International Journal of Approximate Reasoning 143 (2022) 192–215

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Belief functions and rough sets: Survey and new insights

Andrea Campagner [a,*], Davide Ciucci [a], Thierry Denœux [b,c]

[a] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy
[b] Université de technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, Compiègne, France
[c] Institut universitaire de France, Paris, France

ARTICLE INFO

Article history:
Received 2 November 2021
Received in revised form 24 January 2022
Accepted 24 January 2022
Available online 31 January 2022

Keywords:
Rough set theory
Evidence theory
Belief functions
Uncertainty representation
Knowledge representation
Machine learning

ABSTRACT

Rough set theory and belief function theory, two popular mathematical frameworks for uncertainty representation, have been widely applied in different settings and contexts. Despite different origins and mathematical foundations, the fundamental concepts of the two formalisms (i.e., approximations in rough set theory, belief and plausibility functions in belief function theory) are closely related. In this survey article, we review the most relevant contributions studying the links between these two uncertainty representation formalisms. In particular, we discuss the theoretical relationships connecting the two approaches, as well as their applications in knowledge representation and machine learning. Special attention is paid to the combined use of these formalisms as a way of dealing with imprecise and uncertain information. The aim of this work is, thus, to provide a focused picture of these two important fields, discuss some known results and point to relevant future research directions.

1. Introduction

Many mathematical formalisms to represent and manage uncertainty have been proposed and studied in the literature, starting from probability theory and including, among others, imprecise probabilities [4,102], fuzzy set theory [122], possibility theory [123], rough set theory [64] and belief functions [22,74]. These latter two, that is, rough set theory and belief function theory, represent particularly important examples.

Rough set theory, originally proposed by Pawlak [64], is a set-theoretic uncertainty model that allows one to study different forms of uncertainty (including: indiscernibility, granularity and ambiguity), as induced by a relation among objects of interest representing their mutual indiscernibility (i.e., equality, similarity). By contrast, belief function theory, first introduced by Dempster [22] and then formalized by Shafer [74], is a framework for representing and reasoning with uncertain information based on non-additive measures, which can be seen as an extension of set theory and probability theory [28].

Despite having had more or less orthogonal developments, the formalisms of rough set theory and belief functions are strongly related. Indeed, a theoretical relationship between the two formalisms was already noted by Pawlak [65], and further studied during the following years [46,88,106,117,126], leading to potential applications (mainly driven by the translation of properties and tools from belief functions to rough sets), especially in the fields of knowledge representation and machine learning.

* Corresponding author.
E-mail address: a.campagner@campus.unimib.it (A. Campagner).

https://doi.org/10.1016/j.ijar.2022.01.011
0888-613X/

In this survey article, we present a review of the most relevant contributions at the intersection of rough set and belief function theories focusing, in particular, on the theoretical contributions linking these two formalisms, as well as on the applications (both theoretical and practical) of such results in the areas of knowledge representation and machine learning. With respect to knowledge representation, we highlight the contributions providing theoretical links and relationships among the mathematical formalisms of rough set and belief function theories, as well as results highlighting the combined use of the two theories to represent uncertain data. From the machine learning viewpoint, we highlight the most relevant contributions providing a direct application of the above mentioned theoretical results, with special emphasis on the applications of the two formalisms to clustering and learning from uncertain or imprecise data.

The rest of the paper is structured as follows. In Section 2 we quickly review the basics of belief function and rough set theories. In Section 3 we recall the main theoretical relationships among the two formalisms and their application to knowledge representation. In particular, in Section 3.1 we discuss results related to Pawlak's rough set model, in Section 3.2 we consider the extensions of the previous results to more general rough set models, while in Section 3.3 we discuss the use of rough sets and belief functions to model uncertain and imprecise data. In Section 4 we focus on applications in Machine Learning: in particular, the relationships between evidential and rough set-based clustering models are discussed in Section 4.1, where we also provide new results on the link between credal partitions and three-way clustering. In Sections 4.2 and 4.3, we discuss applications to the management of data affected by uncertainty, respectively, in the conditional and decision attributes.

2. Background

In this section, we provide the necessary background material on belief functions and rough sets. In particular, we recall the basic notions of belief function theory in Section 2.1, including the definition of belief and plausibility functions, the relationship of the model with other uncertainty representation frameworks, the definitions of the main combination and transformation methods, as well as the basic concepts concerning the application of belief function theory to decision-making and clustering. In Section 2.2, we first introduce Pawlak's classical rough set model and its main extensions; we then summarize the applications of rough set theory in feature selection and classification, as well as the basic concepts on rough clustering and related soft clustering models.

2.1. Basic notions on belief functions

Belief function theory (also referred to as Dempster-Shafer theory or evidence theory) is a general mathematical framework for uncertainty representation and management, first proposed by Dempster and Shafer [22,74] (see also [28] for a recent introduction to belief function theory). Formally, we have the following definitions.

Definition 2.1. Let X be a finite set and 2^X the corresponding power set. A *basic belief assignment* (bba) is a function $m : 2^X \mapsto [0,1]$ s.t. $\sum_{A\in 2^X} m(A) = 1$. If $m(\emptyset) \neq 0$, then we say that the bba is *unnormalized*.

Starting from a bba, we can define two other set functions, namely the *belief* and *plausibility* functions:

Definition 2.2. Let m be a bba. Then, the belief and plausibility functions, $Bel : 2^X \mapsto [0,1]$ and $Pl : 2^X \mapsto [0,1]$ are defined, respectively, as:

$$Bel(A) = \sum_{B:\emptyset\neq B\subseteq A} m(B) \tag{1}$$

$$Pl(A) = \sum_{B:B\cap A\neq\emptyset} m(B). \tag{2}$$

It is easy to observe that Bel and Pl are dual of each other, that is $Bel(A) = 1 - m(\emptyset) - Pl(A^c)$ and $Pl(A) = 1 - m(\emptyset) - Bel(A^c)$. Given a bba, we can define the collection of *focal sets* $\mathcal{F}_m = \{A \in 2^X : m(A) \neq 0\}$, i.e., the sets for which the basic belief assignment is greater than 0. If $m(A) = 1$ for some $A \subseteq X$, then m is said to be *logical*.

Example 2.1. Let $X = \{a, b, c\}$ and let $m : 2^X \mapsto [0,1]$ be a bba defined by $m(\{a\}) = 0.1$, $m(\{b\}) = 0.2$, $m(\{b,c\}) = 0.3$ and $m(X) = 0.4$. The focal sets of m are $\mathcal{F}_m = \{\{a\}, \{b\}, \{b,c\}, X\}$. Then, taking for example $A = \{a,c\}$, we can compute the belief $Bel(A) = 0.1$, and plausibility $Pl(A) = 0.8$.

Given two bbas, m_1 and m_2, we can define their combination (also, orthogonal sum) through the so-called *Dempster's rule of combination*, namely:

$$m_1 \oplus m_2(A) = \frac{1}{1 - \mathcal{K}(m_1, m_2)} \sum_{B,C:B\cap C=A} m_1(B)\cdot m_2(C), \tag{3}$$

where $\mathcal{K}(m_1, m_2) = \sum_{A,B:A\cap B=\emptyset} m_1(A) \cdot m_2(B)$ is the *conflict* between m_1 and m_2. If $\mathcal{K}(m_1, m_2) = 1$, then the combination $m_1 \oplus m_2$ is undefined. The rule of combination $\oplus$ assumes that both sources of information m_1, m_2 can be considered as reliable. This assumption is weakened in the *disjunctive rule* [23,85], which only assumes that at least one of the two bbas is reliable:

$$m_1 \bigcirc m_2(A) = \sum_{B,C:B\cup C=A} m_1(B) \cdot m_2(C). \tag{4}$$

Example 2.2. Consider the bba defined in Example 2.1, and let m' be defined as $m'(\{a, c\}) = 0.6$ and $m(X) = 0.4$. Then, $m^* = m \oplus m'$ is defined as $m^*(\{a\}) = 0.11$, $m^*(\{b\}) = 0.09$, $m^*(\{c\}) = 0.21$, $m^*(\{a, c\}) = 0.27$, $m^*(\{b, c\}) = 0.14$ and $m^*(X) = 0.18$. On the other hand, $m^\dagger = m \bigcirc m'$ is defined as $m^\dagger(\{a, c\}) = 0.06$ and $m^\dagger(X) = 0.94$.

Discussing the connections of belief function theory with other uncertainty representation formalisms, we recall that if the focal sets are all singletons, then m is said to be *Bayesian*: in this case, m is equivalent to a probability distribution and, for each $A \subseteq X$, it holds that $Bel(A) = Pl(A)$. In this sense, belief function theory can be understood as a generalization of probability theory. Indeed, if m_1 is a Bayesian bba and m_2 is logical, it can be easily seen that the result of $m_1 \oplus m_2$ is equivalent to probabilistic conditioning.

Furthermore, it can also be shown that belief functions are related to imprecise probabilities. Indeed, any belief function *Bel* (and the corresponding bba m) can be uniquely associated to the (convex) set of probabilities that dominate (i.e., are point-wise greater than) *Bel*. Namely,

$$\mathcal{P}(m) = \{P : \forall A \subseteq X, P(A) \geq Bel(A)\}. \tag{5}$$

Thus, in the particular case where *Bel* is Bayesian, the set $\mathcal{P}_m$ contains exactly *Bel* as unique element.

If, on the other hand, the focal sets are nested (i.e. for each $A, B \in \mathcal{F}_m$, either $A \subseteq B$ or $B \subseteq A$) then m is said to be *consonant*, and it can be shown that *Bel* is a *necessity measure* and *Pl* is a *possibility measure* [28]. Thus, belief function theory can also be understood as a generalization of possibility theory [123]. However, we note that the two theories differ in their semantics and in certain formal aspects [37]. In particular, Dempster's rule (3) does not preserve consonance and is, therefore, not compatible with possibility theory.

Starting from a bba m, we can define mappings that transform m into a probability distribution. In particular, the *pignistic probability* [83,85], P_{Bet_m}, based on m is defined as:

$$P_{Bet_m}(x) = \sum_{A:x\in A} \frac{m(A)}{|A|}, \tag{6}$$

for all $x \in X$. Similarly, the *plausibility probability function* [75,101], P_{Pl_m}, is defined as:

$$P_{Pl_m}(x) = \frac{Pl(\{x\})}{\sum_{y\in X} Pl(\{y\})}. \tag{7}$$

Example 2.3. Let m be the bba defined in Example 2.1. Then $P_{Bet_m}(a) = 0.2\bar{3}$, $P_{Bet_m}(b) = 0.48\bar{3}$ and $P_{Bet_m}(c) = 0.28\bar{3}$. On the other hand, $P_{Pl_m}(a) = 0.24$, $P_{Pl_m}(b) = 0.43$ and $P_{Pl_m}(c) = 0.33$.

The two transformations generally yield different results and satisfy different properties. For example, it has been shown that the plausibility transform is the only probabilistic transformation of belief functions compatible with Dempster's rule [19], in the sense that

$$P_{Pl_{m_1\oplus m_2}} = P_{Pl_{m_1}} \otimes P_{Pl_{m_2}}, \tag{8}$$

where, for any two probability distribution P_1, P_2, the orthogonal product is defined as:

$$P_1 \otimes P_2(x) = \frac{P_1(x) \cdot P_2(x)}{\sum_{x\in X} P_1(x) \cdot P_2(x)}.$$

On the other hand, the pignistic transform is the only transformation from belief functions to probability distributions which is linear [84], i.e.,

$$P_{Bet_{\alpha m_1+(1-\alpha)m_2}} = \alpha P_{Bet_{m_1}} + (1-\alpha) P_{Bet_{m_2}} \tag{9}$$

Table 1
An example of utility matrix.

F \ D	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

Table 2
An example of extended utility matrix.

F \ D	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1
{1, 2}	0.75	0.75	0
{1, 3}	0.75	0	0.75
{2, 3}	0	0.75	0.75
{1, 2, 3}	0.67	0.67	0.67

2.1.1. Decision-making with belief functions

In this section, we recall some basic concepts of decision-making based on belief function theory [24]. Assume $D = \{d_1, \ldots, d_n\}$ is a set of *states of nature* or *classes*, and $F = \{f_1, ..., f_k\}$ is a set of *acts*. In the following, we assume that $F \subseteq 2^D$, where $f \in F$ with $|f| > 1$ represents a partial classification (or, partial assignment). We consider a utility function $U : D \times F \mapsto [0, 1]$, where the entry $U(d, f)$ corresponds to the utility of selecting act f when the true class is d. We will assume that, for each $d \in D$, $U(d, \{d\}) = k$, and, for $d' \neq d$, $U(d, \{d\}) > U(d, \{d'\})$. The information about the utility function can be expressed in matrix notation, as shown in the following example.

Example 2.4. Let $D = \{1, 2, 3\}$ and $F = \{1, 2, 3\}$. Then, Table 1 represents the utility function defined by $U(d, d) = 1$ and, for $d' \neq d$, $U(d, d') = 0$.

Starting from a basic utility function $U : D \times D \mapsto [0, 1]$, if we want to consider also the utility of partial assignments U has to be extended to $D \times 2^D$. Various ways for doing so have been considered in the literature: Campagner et al. [10] considered an approach based on arbitrary anti-tone functions, while Ma et al. [57] considered an approach based on Ordered Weighted Average (OWA) operators [110]. According to this latter approach, having fixed the utility function for the singleton acts, the extended utility function is controlled by a single parameter α, called *imprecision tolerance*, which interpolates between the maximum and the minimum OWA operators (see [57] for further details).

Example 2.5. Starting from Table 1, we want to extend the corresponding utility function to $D \times 2^D$. Assume that we adopt the OWA-based approach and set $\alpha = 0.75$ (thus, we consider the OWA operator which is half-way between the maximum and the average), then the extended utility function is represented in Table 2.

In any case, as long as $d \in f$, it must hold that $U(d, f) \geq \frac{1}{|f|}$ (otherwise, a rational decision-making agent would always prefer choosing a precise act at random, rather than a partial assignment).

