Belief Theory for Large-Scale Multi-Label Image Classification

Amel Znaidia & Hervé Le Borgne & Céline Hudelot

Abstract

Classifier combination is known to generally perform better than each individual classifier by taking into account the complementarity between the input pieces of information. Dempster-Shafer theory is a framework of interest to make such a fusion at the decision level, and allows in addition to handle the conflict that can exist between the classifiers as well as the uncertainty that remains on the sources of information. In this contribution, we present an approach for classifier fusion in the context of large-scale multi-label and multi-modal image classification that improves the classification accuracy. The complexity of calculations is reduced by considering only a subset of the frame of discernment. The classification results on a large dataset of $18,000$ images and 99 classes show that the proposed method gives higher performances than of those classifiers separately considered, while keeping tractable computational cost.

Keywords Demspster-Shafer theory, multi-label classification, multi-modal classification, classifier fusion.

1 Introduction

Image annotation consists in describing an image content according to a finite number of concepts. This problem is usually posed as a set of binary classification tasks,

Amel Znaidia
CEA, LIST, Laboratory of Vision and Content Engineering, e-mail: amel.znaidia@cea.fr, This work was supported by grants from DIGITEO and Région Ile-de-France.

Hervé Le Borgne
CEA, LIST, Laboratory of Vision and Content Engineering e-mail: herve.le-borgne@cea.fr

Céline Hudelot
MAS Laboratory, Ecole Centrale de Paris e-mail: celine.hudelot@ecp.fr

which means to address both image description and visual concept learning. Concerning the first step, images are commonly described using only visual content such as color, texture or shape etc. However, in practice an important gap remains between visual descriptors and the semantic content of images [12].

Therefore, the use of multiple classifiers trained on different modalities (visual, textual ...) and features becomes more popular due to the fact that classifiers are different and informative [5, 7]. Thus, the fusion of their decisions can yield to higher performance than the best individual classifier [4].

Most commonly, straightforward fusion approaches, such as majority voting, maximum and averaging [13] have been used in the literature. According to Tax *et al.* [13] simple average is the optimal linearly combining rule, only if the individual classifiers exhibit both identical performances and correlations between estimation errors. Otherwise, Dempster-Shafer theory [11] is particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers. However, it suffers from a high computational cost, in particular when the number of classes (*i.e* the frames of discernment) is large. To encounter this limitation, Denoeux *al.* [2] proposed a method to reduce the complexity of manipulating and combining mass functions, when belief functions are defined over a suitable subset of the frame of discernment equipped with a lattice structure. This approach was applied for multi-label classification based on the Evidential KNN classifier. For a problem with $\mathbf{C}$ classes, this method reduces the complexity from $2^{2^{\mathbf{C}}}$ to $3^{\mathbf{C}}+1$. Althougth such a reduction is impressive, the problem remains intractable when C is above 10, that is quite common for a multimedia classification problem, for which $\mathbf{C}$ can reach 100 or 1000.

The most similar prior work is [9], which combine textual and visual classifiers based on Dempster's rule to improve the classification accuracy. However, their system was applied for single-label classification task, for a small dataset ($\approx 1,200$ images) and only for *six* classes of emotions.

In this work, we aim at improving the classification accuracy based on classifier fusion in the Dempster-Shafer theory to handle the uncertainty and the conflict that can exist between different classifiers and to assess the discrepancy between various sources of information. The major difference between our work and aforementioned efforts is that we address the problem of combination in a multi-label classification task for a large problem: to the best of our knowledge, this is the first attempt to apply Dempster theory for a multimodal multi-label image classification for a large dataset ($\approx 18,000$ images) and a large variety of categories simultaneously (scene, event, objects, image quality and emotions ≈ 99 concepts). First, we convert the classifier output probabilities into consonant mass functions using the inverse pignistic transform [3]. Secondly, these mass functions are combined in the belief theory using Dempster's rule [11]. Since Average rule has been widely used in the literature, and it outperforms other conventional methods (Maximum, Product, Majority voting), we use it as a baseline to compare with the Dempster's rule.

