Feature Reduction in Superset Learning Using Rough Sets and Evidence Theory

Andrea Campagner[1], Davide Ciucci[1](✉), and Eyke Hüllermeier[2]

[1] Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano–Bicocca, viale Sarca 336, 20126 Milan, Italy
davide.ciucci@unimib.it

[2] Department of Computer Science, Paderborn University, Paderborn, Germany

Abstract. Supervised learning is an important branch of machine learning (ML), which requires a complete annotation (labeling) of the involved training data. This assumption, which may constitute a severe bottleneck in the practical use of ML, is relaxed in weakly supervised learning. In this ML paradigm, training instances are not necessarily precisely labeled. Instead, annotations are allowed to be imprecise or partial. In the setting of superset learning, instances are assumed to be labeled with a set of *possible* annotations, which is assumed to contain the correct one. In this article, we study the application of *rough set theory* in the setting of superset learning. In particular, we consider the problem of feature reduction as a mean for *data disambiguation*, i.e., for the purpose of figuring out the most plausible precise instantiation of the imprecise training data. To this end, we define appropriate generalizations of decision tables and reducts, using information-theoretic techniques based on evidence theory. Moreover, we analyze the complexity of the associated computational problems.

Keywords: Feature selection · Superset learning · Rough sets · Evidence theory

1 Introduction

In recent years, the increased availability of data has fostered the interest in machine learning (ML) and knowledge discovery, in particular in *supervised learning* methodologies. These require each training instance to be annotated with a target value (a discrete label in classification, or a real number in regression). The annotation task is a fundamental component of the ML pipeline, and often a bottleneck in terms of cost. Indeed, the high costs caused by the standard annotation process, which may require the involvement of domain experts, have triggered the development of alternative annotation protocols, such as those based on *crowdsourcing* [4] or (semi-)*automated annotation* [12].

A different approach, which has attracted increasing attention in the recent years, is the combination of supervised and unsupervised learning techniques,

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1237, pp. 471–484, 2020.
https://doi.org/10.1007/978-3-030-50146-4_35

sometimes referred to as *weakly supervised learning* [30]. In this setting, training instances are not necessarily labeled precisely. Instead, annotations are allowed to be imprecise or partial.

A specific variant of weakly supervised learning is the setting of *superset learning* [9,16,18], where an instance x is annotated with a set S of (precise) candidate labels that are deemed *possible*. In other words, the label of x cannot be determined precisely, but is known to be an element of S. For example, an image could be tagged with $\{\text{horse}, \text{pony}, \text{zebra}\}$, suggesting that the animal shown on the picture is one of these three, though it is not exactly known which of them. Superset learning has been widely investigated under the classification perspective [10,15], that is, with the goal of training a predictive model that is able to correctly classify new instances, despite the weak training information. Nevertheless, the task of *feature selection* [6], which is of critical importance for machine learning in general, has not received much attention so far.

In this article, we study the application of *rough set theory* in the setting of superset learning. In particular, we consider the problem of feature reduction as a mean for *data disambiguation*, i.e., for the purpose of figuring out the most plausible precise instantiation of the imprecise training data. Broadly speaking, the idea is as follows: An instantiation that can be explained with a simple model, i.e., a model that uses only a small subset of features, is more plausible than an instantiation that requires a complex model. To this end, we will define appropriate generalizations of decision tables and reducts, using information-theoretic techniques based on evidence theory. Moreover, we analyze the complexity of the associated computational problems.

2 Background

In this section, we recall basic notions of rough set theory (RST) and evidence theory, which will be used in the main part of the article.

2.1 Rough Set Theory

Rough set theory has been proposed by Pawlak [19] as a framework for representing and managing uncertain data, and has since been widely applied for various problems in the ML domain (see [2] for a recent overview and survey). We briefly recall the main notions of RST, especially regarding its applications to feature reduction.

A decision table (DT) is a triple $DT = \langle U, Att, t\rangle$ such that U is a universe of objects and Att is a set of *attributes* employed to represent objects in U. Formally, each attribute $a \in Att$ is a function $a : U \to V_a$, where V_a is the domain of values of a. Moreover, $t \notin Att$ is a distinguished *decision* attribute, which represents the target decision (also labeling or annotation) associated with each object in the universe. We say that DT is *inconsistent* if the following holds: $\exists x_1, x_2 \in U, \forall a \in Att, a(x_1) = a(x_2)$ and $t(x_1) \neq t(x_2)$.

