An Adaptive Framework for Multistream Classification

Swarup Chandra, Ahsanul Haque, Latifur Khan and Charu Aggarwal*
University of Texas at Dallas
*IBM Research

This material is based upon work supported by
Data Stream Classification

- Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. ICDM 2008: 929-934
Data Stream Classification

Sources

Model

Time t

Time t+1

Label Prediction
Data Stream Analytics

- Ahsanul Haque, Latifur Khan, Michael Baron, Bhavani M. Thuraisingham, Charu C. Aggarwal: Efficient handling of concept drift and concept evolution over Stream Data. ICDE 2016: 481-492
Data Stream Analytics

Sources

Model

Semi-supervised or Active Learning

Label Prediction

Evaluation

Concept Drift detection
Motivation

What if we do not find a good training set?

Biased training data selection mechanism.

Example Scenario

Small set of users

Population

Biased

Labeled training data

Unlabeled test data

Affects Classifier Accuracy
Problem (Multistream Classification)

Two types of data stream (independent).

Source Stream

Target Stream

Labeled Data

Unlabeled Data

Create training data

Label Prediction

Create

training
data

Label
Prediction
Problem (Multistream Classification)

Two types of data stream (independent).

Source Stream
- Labeled Data

Target Stream
- Unlabeled Data

Create training data

Label Prediction

Concept Drift detection

\[ t \rightarrow t+1 \]
Potential Applications

Domain Adaptation and Transfer Learning over data streams

Text Classification

Sensor-based location estimation

Collaborative filtering
Outline

Challenges
Solution Overview (MSC)
Framework Details
Empirical Evaluation
Conclusion
Challenges

- Leveraging labeled and unlabeled data
  - bias-corrected training set.
- Asynchronous concept drift in source and target stream.
  - Drift detection
  - Drift correction
Challenges

• Can the two streams be combined?
  • Data distributions are different.
  • Combination represent same distribution
  • Separate representation has advantages when multiple sources are present.
Solution Overview

Target Stream (Unlabeled)

Non-stationary Domain

Source Stream (Labeled)

Output

Class Prediction

Target Classifier

Drift Detection (CDT)

Ensemble Update

Source Classifier
Design Overview

- Two data streams
Design Overview

- Two data streams
- To address asynchronous concept drift.
Design Overview

- Two data streams
- To address asynchronous concept drift.
Solution Overview

• Data in source and target occur simultaneously.
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data ...

```plaintext
- Target Stream (Unlabeled)
- Source Stream (Labeled)
- Non-stationary Domain
- Target Classifier
- Source Classifier
- Ensemble Update
- Drift Detection (CDT)
- Output
- Class Prediction
- Update
- Output

1 2 3 4a 4b 5a 5b
- Data in source and target occur simultaneously.
- In the case of source data ...
```
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data...
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data, drift detection output used to update source classifier.
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data, drift detection output used to update source classifier.
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data ...
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data...

Diagram:
- Target Stream (Unlabeled)
  - Non-stationary Domain
- Source Stream (Labeled)
  - Class Prediction
    - Drift Detection (CDT)
    - Ensemble Update
  - Target Classifier
    - Output
      - 1
        - 2
          - 3
            - 6
              - 4a
                - 5a
                  - 5b
    - Source Classifier
      - 4b
        - 4a
            - 2
              - 3
                - 6
                  - 4a
                    - 5a
                      - 5b
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data ...
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data, drift detection output used to update target classifier.
- Target classifier corrects bias between source and target stream at time t.
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data, drift detection output used to update target classifier.
- Target classifier corrects bias between source and target stream at time t.
Classifier

• **Source Classifier**
  • Typical classifier using training data from source stream.
  • Predict labels of newly occurring source stream data.