The remainder of the paper is organized as follows. The background on belief functions is first recalled in section 2. The proposed approach for large scale multi-label image classification is presented in section 3, and experimental results are reported and discussed in section 4. Section 5 concludes this paper.

2 Basics of Dempster-Shafer Theory

In Dempster-Shafer (DS) theory [11], a *frame of discernment* Ω is defined as the set of all hypothesis in a certain domain. A basic belief assignement (BBA) is a function m that defines the mapping from the power set of Ω to the interval $[0,1]$ and verifies:

$$m : 2^{\Omega} \rightarrow [0,1] \tag{1}$$

$$\sum_{A \in 2^{\Omega}} m(A) = 1 \tag{2}$$

The quantity $m(A)$ can be interpreted as a measure of the belief that is commited exactly to A, given the available evidence. A subset $A \in 2^{\Omega}$ with $m(A) > 0$ is called a *focal element* of m. In DS theory, two functions of evidence can be deduced from m and its associated focal elements, belief function Bel and plausibility function Pl. $Bel(A)$ is the measure of the total belief committed to a set A. The belief function is defined as a mapping $Bel : 2^{\Omega} \rightarrow [0,1]$ that satisfies $Bel(\emptyset) = 0, Bel(\Omega) = 1$ and for each focal element A, we have:

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \tag{3}$$

The *plausibility* of A, $Pl(A)$, represents the amounts of belief that could potentially placed in A and defined as:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{4}$$

2.1 Dempster's combination rule

When there are many sources of information defined on the same frame of discernment, the mass functions from different sources are combined under the normalized Dempster's combination rule [11].

$$m_{1-2}(A) = m_1 \oplus m_2 = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}, \ \forall \ A \subseteq \ \Omega, \ A \neq \emptyset \\ 0 \qquad\qquad\qquad\qquad if \ \ A = \emptyset \end{cases} \tag{5}$$

where $k = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ represents the degree of conflict between the two sources. If $k = 1$ the two evidences are in conflict and they can not be combined.

3 Proposed Multi-Label Classification System

In the context of multi-label and multi-modal classification problem, each image can belongs to one or more than one class. Formally, let $\Omega = \{w_1,...,w_C\}$ be the set of labels or classes. The frame of discernment of the multi-label extended DS theory is not the set of all possible single hypotheses but its power set $\Theta = 2^{|\Omega|}$. Given a training set $T = \{(X_1,Y_1),...,(X_N,Y_N)\}$ of N labelled images, where $X_i = \{x_i^1 \dots x_i^L\}$ represents the feature vector of image I_i extracted from L modalities and Y_i the corresponding set of labels, our goal is to predict the set of lables that describe the image content. The flowchart of the proposed system is presented in Figure 1.

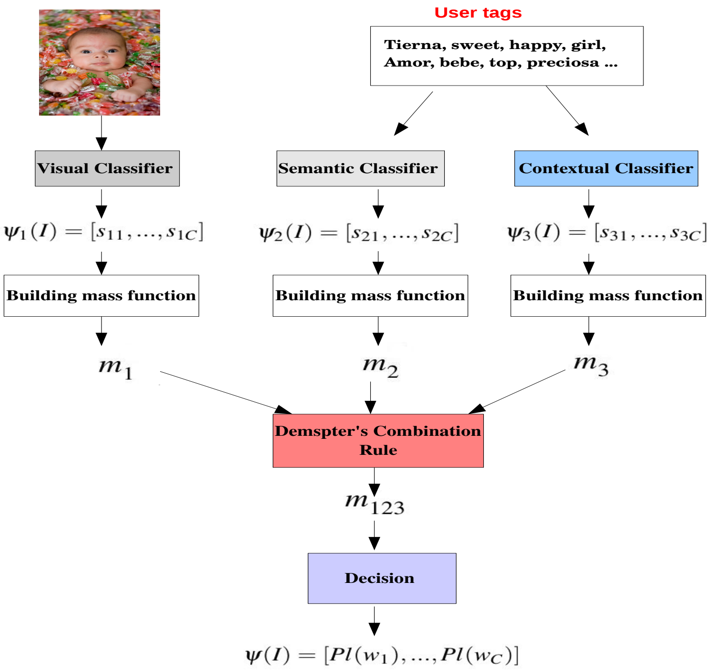
User tags
Tierna, sweet, happy, girl,
Amor, bebe, top, preciosa ...
Visual Classifier
Semantic Classifier
Contextual Classifier
$\psi_1(I) = [s_{11},...,s_{1C}]$
$\psi_2(I) = [s_{21},...,s_{2C}]$
$\psi_3(I) = [s_{31},...,s_{3C}]$
Building mass function
Building mass function
Building mass function
m_1
m_2
m_3
Demspter's Combination Rule
m_{123}
Decision
$\psi(I) = [Pl(w_1),...,Pl(w_C)]$