Given $B \subseteq Att$ we can define the *indiscernibility partition* with respect to B as $\pi_B = \{[x]_B \subset U \mid \forall x' \in [x]_B, \forall a \in B,\ a(x') = a(x)\}$. We say that $B \subseteq Att$ is a *decision reduct* for DT if $\pi_B \leq \pi_t$ (where the order $\leq$ is the refinement order for partitions, that is, π_t is a coarsening of π_B) and there is no $C \subsetneq B$ such that $\pi_C \leq \pi_t$. Then, evidently, a reduct of a decision table DT represents a set of non-redundant and necessary features to represent the information in DT. We say that a reduct R is *minimal* if it is among the smallest (with respect to cardinality) reducts.

Given $B \subseteq Att$ and a set $S \subseteq U$, a *rough approximation* of S (with respect to B) is defined as the pair $B(S) = \langle l_B(S), u_B(S)\rangle$, where $l_B(S) = \bigcup\{[x]_B \mid [x]_B \subseteq S\}$ is the *lower approximation* of S, and $u_B(s) = \bigcup\{[x]_B \mid [x]_B \cap S \neq \emptyset\}$ is the corresponding *upper approximation.*

Finally, given $B \subseteq Att$, the *generalized decision* with respect to B for an object $x \in U$ is defined as $\delta_B(x) = \{t(x') \mid x' \in [x]_B\}$. Notably, if DT is not inconsistent and B is a reduct, then $\delta_B(x) = t(x)$ for all $x \in U$.

We notice that in the RST literature, there exist several definitions of reduct. We refer the reader to [25] for an overview of such a list and a study of their dependencies. We further notice that, given a decision table, the problem of finding the minimal reduct is in general Σ_2^P-complete (by reduction to the *Shortest Implicant* problem [28]), while the problem of finding a reduct is in general NP-complete [23]. We recall that Σ_2^P is the complexity class defined by problems that can be verified in polynomial time given access to an oracle for an NP-complete problem [1].

2.2 Evidence Theory

Evidence theory (ET), also known as Dempster-Shafer theory or belief function theory, has originally been introduced by Dempster in [5] and subsequently formalized by Shafer in [21] as a generalization of probability theory (although this interpretation has been disputed [20]). The starting point is a *frame of discernment* X, which represents all possible states of a system under study, together with a *basic belief assignment* (bba) $m : 2^X \rightarrow [0,1]$, such that $m(\emptyset) = 0$ and $\sum_{A \in 2^X} m(A) = 1$. From this bba, a pair of functions, called respectively *belief* and *plausibility*, can be defined as follows:

$$Bel_m(A) = \sum_{B : B \subseteq A} m(B) \tag{1}$$

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

As can be seen from these definitions, there is a clear correspondence between belief functions (resp., plausibility functions) and lower approximations (resp., upper approximations) in RST; we refer the reader to [29] for further connections between the two theories.

Starting from a bba, a probability distribution, called *pignistic probability*, can be obtained [26]:

$$P^m_{Bet}(x) = \sum_{A:x\in A} \frac{m(A)}{|A|} \tag{3}$$

Finally, we recall that appropriate generalizations of information-theoretic concepts [22], specifically the concept of *entropy* (which was also proposed to generalize the definition of reducts in RST [24]), have been defined for evidence theory. Most relevantly, we recall the definition of *aggregate uncertainty* [7]

$$AU(m) = \max_{p\in\mathcal{P}(m)} H(p), \tag{4}$$

where $H(p) = -\sum_{x\in X} p(x) log_2 p(x)$ is the Shannon entropy of p and $\mathcal{P}(m)$ the set of probability distributions p such that $Bel_m \leq p \leq Pl_m$; and the definition of *normalized pignistic entropy* (see [13] for the un-normalized definition)

$$H_{Bet}(m) = \frac{H(P^m_{Bet})}{H(\hat{p}_m)}, \tag{5}$$

where $\hat{p}_m$ is the probability distribution that is uniform on the support of $P^m_{Bet}(x)$, i.e., on the set of elements $\{x \mid P^m_{Bet}(x) > 0\}$.

3 Superset Decision Tables and Reducts

In this section, we extend some key concepts of rough set theory to the setting of superset learning.

3.1 Superset Decision Tables

In superset learning, each object $x \in U$ is not associated with a single annotation $t(x) \in V_t$, but with a set S of candidate annotations, one of which is assumed to be the true annotation associated with x. In order to model this idea in terms of RST, we generalize the definition of a decision table.

