An Evidence-Theoretic *k*-Nearest Neighbor Rule for Multi-Label Classification

Zoulficar Younes, Fahed abdallah, Thierry Denœux

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

Abstract. In multi-label learning, each instance in the training set is associated with a set of labels, and the task is to output a label set for each unseen instance. This paper describes a new method for multi-label classification based on the Dempster-Shafer theory of belief functions to classify an unseen instance on the basis of its k nearest neighbors. The proposed method generalizes an existing single-label evidence-theoretic learning method to the multi-label case. In multi-label case, the frame of discernment is not the set of all possible classes, but it is the powerset of this set. That requires an extension of evidence theory to manipulate multi-labelled data. Using evidence theory makes us able to handle ambiguity and imperfect knowledge regarding the label sets of training patterns. Experiments on benchmark datasets show the efficiency of the proposed approach as compared to other existing methods.

1 Introduction

Traditional *single-label classification* assigns an object to exactly one class, from a set of Q disjoint classes. In contrast, *Multi-label classification* is the task of assigning an instance to one or multiple classes simultaneously. In other words, the target classes are not exclusive: an object may belong to an unrestricted set of classes instead of exactly one. This task makes multi-label classifiers more difficult to train than traditional single-label classifiers. Recently, multi-label classification methods have been increasingly required by modern applications where it is quite natural that some instances belong to several classes at the same time. In text categorization, each document may belong to multiple topics, such as arts and humanities [8]. In natural scene classification, each image may belong to several image types at the same time, such as sea and sunset [1]. In classification of music into emotions, music may evoke more than one emotion at the same time, such as relaxing and sad [7].

Few algorithms have been proposed for multi-label learning. A first family of algorithms transforms the multi-label classification problem into a set of binary classification problems; each binary classifier is then trained to separate one class from the others [1]. A second family consists in extending common learning algorithms and making them able to manipulate multi-label data directly. In [14] and [15], a *Bayesian* approach based on multi-label extension of

the k-nearest neighbor (k-NN) rule is presented. In the literature, there also exist multi-label extensions of neural networks [2], support vector machine [6], and boosting learning algorithms [9].

In this paper, we present a new method for multi-label classification based on the Dempster-Shafer theory of belief functions to classify an unseen instance on the basis of its k nearest neighbors.

The Dempster-Shafer (D-S) theory [10] is a formal framework for representing and reasoning with uncertain and imprecise information. Different approaches for pattern classification in the framework of evidence theory have been presented in the literature [4] [5]. In [3], A k-NN classification rule based on D-S theory is presented. Each neighbor of an instance to be classified is considered as an item of evidence supporting certain hypotheses regarding the class membership of that instance. The degree of support is defined as a function of the distance between the two samples. The evidence of the k nearest neighbors is then pooled by means of Dempster's rule of combination.

The proposed method generalizes the k-NN classification rule based on the D-S theory to the multi-label case. This generalization requires an extension of the D-S theory in order to handle multi-labelled data. In mono-labelled data case, the uncertainty is represented by evidence on multiple hypotheses where each hypothesis is a label to be assigned or not to an unseen instance. In contrast, when the data is multi-labelled, each hypothesis represents a set of labels and the uncertainty is then expressed by evidence on sets of label sets. The proposed algorithm is called $EML-kNN$ for Evidential Multi-Label k-Nearest Neighbor.

The remainder of the paper is organized as follows. Section 2 recalls the basics of the D-S theory and the principle of the single-label evidence-theoretic k-NN rule [3]. Section 3 introduces the extension of the D-S theory to the multi-label case and describes the proposed algorithm for multi-label learning that consists in applying the D-S multi-label extended theory using the k-NN rule. Section 4 presents experiments on two real datasets and shows the effectiveness of the proposed algorithm as compared to a recent high-performance method for multi-label learning based on k-NN rule, referred to as $ML-kNN$ [15]. Finally Section 5 summarizes this work and makes concluding remarks.

2 Single-Label Classification

2.1 Basics of Dempster-Shafer Theory

In D-S theory, a *frame of discernment* Ω is defined as the set of all hypotheses in a certain domain, e.g., in classification Ω is the set of all possible classes. A *basic belief assignment* (BBA) is a function m that defines a mapping from the power set of Ω to the interval $[0,1]$ verifying:

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

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

Given a certain piece of evidence, the value of the BBA for a given set A expresses a measure of belief that one is willing to commit exactly to A. The quantity $m(A)$ pertains only to the set A and makes no additional claims about any subsets of A. If $m(A) > 0$, then the subset A is called a *focal element* of m.

