Multi-label Classification: A Comparative Study on Threshold Selection Methods

Reem Al-Otaibi[1,2], Peter Flach[1], and Meelis Kull[1]
ra12404, peter.flach, meelis.kull@bristol.ac.uk

[1]Intelligent System Laboratory, Computer Science, University of Bristol
Bristol, United Kingdom
[2]King Abdulaziz University, Saudi Arabia

Abstract. Dealing with multiple labels is a supervised learning problem of increasing importance. However, in some tasks, certain learning algorithms produce a confidence score vector for each label that needs to be classified as relevant or irrelevant. More importantly, multi-label models are learnt in training conditions called operating conditions, which most likely change in other contexts. In this work, we explore the existing thresholding methods of multi-label classification by considering that label costs are operating conditions. This paper provides an empirical comparative study of these approaches by calculating the empirical loss over range of operating conditions. It also contributes two new methods in multi-label classification that have been used in binary classification: score-driven and one optimal.

Keywords: Multi-label learning, Threshold, Cost-sensitive, Context, Operating condition, Label cost

1 Introduction

Multi-label classification differs from traditional single-label classification in that the model needs to predict multiple labels for each instance. Several tasks are relevant to multi-label classification: ranking, bipartition or both. Some multi-label learning models output a score vector for each label and employ one thresholding method in order to be able to output bipartitions [3].

In machine learning and data mining, a model is learnt from data, which is called a training set in the training process, and then this model is applied to new data. However, because such a model is obtained under training conditions, which are called operating conditions, it most probably changes in other contexts. Thus, the model cannot be applied when changes occur and re-training a new model is necessary.

The rest of this paper is organised as follows. Section 2 draws attention to the motivation for this work. The notations used in this paper are introduced in Section 3. Section 4 presents multi-label classification thresholding methods, while Section 5 analyses the performance of different thresholding approaches. Section 6 concludes the paper.

2 Motivation

We summarise the motivation for this work into two main points. First, many thresholding methods have been introduced in the literature. Some of these methods are extensions of single-label methods while others are specifically designed for multi-label classifiers. The threshold can be adjusted in many different ways: label-wise, instance-wise or globally (one threshold). Moreover, most of the multi-label classifiers are meta-classifiers using a single-label classifier as a base classifier. In some thresholding methods, it is important to predict well-calibrated probabilities, as we will see later.

Second, cost-sensitivity is an important research area with many real-world applications. Making better decisions depends not only on minimising the expected errors, but also on minimising the total cost associated with those decisions. Current learning research has focused on binary or multi-class cost-sensitive classification, but not on multi-label classification [5]. A multi-label model can be learnt in a context with training label costs but deployed in other contexts where these costs have been changed.

In this work, we will consider changing label costs between training and deployment; costs can be similar for all labels or they can be varied. In addition, we will focus on applying different thresholding methods and analyse the differences between these methods.

3 Notation

Let D be a multi-label data set where each instance is associated with a set of labels. Let $\mathscr{X}$ be the instance space and $\mathscr{Y}$ be the label space: then $D \subseteq \mathscr{X} \times 2^{\mathscr{Y}}$. L is the number of labels in $\mathscr{Y}$.

Each label l_j has class ratio π_{j_0} and π_{j_1} denoting positive and negative classes. The misclassification for positive and negative classes are represented by c_{j_0} and c_{j_1}, respectively. Total misclassification for a label is $b_j = c_{j_0} + c_{j_1}$. Thus, the relative cost for l_j is $c_j = c_{j_0}/b_j$. Clearly, both c_{j_0} and c_{j_1} can be written in terms of c_j. We assume $b_j = 2$ for all labels; justifications are explained in [2].