Definition 1. *A* superset decision table *(SDT) is a tuple* $SDT = \langle U, Att, t, d\rangle$, *where* $\langle U, Att, t\rangle$ *is a decision table, i.e.:*

- *U is a universe of objects of interest;*
- *Att is a set of attributes (or features);*
- *t is the decision attribute (whose value, in general, is not known);*

and d, with $\{d\} \cap Att = \emptyset$, is a set-valued decision attribute, that is, $d : U \to \mathcal{P}(V_t)$ such that the superset property *holds: For all $x \in U$, the real decision $t(x)$ associated with x is in $d(x)$.*

The intuitive meaning of the set-valued information d is that, if $|d(x)| > 1$ for some $x \in U$, then the real decision associated with x (i.e. $t(x)$) is not known precisely, but is known to be in $d(x)$. Notice that Definition 1 is a proper generalization of decision tables: if $|d(x)| = 1$ for all $x \in U$, then we have a standard decision table.

Remark 1. In Definition 1, a set-valued decision attribute is modelled as a function $d : U \to \mathcal{P}(V_t)$. While this mapping is formally well-defined for a concrete decision table, let us mention that, strictly speaking, there is no functional dependency between x and $d(x)$. In fact, $d(x)$ is not considered as a property of x, but rather represents *information* about a property of x, namely the underlying decision attribute $t(x)$. As such, it reflects the epistemic state of the decision maker.

Definition 2. *An* instantiation *of an SDT $\langle U, Att, t, d\rangle$ is a standard DT $\langle U, Att, t'\rangle$ such that $t'(x) \in d(x)$ for all $x \in U$. The set of instantiations of SDT is denoted $\mathcal{I}(SDT)$.*

Based on the notion of SDT, we can generalize the notion of inconsistency.

Definition 3. *Let $B \subset Att$, then SDT is B-inconsistent if*

$$\exists x_1, x_2 \in U, \forall a \in B, a(x_1) = a(x_2) \text{ and } d(x_1) \cap d(x_2) = \emptyset. \tag{6}$$

We call such a pair x_1, x_2 inconsistent, otherwise it is consistent.

Thus, inconsistency implies the existence of (at least) two indiscernible objects with non-overlapping superset decisions. We say that an instantiation I is *consistent with a SDT S* (short, is consistent) if the following holds for all x_1, x_2: if x_1, x_2 are consistent in S, then they are also consistent in I.

3.2 Superset Reducts

Learning from superset data is closely connected to the idea of *data disambiguation* in the sense of figuring out the most plausible instantiation of the set-valued training data [8,11]. But what makes one instantiation more plausible than another one? One approach originally proposed in [9] refers to the principle of simplicity in the spirit of *Occam's razor* (which can be given a theoretical justification in terms of *Kolmogorov complexity* [14]): An instantiation that can be explained by a simple model is more plausible than an instantiation that requires a complex model. In the context of RST-based data analysis, a natural measure of model complexity is the size of the reduct. This leads us to the following definition.

Definition 4. *A set of attributes $R \subseteq Att$ is a* superset reduct *if there exists a consistent instantiation $\mathcal{I} = \langle U, Att, t\rangle$ such that R is a reduct for $\mathcal{I}$. We denote with RED_{super} the set of superset reducts. The* minimum description length (MDL) instantiation *is one of the consistent instantiations of SDT that admit a reduct of minimum size compared to all the reducts of all possible consistent instantiations. We will call the corresponding reduct* MDL reduct*.*

First of all, we briefly comment on the fact that the definition of MDL reduct generalizes the standard definition of (minimal) reduct. Indeed, in a classical decision table, there is only one possible instantiation, hence the MDL reduct is

Algorithm 1. The brute force algorithm for finding MDL reducts of a superset decision table S.

```
procedure BRUTE-FORCE-MDL-REDUCT($S$: superset decision table)
    $reds \leftarrow \emptyset$
    $l \leftarrow \infty$
    $ists \leftarrow enumerate\text{-}instantiations(S)$
    for all $i \in ists$ do
        $tmp\text{-}reds \leftarrow find\text{-}shortest\text{-}reducts(i)$
        $len \leftarrow |red|$ where $red \in tmp\text{-}reds$
        if $len < l$ then
            $reds \leftarrow tmp\text{-}reds$
            $l \leftarrow len$
        else if $len = l$ then
            $reds \leftarrow reds \cup tmp\text{-}reds$
        end if
    end for
    return $reds$                                  ▷ The MDL reducts for S
end procedure
```

exactly (one of) the minimal reducts of the decision table. Further, if we denote by RED_{MDL} the set of MDL reducts, then evidently $RED_{MDL} \subsetneq RED_{super}$.