Fig. 1 Flowchart of the proposed system. First, the classifier output scores ψ_i are normalized to sum to one. Secondly, the obtained probabilities are transformed into mass function using the inverse pignistic transform. A combination is performed to obtain the final mass function, used to compute the plausibility fo decision making.

Assume that we have Q classifiers, denoted by $\psi_1, \psi_2, \dots \psi_Q$ to be combined. Given an input image I, each classifier ψ_i produced an output $\psi_i(I)$ defined as :

$$\psi_i(I) = [s_{i1},...,s_{iC}] \tag{6}$$

where s_{ij} indicates the degree of confidence in saying that 'image I belongs to class w_j according to classifier ψ_i'. First, classifier output are normalized to obtain a probability distribution p_i over Ω as follows:

$$p_i(w_j) = \frac{s_{ij}}{\sum_{k=1}^{C} s_{ik}}, \;\; for \;\; j = 1, ..., C \tag{7}$$

For each classifier ψ_i, the element of Ω are ranked by decreasing probabilities such that $p(w_1) \geq p(w_2) \geq ... \geq p(w_{|\Omega|})$. The class label of an instance may be represented by a variable Y taking values in $\Theta = 2^{|\Omega|}$. Thus, expressing partial knowledge of Y in the Dempster-Shafer framework may imply storing 2^{2^C} numbers. Based on this ordering, instead of considering the whole power set of Θ, we will focus on a smaller subset $R(\Omega)$ defined by:

$$R(\Omega) = \{A_k = \{w_1, ..., w_{k+1}\}, \forall \, k = 1, ..., |\Omega| - 1\} \tag{8}$$

The size of this subset is $|\Omega| - 1$, it is thus much smaller than 2^{2^C} while being rich enough to express evidence because we consider only the most probable subsets. Secondly, we convert the obtained probabilities into consonant mass functions using the inverse pignistic transform [3]. The consonant mass function derived from these probabilities verifies :

$$m : 2^{\Omega} \rightarrow [0,1], \sum_{A_k \in 2^{\Omega}} m(A_k) = 1 \tag{9}$$

$$\begin{aligned} m(\{w_1, w_2, ..., w_i\}) &= i \times \; [p(w_i) - p(w_{i+1})] \; \forall \, i \; < \; |\Omega| \\ m(\{w_1, w_2, ..., w_{|\Omega|}\}) &= |\Omega| \times p(w_{|\Omega|}) \\ m(X) &= 0 \;\; \forall \, X \notin R(\Omega). \end{aligned} \tag{10}$$

In this work, we choose to combine the obtained consonant mass functions from different classifiers using the normalized Dempster's rule [11]. Other combination rules can be used [10]. Let m_i be the mass function of the source i, the combination of n mass function (corresponding to n classifiers) is defined according to Dempster's combination rule as follows:

$$m_{1-n}(A) = \begin{cases} \dfrac{\sum\limits_{\cap_{k=1}^{n} b_k = A} \prod\limits_{i=1}^{n} m_i(b_i)}{1 - \sum\limits_{\cap_{k=1}^{n} b_k = \emptyset} \prod\limits_{i=1}^{n} m_i(b_i)}, \; \forall A \subseteq \; \Omega, \; A \neq \emptyset, b_k \in R_k(\Omega) \\ \qquad\qquad 0 \qquad\qquad if \;\; A = \emptyset \end{cases} \tag{11}$$