The cost-based loss function in multi-label classification is the average loss over all labels:

$$\text{Multi-label Loss Function} \stackrel{\Delta}{=} \frac{1}{L}\sum_{j=1}^{L} Q(t;c_j) \tag{1}$$

$$= \frac{1}{L}\sum_{j=1}^{L} 2\{c_j\pi_{j_0}(1-F_{j_0}(t)) + (1-c_j)\pi_{j_1}F_{j_1}(t)\} \tag{2}$$

where; $F_{j_0}(t)$, $1-F_{j_0}(t)$ and $F_{j_1}(t)$ represent the true positive, false negative and false positive rates for that label at threshold t, respectively.

For simplicity, we will use π_j for label l_j to define the proportion of instances that have this label as true rather than π_{j_0} and π_{j_1}. In addition, π will be used as the average number of labels of instances on D, which is also known as the cardinality of D.

4 Thresholding Approaches

This section presents some methods used to adjust the thresholds in multi-label classification. A link to binary classification will then be introduced.

4.1 Multi-label Classification

Multi-label thresholding methods are grouped into score-based and rank-based methods. There are various ways to select the threshold, such as global threshold or multi-thresholds. A global threshold means that only one threshold applies for all labels, whereas multi-thresholds can be equal to the number of labels (label-wise) or the number of instances (instance-wise).

4.1.1 Score-based Methods Score-based methods tune the threshold based on the score produced by the classifier for each instance-label pair.

Fixed: A global fixed threshold can be used of all labels or different fixed thresholds, one for each label.

Definition 1. *The label-wise fixed threshold choice method is defined as*

$$T_j^{fixed}(c_j) \stackrel{\Delta}{=} t_j \tag{3}$$

SCut Score-based: This method can be a global or label-wise method that adjusts the threshold for each label to achieve a specific loss function using a validation set or cross validation [7].

Definition 2. *Given a label l_j and loss function Q, the label-wise SCut threshold choice method is defined as*

$$T_j^{SCut}(c_j) \stackrel{\Delta}{=} \arg\min_{t_j} Q(t_j; c_j) \tag{4}$$

Score-Driven: Score-driven was used in [2] for binary classification but no multi-label version of this method is found in the literature. We define score-driven in a multi-label setting, which can be either label-wise or global.

Definition 3. *The label-wise score-driven threshold choice method is defined as*

$$T_j^{sd}(c_j) \stackrel{\Delta}{=} c_j \tag{5}$$

The global versions of Fixed, SCut and Score-driven are similar to the binary classification in [2]. It finds one global threshold by putting all labels in one bin and treating them as in a binary classification.

4.1.2 Rank-based Methods Rank-based methods adjust the threshold using the ranking, which can be a label-wise rank or an instance-wise rank.

PCut: The Proportion Cut (PCut) method can be a label-wise or global method that calibrates the threshold(s) from the training data globally or per label. Label-wise PCut sets different thresholds for each label, which guarantees that the predicted positive rate R_j for this label is equal to the training positive rate [7].

Definition 4. *Given a label l_j with positive proportion π_j, the label-wise PCut threshold choice method is defined as (assuming R_j is invertible and $c_j = \pi_j$)*

$$T_j^{PCut}(c_j) \stackrel{\Delta}{=} R_j^{-1}(c_j) \tag{6}$$

The label-wise PCut method is similar to the rate-driven method used in binary classification [2]. In [6], the authors proposed a global PCut, which sets one global threshold that leads to the closest approximation of the average number of labels π between training and deployment.

RCut: The Rank Cut (RCut) method is an instance-wise strategy, which outputs the k labels with the highest scores for each instance at the deployment [7].

Definition 5. *Given an instance x_i, $\hat{y}_i$ is the sorted score list for this instance, the RCut threshold choice method is defined as (assuming F_i is invertible)*

$$T_i^{RCut} \stackrel{\Delta}{=} F_i^{-1}(max_k\{\hat{y}_i\}),\ k = 1,...,L \tag{7}$$

It is important to emphasise that even though k is a fixed parameter, the thresholds using RCut are not. Each instance will have its own threshold that guarantees k labels to be relevant.

MCut: The Maximum Cut (MCut) automatically determines a threshold for each instance that selects a subset of labels with higher scores than others. This leads to the selection of the middle of the interval defined by these two scores [4] as the threshold.

