Multilabel classification using heterogeneous ensemble of multi-label classifiers

Muhammad Atif Tahir , Josef Kittler , Ahmed Bouridane

a Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK

b School of Computing, Engineering and Information Sciences, University of Northumbria, Newcastle upon Tyne NE2 1XE, UK

Josu Maiora

Outline

1. Introduction

- A conventional multi-class classification system assigns each instance x a single label l from a set of disjoint labels L.
- In this paper each instance is to be assigned to a subset of labels $Y \subseteq L$. This problem is known as multi-label learning.
- They include highly imbalanced training sets, as very limited data is available for some labels, and capturing correlation among classes.
- In this paper, we focus on highly imbalanced data distributions using ensemble of multi-label classifiers.

1. Introduction

- Ensemble techniques are becoming increasingly important → they improve the accuracy with highly imbalanced data.
- Ensembles can be homogeneous (every base classifier using the same algorithm), or heterogeneous (different algorithms).
- The aim of this paper is to use **heterogeneous** ensembles of **multi-label** learners to improve the performance.

1. Introduction

- The proposed ensemble multi-label learning approach (EML)[1] is applied to six publicly available multi-label data sets

2. Related work

- Multi-label classification
  - Problem transformation methods (one or more single-label)
  - Algorithm adaptation methods (extend traditional classifier to handle multi-label)
- RaKEL (Tsoumakas and Vlahavas, 2007),
- Calibrated Label Ranking (CLR) (Furnkranz et al., 2008),
- Multi-label KNN (MLKNN) (Zhang and Zhou, 2007),
- Instance Based Logistic Regression (IBLR) (Chengand Hullermeier, 2009) and
- Ensemble of Classifier Chains (ECC)(Read et al., 2009).

3. Ensemble of multi-label classifiers (EML)

- Let X denote a set of instances (sample) and let Y = {1, 2, . . . ,N} be a set of labels

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

- The goal is to design a multi-label classifier H that predicts a set of labels for an unseen example.

- Ensemble of multi-label classifier train q multi-label classifiers $H_1, H_2, \ldots, H_q$. For an unseen instance *x*, each k^{th} individual model (of q models) produces an N-dimensional vector

  $P_k = [p_{1k}, p_{2k}, \ldots, p_{Nk}]$,

  where the value p_{bk} is the probability of the b^{th} class label assigned by classifier *k* being correct.

3. Ensemble of multi-label classifiers (EML)

- MEAN, MAX, MIN are the simplest and most popular way to combine the scores of classifiers (Kuncheva, 2004) → These combiners have no extra parameters to be trained

- Weighted voting methods also have the potential to make the multiple classifier systems more robust to the choice of individual classifiers

3. Ensemble of multi-label classifiers (EML)

1. Average of probabilities (EML_A)

$$\mu_b(x) = \frac{1}{q}\sum_{k=1}^{q} p_{bk}(x)$$

3. Ensemble of multi-label classifiers (EML)

2. Average of probabilities and threshold selection via multi-labelled-ness (EML_T)

- Properly adjusting the decision thresholds (instead of the traditional value of 0.5) can improve the performance of a multi-label classifier.
- Let X_T be the training set and X_S the test set. A threshold t is then selected using Eq. (2) to choose the final predicted multi-label set Z.

$$t = \arg \min_{\{t \in 0.00, 0.001, \ldots, 1.00\}} |LCard(X_T) - LCard(H_t(X_s))|$$

$$LCard(X) = \frac{\sum_{i=1}^{|X|} |E_i|}{|X|}$$

where Ei is the actual set of labels for the training set and a predicted set of labels under threshold t for the test set

3. Ensemble of multi-label classifiers (EML)

3. Static weighting by N-Fold Cross Validation (EML_S)

In static weighting, the weights for each classifier are computed in the training phase. In this paper, the weights for each classifier are learnt via N-Fold Cross Validation (N = 5)

