Multi Instance Multi Label

Lorenzo Niccolai
lorenzo.niccolai3@stud.unifi.it

Fabio Vittorini
fabio.vittorini@stud.unifi.it

Machine Learning

University of Florence, Department of Information Engineering

July 15, 2017

Overview

Binary classification

Goal: To produce a classifier able to decide whether an object belongs to one or more classes.

Idea: Supervised Learning: Given a dataset of already classified examples, the classifier *learns* a function that solves classification problem.

Notation

- A vector $x \in \mathbb{R}^f$ represents an object using f *relevant* features.
- A vector $y \in \{-1, +1\}^l$ indicates wether the example belong to each of the l label classes.

The input of a classification problem is a dataset $D = \{X, Y\}$ where $X \in \mathbb{R}^{x \times f}$ is a set of examples and $Y \in \mathbb{R}^{n \times l}$ is a set of labels.
While learning the target function, the dataset is divided in *training set* and *test set*.

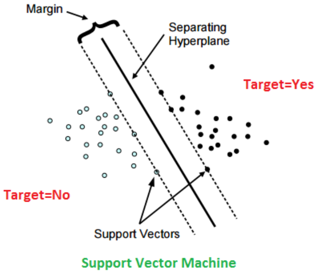
SVM idea

- For 1-class problems we have to compute the *maximum-margin hyperplane* $w^T x + b$ which best separates positive examples from negative examples.

Optimization problem is:

$$argmin_w \frac{1}{2}||w||^2$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 \ \forall i \in [1, n]$$



SVM idea

- The examples may not be linearly serparable and so the problem would not have any solutions because constraints are not satisfied. Then we introduce slack variables ξ

Optimization problem becames:

$$argmin_{w,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1} n\xi^{(i)}$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi^{(i)} \ \forall i \in [1, n]$$

$$\xi^{(i)} \geq 0 \ \forall i \in [1, n]$$

Multi instance classification

Motivation:

- Sometimes a complex item can be well represented by a set of *instances*
- A single instance may belong or not to a class
- An example is positive if at least one of its instances is positive, it's negative otherwise
- Dataset labels are assigned to examples, not to instances
- We have a *semi-supervised learning* problem

Notation

Dataset is now a set of bags, where each bag is a set of instances:

$$D = \{(X^{(i)}, Y^{(i)}) | i \in [1, n]\}$$

$$X^{(i)} = \{x_k^{(i)} | k \in [1, k_i], x_k \in \mathbb{R}^f\}$$

Notice that each bag can be made of any number of instances, but every instance has a fixed number of features f.

SIL

The first naive approach makes the following label assignment:

- If an instance belongs to a negative bag, sets its label to -1
- If an instance belongs to a positive bag, sets its label to $+1$

The resulting problem can be solved using a regular SVM, treating each instance as a whole document.

Using this approach makes almost useless multi-instance formulation.

mi-SVM

Instances label assignment:

- If an instance belongs to a negative bag we can say that its label is -1
- If an instance belongs to a positive bag we don't know for sure its label

This leads to 2 new constraints in SVM problem:

$$y_k^{(i)} = -1 \ if \ \ Y^{(i)} = -1$$

$$\sum_{k=1}^{k_i} \frac{y_k^{(i)} + 1}{2} \geq 1 \ if \ \ Y^{(i)} = +1$$

mi-SVM

Our SVM problem becames the following:

$$min_Y \, min_{w,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi^{(i)}$$

$$y_k^{(i)}(w^T x_k^{(i)} + b) \geq 1 - \xi^{(i)} \; \forall i \in [1, n], k \in [1, k_i]$$

$$\xi^{(i)} \geq 0 \; \forall i \in [1, n]$$

$$y_k^{(i)} = -1 \; if \;\; Y^{(i)} = -1$$

$$\sum_{k=1}^{k_i} \frac{y_k^{(i)} + 1}{2} \geq 1 \; if \;\; Y^{(i)} = +1$$

That is an intractable mixed optimization problem

Algorithm

A feasible algorithm that finds a non optimal solution is the following:

MI-SVM (X, Y)

$y_k^{(i)} = -1$ *if* $Y^{(i)} = -1$
$y_k^{(i)} = +1$ *if* $Y^{(i)} = +1$
do
    Solve regular SVM finding w, b
    $y_k^{(i)} = sign(w^T x_k^{(i)} + b)$ *if* $Y^{(i)} = +1$
    Adjust each positive bag to satisfy constraints
while ($y_k^{(i)}$ change)

MI-SVM

This approach uses directly the dataset in its bag form:

$$argmin_{w,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi^{(i)}$$

$$y^{(i)}(max_k w^T x_k^{(i)} + b) \geq 1 - \xi^{(i)} \ \forall i \in [1, n]$$

$$\xi^{(i)} \geq 0 \ \forall i \in [1, n]$$

This is possible by selecting a *witness* from each bag instances.