Definition 6. *Given an instance x_i, $\hat{y}_i$ is the sorted score list for this instance, and the MCut threshold choice method is defined as*

$$T_i^{MCut} \stackrel{\Delta}{=} \frac{\hat{y}_i(d) + \hat{y}_i(d+1)}{2}\ ,\ d = \arg\max\{[\hat{y}_i(l) - \hat{y}_i(l+1)], l = 1,...,L\} \tag{8}$$

4.2 Discussion

In summary, some of the abovementioned methods use information about the context to adjust the threshold, whereas other methods do not. For example, PCut1, RCut, MCut do not consider the operating condition when assigning the threshold. Thus, all these methods can be seen as similar to the fixed threshold that uses a fixed number.

Label-wise versions of PCut and SCut compare example scores by fixing the label, while RCut and MCut compare the label scores by fixing the example. MCut assigns a different number of labels per example, whereas RCut assigns the same number of labels for all examples. RCut requires user-specified parameters, but the others do not

Table 1. The Link between Multi-label and Binary ClassificationThresholding Methods

Multi-Label	Applied	Number of Thresholds	Cost-Based	Link to Binary Classification
Fixed	Globally	1	No	Fixed
	Label-wise	L		
PCut	Globally	1	No	-
	Label-wise	L	Yes	Rate-Driven
SCut	Globally	1	Yes	Optimal
	Label-wise	L	Yes	
Score-driven	Globally	1	Yes	Score-Driven
	Label-wise	L	Yes	
RCut	Instance-wise	N	No	-
MCut	Instance-wise	N	No	-

need any parameterisation. The rate-driven and optimal methods in binary classification are similar to PCut and SCut in multi-label classifications, respectively.

Table 1 summarises the differences among all these methods and their link to the binary classification thresholding methods presented in [2].

To clarify the differences among the thresholding methods, the six benchmarks used have been retrieved from the Mulan repository. The key statistics for these data sets are available in Mulan. We here consider two possibilities: all labels have an equal cost or there are different costs for each label. Cost curve analysis is used to understand the differences between all the above methods. It draws the loss on the y-axis against cost on the x-axis [1]. Scatter diagrams are used to determine the loss associated using different costs, in addition to cost curves for identical costs.

Figure 1 compares the cost curves of the global methods in two cases: equal and different costs. A global method assigns only one threshold for the data at deployment. The linear relationship between the losses and costs are represented by the black and green lines, respectively. The black line provides the loss associated with the global PCut, which assigns one threshold to achieve $\pi_{deploy} = \pi_{train}$ independent from the operating condition. The global PCut is also called fixed, but it is fixed to a calibrated number. In B, C and D, the false positive and false negative rates are very close, which justifies the horizontal PCut lines in these plots. The score-driven curves shown in red are based on the scores. In A, B and D, scores are extreme at 1 or 0, whereas in C, they are uniformly distributed.

The pattern in the scatter plots is that if the equal cost curve is (almost) a line, the unequal cost cloud is close to the line; if for equal costs, there is a curve, then the clouds somehow average over the curve, and therefore, end up in the middle under the curve. Two scatter plots were analysed in detail by looking at the behaviour of each label separately. We will first consider the red cloud in the D ”Scene” that shows the loss for the score-driven method when the threshold is equal to the average label cost. We notice that five out of six labels exhibit similar behaviour; the false positive rate is

higher than the false negative rate over the range of the operating conditions, leading to the same result on average. However, in most cases, by setting the threshold to the average cost, the false positive and false negative rates remain constant. This is because the scores are poorly calibrated; most label scores are close to 1 or 0. In contrast, the red cloud in C is scattered in the middle of the red curve. First, we notice that some points have the same average but different loss due to the change in the average of false positive and false negative rates. The labels in the "Yeast" data set behave differently toward the threshold depending on the scores, which are uniformly distributed between 0 and 1.

