Evidential Multi-Label Classification Approach to Learning from Data with Imprecise Labels

Zoulficar Younes, Fahed Abdallah, and Thierry Denœux

UMR CNRS 6599 Heudiasyc,
Université de Technologie de Compiègne, France
{firstname.lastname}@hds.utc.fr

Abstract. Multi-label classification problems arise in many real-world applications. Classically, in order to construct a multi-label classifier, we assume the existence of a labeled training set, where each instance is associated with a set of labels, and the task is to output a label set for each unseen instance. However, it is not always possible to have perfectly labeled data. In many problems, there is no ground truth for assigning unambiguously a label set to each instance, and several experts have to be consulted. Due to conflicts and lack of knowledge, labels might be wrongly assigned to some instances. This paper describes an evidence formalism suitable to study multi-label classification problems where the training datasets are imperfectly labelled. Several applications demonstrate the efficiency of our apporach.

1 Introduction

In multi-label classification problems, each object may belong simultaneously to several classes, contrary to standard single-label problems where objects belong to only one class. Multi-label classification methods have been increasingly required by modern applications where the target classes are not exclusive and an object may belong to an unrestricted set of classes instead of exactly one. For instance, in natural scene classification, each image may belong to several semantic classes, such as sea and sunset [1].

Several methods have been proposed for multi-label learning. These methods can be categorized into two groups. A first group contains the *indirect* methods that transform the multi-label classification problem into a set of binary classification problems (Binary relevance approach (BR): a binary classifier for each class or pairwise classifiers) [12] [11] [6] or into multi-class classification problem (Label powerset approach (LP): each subset of classes is considered as a new class) [9]. A second group consists in extending common learning algorithms and making them able to manipulate multi-label data *directly* [10].

Usually, multi-label classification tasks are based on training datasets where each instance is associated with a perfectly known set of labels. In practice, gathering such high quality information is not always feasible at a reasonable cost. In many problems, however, there is no ground truth for assigning unambiguously a label set to each instance, and the opinions of one or several experts have to be

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, LNAI 6178, pp. 119–128, 2010.

elicited. Typically, an expert may express lack of confidence for assigning exactly one label set. If several experts are consulted, some conflicts will inevitably arise. This again will introduce some uncertainties in the labeling process.

In [10] and [4], an evidential formalism for handling uncertainty on the classification of multi-labeled data has been presented. This formalism extends all the notions of Dempster-Shafer (D-S) theory [7] to the multi-label case with only a moderate increase in complexity as compared to the classical case. Based on this formalism, an evidence-theoretic k-NN rule for multi-label classification has been presented. The proposed method, called EML-kNN for Evidential Multi-Label k-Nearest Neighbor, generalizes the single-label evidence-theoretic k-NN rule [2] to the multi-label case. Thus, an unseen instance is classified on the basis of its k nearest neighbors under the D-S framework.

In [10], we applied our method on several benchmark datasets where all instances were perfectly labelled. We also noticed that our evidential formalism for set-valued variables allows us to express ambiguities and uncertainties when the available data used to train the multi-label classifier are imprecisely labelled. As far as our knowledge, such imprecise data are not available from real-world problems. Thus, in order to show the performance of EML-kNN in such cases and demonstrate its effectivness, we propose a labeling process to randomly simulate imprecise multi-labelled data.

The remainder of the paper is organized as follows. Section 2 describes the evidence formalism for multi-label case. Section 3 recalls the evidence-theoretic k-NN rule for multi-label classification. Section 4 presents experiments on some real datasets and shows the effectiveness of our approach to handle imprecise data. Finally, Section 5 makes concluding remarks.