The BBA m and its associated focal elements define a *body of evidence*, from which a belief function Bel and a plausibility function Pl mapped from 2^{Ω} to $[0, 1]$ can be deduced. For a set A, $Bel(A)$, called *belief* in A or *credibility* of A, represents a measure of the total belief committed to the set $A \subseteq \Omega$. $Bel(A)$ is defined as the sum of all the BBAs of the non-empty subsets of A.

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

$Pl(A)$, called *plausibility* of A, represents the amount of belief that could potentially be placed in A, if further information became available [3]. $Pl(A)$ is defined as the sum of all the BBAs of the sets that intersect A.

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

From the definitions of belief and plausibility functions, it follows that:

$$Pl(A) = Bel(\Omega) \ - \ Bel(\bar{A}) \tag{5}$$

where $\bar{A}$ is the complement of A.

Given the belief function Bel, it is possible to derive the corresponding BBA as follows:

$$m(\emptyset) = 1 - Bel(\Omega), \tag{6}$$

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} \ Bel(B), \ A \neq \emptyset \tag{7}$$

where $|A \backslash B|$ is the cardinality of the complement of B in A.

As a consequence of (5), (6) and (7), given any one of the three functions m, Bel and Pl it is possible to recover the other two.

The unnormalized *Dempster's rule of combination* [10] [11] is an operation for pooling evidence from a variety of sources. This rule aggregates two independent bodies of evidence defined within the same frame of discernment into one body of evidence. Let m_1 and m_2 be two BBAs. Let m_{12} be the new BBA obtained by combining m_1 and m_2 using the unnormalized Dempster's rule of combination. m_{12} is the *orthogonal sum* of m_1 and m_2 denoted as $m_{12} = m_1 \textcircled{\cap} m_2$. The aggregation is calculated in the following manner:

$$m_{12}(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \ A \subseteq \Omega. \tag{8}$$

This rule is commutative and associative, and admits the *vacous* BBA ($m(\Omega) = 1$) as neutral element.

2.2 Evidence-Theoretic *k*-NN Rule

Let $\mathcal{X} = R^P$ denote the domain of instances and let $\mathcal{Y} = \{1, 2, \ldots, Q\}$ be the finite set of classes, also called labels or categories. The available information is assumed to consist in a training set $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M)\}$ of M single-labelled samples, where $\mathbf{x}_i \in \mathcal{X}$ and the corresponding class label y_i takes value in $\mathcal{Y}$, for each $i = 1, \ldots, M$.

Let $\mathbf{x}$ be a new instance to be classified on the basis of its nearest neighbors in $\mathcal{T}$. Let $\mathcal{N}_\mathbf{x} = \{(\mathbf{x}_i, y_i) | i = 1 \ldots, k\}$ be the set of the k-nearest neighbors of $\mathbf{x}$ in $\mathcal{T}$ based on a certain distance function $d(.,.)$, e.g, the Euclidean distance. Each pair $(\mathbf{x}_i, y_i)$ in $\mathcal{N}_\mathbf{x}$ constitutes a distinct item of evidence regarding the class membership of $\mathbf{x}$. If $\mathbf{x}$ is *close* to $\mathbf{x}_i$ according to the distance function d, then one will be inclined to believe that both instances belong to the same class, while when $d(\mathbf{x}, \mathbf{x}_i)$ increases, this belief decreases and that yields to a situation of almost complete ignorance concerning the class of $\mathbf{x}$. Consequently, each pair $(\mathbf{x}_i, y_i)$ in $\mathcal{N}_\mathbf{x}$ induces a basic belief assignment m_i over $\mathcal{Y}$ defined by:

$$m_i(\{y_i\}) = \alpha\phi(d_i) \tag{9}$$

$$m_i(\mathcal{Y}) = 1 - \alpha\phi(d_i) \tag{10}$$

$$m_i(A) = 0, \ \forall A \in 2^\mathcal{Y} \backslash \{\mathcal{Y}, \{y_i\}\} \tag{11}$$

where $d_i = d(\mathbf{x}, \mathbf{x}_i)$, α is a parameter such that $0 < \alpha < 1$ and ϕ is a decreasing function verifying $\phi(0) = 1$ and $\lim_{d \to \infty} \phi(d) = 0$. In [3], the author suggests to choose the function ϕ as:

$$\phi(d) = \exp(-\gamma d^\beta) \tag{12}$$

where $\gamma > 0$ and $\beta \in \{1, 2, \ldots\}$. As explained in [3], parameter β has been found to have very little influence on the performance of the method, and can be arbitrarily fixed to a small value (1 or 2). The most influential parameter on the performance of the classifier is γ. In [3], a distinct parameter γ_q was associated for each class $q \in \mathcal{Y}$. When considering the item of evidence $(\mathbf{x}_i, y_i)$ for the class membership of $\mathbf{x}$, if $y_i = q$, using (12), $\phi(d_i)$ in (9) and (10) was replaced by $\gamma_q d_i^\beta$. The values of α and $\gamma_q, \ q = 1, \ldots, Q$ were fixed via heuristics [3].

As a result of considering each training instance in $\mathcal{N}_\mathbf{x}$ as an item of evidence, we obtain k BBAs that can be pooled by means of the unnormalized Dempster's rule of combination yielding to the aggregated BBA m synthesizing one's final belief regarding the class membership of $\mathbf{x}$:

$$m = m_1 \textcircled{\cap} \ldots \textcircled{\cap} m_k. \tag{13}$$

For making decisions, functions Bel and Pl can be derived from m using (3) and (4) respectively, and the test instance $\mathbf{x}$ is assigned to the class q that corresponds to the maximum credibility or the maximum plausibility.

3 Multi-Label Classification

3.1 Multi-Label Extension of Dempster-Shafer Theory

In Sect. 2.1, we have recalled the basics of D-S theory used to handle uncertainty in problems where *only one single hypothesis* is true. Moreover, there exist problems where *more than one hypothesis* is true at the same time, e.g., the multi-label classification task. To handle such problems, we need to extend the classical D-S framework. The frame of discernment of the multi-label extended D-S theory is not the set Ω of all possible single hypotheses but its power set $\Theta = 2^{\Omega}$. A basic belief assignment is now defined as a mapping from the power set of Θ to the interval $[0, 1]$. Instead of considering the whole power set of Θ, we will focus on the subset $\mathcal{C}(\Omega)$ of 2^{Θ} defined as:

$$\mathcal{C}(\Omega) = \{\varphi(A, B) |\ A \cap B = \emptyset\} \cup \{\emptyset_{\Theta}\} \tag{14}$$

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:

$$\varphi(A, B) = \{C \subseteq \Omega |\ C \supseteq A \ and \ C \cap B = \emptyset\}. \tag{15}$$

The size of the subset $\mathcal{C}(\Omega)$ of 2^{Θ} is equal to $3^{|\Omega|} + 1$, it is thus much smaller than the size of 2^{Θ} ($|2^{\Theta}| = 2^{2^{|\Omega|}}$), while being rich enough to express evidence in many realistic situations. That reduces the complexity of such problems.

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 $\varphi(A, B)$, one can deduce that:

$$\varphi(\emptyset, \emptyset) = \Theta, \tag{16}$$

$$\forall A \subseteq \Omega,\ \varphi(A, \bar{A}) = \{A\}, \tag{17}$$

$$\forall A \subseteq \Omega,\ A \neq \emptyset,\ \varphi(A, A) = \emptyset_{\Theta}. \tag{18}$$

By convention, we will note $\emptyset_{\Theta}$ by $\varphi(\Omega, \Omega)$ in the rest of the paper.

Example 1. Let $\Omega = \{a, b\}$ be a frame of discernment. The corresponding subset $\mathcal{C}(\Omega)$ of 2^{Θ}, where Θ is the power set of Ω, is:

$$\begin{aligned}\mathcal{C}(\Omega) = \{&\varphi(\emptyset, \emptyset), \varphi(\emptyset, \{a\}), \varphi(\emptyset, \{b\}), \varphi(\emptyset, \Omega), \varphi(\{a\}, \emptyset),\\ &\varphi(\{b\}, \emptyset), \varphi(\Omega, \emptyset), \varphi(\{a\}, \{b\}), \varphi(\{b\}, \{a\}), \varphi(\Omega, \Omega)\}.\end{aligned}$$

For instance, $\varphi(\{a\}, \emptyset) = \{\{a\}, \Omega\}$ and $\varphi(\{a\}, \{b\}) = \{\{a\}\}$.