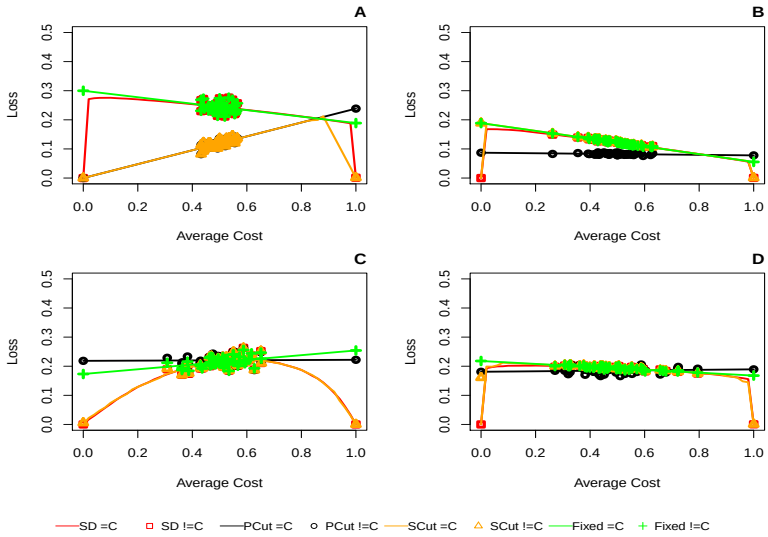

Fig. 1. Cost Curves for Global Threshold Methods using Equal ($= c$) and Different ($! = c$) Costs for all Labels using Logistic Regression as a Base Classifier: A) Enron B) Birds C) Yeast D)Scene

Figure 2 shows the cost curves by applying multiple threshold methods, both label-wise and instance-wise. Regarding both the purple and blue lines, specified as MCut and RCut, respectively, there is also a linear relationship between loss and cost. Although these two methods calibrate the threshold from the data set, both are independent of the cost. The fixed (0.5) and global PCut methods are also called fixed.

5 Performance Evaluation

Table 2 represents the empirical loss of different thresholding methods over range of uniform cost(s) parameter(s). First, we assume that the cost for all labels is equal and measure the empirical loss for all methods. Then, we change the cost between labels so that each label has a different cost. In both cases, we use the binary relevance (BR) using logistic regression as a base classifier.

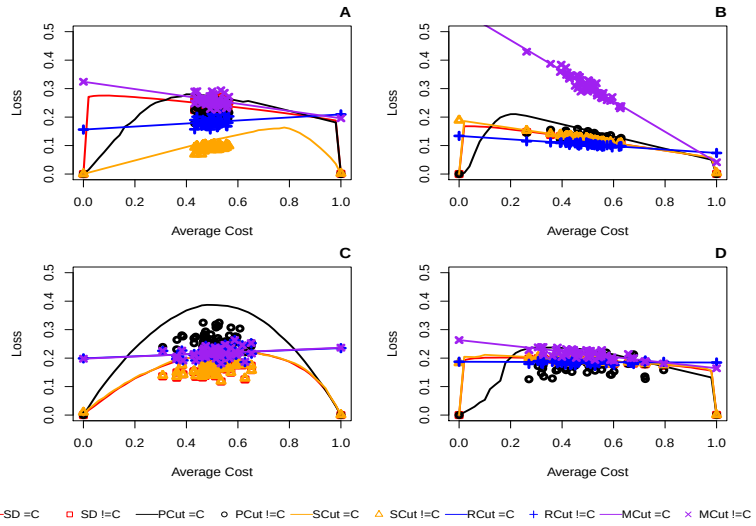

Fig. 2. Cost Curves for Multiple Threshold Methods using Equal ($= c$) and Different ($! = c$) Costs for all Labels using Logistic Regression as a Base Classifier: A) Enron B) Birds C) Yeast D)Scene

As seen in Table 2, if labels have equal costs, global score-driven threshold achieves the lowest loss among the others. On the other hand, the label-wise score-driven threshold leads to better performance if the costs are different. Method that uses the prior frequencies of the labels observed in the trained model such as PCut can lead to the lowest loss when data sets are sparse as seen in "Birds" data set. In general, MCut that assigns the threshold based on the maximum difference between scores has the highest loss.