3. Ensemble of multi-label classifiers (EML)

4. Dynamic weighting using Dudani rule (EML_D)

- A weighted k-NN rule is proposed for classifying new patterns
- The main idea is to weight a neighbor with smaller distance more heavily than the one with a greater distance

$$w_j = \begin{cases} \frac{d_k - d_j}{d_k - d_1} & \text{if } d_k \neq d_1 \\ 1 & \text{otherwise} \end{cases}$$

d_k furthest neighbor
d_1 nearest neighbor

- For multi-label classifier

$$weight(M_j) = \begin{cases} \frac{d_q - d_j}{d_q - d_1} & \text{if } d_q \neq d_1 \\ 1 & \text{otherwise} \end{cases}$$

- w(Mj) is the weight of multi-label classifier j for instance x

3. Ensemble of multi-label classifiers (EML)

5. Dynamic weighting using Shepard rule (EMLP)

- Shepard:

  ‘‘the relevance of a previous stimulus for the generalization to a new stimulus is an exponentially decreasing function of its distance in psychological space’’.

$$weight(M_j) = e^{-\alpha d_j^{\beta}}$$

α y β constants

4. Experiments and results

- Datasets: six multi-label datasets from a variety of domains
- Features: publicly available feature vectors are used for all datasets
- Evaluation measures: Hamming Loss, Accuracy, F1, and Classification Accuracy from the example-based category, and Micro/Macro F1/AUC from the label-based group. Additionally, we use One-error, Coverage, Ranking Loss and Average Precision from the ranking-based group

Table 1
Standard and multilabel statistics for the data sets used in the experiments.

Datasets	Domain	Samples	Features	Labels	LCard
Enron	Text	1702	1001	53	3.38
Medical	Text	978	1449	45	1.25
Scene	Vision	2407	294	6	1.07
Pascal07	Vision	9963	500	20	1.44
Yeast	Biology	2417	103	14	4.24
Emotions	Music	593	72	6	1.87

