MIProblems – Datasets for Multiple Instance Learning

Veronika Cheplygina

This file provides the descriptions of datasets previously stored at miproblems.org. As I am now longer maintaining the website, I moved the datasets to Figshare.

If you use these datasets, please cite our paper:

Cheplygina, V., Tax, D. M., & Loog, M. (2015). Multiple instance learning with bag dissimilarities. *Pattern Recognition*, *48*(1), 264-275.

```
@article{cheplygina2015multiple,
  title={Multiple instance learning with bag dissimilarities},
  author={Cheplygina, Veronika and Tax, David MJ and Loog, Marco},
  journal={Pattern Recognition},
  volume={48},
  number={1},
  pages={264--275},
  year={2015},
  publisher={Elsevier}
}
```

You might also be interested in this page of results of MIL classifiers: http://homepage.tudelft.nl/n9d04/milweb/index.html

Biocreative

Description

Biocreative is a text categorization problem. From the README file: the task is to decide whether a given pair should be annotated with some Gene Ontology (GO) code. As input, we have paragraphs of documents, each paragraph described by a feature vector. Features used are word

occurrence frequencies and some statistics about the nature of the protein-GO code interaction for each paragraph. Each document corresponds to a bag and each paragraph to an instance in a bag. The hypothesis is that a bag should be annotated with a GO code iff there exists a paragraph in it that supports this annotation. Conversely, if no paragraph supports such an annotation, the document should not be annotated.

Original source

The original data that this dataset is based on can be found here: http://www.biocreative.org/tasks/biocreative-i/task-2-functional-annotations/. This dataset has been represented as a MIL problem (in C4.5 format) by Dr. Soumya Ray. For more details about the creation of the dataset please refer to:

```
@article{ray2005learning,
title={Learning statistical models for annotating proteins with function
information using biomedical text},
author={Ray, Soumya and Craven, Mark},
journal={BMC bioinformatics},
volume={6},
number={Suppl 1},
pages={S18},
year={2005},
publisher={BioMed Central Ltd}
}
```

```
@inproceedings{ray2005supervised,
title={Supervised versus multiple instance learning: An empirical comparison},
author={Ray, Soumya and Craven, Mark},
booktitle={Proceedings of the 22nd international conference on Machine learning},
pages={697--704},
year={2005},
organization={ACM}
}
```

The data was then converted to Matlab format with a parser by Gary Doran.

Files

biocreative.zip – This file contains three different .MAT files for different tasks (component, function, process) in the dataset. Each .MAT file contains a training and a test set.

Birds

Description

The bird songs problem is originally a multi-label, multi-class problem. Each bag is a recording of one or more birds. The bag inherits all the labels of the birds present in the recording. This can be converted to a conventional binary MI problem by choosing a bird “target class”. The bird classes are:

BRCR – Brown Creeper
WIWR – Winter Wren
PSFL – Pacific-slope Flycatcher
RBNU – Red-breasted Nuthatch
DEJU – Dark-eyed Junco
OSFL – Olive-sided Flycatcher
HETH – Hermit Thrush
CBCH – Chestnut-backed Chickadee
VATH – Varied Thrush
HEWA – Hermit Warbler
SWTH – Swainson's Thrush
HAFL – Hammond's Flycatcher
WETA – Western Tanager

Each bag (10 second recording) is converted to a spectrogram, and a segmentation procedure is applied. An instance is represented by a segment of a spectrogram and is described by 38 features (shape of the segment, time and frequency profile statistics, histogram of gradients).

Original source

Thanks to Forrest Briggs for his permission to distribute these datasets.

BibTeX:

```
@inproceedings{briggs2012rank,
 title={Rank-loss support instance machines for MIML instance annotation},
 author={Briggs, F. and Fern, X.Z. and Raich, R.},
 booktitle={Proceedings of the 18th ACM SIGKDD international conference on
Knowledge discovery and data mining},
 pages={534--542},
 year={2012},
 organization={ACM}
 }
```

Files

birds.zip – This file contains a MIL dataset x with 13 different label lists. The default label list is for Brown Creeper, but you can switch to a different version of the dataset by doing (use the abbreviations in the list above):

```
x = changelablist(x,'WIWR');
```

You need the MIL toolbox to load this version of the dataset correctly. If you do not want to use the toolbox, just load the .MAT file. You can access the data and the label lists by:

```
data=x.data;

labels=x.nlab;  %Get the k'th label list by labels(:,k)
```