Assume, then, that the knowledge of a decision-making agent is expressed through a bba $m : 2^D \mapsto [0, 1]$ on the power set of the classifications. In the standard decision-theoretic setting, a preference among acts can be determined by means of the expected utility criterion. Different generalizations of this criterion to the setting of belief function theory [24] have been considered, here we recall the following:

- *Lower Expected Utility*: $\underline{E}(f) = \sum_{B \subseteq 2^D} m(B) \min_{d \in B} U(d, f)$;
- *Upper Expected Utility*: $\overline{E}(f) = \sum_{B \subseteq 2^D} m(B) \max_{d \in B} U(d, f)$;
- *Hurwicz Expected Utility*: $E_\beta(f) = \beta \underline{E}(f) + (1 - \beta)\overline{E}(f)$, with $\beta \in [0, 1]$

Each of these criteria determines a total preorder over the acts: namely, given an expected utility function E among $\{\underline{E}, \overline{E}, E_\beta\}$, $f_1 \geq f_2$ iff $E(f_1) \geq E(f_2)$. A decision rule can then be obtained by selecting (one of) the acts with maximal expected utility.

Example 2.6. Consider the utility function given in Table 2 and let m be the bba defined by $m(1) = 0.2$, $m(\{1, 3\}) = 0.4$, $m(\{2, 3\}) = 0.2$, $m(D) = 0.2$. Based on the lower and upper expected utility criteria we get that:

- $\underline{E}(1) = 0.2$, $\overline{E}(1) = 0.8$;

- $\underline{E}(2) = 0$, $\overline{E}(2) = 0.4$;
- $\underline{E}(3) = 0$, $\overline{E}(3) = 0.8$;
- $\underline{E}(\{1,2\}) = 0.15$, $\overline{E}(\{1,2\}) = 0.75$;
- $\underline{E}(\{1,3\}) = 0.45$, $\overline{E}(\{1,3\}) = 0.75$;
- $\underline{E}(\{2,3\}) = 0.15$, $\overline{E}(\{2,3\}) = 0.75$;
- $\underline{E}(D) = 0.67$, $\overline{E}(D) = 0.67$;

So, according to the lower expected utility criterion we have:

$$D > \{1,3\} > 1 > \{1,2\} = \{2,3\} > 2 = 3$$

while, based on the upper expected utility criterion we have:

$$1 = 3 > \{1,2\} = \{1,3\} = \{2,3\} > D > 2$$

Thus, according to the Hurwicz expected utility criterion we get that 1 is the optimal act as long as $\beta \leq \frac{2}{9}$, while if $\beta > \frac{2}{9}$ the optimal act is D (i.e., the most imprecise classification).

2.1.2. Evidential clustering

In this section we recall the basic definitions of evidential clustering. Let $X = \{x_1, \ldots, x_N\}$ be a set of objects, and $C = \{c_1, ..., c_n\}$ a set of clusters. Then, a *credal partition* [34] is a collection of bbas $\{m_x\}_{x \in X}$, where for each x it holds that $m_x : 2^C \mapsto [0,1]$ and $\sum_{A \in 2^C} m_x(A) = 1$. If for some $x \in X$ it holds that $m_x(\emptyset) \neq 0$, then we say that the credal partition is *unnormalized*. Irrespective of the specific evidential algorithm used (see e.g. [3,25,27,31,34,51,59,60,127], or also the recent review [30]), the main goal of evidential clustering models is to induce a credal partition from the data [29]. This credal partition can be interpreted as a *soft clustering* of the data, i.e., the bbas represent the uncertain assignment of the objects to the clusters.

Example 2.7. Let $X = \{x_1, \ldots, x_5\}$ and $C = \{c_1, c_2, c_2\}$. Then, an example of evidential clustering is the following:

- $m_{x_1}(\{c_1\}) = 1$;
- $m_{x_2}(\{c_1\}) = 0.8, m_{x_2}(C) = 0.2$;
- $m_{x_3}(\{c_2, c_3\}) = 1$;
- $m_{x_4}(\{c_2\}) = 1$;
- $m_{x_5}(\{c_3\}) = 0.4, m_{x_5}(C) = 0.6$.

It has been shown [29] that evidential clustering is a very general framework, extending other clustering approaches, obviously including *hard clustering*, but also many other soft clustering approaches like *fuzzy clustering* [7], *rough clustering* (see Section 2.2.3) and *possibilistic clustering* [49]. In Section 4.1, we will explore the relationships among evidential clustering and rough clustering, as well as related approaches.

2.2. Basic notion on rough sets

The classical model of rough set theory, first defined by Pawlak [64], is based on the study of so-called *information tables*, namely:

Definition 2.3 *([64,66]).* An *information table* is a tuple $S = (X, Att, V, F)$ where X is a finite set of objects, Att is a finite set of attributes, $V = \cup_{a \in Att} V_a$ where V_a is the set of values of attribute a and $|V_a| > 1$, F is a function mapping objects and attributes to values, that is $F : X \times A \mapsto V$.

A *decision table* is an information table where the attributes are divided into two groups $C \cup D$, with C *condition attributes*, which represent the covariates or predictors, and D *decision attributes*, which represent the target features.

Example 2.8. An example of decision table is given in Table 3, where condition attributes are {Temperature, Pressure, Headache, Muscle Pain} and the decision attribute is Disease.

Given an information (or decision) table, the most basic notion in rough set theory is that of an indiscernibility relation $\mathcal{I}_B$ with respect to a set of attributes $B \subseteq Att$. Such a relation intuitively represents the fact two distinct objects may be indistinguishable as a consequence of the chosen representation language (i.e., the selected set of attributes B). Thus, this can be formally defined as

$$\forall x, y \in X, \quad x \mathcal{I}_B y \quad \text{iff} \quad \forall a \in B, \ F(x,a) = F(y,a).$$

Table 3
Example of decision table.

Patient	Temperature	Pressure	Headache	Muscle Pain	Disease
p_1	very high	2	yes	yes	A
p_2	high	3	no	yes	B
p_3	normal	1	yes	no	NO
p_4	high	2	yes	yes	NO
p_5	high	2	yes	yes	A

This relation is an equivalence one, which partitions X into equivalence classes $[x]_B$ that constitute *granules* of information. Due to lack of knowledge, we are not able to distinguish objects inside the granules, thus, it can happen that not all subsets of X can be precisely characterized in terms of the available attributes B.

Example 2.9. Let us consider the table in Example 2.8 and let

$$B = \{Pressure, Headache\}.$$

Then, we have three equivalence classes: $\{p_1, p_4, p_5\}$, $\{p_2\}$ and $\{p_3\}$. The set of patients $\{p_1, p_2\}$ cannot be completely characterized, in the sense that it is not the union of equivalence classes. That is to say, the two attributes in B are not sufficient to describe the two patients p_1 and p_2.

However, any set $H \subseteq X$ can be approximated by a lower and an upper approximation. Intuitively, the lower approximation $L(H)$ of a set H collects all objects which are surely contained in H: that is, $L(H)$ contains all objects whose equivalence class is inside H. On the other hand, the upper approximation $U(H)$ collects all objects which are compatible with H: that is, $U(H)$ contains all objects whose equivalence class is not disjoint from H. Formally, the two approximations can be defined as

$$L_B(H) = \{x : [x]_B \subseteq H\}, \tag{10a}$$

$$U_B(H) = \{x : [x]_B \cap H \neq \emptyset\}. \tag{10b}$$

The pair $(L_B(H), U_B(H))$ is called a *rough set*.[1]

Clearly, we have $L(H) \subseteq H \subseteq U(H)$, which justifies the names lower/upper approximations. Moreover, the boundary is defined as the collection of objects belonging to the upper approximation but not to the lower approximation, i.e., $Bnd(H) = U(H) \setminus L(H)$, while the exterior is the collection of objects not belonging to the upper approximation: $E(H) = U^c(H)$. The interpretation attached to these regions is that the objects in the lower approximation surely belong to H, the objects in the exterior surely do not belong to H and the objects in the boundary possibly belong to H. Hence, the boundary represents the uncertainty on the domain we are describing due to the insufficient ability of the attributes to discern among objects.

Other forms of imprecision arise in decision tables, when considering the decision attributes D. Indeed, it may happen that two objects with same conditions have different decision. In this case the decision table is said to be *non-deterministic*, and it is useful to introduce the *generalized decision*. Assume for simplicity that we have a single decision attribute, that is $D = \{d\}$. Then, the generalized decision for an object x, based on the attribute set $B \subseteq Att$, is the set of all decisions that have been associated to objects that are B-indiscernible from x, that is they have the same values as x on all attributes in B. Formally, the generalized decision is defined as:

$$\delta_B(x) = \{i \in V_d : \exists y \in X \text{ s.t. } F(y, d) = i \text{ and } x \mathcal{I}_B y\},$$

that is, for a given set of conditions, it collects all the possible decisions.

Example 2.10. The generalized decision of Table 3 with respect to the whole collection of condition attributes is $\delta(x_1) = \{A\}$, $\delta(x_2) = \{B\}$, $\delta(x_3) = \{NO\}$, $\delta(x_4) = \delta(x_5) = \{A, NO\}$.

Thus, in a non-deterministic situation, only a subset of objects can be precisely classified. These objects form the so-called *positive region* of the decision table, defined as

$$POS_B(X, d) = \bigcup_{x \in X} L_B([x]_{\{d\}}) = \{x \in X : |\delta_B(x)| = 1\}. \tag{11}$$

[1] Notice that sometimes with rough set it is meant a set H which cannot be described by means of the equivalence classes, in contrast with an *exact set* K such that $L_B(K) = K = U_B(K)$ [16].

Example 2.11. Let us consider objects p_4, p_5 in Table 3. Clearly, using the available symptoms (the condition attributes), we are not able to distinguish them, but on the other hand it is known that they have a different disease (the decision attribute). The positive region captures this uncertainty by not considering p_4 and p_5. Indeed, once fixed $X = \{p_1, \ldots, p_5\}$ and Att the set of all condition attributes, we have $POS_{Att}(X, d) = \{p_1, p_2, p_3\}$.

Finally, the Coefficient of Dependence of a decision d, given a set of attributes $B \subseteq Att$ represents the relative size of the positive region, thus providing a numeric representation of the granularity of the indiscernibility relation. This is defined as:

$$Dip(B, d) = \frac{|POS_B(X, d)|}{|X|}. \tag{12}$$

2.2.1. Rough sets: generalizations

Several generalizations of the basic rough set model have been introduced in the literature [111,112]. They are motivated by applications, e.g., to weaken the requirement of equivalence when generating the approximations, or by theoretical consideration. For instance, by interpreting the approximations as operators in a modal logic, we can obtain in a straightforward manner several new rough set models [113]. In the following, we recall the models that are the most relevant to the forthcoming discussion.

Generalized relation The classical model recalled in Section 2.2 is based on an equivalence relation, i.e. a reflexive, symmetric and transitive relation. By weakening or getting rid of some of these properties we obtain generalized models [48,113,128]. Particularly interesting are similarity (reflexive and symmetric) relations [99], which can account for not exactly equal instances but similar ones. For instance, we can group objects with at least 80% of equal attributes or we can impose that two objects are similar if their distance (on a set of numerical features) is less than a fixed threshold. Moreover, dominance (reflexive and transitive) relations give rise to the so-called *Dominance Based Rough Set* approach (DRSA) [82], which can handle ordinal attributes. In general, given any binary relation $R \subseteq X \times X$, the *granule* associated to $x \in X$ is defined as $R(x) = \{y \in X : (x, y) \in R\}$. Thus, $R(x)$ collects all objects that are R-related to x. Then, lower and upper approximation operators $L, U : 2^X \mapsto 2^X$ are defined as:

$$L(H) = \{x \in X : R(x) \subseteq H\}, \tag{13}$$

$$U(H) = \{x \in X : R(x) \cap H \neq \emptyset\}. \tag{14}$$

Interval rough sets Interval structures (also called *interval algebras*) represent another possible generalization of the classical model [104]. In interval rough sets, which are directly inspired by the multi-valued mapping in Dempster's original study of belief functions [22], the approximation operators (i.e., the lower and upper approximations) are determined by a multi-valued mapping between two universe sets, or, equivalently, by a compatibility relation on the same two universes. Namely, let X, Y be two sets and $R \subseteq X \times Y$ be a binary relation. Relation R defines a multi-valued mapping $r : X \mapsto 2^Y$ s.t., for each $x \in X$, $r(x) = \{y \in Y : (x, y) \in R\}$. Approximation operators $L, U : 2^Y \mapsto 2^X$ can then be defined as:

$$L(H) = \{x \in X : r(x) \subseteq H\},$$

$$U(H) = \{x \in X : r(x) \cap H \neq \emptyset\},$$

where $H \subseteq Y$. Usually, it is required that relation R is at least serial, in order to be consistent with models of rough sets based on generalized relations. In particular, it is easy to observe that the interval rough set model extends the model based on generalized relations: in particular, generalized relation models coincide with interval rough sets when the two universe sets are the same, i.e. $X = Y$ (in which case R is a binary relation on X).

Approximation spaces An approximation space [15,14] is an abstraction of the environment of rough set theory. Basically, the granulation of the universe is given for granted and it is the starting point, instead of being built from data. It is no longer necessarily a partition, but it can be a covering or a partial covering, or even a more complex structure. Generally, it comes with a set of axioms that the granules or the induced approximations should satisfy.

Definition 2.4. An approximation space is a pair $(X, G(X))$, where $G(X)$ is a granulation of the universe, i.e., a collection of sets

$$G(X) = \{G(x) : x \in X\},$$

such that $\bigcup_{x \in X} G(x) = X$.

Obviously, the approximation space and generalized relation models are related. Indeed, the neighborhood generated by any relation forms a granulation. In the particular case where the approximation space is generated by a binary relation R, we will denote it as (X, R).

While approximation spaces represent a generalization of information tables, the same idea can be applied to generalize decision tables instead:

Definition 2.5. A decision approximation space is a triple $(X, G(X), D(X))$, where $(X, G(X))$ is an approximation space and $D(X)$ is a partition of X into decision classes. That is, $D(X) = \{X_1, ..., X_d\}$, where the decision classes $X_1, \dots X_d$ are pairwise disjoint and cover X.

Compared to approximation spaces, in decision approximation spaces only the sets corresponding to the decision classes are deemed relevant.