• **Target Classifier**
  • *Bias corrected* source stream data for training.
  • Predict labels of newly occurring target stream data.
Target Classifier

- Training: Sampling bias correction via Kernel Mean Matching
  - Minimize mean discrepancy between labeled source and unlabeled target distribution.

\[
\beta(t)^* \approx \min_{\beta(t)} \frac{1}{2} \beta^T K \beta - \kappa^T \beta
\]

subject to \( \beta_i \in [0, B_{kmm}] \) & \[ \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_i - 1 \] \( \leq c_{kmm} \)

Source data instance weight: \( \beta(B_S) = \frac{P_T(B_T)}{P_S(B_S)} \)

\( B_S \) : Source window

\( B_T \) : Target window

Matrices of kernel in RKHS:

\[
K_{ij} = k(x_{tr}^{(i)}, x_{tr}^{(j)})
\]

\[
\kappa_i = \frac{n_{tr}}{n_{te}} \sum_{j=1}^{n_{te}} k(x_{tr}^{(i)}, x_{te}^{(j)})
\]
Label Prediction

- Finite dynamic size window for incoming source and target data.
- Weighted hybrid ensemble
  - Fixed number of classifiers.
  - Contains both source and target classifiers.
  - Source classifier weight based on classifier error.
  - Target classifier weight based on classifier confidence on unlabeled target data.

\[ w_S : \frac{1}{2} \ln \frac{1-n}{n} \]
Concept Drift Detection

- **Source classifier error window**
  - Contain binary values.
  - Follow Bernoulli distribution.

- **Target classifier confidence window**
  - Contain confidence value between 0 and 1.
  - Follow Beta distribution.

CUSUM-type change point detection to detect change point at element $q$ of window $W$.

Sequential sub-window

\[
\begin{align*}
W_h^b &= W[1 : q] \\
W_h^a &= W[q + 1 : n]
\end{align*}
\]

Likelihood ratio score at point $q$:

\[
s(q, n) = \sum_{i=q+1}^{n} \log \left( \frac{P(W_h[i] | \theta_a)}{P(W_h[i] | \theta_b)} \right)
\]

Change point is at $q$ if:

\[
\omega_n = \max_{\gamma \leq q \leq n-\gamma} s(q, n) > \text{Threshold}
\]
Drift Adaptation

- Why not train both types of classifiers once a drift is detected on either stream?
- Sampling bias correction if target stream has a concept drift.

Case 1
Source only drift
Source Adaptation not required

Case 2
Target only drift
Source Adaptation required

Case 3
Source & Target drift
Source Adaptation required
Empirical Evaluation

### Real World
- ForestCover
  - # features: 53
  - # classes: 7
  - # instances: 146,438
- Sensor
  - # features: 5
  - # classes: 58
  - # instances: 150,000
- SEA
  - # features: 3
  - # classes: 3
  - # instances: 58,000

### Synthetic
- SynEDC
  - # features: 40
  - # classes: 20
  - # instances: 98,816
- SynRBF@00
  - # features: 70
  - # classes: 7
  - # instances: 98,000
- SynRBF@00
  - # features: 70
  - # classes: 7
  - # instances: 98,686

Divide dataset into Source and Target Stream, with bias in source stream data selection according to: $e^{-|x - \bar{x}|^2}$
Empirical Evaluation

- SVM as base classifier
  - Source Classifier : Typical multiclass SVM.
  - Target Classifier : Weighted SVM

- Classifier confidence:
  - Distance of test data to hyperplane.
Empirical Evaluation

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sKMM</td>
<td>Single target classifier without update.</td>
</tr>
<tr>
<td>mKMM-5k</td>
<td>Single target classifier with update every 5k instances.</td>
</tr>
<tr>
<td>srcMSC</td>
<td>CPD with source classifier only. No bias correction.</td>
</tr>
<tr>
<td>trgMSC</td>
<td>CPD with target classifier only. No source drift adaptation.</td>
</tr>
<tr>
<td>MSC</td>
<td>Proposed method with hybrid ensemble.</td>
</tr>
<tr>
<td>MSC2</td>
<td>Proposed method with separate source and target ensemble.</td>
</tr>
</tbody>
</table>
Results

ForestCover Dataset

MSC is better

Sensor Dataset

MSC2 is better

MSC baselines also good, but ..
Results

SynRBF@002 Dataset

MSC2 is better

SynRBF@003 Dataset

MSC2 is better
Conclusion

• Introduce a new data stream mining setting with bias labeled data
• Propose a framework to address new challenges of concept drift in this setting.
• Empirical results achieve significantly better accuracy than baseline.

• Future work: Multi-source setting and Semi-supervised target stream classification.
Thank you

Q & A