Evidential Multi-label classification using the Random k-Label sets approach

Sawsan Kanj, Fahed Abdallah and Thierry Denœux

Abstract Multi-label classification deals with problems in which each instance can be associated with a set of labels. An effective multi-label method, named RA*k*EL, randomly breaks the initial set of labels into smaller sets and trains a single-label classifier in each of this subset. To classify an unseen instance, the predictions of all classifiers are combined using a voting process. In this paper, we adapt the RA*k*EL approach under the belief function framework applied to set-valued variables. Using evidence theory makes us able to handle lack of information by associating a mass function to each classifier and combining them conjunctively. Experiments on real datasets demonstrate that our approach improves classification performances.

1 Introduction

Multi-label classification considers problems in which an object may belong simultaneously to multiple classes [4, 5, 10]. Several applications may be subscribed under the multi-label classification problem. In semantic scene classification, each image can be separated into semantic classes as beaches, sunsets or parties [1]. In text categorization, each document may belong to multiple categories such as government, arts and health [6]. In music classification, each song can evoke more than one emotion at the same time, such as amazed, happy, excited, etc. [7].

A lot of algorithms have been proposed for multi-label learning. The existing methods can be categorized into two groups: the *indirect* methods and the *direct* ones [8]. The former one transforms the multi-label classification problem into one or more single-label classification problems, while the latter handles directly the multi-label classification problem.

Sawsan Kanj, Fahed Abdallah and Thierry Denœux
Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, France, e-mail: firstname.lastname@hds.utc.fr

This paper focuses on an effective multi-label learning method introduced in [9]. This method, named RA*k*EL (RAndom-*k*-labEL sets), aims at solving the multi-label classification problem while taking into consideration the correlation between labels. It randomly breaks the set of labels into smaller sets and learns a single-label classifier for each subset. To make a decision, the different predictions for each label are aggregated via voting. In this approach, the user has to identify the number of random label sets, the size of these sets and an adequate threshold in the voting process.

Our goal in this paper is to alleviate the loss of information inherent in the RA*k*EL method (as each base classifier only considers a subset of labels) while accounting for label correlation in a more efficient way. For this purpose, we propose to retain the basic principle of the RA*k*EL approach but to combine the different classifiers in the belief function framework. In [3], a formalism for representing uncertain information has been proposed for manipulating knowledge about set-valued variables. We use this formalism in order to represent and combine information about an unseen instance and to predict its set of labels. To show the effectiveness of this strategy even when using simple classifiers structure, we use Linear Discriminant Analysis (LDA) as the base-level learning method for each classifier. In LDA, each classifier provides information about the object to classify on the form of estimated posterior probabilities. Due to the fact that these outputs can be expressed as set-valued variables, we encode them as mass functions and combine them conjunctively. To make a final decision, we compute the belief function for each label or the maximum of commonality in order to find the whole set of labels to be assigned. The proposed method, called *Evidential-Rakel-LDA* has the advantage of reducing the number of parameters since the decision making process is automatically performed under the belief function framework.

The rest of this paper is organized as follows. Section 2 recalls the background on belief functions for set-valued variables. Section 3 introduces the *Rakel-LDA* method. Section 4 presents experiments on two real datasets and discusses the results. Finally, section 5 concludes the paper.

2 Belief functions on set-valued variables

Let X be a variable taking zero, one or several values in a finite set Ω. Such a variable is said set-valued [3].

To express partial knowledge about a set-valued variable X, we may specify a set A of values that are *certainly* taken by X and a set B of values that are *certainly not* taken by X. The set of subsets of Ω that contain A and have an empty intersection with B is denoted by $\varphi(A,B)$. Let $C(\Omega)$ be the set of all subsets of $\Theta = 2^{\Omega}$ of the form $\varphi(A,B)$, completed by the empty set of Θ.

The theory of belief functions can be applied to describe partial knowledge about set-valued variables by defining a mass function on $\Theta = 2^{\Omega}$. It is clear that the cardinality of $C(\Omega)$ is equal to $3^K + 1$.

