A novel k-NN approach for data with uncertain attribute values

Asma Trabelsi[1,2], Zied Elouedi[1], and Eric Lefevre[2]

[1] Université de Tunis, Institut Supérieur de Gestion de Tunis, LARODEC , Tunisia
trabelsyasma@gmail.com,zied.elouedi@gmx.fr
[2] Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), Béthune, F-62400, France
eric.lefevre@univ-artois.fr

Abstract. Data uncertainty arises in several real world domains, including machine learning and pattern recognition applications. In classification problems, we could very well wind up with uncertain attribute values that are caused by sensor failures, measurements approximations or even subjective expert assessments, etc. Despite their seriousness, these kinds of data are not well covered till now. In this paper, we propose to develop a machine learning model for handling such kinds of imperfection. More precisely, we suggest to develop a new version of the well known k-nearest neighbors classifier to handle the uncertainty that occurs in the attribute values within the belief function framework.

Keywords: Evidential k-nearest neighbors, uncertainty, belief function theory, classification.

1 Introduction

The k Nearest Neighbor (k-NN) classifier, firstly proposed by Fix and Hodges [4], is regarded as one of the well commonly used classification techniques in the fields of machine learning and pattern recognition. The original k-NN version consists of assigning a query pattern to the majority class of its k nearest neighbors. The major shortcoming of this technique arises from learning a k-NN classifier with skewed class distributions, meaning that training instances with the most prevalent class may dominate the prediction of new query patterns due a large value of k. From this, numerous researchers have proven that the uncertainty about the class label of a given test pattern can be modeled through various uncertainty theories such as the possibilistic theory [8], the fuzzy theory [15], the belief function theory [9], etc. This latter, also referred to as evidence theory, has shown a great success in several pattern recognition problems, notably for representing and managing the uncertainty relative to the label class of new patterns to be classified. In [2], Denoeux has proposed an evidence theoretic k-NN (Ek-NN) method relied on the belief function theory where each neighbor of a pattern to be classified is regarded as a piece of evidence supporting some hypothesis concerning its class membership. The basic belief assignments obtained

by all the k nearest neighbors are then merged through the Dempster rule to identify the class label relative to each test pattern. An extended version of the Ek-NN, denoted by EEk-NN, has been introduced in [5], where the label class of each training instance will be represented by an evidential label to handle the uncertainty that occurs in the training data. Its is worth noting that, in several real world data, the attribute values may also contain some noise and outliers that can make erroneous classification results. Thus, evidential databases where attributes' values are represented using the evidence theory have been introduced over the past few years. Despite their accuracy, neither the Ek-NN nor the EEk-NN are able to handle such kinds of data. Inspired from both Ek-NN and EEk-NN, in this paper, we suggest to develop a new k-NN version for dealing with data described by uncertain attribute values, particularly where the uncertainty is represented within the belief function framework. The reminder of this paper is organized as follows: Section 2 is devoted to highlighting the basic concepts of the belief function theory as explained by the Transferable Belief Model framework, one interpretation of the belief function theory. In Section 3, we present our novel k-NN version for handling evidential databases. Our experimentation on several synthetic databases are described in Section 4. Finally, in Section 5, we draw our conclusion and our main future work directions.

2 Belief function theory: background

The belief function theory, originally pointed out by Dempster [1] and Shafer [9], has shown a great success for modeling uncertain knowledge. In what follows, we recall the main concepts of this theory.

2.1 Frame of discernment

The frame of discernment, denoted by Θ, is the set of all possible answers for a given problem which should be mutually exhaustive and exclusive:

$$\Theta = \{H_1, \ldots, H_N\} \tag{1}$$

From the frame of discernment Θ, one can deduce the set 2^Θ containing all subsets of Θ:

$$2^\Theta = \{\emptyset, H_1, H_2, \ldots, H_N, H_1 \cup H_2, \ldots, \Theta\} \tag{2}$$

2.2 Basic Belief Assignment

A basic belief assignment (bba), denoted by m, is a mapping function m: $2^\Theta \rightarrow [0, 1]$, such that:

$$\sum_{A \subseteq \Theta} m(A) = 1 \tag{3}$$

Each subset A of 2^Θ fulfilling $m(A) > 0$ is called a focal element.