2 Evidence Formalism

The Dempster-Shafer (D-S) theory is a formal framework for representing and reasoning with uncertain and imprecise information. Different approaches to single-label classification in the framework of evidence theory have been presented in the literature [3] [2]. This theory is usually applied to handle uncertainty in problems where *only one single hypothesis* is true. However, there exist problems where *more than one hypothesis* are true at the same time, e.g., the multi-label classification task. Let Ω denote the set of all hypotheses in a certain domain, e.g., in classification, Ω is the set of all possible classes. The frame of discernment of the evidence formalism for multi-label case is not Ω, as in the single label classification problem, but its power set $\Theta = 2^{\Omega}$. A mass function m is thus defined as a mapping from the power set of Θ to the interval $[0, 1]$. As proposed in [4], instead of considering the whole power set of Θ, we will focus on the subset $\mathcal{C}(\Omega)$ of 2^{Θ} defined as:

$$\mathcal{C}(\Omega) = \{\varphi(A, B) |\ A \cap B = \emptyset\} \cup \{\emptyset_{\Theta}\} \tag{1}$$

where $\emptyset_\Theta$ represents the conflict in the frame 2^Θ, and for all $A,\ B \subseteq \Omega$ with $A \cap B = \emptyset$, $\varphi(A, B)$ is the set of all subsets of Ω that include A and have no intersection with B:

$$\varphi(A,B) = \{C \subseteq \Omega |\ C \supseteq A\ and\ C \cap B = \emptyset\}. \tag{2}$$

The size of the subset $\mathcal{C}(\Omega)$ of 2^Θ is equal to $3^{|\Omega|} + 1$ and is thus much smaller than the size of 2^Θ ($|2^\Theta| = 2^{2^{|\Omega|}}$). Consequently, this formulation reduces the complexity of multi-label problems, while being rich enough to express evidence in many realistic situations. The chosen subset $\mathcal{C}(\Omega)$ of 2^Θ is closed under intersection, i.e., for all $\varphi(A, B)$, $\varphi(A', B') \in \mathcal{C}(\Omega)$, $\varphi(A, B) \cap \varphi(A', B') \in \mathcal{C}(\Omega)$. Based on the definition of $\varphi(A, B)$, we can deduce that:

$$\varphi(\emptyset, \emptyset) = \Theta, \tag{3}$$

$$\forall A \subseteq \Omega,\ \varphi(A, \bar{A}) = \{A\}, \tag{4}$$

$$\forall A \subseteq \Omega,\ A \neq \emptyset,\ \varphi(A, A) = \emptyset_\Theta. \tag{5}$$

By convention, $\emptyset_\Theta$ will be represented by $\varphi(\Omega, \Omega)$.

For any $\varphi(A, B)$, $\varphi(A', B') \in \mathcal{C}(\Omega)$, the intersection operator over $\mathcal{C}(\Omega)$ is defined as follow:

$$\varphi(A,B) \cap \varphi(A',B') = \begin{cases} \varphi(A \cup A', B \cup B') & if\ A \cap B' = \emptyset\ and\ A' \cap B = \emptyset \\ \varphi(\Omega, \Omega) & otherwise, \end{cases} \tag{6}$$

and the inclusion operator over $\mathcal{C}(\Omega)$ is defined as:

$$\varphi(A,B) \subseteq \varphi(A',B') \iff A \supseteq A'\ and\ B \supseteq B'. \tag{7}$$

A mass function m on $\mathcal{C}(\Omega)$ can be represented with the following two equations:

$$m : \mathcal{C}(\Omega) \longrightarrow [0,1] \tag{8}$$

$$\sum_{\varphi(A,B) \in \mathcal{C}(\Omega)} m(\varphi(A,B)) = 1. \tag{9}$$

For convenience of notation, $m(\varphi(A, B))$ will be simplified to $m(A, B)$. For any $\varphi(A, B) \in \mathcal{C}(\Omega)$, the belief and plausibility functions are now defined as:

$$bel(A,B) = \sum_{\varphi(\Omega,\Omega) \neq \varphi(A',B') \subseteq \varphi(A,B)} m(A',B'), \tag{10}$$

and

$$pl(A,B) = \sum_{\varphi(A',B') \cap \varphi(A,B) \neq \varphi(\Omega,\Omega)} m(A',B'). \tag{11}$$