Algorithm

A feasible algorithm that finds a solution is the following:

$\text{MI-SVM}\,(X, Y)$

```
x_s^(i) = avg(x_k^(i)) ∀i ∈ [1, n]
do
    Solve regular SVM finding w, b, balancing lagrange multipliers
    Find new x_s^(i) by selecting the best one for each positive bag
while (witnesses change)
```

Multi label classification

Motivation:

- Sometimes a complex item can be well represented by a set of *labels*
- Helps single label classification when the concept is more complicated or general

Solutions:

- Problem transformation
- Algorithm adaptation

Notation

A set of labels $L = \{y_1, y_2, ...y_l\}$ is given.
Each object contained in the dataset is associated with a set of labels:

$$D = \{(X_i, Y_i | i \in [1, n]\}$$

$$X_i \in \mathbb{R}^f$$

$$Y_i = \{y_{i,h} | h \in [1, h_i], y_{i,h} \in L, h_i \leq l\}$$

Problem transformation

Attempt to convert the multilabel problem in a regular binary task.
Two lossy methods:

- Randomly discard each label information except one from each instance
- Remove instances that have actually more than one label

Other solutions:

- Train a binary classifier for each existing combination of labels
- Train a binary classifier for each label (used in this work)

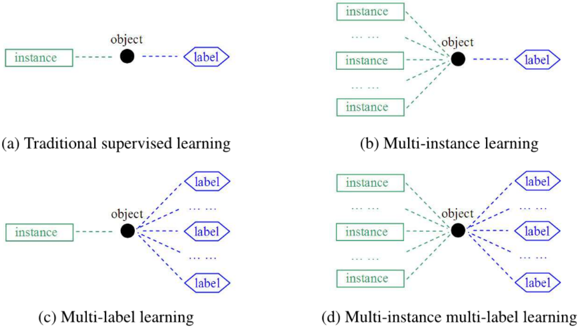
Algorithm adaptation

Regular algorithms are modified to support multi-label tasks.
Sometimes they use problem transformation at the core.
An example using SVM-related approach based on ranking and label set size prediction.

MIML problems combine motivations of multi instance and multi label ones.
Given a set of labels $L = \{y_1, y_2, ...y_l\}$

$$X_i = \{x_{i,k} | k \in [1, k_i], x_{i,k} \in \mathbb{R}^f\} Y_i = \{y_{i,h} | h \in [1, h_i], y_{i,h} \in L, h_i \leq l\}$$

$$D = \{(X_i, Y_i) | i \in [1, n]\}$$



(a) Traditional supervised learning

(b) Multi-instance learning

(c) Multi-label learning

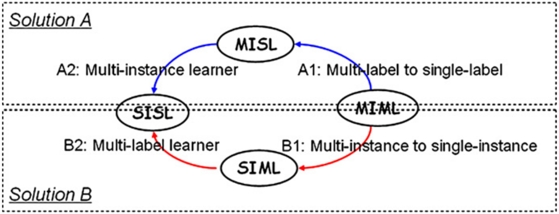
(d) Multi-instance multi-label learning

SVM Solution

To allow regular SVMs to solve this problem, we use *problem transformation*.

There are 2 possibilities:

- MIML → MISL → SISL (used in this work)
- MIML → SIML → SISL



Multi label to single label

Excluding lossy approaches, the idea is to train a multi-instance (single label) classifier for each label.
Given a MIML dataset $D = \{(X_i, Y_i) | i \in [1, n]\}$ we produce l datasets as follows:

$$D_{y_j} = \{(X_i, Y_{y_j}) | i \in [1, n]\} \; \forall j \in [1, L]$$

Where

$$Y_{y_j} = \begin{cases} +1 \; if \;\; y_j \in Y_i \\ -1 \; otherwise \end{cases}$$

Then we train L regular multi-instance SVMs and collect their results.

Multi instance to single instance

Given one of MISL datasets produced at previous step, we compared the 3 methods previously exposed:

- SIL
- MI-SVM
- mi-SVM

They all use a standard SISL SVM as subroutine.