MineClass: A Synergy of Data Stream Classification and Novel Class Detection
Mohammad Meheyd Masud†, Jing Gao‡, Latifur Khan†, Jiawei Han‡ and Bhavani Thuraisingham†
†University of Texas at Dallas, ‡University of Illinois at Urbana Champaign

Introduction
Data Stream Classification faces three major problems:
• Infinite length: solution – incremental learning
• Concept-drift: solution – adapting to the most recent concept
• Novel class: solution - MineClass

Figure 1. Overview of MineClass

Applications of novel class detection:
• Intrusion detection: detecting new kind of attack
• Text classification: detecting new category of text
• Fault detection: detecting new kind of fault

Basic idea
Assumption: A data point should be closer to the data points of its own class (cohesion) and farther apart from the data points of any other classes (separation).

Figure 2. A decision tree (a), corresponding feature-space partitioning (b), and arrival of a novel class in the unused portions of the feature space (c).

Novel class detection
Steps:
1. Save inventory of used spaces during training
2. Outlier detection and filtering
3. Measuring cohesion and separation

Figure 3. Saving inventory of used spaces as micro-clusters

Figure 4. Outlier detection and filtering

N-NSC(x) = \frac{b_{min}(x) - a(x)}{\max(b_{min}(x), a(x))}

Figure 5. Measuring cohesion and separation by computing (a) λ neighborhood for N=3, and (b) N-neighborhood silhouette coefficient (N-NSC). A novel class is declared if N-NSC(\lambda) is positive for at least N instances

Figure 5. Error comparison on (a) SynCN, (b) SyncN, (c) KDD and (d) Forest Cover.

Table I. Run-time comparison on all datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Time/sec/chunk</th>
<th>Points/sec</th>
<th>Speed gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>SynC</td>
<td>0.18</td>
<td>0.831</td>
<td>5.346</td>
</tr>
<tr>
<td>NycCN</td>
<td>0.27</td>
<td>0.651</td>
<td>3.825</td>
</tr>
<tr>
<td>KDD</td>
<td>0.95</td>
<td>199.5</td>
<td>22.255</td>
</tr>
<tr>
<td>Forest</td>
<td>2.11</td>
<td>21.31</td>
<td>10.499</td>
</tr>
</tbody>
</table>

Experiments and results
We evaluated our approach on two synthetic and two real datasets
Baseline: MineClass (MC), WCE-OILINDDA_PARALLEL(W-OP): WCE-OILINDDA_SINGLE(W-OS)
WCE-OILINDDA is a combination of the Weighted Classifier Ensemble (WCE) [1] and novel class detector OLINDA [2]

Conclusion
MineClass handles all three problems in data stream classification:
• Handles infinite memory and concept-drift problem using ensemble classification
• Novel class detection problem using “class mining”
• It is an unique combination of stream classification and novel class detection

References