Outlier detection for simple default theories

Fabrizio Angiullia, Rachel Ben-Eliyahu-Zoharyb,∗,1, Luigi Palopoli

a DEIS, Università della Calabria, Via Pietro Bucci 41C, 87036 Rende (CS), Italy
b Communication Systems Engineering Dept., Ben-Gurion University, Beer-Sheva, Israel

1. Introduction

Default logic was originally introduced by Reiter [18] as a tool for working with incomplete knowledge. However, as default rules allow one to describe a normal behavior of a system and draw conclusions, they can also be exploited for detecting outliers, that are facts (also called observations in the following) that are indeed unexpected to hold according to the default theory at hand. This is the basic idea behind the research that has been conducted in the last few years on outlier detection using default reasoning [1] and logic programming [3]. According to this line of research, outliers are sets of observations that demonstrate some properties contrasting with those that can be logically “justified” according to the given knowledge base. Along with outliers, their “witnesses” are singled out — which are sets of observations encoding the unexpected properties associated with outliers. The authors of [3,1] show several application cases for outlier detection in diverse areas. One of the examples they present can be summarized as follows. A well-known center for rare diseases is located in the small city of Lamezia in Calabria, Italy. One hot summer day you are walking along the pleasant streets of Lamezia when you notice a young man wearing a heavy coat going in the same direction. In this situation, if you are a student in a school of medicine interested in genetic diseases, you are curious about his rare illness. Another way to put it is to say that the fact that the man is wearing a coat on a hot summer day makes him an outlier, and one of the probable explanations at that time and place for such behavior is that this man has a rare genetic disease.

Note that, according to the “outlier detection using default reasoning” approach [1], exceptions are not explicitly recorded in the knowledge base as “abnormals,” as is the case for logical-based abduction [17,9,10]. Rather, their “abnormality” is singled out precisely because some of the properties characterizing them cannot be justified within the given theory. Nonetheless, outliers and outlier witnesses are mined from explicitly observed facts, since of the main rationales is to define outlier detection as to embody a data mining technique.

Within the framework of default reasoning, outliers were defined in both the related formalisms of Reiter’s default logic and extended disjunctive logic programming (EDLP). Unfortunately, computing answers to problems (a.k.a., queries) related to outlier detection in default languages is a formidable task. Indeed, it has been shown [1] that even the computationally simplest variant of the outlier detection problem, that is, given a default knowledgebase KB and two sets of literals L and S among those explicitly declared true in KB, is S a witness for the outlier set L in KB? is D2-complete for KB encoding a general theory, and still as difficult as D0-complete for KB restricted to disjunction-free theories. It is therefore consequent to ask whether it is possible to single out some significant fragment of default logics for which outlier detection turns out to be tractable, with natural candidate fragments being the ones for which the entailment problem is polynomial-time solvable.

Such an analysis constitutes the subject of this note and our result provides an essentially negative answer to the question at hand. To fulfill its purposes, the paper investigates outlier detection problems defined over two very simple forms of default theories, namely (dual) normal unary theories, for both of which the entailment problem is solvable in polynomial time. We prove that although these logics are very simple, the associated outlier detection problems remain intractable, the only exception being the simplest problem form, for which tractability can actually be attained. The scenario depicted above suggests that, in fact, outlier detection problems in default logics can be overall looked at as an inherently hard problem.

We proceed as follows. Section 2 recalls the syntax and semantics of Reiter’s Default Logic, the definition of the fragments which are of interest here and the definitions of outlier detection problems in default reasoning. Section 3 presents our complexity results. Section 4 closes the note summarizing contributions and indicating open problems.

2. Preliminaries

2.1. Default logic

We begin by recalling the basic facts about the propositional fragment of default logic. For T a propositional theory and S a set of propositional formulae, T = denotes the logical closure of T and ¬S the set {¬s | s ∈ S}. A set of literals L is inconsistent if ¬t ∈ L for some literal t ∈ L. Given a literal t, letter(t) denotes the letter in the literal t. Given a set of literals L, letter(L) denotes the set {A | A = letter(t), t ∈ L}.

