Introduction

- **Multi-label Data**
  - Each instance is associated with multiple labels
  - Given instances \( x_1, x_2, \ldots \) and a predefined set of labels \( L \):
    - single-label data: \( (x_i, l_i) \) where \( i \in L \)
    - multi-label data: \( (x_i, S_i) = (x_i, l_i, \ldots, l_j) \) where \( S_i \subseteq L \)
- **Data Streams**
  - theoretically infinite stream
  - potentially large amount of data
- **Examples of multi-label data streams**:
  - news, new feeds
  - forums, newsgroups
  - social networking sites
  - e-mail
  - scene and video classification
- **Why Generate Synthetic Multi-label Data Streams?**
  - create more multi-label data stream (very few real world datasets)
  - allow a theoretically infinite stream
  - analyse certain algorithm properties

How to Generate Synthetic Multi-label Data Streams

- Using existing single-label data stream generators
- Combine the label space and feature space of single-label examples to create multi-label examples.

\[(x_1, l_1), (x_2, l_2), (x_3, l_3) \rightarrow (x_1 \oplus x_2 \oplus x_3, \{l_1, l_2, l_3\}) \rightarrow (x', \{l_1, l_2, l_3\})\]

Label Skew, Label Cardinality and Label Distribution

- **Label skew**: the overall frequency of each label
  - in multi-label data: more than one label can be relevant to over 50% of examples
  - data naturally skewed when combining single-label instances
- **Label cardinality**: the average number of labels per example
- **Two types of label distribution**:
  - **Type A** Multiple labels to resolve ambiguities. E.g. newsgroups, Slashdot
  - **Type B** Label set chosen specifically for a multi-labelling task. E.g. Enron, Media Mill
- Can be approximated by a Poisson function: \( \text{POIS}(k, \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \)

Label Relationships

- **Multi-label data exhibits relationships between labels**.
  - For example (Economy, Politics) more likely than (Economy, Sports)
- **These relationships can be represented in the form of a contingency matrix**:
  - \( m[l \mid j] = Pr(l \mid j) \) (relationship)
  - \( m[l \mid j] = Pr(l) \) (frequency)

Adding Concept Drift

- **Stream data is affected by concept drift**.
  - **Feature space concept drift** (a)
  - **Label space concept drift** (b) (multi-label specific)

- **Synthetic concept drift can be approximated with a sigmoid function**
  \[ \text{sig}(d) = \frac{1}{1 + e^{-d}} \]
  - Applied to either label space or feature space.

Conclusions

- Analysis of multi-label data, and concept drift
  - A framework for creating synthetic multi-label data streams
  - Software: http://cs.waikato.ac.nz/~jmr30/#software
  - Contact: {jmr30, bernhard, geoff}@waikato.ac.nz