Given two independent bodies of evidence over the same frame of discernment like $\mathcal{C}(\Omega)$, the aggregated mass function, denoted by m_{12}, obtained by combining the mass functions m_1 and m_2 of the two bodies of evidence using the unnormalized Dempster's rule is calculated in the following manner:

$$m_{12}(A,B) = \sum_{\varphi(A',B') \cap \varphi(A'',B'') = \varphi(A,B)} m_1(A',B') m_2(A'',B''). \tag{12}$$

This rule is commutative and associative, and has the vacuous mass function ($m(\emptyset, \emptyset) = 1$) as neutral element.

3 Evidential Multi-Label *k*-NN

Problem. Let $\mathcal{X} = R^P$ denote the domain of instances and let $\Omega = \{\omega_1, \ldots, \omega_Q\}$ be the finite set of labels. The multi-label classification problem can now be formulated as follows. Given a set $\mathcal{S} = \{(\mathbf{x}_1, A_1, B_1), \ldots, (\mathbf{x}_M, A_M, B_M)\}$ of M training examples, where $\mathbf{x}_i \in \mathcal{X}$, $A_i \subseteq \Omega$ denotes a set of classes that surely apply to instance i, and $B_i \subseteq \Omega$ is a set of classes that surely do not apply to the same instance. For instance, assume that instances are songs and classes are emotions generated by these songs. Upon hearing a song, an expert may decide that this song certainly evokes happiness and certainly does not evoke sadness, but may be undecided regarding the other emotions (such as quietness, anger, surprise, etc.). In that case, the song cannot be assigned to a single label set, but one can associate to it the set of all label sets containing "happiness" and not containing "sadness". The goal of the learning system is to build a multi-label classifier $\mathcal{H} : \mathcal{X} \rightarrow 2^{\Omega}$ that associates a label set to each unseen instance.

To determine the multi-label classifier, the evidential multi-label kNN rule introduced in [10] can be used. Hereafter, we recall the principle of this method.

EML-*k*NN. Let $\mathbf{x}$ be an unseen instance, Y its unknown label set, and $\mathcal{N}_{\mathbf{x}}$ its k nearest neighbors in $\mathcal{S}$ based on a certain distance function d(., .), usually the Euclidean one. Each element $(\mathbf{x}_i, A_i, B_i)$ in $\mathcal{N}_{\mathbf{x}}$ constitutes a distinct item of evidence regarding the label set of $\mathbf{x}$.

The mass function m_i over $\mathcal{C}(\Omega)$ induced by the item of evidence $(\mathbf{x}_i, A_i, B_i)$ regarding the label set of $\mathbf{x}$ is defined as:

$$m_i(A_i, B_i) = \alpha \exp(-\gamma d_i) \tag{13}$$

$$m_i(\emptyset, \emptyset) = 1 - \alpha \exp(-\gamma d_i) \tag{14}$$

where $d_i = d(\mathbf{x}, \mathbf{x}_i)$, $0 < \alpha < 1$ and $\gamma > 0$. Parameter α is usually fixed at a value close to 1 such as $\alpha = 0.95$ [2], whereas γ should depend on the scaling of distances and can be fixed by cross-validation [10].

After considering each item of evidence in $\mathcal{N}_{\mathbf{x}}$, we obtain k mass functions m_i, $i = 1, \ldots, k$ that can be combined using the multi-label extension of the unnormalized Dempster's rule of combination (12) to form the resulting mass function m.