A propositional default theory Δ is a pair (D, W) where W is a set of propositional formulae and D is a set of default rules. We assume that both sets D and W are finite. A default rule δ is α : β1,...,βm/γ, where α (the prerequisite), βi, 1 ≤ i ≤ m (the justifications) and γ (the consequent) are propositional formulae. For δ a default rule, pre(δ), just(δ) and concl(δ) denote the prerequisite, justification, and consequent of δ, respectively. Analogously, given a set of default rules D = {δ1,...,δn}, pre(D), just(D), and concl(D) denote, respectively, the sets {pre(δ1),...pre(δn)}, {just(δ1),...just(δn)}, and {concl(δ1),...concl(δn)}. Whereas, the prerequisite may be missing, the justification and the consequent are required (an empty justification denotes the presence of the identically true literal true specified therein).

A default theory is normal unary (short, NU) (resp., dual normal unary — short, DNU) if W is a set of literals and the set D only contains defaults of the form α : β/β, where α is either empty or a positive literal (resp., either empty or a negative literal) and β is a literal.

The informal meaning of a default rule δ is as follows: if pre(δ) is known to hold, and if it is consistent to assume just(δ), then infer concl(δ). The formal semantics of a default theory Δ is defined in terms of extensions [18]. A set E is an extension for a theory Δ = (D, W) if it satisfies the following set of equations:

- E0 = W,
- for i ≥ 0, Ei+1 = Ei \cup {γ | αβ1,...,βm/γ ∈ D, α ∈ Ei, ¬β1 ∈ E, ..., ¬βm ∈ E},
- E = \bigcup_{i=0}^{∞} Ei.

A default theory may not have any extensions (an example is the theory ((\{p\}, \emptyset))). Then, a default theory is called coherent if it has at least one extension, and incoherent otherwise. Normal default theories are always coherent. A coherent default theory Δ = (D, W) is called inconsistent if it has just one extension which is inconsistent. By Theorem 2.2 of [18], the theory Δ is inconsistent iff W is inconsistent.

The entailment problem for default theories is as follows: Given a default theory Δ and a propositional formula φ, does every extension of Δ contain φ? In the affirmative case, we write Δ |= φ. For a set of propositional formulae S, we analogously write Δ |= S to denote (∀φ ∈ S)(Δ |= φ).
2.2. Outlier detection in default reasoning

The issue of outlier detection in default theories is extensively described in [1]. The formal definition of outlier there proposed is recalled next. For a given set \( W \) and a list of sets \( S_1, \ldots, S_n \), \( W_{S_1, \ldots, S_n} \) denotes the set \( W \setminus (S_1 \cup S_2 \cup \cdots \cup S_n) \).

**Definition 2.1 (Outlier and outlier witness set).** (See [1].) Let \( \Delta = (D, W) \) be a propositional default theory and let \( L \subseteq W \) be a set of literals. If \( (\exists S \subseteq W_L) (S \neq \emptyset) \) such that:

(i) \((D, W_S) \models \neg S\), and

(ii) \((D, W_{S,L}) \not\models \neg S\)

then \( L \) is an outlier set in \( \Delta \) and \( S \) is an outlier witness set for \( L \) in \( \Delta \).

The intuitive explanation of the different roles played by an outlier and its witness is as follows. Condition (i) of Definition 2.1 states that the outlier witness set \( S \) denotes something that does not agree with the knowledge encoded in the defaults. Indeed, by removing \( S \) from the theory at hand, we obtain \( \neg S \). In other words, if \( S \) had not been observed then, according to the given defaults, we would have concluded the exact opposite. Moreover, condition (ii) of Definition 2.1 states that the outlier \( L \) is a set of literals that, when removed from the theory, makes such a disagreement disappear. Indeed, by removing both \( S \) and \( L \) from the theory, \( \neg S \) is no longer obtained. Otherwise stated, that disagreement for \( S \) is a consequence of the presence of \( L \) in the theory. Summarizing, the set \( S \) witnesses that the piece of knowledge denoted by \( L \) behaves, in a sense, exceptionally. This tells that \( L \) is an outlier set and \( S \) is its associated outlier witness set.

The intuition here is better illustrated by referring to the example on rare diseases given in the Introduction. A default theory \( \Delta = (D, W) \) that describes such an episode might be as follows: \( \Delta = (D, W) \), where:

- \( D = \{ \text{WarmDay} \vdash \neg \text{WearCoat} \} \),
- \( W = \{ \text{WarmDay}, \text{WearCoat} \} \).