Several definitions of approximation have been studied on approximation spaces. Only considering covering-based rough sets (i.e., approximation space models in which the granulation is a covering), more than 30 approximations are known [115]. In the following, we will not assume a particular definition of approximation, but we will require that the approximation operators satisfy the following three axioms:

(R1) $L(H^c) = [U(H)]^c$;
(R2) $L(H) \subseteq X$;
(R3) If $H \subseteq K$ then $L(H) \subseteq L(K)$.

Fuzzy rough sets Fuzzy rough sets [38] represent another possible generalization of the Pawlak's equivalence relation-based model. In the Fuzzy Rough Set model, the relation R is assumed to be a *fuzzy relation*, that is a function $R : X \times Y \mapsto [0, 1]$, and then the triple $F = (X, Y, R)$ is called a *fuzzy approximation space*. Since X and Y are allowed to be different, fuzzy approximation spaces can be understood as a generalization of the previously described Interval Rough Set model. Based on a fuzzy approximation space F, the lower and upper approximations of any fuzzy set $H : Y \mapsto [0, 1]$ can be defined as [107,108]:

$$L(H)(x) = \bigwedge_{y \in Y} (R(x, y) \implies H(y)) \tag{15}$$

$$U(H)(x) = \neg L(\neg H)(x) \tag{16}$$

where $\implies$ is a fuzzy implication, $\wedge$ is a t-norm and $\neg$ is a negation operator. For each fuzzy set H, the pair $(L(H), U(H))$ is called a *fuzzy rough set*. Fuzzy rough sets keep many relevant properties of classical rough set models, and have been subject to a rich theoretical development in the recent years. Though we do not provide further details in this sense, as such a topic could be the subject of another review due to its scope, we refer the interested reader to [21,100,107,108] for further details and developments. As a final note, we remark that the notion of fuzzy rough sets generalizes the definition of an interval rough set. Indeed, if H is a crisp set (i.e., for all $x \in X$, $H(x) \in \{0, 1\}$) and R is a crisp relation, then the two definitions coincide.

2.2.2. Feature selection and classification

Feature selection in rough set theory is based on the central notion of reduct. A reduct of a set of attributes (that is, features) is a subset of attributes that conveys the same information as the whole set, specifically, the same equivalence relation in the basic rough set model. Thus, a reduct can help in understanding which feature is relevant and which not and, as such, to perform feature selection and dimensionality reduction.

Definition 2.6 *(Reduct in information table).* Given an information table, a set of attributes B_1 is a *reduct* of B_2, with $B_1 \subseteq B_2$ if:

(R1) $x\mathcal{I}_{B_1} y \implies x\mathcal{I}_{B_2} y$, i.e. the set of attributes B_1 gives an indiscernibility relation R which is the same as (or finer than) the one given by B_2;
(R2) A minimality condition holds: $\nexists C \subset B_1$ s.t. C satisfies condition $(R1)$.

In general, multiple reducts may exist, but it can happen that some attributes belong to all the reducts. This set of attributes is named the *core* and it represents the set of *indispensable features*. We can thus divide the features in three categories: indispensable (the core), important but dispensable (belonging to some reducts but not to the core), useless (not belonging to any reduct). However, this clear-cut distinction may not be satisfying in many situations, where we may tolerate some error and thus define a grade of *dispensability* or otherwise stated *importance* of the different features. This leads to what are called *approximate reducts* [63]. Let us remark the following points on reducts:

- Computing the shortest reduct is a NP - hard problem, while computing the set of all reducts is a NP^{NP}-hard problem [79,98]. Hence, several heuristic and approximation algorithms have been proposed to address the problem of finding reducts, see, e.g., the surveys [6,90].
- Several generalized definitions have been given, in particular in cases of inconsistent [125] and incomplete data [50]. For a recent survey, see [43].

When the available information is in the form of a decision table, reducts can be defined in different ways, e.g., based on the generalized decision, or on the positive region:

Definition 2.7 *(Reduct in decision table - generalized decision).* Given a decision table, a set of attributes B_1 is a *reduct* of B_2, with $B_1 \subseteq B_2$, if

(RG1) B_1 and B_2 generate the same generalized decision: for all objects $x \in X$, $\delta_{B_1}(x) = \delta_{B_2}(x)$;
(RG2) Minimality: $\nexists C \subset B_1$ s.t. $\delta_C = \delta_{B_1} = \delta_{B_2}$.

Definition 2.8 *(Reduct in decision table - positive region).* Given a decision table, a set of attributes B_1 is a *reduct* of B_2, with $B_1 \subseteq B_2$, if

(RP1) $Dip(B_1, d) = Dip(B_2, d)$;
(RP2) Minimality: $\nexists C \subset B_1$ such that $Dip(C, d) = Dip(B_1, d)$.

Example 2.12. Consider the decision table given in Table 3. The set of attributes $R = \{Temperature, Pressure\}$ is a generalized decision-based reduct of Att. Indeed, for Att we have that $\delta_{Att}([p_1]_{Att} = \{p_1\}) = \{A\}, \delta_{Att}([p_2]_{Att} = \{p_2\}) = \{B\}, \delta_{Att}([p_3]_{Att} = \{p_3\}) = \{NO\}, \delta_{Att}([p_4]_{Att} = \{p_4, p_5\}) = \{NO, A\}$. The same equivalences can be shown to hold also for R. More generally, it can be shown that the only other reduct is $\{Temperature, Headache\}$. Thus, in particular, the *core* is equal to $\{Temperature\}$.

The same conclusion also holds for positive region-based reducts. Indeed, the positive region determined by Att is $POS_{Att}(X, d) = \{p_1, p_2, p_3\}$. Thus, the set of attributes $R = \{Temperature, Pressure\}$ is a reduct, since $POS_R(X, d) = POS_{Att}(X, d) = \{p_1, p_2, p_3\}$. Similarly, the set $\{Temperature, Headache\}$ is also a reduct and thus, also in this case, the *core* is equal to $\{Temperature\}$.

Many other definitions of a reduct exist [43,81]: the relationships among these different models have recently been explored in [81].

Finally, we note that reducts can be useful not only for feature selection but also to perform classification [6,86]. Indeed, classification rules can be deduced from a reduct. Let us consider a reduct $R = \{a_1, a_2, \ldots, a_n\}$. Then, for any object equivalence class $[x]_R$, we can define the rule

$$\text{If } (a_1 = F(x, a_1)) \text{ and } \ldots \text{ and } (a_n = F(x, a_n)) \text{ then } d \in \delta_R(x).$$

Intuitively, such a rule can be used for classifying new objects: if the new object x to be classified fits the pattern described by the antecedent of a given rule (i.e., x belongs to some equivalence class $[y]_R$, where R is a reduct), then its decision should be one of those in the associated generalized decision (i.e., the correct classification for x lies in $\delta_R(y)$).

Example 2.13. Consider the decision table in Table 3. As shown in Example 2.12, the two reducts of the decision table are $R_1 = \{Temperature, Pressure\}$ and $R_2 = \{Temperature, Headache\}$. Thus, R_1 induces the following rules:

$$\text{If } Temperature = \text{very high and } Pressure = 2 \text{ then } A$$
$$\text{If } Temperature = high \text{ and } Pressure = 3 \text{ then } B$$
$$\text{If } Temperature = normal \text{ and } Pressure = 1 \text{ then } NO$$
$$\text{If } Temperature = high \text{ and } Pressure = 2 \text{ then } A \text{ or } NO$$

On the other hand, R_2 induces the following set of rules:

$$\text{If } Temperature = \text{very high and } Headache = yes \text{ then } A$$
$$\text{If } Temperature = high \text{ and } Headache = no \text{ then } B$$
$$\text{If } Temperature = normal \text{ and } Headache = no \text{ then } NO$$
$$\text{If } Temperature = high \text{ and } Headache = yes \text{ then } A \text{ or } NO$$

2.2.3. Rough clustering

The concept of rough clustering [72] is a generalization of classical clustering, incorporating some of the principles of rough set theory. In this clustering model, each cluster c is approximated by a *lower cluster* $l(c)$, i.e., the elements that

"certainly" belong to the cluster, and an *upper cluster* $u(c)$, i.e., the elements that "possibly" belong to the cluster. Thus, rough clustering (similarly to evidential clustering reviewed in Section 2.1.2) is a soft clustering model, since the assignment of objects to clusters is not necessarily clear-cut.

Multiple rough clustering approaches have been considered in the literature [52,53,62,69,70,76,97], see also the section on clustering in the review by Bello and Falcon [6]. Nonetheless, the rough clustering model itself has been formalized by Lingras and Peters in [53], using the following three axioms:

(RC1) Any instance x belongs to at most one lower approximation:

$$\forall x \in X,\ \exists \langle l(c), u(c)\rangle \text{ s.t. } x \in l(c) \Rightarrow \exists! \langle l(c), u(c)\rangle \text{ s.t. } x \in l(c);$$

(RC2) For each cluster, the lower approximation is contained in the upper one:

$$\forall \langle l(c), u(c)\rangle,\ l(c) \subseteq u(c);$$

(RC3) If an object x does not belong to any lower approximation, then it belongs to at least two upper approximations:

$$\forall x \in X,\ \nexists \langle l(c), u(c)\rangle \text{ s.t. } x \in l(c) \Rightarrow$$
$$\exists \langle l(c_i), u(c_i)\rangle, \langle l(c_j), u(c_j)\rangle, i \neq j \text{ s.t. } x \in u(c_i), u(c_j).$$

Example 2.14. Let $X = \{x_1, ..., x_5\}$ and $C = \{c_1, c_2, c_2\}$. Then $l(c_1) = \{x_1, x_2\}, u(c_1) = \{x_1, x_2, x_3\}$, $l(c_2) = \{x_4\}, u(c_2) = \{x_3, x_4, x_5\}$ and $l(c_3) = \emptyset, u(c_3) = \{x_5\}$ is a rough clustering.

Yao et al. in [114] introduced *interval–set clustering* and Yu in [118] introduced *three–way clustering*. As in rough clustering, in both interval–set clustering and three–way clustering (which are equivalent, from a formal point of view) each cluster c is approximated by a lower and an upper approximation $[c_l, c_u]$. However, they impose different conditions on the clusters:

(IC1) The lower approximation of any cluster cannot be empty: $\forall c,\ c_l \neq \emptyset$;
(IC2) The upper approximations are a covering of the universe: $\bigcup_c c_u = X$;
(IC3) All lower approximations are disjoint: $\forall i, j, i \neq j \Rightarrow c_l^i \cap c_l^j = \emptyset$.

Example 2.15. Let $X = \{x_1, ..., x_5\}$ and $C = \{c_1, c_2, c_2\}$. Consider the rough clustering defined in Example 2.14. Then, the latter one is not a three-way clustering, since $l(c_3) = \emptyset$. On the other hand, consider $c_l^1 = \{x_1, x_2\}, c_u^1 = \{x_1, x_2, x_3\}$, $c_l^2 = c_u^2 = \{x_4\}$ and $c_l^3 = c_u^3 = \{x_5\}$. Then, this is a three-way clustering. Notice, however, that this is not a rough clustering: indeed x_3 belongs only to a single boundary set, namely c_u^1.

The two sets of axioms are related. Indeed, obviously, $IC3$ is equivalent to axiom $RC1$. Also, it can be easily seen that axioms $IC1, IC2, IC3$ imply axiom $RC2$ while, conversely, axioms $RC1, RC2, RC3$ imply axiom $IC2$. However, the two systems are not equivalent. Indeed, rough clustering allows lower approximations to be empty (and in this case, the corresponding objects are required to belong to at least two upper approximations), which is not allowed in three–way clustering due to axiom $IC1$. On the other hand, three–way (and interval–set) clustering allows objects to be in only one upper approximation, which is not allowed in rough clustering due to axiom $RC3$. Nonetheless, in [11], it has been shown that any three–way clustering can be represented as an equivalent rough clustering by including a *noise cluster*, i.e., a cluster to which the object that belong to only a single upper approximation is assigned. For recent work on interval-set and three-way clustering see [1,2,103,119–121], as well as the section on clustering in the recent review by Campagner et al. [9].

3. Knowledge representation

In this section, we review the most relevant results concerning the theoretical relationship between rough set belief function theories, focusing, in particular, on the knowledge representation aspect of these two frameworks. In Section 3.1 we recall the basic results relating Pawlak's rough set model with belief functions, while in Section 3.2 we consider extensions of these results to more general rough set models. Finally, in Section 3.3, we consider knowledge representation models that combine rough sets and belief functions to represent uncertain and imprecise data.

3.1. Pawlak rough sets

The link between Pawlak rough sets and belief functions is evident from their definitions, and it was already recognized in the first years after Pawlak seminal work on rough sets. Indeed, in the introduction of "Rough sets and probability" [65], we can read:

the inner and outer probabilities considered in this paper may be viewed as a special case of lower and upper probabilities introduced by Dempster and as a special case of Shafer's belief theory.

Formally, given the lower and upper approximations, defined as in Pawlak's rough set model, we can define a pair of belief and plausibility functions, as stated in the following proposition.

Proposition 3.1 *([77]). Let L, U be the lower and upper approximations induced by an equivalence relation, according to (10), on a universe X. Then, for all $H \subseteq X$,*

$$Bel(H) = \frac{|L(H)|}{|X|}, \tag{17a}$$

$$Pl(H) = \frac{|U(H)|}{|X|} \tag{17b}$$

are a pair of dual belief and plausibility functions. The corresponding bba is:

$$m(H) = \begin{cases} \frac{|H|}{|X|} & \text{if } H \text{ is an equivalence class} \\ 0 & \text{otherwise.} \end{cases}. \tag{18}$$

Conversely, given a pair of belief and plausibility functions we can give the conditions in order to obtain a pair of lower and upper approximations.