An algorithmic solution to the problem of finding the MDL reduct for an SDT can be given as a brute force algorithm, which computes the reducts of all the possible instantiations, see Algorithm 1. It is easy to see that the worst case runtime complexity of this algorithm is exponential in the size of the input. Unfortunately, it is unlikely that an asymptotically more efficient algorithm exists. Indeed, if we consider the problem of finding *any* MDL reduct, then the number of instantiations of S is, in the general case, exponential in the number of objects, and for each such instantiation one should find the shortest reduct for the corresponding decision table, which is known to be in Σ_2^P. Interestingly, we can prove that the decisional problem MDL-reduct related to finding MDL-Reducts is also in Σ_2^P. That is, finding an MDL-Reduct is no more complex than finding a minimal reduct in standard decision tables.

Theorem 1. *MDL-Reduct is Σ_2^P-complete.*

Proof. We need to show that there is an algorithm for verifying instances of MDL-Reduct whose runtime is polynomial given access to an oracle for an NP-complete problem. Indeed, a certificate can be given by an instantiation I (whose size is clearly polynomial in the size of the input SDT) together with a reduct R for I, which is an MDL-reduct. Verifying whether R is a minimal reduct for I can then be done in polynomial time with an oracle for NP, hence the result. Further, as finding the minimal reduct for classical decision tables is Σ_2^P-complete (by reduction to the Shortest Implicant problem), MDL-Reduct is also complete.

While heuristics could be applied to speed up the computation of reducts [27] (specifically, to reduce the complexity of the *find-shortest-reducts* step in

Algorithm 1) the approach described in Algorithm 1 still requires enumerating all the possible instantiations. Thus, in the following section we propose two alternative definitions of reduct in order to reduce the computational costs.

4 Methods

In this section, we present the main results concerning the application of rough set and evidence theory towards feature reduction in the superset learning setting.

4.1 Entropy Reducts

We begin with an alternative definition of reduct, based on the notion of entropy [24], which simplifies the complexity of finding a reduct in SDT. Given a decision d, we can associate with it a pair of belief and plausibility functions. Let $v \in V_t$ and $[x]_B$ for $B \subseteq Att$ an equivalence class, then:

$$Bel_S(v|[x]_B) = \frac{|\{x' \in [x]_B : d(x') = \{v\}\}|}{|[x]_B|}$$
$$Pl_S(v|[x]_B) = \frac{|\{x' \in [x]_B : v \in d(x')\}|}{|[x]_B|}$$

For each $W \subseteq V_t$, the corresponding basic belief assignment is defined as

$$m(W|[x]_B) = \frac{|\{x' \in [x]_B : d(x') = W\}|}{|[x]_B|}. \tag{7}$$

Given this setting, we now consider two different entropies. The first one is the pignistic entropy $H_{Bet}(m)$ as defined in (5). As regards the second definition, we will not directly employ the AU measure (see Eq. (4)). This measure, in fact, corresponds to a quantification of the degree of conflict in the bba m, which is not appropriate in our context, as it would imply finding an instantiation which is maximally inconsistent. We thus define a modification of the AU measure that we call *Optimistic Aggregate Uncertainty* (OAU). This measure, which has already been studied in the context of superset decision tree learning [9] and soft clustering [3], is defined as follows:

$$OAU(SDT) = \min_{I \in \mathcal{I}(SDT)} H(p(I)), \tag{8}$$

where $p(I)$ is the probability distribution over the decision attribute induced by the instantiation $I \in \mathcal{I}$.

Let $B \subseteq Att$ be a set of attributes and denote by $IND_B = \{[x]_B\}$ the equivalence classes (granules) with respect to B. Let $d_{[x]_B}$ be the restriction of d on the equivalence class $[x]_B$. The entropy of d, conditional on B, is defined as

$$H_{Bet}(d|B) = \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} H_{Bet}(d_{[x]_B}) = \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} \frac{H(P^m_{Bet}(d_{[x]_B}))}{H(\hat{p}_m(d_{[x]_B}))} \tag{9}$$

$$OAU(d|B) = \sum_{[x]_B \in IND_B} \frac{|[x]_B|}{|U|} OAU(d_{[x]_B}) \tag{10}$$