6 Concluding Remarks

There is a great deal of literature on multi-label learning. To our knowledge, none of these works considers changing label costs between training and deployment. In this paper, we explored the multi-label threshold choice methods: fixed, PCut, SCut, RCut and MCut. In addition, we introduced two new thresholding methods for multi-label methods: score-driven and one optimal threshold. More research on this topic needs to be undertaken to clarify the association between label costs, loss and threshold choice method. Further work should investigate the influence when the operating conditions change for some labels while remaining the same for others.

Acknowledgments

The first author is a PhD student who is sponsored by King Abdulaziz University, Saudi Arabia. This work is also supported by the REFRAME project granted by the European

Table 2. Empirical Losses of Different Thresholding Methods on Multiple Data Sets over range of Operating Conditions (Cost)

		Global				Label-wise			Instance-wise	
		Fixed	PCut	Score-Driven	$SCut_{train}$	PCut	Score-Driven	$SCut_{train}$	RCut	MCut
Loss of Equal Cost	Enron	0.244	0.174	0.231	0.102	0.206	0.231	**0.091**	0.210	0.259
	Birds	0.122	**0.082**	0.114	0.117	0.126	0.114	0.117	0.103	0.309
	Yeast	0.213	0.220	**0.151**	0.152	0.249	**0.151**	0.154	0.217	0.217
	Flags	0.263	0.272	**0.192**	0.195	0.237	**0.192**	0.207	0.329	0.327
	Emotions	0.244	0.244	**0.179**	0.180	0.188	**0.179**	0.184	0.240	0.266
	Scene	0.192	0.185	**0.181**	0.182	0.186	**0.181**	0.183	0.185	0.213
Loss of Different Cost	Enron	0.249	0.173	0.232	0.115	0.116	0.228	**0.089**	0.213	0.260
	Birds	0.125	**0.083**	0.115	0.122	0.127	0.112	0.122	0.104	0.310
	Yeast	0.219	0.221	0.204	0.206	0.252	**0.146**	0.157	0.218	0.219
	Flags	0.267	0.270	0.249	0.253	0.244	**0.195**	0.208	0.329	0.327
	Emotions	0.253	0.248	0.230	0.232	0.191	**0.184**	0.191	0.244	0.261
	Scene	0.197	0.186	0.183	0.189	**0.175**	0.180	0.188	0.187	0.215

Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Engineering and Physical Sciences Research Council in the UK under grant EP/K018728/1.

References

1. Drummond, C., Holte, R.C.: Cost curves: An improved method for visualizing classifier performance. Machine Learning **65**(1) (2006) 95–130
2. Hernández-Orallo, J., Flach, P., Ferri, C.: A unified view of performance metrics: Translating threshold choice into expected classification loss. J. Mach. Learn. Res. **13**(1) (October 2012) 2813–2869
3. Ioannou, M., Sakkas, G., Tsoumakas, G., Vlahavas, I.: Obtaining Bipartitions from Score Vectors for Multi-Label Classification. (October 2010) 409–416
4. Largeron, C., Moulin, C., Gry, M.: Mcut: A thresholding strategy for multi-label classification. In Hollmn, J., Klawonn, F., Tucker, A., eds.: IDA. Volume 7619 of Lecture Notes in Computer Science., Springer (2012) 172–183
5. Lo, H.Y., Wang, J.C., Wang, H.M., Lin, S.D.: Cost-sensitive multi-label learning for audio tag annotation and retrieval. IEEE Transactions on Multimedia **13**(3) (2011) 518–529
6. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II. ECML PKDD '09, Berlin, Heidelberg, Springer-Verlag (2009) 254–269
7. Yang, Y.: A study of thresholding strategies for text categorization. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '01, New York, NY, USA, ACM (2001) 137–145