Neurocomputing 205 (2016) 439–454

Contents lists available at ScienceDirect

Neurocomputing

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

Feature ranking for multi-label classification using Markov networks

Paweł Teisseyre

Institute of Computer Science, Polish Academy of Sciences Jana Kazimierza 5, 01-248 Warsaw, Poland

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form
7 January 2016
Accepted 12 April 2016
Communicated by Deng Cheng
Available online 12 May 2016

Keywords:
Feature selection
Multi-label learning
Markov networks
Ising model

ABSTRACT

We propose a simple and efficient method for ranking features in multi-label classification. The method produces a ranking of features showing their relevance in predicting labels, which in turn allows us to choose a final subset of features. The procedure is based on Markov networks and allows us to model the dependencies between labels and features in a direct way. In the first step we build a simple network using only labels and then we test how much adding a single feature affects the initial network. More specifically, in the first step we use the Ising model whereas the second step is based on the score statistic, which allows us to test a significance of added features very quickly. The proposed approach does not require transformation of label space, gives interpretable results and allows for attractive visualization of dependency structure. We give a theoretical justification of the procedure by discussing some theoretical properties of the Ising model and the score statistic. We also discuss feature ranking procedure based on fitting Ising model using l_1 regularized logistic regressions. Numerical experiments show that the proposed methods outperform the conventional approaches on the considered artificial and real datasets.

1. Introduction

Multi-label classification (MLC) has recently attracted a significant attention, motivated by an increasing number of applications. Examples include text categorization [1–5], image classification [6–8], video classification [9,10], music categorization [11], gene and protein function prediction [12–14], medical diagnosis [15,16], chemical analysis [17,18], social network mining [19,20] and direct marketing [21]. More examples can be found in [22–24]. The key problem in multi-label learning is how to utilize label dependencies to improve the classification performance, motivated by which number of multi-label algorithms have been proposed in recent years (see [25] for extensive comparison of several methods). The recent progress in MLC is summarized in [26,22]. In MLC, each object of our interest (e.g. text, image, patient, etc.) is described by a vector of p features $\mathbf{x} = (x_1, \ldots, x_p)^T$ and a vector of K binary labels $\mathbf{y} = (y_1, \ldots, y_K)^T$. The main objective is to build a model (using some training examples) which predicts $\mathbf{y}$ based on $\mathbf{x}$.

One of the trending challenges in MLC is a dimensionality reduction of the feature space [22], i.e. reducing the dimensionality of the vector $\mathbf{x}$. Usually only some features affect $\mathbf{y}$. The issue is very important as in practical applications, the dimensionality of feature space can be very large. For example in text categorization a standard approach is to use so-called *bag-of-words model* in which frequencies of occurrence of words in a corpora are taken as features. This method generates thousands of features. Moreover, one can also take into account higher degree n-grams (bigrams, trigrams, etc.) and many other types of features (e.g. stylistic features like averaged word length), which further increases the dimensionality of feature vector. Elimination of redundant features is essential for the following reasons. First, it allows us to reduce the computational burden of MLC procedures. Secondly, it improves a prediction accuracy of MLC methods. Fitting many MLC models includes estimation of large number of parameters. It is well known that fitting models with many spurious features increases the variance of estimators and thus decreases the prediction accuracy of the model (see e.g. Chapter 7 in [27]). Finally, feature selection methods are used to discover dependency structure in data. This allows us to understand how features affect the labels, which is particularly important in biological and medical applications. For example, in multi-morbidity (co-occurrence of two or more chronic medical conditions in one person) it is crucial to discover which characteristics of the patient influence the co-occurrence of diseases [28]. Moreover, it would be interesting to know which diseases are likely to occur simultaneously given some characteristics of the patient (for example age, gender and previous diseases). We discuss different approaches of dimensionality reduction in MLC in Section 2.

In this paper we focus on feature ranking (FR) methods (sometimes also called filters). Although the MLC attracted a significant attention in machine learning community, only a few works address the feature ranking problem in multi-label setting. Feature ranking

E-mail address: teisseyrep@ipipan.waw.pl
URL: http://www.ipipan.eu/~teisseyrep/

http://dx.doi.org/10.1016/j.neucom.2016.04.023
0925-2312/

(FR) methods are mainly used to assess the individual relevance of available features. More precisely, they allow us to order features with respect to their relevance in predicting labels, which in turn allows us to remove the least significant features and build a final classification model using the most significant features. Although usually in this approach neither the possible redundancy between features nor their joint relevancy is taken into account, the main advantage is a low computational cost, which allows us to compute the importance of thousands of features relatively fast. This is crucial in many domains, like text categorization or functional genomics. Moreover, in some applications it is important to evaluate the individual relevance of features, not only their joint relevance. Some authors use FR methods as an initial step to filter out spurious features and then use more sophisticated selection methods on the remaining set of features (see e.g. sure independence screening procedure proposed by [29]). We also discuss FR method, which incorporates all features simultaneously.