For decision making, different procedures can be used. The following simple and computationally efficient rule was implemented. Let $\widehat{Y}$ be the predicted label set for instance $\mathbf{x}$ to differentiate it from the ground truth label set Y of $\mathbf{x}$. To decide whether to include each class $\omega_q \in \Omega$ or not, we compute the degree of belief $bel(\{\omega_q\}, \emptyset)$ that the true label set Y contains ω_q, and the degree of belief $bel(\emptyset, \{\omega_q\})$ that it does not contain ω_q. We then define $\widehat{Y}$ as

$$\widehat{Y} = \{\omega_q \in \Omega \mid bel(\{\omega_q\}, \emptyset) \geq bel(\emptyset, \{\omega_q\})\}. \tag{15}$$

4 Experiments

4.1 Datasets

Three datasets were used in our experiments: the emotion, scene and yeast datasets[1]. Each one was split into a training set and a test set. Table 1 summarizes the characteristics of the datasets used in the experiments. The label cardinality of a dataset is the average number of labels of the instances, while the label density is the average number of labels of the instances divided by the total number of labels [8].

Table 1. Characteristics of datasets

Dataset	Number of instances	Feature vector dimension	Number of labels	Training instances	Test instances	Label cardinality	Label density	maximum size of a label set
emotion	593	72	6	391	202	1.869	0.311	3
scene	2407	294	6	1211	1196	1.074	0.179	3
yeast	2417	103	14	1500	917	4.237	0.303	11

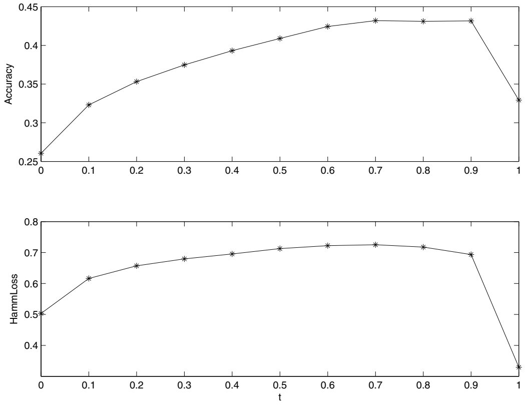

Fig. 1. *Accuracy* and *HammLoss* for EML-kNN on the emotion dataset for different values of the confidence threshold t

4.2 Imprecise Labeling Process

To simulate imprecise labeling by an expert, the following procedure was used. Let Y_i be the true label set for instance i, and let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iQ})$ be the vector of $\{-1, 1\}^Q$ such that $y_{iq} = 1$ if $\omega_q \in Y_i$ and $y_{iq} = -1$ otherwise. For each instance i and each class ω_q, we generated a probability of error $p_{iq} = p'_{iq}/2$, where p'_{iq} was taken from a beta distribution with parameters $a = b = 0.5$ (this consists on a bimodal distribution with modes at 0 and 1), and we changed y_{iq}

[1] http://mlkd.csd.auth.gr/multilabel.html

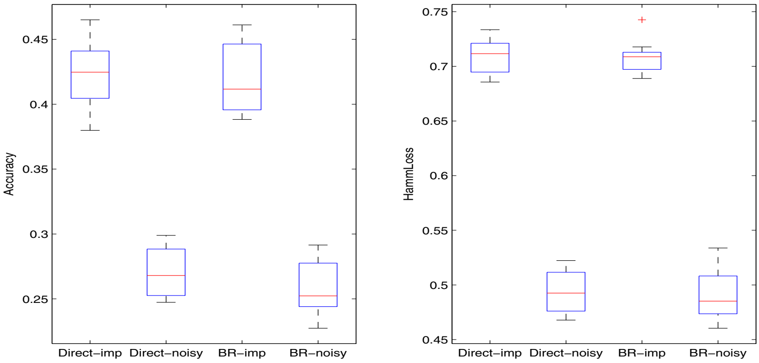

Fig. 2. Comparison between direct and BR versions of EML-kNN over 10 trials on imprecise and noisy labels generated from the emotion dataset

to $-y_{iq}$ with probability p_{iq}, resulting in a *noisy label vector* $\mathbf{y}'_i$. Each number p_{iq} represents the probability that the membership of instance i to class ω_q will be wrongly assessed by the experts. This number may be turned into a degree of confidence c_{iq} by the transformation:

$$c_{iq} = 1 - 2p_{iq}, \tag{16}$$

where $c_{iq} = 1$ means that the expert is totally sure about the membership ($y_{iq} = 1$) or not ($y_{iq} = -1$) of instance i to class ω_q, while $c_{iq} = 0$ means that he is totally undecided about this membership. In practice, these numbers can be provided by the experts.

