Multi Label Classification using Label Clustering

Pranav Gupta IIT.G.PRANAV@GMAIL.COM
Ashish Anand ANAND.ASHISH@IITG.ERNET.IN
Indian Institute of Technology Guwahati, Guwahati - 781039

Abstract

Multi Label Classifier classifies a given instance into *one or more* labels from a set of labels. In literature, such classifiers have been classified into two categories: *Problem Transformation* (PT) and *Algorithm Adaptation* (AA). Because of simplicity and competitive performance of PT-methods, we explore a PT method, namely Label Powerset (LP). Existing LP methods are either too slow or tend to underutilize multi label information. We propose a novel LP approach, achieving competitive performance with respect to Hamming Loss and F1-measure, in relatively less time.

1. Introduction

Algorithm Adaptation Methods modify single label classification algorithms to handle multi label data and generate multi label output(Madjarov et al., 2012). ML-C4.5 modifies entropy formula of C4.5(Madjarov et al., 2012). ML-kNN extends kNN(Madjarov et al., 2012). PT methods(Madjarov et al., 2012) reduce the problem to one or more single label classification problems. Though it simplifies the problem, it requires careful exploitation of correlation information among labels for better prediction. Due to its simplicity and competitiveness across several datasets(Madjarov et al., 2012), we explore PT methods.

A popular PT method, 'Binary Relevance' (BR)(Madjarov et al., 2012), builds a binary classifier using 'one vs one' or 'one vs all' method. It, however, assumes the labels to be independent and quality of prediction suffers. Another PT method, 'Label Powerset' (LP)(Madjarov et al., 2012) forms a set L' of multi labels l seen in training data and trains a single label classifier S treating l as single class, implicitly incorporating label correlation information. However, firstly, the size of L' can be large. Secondly, many multi labels appear infrequently and cause class imbalance for S. Thirdly, unseen multi labels cannot be identified directly during classification. In this paper, we review popular LP methods which try to solve these challenges, discuss their disadvantages and propose a new algorithm to improve upon them.

Appearing in Proceedings of the 1st Indian Workshop on Machine Learning, IIT Kanpur, India, 2013.

2. Related Work

Pruned Sets (PS) (Read et al., 2008), before using LP, removes all training instances (**x**,**y**) where **y** is an infrequently occurring multi label. It then reproduces this (**x**,**y**) as (**x**,**y'**) ∀**y'** such that **y'** ⊂ **y** and **y'** is frequently occurring. It thus reduces number of multi labels and removes class imbalance, improving over naive LP. But it usually throws away crucial information, missing out on important multi labels. The issue with unseen multi labels during classification remains unresolved.

RAndom k-labELsets (RAkEL) (Tsoumakas & Vlahavas, 2007) detects unseen multi labels by creating an ensemble of m LP classifiers, each trained on a k-sized subset of labels. During classification, each classifier gives a binary prediction for the set of labels it was trained with. Averaging over all predictions, labels with score greater than a threshold are finally selected. However, the training time for RAKEL is very large.

Ensemble of Pruned Sets (EPS) (Read et al., 2008) creates an ensemble of m PS classifiers by choosing random subsets of the training data. Classification follows RAKEL. EPS, however, loses a lot of multi label information before identifying unseen multi labels and has a fairly large training time.

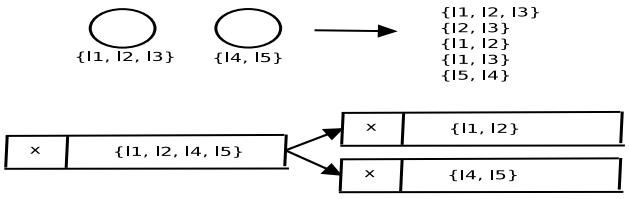

Figure 1. AlgL: Modification of Training Data

Algorithm 1 AlgL: Modification of Training Data

Input: Training Instances T, new multi labels C
for $(\mathbf{x}, \mathbf{y}) \in T$ **do**
  for $\mathbf{l} \subset \mathbf{y}$ **do**
    if $l \in C$ **then**
      $T \leftarrow T \cup \{(\mathbf{x}, \mathbf{l})\}$
    end if
  end for
end for
Output: T

3. Proposed Algorithm (AlgL)

AlgL identifies important unseen multi labels before pruning infrequent multi labels. It assumes correlated labels have high probability of forming multi labels. It clusters labels based on training data. Each cluster contains correlated labels. Each label is an N-dimensional boolean vector. Its i^{th} dimension is 1 if it is present in the i^{th} training record, else 0. The labels are clustered hierarchically using simple k-means with k=2 at each level in hierarchy. New multi labels are formed and incorporated in training data as follows:

1. For every cluster c, add all $\mathbf{y} \subseteq c$ to the list of new multi labels, if $|\mathbf{y}| \geq t$, where t is a tunable parameter. In Figure 1, $t = 2$.
2. Modify training data to contain instances labelled with the new multi labels as described in Algorithm 1 and Figure 1.
3. Train a PS classifier with the modified data.

4. Experiments

We use diverse datasets(Madjarov et al., 2012), namely Yeast, Enron, Scene, Medical, Corel5K. Due to space constraints, we present results only on datasets of Table 1. Experimental setup is consistent with (Read et al., 2008): We use the train and test splits originally provided with the datasets. Parameter tuning is done using 5 Fold Cross Validation on the training data. Models are evaluated on Hamming Loss and F_1 Measure. We compare AlgL against PS and EPS with respect to Hamming Loss and F_1 Measure on independent test data and training-time.

Table 1. Multi Label Datasets

Data set	Domain	tr.e.	t.e.	L[a]	LC[b]
Enron	Text	1123	579	53	3.38
Medical	Text	645	333	45	1.25
Corel5K	Multimedia	4500	500	374	3.52

[a]Number of Labels
[b]Label Cardinality

Table 2. Results: F_1 Measure, Hamming Loss, Build Time

Data set	EPS	AlgL
Enron	0.06/**0.54**/60.40	**0.05**/0.50/**20.05**
Medical	0.02/**0.77**/7.27	**0.01**/0.70/**2.95**
Corel5K	0.01/**0.17**/**47.67**	**0.01**/0.15/110.01

5. Results and Analysis

Both EPS and AlgL outperform PS on all datasets. The results for EPS and AlgL are presented in Table 2. Each cell resembles Hamming Loss/F_1 measure/Training Time. AlgL has a smaller build time because label clustering is usually much cheaper than an ensemble. This may not be true for large number of labels, as in Corel5k. For this reason, we plan to experiment with balanced clustering.

References

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Deroski, S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45 (9):3084 – 3104, 2012.

Read, J., Pfahringer, B., and Holmes, G. Multi-label classification using ensembles of pruned sets. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pp. 995–1000, 2008.

Tsoumakas, G and Vlahavas, I. Random k-labelsets: An ensemble method for multilabel classification. In *Machine Learning: ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, pp. 406–417. 2007.