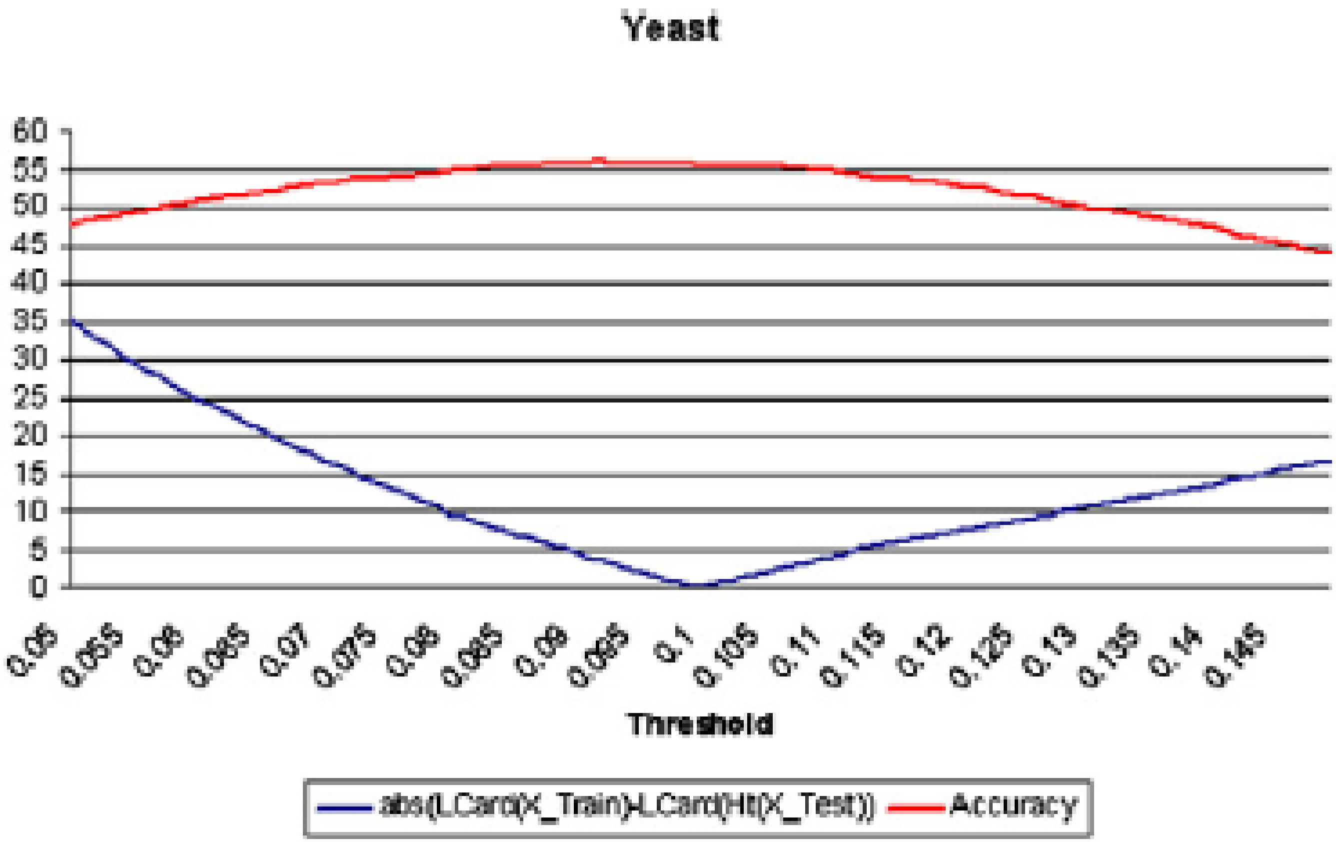
4. Experiments and results

❑ Benchmark methods

Table 2
Comparison of the proposed ensemble method (EML) with the state-of-the-art multi-label classifiers for Yeast. For each evaluation criterion, ↓ indicates "the smaller the better" while ↑ indicates "the higher the better". $*$ means significantly better than all other methods except those which are marked as $+$. Bold values indicate the best performance among all including ensemble classifiers while underscore values indicate the best performance among Individual Classifiers.

	MLkNN	IBLR	RAkEL	CLR	ECC	EML_A	EML_T	EML_S	EML_P	EML_D
Hamming Loss ↓	0.198	0.198	0.229	0.211	0.216	**0.193***	0.199	0.198	0.197	0.198
Accuracy ↑	0.499	0.510	0.475	0.491	0.529	0.540	0.549	0.549	**0.553***	0.552^{+}
Fmeasure ↑	0.635	0.644	0.621	0.633	0.659	0.671	0.681^{+}	0.681^{+}	**0.683***	0.682^{+}
ClassAcc ↑	0.165	0.186	0.102	0.134	0.183	0.196^{+}	0.188	0.188	**0.200***	0.196
Micro F_1 ↑	0.633	0.641	0.615	0.629	0.649	0.665	0.672^{+}	0.672^{+}	**0.674***	0.672^{+}
Macro F_1 ↑	0.352	0.375	0.400	0.390	0.415	0.399	0.486^{+}	0.486^{+}	**0.489***	0.488
Micro AUC ↑	0.835	0.838	0.792	0.815	0.805	**0.846***	0.842	0.842	0.843	0.843
Macro AUC ↑	0.668	0.685	0.628	0.656	0.652	**0.705***	0.693	0.693	0.695	0.693
One-error ↓	0.238	0.237	0.279	0.239	0.259	0.227^{+}	0.227^{+}	0.226^{+}	**0.224***	0.225^{+}
Coverage ↓	6.370	6.331	7.641	6.632	7.086	6.241^{+}	6.241^{+}	6.236^{+}	**6.173***	6.190^{+}
Ranking Loss ↓	0.173	0.170	0.223	0.181	0.214	0.162^{+}	0.162^{+}	0.162^{+}	**0.161***	0.162^{+}
AvgPrecision ↑	0.756	0.759	0.710	0.750	0.735	0.768^{+}	0.768^{+}	0.768^{+}	**0.769***	0.768^{+}
# Wins (Ind)	1/12	8/12	0/12	0/12	4/12	–		–	–	–
# Wins (All)	0/12	0/12	0/12	0/12	0/12	3/12	0/12	0/12	9/12	0/12

4. Experiments and results



In order to demonstrate the effectiveness of the threshold selection method discussed in Section 3.2, Fig. 1 shows the graphs for different values of threshold t in the X-axis and two curves in the Y-axis (jLCard (XT) – LCard (Ht(Xs))j, Accuracy) for the various data sets.

