Information Processing and Management 54 (2018) 359–369

Contents lists available at ScienceDirect

Information Processing and Management

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

Correlation analysis of performance measures for multi-label classification

Rafael B. Pereira[*,a], Alexandre Plastino[*,a], Bianca Zadrozny[b], Luiz H.C. Merschmann[c]

[a] Universidade Federal Fluminense (UFF). Instituto de Computação - Campus da Praia Vermelha. Av. Gal. Milton Tavares de Souza, s/n°, São Domingos, Niterói - RJ, 24210-346, Brazil
[b] IBM Research. Av. Pasteur, 146, Botafogo, Rio de Janeiro - RJ, 22296-900, Brazil
[c] Universidade Federal de Lavras (UFLA). Departamento de Ciência da Computação. Campus Universitário, Caixa Postal 3037, CEP 37200-000, Lavras - MG, Brazil

ARTICLE INFO

Keywords:
Multi-label classification
Evaluation measures

ABSTRACT

In many important application domains, such as text categorization, scene classification, biomolecular analysis and medical diagnosis, examples are naturally associated with more than one class label, giving rise to multi-label classification problems. This fact has led, in recent years, to a substantial amount of research in multi-label classification. In order to evaluate and compare multi-label classifiers, researchers have adapted evaluation measures from the single-label paradigm, like Precision and Recall; and also have developed many different measures specifically for the multi-label paradigm, like Hamming Loss and Subset Accuracy. However, these evaluation measures have been used arbitrarily in multi-label classification experiments, without an objective analysis of correlation or bias. This can lead to misleading conclusions, as the experimental results may appear to favor a specific behavior depending on the subset of measures chosen. Also, as different papers in the area currently employ distinct subsets of measures, it is difficult to compare results across papers. In this work, we provide a thorough analysis of multi-label evaluation measures, and we give concrete suggestions for researchers to make an informed decision when choosing evaluation measures for multi-label classification.

1. Introduction

A large body of research in supervised learning deals with the analysis of single-label data, where instances are associated with a single label from a set of class labels. More specifically, the single-label classification problem can be stated as the process of predicting the single class label of a new instance, which is described by its feature values.

However, in many important data mining applications, the instances are associated with more than one class label. This characterizes the multi-label classification problem, a relevant topic of research, which has become a very common real-world task (Zhang & Zhou, 2007).

In order to evaluate the performance of multi-label classifiers, some measures were adapted from the single-label paradigm, like Precision and Recall; and many were developed specifically for the multi-label paradigm, like Hamming Loss and Subset Accuracy. However, these measures have been used in multi-label experiments without an objective analysis of correlation or bias.

For instance, in Zhang and Zhou (2007), Cheng and Hüllermeier (2009) and Zhang, Peña, and Robles (2009), the adopted measures for evaluating the proposed algorithms were: Hamming Loss, One Error, Coverage, Ranking Loss and Average Precision. In

* Corresponding authors at: Av. Gal. Milton Tavares de Souza, s/n, Campus da Praia Vermelha, Boa Viagem, Niterói, RJ 24210-346, Brazil.
E-mail addresses: rbarros@ic.uff.br (R.B. Pereira), plastino@ic.uff.br (A. Plastino).

https://doi.org/10.1016/j.ipm.2018.01.002
Received 23 September 2014; Received in revised form 19 August 2017; Accepted 4 January 2018
Available online 16 January 2018
0306-4573/

Tsoumakas and Vlahavas (2007), the measures were Hamming Loss and Example-Based F-Measure. In Chen, Yan, Zhang, Chen, and Yang (2007), Micro-Averaged F-Measure and Macro-Averaged F-Measure. In Trohidis, Tsoumakas, Kalliris, and Vlahavas (2008), the measures were Hamming Loss and Example-Based Accuracy; and in Spolaôr, Cherman, Monard, and Lee (2013), just the Micro-Averaged F-Measure was used to report the results. In Madjarov, Kocev, Gjorgjevikj, and Dvzeroski (2012), a total of 16 multi-label measures were used to evaluate a large number of multi-label classifiers.

The adoption of arbitrary measures without an objective analysis of correlation or bias can lead to misleading conclusions, as an experiment evaluated with a subset of measures may appear to perform differently than when evaluated with another subset. Also, as different papers in the area currently employ distinct subsets of measures, it is difficult to compare results across papers. In this work, we provide a thorough analysis of multi-label evaluation measures, and we give concrete suggestions for researchers to make an informed decision when choosing evaluation measures for multi-label classification.