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

$$\varphi(A, B) \cap \varphi(A', B') = \begin{cases} \varphi(A \cup A', B \cup B') & if\ A \cap B' = \emptyset \ and \ A' \cap B = \emptyset \\ \varphi(\Omega, \Omega) & otherwise, \end{cases} \tag{19}$$

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

$$\varphi(A,B) \subseteq \varphi(A',B') \iff A \supseteq A' \text{ and } B \supseteq B'. \tag{20}$$

The description of a BBA m on $\mathcal{C}(\Omega)$ can be represented with the following two equations:

$$m : \mathcal{C}(\Omega) \longrightarrow [0,1] \tag{21}$$

$$\sum_{\varphi(A,B)\in\mathcal{C}(\Omega)} m(\varphi(A,B)) = 1. \tag{22}$$

In the following, the 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 defined as:

$$Bel(A,B) = \sum_{\varphi(\Omega,\Omega)\neq\varphi(A',B')\subseteq\varphi(A,B)} m(A',B'), \tag{23}$$

and

$$Pl(A,B) = \sum_{\varphi(A',B')\cap\varphi(A,B)\neq\varphi(\Omega,\Omega)} m(A',B'). \tag{24}$$

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

$$m_{12}(A,B) = \sum_{\varphi(A',B')\cap\varphi(A'',B'')=\varphi(A,B)} m_1(A',B')m_2(A'',B''). \tag{25}$$

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

3.2 Evidential Multi-Label *k*-NN

Problem. As in Sect. 2.2, let $\mathcal{X} = R^P$ denote the domain of instances and let $\mathcal{Y} = \{1, 2, \ldots, Q\}$ be the finite set of labels. The multi-label classification problem can be formulated as follows. Given a set $\mathcal{S} = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_M, Y_M)\}$ of M training examples drawn from $\mathcal{X} \times 2^{\mathcal{Y}}$, and identically distributed, where $\mathbf{x}_i \in \mathcal{X}$ and $Y_i \subseteq \mathcal{Y}$, the goal of the learning system is to output a multi-label classifier $\mathcal{H} : \mathcal{X} \to 2^{\mathcal{Y}}$ that optimizes some pre-defined criteria.

The method. Let $\mathbf{x}$ be an unseen instance that we search to estimate its set of labels Y on the basis of its k nearest neighbors in $\mathcal{S}$ represented by $\mathcal{N}_{\mathbf{x}}$ using the multi-label extension of the D-S theory introduced in Sect. 3.1. The frame of discernment of the multi-label classification problem is the powerset of $\mathcal{Y}$.

Each pair $(\mathbf{x}_i, Y_i)$ in $\mathcal{N}_\mathbf{x}$ constitutes a distinct item of evidence regarding the label set of $\mathbf{x}$. Regarding the label set Y_i, we can conclude either that Y must include all the labels in Y_i, or that Y must contain at least one of the labels that belong to Y_i, or that Y does not contain any label not belonging to Y_i. Let $\varphi(A_i, B_i)$ be the set of label sets that corresponds to the item of evidence $(\mathbf{x}_i, Y_i)$, where A_i, $B_i \subseteq \mathcal{Y}$. We recall that the set $\varphi(A_i, B_i)$ contains all the label sets that include A_i and having no intersection with B_i. There exist different ways to express our beliefs about the labels to be assigned to the instance $\mathbf{x}$ based on the item of evidence $(\mathbf{x}_i, Y_i)$. This leads to different versions of our proposed method $EML-kNN$:

- Version 1 (V1): the mass is attributed to the set Y_i, thus $\varphi(A_i, B_i) = \varphi(Y_i, \bar{Y_i})$.
- Version 2 (V2): the mass is attributed to the set Y_i and all its supersets, thus $\varphi(A_i, B_i) = \varphi(Y_i, \emptyset)$.
- Version 3 (V3): the mass is attributed to the set Y_i and all its subsets, thus $\varphi(A_i, B_i) = \varphi(\emptyset, \bar{Y_i})$.

The BBA m_i over $\mathcal{C}(\mathcal{Y})$ induced by the item of evidence $(\mathbf{x}_i, Y_i)$ regarding the label set of $\mathbf{x}$ can then be defined as:

$$m_i(A_i, B_i) = \alpha\phi(d_i) \tag{26}$$

$$m_i(\emptyset, \emptyset) = 1 - \alpha\phi(d_i) \tag{27}$$

where $d_i = d(\mathbf{x}, \mathbf{x}_i)$, ϕ is the decreasing function introduced in Sect. 2.2 (see (12)).