4. Experiments and results

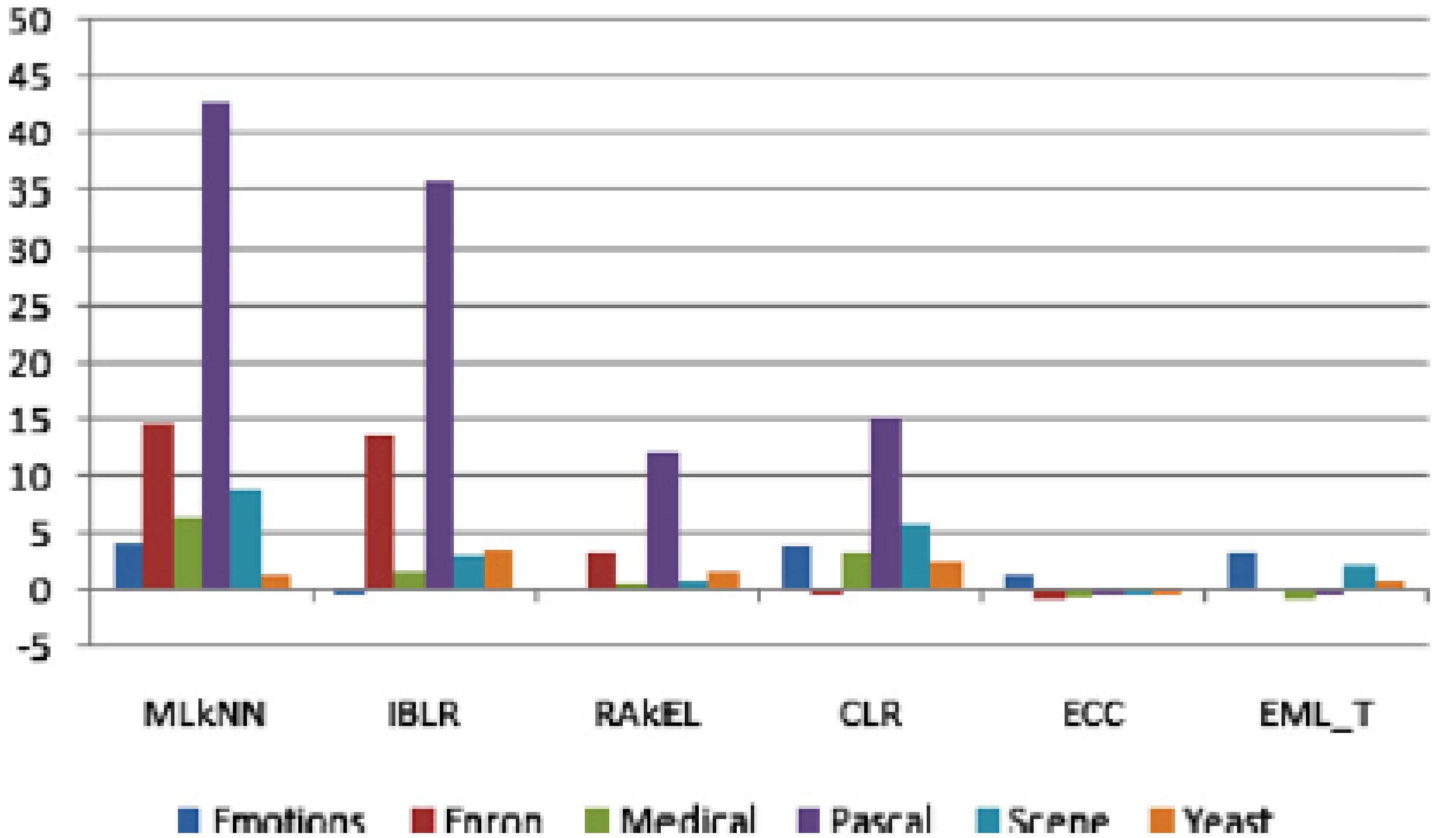


Fig. 2. Performance gains obtained by multi-label classifiers when the threshold is optimized using Micro F_1 and the performance is evaluated using Micro F_1.