By fixing a threshold t for confidence values (intuitively, $t > 0.5$), we then define the imprecise label vector as $\mathbf{y}''_i = (y''_{i1}, \ldots, y''_{iQ})$ with

$$y''_{iq} = \begin{cases} y'_{iq} & \text{if } c_{iq} \geq t, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

As shown in Section 2, such a vector of $\{-1, 0, 1\}^Q$ can be represented by $\varphi(A_i, B_i)$, the set of subsets of Ω, such that:

$$\begin{cases} A_i = \{\omega_q \in \Omega \mid y''_{iq} = 1\}, \\ B_i = \{\omega_q \in \Omega \mid y''_{iq} = -1\}. \end{cases} \tag{18}$$

The set A_i then contains the classes ω_q that can be definitely assigned to instance i with a high degree of confidence ($c_{iq} \geq t$), while B_i is the set of classes which are definitely *not* assigned to instance i. The remaining set $\Omega \setminus (A_i \cup B_i)$ contains those classes about which the expert is undecided ($c_{iq} < t$). We recall that $\varphi(A_i, B_i)$ contains all the label sets including A_i and non intersecting B_i.

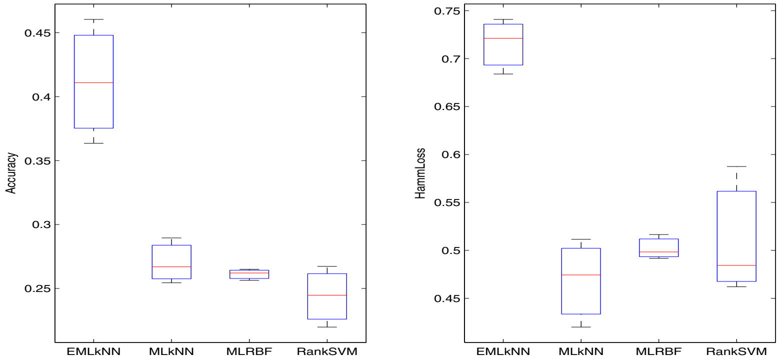

Fig. 3. *Accuracy* and *HammLoss* box plots over 10 trials for the emotion dataset with the following methods: EML-kNN with imprecise labels, ML-kNN, ML-RBF and Rank-SVM with noisy labels

4.3 Evaluation Metrics

Let $\mathcal{D} = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_N, Y_N)\}$ be a multi-label evaluation dataset containing N labeled examples. Let $\widehat{Y}_i = \mathcal{H}(\mathbf{x}_i)$ be the predicted label set for the pattern $\mathbf{x}_i$, while Y_i is the ground truth label set for $\mathbf{x}_i$.

A first metric called "Accuracy" gives an average degree of similarity between the predicted and the ground truth label sets of all test examples [8]:

$$Accuracy(\mathcal{H}, \mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|}. \tag{19}$$

A second metric is the "Hamming loss" that counts prediction errors (an incorrect label is predicted) and missing errors (a true label is not predicted). In order to be consistent with the above measure, we report 1-Hamming loss [6]:

$$HamLoss(\mathcal{H}, \mathcal{D}) = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{Q}|Y_i \bigtriangleup \widehat{Y}_i|, \tag{20}$$

where $\bigtriangleup$ is an operator to compute the symmetric difference of two sets.

The values of these evaluation criteria are in the interval $[0, 1]$. Larger values of these metrics correspond to higher classification quality.