2.3 Combination operators

Several combination rules have been introduced to merge reliable independent information sources issued from independent information sources. The conjunctive operator, proposed within the Transferable Belief Model (TBM) [11], is a well known one. For two information sources S_1 and S_2 having respectively the bbas m_1 and m_2, the conjunctive rule, denoted by $\textcircled{\cap}$, will be written in the following form:

$$m_1 \textcircled{\cap} m_2(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Theta. \tag{4}$$

The belief committed to the empty set is called conflictual mass. A normalized version of the conjunctive operator, proposed by Dempster [1], manages the conflict by redistributing the conflictual mass over all focal elements. The Dempster rule is defined as follows:

$$m_1 \oplus m_2(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Theta \tag{5}$$

where K ($K \neq 1$), representing the conflictual mass between the two bbas m_1 and m_2, is set as:

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{6}$$

2.4 Decision making

To make decisions within the belief function framework, Smets, in [10], has proposed the so-called pignistic probability denoted by $BetP$ which transforms the beliefs held into probability measures as follows:

$$BetP(A) = \sum_{B \cap A = \emptyset} \frac{|A \cap B|}{|B|} m(B), \quad \forall A \in \Theta \tag{7}$$

2.5 Dissimilarity between bbas

In the literature, there have been several measures allowing the computation of the degree of dissimilarity between two bodies of evidence [6, 12]. One of the commonly used measures is the Jousselme distance which is set as follows for two given bbas m_1 and m_2:

$$dist(m_1, m_2) = \sqrt{\frac{1}{2}(m_1 - m_2)^T D (m_1 - m_2)} \tag{8}$$

where D is the Jaccard similarity measure defined by:

$$D(X,Y) = \begin{cases} 1 & \text{if } X{=}Y{=}\emptyset \\ \dfrac{|X \cap Y|}{|X \cup Y|} & \forall\ X,Y \in 2^{\Theta} \end{cases} \tag{9}$$

3 Nearest Neighbor classifiers for uncertain data

In what follows, we address classification problems with uncertain data. More precisely, we get inspired from the Evidential k-NN classifier and its extended version [2, 5] to handle the uncertainty that occurs in the attribute values and is represented within the belief function framework. Let $X = \{x^i = (x^i_1,\ldots,x^i_n)|\text{i} = 1,\ldots,N\}$ be a set of N n-dimensional training samples, and let $\Theta = \{H_1,\ldots,H_M\}$ be a set of M classes. Each sample x^i is described by n uncertain attribute values represented within the belief function framework and a class label $L^i \in \{1,\ldots,M\}$ expressing with certainty its membership to one class in Θ. Assume that L is the set of labels, we denote by $T = \{(x^1, L^1), ...(x^N, L^N)\}$ the training set that will be used to classify new objects. Suppose that y is a new pattern to be classified based on the information contained in the training set T. The idea consists of computing the distance between the test pattern y and each pair (x^i, L^i) in T using a distance metric $d_{y,i}$ which is calculated as the sum of the absolute differences between the attribute values. More specifically, we have resorted to the Jousselme distance metric to cope with the uncertainty that arises in the attribute values. Thus, $d_{y,i}$ is set as follows:

$$d_{y,i} = \sum_{j=1}^{n} \sqrt{\frac{1}{2}(x^i_j - y_j)^T D_j (x^i_j - y_j)} \tag{10}$$

where D_j is the Jaccard similarity measure defined by:

$$D(X,Y) = \begin{cases} 1 & \text{if } X{=}Y{=}\emptyset \\ \dfrac{|X \cap Y|}{|X \cup Y|} & \forall\ X,Y \in 2^{\Theta_j} \end{cases} \tag{11}$$

