Multilabel Classification Techniques

Amit Nandi(CMS1503)[1]
guided by
Prof. VK Jayaraman
and
Dr.B.S.Pujari

[1]Center for Modeling and Simulation
University of pune

Mid-sem Final-project Presentation, 2017

Outline

Outline

What is Multilabel Data?

Introduction

Single label classification: Is this a picture of a beach?

$$\epsilon\{yes, no\}$$

Multi label classification[1]**:** Which labels are relevant to this picture?

$$\subseteq \{beach, foilage, field, mountain\}$$

i.e. each example could belongs to more than one label

So, for a given image we could express above problem into below format:

X	$Beach$	$Foilage$	$Field$	$Mountain$
$Ex1$	1	0	0	1

[1] Multi-label classification by Jasse Read

Multilabel data

Types of Method

1. Problem Transformation Method[2]
2. Algorithm Adaptation Method

[2] Random k-Labelsets for Multi-Label Classification

Project Focus Area

Problem Transformation Method

- Problem Transformation Method.
- Project work involves learning and implementing various algorithm to tackle given data and produce result to validate the algorithm.

Outline

Literature Survey

1. Major Algorithm[3]:
   - Binary Relevance
     a) Chain Classifier
     b) Two Level classifier
   - RAkEL
     a) Disjoint
     b) Overlap
   - Label Power-set

[3] Tsoumakas and Katakis, 2007;

Outline

Frame Work

Algorithms - Examples

Ex1: Beach

Ex2: Foliage

Ex3: Beach + Mountain

Ex4: Field + Mountain

Frame Work

Algorithms - Binary Relevance

Creates L separate binary problems (L = no. of labels)

Generate 4 models for our 4 label problem (considering a single label at a time)