The remainder of this paper is organized as follows. In Section 2, we revisit the multi-label classification problem. In Section 3, we describe the multi-label measures most frequently used in the literature. In Section 4, we describe the experiments conducted to assess the correlation between multi-label measures. Finally, in Section 5, we make our concluding remarks and point to directions for future research.

2. Multi-label classification

In the multi-label classification task, each data instance may be associated with multiple labels. Multi-label classification is suitable for many domains such as text categorization, scene classification, medical diagnosis, microbiology (Read, 2010), and it is also a challenging problem in bioinformatics (Li, You, Ge, Yang, & Yang, 2010). In all these cases, the task is to assign for each unseen instance a label set whose size is unknown a priori (Zhang & Zhou, 2007).

The strategies proposed to deal with multi-label classification rely mainly on problem transformation, where the multi-label problem is transformed into a set of one or more single-label problems, and on algorithm adaptation, where the single-label learning algorithm is adapted to handle multi-label data directly (de Carvalho & Freitas, 2009; Tsoumakas, Katakis, & Vlahavas, 2010; Zhang & Zhou, 2014).

The simplest way to apply a classification strategy to a multi-label data set is to transform it into single-label data sets. Then a traditional classification technique – like *k*-NN or a decision tree – can be employed to perform the classification task. This way, the transformation technique allows the usage of single-label classification algorithms, which have been thoroughly studied and perfected over the last decades.

A popular transformation is Label Powerset (LP), which creates one label for each different subset of labels that exists in the multi-label training data set. Thus, the new set of labels corresponds to the powerset of the original set of labels. After this transformation process, a single-label classification algorithm can handle the transformed data set. This classifier can then be used to assign one of these new labels to new instances, which can then be mapped back to the corresponding subset of the original labels (Tsoumakas & Vlahavas, 2007).

Label powerset is recommended only for data sets with a small number of labels, as the possible powerset combinations are 2^L, where L is the number of distinct labels in the data set. For data sets with a large number of labels, the resulting powerset data tends to become sparse and therefore make it harder for the classifier to work (Dembczyński, Waegeman, Cheng, & Hüllermeier, 2012). The original label powerset technique has been extended and improved in subsequent work. Three variations are: Pruned Problem Transformation (PPT), proposed in Read (2008), which prunes away label sets that occur fewer times than a certain minimum threshold value; Random *K*-labelsets (RAKEL), proposed in Tsoumakas and Katakis (2007), which constructs an ensemble of LP classifiers trained using different and small random subsets of the set of labels (Tsoumakas et al., 2010); and HOMER (Tsoumakas, Katakis, & Vlahavas, 2008), which constructs a hierarchy of multi-label classifiers, each one dealing with a smaller set of labels.

Binary Relevance (BR) is a transformation technique that produces a binary classifier for each different label of the original data set. The method is called “binary relevance”, because each label is classified as relevant or non-relevant for an instance to be classified. The data transformation is applied to the multi-label data set, generating L single-label data sets, where L is the number of distinct labels in the original data set. Each single-label classifier yields a positive or negative result for each instance. The BR classification result is the union of the labels that are positively predicted by each classifier, as each one is capable of predicting one single label. As binary relevance learns a single binary model for each different label, it has linear complexity with respect to the number of labels (Tsoumakas et al., 2009).

Binary relevance does not take into account label correlations. Without this information, some relevant label dependences will not be considered (Tsoumakas et al., 2009). In order to minimize this drawback, several techniques have been proposed to extend and improve the binary relevance technique, such as: Ranking by Pairwise Comparison (Hüllermeier, Fürnkranz, Cheng, & Brinker, 2008); Calibrated Ranking by Pairwise Comparison (CRPC) (Fürnkranz, Hüllermeier, Loza Mencía, & Brinker, 2008); Classifier Chains (CC) (Read, Pfahringer, Holmes, & Frank, 2009) and Ensembles of Classifier Chains (ECC) (Read, Pfahringer, Holmes, & Frank, 2011). In general, these BR variations train more than L binary models or evaluate them in a specific order.

Besides using LP, BR or other popular transformation methods, another way to handle the multi-label data is to adapt well-known single-label classification algorithms. This is called algorithm adaptation, and most traditional classification algorithms employed in single-label problems have been adapted to the multi-label paradigm (Tsoumakas et al., 2010). The C4.5 decision-tree learning algorithm has been adapted to handle multi-label data in Clare and King (2001), by allowing multiple labels in the leaves of the tree. An SVM algorithm that minimizes the ranking loss metric has been proposed in Elisseeff and Weston (2001). A multi-label adaptation