Synthetic Multi-label Data Streams

Jesse Read, Bernhard Pfahringer, Geoff Holmes

University of Waikato, Hamilton, New Zealand

Introduction

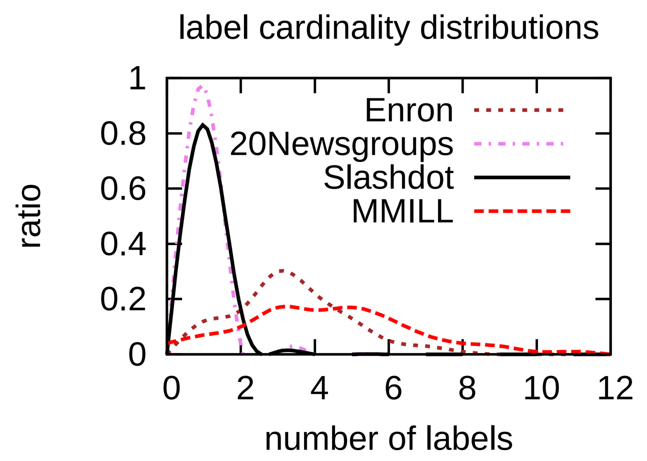
- Multi-label Data
  - Each instance is associated with *multiple* labels
  - Given instances $x_1, x_2, \cdots$ and a predefined set of labels L:
    - single-label data: $(x_1, l_1), (x_2, l_2), \cdots$ where each $l_i \in L$
    - multi-label data: $(x_1, S_1), (x_2, S_2), \cdots$ where each $S_i \subseteq L$
- Data Streams
  - theoretically infinite stream
  - potentially large amount of data
- Examples of multi-label data streams:
  - news, news feeds
  - forums, newsgroups
  - social networking sites
  - e-mail
  - scene and video classification
- Why Generate Synthetic Multi-label Data Streams?
  - create more multi-label stream data (very few real world datasets)
  - allow a theoretically infinite data stream
  - analyse certain algorithm properties

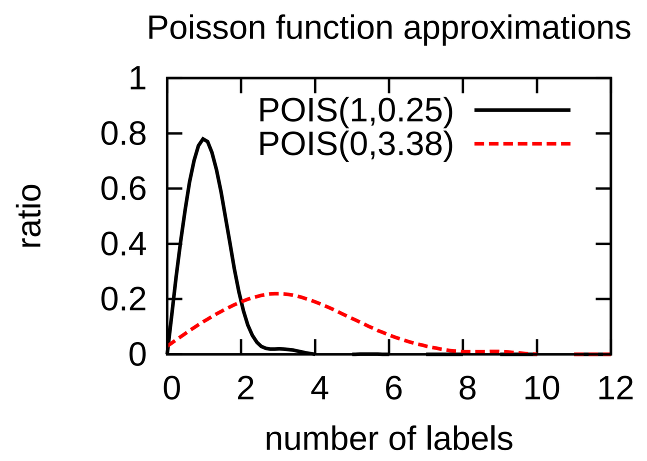
How to Generate Synthetic Multi-label Data Streams

- 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. *20 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: $POIS(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

