Pattern Recognition 63 (2017) 45–55

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Multi-label methods for prediction with sequential data

Jesse Read [a,b,*], Luca Martino [c,d], Jaakko Hollmén [b]

[a] Computer Science and Networks Department Télécom ParisTech, Université Paris-Sarclay, France
[b] Department of Computer Science, Aalto University and HIIT, Helsinki, Finland
[c] Institute of Mathematical Sciences and Computing, São Carlos, Brazil
[d] Image and Signal Processing Group, Universitat de València, Spain

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form
23 August 2016
Accepted 19 September 2016
Available online 21 September 2016

Keywords:
Multi-label classification
Problem transformation
Sequential data
Sequence prediction
Markov models

ABSTRACT

The number of methods available for classification of multi-label data has increased rapidly over recent years, yet relatively few links have been made with the related task of classification of sequential data. If labels indices are considered as time indices, the problems can often be seen as equivalent. In this paper we detect and elaborate on connections between multi-label methods and Markovian models, and study the suitability of multi-label methods for prediction in sequential data. From this study we draw upon the most suitable techniques from the area and develop two novel competitive approaches which can be applied to either kind of data. We carry out an empirical evaluation investigating performance on real-world sequential-prediction tasks: electricity demand, and route prediction. As well as showing that several popular multi-label algorithms are in fact easily applicable to sequencing tasks, our novel approaches, which benefit from a unified view of these areas, prove very competitive against established methods.

1. Introduction

Multi-label classification is the supervised learning problem where an instance is associated with multiple class variables (i.e., *labels*), rather than with a single class, as in traditional classification problems. See [1] for a review. The typical argument is that, since these labels are often strongly correlated, modelling the dependencies between them allows methods to obtain higher performance than if labels were modelled independently – at the expense of an increased computational cost.

The case of binary labels is most common, where a positive class value denotes the relevance of the label (and the negative or null class denotes irrelevance). Typical examples of binary multi-label classification involve categorizing text documents and images, which can be assigned any subset of a particular label set. For example, an image can be associated with both labels `beach` and `sunset`. This is usually represented in vector form, such that, given a set of labels[1] $\mathcal{L} = \{beach, urban, foliage, sunset, mountains, fields\}$, then an associated label vector is

$$\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6] = [\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}]$$

which indicates that the first and fourth labels (`beach` and `sunset`) are relevant. The image itself can be represented by feature vector $x = [x_1, \ldots, x_D]$, and thus the pair x, y represents an image and its associated labels. The multi-label classification paradigm has been successfully considered also in many other domains, such as text, video, audio, and bioinformatics – see [1] and references therein for further examples.

Although binary labels (representing relevance and irrelevance) are enough to represent a huge number of practical problems, the generalization where each label can take multiple values – variously called multi-target, multi-output, or multi-dimensional classification – has also been investigated in the literature (see [3–5]). In this case each t-th 'label' ($t = 1, \ldots, T$) can take on up to L values such as a rating $y_t \in \{1, 2, 3, 4, 5\}$ (where $L=5$), hour of day $y_t \in \{0, \ldots, 23\}$ (where $L=24$) and so on, rather than the simple relevance/irrelevance case ($L=2$). In practice many multi-label algorithms can be applied directly to the general multi-output case, and are always applicable indirectly, following from the fact that any binary number can be represented as any decimal number and vice versa. Fig. 1 shows the relationship between these paradigms. Throughout this work, we will continue to use the term multi-label classification for the general case..

Sequential data applications deal with a changing *state* over time, for example of an object or scenario at a particular time index. Approaches to modelling in relevant domains are frequently based on some variety of Markov model, of which detailed overviews are given by [6] and [7].

* Corresponding author at: Computer Science and Networks Department of Télécom ParisTech, Université Paris-Sarclay, France; Laboratory of Informatics, École Polytechnique, France.
E-mail address: jesse.read@polytechnique.edu (J. Read).

[1] Such as those in the Scene dataset, see [2].

http://dx.doi.org/10.1016/j.patcog.2016.09.015
0031-3203/

	$L = 2$	$L > 2$
$T = 1$	**binary**	**multi-class**
$T > 1$	**multi-label**	**multi-label**†

† also known as multi-output, multi-target, multi-dimensional.

Fig. 1. Different classification paradigms: T is the number of class *labels* (or target variables), and L is the number of *values* that each label variable can take.

For example, a traveller's movements among waypoints in a city can be modelled as a series of references to these points, where we can consider y_t as indicating the waypoint at time t, then an example of a short path among four points under typical notation[2]

$$y_{1:4} = y_1, y_2, y_3, y_4 = 3, 8, 17, 5$$

where the numbers are unique to each node. The difference in real time between each t and $t + 1$ depends on the application (it could be seconds, or minutes, for example). The observation (known often emission) available at time point t is represented as vector $\mathrm{x_t}$.