	Beach	Foliage	Field	Mountain
Ex1	1	0	0	0
Ex2	0	1	0	0
Ex3	1	0	0	1
Ex4	0	0	1	1

	Beach
Ex1	1
Ex2	0
Ex3	1
Ex4	0

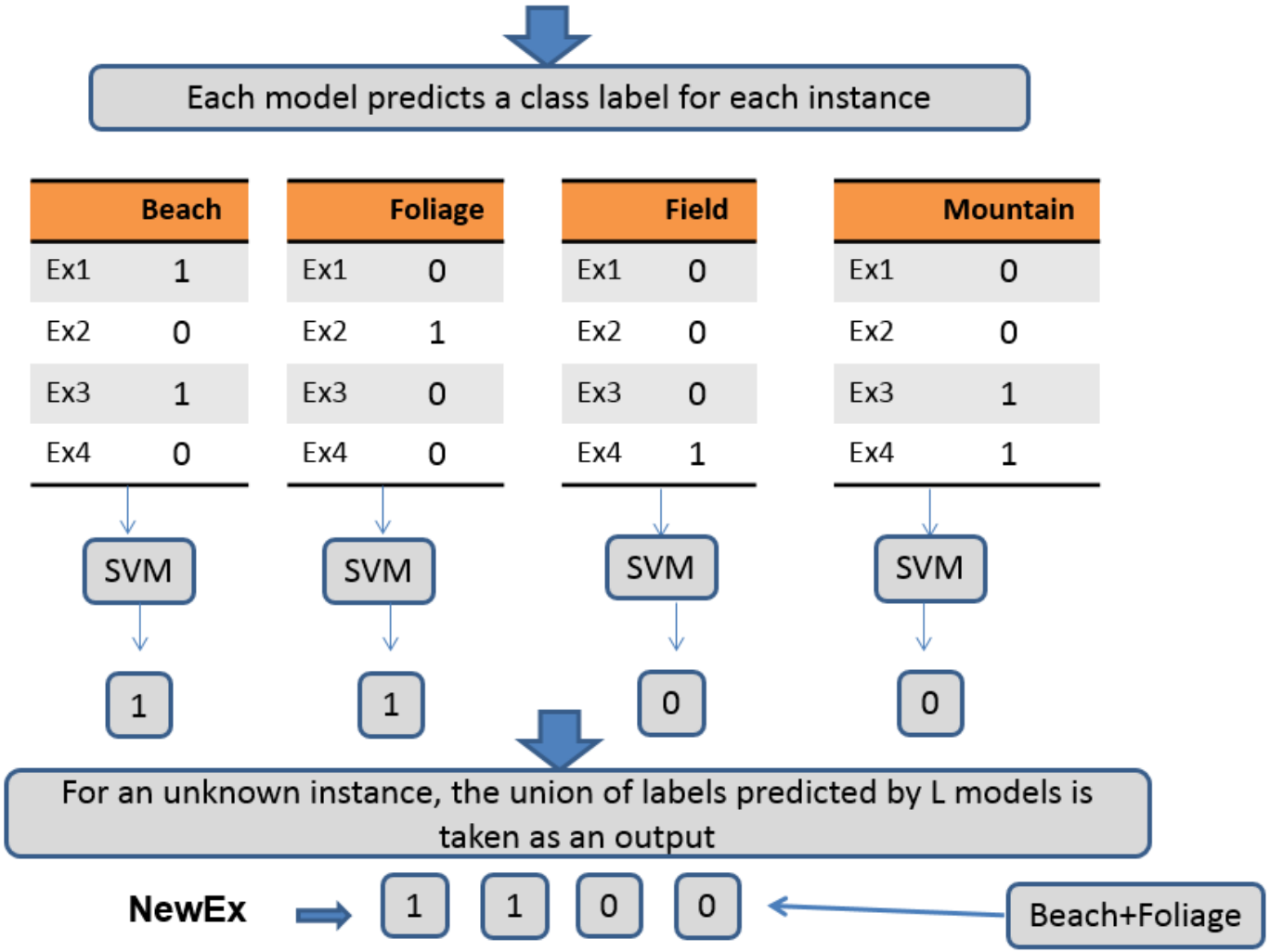
	Foliage
Ex1	0
Ex2	1
Ex3	0
Ex4	0

	Field
Ex1	0
Ex2	0
Ex3	0
Ex4	1

	Mountain
Ex1	0
Ex2	0
Ex3	1
Ex4	1

Frame Work

Algorithms - Binary Relevance



Frame Work

Algorithms - Label Powerset

Here, every distinct labelset in the original multi-label data is considered as a new class

The multi-label problem is converted into a multiclass (single label) problem with 2^L possible class values (L = no. of labels)

	Beach	Foliage	Field	Mountain
Ex1	1	0	0	0
Ex2	0	1	0	0
Ex3	1	0	0	1
Ex4	0	0	1	1

Ex1	1000
Ex2	0100
Ex3	1001
Ex4	0011

Beach

Foliage

Beach+Mountain

Field+Mountain

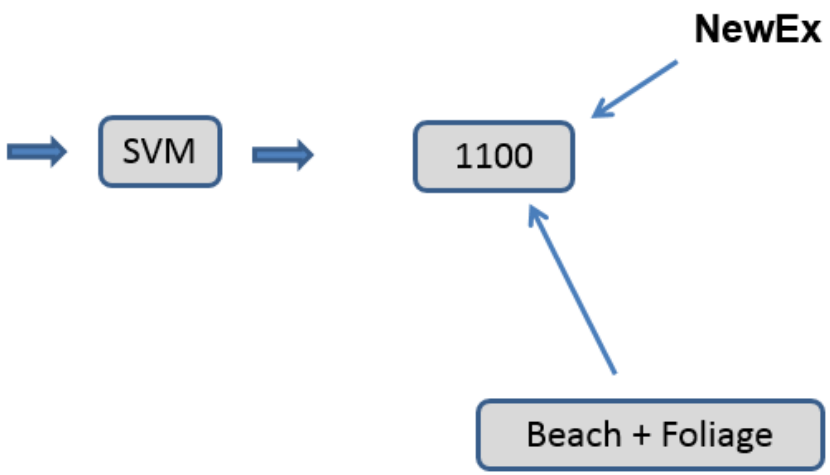
Frame Work

Algorithms - Label Powerset

Then use SVM for training the transformed multiclass data and generating models

For an unknown instance, a class label from multiclass data is predicted which is then mapped back as a set of labels in the original data

Ex1	1000
Ex2	0100
Ex3	1001
Ex4	0011



Frame Work

Algorithms - RAkEL

Randomly select a value of k (size of a label subset) and build *m* Label-Powerset classifiers

2^k problems are considered at a time rather than 2^L problems like in LP (L = no. of labels)