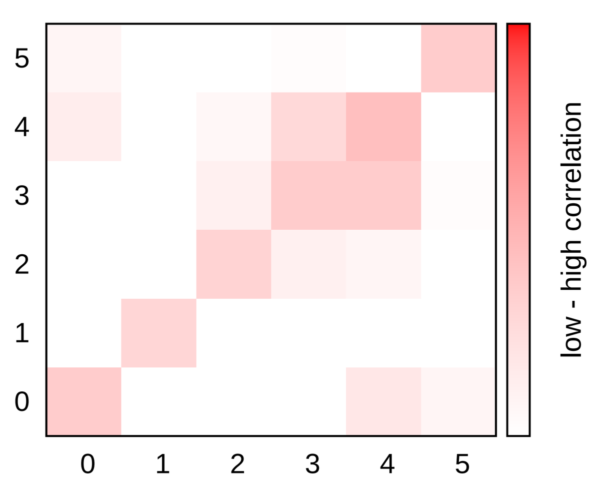


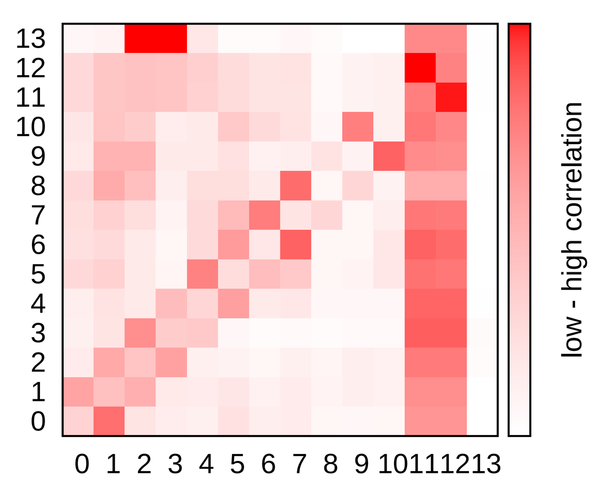


Label Relationships

- 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:
  - $m[k][j] = Pr(l_k|l_j)$ (relationship)
  - $m[k][k] = Pr(l_k)$ (frequency)

Figure: *Scene* ($|L| = 6$) and *Yeast* ($|L| = 14$).





- A synthetic matrix with similar properties to real-world data controls the *label space* of multi-label examples.

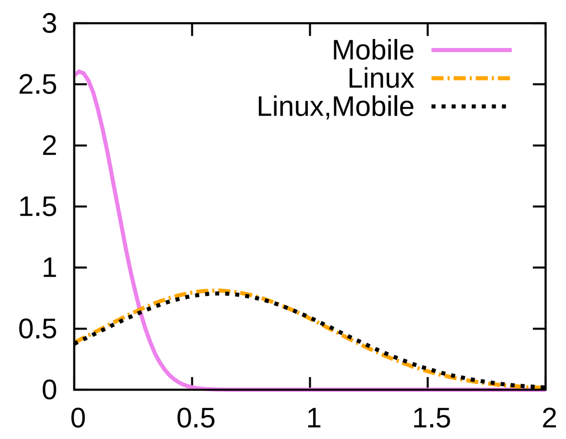
Feature Space

- Three main *feature-label effects* in real world multi-label data:
  - Feature-Label effect: A feature identifies a label, e.g. linux, mobile, phone
  - Feature-Combination effect: A feature identifies a combination of labels, e.g. netbook
  - Random effect: A feature does not identify anything, e.g. anonymous

Table: Most frequent word features from *Slashdot* for labels `Linux` and `Mobile` and the combination {`Linux`,`Mobile`}.

`Linux`	`Mobile`	{`Linux`,`Mobile`}
linux	mobile	linux
ubuntu	iphone	open
source	anonymous	windows
open	reader	phone
released	phone	netbook
anonymous	android	source
kernel	apple	mobile
software	phones	free

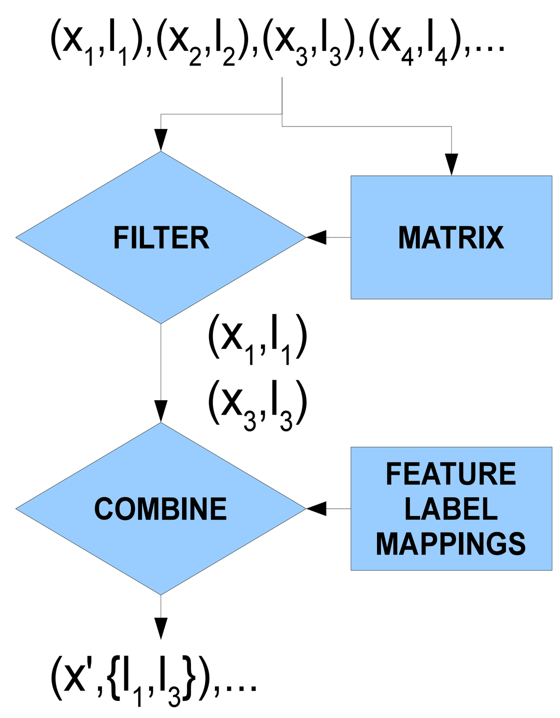
Figure: Frequencies of the feature 'linux' as normal distributions.



