbo Zhou, Anthony K. H. Tung, Wei Wu, Wee Siong Ng; "A Semi-Lazy Approach to Probabilistic Path Prediction in Dynamic Environments"; Proc. of 2013 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2013).

Jingbo Zhou, Anthony K. H. Tung, Wei Wu, Wee Siong Ng; "R2-D2: a System to Support Probabilistic Path Prediction in Dynamic Environments"; Proc. of 2013 Int. Conf. on Very Large Databases (VLDB 2013 Demo).
Can we predict a mobile user based on his/her interaction with his smartphone?

Place Semantics:
Where is the mobile user now, the semantic meaning of the current place?
Home, office, friend’s home, transportation location, … ?

User Demographics:
Is the mobile user a male or female? How old is he/she? What types of job is he/she doing? …

Next place:
Where is the next place that a smartphone user would go?
Background

• Personalization:
  – involves using technology to accommodate the differences between individuals (Wikipedia)
  – to meet an individual's specifications, needs, or preferences (Dictionary.com)

• Mobile user analytics:
  – identify the differences between mobile user groups
  – using smartphone’s data to understand, model and intelligently predict the mobile users;
  – smartphones have rich interaction data between the users, the phone and the environment.
Nokia’s Mobile Challenge Data

• Collected from more than 100 users for 1 year
• Comprehensive data in 14 different tables
  – Accelerometer, application, calendar, contact, call-log, wlan, bluetooth, gsm, mediaplay, process, mediaplay, visits, …
• Big data: About 50Gb for training dataset.
Prediction of Semantic Places

• Ten semantic classes:
  – Home, friend’s home, workplace, transportation location, friend’s workplace/school, etc.

• Our cross-validation accuracy (Weka tool):

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62.8%</td>
<td>62.8%</td>
<td>61.1%</td>
<td>59.2%</td>
<td>58.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Place Classification</td>
<td>AttributeSelectedClassifier</td>
<td>AdaBoostM1</td>
<td>Decorate</td>
<td>RandomSubspace</td>
<td></td>
</tr>
<tr>
<td></td>
<td>67.3%</td>
<td>65.5%</td>
<td>63.7%</td>
<td>61.9%</td>
<td></td>
</tr>
</tbody>
</table>

• AttributeSelectedClassifier achieved our best result 67.3%.
• Our result (67.3%) is better than that (65.3%) of MDC Zhu2012.
Prediction Tasks of User Demographics

- **Gender:**
  - 1: female
  - 2: male

- **Age group:**
  - 1: 16-21
  - 2: 22-27
  - 3: 28-32
  - 4: 33-38
  - 5: 39-44
  - 6: 45+

- **Marital status:**
  - Single or Divorce
  - In a relationship
  - Married or living together with my partner

- **Job:**
  - Training
  - PhD student
  - Employee without executive function
  - Employee exercising executive functions

- **Number of people in the household:**
  - 1, 2, 3, 4, 5 (4+)
## Our Results:

### Accuracy for Gender Classification via 10-fold Cross Validation

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th>JRip</th>
<th>NaiveBayes</th>
<th>NNge</th>
<th>RandomForest</th>
<th>HyperPipes</th>
<th>BayesNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td>83.3</td>
<td>89.7%</td>
<td>88.5%</td>
<td>89.7%</td>
<td>87.2%</td>
<td>92.3%</td>
</tr>
</tbody>
</table>

### Accuracy for Marital Status Classification via 10-fold Cross Validation

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th>NNge</th>
<th>RandomForest+J48</th>
<th>BayesNet</th>
<th>NaiveBayes</th>
<th>RotationForest+NaiveBayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marital</td>
<td></td>
<td>59.5%</td>
<td>53.2%</td>
<td>53.2%</td>
<td>60.8%</td>
<td>63.3%</td>
</tr>
</tbody>
</table>

### Root Mean Square Error for Age Group Regression via 10-fold Cross Validation

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th>LinearReg</th>
<th>REPTree</th>
<th>DecisionTable</th>
<th>RBFNetwork</th>
<th>GaussianProcess</th>
<th>SMOReg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Grp</td>
<td></td>
<td>1.19</td>
<td>1.35</td>
<td>1.30</td>
<td>1.15</td>
<td>1.15</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### Accuracy for Job Prediction via 10-fold Cross Validation

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th>REPTree</th>
<th>RandomForest+J48</th>
<th>J48</th>
<th>NaiveBayes</th>
<th>J48Graft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job</td>
<td></td>
<td>50.7%</td>
<td>54.8%</td>
<td>56.2%</td>
<td>50.7%</td>
<td>57.5%</td>
</tr>
</tbody>
</table>

### Root Mean Square Error for Number of People in a Household Regression via 10-fold Cross Validation

<table>
<thead>
<tr>
<th>Task</th>
<th>Learning Methods</th>
<th>M5Rules</th>
<th>LinearRegression</th>
<th>RBFNetwork</th>
<th>GaussianProcess</th>
<th>SMOReg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. People</td>
<td></td>
<td>0.962</td>
<td>1.189</td>
<td>1.083</td>
<td>0.970</td>
<td>0.698</td>
</tr>
</tbody>
</table>

- We achieved a good accuracy of 92.3% for gender prediction.
- SMOReg is highly effective for regression tasks.
Comparison with the Best of MDC Winners