After considering each item of evidence in $\mathcal{N}_\mathbf{x}$, we obtain the BBAs m_i, $i = 1, \ldots, k$ that can be combined 2 by 2 using the multi-label extension of the unnormalized Dempster's rule of combination presented in Sect. 3.1 (see (25)) to form the resulting BBA m.

Let $\hat{Y}$ denote the estimated label set of the instance $\mathbf{x}$ to differentiate it from the ground truth label set Y of $\mathbf{x}$. One of the methods to determine $\hat{Y}$ that we have adopted in this paper is to assign $\mathbf{x}$ to the set $C \subseteq \mathcal{Y}$ that corresponds to the maximum plausibility. Thus, the estimated label set of $\mathbf{x}$ is:

$$\hat{Y} = \max_{C \subseteq \mathcal{Y}} Pl(C, \bar{C}). \tag{28}$$

The plausibility function Pl derived from the aggregated BBA m is determined using (24).

4 Experiments

4.1 Datasets

Two datasets are used for experiments: the emotion and the scene datasets.

Emotion Dataset. This dataset contains 593 songs, each represented by a 72-dimensional feature vector (8 rhythmic features and 64 timbre features) [7]. The emotional labels are: *amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely and angry-fearful.*

Scene Dataset. This dataset contains 2000 natural scene images. Each image is associated with some of the six different semantic scenes: *sea, sunset, trees, desert and mountains*. For each image, spatial color moments are used as features. Images are divided into 49 blocks using a 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 [1]. Each image is then transformed into a $49 \times 3 \times 2 = 294$-dimensional feature vector.

Each dataset 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 [12].

4.2 Evaluation metrics

Let $\mathcal{D} = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_N, Y_N)\}$ be a multi-label evaluation dataset containing N labelled 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:

$$Accuracy(\mathcal{H}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|}. \tag{29}$$

Two other metrics called *Precision* and *Recall* are also used in the literature to evaluate a multi-label learning system. The former computes the proportion of correct positive predictions while the latter calculates the proportion of true labels that have been predicted as positives:

$$Precision(\mathcal{H}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|\widehat{Y}_i|}, \tag{30}$$

$$Recall(\mathcal{H}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i|}. \tag{31}$$

These metrics have been cited in [12].

Another evaluation criterion is the $F1$ measure that is defined as the harmonic mean of the *Precision* and *Recall* metrics [13]:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}. \tag{32}$$

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

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	1.868	0.311	3
scene	2407	294	6	1211	1196	1.074	0.179	3

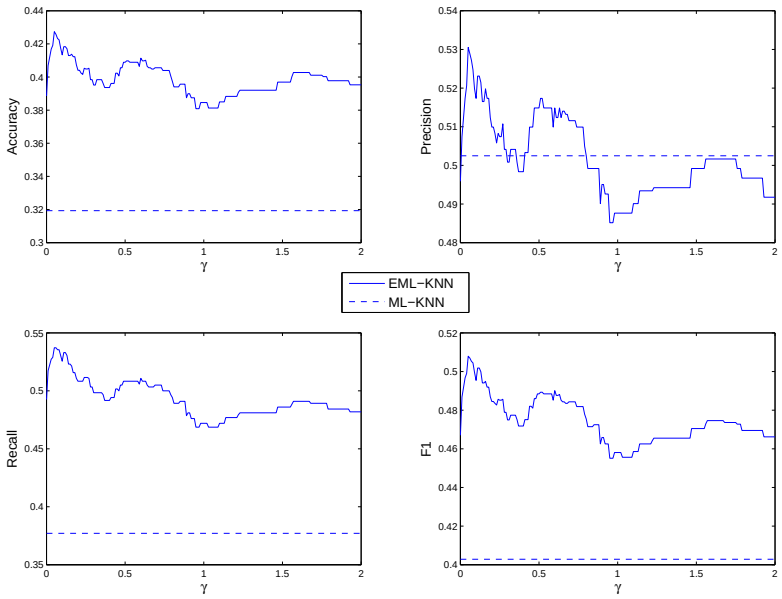

Fig. 1. *Accuracy*, *Precision*, *Recall* and $F1$ measures for $EML-kNN$ (V1) and $ML-kNN$ algorithms as a function of γ on the emotion dataset, for $k = 10$.