Let $\hat{Y}$ be the predicted label set for instance x. To decide whether to include each class or not, we compute the degree of plausibility $Pl(w_j)$ that the true label set Y contains the label w_j, and the degree of plausibility $Pl(\bar{w}_j)$ that it does not contain the label w_j using formula (4). We then define $\hat{Y}$ as:

$$\hat{Y} = \{w_j \in \Omega | Pl(w_j) \geq Pl(\bar{w}_j)\} \tag{12}$$

4 Experimental Results

4.1 Dataset & Experimental setup

The Dataset used in our experiments is the MIR Flickr dataset [6] containing 8,000 images for training and 10,000 for testing belonging to 99 concept classes. These concepts describe the scene ’indoor, outdoor, landscape...‘, depicted objects ’car, animal, person...‘, the representation of image content ’portrait, graffiti, art...‘, events ’travel, work...‘, or quality issues ’overexposed, underexposed, blurry...‘ and emotions ’funny, cute, nice, scary ... ‘. Figure 2 shows samples of images taken from the dataset with their annotated concepts.

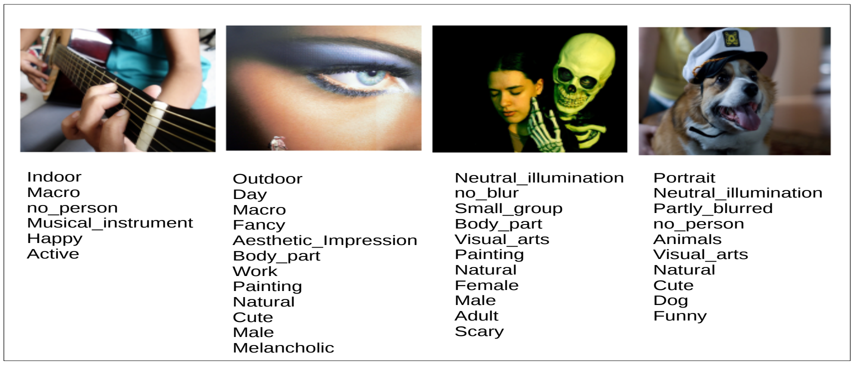

Fig. 2 Samples of images taken from the dataset with their annotated concepts.

Features We used two textual descriptors and one visual descriptor. The textual descriptor is based on semantic similarity between tags and visual concepts. Two distances were used: one based on the Wordnet ontology and one based on social networks. Each feature vector is of size 99 (the number of concepts). The visual component considers various local and global features, such as colour and edge features. The visual feature vector is of size 890. More details about the used features can be found in [14]. Each feature vector was used to train a classifier using the Fast Shared Boosting algorithm [8]. Three measures are used to test the performance of the individual classifiers and the different combinations: Mean Average Precision (MAP), Equal Error Rate (ERR) and Area Under Curve (AUC).

4.2 Results and discussions

Table 1 displays the performances of individual classifiers and the two considered combination rules in terms of MAP, ERR and AUC. These results show that individual classifiers exhibit identical performances with a small superiority to the contextual classifier. Since Average rule has been widely used in the literature, and it outperforms other conventional methods (Maximum, Product, Majority voting), we will use it as a baseline to compare to the Dempster's rule.

Classifier	Visual Classifier	Contextual Classifier	Semantic Classifier	Dempster's rule	Average rule
MAP	29.86	32.13	29.24	39.05	**40.21**
EER	28.93	31.50	35.69	26.21	**24.64**
AUC	77.59	74.32	68.44	80.79	**82.29**

Table 1 Comparative Performance of individual classifiers in terms MAP, ERR and AUC.