Definition 5. *We say that $B \subseteq Att$ is*

- *an OAU super-reduct (resp., H_{Bet} super-reduct) if $OAU(d \mid B) \leq OAU(d \mid Att)$ (resp., $H_{Bet}(d \mid B) \leq H_{Bet}(d \mid Att)$);*
- *an OAU reduct (resp., H_{Bet} reduct) if no proper subset of B is also a super-reduct.*

Definition 6. *We say that $B \subseteq Att$ is*

- *an OAU ϵ-approximate super-reduct (resp., H_{Bet} ϵ-approximate super-reduct), with $\epsilon \in [0, 1)$, if $OAU(d \mid B) \leq OAU(d \mid Att) - log_2(1 - \epsilon)$ (resp., $H_{Bet}(d \mid B) \leq H_{Bet}(d \mid Att) - log_2(1 - \epsilon)$);*
- *an OAU ϵ-approximate reduct (resp., H_{Bet} ϵ-approximate reduct) if no proper subset of B is also an ϵ-approximate super-reduct.*

Let $[x]_B$ be one of the granules with respect to an OAU-reduct. Then, the *OAU instantiation* with respect to $[x]_B$ is given by

$$dec_{OAU(B)}([x]_B) = \arg\max_{v \in V_t} \Big\{ p(v) \mid p = \arg\min_{p \in P_{Bel}} H(p) \Big\}, \tag{11}$$

that is, the most probable among the classes under the probability distribution which corresponds to the minimum value of entropy. Similarly, the H_{Bet} *instantiation* with respect to $[x]_B$ is given by

$$dec_{H_{Bet}(B)}([x]_B) = \arg\max_{v \in V_t} Bet_{Bel}(v) \tag{12}$$

The following example shows, for a simple SDT, the OAU reducts, MDL reducts, and H_{Bet} reducts and their relationships.

Example 1. Consider the superset decision table $SDT = \langle U = \{x_1, ..., x_6\}, A = \{w, x, v, z\}, d \rangle$ given in Table 1. We have $OAU(d \mid A) = OAU(d \mid B) = 0$ for $B = \{x, v\}$. Thus, B is an OAU reduct of SDT, as $OAU(d \mid x) = OAU(d \mid v) > 0$. Notice that $\{z\}$ is also an OAU reduct. The OAU instantiation given by $\{x, v\}$ is $dec_{x,v}(\{x_1, x_2\}) = dec_{x,v}(\{x_3, x_4\}) = 0$, $dec_{x,v}(\{x_5, x_6\}) = 1$, while the one given by $\{z\}$ is $dec_z(\{x_1, x_3, x_6\}) = 0$, $dec_z(\{x_2, x_4, x_5\}) = 1$.

On the other hand, $H_{Bet}(d \mid A) = \frac{1}{2}$, while $H_{Bet}(d \mid \{x, v\}) = 0.81$. Therefore, $\{x, v\}$ is not an H_{Bet} reduct. Notice that, in this case, there are no H_{Bet} reducts (excluding A). However, it can easily be seen that $\{x, v\}$ is an H_{Bet} approximate reduct when $\epsilon \geq 0.20$.

The MDL instantiation is $dec_{MDL}(\{x_1, x_3, x_6\}) = 0$, $dec_{MDL}(\{x_2, x_4, x_5\}) = 1$, which corresponds to the MDL reduct $\{z\}$. Thus, in this case, the MDL reduct is equivalent to one of the OAU reducts.

Table 1. An example of superset decision table

	w	x	v	z	d
x_1	0	0	0	0	0
x_2	0	0	0	1	$\{0, 1\}$
x_3	0	1	1	0	0
x_4	0	1	1	1	$\{0, 1\}$
x_5	0	1	0	1	1
x_6	0	1	0	0	$\{0, 1\}$

In Example 1, it is shown that the MDL reduct is one of the OAU reducts. Indeed, we can prove that this holds in general.

Theorem 2. *Let R be an MDL reduct whose MDL instantiation is consistent. Then R is also an OAU reduct.*

Proof. As the instantiation corresponding to R is consistent, $OAU(d \,|\, R) = 0$. Thus R is an OAU reduct.

Concerning the computational complexity of finding the minimal OAU or one OAU, we have the following results.

Proposition 1. Finding the minimal OAU reduct for a consistent SDT is Σ_2^P-complete.

Proof. As any MDL reduct of a consistent SDT is also an OAU reduct and MDL reducts are by definition minimal, the complexity of finding a minimal OAU reduct is equivalent to that of finding MDL reducts, hence is Σ_2^P-complete.