- Map each feature to one label effect, i.e. create feature-label mappings, to combine the feature space of single-label examples.

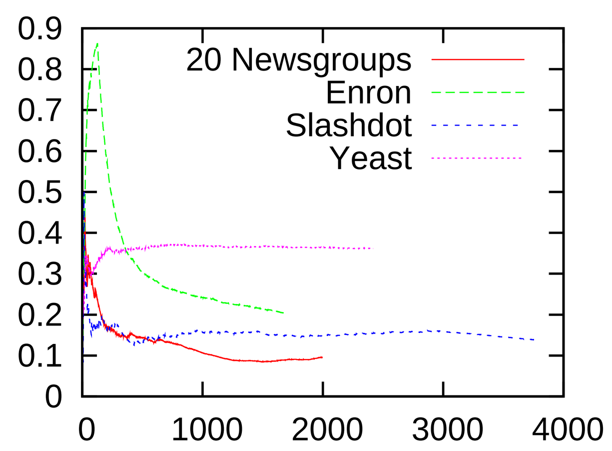
Process

- Setup:
  1. initialise a single-label data stream
  2. calculate skew and create a contingency **MATRIX**
  3. assign **FEATURE-LABEL MAPPINGS**
- Process:
  1. select a single-label example
  2. **FILTER** more examples according to label relationship **MATRIX**
  3. **COMBINE** the label and feature spaces into a multi-label example
  4. repeat

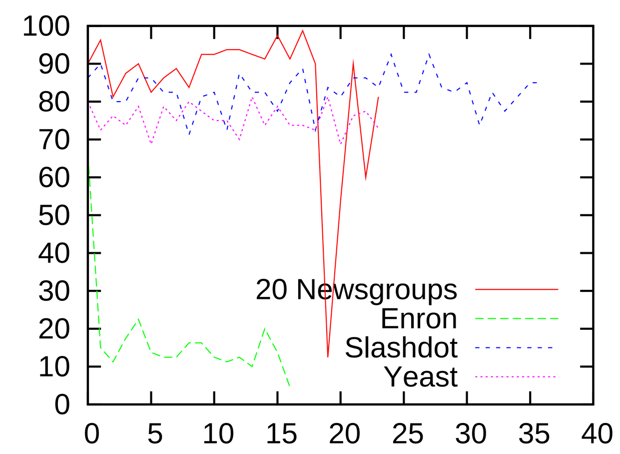


Adding Concept Drift

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

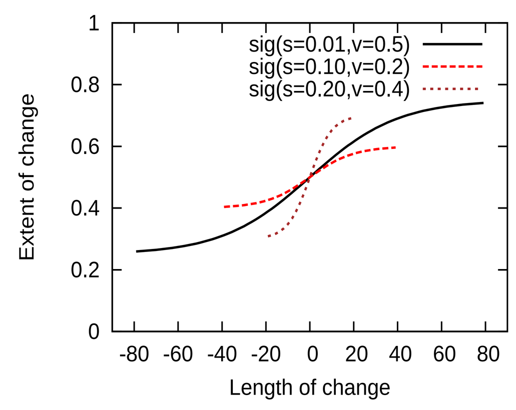


(a) accuracy over time



(b) label combinations over time

- Synthetic concept drift can be approximated with a sigmoid function $sig(d) = \frac{1}{(\Delta x + e^{-s(d-d_0)})}$
  - 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`