Scalable Multi-label Classification

Jesse Read
Supervised by Bernhard Pfahringer and Geoff Holmes
Machine Learning Group
University of Waikato
New Zealand

September 8, 2010

Introduction

- Supervised Classification: given *training data* (consisting of examples of input instances associated with an output e.g. labels) train a *classifier* that can predict output *automatically* for new instances.

Introduction

- Supervised Classification: given *training data* (consisting of examples of input instances associated with an output e.g. labels) train a *classifier* that can predict output *automatically* for new instances.

- Multi-class (Single-label) Classification: predict a class label; e.g. $\in \{\texttt{Beach}, \texttt{Forest}, \texttt{Urban}, \texttt{Sunset}\}$

Introduction

- Supervised Classification: given *training data* (consisting of examples of input instances associated with an output e.g. labels) train a *classifier* that can predict output *automatically* for new instances.

- Multi-class (Single-label) Classification: predict a class label; e.g. $\in \{\texttt{Beach}, \texttt{Forest}, \texttt{Urban}, \texttt{Sunset}\}$
- Multi-label Classification: predict (potentially multiple) labels e.g. $\subseteq \{\texttt{Beach}, \texttt{Forest}, \texttt{Urban}, \texttt{Sunset}\}$

Introduction

- Supervised Classification: given *training data* (consisting of examples of input instances associated with an output e.g. labels) train a *classifier* that can predict output *automatically* for new instances.

- Multi-class (Single-label) Classification: predict a class label; e.g. $\in \{\texttt{Beach}, \texttt{Forest}, \texttt{Urban}, \texttt{Sunset}\}$
- Multi-label Classification: predict (potentially multiple) label*s* e.g. $\subseteq \{\texttt{Beach}, \texttt{Forest}, \texttt{Urban}, \texttt{Sunset}\}$

Multi-label classification is the supervised classification task where each data instance may be associated with *multiple* class labels.

Notation

- Input space: $\mathcal{X} = \mathbb{R}^d$
- Instance $\mathbf{x} = [x_1, \ldots, x_d]$
- Output space: $\mathcal{Y} = \{0, 1\}^L$
- Labels: $\mathbf{y} = [y_1, \ldots, y_L]$ where $y_j = 1$ if jth label relevant to $\mathbf{x}$ (else 0)
- Training examples: $\{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, \ldots, N\} \subset (\mathcal{X} \times \mathcal{Y})$
- Classification: $\mathcal{X} \rightarrow \mathcal{Y}$
- Prediction: $\hat{\mathbf{y}} = \mathbf{h}(\mathbf{x})$
- Evaluation:
  - $\hat{\mathbf{y}}_i = \mathbf{y}_i$?
  - $\hat{y}_{ij} = y_{ij}$?

Example Applications ($\mathcal{X} \to \mathcal{Y}$)

Multi-label classification is relevant to many domains:

- Text
  - text documents $\to$ subject categories
  - e-mails $\to$ labels
  - medical description of symptoms $\to$ diagnoses
- Vision
  - images/video $\to$ scene concepts
  - images/video $\to$ objects identified/recognised
- Audio
  - music $\to$ genres / moods
  - sound signals $\to$ events / concepts
- Bioinformatics
  - genes $\to$ biological functions
- Robotics
  - sensor inputs $\to$ states / error diagnosis

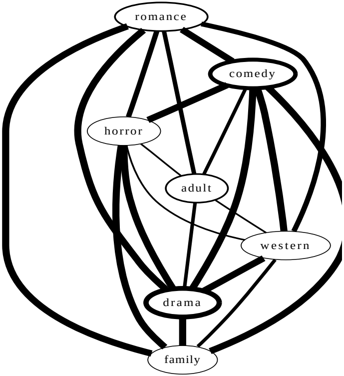
Datasets and Statistics

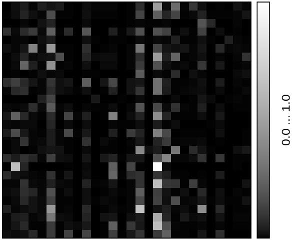
	N	L	$(\sum \mathbf{y})/N$	$uniq.\mathbf{y}$	Type
Music	593	6	1.87	0.046	media
Scene	2407	6	1.07	0.006	media
Yeast	2417	14	4.24	0.082	biology
Genbase	661	27	1.25	0.048	biology
Medical	978	45	1.25	0.096	medical text
Slashdot	3782	22	1.18	0.041	news
Lang.Log	1460	75	1.18	0.208	forum
Enron	1702	53	3.38	0.442	e-mail
Reuters(avg)	6000	103	1.46	0.147	news
OHSUMED	13929	23	1.66	0.082	medical text
tmc2007	28596	22	2.16	0.047	text
Media Mill	43907	101	4.38	0.149	media
Bibtex	7395	159	2.40	0.386	text
IMDB	95424	28	1.92	0.036	text
del.icio.us	16105	983	19.02	0.981	text