4.3 Results and discussions

The proposed algorithm was compared to a *Bayesian* method for multi-label classification based on the k-NN rule named $ML-kNN$ [15].

The model parameters for $EML-kNN$ are : The number of neighbors k, and the parameters for the induced BBAs, α, β and γ. $ML-KNN$ has only one parameter that needs to be optimized, which is k. As in [3], α was fixed to 0.95 and β to 1. For all experiments, $EML-kNN$ and $ML-kNN$ were trained on the training data and evaluated on the test data of each of the two datasets.

To take an idea about the influence of the parameter γ on the performance of the proposed algorithm, we evaluated version 1 of our method on the emotion dataset where k was fixed to 10 and γ was varied from 0 to 2 with 0.01 steps. Figure 1 shows the results. For the different values of γ, our algorithm performs better than $ML-KNN$ for all criteria except *Precision*. Based on (26) and (27),

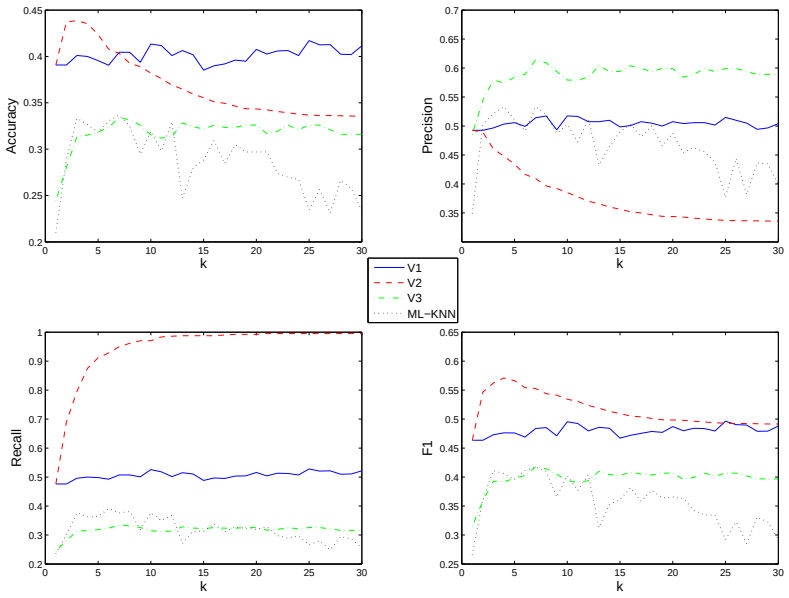

Fig. 2. *Accuracy*, *Precision*, *Recall* and $F1$ measures for the three versions of $EML-kNN$ and $ML-kNN$ algorithms as a function of k on the emotion dataset, for $\gamma = 0.1$.

we can notice that for small values of γ, we favor the allocation of mass to the set of label sets $\varphi(A_i, B_i)$ that corresponds to the item of evidence $(\mathbf{x}_i, Y_i)$. In contrast, for larger values of γ, a larger fraction of the mass is assigned to the ignorance set $\varphi(\emptyset, \emptyset)$.

In a second step, γ was fixed to 0.1 and k was varied from 1 to 30. Figures 2 and 3 show the performance of the three versions of $EML - kNN$ (denoted by V1, V2 and V3) and the $ML - kNN$ algorithms on the emotion and scene datasets, respectively. Algorithm V1 yields the better performance on the emotion dataset based on all criteria. On the scene dataset, algorithms V1 and V3 yield similar results and both outperform $ML - KNN$ for the different values of k and for all evaluation measures. Algorithm V2 yields poor results for all values of k and for all metrics except *Recall*. We recall that for version 2 of $EML - kNN$, given an item of evidence $(\mathbf{x}_i, Y_i)$, the belief is allocated to the set Y_i and all its supersets. For higher values of k, given an unseen instance $\mathbf{x}$, the most plausible label set after pooling the BBAs induced by the k nearest neighbors will be the set of all labels, i.e., the predicted label set will be $\hat{Y} = \mathcal{Y}$. That explains the fact that the *Recall* measure tends to 1 while the *Precision* measure decreases when the value of k increases.