These two problems (the one of multi-label and sequential prediction), have until now mostly received attention as different areas of research. However, they can often be seen not just as related problems, but in fact as identical problems, where the terms 'time index', 'state', 'observation', and 'path' can be interchanged with terms like 'label index', 'label', 'input', and 'label vector', respectively.

We were motivated to take a unified view of these two tasks – multi-label classification and sequential prediction – in a framework that allows the natural application of one to the other. This allows us to apply and further develop suitable techniques from multi-label classification to the domain of sequential prediction, in the form of novel methods that overcoming the disadvantages of hidden Markov models and related approaches by allowing the simultaneous prediction of multiple values across time.

In the first contribution of this work, we compare and contrast typical approaches for modelling of multi-label and sequential data, then draw strong connections between these areas (Section 2). We show that many (if not, most) methods are directly applicable from one problem to the other, and that all methods are applicable in some way, usually only with small modifications to the way the data is preprocessed. We analyze and discuss the relative advantages and disadvantages of each method. Furthering this, we provide a unified view (in Section 3) describing a common framework for multi-label and sequential-data algorithms. We look particularly at the applicability of multi-label methods for obtaining competitive performance and necessary scalability characteristics for sequential prediction. In a novel manner we adapt a Markov-based methodology for multi-label data to create a new method (Viterbi Classifier Chains, Section 4), and discuss its suitability in both domains. This leads us to formulate a further novel approach (in Section 5): Sequential Increasingly-sized Chained Labelsets (SICL), which casts a combination of chain-based and set-based approaches to the sequential problem by taking into account the decay of confidence for points relatively further in the future. In Section 6 we compare against a number of competitive multi-label and sequential methods in empirical evaluations on some real-world sequential-data problems. We find that our novel schemes are competitive and scalable. Finally, in Section 7 we discuss the results, summarize our contributions, draw conclusions and mention promising future work in both areas.

[2] Although, in typical Markov-model notation, y is often used to denote the observation or emission, rather than the state.

2. Connections between multi-label and sequential classification problems

In this work we study the supervised classification task, where a series of inputs is mapped to a series of outputs by a model trained on similar labelled examples (i.e., a training set is available). In the sequential task, classification of the future is often specifically referred to as *prediction* (as opposed to the *estimation* of a current state). In the multi-label context, there is no explicit time context, and therefore the term prediction/estimation are used interchangeably for all outputs.

It should be noted that Markov methods are also used frequently in an unsupervised fashion, which is analogous to clustering in non-sequential data. Although this is also a major task, it is not one that we are directly concerned with in this work.

Also, if the state variable is continuous (i.e., $y_t \in R$), a natural extension of Markov models are the Kalman and particle filters, which is analogous to multi-output regression. We do not specifically address this case, although many of the connections we look at transfer also easily to the scenario of real-valued outputs.

In Table 1 we outline the parallels between the terminology used in research dealing with the areas of sequential and multi-label data. To the best of our knowledge connection has not been documented to such as extent in the literature. We will start with a discussion on models (Section 2.1) for sequential data, and refer back to these models thereafter as we draw connections from multi-label data (Section 2.2).

2.1. Models for sequential data

Applications of classification in sequential data abound in the real world and this is echoed in a wealth of scientific literature. Applications include speech, handwriting, and gesture recognition, part-of-speech tagging, daily activity and medical monitoring, fraud detection [8], and traveller modes and movements in an urban setting [9–11].

A prominent methodology is that of *Markov model*s, both for estimation and prediction, where the sequence of states generates a corresponding sequence of observations. The classification task can be carried out retrospectively or in real time.

Recal the notation outlined in Table 1 where each state $y_t \in \{1, \ldots, L\}$ at time t is a discrete variable taking one of L values. In a *hidden Markov model* (HMM), each state y_t at time t is seen as generating an observation/emission $\mathrm{x_t}$, in addition to the following state y_{t+1}, such that

Table 1
Notation, and comparison of typical terms in the literature dealing with sequential and multi-label data. Note that indexing with t is more typical of the former, whereas j, k, or ℓ are used to index labels. As the target application of this work involves sequential data, we use the t index henceforth throughout. On the other hand, we use y_t to indicate an output label, and $\mathbf{x}_t$ the inputs, as per multi-label convention and in contrast to many uses in sequential-algorithms, particularly Markov models.

symbol	sequential data	multi-label data
$t = 1, \ldots, T$	time index	label index
L	number of states	number of values per label
T	sequence length	total number of labels
$y_t \in \{1, \ldots, L\}$	state at time t	value of t-th label
$\mathbf{x}_{1:T} \equiv \mathbf{x}$	full emissions	input feature vector
$\mathbf{x}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_D]$	emission at time t	input subset
$y_{1:T} = y_1, \ldots, y_T$ $\equiv \mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$	sequence/path	label vector ($\mathbf{y}$)
$\{\mathbf{y}^{(i)}\}_{i=1}^N$	N sequences	label vectors
$\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N$	training data	training data / dataset