On the other hand, as both OAU [3,9] and H_{Bet} can be computed in polynomial time, the following result holds for finding OAU (resp. H_{Bet}) reducts.

Theorem 3. *Finding an OAU (resp. H_{Bet}) reduct is NP-complete.*

On the other hand, as shown in Example 1, the relationship between MDL reducts (or OAU reducts) and H_{Bet} reducts is more complex as, in general, an OAU reduct is not necessarily a H_{Bet} reduct. In particular, one could be interested in whether an H_{Bet} exists and whether there exists an H_{Bet} reduct which is able to disambiguate objects that are not disambiguated when taking in consideration the full set of attributes Att. The following two results provide a characterization in the binary (i.e., $V_t = \{0, 1\}$), consistent case.

Theorem 4. *Let $B \subseteq Att$ be a set of attributes, $[x_1]_{Att}, [x_2]_{Att}$ be two distinct equivalence classes (i.e., $[x_1]_{Att} \cap [x_2]_{Att} = \emptyset$) that are merged by B (i.e., $[x_1]_B = [x_1]_{Att} \cup [x_2]_{Att}$), that are not inconsistent and such that $|[x_1]_{Att}| = n_1 + m_1$, $|[x_2]_{Att}| = n_2 + m_2$, where the n_1 (resp., n_2) objects are such that $|d(x)| = 1$ and the m_1 (resp., m_2) objects are such that $|d(x)| = 2$. Then $H_{Bet}(d \,|\, B) \geq H_{Bet}(d \,|\, Att)$, with equality holding iff one of the following two holds:*

1. $m_1 = m_2 = 0$ and $n_1, n_2 > 0$;
2. $m_1, m_2 > 0$ and $n_1 \geq 0$, $n_2 = \frac{m_2 n_1}{m_1}$ (and, symmetrically when changing n_1, n_2).

Proof. A sufficient and necessary condition for $H_{Bet}(d \,|\, B) \geq H_{Bet}(d \,|\, Att)$ is:

$$\frac{n_1 + \frac{m_1+m_2}{2} + n_2}{n_1 + m_1 + n_2 + m_2} \geq \max\left\{\frac{n_1 + \frac{m_1}{2}}{n_1 + m_1}, \frac{\frac{m_2}{2} + n_2}{n_2 + m_2}\right\} \tag{13}$$

under the constraints $n_1, n_2, m_1, m_2 \geq 0$, as the satisfaction of this inequality implies that the probability is more peaked on a single alternative. The integer solutions for this inequality provide the statement of the Theorem. Further, one can see that the strict inequality is not achievable.

Corollary 1. *A subset $B \subseteq Att$ is an H_{Bet} reduct iff, whenever it merges a pair of equivalence classes, the conditions expressed in Theorem 4 are satisfied.*

Notably, these two results also provide an answer to the second question, that is, whether an H_{Bet} reduct can disambiguate instances that are not disambiguated when considering the whole attribute set *Att*. Indeed, Theorem 4 provides sufficient conditions for this property and shows that, in the binary case, disambiguation is possible only when at least one of the equivalence classes (w.r.t. *Att*) that are merged w.r.t. the reduct is already disambiguated. On the contrary, in the general n-ary case, disambiguation could happen also in more general situations. This is shown by the following example.

Example 2. Let $SDT = \langle U = \{x_1, ..., x_{10}\}, Att = \{a, b\}, d\rangle$ such that $\forall i \leq 5$, $d(x_i) = \{0, 1\}$ and $\forall i > 5, d(x_i) = \{1, 2\}$. Then, assuming the equivalence classes are $\{x_1, ..., x_5\}, \{x_6, ..., x_{10}\}$, it holds that $H_{Bet}(d \,|\, Att) = 1$.

Suppose further that $\pi_a = \{U\}$. Then $H_{Bet}(d \,|\, a) < 0.95 < H_{Bet}(d \,|\, Att)$ and hence a is a H_{Bet} reduct. Notice that *Att* is not able to disambiguate since

$$dec_{H_{Bet}(Att)}([x_1]_{Att}) = \{0, 1\}$$
$$dec_{H_{Bet}(Att)}([x_6]_{Att}) = \{1, 2\}.$$

On the other hand, $dec_{H_{Bet(a)}}(x_i) = 1$ for all $x_i \in U$. Notice that, in this case, $\{a\}$ would also be an OAU reduct (and hence a MDL reduct, as it is minimal).

A characterization of H_{Bet} reducts in the n-ary case is left as future work.