The FR task in multi-label setting is much more challenging than in a single-label case. In traditional classification with only one target variable, FR methods aim to model the dependence between target variable y and a single feature x_j using different variable importance measures. Then the procedure is repeated for all possible features. The most popular measures are: information gain [30], the chi-squared statistic and simple statistics based on univariate logistic regression [31], among others. On the other hand, in MLC feature x_j may affect targets $y_1, \ldots, y_K$ in different ways. First, it may happen that x_j influences only some of labels, while others are independent from x_j. More importantly, since in MLC methods dependencies between labels are usually considered, we should verify how x_j affects a given label y_k, in a presence of the remaining labels. It may happen that x_j is independent from y_k, while x_j becomes dependent on y_k, when conditioned on other labels. Finally, feature x_j can influence only the interactions between labels, while the marginal dependencies are not present. Examples of such situations are provided in Sections 3.1 and 3.3. A desirable FR method should take into account all the above aspects.

The main limitation of recent FR methods is that they require problem transformation methods: binary relevance (BR) or label powerset (LP) transformation for evaluating the relevance of given features. Unfortunately, both transformations suffer from many serious drawbacks, discussed in more detail in Section 2. To propose a desirable FR method, we make an effort to take into account the following aspects.

- The method should not use BR or LP transformation.
- The method should take into account specificity of multi-label setting, i.e. it should measure the dependence between feature x_j and label y_k, given the remaining labels.
- The method should give interpretable results to see which labels (or interactions between labels) and how are influenced by feature.
- The computational cost of the procedure should be low.

To take into account the above postulates, we propose a novel approach which is based on Markov networks. Markov network (see e.g. [32], Section 8) can be represented as a graph, with node set representing random variables (in our case labels and features) and edge set representing dependencies between variables. Existing edge between two variables means that they are conditionally dependent given the rest of the graph. The main advantage of Markov networks is that they allow us to model the pairwise dependencies between labels and features in a direct way. Although, Markov networks have already been applied in MLC (see e.g. [33] or [34]), they have not been used as a feature ranking method. Our approach is based on the following idea. We initially build a Markov network containing only labels, which allows us to model the dependencies among the labels. In the second step, we test how much adding a single feature x_j affects the initial network. This allows us to test the dependence strength between a given feature x_j and a given label y_k, conditioning on the remaining labels. The procedure is repeated for all available features, which yields the final ranking. Specifically, in our method we use the Ising model [35,36] which is a simple example of Markov network. It turns out that for the Ising model, building an initial network containing only labels can be done relatively simply, especially for moderate number of labels. See Section 3.5 for deeper justification of using the Ising model. In a second step we propose to use the score statistic [37], which is very computationally efficient in this case. Namely, it is not necessary to refit an initial network when we add feature x_j. This allows us to test a significance of added features very quickly which is crucial in FR methods. The details of the procedure are given in next sections. Fig. 1 shows networks corresponding to the most and the least significant features for `scene` dataset, in which the task is to predict six labels (beach, sunset, field, fall, mountain, and urban). Numbers over edges $u_1, \ldots, u_6$ are the score statistics which reflect the conditional dependences between feature x_j and labels (given the remaining labels). The higher the value of the score statistic, the larger is the influence of x_j on the given label, in the presence of remaining labels. The score statistics for a given feature x_j are added together, which gives an importance measure for x_j. The final ranking is based on these importances. We also discuss FR procedure based on fitting Ising model using l_1 regularized logistic regressions.

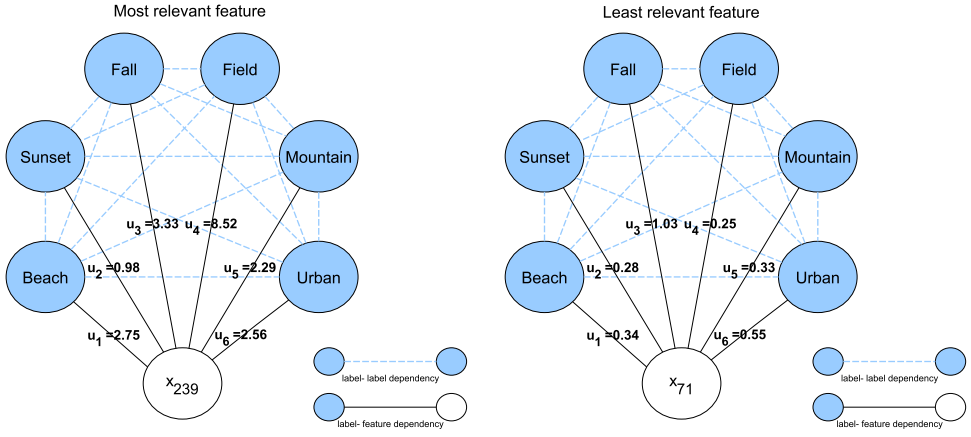

Fig. 1. Markov networks corresponding to the most (x_{239}) and the least (x_{71}) significant features for `scene` dataset. The numbers over edges are scores statistics describing importances of features.