Common multilabel formulae

...under a common notation

David Charte

Abstract

Multilabel Classification is a branch of Data Mining in which many different metrics and equations are defined. Each author uses different notation and basic definitions, thus difficulting the coherence of future papers where some of these formulas are used. This is a document where the multilabel scenario is presented with simple basic definitions, and a list of commonly used formulas are rigurously adapted to this notation, in an attempt to introduce them in a clear way.

Definitions

Let $A_1, A_2, \dots A_f$ be arbitrary sets. We will call them *input attributes* or simply *attributes*. An instance will take a certain value on each of these sets, that is, we will be working with elements of their cartesian product, $A_1 \times A_2 \times \cdots \times A_f$.

Let L be a finite set. This will be the set of all possible labels. Each instance of a dataset will then have a subset of active labels or *labelset*, $y \subset L$.

Let D be a finite subset of $A_1 \times A_2 \times \cdots \times A_f \times \mathcal{P}(L)$, where $\mathcal{P}(L)$ is the powerset of L, that is, the set of all possible combinations of labels. We will call D a *dataset* and each of its elements an *instance*: $(x, y) \in D$ where $x = (x_1, x_2, \dots x_f) \in A_1 \times A_2 \times \cdots \times A_f$ and $y = \{l_1, \dots l_k\} \in \mathcal{P}(L)$.

Note: Since D is a set, we will be assuming no two instances are identical.

Basic measures

In the following we will assume D is a fixed set. Otherwise, many of the measures would be dependent on D. Some basic data directly extracted from the definitions are:

- Number of input attributes, f
- Number of labels, $|L|$
- Number of instances, $|D|$

- Number of distinct labelsets, $|\{Y : (X, Y) \in D\}|$

Many of the measures specific to multilabel classification are related to the labels themselves. For instance, we can find out the mean number of labels in a labelset (*Card*, 1) and its proportion as to the total number of labels (*Dens*, 2):

$$Card = \frac{1}{|D|} \sum_{(x,y) \in D} |y| \tag{1}$$

$$Dens = \frac{1}{|D|\,|L|} \sum_{(x,y) \in D} |y| \tag{2}$$

Imbalance and concurrence