Here, the person encountered might be suffering from a strange disease, for otherwise he would not be wearing a coat. Accordingly, \( L = \{ \text{WarmDay} \} \) is an outlier set here, and \( S = \{ \text{WearCoat} \} \) is the associated witness set. This reasoning agrees with our intuition that an outlier is abnormal in some sense and that the corresponding witness testifies to it.

To analyze the computational complexity underlying outlier detection, some basic decision problems, also called queries, were defined in [1]. These are recalled next. Let \( \Delta = (D, W) \) be a given default theory:

- **Outlier**: Given \( \Delta \), does there exist at least one outlier set in \( \Delta \)?
- **Outlier[k]**: Let \( k \) be a constant positive integer. Given \( \Delta \), does there exist at least one outlier set with cardinality at most \( k \) in \( \Delta \)?
- **Outlier(L)**: Given \( \Delta \) and a set of literals \( L \subseteq W \), is \( L \) an outlier in \( \Delta \)?
- **Outlier[k](S)**: Let \( k \) be a constant positive integer. Given \( \Delta \) and a set of literals \( S \subseteq W \), is \( S \) a witness for any outlier set \( L \) with cardinality of at most \( k \) in \( \Delta \)?
- **Outlier(S)(L)**: Given \( \Delta \), a set of literals \( S \subseteq W \), and a set of literals \( L \subseteq W \), is \( L \) an outlier with witness \( S \) in \( \Delta \)?

The known complexity results associated with outlier detection in general default theories are summarized in Table 1 [1]. The purpose of this note is to draw the tractability/intractability border associated with outlier detection queries. To this end, we analyze the complexity of outlier detection in simpler fragments of default logics, namely NU and DNU theories.

\[ \text{Table 1} \]

<table>
<thead>
<tr>
<th>Query</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier</td>
<td>( \Sigma^P_2 )-complete</td>
</tr>
<tr>
<td>Outlier[k]</td>
<td>( \Sigma^P_2 )-complete</td>
</tr>
<tr>
<td>Outlier(L)</td>
<td>( \Sigma^P_2 )-complete</td>
</tr>
<tr>
<td>Outlier<a href="S">k</a></td>
<td>( \Sigma^P_2 )-complete</td>
</tr>
<tr>
<td>Outlier(S)(L)</td>
<td>( \Sigma^P_2 )-complete</td>
</tr>
</tbody>
</table>

---

2 Even if the definition is applicable to general default theories, in this work we will limit our attention to NU and DNU propositional default theories only.
3. Complexity results for NU and DNU theories

In this section we present our new results which show that when posed on very simple fragments of default logics, most queries become easier but they still remain intractable. Some definitions are needed first.

Let $T$ be a truth assignment to the set $x_1, \ldots, x_n$ of boolean variables. Then $\text{Lit}(T)$ denotes the set of literals $\{\ell_1, \ldots, \ell_n\}$, such that $\ell_i$ is $x_i$ if $T(x_i) = \text{true}$ and is $\neg x_i$ if $T(x_i) = \text{false}$, for $i = 1, \ldots, n$.

Let $L$ be a consistent set of literals. Then $T_L$ denotes the truth assignment to the set of literals (boolean variables) occurring in $L$ such that, for each positive literal $p \in L$, $T_L(p) = \text{true}$, and for each negative literal $\neg p \in L$, $T_L(p) = \text{false}$.

Next we report our complexity results on outlier detection in NU and DNU default theories. A known proposition is recalled first.

**Proposition 1.** (Proved in [13,5].) Let $\Delta$ be an NU or a DNU propositional default theory and let $L$ be a set of literals. Deciding whether $\Delta \models L$ is $\mathcal{O}(n^2)$, where $n$ is the size of the theory $\Delta$.

We begin by showing that the basic outlier detection query is polynomial time solvable for NU and DNU propositional default theories.

**Theorem 3.1.** Outlier$(S)(L)$ on NU and DNU propositional default theories is in P.

**Proof.** Answering query Outlier$(S)(L)$ amounts to checking if (1) theory $(D, W_S)$ entails the set of literals $\neg S$, and (2) theory $(D, W_S, L)$ does not entail the set of literals $\neg S$. Since the sets $S$ and $L$ are provided in input together with the NU (or DNU) default theory $(D, W)$, by Proposition 1 the aforementioned check can be accomplished in time $O(n^2)$.

