Multi-Label Classification

M

Jesse Read
Universidad Carlos III, Spain

Albert Bifet
Yahoo! Research Barcelona, Spain

INTRODUCTION

Humans naturally associate an object with more than one concept, i.e., *label* (also known as a category, tag, or genre). Adding labels to data collections can greatly facilitate retrieval and organization. This is a vital task with the ever-increasing collections of data being created and accessed each day. In 2010 it was estimated that we were creating the same quantity of data every two days, as we created from the dawn of time up until 2003. Multi-label classification is a way to learn label multi-label associations and, importantly, assign labels to new data automatically.

BACKGROUND

Multi-label classifications come naturally. For example, a movie can be assigned to both genres "Action" and "Comedy." In the physical world, this often involves physical duplication, for example, a video store has to put copies of such a movie in two different sections -- "Action" and "Comedy" -- or create a single-category ("Action-Comedy"). In a digital environment there are no such restrictions, and it is likely for this reason that there has been a rapid rise in the popularity of the multi-label concept in everyday applications. Consider how the once-familiar 'folder' metaphor is being replaced by the label/tag term in many everyday applications: email, picture, document, and media collections, and so forth.

The typical goal of multi-label learning is to learn a model to predict or recommend labels for data points automatically, and thus reduce or even eliminate the need for time-consuming manual labeling of data and document collections. Such collections may include e-mails and text documents, images, audio and video collections, or even certain biological applications such as where genes can be associated with multiple functions. For a detailed introduction and review of multi-label classification, see for example (Tsoumakas, G., Katakis, I., & Vlahavas, 2010).

Multi-label classification has borrowed heavily from the already-existing domain of traditional single-label classification, of assigning a single 'class' to each data element. Methods can be grouped roughly into two categories:

Problem/Data Transformation

In this approach, the multi-labeled data is transformed into one or a series of single-label problems. Standard off-the-shelf single-label classifiers can then be applied. A typical approach is to create one binary problem for each label (to predict if the label is relevant or not), as in (Read et al., 2011), or a multi-class problem where combinations of labels are represented as atomic mutually-exclusive classes ("Action-Comedy," in the movie example, would be considered a

DOI: 10.4018/978-1-4666-5202-6.ch142

single class), as in (Tsoumakas, G., Katakis, I., & Vlahavas, P., 2011). Perhaps the earliest work of this type with specific reference to multi-label classification is (Boutell et al., 2004) where images were assigned scene labels.

Algorithm Adaptation

In this approach, existing single-label algorithms are adapted to deal with the multi-label problem, for example, in artificial neural networks that have multiple outputs (Zhang, M.-L. & Zhou, Z.-H., 2006), and in decision trees that predict multiple labels in the leaves (Vens, C., et al., 2008; and more recently, Kocev, D., 2013.), among many other adaptations.

Multi-label classification software with a variety of methods for multi-label learning and evaluation include:

- **MULAN:** http://mulan.sourceforge.net
- **MEKA:** http://meka.sourceforge.net

The selection of method depends on the size of the data collection, the number of labels, and the peculiarities of the data, i.e. the application domain. A good recent review of some of the most well-known methods so far is given in (Zhang, M.-L. & Zhou, Z.-H., 2013).

MAIN FOCUS

The main focus of methods for multi-label classification is modeling the dependencies between labels, and doing this efficiently enough to be practical on large collections. The naïve approach of predicting labels independently may not yield best results (by not taking into account label dependencies, such as when two labels should not occur together). Most modern multi-label methods will do this, although not all are applicable in scenarios where large numbers of labels and/or instances are involved.

Dealing with Complexity

Assigning one of 10 labels (single-label association) implies 10 possibilities whereas, in the multi-label case, any combination of labels is, in principle, possible for each data point, implying (given the same set of 10 labels) 1024 possible label assignments (i.e., exponential with respect to the number of labels). However, clearly, some combinations are more likely than others, and many may not occur at all.

For more than about 50 labels, many methods may need to use special adaptations for efficiency. A good example is the method proposed by (Tsoumakas, G., 2008), that clusters labels hierarchically, making a number of sub problems (having a subset of the total number of labels) that are easier to solve. Most multi-label problems typically range from having 6 to 100 or more labels. If there are over 1000 or so labels, this involves a qualitative difference, and is more likely to be a keyword-assignment or unstructured-tagging problem. This problem is identical to the multi-label problem with the exception that labels are not selected from a known predefined set, but rather may be created as seen fit, often as in a *folksonomy*.

Dealing with Imbalanced Data

If transforming the multi-label problem into a series of sub problems, care must be taken to avoid or deal with issues associated with imbalanced data. That is to say data with an overwhelming representation of one type of class assignment. This frequently occurs because labeling for most problems is quite sparse; i.e., each data point usually has a relatively small number of labels associated with it. For example, in a database of movies, the number negative examples of "Documentary" (i.e., any movie which is *not* labeled documentary) may be much greater than positive examples.