<table>
<thead>
<tr>
<th>Task</th>
<th>MDC Best (Mo2012)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>89.7% (LR)</td>
<td>93.6%</td>
</tr>
<tr>
<td>Age</td>
<td>.864 (SVM)</td>
<td>0.68</td>
</tr>
<tr>
<td>Marital</td>
<td>78.5% (SVM)</td>
<td>71.3%</td>
</tr>
<tr>
<td>Job</td>
<td>78.1% (SVM)</td>
<td>80%</td>
</tr>
<tr>
<td>Num Ppl</td>
<td>.721 (SVM)</td>
<td>0.692</td>
</tr>
</tbody>
</table>

Bold face indicate the better comparative result

- Our results are generally better than the best results of the MDC winners (Zhu2012).
Next Place Prediction

- Demographics
- Semantic labeling
- Next place prediction
Next Place Prediction

- Data: Stream of user locations over time.
- Classification problem: from current location predict the next location.

Concept Drift in the data
Red location changes to Black.
Given the time of visit, this location represents home.
Data Analytics @ I2R
High-Speed Data Streams on the Cloud, Privacy Preserving Analytics

Large-Scale, Distributed Analytics for BIG Data, Mobile Data Mining

From Data to Knowledge to $ - Extracting Hidden Insight

Machine Learning

Information Extraction from Text, Natural Language Processing

Semantic Services, Reasoning, Cloud / Hadoop Resources Allocation

30+ Data Scientists and 10+ R&D Engineers
Publications (2011-2013)

• **Over 50 Tier 1 Publications**
  - ICDM
  - KDD
  - IJCAI
  - AAAI
  - IEEE Transactions (TKDE, SMCB, Neural Networks etc.)
  - Journal of Machine Learning
  - SIGKDD Explorations
  - ICDE
  - VLDB
  - Ubicomp
  - Journal of Computer Systems and Science
  - ACL
Awards & Benchmarking (2011-2013)

• First Place in GE Flight Quest Challenge – 2013
• Third Place for EC2BargainHunter in IEEE Cloud Cup – 2013
• First Place in PAKDD Churn Prediction - 2012
• First Place in Fraud Detection in Mobile Advertising - 2012
• First Place in Mobile Activity Recognition Challenge – 2011
• Third Place in Time-Series Forecasting Competition - 2012
• Third Place in IEEE Services Cup - 2012
• Fifth Place in IEEE International Conference on Data Mining (ICDM) Contest - 2012
• Top 5 Innovative Ideas in the Urban Prototyping Challenge @ World Cities Summit - 2012
• Second Place in NIST Entity Linking Competition - 2011
Industry Engagement (2011-2013)

- Financial Analytics – VISA
- Condition Monitoring & Predictive Asset Maintenance
  - Boeing
  - Rolls Royce
  - EADS
- Energy Analytics
  - Singapore Powergrid
  - Power Automation
- Customer Insights
  - SingTel
  - Air Asia
  - STEE
  - OneEmpower
- Text Mining
  - EADS
  - SQLView
GE Flight Quest Challenge

No. 1 in Flight Arrival Time Predictions

First place out of 180 teams in predicting runway and gate arrival times of US domestic flights using flight data

Flight Quest
Make flying more efficient

Flight Data
- **Historical data** for 26,000 flights x 87 days.
- **Flight details**
  - Airline, aircraft, scheduled times, actual departure time, gate
- **Flight route plans**
- **Weather**
  - wind and temperature at various altitudes, cloud cover, adverse phenomena, turbulence, icing, dewpoint...

I2R Algorithm

Estimation of gate & runway arrival times

- 40% and 45% improvements over the standard industry benchmark estimates
- Average errors of 4.2 and 3.2 minutes for gate and runway arrivals

http://www.gequest.com/c/flights
The Flight Data

- Big with over 300 GB uncompressed;
- 87 days of concurrent flight data for the entire set of US domestic flights;
- Over 26,000 flights per day;
- Each flight associated with large amount of data in different types:
  - Flight history: scheduled/actual flights stats and events;
  - ASDI: flight route plan, actual trajectories, etc.
  - ATSCC: airport delays, deicing, ground delays, etc.
  - METAR: Weather reports by the weather stations
  - Other Weather: weather phenomena, wind, turbulence, etc.
  - Info. on weather stations
GE Flight Quest - Phase 1 (179 teams, 242 participants, 3073 entries)

I²R’s Winning Formula

Competition Data: 26,000 US domestic flights and weather data x 87 days

Feature Engineering:
- 258 Features Extracted

Feature Selection:
- 84 Features Selected
  - Only 58 used for predicting Runway Arrival Time

Machine Learning

Winning Results:
- Average 40%-45% less errors for gate and runway arrival time, respectively, compared to the standard industry benchmark

Prediction of Runway and Gate Arrival Time on Test Data (unseen)

Mixture of Prediction Models: GBM and RF models
Publicity and Potential Impact

S’pore team tops in predicting flight timings

It beats 120 teams with solution that could help airlines save millions of dollars

Cruz Chua
Senior Correspondent

Lines reduce gate congestion and manage crews more efficiently.

Each minute reduced per flight could also save US$1.2 million in annual crew costs and US$5 million in annual fuel savings for a mid-sized airline, it added.

Multiply the savings across hundreds of airlines around the world and the potential savings are huge, besides greater efficiency for airlines and convenience for passengers.

The challenge was issued last November. To come up with a solu-
Student Internships @ I2R

- Under-graduate
- Post-graduate
- Doctoral

- Topics: From theoretical to applied….and often at the intersection