Finally, we notice that, while the complexity of finding OAU (resp. H_{Bet}) reducts is still NP-complete, even in the approximate case, these definitions are more amenable to optimization through heuristics, as they employ a quantitative measure of quality for each attribute. Indeed, a simple greedy procedure can be implemented, as shown in Algorithm 2, which obviously has polynomial time complexity.

Algorithm 2. An heuristic greedy algorithm for finding approximate entropy reducts of a superset decision table S.

```
procedure HEURISTIC-ENTROPY-REDUCT($S$: superset decision table, $\epsilon$: approximation level, $E \in \{OAU, H_{Bet}\}$)
    $red \leftarrow Att$
    $Ent \leftarrow E(d \,|\, red)$
    $check \leftarrow True$
    while check do
        Find $a \in red$ s.t. $\begin{cases} E(d \,|\, red \setminus \{a\}) \leq E(d \,|\, Att) - log_2(1-\epsilon) \\ E(d \,|\, red \setminus \{a\}) \text{ is minimal} \end{cases}$
        if $a$ exists then
            $red \leftarrow red \setminus \{a\}$
        else
            $check \leftarrow False$
        end if
    end while
    return $red$
end procedure
```

5 Conclusion

In this article we investigated strategies for the simultaneous solution of the feature reduction and disambiguation problems in the superset learning setting through the application of rough set theory and evidence theory. We first defined a generalization of decision tables to this setting and then studied a purely combinatorial definition of reducts inspired by the Minimum Description Length principle, which we called MDL reducts. After studying the computational complexity of finding this type of reducts, which was shown to be NP-hard, harnessing the natural relationship between superset learning and evidence theory, we proposed two alternative definitions of reducts, based on the notion of entropy. We then provided a characterization for both these notions in terms of their relationship with MDL reducts, their existence conditions and their disambiguation power. Finally, after having illustrated the proposed notions by means of examples, we suggested a simple heuristic algorithm for computing approximate entropy reducts under the two proposed definitions.

While this paper provides a first investigation towards the application of RST for feature reduction in the superset learning setting, it leaves several interesting open problems to be investigated in future work:

- In Theorem 2, we proved that (in the consistent case) $RED_{MDL} \subset RED_{OAU}$, that is, every MDL reduct is also an OAU reduct. In particular, the MDL reducts are the minimal OAU reducts. As $RED_{MDL} \subseteq RED_{super}$, the relationship between the OAU reducts and the superset reducts should be investigated in more depth. Specifically we conjecture the following:

Conjecture 1. For each SDT, $RED_{super} = R_{OAU}$.

While the inclusion $RED_{super} \subseteq RED_{OAU}$ is easy to prove in the consistent case, the general case should also be considered.

- In Theorem 4, we provided a characterization of H_{Bet} reducts in the binary consistent case, however, the behavior of this type of reducts should also be investigated in the more general setting, specifically with respect to the relationship between RED_{OAU} and $RED_{H_{Bet}}$.
- Given the practical importance of the superset learning setting, an implementation of the presented ideas and algorithms should be developed, in order to provide a computational framework for the application of the rough set methodology also to these tasks, in particular with respect to the implementation of algorithms (both exact or heuristic) for finding MDL or entropy reducts.

In closing, we would like to highlight an alternative motivation for the superset extension of decision tables in general and the search for reducts of such tables in particular. In this paper, the superset extension was motivated by the assumption of imprecise labeling: The value of the decision attribute is not known precisely but only characterized in terms of a set of possible candidates. Finding a reduct is then supposed to help disambiguate the data, i.e., figuring out the most plausible among the candidates. Instead of this "don't know" interpretation, a superset S can also be given a "don't care" interpretation: In a certain context characterized by x, all decisions in S are sufficiently good, or "satisficing" in the sense of March and Simon [17]. A reduct can then be considered as a maximally simple (least cognitively demanding) yet satisficing decision rule. Thus, in spite of very different interpretations, the theoretical problems that arise are essentially the same as those studied in this paper. Nevertheless, elaborating on the idea of reduction as a means for specifically finding satisficing decision rules from a more practical point of view is another interesting direction for future work.