4.4 Results and Discussions

Figure 1 shows the performance of EML-kNN over the two evaluation criteria *Accuracy* and *HammLoss* for different values of confidence threshold t after 10-fold cross validation on imprecise labels generated from the training emotion dataset. The best results were obtained for $t \in [0.5, 0.9]$. In the following, the

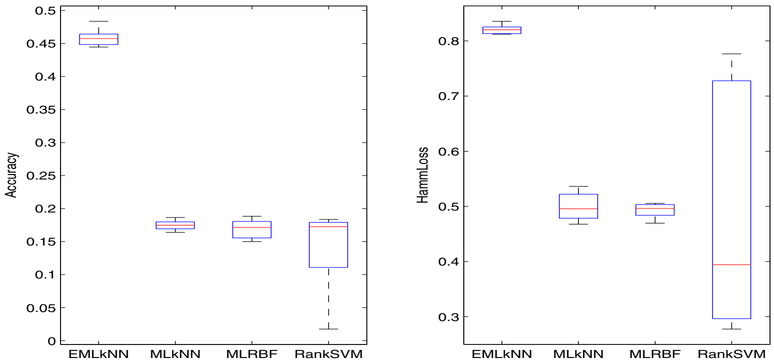

Fig. 4. *Accuracy* and *HammLoss* box plots over 10 trials for the scene dataset with the following methods: EML-kNN with imprecise labels, ML-kNN, ML-RBF and Rank-SVM with noisy labels

value of t was fixed to 0.6. Note that, for EML-kNN, γ was fixed to 0.5 and k to 10. The values of these two parameters can easily be determined by cross validation, but here, they are fixed manually to moderate values.

EML-kNN was originally developed in order to construct a multi-label learning system able to handle multi-labeled data directly. However, it can be also used when transforming the multi-label leaning problem into single-label one, which is referred to as indirect approach. To get an idea about the performance of each approach, the original EML-kNN (direct version) and the BR version (binary learning for each label) were applied to imprecise and noisy labeled data generated from the emotion dataset. Figure 2 shows the results over 10 trials. First, we notice the improved performances of our leaning system when applied to imprecise labels. This result demonstrates the usefulness of our evidence formalism. Secondly, we remark that the performances of the direct and BR versions of our method are very close, with a slight advantage for the direct approach. Note that, in terms of execution time, the direct approach is much faster. In the next experiments, the originial version (direct) of EML-kNN was used.

EML-kNN was compared to three existing multi-label classification methods that were shown to exhibit good performances: ML-kNN [13] that is the closest to our method as both are based on k-NN rule, ML-RBF [12] derived from radial basis function neural networks, and Rank-SVM [5] that is based on the traditional support vector machine. For ML-kNN, k was fixed to 10 as in [13]. As used in [12], the fraction parameter for ML-RBF was set to 0.01 and the scaling factor to 1. For Rank-SVM, the best parameterization reported in [5], i.e. polynomial kernels with degree 8, was used.

After performing the labeling process explained in Section 4.2, noisy labels and imprecise labels were generated for instances from each dataset. EML-kNN was applied to imprecise labels ($\mathbf{y}_i''$ corresponding to $\varphi(A_i, B_i)$ in the multi-label

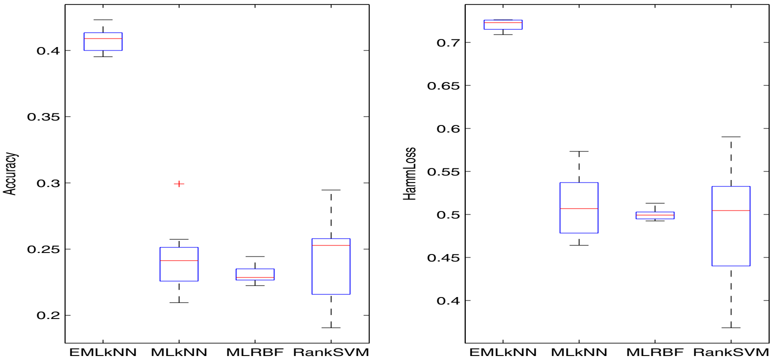

Fig. 5. *Accuracy* and *HammLoss* box plots over 10 trials for the yeast dataset with the following methods: EML-kNN with imprecise labels, ML-kNN, ML-RBF and Rank-SVM with noisy labels

evidence formalism), while the ML-kNN, ML-RBF and Rank-SVM algorithms were applied to noisy labels ($\mathbf{y}'_i$), as it is not clear how imprecise labels could be handled using these methods.