Corel

Description

Corel is a 20-class image classification problem. One of the classes is assigned to be the positive class. The classes are:

'African', 'Horses', 'Cars', 'Beach', 'Mountains', 'Waterfalls', 'Historical', 'Food', 'Antique', 'Buses', 'Dogs', 'Battleships', 'Dinosaurs', 'Lizards', 'Skiing', 'Elephants', 'Fashion', 'Desserts', 'Flowers', 'Sunset'

Original source

Thanks to professor James Wong for permission to distribute this version of the dataset. The data (including the thumbnails of the images) can be found here.

The related publication is:

```
@article{chen2006miles,
title={MILES: Multiple-instance learning via embedded instance selection},
author={Chen, Yixin and Bi, Jinbo and Wang, James Z},
journal={Pattern Analysis and Machine Intelligence, IEEE Transactions on},
volume={28},
number={12},
pages={1931--1947},
year={2006},
publisher={IEEE}
}
```

Each bag is an image, and the instances are image segments. Each segment is represented by the mean of the 4×4 patch features:

1. three average LUV color components

2. three (sqrt) energy components in the high frequency bands of the wavelet transform

3. three shape components with normalized inertia of order 1,2,3

Files

Files

corel.zip – This file contains a MIL dataset x with 20 different label lists. The default label list is for African, but you can switch to a different version of the dataset by doing (use the labels above):

```
x = changelablist(x,'Cars');
```

You need the MIL toolbox to load this version of the dataset correctly. If you do not want to use the toolbox, just load the .MAT file. You can access the data and the label lists by:

```
data=x.data;

labels=x.nlab;  %Get the k'th label list by labels(:,k)
```

Fox, Tiger, Elephant

Description

There are three datasets, Fox, Tiger and Elephant. The bags are images, and the instances are image segments. For each category, positive bags are images that contain the animal, and negative bags are images that contain other animals (also from other categories, not just from the three categories here).

Original source

BibTeX:

```
@article{andrews2002support,
  title={Support vector machines for multiple-instance learning},
  author={Andrews, Stuart and Tsochantaridis, Ioannis and Hofmann, Thomas},
  journal={Advances in neural information processing systems},
  volume={15},
  pages={561--568},
  year={2002}
}
```

The extracted features for the data were obtained here

Files

fte.zip– This file contains three different .MAT files for the Fox, Tiger and Elephant problems. You need the MIL toolbox to load this version of the dataset correctly.

Messidor

Description

Messidor is an image classification problem. The data consists of 1200 eye fundus images from 654 diabetes and 546 healthy patients. Each image from the original data is rescaled to 700×700 pixels and split up into patches of 135×135 pixels. Patches which do not have a sufficient amount of foreground are discarded. The features used are: intensity histogram of RGB channels for 26

bins, mean of local binary pattern histograms of 20×20 pixel grids, mean of SIFT descriptors, and box count for grid sizes 2 to 8. Some of the features return NaNs, replacing by zero is advised.

Original source

The original data is kindly provided by the Messidor program partners
(see http://messidor.crihan.fr/download.php).

This dataset has been represented as a MIL problem by Dr. Melih Kandemir.

When using the dataset, please cite the following papers:

```
@article{decenciere2014feedback,
title={Feedback on a publicly distributed image database: the {M}essidor
database},
author={Decenci{\`e}re, Etienne and Zhang, Xiwei and Cazuguel, Guy and Lay, Bruno
and Cochener, B{\'e}atrice and Trone, Caroline and Gain, Philippe and Ordonez,
Richard and Massin, Pascale and Erginay, Ali and Charton, B{\'e}atrice and Klein,
Jean-Claude},
journal={Image Analysis and Stereology},
pages={231--234},
year={2014}
}
```

```
@article{kandemir2014computer,
title={Computer-aided diagnosis from weak supervision: A benchmarking study},
author={Kandemir, Melih and Hamprecht, Fred A},
journal={Computerized Medical Imaging and Graphics, in press},
year={2014},
doi={10.1016/j.compmedimag.2014.11.010}
}
```

Files

messidor.zip

Musk

Description

There are two datasets, Musk1 and Musk2. Both are about predicting whether a molecule has a musky smell or not. A molecule is described by the different shapes it can fold into (conformers), each bag corresponds to a molecule and each instance to one of its conformers. Conformers are responsible for the properties of a molecule, i.e. its smell. If at least one of the conformers can cause a molecule to smell musky, the molecule is positive for the musky class. If none of the conformers have this property, the molecule is negative.