The belief and commonality functions are defined, respectively, as follows:

$$bel(A,B) = \sum_{\varphi(C,D)\subseteq\varphi(A,B)} m(C,D) - \emptyset_\Theta, \tag{1}$$

$$q(A,B) = \sum_{\varphi(C,D)\supseteq\varphi(A,B)} m(C,D), \tag{2}$$

where $m(A,B)$ is a notation for $m(\varphi(A,B))$.

As shown in [3], Dempster's rule can be expressed as follows:

$$(m_1 \oplus m_2)(A,B) = \frac{\sum_{\varphi(C,D)\cap\varphi(E,F)=\varphi(A,B)} m_1(C,D)m_2(E,F)}{\sum_{\varphi(C,D)\cap\varphi(E,F)\neq\emptyset_\Theta} m_1(C,D)m_2(E,F)}. \tag{3}$$

Even if the evidential approach reduces the number of focal elements to 3^K+1, this method still has high complexity for large numbers of labels. As an example, if we have 20 labels in the multi-label problem, we may have to handle up to $3.4868e+009$ focal elements. The method proposed in the next section aims to overcome this problem by applying the Evidential formalism to several partitions of the label set and to combine the results under the belief functions framework.

3 Evidential-Rakel-LDA

Let $\mathscr{X} = \mathbb{R}^d$ denote the input space, and let $\Omega = \{\omega_1, \omega_2, \dots, \omega_Q\}$ be the finite set of labels. The multi-label classification problem can be described as follows. Given a training set $\mathscr{D} = \{(x_1,Y_1),\dots,(x_N,Y_N)\}$, of N instances drawn from $\mathscr{X} \times 2^\Omega$, and identically distributed, where x_i is a feature vector describing instance i, and $Y_i \subseteq \Omega$ is the set of labels for that instance, the goal of the multi-label learning is to find a multi-label classifier $\mathscr{H} : \mathscr{X} \rightarrow 2^\Omega$ that can associate a set of labels to each unseen instance.

As in the standard RA*k*EL method, we randomly split the initial set of labels Ω into a number of smaller label sets Ω_j. For each one, the training set of instances, denoted $\mathscr{D}_j$, is deduced from the original dataset $\mathscr{D}$ by replacing the label sets of training instances by their intersections with Ω_j. Inside $\mathscr{D}_j$, each combination of labels is considered as a new class (or group of classes).Using $\mathscr{D}_j$, we train an LDA classifier, denoted h_j (here h_j is a single-label classifier). Note that LDA is used to generate a set of linear functions, one for each group. These functions are built by maximizing the ratio of the between-class variance to the within-class variance. In order to make a decision for an unseen instance x, LDA estimates the posterior probability for each group of the set Ω_j.

In the frame of discernment Ω, the individual classifier outputs are considered as items of evidence. Each output is represented by a mass function on a focal set, noted by $\varphi(A_q,B_q)$ where $A_q,B_q \subseteq \Omega_j$. In other words, A_q is the set of labels assigned to one group and B_q is its complement in Ω_j.

After considering all the items of evidence as items on Ω, we combine them using the Dempster's rule (3) to form the resulting BBA m for an unseen instance. To determine the set of estimated label $\widehat{Y}$ of the unseen instance, we compare the two degrees of belief $bel(\omega, \emptyset)$ and $bel(\emptyset, \omega)$ for each label in Ω [3]:

$$\widehat{Y} = \{\omega \in \Omega / bel(\{\omega\}, \emptyset) \geq bel(\emptyset, \{\omega\})\}. \tag{4}$$

Note here that the decision making process is automatically performed without having to define threshold. As shown by Denœux and Masson [2], we can also calculate the communality function and the maximum of this function can be determined by solving an integer programming problem with non-linear constraints. In this case, another way to calculate $\widehat{Y}$ is to select the set of labels with the largest communality.

