Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Empirical Comparison of Multi-Label Classification Algorithms

Clifford A. Tawiah, Victor S. Sheng
Department of Computer Science, University of Central Arkansas, Conway, Arkansas, USA 72035
ctawiah2@uca.edu, ssheng@uca.edu

Abstract

Multi-label classifications exist in many real world applications. This paper empirically studies the performance of a variety of multi-label classification algorithms. Some of them are developed based on problem transformation. Some of them are developed based on adaption. Our experimental results show that the adaptive Multi-Label K-Nearest Neighbor performs the best, followed by Random k-Label Set, followed by Classifier Chain and Binary Relevance. Adaboost.MH performs the worst, followed by Pruned Problem Transformation. Our experimental results also provide us the confidence of the correlations among multi-labels. These insights shed light for future research directions on multi-label classifications.

Introduction

Multi-label classifications deal with multiple labels being assigned to every instance in a dataset. That is, an instance can be assigned more than one class simultaneously. It is concerned with learning a model that outputs a bipartition of a set of labels into relevant and irrelevant with respect to a query instance. This type of classification differs in some respect from traditional single label classifications in that one of multiple labels is allocated to an instance in the dataset. In single-label classifications each instance is associated with a single label and a classifier learns to associate each new test instance with one of these known labels.

Multi-label classification tasks exist in many real-world applications, such as, gene classification in bioinformatics, medical diagnosis, document classification, music annotation, image recognition, and so on. All these applications require effective and efficient multi-label classification algorithms.

There exist a variety of multi-label classification algorithms. Current existing multi-label classification algorithms are developed based on two basic approaches: algorithm adaptation and problem transformation. Algorithm adaptation is to extend existing traditional classification algorithms to perform multi-label classifications directly, for example, Adaboost.MH (Schapire & Singer 2000) and Multi-label K-Nearest Neighbor (MLKNN) (Zhang & Zhou 2007).

Problem transformation transfers multi-label classifications into multiple traditional single label classifications, specifically, multiple binary classifications (yes or no). After a multi-label classification problem is transferred into multiple binary classification. All the traditional classification algorithms can be applied directly to build a classifier for each binary dataset and make prediction for its correlated test instances. The prediction for a multi-label instance is made by aggregating outputs from autonomous binary classifiers. Binary Relevance (BR), Classifier Chain (Read et al. 2011), Random *k*-Label Set (Tsoumakas et al. 2011), and Pruned Problem Transformation (Jesse 2008) are the examples of classifiers which use problem transformation.

Experiments

We conducted experiments on the above six classification algorithms using the eight datasets (Emotions, Enron, Genbase, Medical, Scene, Yeast, Mediamill, and Bibtex) from MULAN (Tsoumakas et al. 2010). If a dataset is separated into a training and test set already, we only run each classification algorithm once on its training set, and report its performance on its test set. Otherwise, we report the average results of 10 runs. The default base learner is used for the six multi-label classification algorithms. Specifically, J48 is used in conjunction with Binary Relevance and Classifier Chain. LabelPowerset, in conjunction with J48 is used as the base learner for RA*k*EL. J48 pruning tree is used for Pruned Problem Transformation. Adaboost.MH uses AdaBoostM1 as its base learner, which in turn uses decision stump as its base learner. We also introduce five popular performance metrics specifically designed for multi-label classifications,

The work was supported by the National Science Foundation (IIS-1115417).

Hamming Loss, Average Precision, One-Error, Coverage, and Ranking Loss (Tsoumakas et al. 2010).

We have the experimental results for all these datasets. In order to show the general knowledge of the performance of the six multi-label classification algorithms, Table 1 shows the average values of the five performance metrics. We highlight the best in bold, highlight the second in italic, and underline the worst in the table.

We further summarize the comparisons among the six multi-label classification algorithms through ranking over the eight datasets. For each of the eight datasets, we ranked the performance of the six algorithms from 1 (best) to 6 (worst) on each metric. The average rank of the six algorithms on each metric is shown in Table 2. Further, we average the average ranks for each of the six algorithms across the five performance metrics in its last column.