A small value of $d_{y,i}$ reflects the situation that both instances y and x^i have the same label class L^i. On the contrary, a large value of $d_{y,i}$ may reflect the situation of almost complete ignorance concerning the label class of y. The information concerning the label class of the pattern query y can be modeled through the belief function theory. Thus, for the test sample y, each training instance x^i provides an item of evidence $m^{(i)}(.|x^i)$ over Θ as follows:

$$\begin{aligned} &m^{(i)}(H_q|x^i) = \alpha \Phi_q(d_{y,i}) \\ &m^{(i)}(\Theta|x^i) = 1 - \alpha \Phi_q(d_{y,i}) \\ &m^{(i)}(A|x^i) = 0, \forall A \in 2^{\Theta} \backslash \{\Theta, H_q\} \end{aligned} \tag{12}$$

where H_q is the class label of the instance x^i and α is a parameter such that $0 < \alpha < 1$. Author in [2] has proven that setting α to 0.95 can yield good results. The decreasing function Φ_q, verifying $\Phi_q(0){=}1$ and $lim_{d\to\infty}\Phi_q(d) = 0$, should be set as:

$$\Phi_q(d) = exp(-\gamma_q d^2), \tag{13}$$

where γ_q be a positive parameter relative to the class H_q that can be optimized using either an exact method relying on a gradient search procedure for medium

or small training sets or a linearization method for handling large training sets [16]. For both exact and approximated methods, the best values of γ are determined by minimising the mean squared classification error over the whole training set T of size N. The final bba m^y regarding the class of the query pattern y can be obtained by merging the N bbas issued from the different training instances. We ultimately resorted to the Dempster rule, one of the well-known rules used for ensuring fusion. It is set as follows:

$$m^y = m^{(1)}(.|x^1) \oplus m^{(2)}(.|x^2) \oplus \ldots \oplus m^{(N)}(.|x^N) \tag{14}$$

As some training instances may be too far from y, only the k nearest neighbors of the test sample y should be considered to determinate its class membership. The final bba will be set as follows:

$$m^y = m^{(1)}(.|x^1) \oplus m^{(2)}(.|x^2) \oplus \ldots \oplus m^{(k)}(.|x^k) \tag{15}$$

To make a decision about the label class of the query pattern y, the pignistic probability $BetP$ should be computed based on the combined bba m^y as shown in Equation 7. The test pattern is then assigned to the class with the maximum pignistic probability:

$$L^y = argmax_{H_q} BetP(H_q) \tag{16}$$

where $BetP(H_q)$ corresponds to the pignistic probablity of the hypothesis H_q associated to the bba m^y.

4 Experimentations

In this Section, we present our carried out experimentations to assess the performance of our proposed k-NN classifiers.

4.1 Experimentation settings

For checking the performance of our proposed k-NN classifier, we have performed experimentations on several synthetic databases obtained by adding uncertainty to some real world databases acquired from the well known UCI machine learning repository. As we only deal with categorical attributes, in this paper, we have resorted to only symbolic databases. A brief description of these databases is presented in Table 1. We have managed various uncertainty levels according to certain degrees of uncertainty denoted by P:

- No uncertainty: P=0
- Low Uncertainty: $0 < P < 0.4$
- Middle Uncertainty: $0.4 \leq P < 0.7$
- High Uncertainty: $0.7 \leq P \leq 1$

Given a database described by N objects x^i ($i \in \{1, \ldots, N\}$), n attributes x^i_j ($j \in \{1, \ldots, n\}$) for each instance x^i and a specific degree of uncertainty P. Suppose

Table 1: Description of databases

Databases	#Instances	#Attributes	#Classes
Voting Records	435	16	2
Heart	267	22	2
Tic-Tac-Toe	958	9	2
Monks	195	23	2
Balloons	16	4	2
Hayes-Roth	160	5	3
Balance	625	4	3
Lenses	24	4	3

that Θ_j is the frame of discernment relative to the attribute j. Let us denote by $|\Theta_j|$ the cardinality of Θ_j, each attribute value $v^i_{j,t}$ corresponds to an instance x^i such that $v^i_{j,t} \in \Theta_j$ $(t \in \{1, ..., |\Theta_j|\})$ will be represented through the belief function framework as follows:

$$m^{\Theta_j}\{x^i\}(\{v^i_{j,t}\}) = 1 - P \quad \text{and} \quad m^{\Theta_j}\{x^i\}(\Theta_j) = P \tag{17}$$

To evaluate the performance of our proposed k-NN classifier, we have relied on a distance criterion that measures the error rate between the test instance's bba and its real label class. It is set as follows where M corresponds to the number of classes, $P_i = \{BetP_i(H_1), \ldots, BetP_i(H_M)\}$ is the output vector of the pignistic probabilities of the bba obtained by Equation 15 and δ_{iq} equals 1 when L^i represents the real class of the test instance x^i, and 0 otherwise:

$$Distance_i = Distance(P_i, L^i) = \sum_{q=1}^{M} (BetP_i(H_q) - \delta_{iq})^2 \tag{18}$$

Then, we just have to calculate the average distance obtained by all test instances to get a final error rate. Note that the final distance should satisfy the following property:

$$0 \leqslant Distance_i \leqslant 2 \tag{19}$$

In the way, the lower the distance metric the better the classification performance can be obtained.