4 Experiments

4.1 Evaluation metrics

To evaluate the performance of our method, we calculate different metrics used in the multi-label literature [8].

Hamming Loss: The Hamming Loss metric refers to the percentage of labels that are misclassified, i.e., incorrect labels that are predicted or true labels that are not predicted:

$$\mathscr{H}Loss = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \triangle \widehat{Y}_i|}{Q}, \tag{5}$$

where $\triangle$ denotes the symmetric difference between two sets.

Accuracy: Accuracy measures the degree of closeness between the predicted and the ground truth label sets:

$$\mathscr{A}ccuracy = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|}. \tag{6}$$

F_1 measure: The F_1 measure is defined as the harmonic mean of two other metrics called precision and recall. *Precision* is the fraction of predicted labels that are true, while *recall* is the fraction of true labels that are predicted.

$$\mathscr{P}recision = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|\widehat{Y}_i|}, \tag{7}$$

$$\mathscr{R}ecall = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i|}, \tag{8}$$

and

$$\mathscr{F}_1 = 2.\frac{\mathscr{P}recision.\mathscr{R}ecall}{\mathscr{P}recision + \mathscr{R}ecall}. \tag{9}$$

The smaller the value of the *Hamming Loss*, the better the performance. For the other metrics, higher values correspond to better classification quality.

4.2 Datasets

Our method was experimented using the emotions and scene datasets [1].

The Emotion dataset contains 593 songs described by eight rhythmic features and 64 timbre features. There are six classes, and each song can belong to more than one label according to the emotions generated.

The Scene dataset consists of 2407 natural scene images. There are six different semantic classes. Spatial color moments are used as features. Each image is divided into 49 blocks using 7×7 grid. The mean and variance of each band are computed corresponding to a low-resolution image and to computationally inexpensive texture features, respectively. Each image is then described by $49 \times 2 \times 3$ features [1].

4.3 Results and discussions

We compared our method to the classical RA*k*EL approach based on the LDA method with different threshold values. The number k of labels in each subset was fixed to three for all experiments and the number of classifiers was ranging from 2 to $2 * Q$. Experiments on *Rakel-ADL* were done with all meaningful values for the threshold $(0.1,\ 0.5$ and $0.9)$.

Due to randomization of label space, results are very sensitive to the selected combination of labels. To deal with this negative aspect, we grouped results in batches of 10 classifiers calculated for the same value of k, and we computed the average.

Figures 1 to 3 show the box plots for the different metrics obtained for the emotion and scene datasets. From Figure 1, we can notice that our method performs better than *Rakel-ADL* for different values of threshold in term of *Accuracy* on the two datasets.

Figure 2 shows the performance of the F_1 measure metric. As we can see on the scene dataset, the proposed method yields good performances and it is competitive with the two versions of decision. On the emotion dataset, *Rakel-ADL* performs better for a threshold value equal to 0.1. This is due to the fact that the emotion dataset is more labelled than the scene one (the average number of labels per instance is

[1] http://mulan.sourceforge.net/datasets.html

1.87 for the former, while it is 1.07 for the latter). Decreasing the threshold value can result in taking into account all positive true labels and increasing the value of the *recall* metric.

Figure 3 shows the box plot of the minimum *Hamming Loss* for different methods. On the emotion dataset, our approach shows good performances, while on the scene dataset and for a threshold equal to 0.9 we get the best result. This is due to the fact that increasing the threshold is followed by reducing the number of prediction errors (number of incorrect predicted labels), especially with the scene dataset (80% of instances have a single label).

Tables 1 and 2 show that our approach is suitable to multi-label classification problems under the Rakel approach where we have missing information due to lack of knowledge given by each classifier. Note that the intuitive threshold ($t = 0.5$) gives in average better performances on the *Rakel-ADL* over different values of threshold.