Figures 3, 4 and 5 show the box plots for the *Accuracy* and *HammLoss* measures obtained by the applied methods, over ten generations of imprecise and noisy labels, for the emotion, scene and yeast datasets respectively.

Based on the two evaluation criteria and over the three datasets, EML-kNN clearly dominates the remaining methods. These preliminary results demonstrate the ability of our approach to handle imprecise labels in multi-label classification tasks. In fact, when the available learning data have not a ground truth and have been labeled subjectively by a pool of experts, noisy labels will be inevitably assigned to some instances due to conflicts or lack of knowledge. If an expert gives a degree of confidence about each assigned label, by using EML-kNN method based on the evidence formalism explained in Section 2, we are able to reduce the risk of assigning wrongly some labels to an instance i when the degrees of confidence are not high. That explains the good performances of our method.

5 Conclusion

In this paper, we have used the evidence formalism for multi-label learning and the EML-kNN method introduced in [10] to propose a multi-label learning system able to handle complex learning tasks in which the data are imprecisely labeled. In fact, in many real-world problems, there are no ground truth for assigning unambiguously a label set to each instance, and several experts have to be consulted. Due to lack of confidence and conflicts between experts, uncertainties are introduced when labeling instances. To assess the performances of our approach when learning from data with imprecise labels, we have used an algorithm to randomly simulate such data. Experimental results demonstrate the ability of our approach to handle imprecise labels in multi-label classification tasks. EML-kNN dominates state-of-the-art methods in such situations.

References

1. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
2. Denœux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. on Systems, Man and Cybernetics 25(5), 804–813 (1995)
3. Denœux, T., Smets, P.: Classification using Belief Functions, the Relationship between the Case-based and Model-based Approaches. IEEE Trans. on Systems, Man and Cybernetics B 36(6), 1395–1406 (2006)
4. Denœux, T., Younes, Z., Abdallah, F.: Representing uncertainty on set-valued variables using belief functions. Artificial Intelligence 174(7-8), 479–499 (2010)
5. Elisseeff, A., Weston, J.: Kernel methods for multi-labelled classification and categorical regression problems. Advances in Neural Information Processing Systems 14, 681–687 (2002)
6. Fürnkranz, J., Hüllermeier, E., Loza Menca, E., Brinker, K.: Multilabel classification via calibrated label ranking. Machine Learning 73(2), 133–153 (2008)
7. Smets, P.: The combination of evidence in the Transferable Belief Model. IEEE Trans. on Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990)
8. Tsoumakas, G., Katakis, I.: Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining 3(3), 1–13 (2007)
9. Tsoumakas, G., Vlahavas, I.: Random k-Labelsets: An Ensemble Method for Multilabel Classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007)
10. Younes, Z., Abdallah, F., Denœux, T.: An Evidence-Theoretic k-Nearest Neighbor Rule for Multi-Label Classification. In: Godo, L., Pugliese, A. (eds.) SUM 2009. LNCS (LNAI), vol. 5785, pp. 297–308. Springer, Heidelberg (2009)
11. Younes, Z., Abdallah, F., Denœux, T.: Multi-label classification algorithm derived from k-nearest neighbor rule with label dependencies. In: Proc. of the 16th European Signal Processing Conference, Lausanne, Switzerland, August 25-29 (2008)
12. Zhang, M.-L.: ML-RBF: RBF neural networks for multi-label learning. Neural Processing Letters 29(2), 61–74 (2009)
13. Zhang, M.-L., Zhou, Z.-H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–3048 (2007)