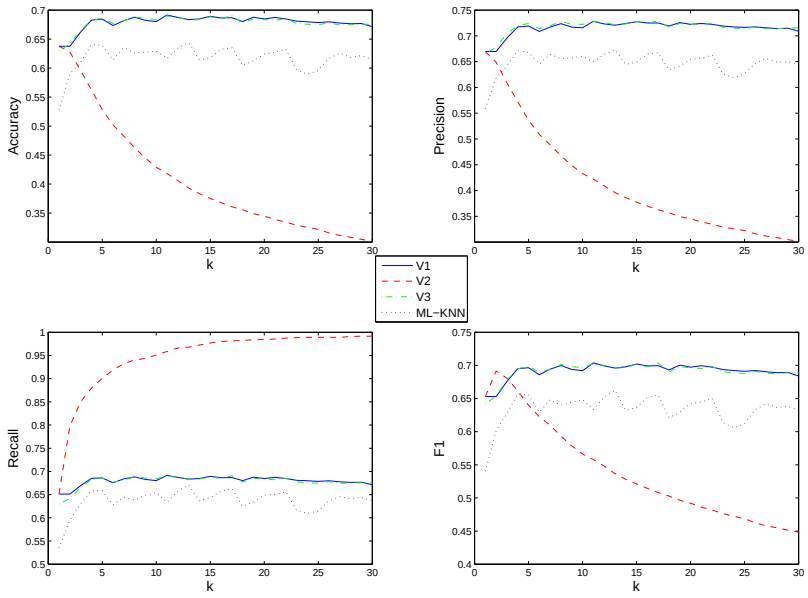

Fig. 3. *Accuracy*, *Precision*, *Recall* and $F1$ measures for the three versions of $EML-kNN$ and $ML-kNN$ algorithms as a function of k on the scene dataset, for $\gamma = 0.1$.

5 Conclusion

In this paper, an evidence-theoretic k-NN rule for multi-label classification has been presented. Using the evidence theory makes us able to handle the ambiguity and making decisions with multiple possible label sets for an unseen instance without having to resort to assumptions about these sets. The proposed method generalizes the single-label evidence-theoretic k-NN rule to the multi-label case. An unseen instance is classified on the basis of its k nearest neighbors. Each neighbor of an instance to be classified is considered as an item of evidence supporting some hypotheses regarding the set of labels of this instance. A first approach consists in supporting the hypothesis that the label set of the unseen instance is identical to the label set of the ith neighbor considered as an item of evidence. A second one consists in supporting the label set of the ith neighbor and all its supersets. The hypotheses supported by a third approach are the label set of the ith neighbor and all its subsets. The experiments on two real datasets demonstrate the effectiveness of the proposed method as compared to state-of-the-art method also based on the k-NN principle. Especially, the first and the third approaches gave better performance than the second one.

Another contribution of this paper is the presentation of an extension of the D-S theory to manipulate multi-labelled data. In the multi-label case, the frame of discernment defined as the set of all hypotheses in a certain domain is not the set of all possible classes but the powerset of this set. Thus, each hypothesis represents a set of labels and the uncertainty is then expressed by evidence on sets of label sets.

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. Crammer, K., Singer, Y.: A Family of Additive Online Algorithms for Category Ranking. Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 151–158 (2002)
3. 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)
4. 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)
5. Denœux, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets and Systems, 122(3), 47–92 (2001)
6. Elisseeff, A., Weston, J.: Kernel methods for multi-labelled classification and categorical regression problems. Advances in Neural Information Processing Systems, 14, 681–687 (2002)
7. Konstantinos, T., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music into emotions. Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA, USA, (2008)
8. McCallum, A.: Multi-Label Text Classification with a Mixture Model Trained by EM. In Working Notes of the AAAI'99 Workshop on Text Learning (1999)
9. Schapire, R.E., Singer, Y.: BoosTexter: A Boosting-based System for Text Categorization. Machine Learning, 39(2-3), 135–168 (2000)
10. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton, N.J. (1976)
11. Smets, P.: The combination of evidence in the Transferable Belief Model. IEEE Trans. on Pattern Analysis and Machine Intelligence, 12(5), 447–458 (1990)
12. Tsoumakas, G., Katakis, I.: Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, 3(3), 1–13 (2007)
13. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1, 78–88 (1999)
14. 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)
15. Zhang, M.-L., Zhou, Z.-H.: ML-KNN:A lazy learning approach to multi-label learning. Pattern Recognition, 40(7), 2038–3048 (2007)