
# Hierarchical partitioning of the output space in multi-label data

Yannis Papanikolaou, Grigorios Tsoumakas, Ioannis Katakis

PII: S0169-023X(17)30451-2

DOI: 10.1016/j.datak.2018.05.003

Reference: DATAK 1652

To appear in: *Data & Knowledge Engineering*

Received Date: 28 September 2017

Revised Date: 19 February 2018

Accepted Date: 5 May 2018

Please cite this article as: Y. Papanikolaou, G. Tsoumakas, I. Katakis, Hierarchical partitioning of the output space in multi-label data, *Data & Knowledge Engineering* (2018), doi: 10.1016/j.datak.2018.05.003.

# Hierarchical Partitioning of the Output Space in Multi-label Data

Yannis Papanikolaou[a,*], Grigorios Tsoumakas[a], Ioannis Katakis[b]

[a] *School of Informatics, Aristotle University of Thessaloniki, Greece.*
[b] *Computer Science Department, University of Nicosia, Cyprus*

**Abstract**

Hierarchy Of Multi-label classifiERs (HOMER) is a multi-label learning algorithm that breaks the initial learning task to several, easier sub-tasks by first constructing a hierarchy of labels from a given label set and secondly employing a given base multi-label classifier (MLC) to the resulting sub-problems. The primary goal is to effectively address class imbalance and scalability issues that often arise in real-world multi-label classification problems. In this work, we present the general setup for a HOMER model and a simple extension of the algorithm that is suited for MLCs that output rankings. Furthermore, we provide a detailed analysis of the properties of the algorithm, both from an aspect of effectiveness and computational complexity. A secondary contribution involves the presentation of a balanced variant of the $k$ means algorithm, which serves in the first step of the label hierarchy construction. We conduct extensive experiments on six real-world data sets, studying empirically HOMER's parameters and providing examples of instantiations of the algorithm with different clustering approaches and MLCs, The empirical results demonstrate a significant improvement over the given base MLC.

*Keywords:* Knowledge discovery, Machine learning, Supervised learning, Text mining

## 1. Introduction

In multi-label learning, training examples are associated with a vector of binary target variables, also known as labels. The goal is to construct models that, given a new instance, predict the values of the target variables (classification), order the target variables from the most to the least relevant one with the given instance (ranking), do both classification and ranking, or even output a joined probability distribution for all target variables.

In the past decade multi-label learning has attracted a great deal of scientific interest. One main reason behind this is that a number of real-world applications can be formulated as multi-label learning problems; functional genomics [1], recommending bid phrases to advertisers [2], image [3] and music [4] classification are some example domains. The other main reason relates to the interesting challenges it poses, such as

*Corresponding author
*Email addresses:* ypapanik@csd.auth.gr (Yannis Papanikolaou), greg@csd.auth.gr (Grigorios Tsoumakas), katakis.i@unic.ac.cy (Ioannis Katakis)