Proposition 3.2 *([117]). Let Bel and Pl be a belief and plausibility functions satisfying the following two conditions:*

1. *The set of focal elements forms a partition;*
2. *$m(H) = \frac{|H|}{|X|}$ for every focal set H.*

Then, there exists an equivalence relation R such that the derived lower and upper approximations satisfy (17). More precisely, R is defined as:

$$R = \{(x, y) \in X^2 : \exists H \in \mathcal{F}_m \text{ s.t. } x, y \in H\}.$$

Example 3.1. Consider the decision table given in Table 3, and let $B = \{Headache, Pressure\}$. Then, the equivalence classes are $[p_1]_B = \{p_1\}$, $[p_2]_B = \{p_2\}$, $[p_3]_B = \{p_3\}$ and $[p_4]_B = \{p_4, p_5\}$. Thus, B determines the mass function m_B s.t. $m_B(\{p_1\}) = m_B(\{p_2\}) = m_B(\{p_3\}) = \frac{1}{5}$ and $m_B(\{p_4, p_5\}) = \frac{2}{5}$. If we let $H = \{p_1, p_2, p_4\}$, it holds that $Bel(H) = \frac{2}{5}$ and $Pl(H) = \frac{4}{5}$.

The interpretation of the relation between, on the one hand, the lower/upper approximations L, U and, on the other hand, the belief/plausibility functions Bel, Pl is particularly intuitive when we consider a decision table. In this setting [78], the following result holds:

Proposition 3.3 *([78]). Let $S = (X, Att, d, V, F)$ be a decision table, let $B \subseteq Att$ and let δ_B be the generalized decision defined by B. Denote with $X_i = \{x \in X : F(x, d) = i\}$ the decision class corresponding to decision i. Then, we can define a bba $m_S^B : 2^{V_d} \mapsto [0, 1]$, s.t. $\forall \Delta \subseteq V_d$:*

$$m_S^B(\Delta) = \frac{|\{x \in X : \delta_B(x) = \Delta\}|}{|X|} \tag{19a}$$

$$Bel_S^B(\Delta) = \frac{|L(\bigcup_{i \in \Delta} X_i)|}{|X|} \tag{19b}$$

$$Pl_S^B(\Delta) = \frac{|U(\bigcup_{i \in \Delta} X_i)|}{|X|} \tag{19c}$$

Thus, the belief (resp., plausibility) of a set of decision values can be interpreted naturally as the probability of its lower (resp., upper) approximation, while the bba can be naturally interpreted as assigning to each possible set of decision values the corresponding probability (or, frequency). Furthermore, the authors of [78] also show an interesting relationship between Dempster's rule of aggregation and the independent product of two decision tables defined as follows:

Definition 3.1. Let $S_1 = (X_1, Att_1, d_1, V_1, F_1), S_2 = (X_2, Att_2, d_2, V_2, F_2)$ be two decision tables. The independent product $S_1 \odot S_2$ is defined as the decision table $S = (X, Att, d)$ given by:

(i) $X = (X_1 \times X_2) \setminus (X_1 \otimes X_2)$;
(ii) $d((x_1, x_2)) = \delta_{Att_1}(x_1) \cap \delta_{Att_2}(x_2)$;
(iii) $Att = Att_1 + Att_2 = (Att_1 \times \{1\}) \cup (Att_2 \times \{2\})$;
(iv) $V = V_1 \cup V_2$;
(v) For $(a, i) \in Att$ and $(x_1, x_2) \in X$, $F((x_1, x_2), (a, i)) = F_i(x_i, a)$

where $X_1 \otimes X_2 = \{(x_1, x_2) \in X_1 \times X_2 : \delta_{Att_1}(x_1) \cap \delta_{Att_2}(x_2) = \emptyset\}$.

Under the definition given above, the bba obtained from the independent product of two decision tables is shown in [78] to be equivalent to the Dempster combination of the corresponding bbas. That is, the following result holds:

Proposition 3.4 *([78]). Let S_1, S_2 be two decision tables. Then:*

$$m_{S_1 \odot S_2} = m_{S_1} \otimes m_{S_2}.$$

This result implies that, intuitively, Dempster's rule of combination (at least when restricted to bbas that can arise from Pawlak's rough sets, that is bbas for which the collection of focal sets is a partition) can be interpreted as an operation of combination on the underlying sources of information (that is, the decision tables).

Remarks and prospects About the interpretation of the link shown above, Pawlak stated [67]:

> Rough set theory is objective — for a given information table, qualities of corresponding approximations are computed. On the other hand, the Dempster-Shafer theory is subjective — it is assumed that values of belief (or plausibility) are given by an expert.

This comment highlights how the rough set-theoretic approach, at least as originally conceived by Pawlak, can be understood as being of value only insofar as it is directly linked with, and grounded on, the available data, which should guide the application of the theory in practical scenarios. This is contrasted with Dempster-Shafer theory, in which knowledge (as represented by a belief/plausibility function) is elicited by a human expert (hence, subjective). Although this comment is correct in highlighting the fact that rough sets and belief functions may have different interpretations and semantics, it neglects the fact that, in practical applications, belief functions are not necessarily expert-elicited but can also be computed from the data (in the same way as reducts or decision rules in rough set theory). In this sense, the equivalences in [77,117] can be interpreted as a way to compute belief and plausibility functions based on available data, by means of rough set theory. This semantics has been considered, e.g., also in [47,78] where the authors also provide an interpretation of the Dempster's rule of combination, in terms of a combination operator on information tables which is conceptually similar to a join between relations (in the sense of relational algebra). Despite these interesting links and interpretations, we believe that further research should be devoted at understanding the relationship between other fundamental concepts in belief function theory (such as, e.g., other combination rules, marginalization, vacuous extension, etc.) and Pawlak's rough set model.

3.2. Generalized rough sets

Starting from the initial results regarding the connection between Pawlak's model and belief functions, similar results have also been explored for generalized rough set models.

Proposition 3.5 *([117]). Let us consider an approximation space (X, R), where R is at least a serial relation, with the lower and upper approximation defined according to (13)-(14). Then, a pair of dual belief and plausibility functions can be defined as in (17). The corresponding bba is defined as*

$$m(H) = \frac{|\{x \in X : R(x) = H\}|}{|X|}. \tag{20}$$

So, to any serial relation corresponds a belief assignment such that the mass can assume only rational values. Indeed, the converse of the above proposition exactly confirms this relationship.

Proposition 3.6 *([117]). Let Bel and Pl be a belief and plausibility such that for all subsets $H \subseteq X$, $m(H)$ is a rational number with denominator $|X|$. Then, there exists a serial relation R such that the derived lower and upper approximations satisfy (17). In particular, we can define*

$$R(x) = \{y \in X : \exists H \in \mathcal{F}_m \text{ s.t. } \{x, y\} \subseteq H\}.$$

More recently, Zhang et al. [126] also showed that the properties of the relation R uniquely determine the properties of the collection $\mathcal{F}$ of focal sets for the corresponding belief function, and vice-versa. Namely, the following result holds:

Proposition 3.7 *([126]). Let (X, R) be an approximation space, with R being at least serial, and let $\mathcal{F}$ be the collection of focal sets determined by the belief function induced by (X, R) according to (17). Then:*

- *R is symmetric iff $\exists \phi : X \mapsto \mathcal{F}$ s.t. $\forall x, y \in X, x \in \phi(y)$ iff $y \in \phi(x)$;*
- *R is transitive iff $\exists \phi : X \mapsto \mathcal{F}$ s.t. $\forall x, y \in X, x \in \phi(y) \implies \phi(x) \subseteq \phi(y)$;*
- *R is reflexive iff $\exists \phi : X \mapsto \mathcal{F}$ s.t. $\forall x \in X, x \in \phi(x)$.*

In order to consider any rational number, we have to take into account interval rough sets. Indeed, in this case, similar propositions as the previous ones can be obtained by substituting the approximations obtained by a serial relation with the ones obtained by a relation in interval rough sets. Namely, the following result holds:

Proposition 3.8 *([117]). Let (X, Y, R) be an interval structure, with $R \subseteq X \times Y$ being at least a serial relation. Then a pair of dual belief and plausibility functions can be defined as in (17).*

As in the previous cases, also the converse result holds, namely:

Proposition 3.9. *Let Bel and Pl be a belief and plausibility functions such that, for each subset $H \subseteq X$, $m(H)$ is a rational number. Then, there exists an interval structure (X, Y, R) such that the derived lower and upper approximations satisfy (17). In particular, assume that there exists a mapping $\Gamma : X \mapsto 2^Y$ and a probability distribution P on X s.t. for all $A \subseteq Y$,*

$$m(A) = \sum_{x \in X : \Gamma(x) = A} P(x).$$

Then, the compatibility relation R can be defined by $(x, y) \in R \iff y \in \Gamma(x)$.

As already noted in Section 2.2.1, this relationship is not particularly surprising, since the definition of interval rough sets has been directly inspired by belief function theory [22]. The same results have also been extended to general continuous belief function by means of *random information systems* [109], in which a probability distribution P on the objects (or, on the σ-algebra determined by the given relation) is also given.

More recently, Tan et al. [89] also investigated the relationship between general approximation spaces and belief functions.

Definition 3.2. Let $(X, G(X))$ be an approximation space. Given an element $a \in X$, we define:

- The *approximated set of a* as the collection of sets for which a belongs to the corresponding lower approximation, that is $\mathcal{G}(a) = \{A \subseteq X : a \in L(A)\}$;
- The *minimal approximated set of a* as the collection of sets, within the approximated set of a, for which no subset is also in the approximated set of a. That is, $\mathcal{M}_G(x) = \{A \in \mathcal{G}(a) : \forall B \in \mathcal{G}(a),\ B \subseteq A \Rightarrow B = A\}$.

For each set H, we then define $j(H) = \{x : H \in \mathcal{M}_G(x)\}$, that is $j(H)$ is the set of all objects for which H is in the corresponding minimal approximated set.

In case of an approximation space the following proposition can be proved.

Proposition 3.10 *([89]). Given an approximation space and a pair of lower-upper approximations on (X, G) satisfying properties (R1)–(R3),*

1. *We can define a basic belief assignment on the universe X as*

$$m(H) = \begin{cases} 0 & j(H) = \emptyset \\ \frac{1}{|X|} \sum_{x \in j(H)} \frac{1}{|\mathcal{M}_G(x)|} & \text{otherwise;} \end{cases}$$

2. *$\forall x,\ |\mathcal{M}_G(x)| = 1$ iff $\forall H \subseteq X$, $Bel(H) = \frac{|L(H)|}{|X|}$ and $Pl(H) = \frac{|U(H)|}{|X|}$.*

In case of a decision approximation space, we need similar definitions as before, but referred to a decision class.

Definition 3.3. Let $(X, G(X), D(X))$ be a decision approximation space. Given an element $a \in X$, we define

- The *approximated set of a* as the set $\mathcal{G}(a) = \{A \subseteq X : a \in A,\ A \subseteq L(X_i)$ for some decision class $X_i\}$;
- The *minimal approximated set of a* as the set $\mathcal{M}_G(x) = \{A \in \mathcal{G}(a) : \forall B \in \{\mathcal{G}\}(a),\ B \subseteq A \Rightarrow B = A\}$;
- $j(H) = \{x \in X : H \in \mathcal{M}_G(x)\}$.

Then, the following result holds.

Proposition 3.11 *([89]). Given a decision approximation space and a pair lower-upper approximations on $(X, G(X), D(X))$ satisfying properties (R1)–(R3),*

1. *We can define a basic belief assignment on the universe X as*
$$m(H) = \begin{cases} 0 & j(H) = \emptyset \\ \frac{1}{|Pos|} \sum_{x \in j(H)} \frac{1}{|\mathcal{M}_G(x)|} & otherwise \end{cases};$$
2. *For all decision class X_i, $Bel(X_i) = \frac{|L(X_i)|}{|Pos|}$ and $Pl(X_i) = \frac{|U(X_i)|}{|Pos|}$.*

Finally, the relationship between belief function theory and rough set theory has been explored also in the fuzzy setting, by drawing a connection between fuzzy rough set models [38,108] and fuzzy belief functions [8,56,26]. In particular, Wu et al. [108] proved that, given a fuzzy approximation space (X, Y, R) and a fuzzy set H, the lower and upper probabilities of H,

$$\underline{P}(H) = \sum_{x \in X} L(H)(x) \tag{21}$$

$$\overline{P}(H) = \sum_{x \in X} U(H)(x) \tag{22}$$

are exactly the belief and plausibility of the fuzzy event H [8,26]. Furthermore, they show that Propositions 3.8, 3.10 and 3.11 similarly apply to the fuzzy case, and can also be extended to the case where X is infinite, by noting that $\underline{P}(H), \overline{P}(H)$ are always, respectively, a fuzzy monotone Choquet capacity and a fuzzy alternating Choquet capacity of infinite order (hence, they always defined fuzzy belief and fuzzy plausibility functions). Furthermore, the results obtained by Wu et al. [108] have also been more recently extended to more general families of fuzzy approximation spaces, including fuzzy coverings [39].

Remarks and prospects In regard to the interpretation of these results, the same comments made for Pawlak rough sets hold. A particularly important direction of research is to determine in which cases (both practical and theoretical) the above mentioned relationships can be proved useful. So far, results in this sense have been proposed only for generalized relation models: for example, the relationship between similarity-based rough sets (i.e. rough set models in which the underlying relation is symmetric and reflexive) and belief functions has been studied in the setting of feature selection with missing or set-valued data (see Section 4.2). Similar relationships and practical applications for the more general results presented in this section have not been considered in the literature. Consequently, similarly to the case of Pawlak's rough sets model, the semantics and usefulness of the presented results should be further clarified. Further, in this survey we only provided a short account of the results dealing with the fuzzy case: this is a necessary consequence of the broadness of this topic. Nonetheless, due to recent interest in this area, also from an application-oriented point of view (see, e.g., [26,55,41]), further work should be devoted at exploring this rich subject.

3.3. Uncertain decision tables

An uncertain decision table [93] is a decision table where the decision attribute is not exactly known and the uncertainty is expressed in the form of a belief basic assignment on the possible values.

Definition 3.4. An uncertain decision table (UDT) is a structure $(X, Att \cup \{d\}, Val, F, \{m_x\}_{x \in X})$, where $(X, Att \cup \{d\}, Val, F)$ is a decision table according to Definition 2.3 and, for each $x \in X$, $m_x : 2^{Val_d} \mapsto [0, 1]$ is a basic belief assignment on the set of decision values.