Issues and Challenges

Multi-label learning issues / challenges:

- correlations between labels
- dimensionality (output space 2^L instead of L)
- measures of evaluation / loss functions
- an emerging task; no 'standardised' datasets, measures, benchmark methods, etc.





(IMDB dataset: co-occurrences (subset), and conditional probabilities)

Aim

Existing methods:

- very computationally complex (often not applicable in practice);
- very specialised (for a specific domain, dimension, setting); or
- not very competitive (in terms of predictive performance).

Aim

Existing methods:

- very computationally complex (often not applicable in practice);
- very specialised (for a specific domain, dimension, setting); or
- not very competitive (in terms of predictive performance).

The aim of this research was to provide multi-label methods which are:

- scalable
- generally applicable; and
- competitive with state-of-the-art methods

Approach: Problem Transformation

Problem Transformation

- Transform a multi-label problem into single-label problems
- Flexible, general, can be more scalable
- Can use any off-the-shelf single-label classifier (kNN, Decision Trees, SVMs, Naive Bayes, *etc.*)

Approach: Problem Transformation

Problem Transformation

- Transform a multi-label problem into single-label problems
- Flexible, general, can be more scalable
- Can use any off-the-shelf single-label classifier (kNN, Decision Trees, SVMs, Naive Bayes, *etc.*)

For example:

- Label Combination method: each combination becomes a single class-label.
  - $\mathcal{Y} = \mathrm{distinct}(\{\mathbf{y}_1, \ldots, \mathbf{y}_N\})$
  - $\hat{\mathbf{y}} = h(\mathbf{x})$
- Binary Relevance method: each label is a separate binary problem.
  - $\mathcal{Y}_j = \{0, 1\}$
  - $\hat{y}_j = h_j(\mathbf{x})$

Main Contribution 1: The Pruned Sets Method

The Label Combination method (each $\mathbf{y}_i$ is a single class-label):

- Usually good performance, but
- worst-case complexity $\min(2^L, N)$ classes; and
- issues with label sparsity and overfitting.

[1] J. Read, B. Pfahringer, G. Holmes. Multi-label Classification using Ensembles of Pruned Sets. Proc. of IEEE International Conference on Data Mining. 2008.

Main Contribution 1: The Pruned Sets Method

The Label Combination method (each $\mathbf{y}_i$ is a single class-label):

- Usually good performance, but
- worst-case complexity $\min(2^L, N)$ classes; and
- issues with label sparsity and overfitting.

The **Pruned Sets Method** [Read et al., 2008][1]: Prune and subsample *infrequent* label combinations.

- prune where $P(\mathbf{y}_i) < p$, and subsample top s best subsets $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{is}$ (more frequent and more labels = better)
- e.g. $(\mathbf{x}, [1^{beach}1^{urban}1^{forest}0^{sunset}]) \rightarrow (\mathbf{x}, [1100]), (\mathbf{x}, [1010])$
- up to two orders of magnitude faster (with SVMs)
- reduces label sparsity and overfitting

[1] J. Read, B. Pfahringer, G. Holmes. Multi-label Classification using Ensembles of Pruned Sets. Proc. of IEEE International Conference on Data Mining. 2008.

Main Contribution 1: The Pruned Sets Method

The Label Combination method (each $\mathbf{y}_i$ is a single class-label):

- Usually good performance, but
- worst-case complexity $\min(2^L, N)$ classes; and
- issues with label sparsity and overfitting.

The **Pruned Sets Method** [Read et al., 2008][1]: Prune and subsample *infrequent* label combinations.