4.2 Experimentation results

For assessing the results, we have performed the 10-fold cross-validation technique that divides randomly a given dataset into ten equal sized parts where one part is used as a testing set and the remaining parts are used as training sets. This process will be repeated ten times where each part should be used exactly once as a test set. The distance results yielded by our new k-NN classifier are given from Figure 1 to Figure 8 for $k \in$ [1,15]. We can remark from Figure 1 to Figure 8 that our proposed classifier has yielded interesting results for the different uncertainty levels. In fact, the distance results obtained for the mentioned

benchmark data sets with the different uncertainty degrees are almost in the range [0.04,0.775]. For instance, the distance results yielded by the balance-scale database for the best values of k with No, Low, Middle and high uncertainties are respectively equal to 0.405, 0.397, 0.372 and 0.543. As well, there are equal to 0.642, 0.583 0.574 and 0.615 for the worst values of k. These encouraging results may be explained by the fact that our novel k-NN classifier has a great power for predicting the label classes of instances to be classified. We have suggested to evaluate the performance of our novel k-NN classifier against other evidential classifiers dealing also with uncertainty that arises in the attribute values. In our previous works [13, 14], we have proposed extensions of the decision tree classifiers inspired from the belief decision tree paradigm [3] to handle such kind of imperfection. Precisely, we have tackled the case of uncertainty that occurs in both construction and classification phases. For the construction step, we have mainly relied on the ratio $Gain \quad Ratio$ criterion proposed by Quinlan [7] to construct decision trees in [13], while in [14], we have relied on a $Diff$ $Ratio$ criterion based on distance that calculates the difference before and after the partitioning process has been performed using a such attribute. It is worth noting have both versions have yielded interesting results. The performance of our proposed k-NN classification technique for best and worst values of k will then be compared to [13] and [14]. The comparative results are given from Table 2 to Table 5 where $Belief \quad DT \quad Version1$ and $Belief \quad DT \quad Version2$ correspond to our extended decision trees published respectively in [13] and [14]. From the distance results, given in Table 2 to Table 5, we can remark that our proposed k-NN classifier has given mostly distance results smaller than those yielded in [13] and [14] for both best values of k (the values of k that yield the lowest distances) and worst values of k (the values of k that yield the highest distances). From this, we can conclude that our k-NN is the best performance classification technique compared with the two other ones within the framework of uncertain data represented by the evidence theory.