Original source

BibTeX:

```
@article{dietterich1997solving,
  title={Solving the multiple instance problem with axis-parallel rectangles},
  author={Dietterich, Thomas G and Lathrop, Richard H and Lozano-P{\'e}rez,
Tom{\'a}s},
  journal={Artificial Intelligence},
  volume={89},
  number={1},
  pages={31--71},
  year={1997},
  publisher={Elsevier}
}
```

Files

musk.zip – This file contains two different .MAT files for the Musk1 and Musk2 problems. You need the MIL toolbox to load this version of the dataset correctly.

Mutagenesis

Description

Mutagenesis is a drug activity prediction problem. There are two versions, easy (1) and hard (2), of the dataset.

Original source

BibTeX:

```
@inproceedings{srinivasan1995comparing,
  title={Comparing the use of background knowledge by inductive logic programming
systems},
  author={Srinivasan, Ashwin and Muggleton, S and King, RD},
  booktitle={Proceedings of the 5th International Workshop on Inductive Logic
Programming},
  pages={199--230},
  year={1995}
}
```

This dataset has been converted to a MIL problem in ARFF format by Dr. Frank Eibe of http://www.cs.waikato.ac.nz/~ml/

Files

mutagenesis.zip – This file contains two different .MAT files for the Easy and Hard problems. You need the MIL toolbox to load this version of the dataset correctly.

Newsgroups

Description

Newsgroups is a text categorization dataset. A bag is a collection of posts from different newsgroups. There are 20 categories in total. A positive bag for the target category is generated to contain 3% posts from the target category, and 97% of posts, randomly sampled from the other categories.

Source

Thanks to professor Zhi-Hua Zhou for allowing us to provide this data.

BibTeX entry:

```
@inproceedings{zhou2009multi,
  title={Multi-instance learning by treating instances as non-IID samples},
  author={Zhou, Zhi-Hua and Sun, Yu-Yin and Li, Yu-Feng},
  booktitle={Proceedings of the 26th Annual International Conference on Machine Learning},
  pages={1249--1256},
  year={2009},
  organization={ACM}
}
```

Files

newsgroups.zip – This file contains 20 .mat files, one for each target category.

UCSB Breast

Description

UCSB Breast is an image classification problem. The original datasets consists of 58 TMA image excerpts (896 × 768 pixels) taken from 32 benign and 26 malignant breast cancer patients. The learning task is to classify images as benign (negative) or malignant (positive).

Patches of 7×7 size are extracted. The image is thresholded to segment the content from the white background and the patches that contain background more than 75% of their area are

discarded. The features used are 657 features that are global to the patch (histogram, LBP, SIFT), and averaged features extracted from the cells, detected in each patch.

Original source

The original data that this dataset is based on can be found here: http://www.bioimage.ucsb.edu/research/biosegmentation. This dataset has been represented as a MIL problem by Dr. Melih Kandemir. When using the dataset, please cite the following paper:

```
@inproceedings{kandemir2014empowering,
title={Empowering Multiple Instance Histopathology Cancer Diagnosis by Cell Graphs},
author={Kandemir, Melih and Zhang, Chong and Hamprecht, Fred A},
booktitle={MICCAI},
year={2014}
}
```

See also the PDF and the code used in the paper.

Files

ucsb_breast.zip

Web recommendation

Description

The problem is to classify webpage as interesting or not. In total, 9 users rate webpages as such, therefore there are 9 different datasets. A webpage is a bag, and the links on the webpage are the instances. The features are related to word frequency (and therefore very high-dimensional).

Original source

Thanks to professor Zhi-Hua Zhou for allowing us to provide this data.

```
@article{zhou2005multi,
  title={Multi-instance learning based web mining},
  author={Zhou, Zhi-Hua and Jiang, Kai and Li, Ming},
  journal={Applied Intelligence},
  volume={22},
  number={2},
  pages={135--147},
  year={2005},
  publisher={Springer}
}
```

Files

web.zip – This file contains 9 .MAT files, each corresponding to a different user. Each .MAT file has a training set x and testing set z. It is possible to concatenate the train and test set as follows:

```
data = [x;z];
```