Example 3.2. In Table 4, we illustrate an example of an UDT.

Due to this generalized definition, a new definition of decision classes is required. In [93], they are defined according to a tolerance (that is, similarity) relation based on the distance between two bbas. Given a decision value $v \in Val_d$, let m_v be a bba s.t. $m_v(\{v\}) = 1$. Then, given a threshold θ, the θ-tolerance class X_v represents the collection of all objects x whose bba-valued decision is not too dissimilar from the bba associated to the decision v. Formally, this can be defined as

Table 4
Uncertain decision table.

Patient	Temperature	Headache	Disease
p_1	very high	yes	$m(A)=0.8, m(B)=0.05,$ $m(\{A,B\})=0.15$
p_2	high	no	$m(B)=0.7, m(\{A,B\})=0.3$
p_3	normal	no	$m(NO)=1$
p_4	high	yes	$m(A)=0.1, m(NO)=0.8,$ $m(\{A,NO\})=0.1$
p_5	high	yes	$m(A)=0.2, m(B)=0.2,$ $m(NO)=0.2, m(\{A,B,NO\})=0.4$

$$X_v = \{x \in X : \text{dist}(m_v, m_x) < 1-\theta\},$$

where dist denotes Jousselme's distance [45] defined as

$$\text{dist}(m_1, m_2) = \sqrt{\frac{1}{2}(\|\overrightarrow{m_1}\|^2 + \|\overrightarrow{m_2}\|^2 - 2\langle\overrightarrow{m_1}, \overrightarrow{m_1}\rangle)}, \tag{23}$$

with

$$\langle\overrightarrow{m_1}, \overrightarrow{m_1}\rangle = \sum_{i=1}^{|2^{Val_d}|}\sum_{j=1}^{|2^{Val_d}|} m_1(A_i)m_2(A_j)\frac{|A_i \cap A_j|}{|A_i \cup A_j|}, \tag{24}$$

and $A_i, A_j \in 2^{Val_d}$. In (24), the sum is over all the possible subsets of values that the decision attribute can assume.

Example 3.3. Consider the UDT defined in Example 3.2. Then, if we select $\theta = 0.75$, the θ-tolerance classes are defined as: $X_A = \{p_1\}$, $X_B = \{p_2\}$, and $X_{NO} = \{p_3, p_4\}$.

In order to compute the approximations, we first need to combine the bbas inside each equivalence class defined on the condition attributes. In [93], averaging is proposed for performing this operation. Denote the equivalence class defined according to a set B of condition attributes as $[x]_B$ and the aggregated bbas as $m_{[x]_B}$. Then, the lower approximation of the tolerance class X_v collects all the objects for which the equivalence class is surely contained in X_v and, furthermore, the corresponding averaged bba is not too dissimilar from the bba associated to v. Thus, for each decision class X_v, the lower approximation is defined as

$$L(X_v) = \{x \in X : [x]_B \subseteq X_v \quad \text{and} \quad \text{dist}(m_v, m_{[x]_B}) \leq 1-\theta\}. \tag{25}$$

The definition of the upper approximation is unchanged. As we show in Section 4.3, the notion of UDT and the above definitions of approximations have been used by Lingras et al. [93–95] to perform feature selection and classification based on uncertain (in particular, evidential) data.

Example 3.4. Let $C = \{Headache\}$. Then, $m_{[p_1]_C} = (m_{p_1} + m_{p_4} + m_{p_5})/3$ and, for the decision classes, it holds that:

- $m_{[p_1]_C}(A) = 0.67$;
- $m_{[p_1]_C}(B) = 0.08$;
- $m_{[p_1]_C}(\{A,B\}) = 0.05$;
- $m_{[p_1]_C}(NO) = 0.33$;
- $m_{[p_1]_C}(\{A,NO\}) = 0.03$;
- $m_{[p_1]_C}(\{A,B,NO\}) = 0.13$.

On the other hand, $m_{[p2]_C} = (m_{p_2} + m_{p_3})/2$, with:

- $m_{[p2]_C}(B) = 0.35$;
- $m_{[p2]_C}(\{A,B\} = 0.15$;
- $m_{[p2]_C}(NO) = 0.5$.

Therefore $L(X_A) = L(X_B) = L(X_{NO}) = \emptyset$, while $U(X_A) = \{p_1, p_4, p_5\}$, $U(X_B) = \{p_2, p_3\}$ and $U(X_{NO}) = X$.

A related, but different, approach to the representation and management of uncertain decision tables using rough sets and belief functions was also considered in [13]. In this work, the authors consider a specific form of UDT, in which the

decision attribute is set-valued: this implies that the bbas attached to the objects are logical (i.e., they have a single focal set). Based on these definitions, for each equivalence class $[x]_B$ induced by a set of attributes $B \subseteq Att$, we can define a bba whose mass values are simply the relative frequency of each set of decisions, namely

$$m_{[x]_B}(A) = \frac{|\{y \in [x]_B : m_y(A) = 1\}|}{|[x]_B|}. \tag{26}$$

The authors, however, do not provide a generalization of the approximations, focusing on the problem of finding reducts.

Remarks and prospects Interestingly, the definition of approximations in [93] is based on the average combination rule [61] to combine the bbas of the instances in a given equivalence class, rather than Dempster's rule. This alternative rule of combination is also coherent with the semantics of belief functions (indeed, the average of belief functions is still a belief function) and has the effect that the aggregation is always possible, even when two bbas are in conflict. For this reason, the authors suggest that averaging may be better suited to this context than Dempster's rule. Nonetheless, we believe that further research could be devoted to exploring different notions of approximations based on other aggregation operators, such as Dempster's rule or the disjunctive rule (4), which could be useful to account for possible inconsistencies in the UDT. Similarly, further research could investigate how the approximation defined in (25) changes if we consider different distance functions among bbas [45].

4. Applications to machine learning

In this section, we discuss the relationships between belief function and rough set theories in the field of Machine Learning, focusing on applications in which the two formalisms, and particularly the theoretical results shown in Section 3, have been related or combined to solve relevant uncertainty representation or management problems. In Section 4.1, we discuss the connections between rough clustering (and related formalisms, such as three-way clustering) and evidential clustering: we first discuss known results about the relationship between rough and evidential clustering, and we provide an original result extending the previously mentioned connection to three-way clustering. In Section 4.2, we focus on the use of rough set and belief function theories to model uncertainty and data missingness in the covariates, studying, in particular, the problem of feature selection. Finally, in Section 4.3, we continue the review of work related to UDTs (see Section 3.3), focusing on the problems of feature selection and rule induction.

4.1. Clustering

As highlighted in Section 2.1.2, evidential clustering [29] is based on the concept of a *credal* partition, in which each object x_i is assigned a mass function m_i describing an uncertain assignment to clusters. It can easily be seen that rough clustering can be understood as a special case of evidential clustering. Indeed, we have the following result:

Theorem 4.1 *([29]). To each evidential clustering where all mass functions are logical (i.e., $\forall x \in X, \exists A \in 2^C$ s.t. $m_x(A) = 1$) corresponds a unique rough clustering. Vice-versa, to each rough clustering corresponds a unique evidential clustering where all mass functions are logical and normalized (i.e., $\forall x \in X, \exists A \in 2^C \setminus \{\emptyset\}$ s.t. $m_x(A) = 1$).*

More generally, given any evidential clustering, we can obtain from it a rough clustering. For example, a particularly simple approach to perform such a transformation is to apply the *maximum mass rule* [29]. That is, for each object x, we take the set of clusters A to which the bba m_x assigns maximal mass. Then, we say that x is in the upper cluster of every cluster in A: if, furthermore, $|A| \leq 1$, then we also say that x is in the lower cluster of A. Formally, we define $A_x = \arg\max_A m(A)$. Then, if $A_x = \{c\}$, set $x \in l(c)$. Otherwise, for each $c \in A_x$, set $x \in u(c) \setminus l(c)$. Finally, all x s.t. $A_x = \emptyset$ are assigned to the lower cluster of a *noise cluster* (thus, in general the rough clustering may have one additional cluster as compared to the starting evidential clustering).

Example 4.1. Consider the credal partition $\mathcal{M} = \{m_{x_1}, ..., m_{x_5}\}$ introduced in Example 2.7. Applying the maximum mass rule, we obtain the following rough clustering: $l(c_1) = \{x_1, x_2\}, u(c_1) = \{x_1, x_2, x_5\}$, $l(c_2) = \{x_4\}, u(c_2) = \{x_3, x_4, x_5\}$, $l(c_3) = \emptyset, u(c_3) = \{x_3, x_5\}$. Consider, on the other hand, the rough clustering defined in Example 2.14. Then, this can equivalently be represented as the credal partition: $m_{x_1}(\{c_1\}) = m_{x_2}(\{c_1\}) = 1$, $m_{x_3}(\{c_1, c_2\}) = 1$, $m_{x_4}(\{c_2\}) = 1$, $m_{x_5}(\{c_2, c_3\}) = 1$.

However, several different criteria can be applied [24]: in the following example, we show that the decision-theoretic framework described in Section 2.1.1 can be used for this purpose.

Example 4.2. Consider the credal partition introduced in Example 2.7 and let $D = \{1, 2, 3\}$, where classification i corresponds to assigning an object to cluster c_i. Let U be the utility function described in Table 2, corresponding to the OWA operator with imprecision tolerance $\alpha = 0.75$. Suppose we select the lower expected utility criterion, then:

- The optimal act for object x_1 is 1, thus $x_1 \in l(c_1)$;
- The optimal act for object x_2 is 1, thus $x_2 \in l(c_1)$;
- The optimal act for object x_3 is $\{2, 3\}$, thus $x_3 \in u(c_2)$ and $x_3 \in u(c_3)$;
- The optimal act for object x_4 is 2, thus $x_4 \in l(c_2)$;
- The optimal act for object x_5 is D, thus $x_5 \in u(c_1)$, $x_5 \in u(c_2)$ and $x_5 \in u(c_3)$.

By contrast, the problem of finding a relationship between three-way clustering and evidential clustering, to our knowledge, has not been previously studied in the literature. As previously noted, the axioms for rough and three-way clustering are in general incompatible, and this complication obviously arises also in the case of evidential clustering. Nonetheless, in [11], it has been shown that three-way clustering can be represented in terms of rough clustering by adding a so-called *noise cluster*. Based on this relationship, a correspondence between three-way clustering and evidential clustering can also be found. Namely, let $C^* = C \cup \{c_\eta\}$, where c_η represents a noise cluster. If $\mathcal{C} = \{m_x\}_{x \in X}$ is a credal partition where every bba m_x is defined on C^*, then we say that $\mathcal{C}$ is an *extended* credal partition. Then, the following result holds:

Theorem 4.2. *Let $\mathcal{C} = \{m_x\}_{x \in X}$ be an extended credal partition satisfying the following conditions:*

1. $\forall x \in X, \exists A_x \neq \emptyset$ *s.t.* $m_x(A_x) = 1$ *and either* $A_x \subseteq C$, *or* $A_x = \{c, c_\eta\}$ *for some* $c \in C$
2. $\forall c \in C, \exists x \in X$ *s.t.* $m_x(\{c\}) = 1$

Then $\mathcal{C}$ corresponds to a unique three-way clustering. Conversely, if $\mathcal{T}$ is a three-way clustering, then there is a unique extended credal partition $\mathcal{C}$ corresponding to $\mathcal{T}$.

Proof. If $\mathcal{C}$ satisfies the two conditions, then we can construct a three-way clustering $\mathcal{T}_\mathcal{C}$ as follows. First, if $m_x(\{c\}) = 1$, set $x \in c_l$. Condition (2) guarantees that, for each cluster c, at least one such object exists. For any other object x, denote as A_x the set defined as in condition (1). Then, if $|A_x| > 1$, with $A_x \subseteq C$, set $x \in c_u \setminus c_l$ for all $c \in A_x$. Otherwise (i.e., $A_x = \{c, c_\eta\}$ for some $c \in C$), set $x \in c_u \setminus c_l$. The converse follows in a similar way. □

We note that the equivalence expressed in Theorem 4.2 and, more particularly, the transformation from a three-way clustering to the corresponding evidential clustering respects the semantics of belief function theory. Indeed, we can note that, for any given $c \in C$ and $x \in X$, it holds that $Bel_x(\{c\}) = 1$ iff $x \in c_l$, and, similarly, $Pl_x(\{c\}) = 1$ iff $x \in c_u$.

Example 4.3. Consider the three-way clustering defined in Example 2.15. It is easy to see that this can be equivalently represented as the following extended evidential clustering: $m_{x_1}(\{c_1\}) = m_{x_2}(\{c_1\}) = 1$, $m_{x_3}(\{c_1, c_\eta\}) = 1$, $m_{x_4}(\{c_2\}) = 1$, $m_{x_5}(\{c_3\}) = 1$.

In regard to methods for transforming a general evidential clustering to a three-way clustering, we note that the same techniques used for rough clustering cannot be directly used due to the differences between rough clustering and three-way clustering. In particular, in three-way clustering we need to guarantee that, for each cluster c, at least one object $x \in X$ belongs to the lower cluster c_l. Nevertheless, we propose a relatively simple transformation criterion based on the decision-theoretic approach introduced in Section 2.1.1.

Basically, we initially set the imprecision tolerance degree $\alpha = 1$: in this way all objects are assigned to the upper cluster of all clusters. Then, iteratively, we reduce the imprecision tolerance until we obtain an assignment that satisfies the constraints of three-way clustering: namely, we continue decreasing α as long as there is some cluster whose lower cluster is empty. This procedure is guaranteed to stop: in the worst case we reach $\alpha = 0$, in which case we obtain a hard clustering (which, indeed, is a special case of three-way clustering).

We note that, following this procedure, it may happen that even with $\alpha = 0$ some of the clusters end up being empty (i.e., the corresponding lower and upper clusters are both empty). Even though this situation is not problematic from a procedural perspective (i.e., the empty clusters can simply be removed), it may indicate that the original evidential clustering algorithm was not able to find a significant clustering structure. As a consequence, such a situation could suggest to apply again the clustering algorithm with different parameter settings (e.g., reducing the number of clusters).