References

1. Arora, S., Barak, B.: Computational Complexity: A Modern Approach. Cambridge University Press, Cambridge (2009)
2. Bello, R., Falcon, R.: Rough sets in machine learning: a review. In: Wang, G., Skowron, A., Yao, Y., Ślęzak, D., Polkowski, L. (eds.) Thriving Rough Sets. SCI, vol. 708, pp. 87–118. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54966-8_5
3. Campagner, A., Ciucci, D.: Orthopartitions and soft clustering: soft mutual information measures for clustering validation. Knowl.-Based Syst. **180**, 51–61 (2019)
4. Chang, J.C., Amershi, S., Kamar, E.: Revolt: collaborative crowdsourcing for labeling machine learning datasets. In: Proceedings of CHI 2017, pp. 2334–2346 (2017)
5. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. In: Yager, R.R., Liu, L. (eds.) Classic Works of the Dempster-Shafer Theory of Belief Functions, vol. 219, pp. 57–72. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-44792-4_3

6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(Mar), 1157–1182 (2003)
7. Harmanec, D., Klir, G.J.: Measuring total uncertainty in Dempster-Shafer theory: a novel approach. Int. J. Gen. Syst. **22**(4), 405–419 (1994)
8. Hüllermeier, E.: Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int. J. Approximate Reason. **55**(7), 1519–1534 (2014)
9. Hüllermeier, E., Beringer, J.: Learning from ambiguously labeled examples. Intell. Data Anal. **10**(5), 419–439 (2006)
10. Hüllermeier, E., Cheng, W.: Superset learning based on generalized loss minimization. In: Appice, A., Rodrigues, P.P., Santos Costa, V., Gama, J., Jorge, A., Soares, C. (eds.) ECML PKDD 2015. LNCS (LNAI), vol. 9285, pp. 260–275. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23525-7_16
11. Hüllermeier, E., Destercke, S., Couso, I.: Learning from imprecise data: adjustments of optimistic and pessimistic variants. In: Ben Amor, N., Quost, B., Theobald, M. (eds.) SUM 2019. LNCS (LNAI), vol. 11940, pp. 266–279. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35514-2_20
12. Johnson, D., Levesque, S., Zhang, T.: Interactive machine learning system for automated annotation of information in text, 3 February 2005. US Patent App. 10/630,854
13. Jousselme, A.-L., Liu, C., Grenier, D., Bossé, É.: Measuring ambiguity in the evidence theory. IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum. **36**(5), 890–903 (2006)
14. Li, M., Vitányi, P., et al.: An Introduction to Kolmogorov Complexity and Its Applications, 3rd edn. Springer, Heidelberg (2008). https://doi.org/10.1007/978-0-387-49820-1
15. Liu, L., Dietterich, T.: Learnability of the superset label learning problem. In: Proceedings of ICML 2014, pp. 1629–1637 (2014)
16. Liu, L., Dietterich, T.G.: A conditional multinomial mixture model for superset label learning. In: Advances in Neural Information Processing Systems, pp. 548–556 (2012)
17. March, J.G., Simon, H.A.: Organizations. Wiley, New York (1958)
18. Nguyen, N., Caruana, R.: Classification with partial labels. In: Proceedings of the 14th ACM SIGKDD, pp. 551–559 (2008)
19. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**(5), 341–356 (1982)
20. Pearl, J.: Reasoning with belief functions: an analysis of compatibility. Int. J. Approximate Reason. **4**(5–6), 363–389 (1990)
21. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
22. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
23. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Słowiński, R. (ed.) Intelligent Decision Support, vol. 11, pp. 331–362. Springer, Heidelberg (1992). https://doi.org/10.1007/978-94-015-7975-9_21
24. Slezak, D.: Approximate entropy reducts. Fundam. Inform. **53**(3–4), 365–390 (2002)
25. Ślęzak, D., Dutta, S.: Dynamic and discernibility characteristics of different attribute reduction criteria. In: Nguyen, H.S., Ha, Q.-T., Li, T., Przybyła-Kasperek, M. (eds.) IJCRS 2018. LNCS (LNAI), vol. 11103, pp. 628–643. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99368-3_49

26. Smets, P., Kennes, R.: The transferable belief model. Artif. Intell. **66**(2), 191–234 (1994)
27. Thangavel, K., Pethalakshmi, A.: Dimensionality reduction based on rough set theory: a review. Appl. Soft Comput. **9**(1), 1–12 (2009)
28. Umans, C.: On the complexity and inapproximability of shortest implicant problems. In: Wiedermann, J., van Emde Boas, P., Nielsen, M. (eds.) ICALP 1999. LNCS, vol. 1644, pp. 687–696. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48523-6_65
29. Yao, Y.Y., Lingras, P.J.: Interpretations of belief functions in the theory of rough sets. Inf. Sci. **104**(1–2), 81–106 (1998)
30. Zhou, Z.-H.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. **5**(1), 44–53 (2018)