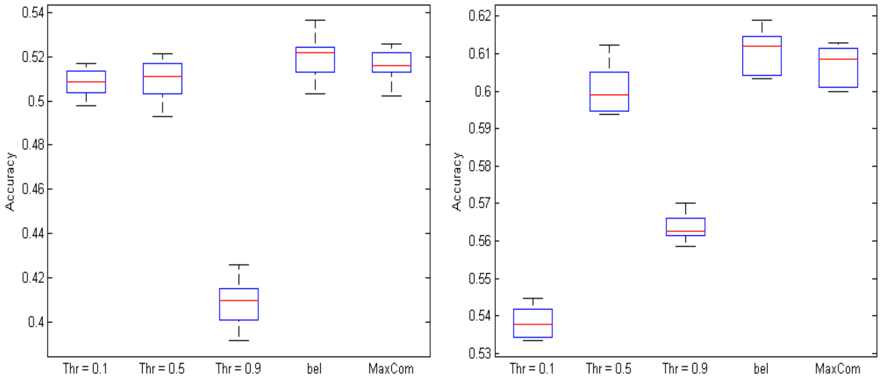

Fig. 1 *Accuracy* box plots with the *Rakel-LDA* method using a threshold values 0.1, 0.5, 0.9, and the *Evidential-Rakel-LDA* method using the belief and the maximum of communality principles. Left figure: for the emotion dataset; right figure: for the scene dataset

	Rakel-LDA **thr =** 0.1	***Rakel-LDA*** **thr =** 0.5	***Rakel-LDA*** **thr =** 0.9	***E-Rakel-LDA*** ***bel***	***E-Rakel-LDA*** ***max of com***
Accuracy	$0.508 \pm 0.006^{(4)}$	$0.509 \pm 0.009^{(3)}$	$0.409 \pm 0.009^{(5)}$	$0.519 \pm 0.009^{(1)}$	$0.516 \pm 0.007^{(2)}$
F_1	$0.621 \pm 0.004^{(1)}$	$0.598 \pm 0.011^{(4)}$	$0.479 \pm 0.011^{(5)}$	$0.607 \pm 0.012^{(2)}$	$0.605 \pm 0.009^{(3)}$
HLoss	$0.301 \pm 0.006^{(5)}$	$0.239 \pm 0.003^{(4)}$	$0.236 \pm 0.003^{(2)}$	$0.235 \pm 0.006^{(1)}$	$0.238 \pm 0.004^{(4)}$

Table 1 Experimental results (mean±std) of the compared algorithms on the emotions dataset

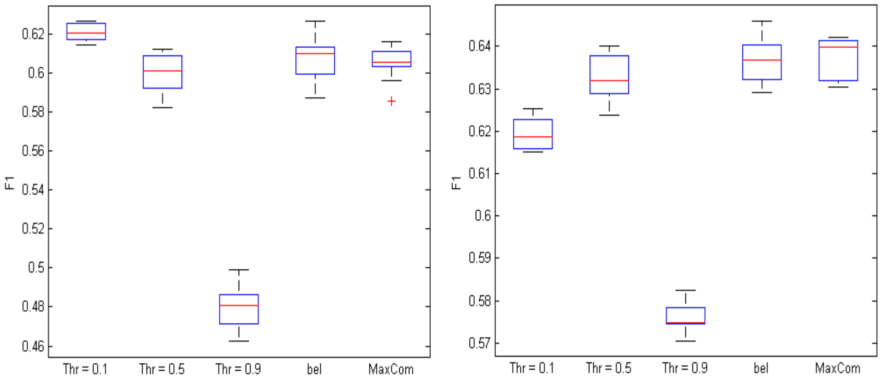

Fig. 2 F_1 box plots with the *Rakel-LDA* method using a threshold values 0.1, 0.5, 0.9, and the *Evidential-Rakel-LDA* method using the belief and the maximum of communality principles. Left figure: for the emotion dataset; right figure: for the scene dataset