4. Experiments and results

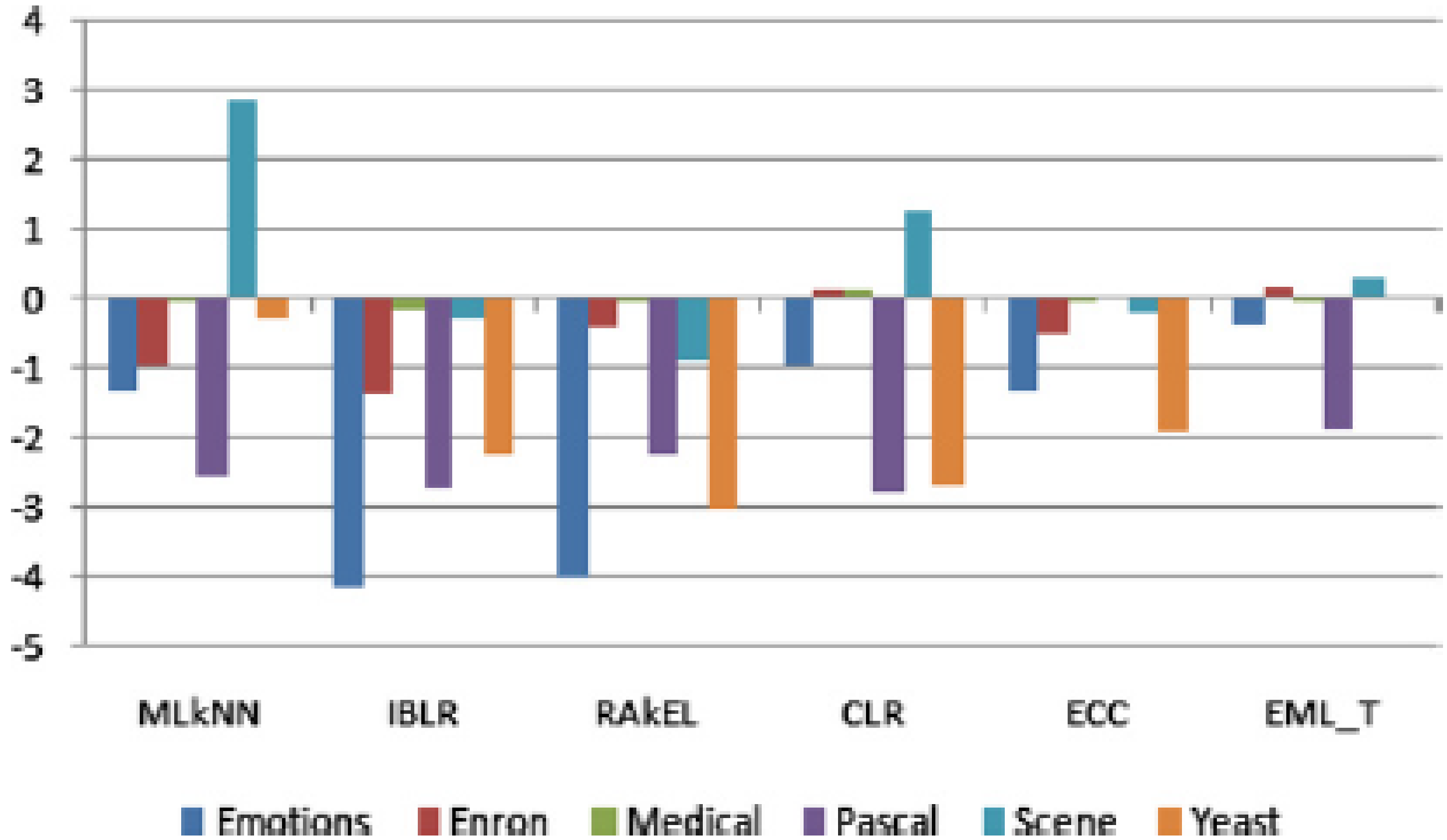


Fig. 3. Difference in performance obtained by multi-label classifiers when the threshold is optimized using Micro F_1 and the performance is evaluated using Hamming loss.