Example 4.4. Consider the credal partition introduced in Example 2.7 and let $D = \{1, 2, 3\}$, where classification i corresponds to assigning an object to cluster c_i. We have shown in Example 4.2 that the value $\alpha = 0.75$ does not result in a three-way clustering, since the lower cluster of cluster c_3 is empty. Thus, we need to select a value of $\alpha < 0.75$. It can easily be shown that for $\alpha = 0.55$ we obtain the following assignment:

- The optimal act for object x_1 is 1, thus $x_1 \in c_l^1$;
- The optimal act for object x_2 is 1, thus $x_2 \in c_l^1$;
- The optimal act for object x_3 is $\{2, 3\}$, thus $x_3 \in c_u^2$ and $x_3 \in c_u^3$;
- The optimal act for object x_4 is 2, thus $x_4 \in c_l^2$;

– The optimal act for object x_5 is 3, thus $x_5 \in c_l^3$.

Thus, in this case, we obtain a three-way clustering since all lower clusters are non-empty. Furthermore, the obtained three-way clustering is not hard, since object x_3 does not belong to any lower cluster. It can be easily shown that for any $\alpha > 0.55$, object x_5 is assigned to the upper cluster of all clusters. Thus, the above described procedure would stop at the optimal value $\alpha = 0.55$.

Remarks and prospects Beyond the formal relationships shown above among evidential, rough and three-way clustering, the relative advantages and strengths of the different approaches have rarely been considered. Joshi and Lingras [44] showed, through a set of illustrative examples on simple toy datasets, that evidential clustering may be more effective at identifying outliers, due the increased flexibility of bbas. By contrast, the main advantage of rough clustering (and, by extension, three-way clustering) is its simplicity [29,71]: rough clustering techniques are usually more computationally efficient, and also more interpretable (e.g., for visualization). Also, rough clustering was found to be more resistant than evidential clustering to the so-called curse of dimensionality [44]. Nonetheless, these comments being based on small toy datasets, we believe that future work should be devoted at comparing rough, three-way and evidential clustering from an empirical perspective on real-world datasets, both in terms of performance, and in terms of the possible relationship between rough partitions induced by rough (resp., three-way) clustering algorithms and the rough partitions obtained as approximations to an evidential clustering. A particularly important problem, in this sense, regards the study and design of general evaluation criteria that can be used to compare different forms of soft clustering, such as those proposed in [11,33,35]. Evaluation criteria can be useful also for the purpose of transforming evidential clustering into rough (resp. three-way) clustering, and vice-versa: namely, given an evidential clustering one could select the rough (resp. three-way) clustering which is maximally similar to it. In particular, we believe that such criteria could be particularly useful for the case of three-way clustering: indeed, we remark that even though the procedure we proposed is conceptually simple, it can be computationally inefficient if implemented naively. While binary search or search heuristics can be used to make searching for the optimal α more efficient, we believe that further research should be devoted at exploring alternative transformation strategies.

4.2. Uncertainty in the conditions: belief reducts

The correspondence between generalized relation-based rough set models and belief functions [117,126] has been exploited in the literature as a foundation for studying the problem of reduct search (i.e., feature selection) in the presence of missing or set-valued data [105]. The starting point, in this setting, is the possibility to define belief and plausibility functions from a similarity relation-based rough approximation. The similarity relation S is used to take into account missing values and partially specified values. Indeed, given an information table such that the condition attributes can be set-valued, and a subset of attributes $B \subseteq Att$, a similarity relation S can be easily defined by declaring two objects $x, y \in X$ to be B-similar when, for each attribute in B, the set values for the two objects are compatible (i.e., they are not disjoint). Formally:

$$S_B = \{(x, y) \in X \times X : \forall a \in B,\ F(a, x) \cap F(a, y) \neq \emptyset\}. \tag{27}$$

Then, belief and plausibility reducts can be defined as follows.

Definition 4.1. Let $B \subseteq Att$ a set of attributes, and S_B the similarity relation defined by B as in Eq. (27). Let Bel_B, Pl_B be the corresponding belief and plausibility functions defined as in (17). Then, given two attribute subsets $B_1 \subseteq B_2 \subseteq Att$, B_1 is a *belief reduct* of B_2 if

1. For all similarity classes $[x]_S$, $Bel_{B_1}([x]_S) = Bel_{B_2}([x]_S)$;
2. A minimality condition holds: $\nexists C \subset B_1$ s.t. $Bel_C([x]_S) = Bel_{B_1}([x]_S)$.

Given two attribute subsets $B_1 \subseteq B_2 \subseteq Att$, B_1 is a *plausibility reduct* of B_2 if

1. For all similarity classes $[x]_R$, $Pl_{B_1}([x]_S) = Pl_{B_2}([x]_S)$;
2. A minimality condition holds: $\nexists C \subset B_1$ s.t. $Pl_C([x]_S) = Pl_{B_1}([x]_S)$.

The following result holds:

Theorem 4.3 *([105]). $B \subseteq Att$ is a reduct iff it is a belief reduct. Furthermore, if $B \subseteq Att$ is a reduct then it is a plausibility reduct.*

Interestingly, the converse of the second statement above in general does not hold. In case of a decision table, the following definition of reducts is given:

Definition 4.2. Given two attribute subsets $B_1 \subseteq B_2 \subseteq Att$, B_1 is a *relative belief reduct* of B_2 if

Table 5
Example of decision table with set-valued conditions.

Patient	Temperature	Pressure	Muscle Pain	Disease
p_1	{very high, high}	2	yes	YES
p_2	{normal, high}	1	no	NO
p_3	high	{1, 2}	yes	NO
p_4	high	2	{no, yes}	YES

1. For all decision classes $[x]_d$, $Bel_{B_1}([x]_d) = Bel_{B_2}([x]_d)$;
2. A minimality condition holds: $\nexists C \subset B_1$ s.t. $Bel_C([x]_d) = Bel_{B_1}([x]_d)$.

An attribute subset $B_1 \subseteq B_2 \subseteq Att$, A is a *relative plausibility reduct* of B_2 if

1. For all decision classes $[x]_d$, $Pl_{B_1}([x]_d) = Pl_{B_2}([x]_d)$;
2. A minimality condition holds: $\nexists C \subseteq B_1$ s.t. $Pl_C([x]_d) = Pl_{B_1}([x]_d)$.

Example 4.5. Consider the decision table given in Table 5. Define $B = \{Temperature, Headache\}$ and let the decision attribute be $d = \{Disease\}$.

Then the similarity relation determined by B is

$$S_B = \{(p_1, p_3), (p_1, p_4), (p_3, p_4), (p_2, p_3)\}^{sr},$$

where, given a relation $R \subseteq X \times X$, R^{sr} represents the symmetric and reflexive closure of R. The similarity classes determined by S_B are $[p_1]_B = \{p_1, p_3, p_4\}$, $[p_2]_B = \{p_2, p_3\}$, $[p_3]_B = \{p_1, p_2, p_3, p_4\}$ and $[p_4]_B = \{p_1, p_3, p_4\}$. By contrast, the similarity relation determined by Att is

$$S_{Att} = \{(p_1, p_3), (p_1, p_4), (p_3, p_4)\}^{sr}.$$

The similarity classes determined by S_{Att} are $[p_1]_{Att} = \{p_1, p_2, p_3\}$ and $[p_2]_{Att} = \{p_2\}$. The partition determined by d is $[p_1]_d = \{p_1, p_4\}$ and $[p_2]_b = \{p_2, p_3\}$. Thus, the similarity relation S_B determines the bba $m_B(\{p_1, p_3, p_4\}) = 1/2$, $m_B(\{p_2, p_3\}) = 1/4$ and $m_B(X) = 1/4$. By contrast, the equivalence relation given by Att determines the bba $m_{Att}(\{p_1, p_3, p_4\}) = 3/4$ and $m_{Att}(\{p_2\}) = 1/4$. Since it holds that $Bel_B(\{p_1, p_4\}) = Bel_{Att}(\{p_1, p_4\}) = 0$, $Bel_B(\{p_2, p_3\}) = Bel_{Att}(\{p_2, p_3\}) = 1/4$, we have that B is a *relative belief reduct*. Furthermore, since $Pl_B(\{p_1, p_4\}) = Pl_{Att}(\{p_1, p_4\}) = 3/4$, $Pl_B(\{p_2, p_3\}) = Pl_{Att}(\{p_2, p_3\}) = 1$, B is also a *relative plausibility reduct*.

In contrast to the case of reducts, for the case of relative reducts the following results hold:

Theorem 4.4. *If the decision table is consistent, then $B \subseteq Att$ is a relative reduct iff it is a relative belief reduct iff it is a relative plausibility reduct. If the decision table is not consistent, then $B \subseteq Att$ is a relative reduct iff it is a relative plausibility reduct.*

The results of [105] have been extended to other generalized rough set models. First, the same authors [109] considered the application of belief and plausibility reducts to random information systems, showing that Theorems 4.3 and 4.4 hold also in this setting. Furthermore, the extension to the case of continuous and interval-valued data has been widely studied [17,68,88], and it has been shown that the equivalence between belief and classical reducts holds also in these settings.

Considering different generalized relations, Du et al. [36] studied the case of dominance-based rough sets, in which the relation induced by the condition attributes is a dominance (i.e., reflexive and transitive) relation, showing that the above described methodology can also be applied to ordinal data. Similarly, Syau et al. [87] considered the extension to reflexive (but not necessarily symmetric) relations, in order to model two different types of missing values (*don't know* vs *doesn't exist*). While in all these models the definitions of belief and plausibility reducts are the same as in (4.1) and (4.2), it has not yet been shown whether Theorems 4.3 and 4.4 similarly hold. Finally, the definition of belief and plausibility reducts has also been extended to the fuzzy case by Yao et al. [116] and Zhang et al. [124]. Remarkably, in this latter setting it has been shown that the extension of Theorem 4.4 to fuzzy information tables and fuzzy approximation spaces holds.

Interestingly, a similar approach has also been proposed for the management of more general forms of uncertainty in the condition attributes. More in particular, Trabelsi et al. [91] studied the problem of reduct search where the uncertain condition attributes are expressed as bbas. In this case, which is inspired by the definition of tolerance relations used in UDTs (see Section 3.3), the similarity relation can be defined by declaring two objects to be B-similar if they have the same decision and, for each attribute in B the corresponding bbas for the two objects are not too distant. Formally, this can be defined as:

$$S_B = \{(x, y) \in X \times X : d(x) \neq d(y) \wedge \forall a \in B \subseteq Att, dist(m_x^a, m_y^a) < 1 - \theta\}, \tag{28}$$

where, for any object $x \in X$ and attribute $a \in Att$, m_x^a is the bba for x corresponding to attribute a, and the distance function is defined as in Eq. (23). We note that the obtained similarity relation is certainly symmetric and reflexive, but it is not, in general, guaranteed to be transitive.

Then, reduct search can be performed through standard techniques based on the discernibility matrix [91], or through belief and plausibility reducts. We note, however, that the theoretical properties of these latter types of reducts have not yet been studied in this setting.

Remarks and prospects Even though reduct search is a practical task, due to its relation with feature selection, all the above mentioned contributions had a primarily theoretical focus. Indeed, none of the mentioned articles investigates the performance and efficacy of algorithms for searching reducts in real applications. Therefore, future work should be devoted at assessing the performance (in terms of reduct size, as well as classification accuracy) of the different notions of evidence theory-based reducts. In particular, it would be interesting to assess the practical implications of Theorems 4.3 and 4.4: namely, since plausibility (resp., relative belief) reducts do not coincide with standard reducts in information (resp., decision) tables, the application of these definitions to real data should be further evaluated. Finally, it would be interesting to evaluate whether Theorems 4.3 and 4.4 hold also for the more general similarity-based approaches proposed in [17,36,87,88], as well as in the bba-based approach proposed in [91].

4.3. Uncertainty in the decision: decision rules in UDT

Feature selection and rule induction tasks have also been studied in the context of UDTs (see Section 3.3). In particular, following the general reduction-rule process, Trabelsi et al. [93] propose the simplification of an UDT and generation of belief decision rules. So, as usual, the first step is reduct generation. The authors of [93] specifically consider the positive region-based reduct model, which is defined in the standard way according to (11) by using the lower approximation of (25). The definition of a reduct is not changed, i.e., it is the same as in Definition 2.8. In particular, to search for the reducts of an UDT (and hence perform feature selection), the same authors provide a heuristic algorithm based on a generalized definition of the discernibility matrix [94]. In [95,96], they also propose a parallel algorithm for application to big data.

Example 4.6. Consider the UDT defined in Example 3.2. Obviously, $S_1 = \{Headache\}$ is not a reduct, since $POS_{S_1}(X, d) = \emptyset$, while $POS_{Att}(X, d) = \{p_1, p_2, p_3\}$. In the same way, $S_2 = \{Temperature\}$ is not a reduct, since $POS_{S_2}(X, d) = \{p_1, p_2\}$. Thus, the only reduct is the full set of features Att.

After performing feature reduction, decision rules can be induced. To do so, redundant objects and attribute values are eliminated. In particular, concerning objects, the bbas of objects that are in the same equivalence class (i.e., they have the same values for the reduced conditional attributes) are aggregated by means of the averaging combination rule. Decision rules can then be generated by reading directly in the simplified table the values of the condition attributes and the corresponding decision, in the form of the averaged bba on Val_D. A major limitation of this definition of rules, is that their accuracy and performance cannot be easily evaluated, as the decision is represented in the form of a bba. For this reason, Trabelsi et al. propose [92,93] to convert the bba for each rule into a probability distribution, by means of the pignistic transform (6): accuracy can then be evaluated as a weighted average or by selecting the single decision with maximal probability.