We note that the tractability result proved in Theorem 3.1 above merely concerns the simplest of the outlier detection queries, that is, the one where both the outlier and the witness sets are fixed in advance. However, such a tractability result could be exploited to bound the complexity of the more complex outlier detection queries from above. The membership result proved in Theorem 3.4 below, concerning the most general of the outlier detection queries, shows that this is indeed the case. However, the corresponding hardness results provide the evidence that the problem complexity cannot be lowered further than NP.

**Lemma 3.2.** Outlier on NU propositional default theories is in NP.

**Proof.** Given the NU default theory $(D, W)$, in order to answer query Outlier, it is necessary to determine whether there exist two disjoint sets of literals $S$ and $L$, whose elements come from the set $W$, satisfying Definition 2.1. This task can be accomplished by a nondeterministic polynomial time Turing machine that first guesses an outlier set $L \subseteq W$ and an outlier witness set $S \subseteq W$, and then verifies whether $S$ and $L$ satisfy Definition 2.1 or not. The former step can be completed in nondeterministic polynomial time, being that the size of both $S$ and $L$ is upper bounded by the size of $W$ while, by virtue of Theorem 3.1, the latter step can be completed in polynomial time, as it amounts to answering query Outlier$(S)(L)$. Since the problem can be solved by a nondeterministic polynomial time Turing Machine, it is in NP.

**Lemma 3.3.** For each boolean formula $\Phi$ there exists an NU propositional default theory $\Delta(\Phi) = (D(\Phi), W(\Phi))$ and a literal $l \in W(\Phi)$ such that $\Phi$ is satisfiable if and only if $|l|$ is an outlier in $\Delta(\Phi)$.

**Proof.** Let $\Phi = f(X)$ be a boolean formula in CNF, where $X = x_1, \ldots, x_n$ is a set of variables, and $f(X) = C_1 \land \cdots \land C_m$, with $C_j = t_{j,1} \lor \cdots \lor t_{j,u_j}$, and each $t_{j,1}, \ldots, t_{j,u_j}$ is a literal on the set $X$, for $j = 1, \ldots, m$. Checking the CNF formulae satisfiability is NP-complete [15].

In order to prove the NP-hardness of the problem Outlier, the NU default theory $\Delta(\Phi) = (D(\Phi), W(\Phi))$ is associated with $\Phi$, where $W(\Phi)$ is the set

$$\{l, x_1, \ldots, x_n, c_1, \ldots, c_{m+1}\}$$

of letters, with $l, c_1, \ldots, c_{m+1}$ being new letters distinct from those occurring in $\Phi$, and $D(\Phi)$ being the set of defaults $D_1 \cup D_2 \cup D_3 \cup D_4$, with:

$$D_1 = \left\{ \delta_{1,1,i} = \frac{x_i \lor x'_i}{x_i', \delta_{1,2,i}} = \frac{x_i}{x_i}, \delta_{1,3,i} = \frac{\neg x_i}{\neg x_i}, \delta_{1,4,i} = \frac{x'_i}{x'_i} \mid i = 1, \ldots, n \right\}.$$  

$$D_2 = \left\{ \frac{f(t_{j,k}) \lor \neg c_j}{\neg c_j} \mid j = 1, \ldots, m; k = 1, \ldots, u_j \right\}.$$
Proof. The membership part of this theorem is analogous to that of Theorem 3.4, the only difference being that now the nondeterministic polynomial time Turing machine guesses an outlier set \( L \) whose size is not greater than \( k \). The hardness part is given in Lemma 3.3 with \( k = 1 \).
Theorem 3.6. Outlier \((L)\) on NU propositional default theories is NP-complete.

Proof. The membership part of this theorem is analogous to that of Theorem 3.4, the only difference being that now the nondeterministic polynomial time Turing machine has to guess only the outlier witness set \(S\). The hardness part is given in Lemma 3.3 by taking \(L = \{l\}\).

The previous intractability results concerning propositional NU theories are applicable also to propositional DNU theories, as explained next.

Theorem 3.7. Outlier, Outlier\([k]\), and Outlier\((L)\) on DNU propositional default theories are NP-complete.

Proof. The membership results of Theorems 3.4, 3.5, and 3.6 also apply to DNU propositional theories by Theorem 3.1. The hardness part of this theorem can be obtained from the hardness parts of Theorems 3.4, 3.5, and 3.6 by replacing the theories \(\Delta(\Phi)\) there employed with those obtained by substituting each literal \(\ell\) occurring in \(\Delta(\Phi)\) with its negated \(\neg\ell\).