For k = 2 and L = 4 ⇒ m = 6 models

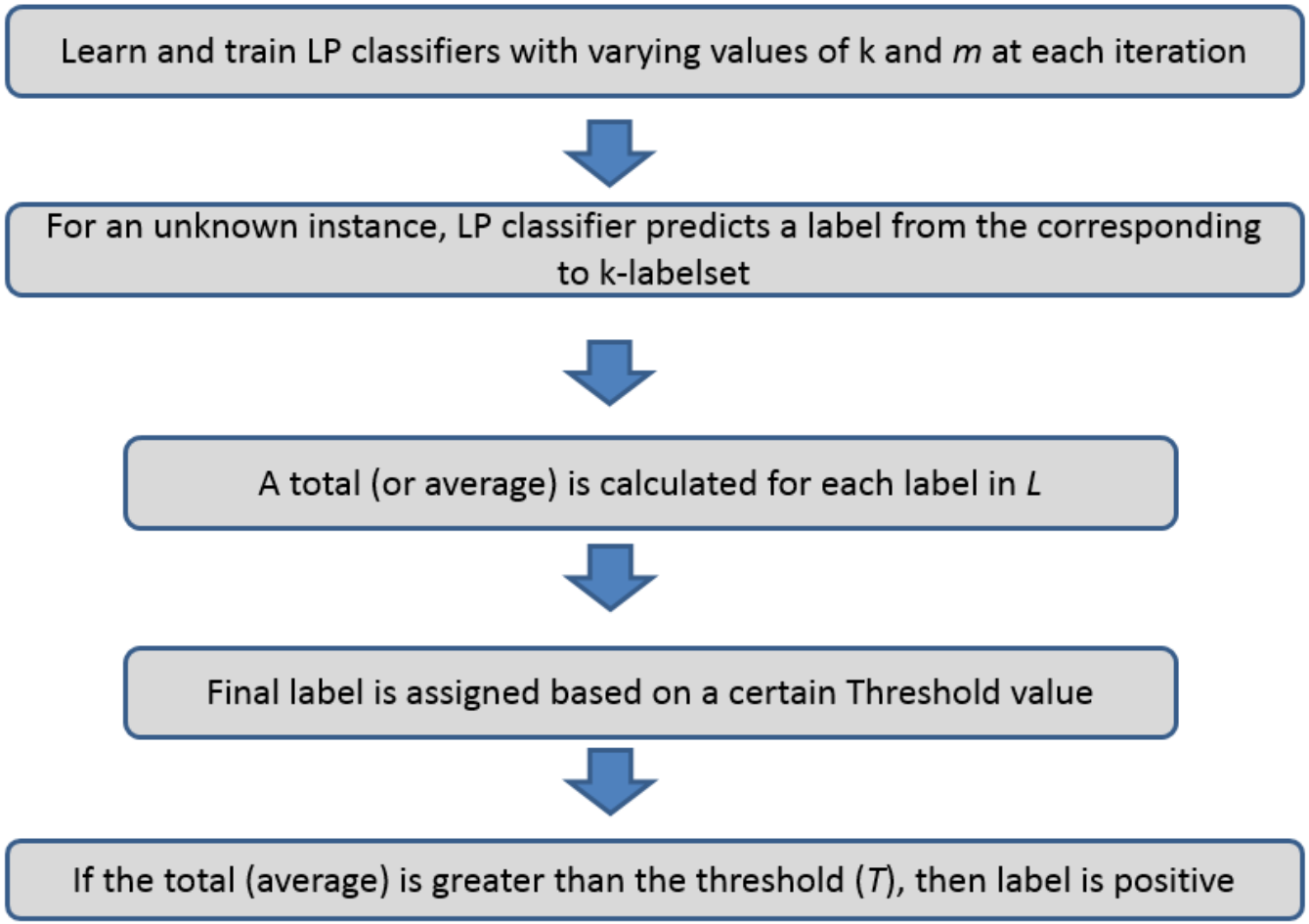
	B	Fo	Fi	M
Ex1	1	0	0	0
Ex2	0	1	0	0
Ex3	1	0	0	1
Ex4	0	0	1	1

⇒

Ex1	10	Ex1	10	Ex1	10	Ex1	00	Ex1	00	Ex1	00
Ex2	01	Ex2	00	Ex2	00	Ex2	10	Ex2	10	Ex2	00
Ex3	10	Ex3	10	Ex3	11	Ex3	00	Ex3	01	Ex3	01
Ex4	00	Ex4	01	Ex4	01	Ex4	01	Ex4	01	Ex4	11

Frame Work

Algorithms - RAkEL



Frame Work

Algorithms - RAkEL

NewEx

	B	Fo	Fi	M
Model1	1	0	0	0
Model2	0	1	0	0
Model3	1	1	0	0
Model4	1	1	1	0
Model5	0	0	0	1
Model6	1	1	1	0
SUM	4	4	2	1

For $T = 3$

1 1 0 0

Beach + Foliage

Outline

Frame Work

Performance Measures

Multilabel Algorithm performance is evaluated with below mentioned measures:

1. Accuracy
   - 1.1 Denotes proportion of Correctly predicted class to total number of class
2. Precision
   - 2.1 Denotes proportion of predicted correct labels to total number of Actual labels, averaged over all instances.
3. Recall
   - 3.1 Denotes proportion of predicted correct labels to total number of Predicted labels
4. F1-measure
   - 4.1 Harmonic mean of Precision and Recall.
5. Hamming Loss
   - 5.1 It is a loss function which calculates the proportion of misclassified labels to the total number of labels

Thank you