By comparing these results, we can see that the combination of classifiers for both Dempster's rule and average rule gives better results than the best individual classifier. We obtain a gain of $\approx 10\%$ in terms of classification accuracy and consequently, reducing the classification error by $\approx 9\%$. For this dataset, we observe that the average rule achieve slightly better performances. These results may be explained by the performance of the individual classifiers which exhibit both identical performances and correlations between estimation errors. In addition, we train individual classifiers with unbalanced data over classes which can generate unreliable confidences (*e.g.* caused by a small training set or by overtraining).

The average rule is hardly ever theoretically optimal, but performs sometimes surprisingly good except for some classes as shown in Table 2. For these challenging classes, Dempster'rule performs much better than the average rule especially when considering ensembles of 'good‘ and 'bad‘ classifiers, then using the average rule to combine the classification results will not be a good choice. We compare Dempster's rule to the ImageClef 2011 Winner [1] for these classes. The proposed method outperforms the state of art [1] for such type of classes. We can notice that the Belief theory seems to offer a significant advantage to such situations. It is particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers.

Classes	Visual	Contextual	Semantic	Dempster	Average	ImageClef 2011 Winner [1]
Travel	18.85	14.78	17.55	**22.12**	14.57	16.72
Technical	08.19	06.37	04.52	**12.85**	07.24	08.51
Boring	07.28	07.78	07.63	**15.88**	08.79	09.94
Bird	17.55	51.71	56.08	**61.52**	58.77	58.71
Insect	14.26	47.84	46.44	**58.08**	53.12	45.21
Airplane	05.36	44.36	42.53	**61.66**	59.32	22.93
Skateboard	00.27	10.29	21.54	**28.42**	11.46	00.56
Scary	18.46	08.31	14.10	**19.02**	11.29	16.39

Table 2 Comparative Performance of individual classifiers, Dempster, Average and the ImageClef 2011 Winner [1] for some challenging classes in terms of Mean Average Precision (MAP).

5 Conclusion

In this paper, we presented a system for combining classifiers using Belief theory for large-scale multi-label image classification. When individual classifiers present similar performances, results have shown that using simple rules such as averaging can be a good choice. While, for conflicting classifiers, the Belief theory seems to be an interesting framework to handle the uncertainty and the conflict that can exist between different classifiers. One direction for future research is to take into account the classifier reliability while combining. An additional direction is to construct mass functions directly in the classifiers.

References

1. A. Binder, W. Samek, M. Kloft, C. Müller, K.-R. Müller, and M. Kawanabe. The joint submission of the tu berlin and fraunhofer first (tubfi) to the imageclef2011 photo annotation task. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.
2. T. Denoeux and M. Masson. Evidential reasoning in large partially ordered sets. *Annals of Operations Research*, May 2011.
3. D. Dubois, H. Prade, and P. Smets. New semantics for quantitative possibility theory. In *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ECSQARU '01, pages 410–421, London, UK, 2001. Springer-Verlag.
4. R. P. W. Duin. The combining classifier: To train or not to train? In *ICPR (2)*, pages 765–770, 2002.
5. M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 902 – 909, jun 2010.
6. M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.
7. M. Kawanabe, A. Binder, C. Muller, and W. Wojcikiewicz. Multi-modal visual concept classification of images via markov random walk over tags. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV)*.
8. H. Le Borgne and N. Honnorat. Fast shared boosting for large-scale concept detection. *Multimedia Tools and Applications*, pages 1–14, 2010.
9. N. Liu, E. Dellandréa, B. Tellez, and L. Chen. Associating textual features with visual ones to improve affective image classification. In *International Conference on Affective Computing and Intelligent Interaction (ACII2011)*, Oct. 2011.
10. B. Quost, M.-H. Masson, and T. Denoeux. Classifier fusion in the dempster–shafer framework using optimized t-norm based combination rules. *Int. J. Approx. Reasoning*, 52:353–374, March 2011.
11. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.
12. A. W. M. Smeulders, S. Member, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.
13. D. M. Tax, M. van Breukelen, R. P. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, Sept. 2000.
14. A. Znaidia, H. L. Borgne, and A. Popescu. Cea list's participation to visual concept detection task of imageclef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.