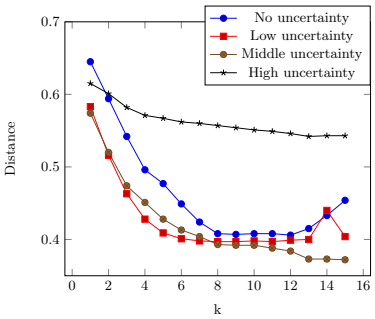

Fig. 1: Distances for Balance database

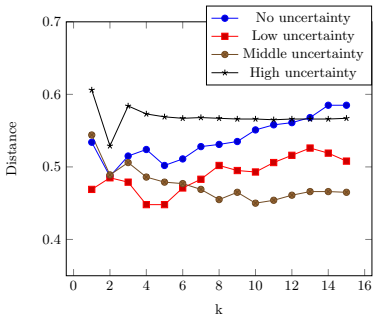

Fig. 2: Distances for Hayes-Roth database

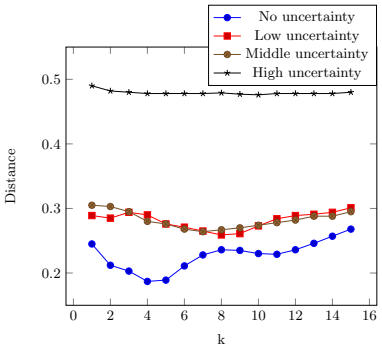

Fig. 3: Distances for Monks database

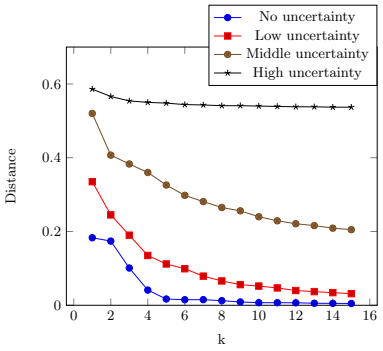

Fig. 4: Distances for Balloons database

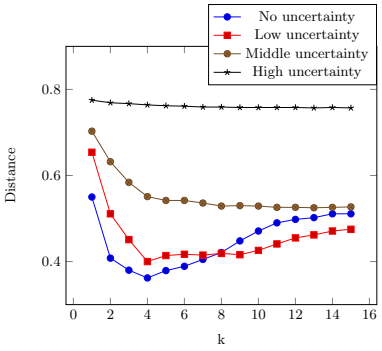

Fig. 5: Distances for Lenses database

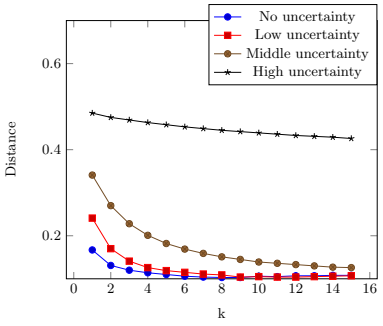

Fig. 6: Distances for Voting Records database

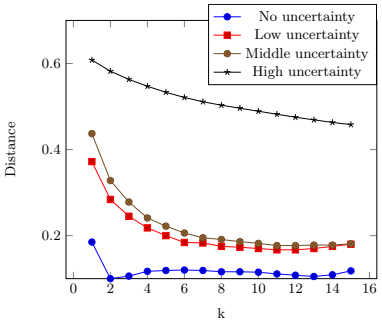

Fig. 7: Distances for Tic-Tac-Toa database

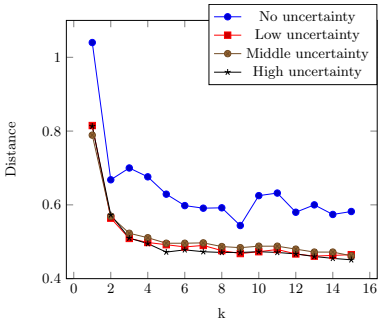

Fig. 8: Distances for Heart database

Table 2: Comparative results: No uncertainty

Bases	New k-NN Best values of k	New k-NN Worst values of k	Belief DT Version 1	Belief DT Version 2
Voting records	0.103 (k=8)	0.167 (k=1)	0.832	1.04
Heart	0.570 (k=14)	1.04 (k=1)	0.649	0.972
Tic-Tac-Toa	0.1 (k=2)	0.185 (k=1)	0.521	1.11
Monks	0.187 (k=4)	0.286(k=15)	0.726	1.18
Balloons	0.0049 (k=15)	0.183 (k=1)	0.468	1.35
Hayes-Roth	0.487 (k=2)	0.586 (k=13)	0.449	1.22
Balance-Scale	0.406 (k=12)	0.645 (k=1)	0.71	1.37
Lenses	0.362 (k=4)	0.511 (k=14)	0.45	1.15

Table 3: Comparative results: Low uncertainty

Bases	New k-NN Best values of k	New k-NN Worst values of k	Belief DT Version 1	Belief DT Version 2
Voting records	0.104 (k=8)	0.241 (k=1)	0.914	1.09
Heart	0.461 (k=13)	0.815 (k=1)	0.713	0.998
Tic-Tac-Toa	0.167 (k=13)	0.372 (k=1)	0.654	1.21
Monks	0.259 (k=8)	0.301 (k=15)	0.817	1.13
Balloons	0.0315 (k=15)	0.335 (k=1)	0.59	1.15
Hayes-Roth	0.469 (k=1)	0.526 (k=13)	0.624	1.16
Balance-Scale	0.397 (k=9)	0.583(k=1)	0.62	1.28
Lenses	0.400 (k=4)	0.654(k=1)	0.581	1.14