Example 4.7. Consider the UDT defined in Table 4. In Example 4.6, we showed that Att is the only reduct. Thus, decision rules are in the form:

$$\text{If } Temperature = \text{very high and } Headache = yes \text{ then}$$
$$m(A) = 0.8, m(B) = 0.05, m(\{A, B\}) = 0.15$$

$$\text{If } Temperature = high \text{ and } Headache = no \text{ then}$$
$$m(B) = 0.7, m(\{A, B\}) = 0.3$$

$$\text{If } Temperature = normal \text{ and } Headache = no \text{ then } m(NO) = 1$$

$$\text{If } Temperature = high \text{ and } Headache = yes \text{ then}$$
$$m(A) = 0.15, m(B) = 0.1, m(NO) = 0.5, m(\{NO, A\}) = 0.05, m(\{NO, A, B\}) = 0.2$$

Based on [92,93], the decisions of the previous rules can be transformed into single-valued decisions by applying the pignistic transform and then selecting the value with maximum probability. The resulting rules are

$$\text{If } Temperature = \text{very high and } Headache = yes \text{ then } A$$
$$\text{If } Temperature = high \text{ and } Headache = no \text{ then } B$$
$$\text{If } Temperature = normal \text{ and } Headache = no \text{ then } NO$$
$$\text{If } Temperature = high \text{ and } Headache = yes \text{ then } NO$$

As previously mentioned in Section 3.3, a different approach to the definition of reducts and rules in UDT is taken in [13], based on the notion of an entropy reduct [80] and the generalized risk minimization [42] principle. In this setting, given a set of attributes $B \subseteq Att$ the entropy of the reduced UDT is computed as

$$H(B) = \sum_{[x]_B} \frac{1}{|[x]_B|} \min_{P \in \mathcal{P}(m_{[x]_B})} H(P),$$

where $m_{[x]_B}$ is the bba determined by the equivalence class of x as defined in (26), $\mathcal{P}(m_{[x]_B})$ is the set of probability measures compatible with $m_{[x]_B}$ as defined in (5), and $H(P) = -\sum_p p \cdot \log(p)$ is the Shannon entropy. Thus, for each equivalence class $[x]_B$, a bba $m_{[x]_B}$ is obtained and the entropy for this bba is simply computed as the lowest possible entropy for all probability distributions that are compatible with it. Then, B is said to be an entropy reduct if $H(B) \leq H(Att)$ and the minimality condition (i.e., $\nexists D \subset B$ s.t. $H(D) \leq H(C)$) holds.

After the reduct search, decision rules can be induced: in contrast to the rule induction procedure studied in [93], the approach proposed in [13] directly provides rules with single-valued decision, by using the maximum plausibility criterion [24]. This approach to feature reduction was compared to state-of-the-art algorithms for feature selection on set-valued data, showing significant improvements on different datasets.

Remarks and prospects While the authors of [95] studied the performance of the discussed feature reduction and classification methods and reported promising empirical results, no comparison with other classifiers for evidential data [5,20,32,40,73] has so far been evaluated in the literature. Furthermore, similarly to the remarks in Section 3.3, also for the case of feature reduction and rule induction it would be interesting to evaluate the behavior and performance of different aggregation as well decision rules. In respect to this latter aspect, while the authors of [95] considered only the pignistic transform, other decision-making criteria such as mentioned in Section 4.1 or *disambiguation* methods [42] should also be evaluated. A final remark regards the relationships among the methods for feature selection proposed in [93] and [13]. While this has not been previously evaluated, it is clear that the two approaches are not equivalent when applied to set-valued data and may thus end up providing different results. Thus, future work should aim at studying possible conditions for the equivalence between the two approaches, as well as generalizations of the approach described in [12,13] to general bbas, and their comparison on real datasets.

5. Conclusion

In this article, we provided a survey of the literature on belief functions and rough sets. While our contribution is not intended to be systematic, we drew a map of the links and cross-fertilizations among these two different research communities, so as to illustrate the most relevant relationships, their interpretation and semantics, as well as to highlight open problems, prospects and directions for future research. To this end, we reviewed the known theoretical results relating the rough set and belief function theories, as well as their applications to knowledge representation and machine learning. We hope that this paper will stimulate further research and investigations at the cross-road of the two research communities. To this purpose, we proposed some particularly relevant open problems related to each covered application, both of a theoretical as well as of an empirical or practical nature. In particular, to summarize, we recall and highlight the following relevant open issues:

- While the mathematical picture connecting the basic notions within rough set theory (i.e., approximations) and belief function theory (i.e., belief and plausibility functions) has been widely studied, the interpretation of these connections is much less clear and should be further investigated. Furthermore, the connections between other relevant concepts in the two theories (e.g., granule refinements or joins of decision tables in rough set theory, Dempster combination rule or coarsenings in belief function theory) should be explored.
- While the theoretical results for Pawlak and generalized relation models have been applied in practical problems (either in knowledge Representation or Machine Learning), the results for more general rough set models (e.g., interval rough sets or general approximation spaces) have not yet seen practical applications. Therefore, it would be interesting to further explore the applications of these results.
- In this review we only provided a brief introduction to the extension of the mentioned relationships between rough set theory and belief function theory to the fuzzy case, due to space constraints. Nonetheless, we believe that more research should be devoted at further exploring this important topic, as well as its possible applications.
- In our summary of theoretical results, we focused on a static picture of the relationship between rough set theory and belief function theory: namely, we assumed the information/decision tables and corresponding belief/plausibility functions to be already given and fixed in time. Nonetheless, it could be interesting to study the relationships between these mathematical structures also in so-called *dynamic* or *incremental* settings, in which the available knowledge evolves and changes with time. Although this issue has been explored within rough set theory [6,18,54], we believe that further attention should be devoted at studying the connections between rough approximation and the corresponding belief/plausibility functions in the dynamic setting.

- The comparative performance and properties (e.g., with respect to interpretability or ease of visualization) of rough and evidential clustering algorithms should be further evaluated, by means of experiments on real-world benchmark datasets. To this end, a particularly interesting open problem regards the definition of appropriate clustering quality metrics that could be applied to evaluate different types of soft clustering [33].
- The practical performance of belief and plausibility reducts (see Section 4.2) has not yet been evaluated in experimental settings. Furthermore, several theoretical problems related to this approach remain open. These include finding theoretical conditions for equivalence between the two reduct definitions, the study of rule induction algorithms, as well as studying the properties of these definitions of reducts in the other generalized relation-based models that have been more recently considered in the literature [17,36,68,87,88,91].
- The relationships between two existing approaches for dealing with UDTs (see Sections 3.3 and 4.3) should be further studied, both in theoretical terms (e.g., when do the two approaches coincide? When they give different results?) and experimental ones. To this end, a particularly interesting open problem regards the extension of the approach proposed in [12,13] to general UDTs.
- Finally, while in this article we reviewed a broad and representative selection of possible applications of the links between rough sets and belief functions, obviously many others have been considered in the literature that could not be covered within the scope of this article. In particular, the connections between belief functions and rule induction in rough set theory seem to be particularly worthy of investigation: indeed, since rough set theory allows the induction of so-called non-deterministic rules [58] (i.e., rules in which either the consequent or antecedent are underspecified), it would be interesting to study the connections between quality measures for such rules (e.g., the support) and measures arising from belief function theory.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.K. Afridi, N. Azam, J. Yao, Variance based three-way clustering approaches for handling overlapping clustering, Int. J. Approx. Reason. 118 (2020) 47–63.
[2] M.K. Afridi, N. Azam, J. Yao, E. Alanazi, A three-way clustering approach for handling missing data using GTRS, Int. J. Approx. Reason. 98 (2018) 11–24.
[3] V. Antoine, B. Quost, M.H. Masson, T. Denoeux, CECM: constrained evidential c-means algorithm, Comput. Stat. Data Anal. 56 (2012) 894–914.
[4] T. Augustin, F.P. Coolen, G. De Cooman, M.C. Troffaes, Introduction to Imprecise Probabilities, John Wiley & Sons, 2014.
[5] N. Bahri, M.A.B. Tobji, B.B. Yaghlane, Rule-based classification for evidential data, in: International Conference on Scalable Uncertainty Management, Springer, 2020, pp. 234–241.
[6] R. Bello, R. Falcon, Rough sets in machine learning: a review, in: G. Wang, A. Skowron, Y. Yao, D. Ślezak, L. Polkowski (Eds.), Thriving Rough Sets: 10th Anniversary - Honoring Professor Zdzisław Pawlak's Life and Legacy & 35 Years of Rough Sets, Springer International Publishing, Cham, 2017, pp. 87–118.
[7] J.C. Bezdek, J. Keller, R. Krisnapuram, N. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Springer Science & Business Media, 1999.
[8] L. Biacino, Fuzzy subsethood and belief functions of fuzzy events, Fuzzy Sets Syst. 158 (2007) 38–49.
[9] A. Campagner, F. Cabitza, D. Ciucci, Three-way decision for handling uncertainty in machine learning: a narrative review, in: International Joint Conference on Rough Sets, Springer, 2020, pp. 137–152.
[10] A. Campagner, F. Cabitza, D. Ciucci, The three-way-in and three-way-out framework to treat and exploit ambiguity in data, Int. J. Approx. Reason. 119 (2020) 292–312.
[11] A. Campagner, D. Ciucci, Orthopartitions and soft clustering: soft mutual information measures for clustering validation, Knowl.-Based Syst. 180 (2019) 51–61.
[12] A. Campagner, D. Ciucci, Feature selection and disambiguation in learning from fuzzy labels using rough sets, in: International Joint Conference on Rough Sets, Springer, 2021, pp. 164–179.
[13] A. Campagner, D. Ciucci, E. Hüllermeier, Rough set-based feature selection for weakly labeled data, Int. J. Approx. Reason. 136 (2021) 150–167.
[14] G. Cattaneo, Abstract approximation spaces for rough theories, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, Physica-Verlag, 1998, pp. 59–98.
[15] G. Cattaneo, D. Ciucci, Investigation about time monotonicity of similarity and preclusive rough approximations in incomplete information systems, in: International Conference on Rough Sets and Current Trends in Computing, Springer, 2004, pp. 38–48.
[16] M.K. Chakraborty, On some issues in the foundation of rough sets: the problem of definition, Fundam. Inform. 148 (2016) 123–132.
[17] D. Chen, W. Li, X. Zhang, S. Kwong, Evidence-theory-based numerical algorithms of attribute reduction with neighborhood-covering rough sets, Int. J. Approx. Reason. 55 (2014) 908–923.
[18] D. Ciucci, Temporal dynamics in information tables, Fundam. Inform. 115 (2012) 57–74.
[19] B.R. Cobb, P.P. Shenoy, On the plausibility transformation method for translating belief function models to probability models, Int. J. Approx. Reason. 41 (2006) 314–330.
[20] E. Côme, L. Oukhellou, T. Denoeux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognit. 42 (2009) 334–348.
[21] C. Cornelis, M. De Cock, A.M. Radzikowska, Fuzzy rough sets: from theory into practice, in: Handbook of Granular Computing, 2008, pp. 533–552.
[22] A. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38 (1967) 325–339.
[23] T. Denœux, Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence, Artif. Intell. 172 (2008) 234–264.
[24] T. Denoeux, Decision-making with belief functions: a review, Int. J. Approx. Reason. 109 (2019) 87–110.
[25] T. Denoeux, Calibrated model-based evidential clustering using bootstrapping, Inf. Sci. 528 (2020) 17–45.
[26] T. Denoeux, Belief functions induced by random fuzzy sets: a general framework for representing uncertain and fuzzy evidence, Fuzzy Sets Syst. 424 (2021) 63–91.