Table 1. Average results of different algorithms

	HL	AP	OE	CV	RL
CC	0.1029	0.6541	0.3534	*19.077*	0.1969
RAKEL	*0.0869*	*0.6967*	**0.2753**	19.781	*0.1608*
AD	0.1147	0.4103	0.5741	29.265	0.3887
PPT	0.1158	0.6231	0.3661	22.254	0.2237
BR	0.0988	0.6458	0.3710	20.399	0.1997
MLKNN	**0.0789**	**0.7003**	*0.2883*	**15.269**	**0.1200**

Table 2. Average ranks of different algorithms

	HL	AP	OE	CV	RL	**Average**
CC	3.375	3.00	3.5	3.125	3.125	3.225
RAKEL	**2.0**	*2.625*	*2.375*	*2.375*	*2.5*	*2.375*
AD	4.5	4.5	4.625	5.75	5.75	5.025
PPT	4.75	4.125	3.625	4.375	4.375	4.25
BR	3.25	3.5	3.625	4.0	3.25	3.525
MLKNN	**2.0**	**2.0**	**2.0**	**1.5**	**1.5**	**1.8**

Table 2 shows that MLKNN performs the best. Its average ranks on all five performance metrics are the best (lowest values). Its overall rank value across the five performance metrics is 1.8, much lower than the second RA*k*EL, whose overall rank value is 2.375. The average performance over the eight datasets shown in Table 1 also supports this. Table 1 shows that MLKNN achieves the lowest values on the lowest-best metrics: Hamming Loss (HL), Coverage (CV), and Ranking Loss (RL), and achieves the highest value on the highest-best metric Average Precision (AP).

Table 2 shows that RA*k*EL is the second on all performance metrics, including the tied best in Hamming Loss (HL). Its overall rank 2.375 is also the second. Thus, we can conclude that RA*k*EL is the second best among the six algorithms. The average performance over the eight datasets shown in Table 1 also supports this. It shows that RA*k*EL achieves the lowest value on the lowest-best metric, One-Error (OE). It also achieves the second lowest in the lowest-best metrics: Hamming Loss (HL) and Ranking Loss (RL), and achieves the second highest value on the highest-best metric Average Precision (AP).

Table 2 shows that Adaboost.MH (AD) performs the worst. Its average ranks on all five performance metrics are the worst (highest values). Its overall rank value across the five performance metrics is 5.025, close to the maximum value 6.0. The average performance over the eight datasets shown in Table 1 also supports this. Table 1 shows that AD achieves the highest values on the lowest-best metrics: One-Error (OE), Coverage (CV), and Ranking Loss (RL), and achieves the lowest value on the highest-best metric Average Precision (AP). Table 1 shows that Classifier Chain (CC), Binary Relevance (BR), and Pruned Problem Transformation (PPT) take the middle positions. We can see that CC is the best, followed by BR, followed by PPT among the three algorithms. Table 1 also shows this relationship.

In summary, our experiments show that the adaptive multi-label learning algorithm MLKNN performs the best, followed by RA*k*EL, followed by Classifier Chain and Binary Relevance. Adaboost.MH performs the worst, followed by Pruned Problem Transformation. This provides the guide for multi-label classification practitioners and saves their time to try and to estimate the possible achievement. This also stimulates us to adapt traditional single label classification algorithms for multi-label problems. Our experimental results also provide us the confidence of the correlations among multi-labels. The multi-label independency assumption is not succeeded in most of datasets. How to utilize the correlations among these labels will shed a light for our future research.

We will continue to evaluate the performance of existing multi-label classification algorithms. In the same time, we are going to design novel algorithms for multi-classifications with the insights found in the experiments.

References

Jesse, R. 2008. A pruned problem transformation method for multi-label classification, Proc. Of the New Zealand Computer Science Research Student Conference (NZCSRS 2008).

Read, J. Pfahringer, B., Holmes, G. and Frank, E. 2011. Classifier chains for multi-label classification. Machine Learning, 85(3):333–359.

Schapire, R.E. & Singer, Y. 2000. Boostexter: a boosting-based system for text categorization. Machine Learning, vol. 39, no. 2/3, pp. 135-168.

Tsoumakas, G., Katakis, I., & Vlahavas, I. 2010. Mining Multi-label Data, Data Mining and Knowledge Discovery Handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd edition.

Tsoumakas, G., Katakis, I., & Vlahavas, I. 2011. Random k-labelsets for multilabel classification, IEEE Transactions on Knowledge and Data Engineering, 23(7): 1079-1089.

Zhang, M. & Zhou, Z. 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40: 2038-2048.