Finally, we consider the case where the outlier witness set \(S\) is fixed in advance and we search for an outlier set \(L\) whose size is not greater than that of a positive constant integer \(k\).

Theorem 3.8. Outlier\([k]\)(S) on NU and DNU default theories is in \(P\).

Proof. Query Outlier\([k]\)(S) can be decided by solving query Outlier(S)(L) for each subset \(L\) of literals in \(W\) having a size of at most \(k\). Since the number of times, that is \(O(|W|^k)\), we call query Outlier(S)(L) is polynomially related to the size of the input, by Theorem 1, this procedure solves Outlier\([k]\)(S) in polynomial time.

All the results above derived are summarized in Table 2 (complexity figures holding for NU and DNU theories can be compared with those of the general case, which are summarized in Table 1).

4. Discussion and conclusions

This note aimed at analyzing the complexity of outlier detection problems in very simple classes of default theories, namely NU and DNU. We showed that, even though for these classes outlier detection is easier, in most cases it remains intractable. These results clearly indicate the inherent intractability of outlier detection problems in default reasoning.

Before concluding, we briefly point out similarities and differences of the approach here pursued with two closely related techniques, that are outlier detection in data and abduction from default logic.

A lot of methods for outlier detection based on statistical or proximity measures have been proposed in the literature [4,12,7]. These approaches can be classified as supervised learning methods, where each example must be labeled as exceptional or not, semi-supervised learning methods, where only examples from the normal class are available, and unsupervised learning methods, where such labels are not required [8,4,14,6,16,2]. In almost all cases, these techniques deal with data organized as a single relational table and often a metrics relating pairs of rows in the table is first required. However, it must be pointed out that little effort has been devoted till now to the design of methods able to exploit domain knowledge in order to improve the process of detecting outliers [3,1]. Indeed, if a description of the domain is available in the form of a knowledgebase encoded in a suitable language for knowledge representation, then more subtle form of anomalies can be singled out with respect to those that can be detected by using statistical-like methods. Outlier detection using default reasoning is one of the first proposals in this context. To further highlight its novelty, we note that the technique here considered is unsupervised. Indeed, it applies to general knowledgebases where the set of potential anomalies is not explicitly identified a priori, as it happens, e.g., with abduction.
In [11], Eiter, Gottlob, and Leone presented a basic model of abduction from default logic and analyzed the complexity of the main associated abductive reasoning tasks. According to [11], a default abduction problem (DAP) is a tuple \( \langle H, M, W, D \rangle \) where \( H \) is a set of ground literals called hypotheses, \( M \) is a set of ground literals called observations, and \( (D, W) \) is a default theory. The goal is to explain observations in \( M \) by using various hypotheses in the context of the default theory \( (D, W) \). The following definition for an explanation is suggested.

**Definition 4.1.** (See [11].) Let \( P = \langle H, M, D, W \rangle \) be a DAP and let \( E \subseteq H \). Then, \( E \) is a skeptical explanation for \( P \) iff

(i) \((D, W \cup E) \models M\), and

(ii) \((D, W \cup E)\) has a consistent extension.

The relationship between outlier detection on normal propositional default theories and skeptical explanations is summarized by the following theorem, taken from [1], which directly applies to DNU and NU theories.

**Theorem 4.2.** (See [1].) Let \( \Delta = (D, W) \) be a normal default theory, where \( W \) is consistent. Let \( L \subseteq W \) and \( S \subseteq W \) be two disjoint sets. Then \( S \) is an outlier witness set for \( L \) in \( \Delta \) if and only if \( L \) is a skeptical explanation for \( \neg S \) in the DAP \( P = \langle L, \neg S, D, W \cup S, L \rangle \).

In sum, it can be said that \( S \) is an outlier witness for \( L \) if \( L \subseteq W \), \( L \) is a skeptical explanation for \( P \), and, hence, \( \neg S \) holds in every extension of the theory.

Despite the duality demonstrated by the above Theorem 4.2, between outlier detection and abduction there is a clear difference. In outlier detection problems the outlier witness set \( S \) (which according to Theorem 4.2 is analogous to the set of observations in abduction problems) has to be guessed, while the set of observations in abduction is part of the input.

**References**