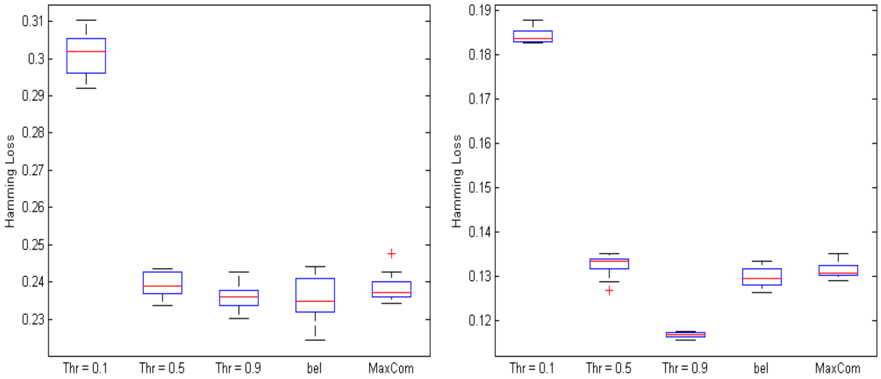

Fig. 3 *Hamming Loss* box plots with the *Rakel-LDA* method using a threshold values 0.1, 0.5, 0.9, and the *Evidential-Rakel-LDA* method using the belief and the maximum of communality principles. Left figure: for the emotion dataset; right figure: for the scene dataset

	Rakel-LDA **thr =** 0.1	***Rakel-LDA*** **thr =** 0.5	***Rakel-LDA*** **thr =** 0.9	***E-Rakel-LDA*** ***bel***	***E-Rakel-LDA*** ***max of com***
Accuracy	$0.538 \pm 0.004^{(5)}$	$0.601 \pm 0.006^{(3)}$	$0.564 \pm 0.004^{(4)}$	$0.611 \pm 0.005^{(1)}$	$0.607 \pm 0.005^{(2)}$
F_1	$0.612 \pm 0.004^{(4)}$	$0.632 \pm 0.006^{(3)}$	$0.576 \pm 0.004^{(5)}$	$0.636 \pm 0.005^{(2)}$	$0.637 \pm 0.005^{(1)}$
HLoss	$0.184 \pm 0.002^{(5)}$	$0.132 \pm 0.003^{(4)}$	$0.117 \pm 0.001^{(1)}$	$0.129 \pm 0.002^{(2)}$	$0.131 \pm 0.002^{(3)}$

Table 2 Experimental results (mean±std) of the compared algorithms on the scene dataset

5 Conclusion

A variant of the RA*k*EL method for multi-label classification has been proposed, based on the theory of belief functions. Our approach uses the formalism developed in [3] to define belief functions for set-valued variables. This framework allows us to combine the outputs from base classifiers in a more efficient way than the voting process used in the reference method. Experimental results demonstrate the effectiveness of the approach.

References

1. M.R. Boutell, J. Shen, and C.M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37:(9):1757–1771, 2004.
2. T. Denœux and M.-H. Masson. Evidential reasoning in large partially ordered sets. Application to multi-label classification, ensemble clustering and preference aggregation. *Annals of Operations Research*, Accepted for publication, 2011. doi:10.1007/s10479-011-0887-2.
3. T. Denoeux, Z. Younes, and F. Abdallah. Representing uncertainty on set-valued variables using belief functions. *Artificial Intelligence*, 174:479–499, 2010.
4. N. Ghamrawi and A. McCallum. Collective multi-label classification. In *14th ACM international conference on Information and knowledge management*, 2005.
5. J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Proc. of the 20th European Conference on Machine Learning (ECML 2009)*, 2009.
6. R. Schapire and Y. Singer. Boostexter: a boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.
7. K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), pp. 325-330*, 2008.
8. G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.
9. G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Proc. 18th European Conference on Machine Learning*, 17-21 September 2007.
10. Z. Younes, F. Abdallah, T. Denoeux, and H. Snoussi. A dependent multilabel classification method derived from the k-nearest neighbor rule. *EURASIP Journal on Advances in Signal Processing*, 2011. Article ID 645964, 14 pages, doi:10.1155/2011/645964.