4. Experiments and results

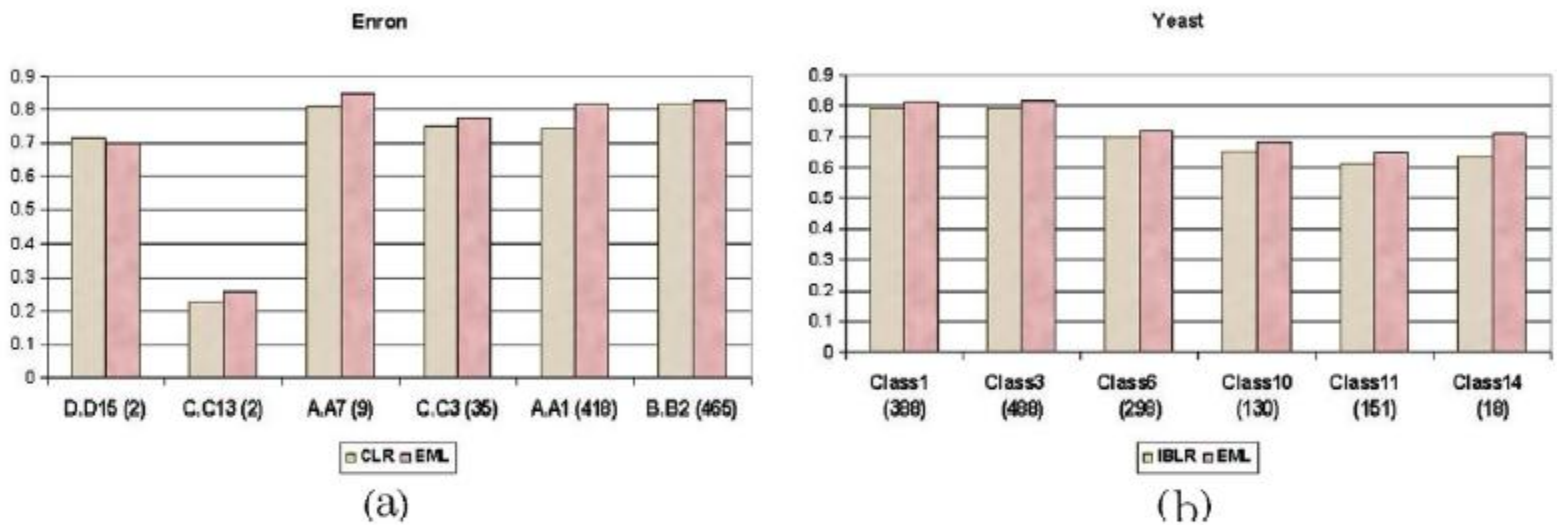


Fig. 4. Performance of individual concepts using AUC. The numbers in bracket shows the total number of samples belong to that concept. (a) Enron. (b) Yeast.

In order to show the effectiveness of the proposed approach in some highly unbalanced concepts, Fig. 4 shows the performance using the Area Under the ROC Curve (AUC) for some highly unbalanced categories in enron and yeast, respectively. The graph clearly indicates that the presented approach (EMLA) has significantly improved the performance in the majority of the highly unbalanced categories. For example, there is an increase of approximatively 3% in performance in categories such as C.C13/A.A7 in Enron.

5. Conclusions

- Heterogeneous ensemble of multi-label learners is proposed to simultaneously tackle both **class imbalance and class correlation problems**

- Ensemble methods are well-known for **overcoming over-fitting** problems and improving the performance of individual classifiers

- It has been shown that the presented approach provides a **very accurate and efficient solution** when compared with the state-of-the-art multi-label methods