- prune where $P(\mathbf{y}_i) < p$, and subsample top s best subsets $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{is}$ (more frequent and more labels = better)
- e.g. $(\mathbf{x}, [1^{beach}1^{urban}1^{forest}0^{sunset}]) \rightarrow (\mathbf{x}, [1100]), (\mathbf{x}, [1010])$
- up to two orders of magnitude faster (with SVMs)
- reduces label sparsity and overfitting

Ensembles of Pruned Sets:

- more robust; competes with state-of-the-art methods

[1] J. Read, B. Pfahringer, G. Holmes. Multi-label Classification using Ensembles of Pruned Sets. Proc. of IEEE International Conference on Data Mining. 2008.

Main Contribution 2: The Classifier Chains Method

The Pruned Sets method worked well, but had limitations:

- difficulty dealing with 'extreme' datasets; and
- worst-case time same as the Label Combination method.

[2] J. Read, B. Pfahringer, G. Holmes, E. Frank. Classifier Chains for Multi-label Classification. In Proc. of European Conference on Machine Learning. 2009.

Main Contribution 2: The Classifier Chains Method

The Pruned Sets method worked well, but had limitations:

- difficulty dealing with 'extreme' datasets; and
- worst-case time same as the Label Combination method.

The Binary Relevance method (L separate problems; $\hat{y}_j = h_j(\mathbf{x})$):

- Relatively robust and good theoretical time complexity; but
- does not explicitly model label correlations (poor prediction).

---

[2] J. Read, B. Pfahringer, G. Holmes, E. Frank. Classifier Chains for Multi-label Classification. In Proc. of European Conference on Machine Learning. 2009.

Main Contribution 2: The Classifier Chains Method

The Pruned Sets method worked well, but had limitations:

- difficulty dealing with 'extreme' datasets; and
- worst-case time same as the Label Combination method.

The Binary Relevance method (L separate problems; $\hat{y}_j = h_j(\mathbf{x})$):

- Relatively robust and good theoretical time complexity; but
- does not explicitly model label correlations (poor prediction).

The **Classifier Chains method** [Read et al., 2009][2]: Pass information between binary classifiers.

- $\hat{y}_j = h_j(\mathbf{x}, \hat{y}_1, \ldots, \hat{y}_{j-1})$; e.g. $?^{forest} = h_3(\mathbf{x}, 1^{beach}, 0^{urban})$
- improves prediction, and approximately as fast

[2] J. Read, B. Pfahringer, G. Holmes, E. Frank. Classifier Chains for Multi-label Classification. In Proc. of European Conference on Machine Learning. 2009.

Main Contribution 2: The Classifier Chains Method

The Pruned Sets method worked well, but had limitations:

- difficulty dealing with 'extreme' datasets; and
- worst-case time same as the Label Combination method.

The Binary Relevance method (L separate problems; $\hat{y}_j = h_j(\mathbf{x})$):

- Relatively robust and good theoretical time complexity; but
- does not explicitly model label correlations (poor prediction).

The **Classifier Chains method** [Read et al., 2009][2]: Pass information between binary classifiers.

- $\hat{y}_j = h_j(\mathbf{x}, \hat{y}_1, \ldots, \hat{y}_{j-1})$; e.g. $?^{forest} = h_3(\mathbf{x}, 1^{beach}, 0^{urban})$
- improves prediction, and approximately as fast

Ensembles of Classifier Chains:

- chain order not an issue (random)
- highly competitive

[2] J. Read, B. Pfahringer, G. Holmes, E. Frank. Classifier Chains for Multi-label Classification. In Proc. of European Conference on Machine Learning. 2009.

Contributions: an example

"*Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains*" [Cheng et al., 2010][3]

- "inspired by the *classifier chain* (CC) ... by [Read et al., 2009]"
- Probabilistic Classifier Chains (PCC): a Bayes optimal way of forming classifier chains. $\mathbf{P}_{\mathbf{x}}(\mathbf{y}) = \prod_{j=1}^{L} h_j(\mathbf{x}, y_1, \ldots, y_{j-1})$.
- some improvement over CC, but ...

[3]Weiwei Cheng and Krzysztof Dembczyński and Eyke Hüllermeier, Bayes optimal multilabel classification via probabilistic classifier chains. 27th International Conference on Machine Learning. 2010

Contributions: an example

"*Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains*" [Cheng et al., 2010][3]

- "inspired by the *classifier chain* (CC) ... by [Read et al., 2009]"
- Probabilistic Classifier Chains (PCC): a Bayes optimal way of forming classifier chains. $\mathbf{P}_{\mathbf{x}}(\mathbf{y}) = \prod_{j=1}^{L} h_j(\mathbf{x}, y_1, \ldots, y_{j-1})$.
- some improvement over CC, but ...
- "PCC has to look at each of the 2^L paths ... which limits applicability to data sets with not more than ... about 15 labels" (they use 10 — ECC deals with 1000 in thesis).
- "the averaging method [of ECC] brings the predictions to the marginals"; "overall good performance of ECC".

