Introduction to the special issue on learning from multi-label data

Grigorios Tsoumakas · Min-Ling Zhang · Zhi-Hua Zhou

In traditional supervised classification, objects belong to only one class out of two or more disjoint classes. However in many real-world applications, objects may belong to more than one class at the same time. For example an article on the Greek debt crisis could among others belong to the following classes of a financial newspaper's taxonomy: *Greece*, *Eurozone*, *Economy* and *Markets*. Such data are called multi-label.

Early research on learning from multi-label data focused on automated document categorization (McCallum 1999; Schapire and Singer 2000), motivated by the need for low-cost annotation of large document collections maintained by news agencies, academic publishers and intellectual property organizations. As our ability to collect and store large amounts of digital content increased in recent years due to technological advances, so did the need for automated annotation of such content, leading to an increased interest in multi-label learning from the communities of image, video, music and multimedia information retrieval (Irie et al. 2010; Su et al. 2011; Lo et al. 2011; Yang and Chen 2011). Furthermore, applications of multi-label learning extend beyond automated content annotation to tag recommendation (Song et al. 2011), query categorization (González-Caro and Baeza-Yates 2011), gene function prediction (Valentini 2011), medical diagnosis (Taylor et al. 2010), drug discovery (Kawai and Takahashi 2009) and direct marketing (Zhang et al. 2006). Multi-label data are clearly ubiquitous.

G. Tsoumakas
Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
e-mail: greg@csd.auth.gr

M.-L. Zhang
MOE Key Laboratory of Computer Network and Information Integration, School of Computer Science and Engineering, Southeast University, 2 Sipailou, Nanjing 210096, China
e-mail: zhangml@seu.edu.cn

Z.-H. Zhou (✉)
National Key Laboratory for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing 210046, China
e-mail: zhouzh@nju.edu.cn

Traditional supervised classification problems (binary or multi-class) can be regarded as special cases of multi-label learning, where each example is confined to have only one label. However, the "generality" of multi-label learning inevitably makes it more difficult to deal with. In the past few years, multi-label learning has become a hot topic in machine learning community. The ECML PKDD'09 tutorial[1] and the two MLD workshops[2] attracted substantial numbers of attendees. Despite the considerable progress achieved in recent years, there are yet many challenging and inspiring problems left to be explored. This special issue is thus organized as a dedicated platform to reflect recent achievements on learning from multi-label data.

A total of 22 submissions were received, 8 of which were finally accepted for this special issue. Each accepted paper has gone through two to four rounds of reviewing, each round with at least three referees. The contents of this special issue cover a variety of aspects on multi-label learning, including theoretical study on label dependency and loss function, novel algorithms derived from techniques such as learning-to-rank mapping, compression-recovery strategy and max-margin formulation, applications to document classification with topic models and gene function prediction with synergism, and more complex learning scenarios such as multi-label data streams and multi-instance multi-label learning (MIML).

Label dependency, or label correlation, is one of the key concepts exploited by most multi-label learning algorithms to achieve good empirical performance, but few works investigate this concept from a theoretical viewpoint. The paper "On Label Dependence and Loss Minimization in Multi-Label Classification" by Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng and Eyke Hüllermeier is an inspiring attempt in this regard within a statistical framework. In this paper, two types of label dependence named conditional and marginal (unconditional) dependence are firstly identified. After that, three different views on multi-label classification are distinguished where connections between the identified label dependence and the loss minimization under each view are established. In addition, concrete theoretical results are presented for the Hamming loss and the subset 0/1 loss, and several state-of-the-art multi-label classification algorithms are revisited in light of exploiting label dependence.

The paper "Multilabel Classification with Meta-Level Features in a Learning-to-Rank Framework" by Yiming Yang and Siddharth Gopal presents a framework that enables learning-to-rank algorithms from the field of information retrieval to be used for learning from multi-label data. This is achieved through the introduction of meta-level features based on the distance of each instance from its k nearest neighbours of each category and from the category's centroid vector. These features are constructed with the aim to discriminate the relevance of category-instance pairs analogous to the relevance of document-query pairs in information retrieval. Empirical results show consistent and significant performance improvements of this framework over state-of-the-art multi-label learning methods.

The paper "Compressed Labeling on Distilled Labelsets for Multi-Label Learning" by Tianyi Zhou, Dacheng Tao and Xindong Wu presents a novel multi-label learning method named compressed labeling by manipulating the binary label matrix with a compression-recovery strategy. In the compression phase, the label matrix of training examples is compressed to the sign matrix of its Gaussian random projections at lower dimensions to improve label balance and independence. Afterwards, one binary classifier is induced for each compressed label by utilizing existing binary classification methods. In the recovery phase, the

---

test example is fed to the induced binary classifiers for prediction. Afterwards, the predicted label vector in the low-dimensional space is reconstructed to the label vector in the original space with a fast recovery algorithm built on distilled labelsets (i.e. frequent label subsets in training set). Comparative studies with several state-of-the-art algorithms on 21 multi-label datasets validate the effectiveness, efficiency and robustness of the proposed method.