[27] T. Denoeux, NN-EVCLUS: neural network-based evidential clustering, Inf. Sci. 572 (2021) 297–330.
[28] T. Denœux, D. Dubois, H. Prade, Representations of uncertainty in ai: beyond probability and possibility, in: A Guided Tour of Artificial Intelligence Research, Springer, 2020, pp. 119–150.
[29] T. Denoeux, O. Kanjanatarakul, Beyond fuzzy, possibilistic and rough: an investigation of belief functions in clustering, in: M.B. Ferraro, P. Giordani, B. Vantaggi, M. Gagolewski, M.Á. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Soft Methods for Data Science, SMPS 2016, Rome, Italy, 12-14 September, 2016, Springer, 2016, pp. 157–164.
[30] T. Denoeux, O. Kanjanatarakul, Evidential clustering: a review, in: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer, 2016, pp. 24–35.
[31] T. Denœux, O. Kanjanatarakul, S. Sriboonchitta, EK-NNclus: a clustering procedure based on the evidential k-nearest neighbor rule, Knowl.-Based Syst. 88 (2015) 57–69.
[32] T. Denoeux, O. Kanjanatarakul, S. Sriboonchitta, A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning, Int. J. Approx. Reason. 113 (2019) 287–302.
[33] T. Denœux, S. Li, S. Sriboonchitta, Evaluating and comparing soft partitions: an approach based on Dempster-Shafer theory, IEEE Trans. Fuzzy Syst. 26 (2017) 1231–1244.
[34] T. Denœux, M.H. Masson, EVCLUS: evidential clustering of proximity data, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 34 (2004) 95–109.
[35] M.R. Depaolini, D. Ciucci, S. Calegari, M. Dominoni, External indices for rough clustering, in: International Joint Conference on Rough Sets, Springer, 2018, pp. 378–391.
[36] W.S. Du, B.Q. Hu, Attribute reduction in ordered decision tables via evidence theory, Inf. Sci. 364 (2016) 91–110.
[37] D. Dubois, H. Prade, Representation and combination of uncertainty with belief functions and possibility measures, Comput. Intell. 4 (1988) 244–264.
[38] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–209.
[39] T. Feng, S.P. Zhang, J.S. Mi, The reduction and fusion of fuzzy covering systems based on the evidence theory, Int. J. Approx. Reason. 53 (2012) 87–103.
[40] X. Geng, Y. Liang, L. Jiao, ARC-SL: association rule-based classification with soft labels, Knowl.-Based Syst. 225 (2021) 107116.
[41] Q. Hu, J.S. Mi, Representation of multigranularity fuzzy rough sets and corresponding belief structure, Comput. Eng. Appl. 53 (2017) 51–54.
[42] E. Hüllermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (2014) 1519–1534.
[43] X. Jia, L. Shang, B. Zhou, Y. Yao, Generalized attribute reduct in rough set theory, Knowl.-Based Syst. 91 (2016) 204–218.
[44] M. Joshi, P. Lingras, Evidential clustering or rough clustering: the choice is yours, in: T. Li, H.S. Nguyen, G. Wang, J.W. Grzymala-Busse, R. Janicki, A.E. Hassanien, H. Yu (Eds.), Rough Sets and Knowledge Technology - 7th International Conference, RSKT 2012, Chengdu, China, August 17-20, 2012. Proceedings, 2012, pp. 123–128.
[45] A.L. Jousselme, P. Maupin, Distances in evidence theory: comprehensive survey and generalizations, Int. J. Approx. Reason. 53 (2012) 118–145.
[46] M.A. Klopotek, S.T. Wierzchon, A new qualitative rough-set approach to modeling belief functions, Lect. Notes Comput. Sci. 1424 (1998) 0346.
[47] M.A. Kłopotek, S.T. Wierzchoń, Empirical models for the Dempster-Shafer theory, in: Belief Functions in Business Decisions, Springer, 2002, pp. 62–112.
[48] M. Kondo, On the structure of generalized rough sets, Inf. Sci. 176 (2006) 589–600.
[49] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (1993) 98–110.
[50] M. Kryszkiewicz, Rough set approach to incomplete information systems, Inf. Sci. 112 (1998) 39–49.
[51] F. Li, S. Li, T. Denoeux, k-CEVCLUS: constrained evidential clustering of large dissimilarity data, Knowl.-Based Syst. 142 (2018) 29–44.
[52] P. Lingras, Evolutionary rough k-means clustering, in: P. Wen, Y. Li, L. Polkowski, Y. Yao, S. Tsumoto, G. Wang (Eds.), Rough Sets and Knowledge Technology, Springer, Berlin, Heidelberg, 2009, pp. 68–75.
[53] P. Lingras, G. Peters, Applying rough set concepts to clustering, in: G. Peters, P. Lingras, D. Slezak, Y. Yao (Eds.), Rough Sets: Selected Methods and Applications in Management and Engineering, Springer, London, 2012, pp. 23–37.
[54] Z. Liu, J. Dezert, G. Mercier, Q. Pan, Belief c-means: an extension of fuzzy c-means algorithm in belief functions framework, Pattern Recognit. Lett. 33 (2012) 291–300.
[55] J. Lu, D.Y. Li, Y.H. Zhai, H.X. Bai, Belief and plausibility functions of type-2 fuzzy rough sets, Int. J. Approx. Reason. 105 (2019) 194–216.
[56] C. Lucas, B.N. Araabi, Generalization of the Dempster-Shafer theory: a fuzzy-valued measure, IEEE Trans. Fuzzy Syst. 7 (1999) 255–270.
[57] L. Ma, T. Denoeux, Partial classification in the belief function framework, Knowl.-Based Syst. 214 (2021) 106742.
[58] B. Marszał-Paszek, P. Paszek, Classifiers based on nondeterministic decision rules, in: Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam, Springer, 2013, pp. 445–454.
[59] M.H. Masson, T. Denoeux, ECM: an evidential version of the fuzzy c-means algorithm, Pattern Recognit. 41 (2008) 1384–1397.
[60] M.H. Masson, T. Denoeux, RECM: relational evidential c-means algorithm, Pattern Recognit. Lett. 30 (2009) 1015–1026.
[61] C.K. Murphy, Combining belief functions when evidence conflicts, Decis. Support Syst. 29 (2000) 1–9.
[62] V.P. Murugesan, P. Murugesan, A new initialization and performance measure for the rough k-means clustering, Soft Comput. 24 (2020) 11605–11619.
[63] H.S. Nguyen, D. Ślezak, Approximate reducts and association rules, in: International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, Springer, 1999, pp. 137–145.
[64] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[65] Z. Pawlak, Rough probability, Bull. Pol. Acad. Sci., Math. 32 (1984) 607–615.
[66] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, vol. 9, Springer Science & Business Media, 1991.
[67] Z. Pawlak, J.W. Grzymala-Busse, R. Slowinski, W. Ziarko, Rough sets, Commun. ACM 38 (1995) 88–95.
[68] Y. Peng, Q. Zhang, Feature selection for interval-valued data based on ds evidence theory, IEEE Access 9 (2021) 122754–122765.
[69] G. Peters, Rough clustering utilizing the principle of indifference, Inf. Sci. 277 (2014) 358–374.
[70] G. Peters, Is there any need for rough clustering?, Pattern Recognit. Lett. 53 (2015) 31–37.
[71] G. Peters, F. Crespo, P. Lingras, R. Weber, Soft clustering-fuzzy and rough approaches and their extensions and derivatives, Int. J. Approx. Reason. 54 (2013) 307–322.
[72] G. Peters, M. Lampart, R. Weber, Evolutionary rough k-medoid clustering, in: J.F. Peters, A. Skowron (Eds.), Transactions on Rough Sets VIII, Springer, Berlin, Heidelberg, 2008, pp. 289–306.
[73] B. Quost, T. Denoeux, S. Li, Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression, Adv. Data Anal. Classif. 11 (2017) 659–690.
[74] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[75] C. Shenoy, P.P. Shenoy, Modeling financial portfolios using belief functions, in: Belief Functions in Business Decisions, Springer, 2002, pp. 316–332.
[76] M. Sivaguru, M. Punniyamoorthy, Performance-enhanced rough k-means clustering algorithm, Soft Comput. 25 (2021) 1595–1616.
[77] A. Skowron, The rough sets theory and evidence theory, Fundam. Inform. 13 (1990) 245–262.
[78] A. Skowron, J.W. Grzymała-Busse, From rough set theory to evidence theory, in: R. Yager, M. Fedrizzi, J. Kacprzyk (Eds.), Advances in the Dempster-Shafer Theory of Evidence, John Wiley & Sons, 1994, pp. 193–236.
[79] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: Intelligent Decision Support, Springer, 1992, pp. 331–362.
[80] D. Slezak, Approximate entropy reducts, Fundam. Inform. 53 (2002) 365–390.

[81] D. Slezak, S. Dutta, Dynamic and discernibility characteristics of different attribute reduction criteria, in: H.S. Nguyen, Q. Ha, T. Li, M. Przybyla-Kasperek (Eds.), Rough Sets - International Joint Conference, IJCRS 2018, Quy Nhon, Vietnam, August 20-24, 2018, Proceedings, Springer, 2018, pp. 628–643.
[82] R. Slowinski, S. Greco, B. Matarazzo, Rough set methodology for decision aiding, in: J. Kacprzyk, W. Pedrycz (Eds.), Springer Handbook of Computational Intelligence, Springer, 2015, pp. 349–370.
[83] P. Smets, The combination of evidence in the transferable belief model, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 447–458.
[84] P. Smets, Decision making in the tbm: the necessity of the pignistic transformation, Int. J. Approx. Reason. 38 (2005) 133–147.
[85] P. Smets, R. Kennes, The transferable belief model, Artif. Intell. 66 (1994) 191–234.
[86] J. Stefanowski, On rough set based approaches to induction of decision rules, in: Rough Sets in Knowledge Discovery, 1998, pp. 500–529.
[87] Y.R. Syau, C.J. Liau, E.B. Lin, On variable precision generalized rough sets and incomplete decision tables, Fundam. Inform. 179 (2021) 75–92.
[88] A. Tan, W. Wu, J. Li, G. Lin, Evidence-theory-based numerical characterization of multigranulation rough sets in incomplete information systems, Fuzzy Sets Syst. 294 (2016) 18–35.
[89] A. Tan, W. Wu, Y. Tao, A unified framework for characterizing rough sets with evidence theory in various approximation spaces, Inf. Sci. 454–455 (2018) 144–160.
[90] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: a review, Appl. Soft Comput. 9 (2009) 1–12.
[91] A. Trabelsi, Z. Elouedi, E. Lefevre, Ensemble enhanced evidential k-nn classifier through rough set reducts, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2018, pp. 383–394.
[92] S. Trabelsi, Z. Elouedi, P. Lingras, Belief rough set classifier, in: Canadian Conference on Artificial Intelligence, Springer, 2009, pp. 257–261.
[93] S. Trabelsi, Z. Elouedi, P. Lingras, Classification systems based on rough sets under the belief function framework, Int. J. Approx. Reason. 52 (2011) 1409–1432.
[94] S. Trabelsi, Z. Elouedi, P. Lingras, Heuristic for attribute selection using belief discernibility matrix, in: T. Li, H.S. Nguyen, G. Wang, J.W. Grzymala-Busse, R. Janicki, A.E. Hassanien, H. Yu (Eds.), Rough Sets and Knowledge Technology - 7th International Conference, RSKT 2012, Chengdu, China, August 17-20, 2012. Proceedings, 2012, pp. 129–138.
[95] S. Trabelsi, Z. Elouedi, P. Lingras, Belief discernibility matrix and function for incremental or large data, in: D. Ciucci, M. Inuiguchi, Y. Yao, D. Slezak, G. Wang (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing - 14th International Conference, RSFDGrC 2013, Halifax, NS, Canada, October 11-14, 2013. Proceedings, 2013, pp. 67–76.
[96] S. Trabelsi, Z. Elouedi, P. Lingras, Exhaustive search with belief discernibility matrix and function, in: Canadian Conference on Artificial Intelligence, Springer, 2013, pp. 162–173.
[97] S. Ubukata, A. Notsu, K. Honda, Objective function-based rough membership c-means clustering, Inf. Sci. 548 (2021) 479–496.
[98] C. Umans, On the complexity and inapproximability of shortest implicant problems, in: Automata, Languages and Programming, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999, pp. 687–696.
[99] D. Vanderpooten, R. Slowinski, Similarity relation as a basis for rough approximations, in: Advances in Machine Intelligence and Soft Computing, vol. 4, 1997, pp. 17–33.
[100] S. Vluymans, L. D'eer, Y. Saeys, C. Cornelis, Applications of fuzzy rough set theory in machine learning: a survey, Fundam. Inform. 142 (2015) 53–86.
[101] F. Voorbraak, A computationally efficient approximation of Dempster-Shafer theory, Int. J. Man-Mach. Stud. 30 (1989) 525–536.
[102] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, 1991.
[103] P. Wang, Y. Yao, CE3: a three-way clustering method based on mathematical morphology, Knowl.-Based Syst. 155 (2018) 54–65.
[104] S.K.M. Wong, L. Wang, Y.Y. Yao, On modeling uncertainty with interval structures, Comput. Intell. 11 (1995) 406–426.
[105] W. Wu, Attribute reduction based on evidence theory in incomplete decision systems, Inf. Sci. 178 (2008) 1355–1371.
[106] W. Wu, Y. Leung, W. Zhang, Connections between rough set theory and Dempster-Shafer theory of evidence, Int. J. Gen. Syst. 31 (2002) 405–430.
[107] W.Z. Wu, Y. Leung, J.S. Mi, On characterizations of (i, t)-fuzzy rough approximation operators, Fuzzy Sets Syst. 154 (2005) 76–102.
[108] W.Z. Wu, Y. Leung, J.S. Mi, On generalized fuzzy belief functions in infinite spaces, IEEE Trans. Fuzzy Syst. 17 (2009) 385–397.
[109] W.Z. Wu, M. Zhang, H.Z. Li, J.S. Mi, Knowledge reduction in random information systems via Dempster-Shafer theory of evidence, Inf. Sci. 174 (2005) 143–164.
[110] R.R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, IEEE Trans. Syst. Man Cybern. 18 (1988) 183–190.
[111] J. Yao, D. Ciucci, Y. Zhang, Generalized rough sets, in: Springer Handbook of Computational Intelligence, 2015, pp. 413–424.
[112] Y. Yao, Generalized rough set models, in: Rough Sets in Knowledge Discovery, vol. 1, 1998, pp. 286–318.
[113] Y. Yao, T.Y. Lin, Generalization of rough sets using modal logics, Intell. Autom. Soft Comput. 2 (1996) 103–119.
[114] Y. Yao, P. Lingras, R. Wang, D. Miao, Interval set cluster analysis: a re-formulation, in: H. Sakai, M.K. Chakraborty, A.E. Hassanien, D. Slezak, W. Zhu (Eds.), Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Berlin, Heidelberg, 2009, pp. 398–405.
[115] Y. Yao, B. Yao, Covering based rough set approximations, Inf. Sci. 200 (2012) 91–107.
[116] Y.Q. Yao, J.S. Mi, Z.J. Li, Attribute reduction based on generalized fuzzy evidence theory in fuzzy decision systems, Fuzzy Sets Syst. 170 (2011) 64–75.
[117] Y.Y. Yao, P. Lingras, Interpretation of belief functions in the theory of rough sets, Inf. Sci. 104 (1998) 81–106.
[118] H. Yu, A framework of three-way cluster analysis, in: L. Polkowski, Y. Yao, P. Artiemjew, D. Ciucci, D. Liu, D. Slezak, B. Zielosko (Eds.), Rough Sets - Proceedings IJCRS 2017, Springer International Publishing, 2017, pp. 300–312.
[119] H. Yu, Z. Chang, G. Wang, X. Chen, An efficient three-way clustering algorithm based on gravitational search, Int. J. Mach. Learn. Cybern. 11 (2020) 1003–1016.
[120] H. Yu, L. Chen, J. Yao, A three-way density peak clustering method based on evidence theory, Knowl.-Based Syst. 211 (2021) 106532.
[121] H. Yu, L. Chen, J. Yao, X. Wang, A three-way clustering method based on an improved DBSCAN algorithm, Physica A 535 (2019) 122289.
[122] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[123] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets Syst. 100 (1999) 9–34.
[124] J. Zhang, X. Liu, Fuzzy belief measure in random fuzzy information systems and its application to knowledge reduction, Neural Comput. Appl. 22 (2013) 1419–1431.
[125] W.X. Zhang, J.S. Mi, W.Z. Wu, Approaches to knowledge reductions in inconsistent systems, Int. J. Intell. Syst. Appl. 18 (2003) 989–1000.
[126] Y.L. Zhang, C.Q. Li, Relationships between relation-based rough sets and belief structures, Int. J. Approx. Reason. 127 (2020) 83–98.
[127] Z.W. Zhang, Z. Liu, A. Martin, Z.G. Liu, K. Zhou, Dynamic evidential clustering algorithm, Knowl.-Based Syst. 213 (2021) 106643.
[128] W. Zhu, Generalized rough sets based on relations, Inf. Sci. 177 (2007) 4997–5011.