[3] Weiwei Cheng and Krzysztof Dembczyński and Eyke Hüllermeier, Bayes optimal multilabel classification via probabilistic classifier chains. 27th International Conference on Machine Learning. 2010

Contributions: a summary

General contributions:

- most extensive empirical analysis in the multi-label literature
- large and varied dataset collection (+ three *new* datasets)
- multiple evaluation measures (+ introduced *log loss*)
- an open-source framework (MEKA[4])

[4] Multi-label wEKA: `http://meka.sourceforge.net/`

Contributions: a summary

General contributions:

- most extensive empirical analysis in the multi-label literature
- large and varied dataset collection (+ three *new* datasets)
- multiple evaluation measures (+ introduced *log loss*)
- an open-source framework (MEKA[4])

Major contributions:

1. Pruned Sets
2. Classifier Chains

- both of which are: scalable, generally applicable, and competitive (with state-of-the-art methods).
- building blocks for other methods, as demonstrated by
  1. Ensembles of Pruned Sets
  2. Ensembles of Classifier Chains
- and have already had impact in the literature: e.g. [Cheng et al., 2010, Zhang and Zhang, 2010].

[4] Multi-label wEKA: `http://meka.sourceforge.net/`

Multi-label Data Streams

Data Streams

- data instances typically arrive continually and rapidly (data labelling often generated by machine)
- update model and predict in real time
- concept drift

Applications

- sensor data
- transactions (e.g. ATM, online)
- network traffic

[5] Massive Online Analysis: `http://moa.cs.waikato.ac.nz/`

Multi-label Data Streams

Data Streams

- data instances typically arrive continually and rapidly (data labelling often generated by machine)
- update model and predict in real time
- concept drift

Applications

- sensor data
- transactions (e.g. ATM, online)
- network traffic

Methods

- MOA[5] framework: existing (single-label) incremental classifiers; concept-change-detection methods, now extended with multi-label classifiers, e.g. Multi-label Hoeffding Tree Classifier with Pruned Sets at the leaves [Read et al., 2010]

---

[5] Massive Online Analysis: `http://moa.cs.waikato.ac.nz/`

Related Tasks

- tag/keyword-assignment
  - more labels, not predefined, more descriptive than categorical

    e.g. $\mathbf{x} \rightarrow$ truck, 4wd, snowing, mountain, cold, trees, fence
- label ranking
  - labels are associated with a rank / real value; $\mathbf{y} \in \mathbb{R}^L$

    e.g. given $\mathcal{Y} = \{beach, people, forest, mountain\}$; $\mathbf{x} \rightarrow [0.7, 0.0, 0.1, 0.2]$
- multi-task learning
  - learning a problem together with other related problems
- transfer learning
  - applying knowledge from one problem to a related problem
- structured outputs
  - labels are structured in some way: graph, hierarchy, coord.s, masks, mappings, bounding boxes, angles, etc.

    e.g. $\mathbf{x} \rightarrow (bird).sits\_on.(truck)$; $\rightarrow bird@[x, y, z]$; $\rightarrow fence@[x_1, y_1][x_2, x_3]$

The End

Thank you for your attention.

References:

Cheng, W., Dembczyński, K., and Hüllermeier, E. (2010).
Bayes optimal multilabel classification via probabilistic classifier chains.
In *ICML '10: 27th International Conference on Machine Learning*, Haifa, Israel. Omnipress.

Read, J., Bifet, A., Holmes, G., and Pfahringer, B. (2010).
Efficient multi-label classification for evolving data streams.
Technical report, University of Waikato, Hamilton, New Zealand.
Working Paper 2010/04.

Read, J., Pfahringer, B., and Holmes, G. (2008).
Multi-label classification using ensembles of pruned sets.
In *ICDM'08: Eighth IEEE International Conference on Data Mining*, pages 995–1000. IEEE.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2009).
Classifier chains for multi-label classification.
In *ECML '09: 20th European Conference on Machine Learning*, pages 254–269. Springer.

Zhang, M.-L. and Zhang, K. (2010).
Multi-label learning by exploiting label dependency.
In *KDD '10: 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 999–1008.