The paper "Efficient Max-Margin Multi-Label Classification with Applications to Zero-Shot Learning" by Bharath Hariharan, S.V.N. Vishwanathan and Manik Varma presents a max-margin formulation for multi-label classification with prior knowledge about densely correlated labels. The primal formulation is developed by incorporating the prior in linear form with an invertible matrix and choosing multi-label loss function decomposable over individual labels. The corresponding dual formulation is optimized with efficient algorithms tuned to both the kernelised and the linear case. Empirical results demonstrate the efficiency of the proposed method as well as its effectiveness for zero-shot learning with proper priors.

The paper "Statistical Topic Models for Multi-Label Document Classification" by Timothy N. Rubin, America Chambers, Padhraic Smyth and Mark Steyvers presents two novel generative models for multi-label classification of documents based on the latent Dirichlet allocation (LDA) framework. The first one, Prior-LDA, accounts for the prior distribution of the labels, while the second one, Dependency-LDA, accounts for dependencies among the labels. Empirical results show that Prior-LDA improves over a simple LDA model, while Dependency-LDA further improves over Prior-LDA. Especially interesting is the focus of this paper to multi-label document corpora exhibiting a power-law distribution of labels, i.e. having a large number of labels, many of which have a very small number of examples. This is the typical case of real-word document corpora, but has not up to now received significant attention. Empirical results show that Dependency-LDA is competitive to or better than Support Vector Machines in such cases, due to performance improvements in the rare labels.

The paper "Synergy of Multi-Label Hierarchical Ensembles, Data Fusion, and Cost-Sensitive Methods for Gene Functional Inference" by Nicolò Cesa-Bianchi, Matteo Re and Giorgio Valentini presents a joint study of three different important issues in whole-ontology and genome-wide gene function prediction: (a) hierarchical relationships of labels, (b) multiple sources of biomolecular data, and (c) class imbalance. Empirical assessment shows that integrating techniques for dealing with all of these issues is a key factor for improved performance in discovering new gene functions.

The paper "Scalable and Efficient Multi-Label Classification for Evolving Data Streams" by Jesse Read, Albert Bifet, Geoff Holmes and Bernhard Pfahringer focuses on streams of evolving multi-label data and presents a framework for learning, evaluation and synthetic data generation as well as new instance-incremental adaptive learning algorithm based on Hoeffding trees with multi-label Pruned Sets classifiers at the leaves. Empirical results show improved performance over the state-of-the-art in learning from multi-label data streams.

MIML, i.e. multi-instance multi-label learning, is a new framework to learn from complicated objects having multiple semantic meanings, where each example is described by multiple instances and associated with multiple class labels. MIML is closely related to multi-label learning where the latter can be regarded as a degenerated version of MIML if each example is only represented by a single instance. The paper "Bayesian Multi-Instance Multi-Label Learning Using Gaussian Process Prior" by Jianjun He, Hong Gu and Zhelong Wang presents a novel approach to MIML in a Bayesian framework. Specifically, instance-level latent functions with Gaussian process prior are assumed for each label to account for the connections between instances and labels as well as the correlations among different labels. Empirical results validate the superior performance of the proposed MIML approach over other state-of-the-art algorithms.

This special issue would not have been possible without the contributions of many people. We wish to sincerely thank all the authors for submitting their work to this special issue. We wish to express our gratitude to all the referees for their expertise and dedication in providing invaluable comments and suggestions. We are also grateful to the previous and current MLJ Editors-in-Chief, Foster Provost and Peter Flach respectively, for their encouraging support, and the editorial office for their consistent help.

References

González-Caro, C., & Baeza-Yates, R. (2011). A multi-faceted approach to query intent classification. In *Proceedings of the 18th international symposium on string processing and information retrieval (SPIRE 2011)* (pp. 368–379), Pisa, Italy.

Irie, G., Satou, T., Kojima, A., Yamasaki, T., & Aizawa, K. (2010). Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE Transactions on Multimedia*, *12*(6), 523–535.

Kawai, K., & Takahashi, Y. (2009). Identification of the dual action antihypertensive drugs using tfs-based support vector machines. *Chem-Bio Informatics Journal*, *9*, 41–51.

Lo, H. Y., Wang, J. C., Wang, H. M., & Lin, S. D. (2011). Cost-sensitive multi-label learning for audio tag annotation and retrieval. *IEEE Transactions on Multimedia*, *13*(3), 518–529.

McCallum, A. (1999). Multi-label text classification with a mixture model trained by em. In *Proceedings of the AAAI' 99 workshop on text learning*.

Schapire, R. E., & Singer, Y. (2000). Boostexter: a boosting-based system for text categorization. *Machine Learning*, *39*(2/3), 135–168.

Song, Y., Zhang, L., & Giles, C. (2011). Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web*, *5*(1), 4:1–4:31. Article 4.

Su, J. H., Chou, C. L., Lin, C. Y., & Tseng, V. S. (2011). Effective semantic annotation by image-to-concept distribution model. *IEEE Transactions on Multimedia*, *13*(3), 530–538.

Taylor, P., Almeida, G., Kanade, T., & Hodgins, J. (2010). Classifying human motion quality for knee osteoarthritis using accelerometers. In *Proceedings of the 32nd annual international conference of the IEEE engineering in medicine and biology society (EMBC 2010)* (pp. 339–343).

Valentini, G. (2011). True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(3), 832–847.

Yang, Y., & Chen, H. (2011). Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 762–774.

Zhang, Y., Burer, S., & Street, W. N. (2006). Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research*, *7*, 1315–1338.