Table 4: Comparative results: Middle uncertainty

Bases	New k-NN Best values of k	New k-NN Worst values of k	Belief DT Version 1	Belief DT Version 2
Voting records	0.130 (k=13)	0.341 (k=1)	0.927	1.17
Heart	0.462 (k=15)	0.789 (k=1)	0.802	1.01
Tic-Tac-Toa	0.177 (k=13)	0.437 (k=1)	0.897	1.32
Monks	0.270 (k=9)	0.303(k=2)	0.901	1.01
Balloons	0.205 (k=15)	0.520(k=1)	0.9304	1.02
Hayes-Roth	0.450 (k=10)	0.544(k=1)	0.946	1.04
Balance-Scale	0.372 (k=15)	0.574(k=1)	0.94	1.25
Lenses	0.526 (k=4)	0.703(k=1)	0.925	1.028

Table 5: Comparative results: High uncertainty

Bases	New k-NN Best values of k	New k-NN Worst values of k	Belief DT Version 1	Belief DT Version 2
Voting records	0.426 (k=15)	0.485 (k=1)	0.987	1.23
Heart	0.45(k=15)	0.813 (k=1)	0.868	1.21
Tic-Tac-Toa	0.485 (k=15)	0.608 (k=1)	0.986	1.35
Monks	0.476 (k=10)	0.490(k=1)	0.985	1
Balloons	0.537 (k=14)	0.586(k=1)	1	1
Hayes-Roth	0.529 (k=2)	0.606(k=1)	0.95	1.03
Balance-Scale	0.543 (k=13)	0.615(k=1)	1	1.08
Lenses	0.757 (k=15)	0.775(k=1)	0.998	1

5 Conclusion

In this paper, we have developed a new version of the well-known k-NN classifier to handle the case of the uncertainty that exists in the attribute values and is represented with the belief function framework. Our novel k-NN technique has been compared to other belief decision tree classifiers that deal with the same kind of uncertainty. The experimental results in terms of the distance criterion have proven the efficiency of our proposed k-NN classifier compared with the two other evidential ones. As a future work, we intend to extend our proposed k-NN in order to handling numerical and mixed databases.

References

[1] A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The annals of mathematical statistics*, pages 325–339, 1967.

[2] T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer Theory. *IEEE transactions on systems, man, and cybernetics*, 25(5):804–813, 1995.

[3] Z. Elouedi, K. Mellouli, and P. Smets. Belief decision trees: theoretical foundations. *International Journal of Approximate Reasoning*, 28(2):91–124, 2001.

[4] E. Fix and J. L. Hodges Jr. Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties. Technical report, DTIC Document, 1951.

[5] L. Jiao, T. Denœux, and Q. Pan. Evidential Editing k-Nearest Neighbor Classifier. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 461–471. Springer, 2015.

[6] A. Jousselme, D. Grenier, and E. Bossé. A new distance between two bodies of evidence. *Information fusion*, 2(2):91–101, 2001.

[7] J. R. Quinlan. Induction of Decision Trees. *Machine learning*, 1(1):81–106, 1986.

[8] A. Sgarro. Possibilistic information theory: a coding theoretic approach. *Fuzzy Sets and Systems*, 132(1):11–32, 2002.

[9] G. Shafer. *A Mathematical Theory of Evidence*, volume 1. Princeton university press Princeton, 1976.

[10] P. Smets. Decision Making in the TBM: the Necessity of the Pignistic Transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005.

[11] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial intelligence*, 66(2):191–234, 1994.

[12] B. Tessem. Approximations for efficient computation in the theory of evidence. *Artificial Intelligence*, 61(2):315–329, 1993.

[13] A. Trabelsi, Z. Elouedi, and E. Lefevre. Handling Uncertain Attribute Values in Decision Tree Classifier Using the Belief Function Theory. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 26–35. Springer, 2016.

[14] A. Trabelsi, Z. Elouedi, and E. Lefevre. New decision tree classifier for dealing with partially uncertain data. In *25eme Rencontres francophones sur la Logique Floue et ses Applications (LFA'2016)*, pages 57–64, 2016.

[15] L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[16] L. M. Zouhal and T. Denoeux. An Evidence-Theoretic k-NN Rule With Parameter Optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(2):263–271, 1998.